Принципы анализа данных в социологии

Ю. Н. Толстова

(Москва)

Коротко формулируются основные методологические принципы анализа данных — положения, выполнение которых необходимо для обеспечения эффективности использования математического аппарата при решении социологических задач. Принципы сгруппированы в соответствии с основными этапами решения задачи.

Ключевые слова: математические методы, методология, анализ данных, измерение, однородность, комплексное применение методов, интерпретация.

Многолетнее применение математических методов анализа данных[1] в социологии не дает возможность говорить об их «триумфальном шествии». Можно привести

---

[1]Термин «анализ данных» в литературе трактуется по-разному: как процесс изучения статистических данных с помощью математических методов, не предполагающих вероятностной модели интересующего исследователя явления; как синоним термина «прикладная статистика», под которой подразумевается научная дисциплина, разрабатывающая и

не так уж много примеров значительных социологических результатов, полученных благодаря лишь этим методам (не стоит говорить о расчете частотных таблиц, который, вообще говоря, нельзя назвать применением математики). Одна из основных причин подобной ситуации, на наш взгляд, состоит в следующем.

Подавляющее большинство методов, используемых социологом для решения той или иной стоящей перед ним задачи, первоначально было разработано для других нужд. Конечно, любой математический аппарат достаточно «безразличен» к природе исходных данных. С его помощью могут решаться совершенно разные по содержанию задачи. Тем не менее, социологические явления обычно столь сложны и уникальны, что не вполне хорошо описываются моделями, заложенными в известных методах анализа данных. Поэтому метод часто оказывается не адекватным характеру исследуемых социологом процессов. Результат его использования не может рассматриваться как модель реальности. Другими словами, методы анализа данных перестают быть средством изучения социальных явлений.

Чтобы избежать таких ситуаций, на наш взгляд, необходима (но, конечно, не достаточна) разработка определенных методологических принципов применения указанного математического инструментария. Их соблюдение позволит последнему стать тем, чем он и должен быть в действительности — эффективным «орудием труда» исследователя. Социолог обязан не «применять факторный (регрессионный или какой-либо другой) анализ», а решать стоящую перед ним задачу: изучать структуру причинно-следственных отношений между наблюдаемыми переменными, выделять типы интересующих его объектов, находить скрытые «пружины» (латентные переменные), определяющие поведение этих объектов и т.д. Одно из назначений методологических принципов — побуждать социолога столь творчески подходить к анализу данных, умело использовать весь набор методов, чтобы получаемые результаты не уводили его в сторону от действительности.(

систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки статистических данных с целью удобного их представления, интерпретации и получения научных и практических выводов; как такие процедуры «свертки» информации, которые не допускают формального алгоритмического подхода. Мы будем толковать его широко: как обиходное понятие, означающее совокупность действий, осуществляемых исследователем в процессе изучения неких данных с целью формирования определенного представления о характере описываемого ими явления.

Он не должен рассматривать те или иные в окончательные, а обязан не раз возвращаться к информации, делать разные предположения о ее анализировать разными способами, сравнив ветствующие результаты и т.д. Другими социологии и математике необходимо взаимодействие. Исследователь-социолог долж «орудием познания», вернее, его частью (друг математический формализм).

Такой подход к анализу известен (см., напр Однако его реализация требует опр конкретизации. Она может иметь разные уровни относиться к решению вполне конкретной социо задачи; группы задач; содержать принципы ре задач.

Мы остановимся на последнем — методол — уровне и попытаемся определить некотор принципы, которые отвечают последовательны применения методов анализа данных в социологи

1. Измерение

Переходя от эмпирической к математическ объектов, т.е. осуществляя измерение в смы измерений [2], необходимо четко представл эмпирической системы и характер мод отношений между ее элементами. А это не все просто сделать в конкретной возникающ социологом ситуации. Приведем примеры.

Тип шкал, заданный «физическим» получения исходной информации, далеко совпадает с тем, который должен учитываться математического метода анализа данных. несовпадение чаще всего возникает из-за того, чт наблюдая признак, фактически интересуется непосредственно не измеряемыми характе которые наблюдаемые величины отражают. словами, исследователь нередко имеет де называемыми признаками-«приборами», служа индикаторами латентных переменных (подробнее например, в [3]),

Рассмотрим пункт практически социологической анкеты — «возраст респон первый взгляд, это «хороший», количественны Однако при углублении в существо решаемых оказывается, что это не так. В таком случае предположить, что разница, допустим, между 10 является очень большой, а между 60 и 70 годами,

Принципы анализа данных в социологии

весьма малой, возможно, равной нулю. Какими же свойствами действительных чисел мы пользуемся при получении шкальных значений? Какие отношения между этими значениями являются образами соответствующих отношений эмпирической системы? Может быть, до 30-40 лет необходимо учитывать порядок как самих чисел, так и их разностей (т.е. считать числа полученными по интервальной шкале), а, скажем, для возраста после 40 лет — оперировать только первым, не используя никаких других отношений (т.е. полагать соответствующие числа измеренными в порядковой шкале)? А, может быть, начиная с какого-то возраста, вообще все значения следует приравнять друг другу? (Здесь мы лишь обрисовываем проблему и не ставим целью давать какие-то конкретные рекомендации).

Другая характеристика — пол респондента. Казалось бы, с ним все ясно: это номинальный дихотомический признак. Но если учитывать роль этой характеристики в решаемых социологом задачах, то положение может оказаться более сложным. В некоторых ситуациях становится естественной постановка вопроса о поиске количественной шкалы для пола респондента. Предположим, что социолог должен решить проблему подбора кадров на ряд должностей, среди которых встречаются руководитель геологической экспедиции и воспитатель ясельной группы детского сада. Прежде всего возникает мысль — рекомендовать на первую должность мужчину, на вторую — женщину. Однако более глубокие размышления показывают, что в первом случае нужен мужчина не по паспорту, а по личным качествам: умению управлять людьми, непритязательности в быту и т.д. Ясно, что такими качествами вполне может обладать и женщина. Во втором случае, напротив, требуется человек с «женскими» качествами: нежностью, терпением, вниманием к бытовым мелочам и т.д. Конечно, не исключено, что подходящими здесь могут оказаться и некоторые мужчины. Возникает гипотеза, что «социологический» пол (как показатель некоторого интересующего исследователя качества) является непрерывной, интервальной переменной. На одном конце отрезка ее изменения находятся индивиды, обладающие в полной мере «м качествами, на другом — «женскими». Ко большинстве случаев «социологический» муж мужчина по паспорту. Подобное справедлив «социологических» женщин. Однако яс соответствующие понятия не идентичны.

Заметим, что при такой интерпретации ном признаков естественно использовать известные мето рования, позволяющие переходить от этих признако ственным (в частности, так называемую оцифровки значений номинального признака [4, дихотомической номинальной характеристики с целью может быть применен номинальный регр анализ [5].

2. Обеспечение однородности выборки

Для конкретного применения любого матем метода требуется определенная од рассматриваемого множества объектов. О необхо обеспечения много говорится в литературе. Однак предлагаются такие определения однородности, к нашему мнению, не являются достаточно по конструктивными с точки зрения пра использования.

Представляется, что можно выделить (хотя в ной мере условно) два этапа обеспечения од Первый этап «готовит почву» для работ математического метода, второй — конкретного ал первом этапе необходимо создать такую ситуацию, элементы изучаемого множества обладали инте исследователя свойствами (при этом часто сложности, связанные, например, с восприятием ре соответствующего вопроса анкеты и ответами на не них был адекватен один и тот же инструмент измер возникают в значительной мере те же сложност предыдущем случае; серьезная роль отводится гипотезам исследователя о факторах, опр интересующее его явление); была возможна интерпретация результатов измерения (скажем, пр бюджетов времени вряд ли уместна одинаковая ин затрат 4-х часов времени на домашнее хозяйство деревне: для селянина это ни о чем не нормальное явление, для горожанина — больш времени, свидетельствующий либо

особых склонностях респондента, либо о неблагоприятных бытовых условиях). Более подробно о перечисленных аспектах однородности идет речь в [6|.

Второй этап построения однородной совокупности — обеспечение самого существования для рассматриваемого множества объектов закономерности, подлежащей изучению с помощью данного математического метода (например, обеспечение возможности применения той или иной вероятностной модели). Естественно, что способ реализации этого этапа связан со спецификой метода и в существенной мере зависит от априорных соображений исследователя о том, для каких именно подсовокупностей объектов его использование осмысленно. Поясним подробнее.

Неоднородность означает, что идентификация параметров искомой модели (чему служат большинство методов анализа данных) становится бессмысленной для совокупности объектов в целом: для разных объектов изучаемая закономерность имеет различный вид. При этом в социологических задачах встречаются два типа неоднородности. Первый тип возникает за счет экзогенных признаков, т.е. таких, которые не участвуют в формировании модели. Получение однородности в этом случае состоит в априорной (до применения самого метода) классификации объектов с целью выделения таких групп, для которых можно ожидать сходство вида искомой закономерности. Универсальных способов решения этой задачи не существует. Из литературы известны примеры, когда на основе выявленной закономерности, представляющей собой зависимость между двумя переменными, выраженную в виде коэффициента парной связи, делались выводы о наличии сильной корреляции во всей совокупности, хотя на отдельных подмножествах она не наблюдалась, и обратно [7]. Но решение задачи отыскания подсовокупностей, для которых вид искомой связи отличается от вида связи на всей совокупности, очень часто зависит от искусства исследователя[1].

[1]Для отдельных случаев разработаны и формальные методы решения. Так, в распространяемом в настоящее время Всесоюзным центром статистических методов и информатики пакете СРСМ благодаря сочетанию кластерного и регрессионного анализа реализуется возможность эффективного поиска подсовокупностей, однородных в смысле адекватности для них определенных регрессионных моделей.

Второй тип неоднородности возникает за счет эндогенных признаков, т.е. задействованных в используемой априорной модели изучаемого явлении. При этом разные значения входящего в модель признака порождают различные варианты изучаемой закономерности. Для некоторых методов разработаны подходы к улавливанию такой неоднородности. Так, в регрессионном анализе в соответствующих случаях в искомую модель включают неаддитивные члены (произведения рассматриваемых переменных). Заметим, что конструктивность этого подхода (а неконструктивность может возникнуть, поскольку заранее не известно, включение каких именно произведений в уравнение регрессии даст хорошие результаты) может быть обеспечена «превращением» количественной переменной в номинальную с последующим определением вида искомого уравнения на основе специального анализа соответствующих таблиц сопряженности (см., например, |7, 8]).

Проблема неоднородности второго тина необычайно актуальна для социологии. Поэтому особое значение имеет использование в процессе решения социологических задач, которые чаще всего содержат качественные переменные, различных методов поиска детерминирующих сочетаний значений заданных признаков (эти признаки-аргументы детерминируют «поведение» объектов в том смысле, что влияют на параметры распределения некоторого признака-функции). Подобных методов известно довольно мною. Среди них немало разработано советскими авторами |7, Ч, Ш, I I] (о некоторых подходах западных ученых см., например, |4, гл. fi)).

3. Комплексное использование различных методов при решении задачи

Существует два направления этого использования: последовательное (применение различных методов на разных этапах исследования) и параллельное (одновременная их реализация).

Потребность в последовательном подходе может быть вызвана несколькими причинами. Во-первых, социологические исследования обычно многоэтапны из-за сложности изучаемого объекта: на каждом этапе требуется свой метод (при построении выборки, подготовке инструментария, конструировании индексов, выявлении связей признаков и т.д.). Во-вторых, до или после

реализации одного метода часто оказывается желательным использование других: для проверки условий применимости исходного (например, до регрессионного анализа можно использовать известные способы проверки некоррелированности независимых переменных), для формирования первичных данных для него (в частности, при построении признакового пространства при классификации [12]); для более эффективной интерпретации результатов его применения (например, результаты многомерного шкалирования могут быть объяснены с помощью методов классификации [4, гл.. 8]). В-третьих, практика показывает, что путем последовательной реализации разных, но решающих одну и ту же задачу алгоритмов, нередко можно эффективно «нащупать» вид искомой закономерности (так, в [11, гл. 6] изложена методика использования целой серии алгоритмов классификации для определения действительной формы «облаков» точек в признаковом пространстве, т.е. для нахождения реальных типов объектов; в [13, гл.2] предлагается последовательность некоторых способов измерения связей между признаками при построении «истинной» их структуры[1]).

Параллельная реализация разных методов также обусловливается рядом причин: появляется возможность сравнения моделей, используемых каждым методом, с точки зрения их адекватности содержанию решаемой задачи; преодолевается узость, односторонность каждого из них (применение алгоритмов, позволяющих рассмотреть изучаемое явление с разных сторон, обеспечивает более глубокий анализ действительности)[2].

4. Интерпретация результатов применения метода

Этот этап — один из самых ответственных. К сожалению, при его реализации не всегда в полной мере учитывается специфика использованного метода, не достаточно глубоко анализируется его смысл как средства познания реальности. В итоге, при

[1]Среди разных способов оценки связей, в частности, имеются «локальные» и «глобальные» подходы: первые используют коэффициенты связи, отражающие зависимость между отдельными альтернативами рассматриваемых признаков, вторые — между признаками в целом; «глобальные» связи являются как бы усреднениями «локальных».

[2]Идея о необходимости обработки одних и тех же данных разными методами является центральной, например, в (14).

интерпретации исследователь часто «накладывает» на полученные результаты свое априорное видение процессов, которые он намеревается изучать с помощью математики. Его выводы — следствие такого видения, а не итог действительного математического анализа реальности. Чтобы этого не происходило, необходимо соблюдение определенных положений, которые мы назовем принципами интерпретации. Сформулируем наиболее существенные.

Интерпретация должна быть согласована с априорной формальной моделью изучаемого явления. Эту модель следует конкретизировать с учетом положений, сформулированных в пп. 1-3. А именно, она должна включать помимо свойств, обеспечиваемых в процессе измерения, свойства, необходимые для однородности, а также применимости выбранного метода и т.д.Формальная модель — результат перевода содержательных представлений исследователя на формальный язык. И она должна учитываться при обратном переходе от формализма на содержательный уровень, которым является интерпретация.

Естественно, что конкретный смысл изложенного принципа тесно связан с сущностью используемого математического формализма. Так, при построении типологии с помощью методов классификации для интерпретации полученных классов нельзя использовать средние арифметические значения рассматриваемых признаков, если алгоритм классификации (который, по предположению, отвечает пониманию искомых типов) предназначался для поиска групп точек, связанных наличием некоторой статистической зависимости между характеризующими их признаками. В таких случаях средние разных классов могут быть близки; напротив, разброс значений какого-либо признака для какого-то класса может оказаться настолько значительным, что потеряет смысл вычисление соответствующего среднего. По существу, использование средних при интерпретации выводит исследователя за рамки априорной модели, за рамки «перевода» содержания на формальный язык. Результатом такого использования может быть нелепость.

Если бы оказалось, что все содержательные представления социолога о сути решаемой задачи так или иначе формализованы при построении априорной формальной модели, то естественная, отвечающая природе метода интерпретация результатов позволила бы получить то самое более глубокое понимание действительности, которое является главной целью применения

метода. Например, использовав формальный алгоритм классификации, исследователь был бы уверен, что обнаружил искомые типы объектов. Однако, как правило, при решении практически любой социологической задачи содержательные представления невозможно полностью отразить в формальном описании исходных данных, предполагаемой методом модели явления и других сторонах априорной модели. Применяя математику, исследователь чаще всего использует подходы, лишь частично приспособленные для решения интересующей его задачи. Для того, чтобы их применение было эффективным, необходимо, чтобы содержательные соображения, не отраженные формально, нашли выход при интерпретации. Это — другой ее главный принцип. Проблема его реализации — основная для процесса интерпретации. Ее решение в значительной мере зависит от интуиции исследователя: здесь больше искусства, чем науки. В частности, при трактовке результатов классификации этот принцип обычно требует определенной перестройки классификации с целью приближения ее к искомой типологии: пользуясь некоторыми, формально не отраженными соображениями, исследователь корректирует классификацию, объединяя одни и разделяя другие классы и т.д. Пример конкретной задачи можно найти в [13, гл.1]. (Подробнее о принципах интерпретации см. [4, гл.1])

Все сформулированные положения должны между собой сочетаться, отражая взаимосвязь перечисленных этапов применения метода, которое всегда процесс с многочисленными обратными воздействиями.

В заключение отметим, что рассмотренные методологические принципы — результат определенного обобщения практического опыта приложения методов анализа данных в социологии. Конечно, для реализации каждого из них необходимы методические разработки, «привязки» к конкретным классам методов и типам решаемых задач. Но хотелось бы надеяться, что уже сама формулировка этих принципов привлечет внимание исследователей к данной проблематике, послужит стимулом к построению эффективного математического аппарата решения задач социологии.

Литература

1. *ТьюкиДж.* Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981.
2. *ПфанцагльИ.* Теория измерений. М.: Мир, 1976.
3. *КлигерСЛ., Косолапов М.С., ТолстоваЮ.Н.* Шкалирование при сборе и анализе социологических данных. М.: Наука, 1978.
4. Интерпретация и анализ данных в социологических исследованиях. М.: Наука, 1987.
5. *Аргунова К.Д.* Качественный регрессионный анализ в социологии. М.: Ин-т социологии АН СССР, 1990.
6. *Толстова Ю.Н.* Обеспечение однородности исходных данных в процессе применения математических методов //Социол. исслед. 1986. №3.
7. *Миркин Б.Г.* Анализ качественных признаков и структур. М.: Статистика, 1980.
8. *Лакутин О.В., Толстова Ю.Н.* Качественная и количественная информация в социологии: диалектика перехода //Математические методы и модели в социологии. Вып. 1. М.: Ин-т социологии АН СССР, 1991.
9. *Лбов Г.С.* Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981.
10. *Чесноков С.В.* Детерминационный анализ социально-экономических данных. М.: Наука, 1982.
11. *Ростовцев П.С.* Алгоритмы анализа структуры прямоугольных матриц «пятна» и «полосы» //Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985.
12. Типология и классификация в социологических исследованиях. М.: Наука, 1982.
13. Математические методы анализа и интерпретация социологических данных. М.: Наука, 1989.
14. *Орлов А.И.* Устойчивость в социально-экономических моделях. М.: Наука, 1979.