МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

О.Б. Божков, Г.А. Козлов, С.В. Сивашинский
(Санкт-Петербург)

АЛГЕБРАИЧЕСКИЙ МЕТОД СРАВНЕНИЯ ГЕНЕАЛОГИЧЕСКИХ СХЕМ[1]

Статья посвящена решению задачи обобщенного описания и классификации генеалогических деревьев. Предлагается метод, предназначенный для оценки и выделения однородных групп объектов (персонажей), входящих в разветвленные описания семейных структур. Родословные структуры имеют вид прямоугольных матриц с неотрицательными элементами, строки которых являются представлениями отдельных персонажей родословных схем. Предлагаемый метод позволяет для любой группы персонажей (в частности, для всей родословной схемы) оценить степень ее однородности, понимаемой в смысле **структурного сходства** (различия) входящих в нее персонажей. Структурное сходство (различие) понимается как степень коллинеарности (параллельности) векторов-строк указанных матриц.

Ключевые слова: родословная схема, персонаж, однородность, матрица, вектор, кластер.

Олег Борисович Божков – старший научный сотрудник Института социологии РАН (СПб филиал), научный руководитель проекта. **E-mail:** olegbozh@hotmail.com.
Герман Адрианович Козлов – кандидат физико-математических наук, старший научный сотрудник СПб экономико-математического института РАН, скоропостижно скончался 16 января 2001 г., когда статья уже была принята журналом, но еще не прошла стадию научного редактирования.
Семен Вульфович Сивашинский – кандидат физико-математических наук, старший научный сотрудник СПб экономико-математического института РАН, один из участников проекта № 98-06-80308, сотрудник и коллега Г.А. Козлова; любезно согласился довести начатую работу до публикации.

[1] Работа выполнена в рамках проекта № 98-06-80308, поддержанного РФФИ.

Постановка задачи

Социологическое изучение проблем социально-культурных изменений делает актуальным не только переосмысление многих концептуальных, теоретических построений, но заставляет также более внимательно относиться и к методологическим вопросам социального познания. Мы уже отмечали [2, с. 79–80], что процессы социально-культурных изменений, как правило, скрыты от обыденного сознания, которое фиксирует лишь то, что существует «здесь и сейчас», т.е. то, что доступно наблюдению на протяжении жизни одного поколения. А социально-культурные изменения протекают неспешно – на протяжении гораздо более длинных промежутков времени. Именно поэтому социологи все чаще обращаются к биографическим текстам и мемуарам, семейным хроникам, художественной литературе, генеалогическим деревьям как к источникам эмпирической информации. Иными словами – к неформализованным, качественным по своей природе, данным.

В течение последних полутора-двух десятилетий в разных областях социологической науки (например – социология семьи, социальная мобильность, социология медицины и т.д.) проявляется интерес к такому нетрадиционному объекту анализа, как генеалогические деревья (родословные схемы, далее – РС). Лет пять назад, сосредоточив внимание именно на родословных схемах как на объекте социологического анализа, мы попытались понять характер этого объекта и прежде всего – природу и специфику содержащейся в них информации. А привлек нас этот объект по той причине, что охватывает подчас очень большие промежутки времени, до полутора и более веков, в силу чего содержит данные не только о ныне живущих, но и о давно ушедших поколениях людей. Традиционно генеалогии были уделом именитых и родовитых персон, царствующих семей и представителей дворянского сословия. К тому времени, когда мы начали заниматься этим объектом, в социологической литературе уже зашла речь о генеалогиях

простых людей или о «массовой» генеалогии. Наши первые опыты работы в этой проблематике также были связаны с самыми обычными городскими семьями – семьями ленинградских школьников.

На первый взгляд генеалогические деревья кажутся достаточно однообразными: треугольники (мужчины) и кружочки (женщины). Различия между генеалогиями имеют, казалось бы, чисто количественный характер: мало персонажей, много или очень много; «низкое» дерево, содержащее мало генерационных (поколенческих) уровней или, напротив, – «высокое», содержащее много таких уровней. Более пристальное вглядывание в рисунки генеалогий позволяет обнаружить и более тонкие различия и тенденции. В частности, в некоторых РС заметно убывание количества детей (в том числе, сиблингов)[1] по мере приближения к современности (т.е. при движении от верхних уровней РС к нижним). В других РС эта тенденция не столь заметна. Или такое наблюдение: если на одном из верхних поколенческих уровнях обнаруживаются персонажи, имеющие более одного брака, то высока вероятность того, что с определенной регулярностью таких персонажей можно найти и на более низких уровнях. В других РС эта тенденция вовсе не просматривается. Еще одно наблюдение – РС отличаются друг от друга по степени сбалансированности отцовской и материнской ветвей персонажа-автора РС. Чаще (во всяком случае на генеалогиях школьников) встречается ситуация заметного «перевеса» материнской ветви над отцовской. Реже – обратная ситуация. Правда, встречаются генеалогии и относительно сбалансированные по этому признаку.

Естественным следствием этих наблюдений стала задача поиска способов оценки сходства и различия между разными РС.

[1]Сиблинги (англ. siblings) – братья и сестры: дети одной родительской пары вне зависимости от их пола. В отличие от сиблингов различаются также единокровные братья и сестры, имеющие общего отца и разных матерей, и единоутробные братья и сестры, имеющие общую мать и разных отцов.

Или – иными словами – задача классификации генеалогических деревьев и, соответственно, выбора или разработки алгоритмов и процедур для решения этой задачи.

Отбор и описание исходных данных

Стоит чуть более подробно остановиться на специфике РС как объекта социологического анализа. Вне зависимости от того, каким образом к социологу попала та или иная генеалогия (РС), кто обозначен автором той или иной РС, это всегда коллективный труд, результат и характеристика коллективной памяти. В определенном смысле автор РС не может быть квалифицирован как обычный респондент. По меньшей мере он – организатор, а чаще всего – исследователь. Он активизирует собственную память, мобилизует память других. Очень часто ему приходится заниматься самостоятельным поиском необходимой информации и обращаться к самым разным источникам (архивным, литературным и т.п.).

Это принципиальный момент для понимания сущности представленной в РС информации и тех параметров генеалогических деревьев, которые мы будем обсуждать далее. Именно факт коллективной (семейной) памяти объясняет наличие или отсутствие тех или иных данных в РС.

Первый шаг к классификации генеалогий – поиск способа описания объекта или определение параметров, на основании которых классификация может быть построена. Реализацию этого шага мы начали с того, что попытались выделить (найти) такие характеристики любой генеалогии, которые описывали бы ее «естественным образом». Оказалось, что это чисто формальные и легко вычисляемые признаки. В определенной мере мы сознательно отказались от таких признаков, которые можно просто ***приписать*** той или иной РС на основании каких-либо теоретических, логических или каких-то еще оснований и соображений.

Второй существенный момент состоял в том, что пришлось пересмотреть собственные представления о структуре РС как целостного объекта и о параметрах (отвечающих первому критерию вычислимости), с помощью которых можно описывать и, соответственно, оценивать генеалогии. Первоначальное представление о структуре РС [1, с. 117–141] оказалось многоуровневым и неоправданно сложным. В качестве относительно автономных объектов мы выделяли, кроме РС в целом, поколенческие (генерационные) уровни, нуклеарные семейные группы и т.п. Соответственно каждый такой объект наделялся своими собственными параметрами. В частности, для описания РС как целого предлагались такие показатели, как общее количество персонажей, число мужчин среди них, число женщин, количество брачных пар, число поколенческих (генерационных) уровней. Описание поколенческих уровней «тянуло за собой» отдельное сложное описание связей между уровнями и т.д.

Поэтому было принято решение максимально «приблизить» описание к графике РС и в качестве элементарной единицы принять каждую вершинку генеалогического древа, т.е. – отдельного персонажа. Это существенно упростило структуру описания, стало отвечать требованиям критерия вычислимости показателей и органично включило все основные связи, т.е. «ребра» (или «ветки») генеалогического дерева. В одном из первых вариантов система признаков включила в себя 13 показателей:

1. Женщина (женщина 1; мужчина 0).
2. Мужчина (женщина 0; мужчина 1).
3. Число персонажей в поколении персонажа.
4. Число сиблингов.
5. Число браков.
6. Число прямых предков.
7. Число родителей.
8. Число старших поколений на схеме по отношению к поколению персонажа.

9. Число прямых предков персонажа.

10. Число прямых потомков.

11. Число детей.

12. Число младших поколений схемы по отношению к поколению персонажа.

13. Число прямых потомков персонажа.

Впоследствии два параметра – а именно, число родственников и свойственников отца персонажа и число родственников и свойственников матери персонажа, – были добавлены. В то же время некоторые параметры мы исключили из рассмотрения, но сохранили в описании. В конечном итоге, каждая РС была представлена в виде матрицы с числом строк, равным числу включенных в нее персонажей, и с фиксированным числом столбцов, равным числу включенных в рассмотрение параметров.

В качестве эмпирического полигона для разработки и отладки алгоритмов построения классификации из имеющейся в нашем распоряжении коллекции РС были отобраны сначала 6 генеалогий, очевидно различающихся при визуальной их оценке, затем 9, а в конечном итоге – 11. На стадии формирования экспериментального массива участники проекта выступили как эксперты. Задача экспертам ставилась относительно простая: в число немногих РС, составляющих экспериментальный массив, требовалось отобрать три-четыре заметно различающихся родословия, а затем найти такие, которые были бы похожи на уже отобранные. Таким образом, предполагалось сформировать три-четыре разные группы, с тем, чтобы каждая группа внутри себя была достаточно однородной. Сам отбор осуществлялся на интуитивном, сугубо качественном уровне по «внутренним», не обязательно жестко артикулируемым критериям. Однако решающим фактором являлось достижение консенсуса относительно как самих отбираемых генеалогий, так и предлагаемой аргументации в их пользу.

Описание метода классификации РС

Ядром всякого формального метода построения классификаций и, тем более, типологий тех или иных объектов является *не формальное*, а содержательное представление о мере близости между ними, вытекающее из природы самих изучаемых объектов: «Выбор метрики (или меры близости) является узловым моментом исследования, от которого решающим образом зависит окончательный вариант разбиения объектов на классы» [3]. Для таких сложнейших структур, какими оказались РС, понятие о мере близости между ними формировалось постепенно через представление о сходстве персонажей *внутри* отдельной РС. Становилось ясным, что это центральный момент проблемы.

Но как оценить близость отдельных персонажей? Меры близости, обычно понимаемые как выраженные так или иначе *расстояния* между объектами в многомерном пространстве, в данном случае не подходят. Они, конечно, сведут в одну группу (кластер) персонажей с близкими значениями показателей, но персонажи, у которых, *с точки зрения социолога,* есть внутреннее сходство (такое, например, как уравновешенность социальных отношений), но значения показателей отличаются значительно, попадут в разные кластеры.

Поэтому мы стали искать, во-первых, формальное выражение сходства далеких с пространственной точки зрения персонажей и, во-вторых, математический измеритель такого сходства (различия). Так мы пришли к идее коллинеарности (параллельности) как выражения сходства между персонажами.

Перейдем к описанию математического метода оценки этого качества.

Он основан на таких базовых понятиях матричной алгебры, как *вектор-строка, вектор-столбец, базис* векторного *пространства, прямоугольная матрица*, *сопряженная* матрица, *произведение* матриц, *собственное число* и *собственный вектор* матрицы, *спектр* матрицы.

Пусть A – исходная прямоугольная матрица персонажей родословной схемы, в которой строк-персонажей n, а столбцов-показателей, их описывающих, – m, и пусть A' – транспонированный экземпляр матрицы A, т.е. A' – матрица, в которой те же персонажи не строки, а столбцы. По правилам матричного умножения может быть построена матрица, являющаяся произведением A' на A, т.е. матрица $K = A' \cdot A$. Согласно указанным правилам, ее размерность – $m \times m$ (т.е. K – квадратная матрица размерности m).

Собственным вектором-строкой матрицы K называется m-элементная строка b, удовлетворяющая равенству $bK = \lambda b$, где bK – произведение вектора на матрицу, выполняемое по тем же правилам, а множитель λ называется собственным числом матрицы K, соответствующим собственному вектору b. Геометрически собственные векторы матрицы – это такие векторы, которые не меняют своих направлений при воздействии на них данной матрицей-оператором. Спектр матрицы – набор всех ее собственных чисел.

Имеют место следующие фундаментальные алгебраические факты, касающиеся матриц вида $K = A' \cdot A$ [4]:

1. Матрица K имеет ровно m собственных векторов-строк и столько же векторов столбцов. Они образуют так называемый базис m-мерного векторного пространства. Это значит, что любой m-мерный вектор может быть (однозначно) выражен в виде линейной комбинации m векторов базиса.

2. Все собственные числа матрицы K неотрицательны, и существует по крайней мере одно положительное собственное число.

3. Если матрица K имеет нулевые собственные числа, то для выражения в виде линейных комбинаций строк-персонажей, составляющих исходную матрицу A, достаточно не m, а меньшего количества базисных векторов, именно – m-l, где l – число нулевых собственных чисел.

4. Наибольшему собственному числу соответствует строго положительный собственный вектор (все его компоненты поло-

жительны – теорема Фробениуса [5]). Собственный вектор с такими свойствами – единственный.

Из этих фактов вытекают важные для нас следствия. Пусть $\lambda_1, \lambda_2, ..., \lambda_m$ – спектр матрицы К, т.е. набор всех ее собственных чисел, среди которых могут быть и равные. Рассмотрим отношение $c = \lambda_{\max} / \sum_{i=1}^{m} \lambda_i$, которое будем называть *коэффициентом общности* (или однородности) *группы* персонажей. Ясно, что *c* не превосходит 1.

Заметим, что указанная в п. 4 уникальность собственного вектора, соответствующего максимальному собственному числу матрицы *K*, позволяет считать его существенным интегральным представителем пространства персонажей соответствующей РС.

Покажем, что чем однороднее группа персонажей, составляющая матрицу *A*, тем коэффициент *c* ближе к его максимально возможному значению 1. Расположим собственные числа в порядке убывания и возьмем *крайний* случай: $\lambda_1 > 0$, $\lambda_2 = 0$, ..., $\lambda_m = 0$, т.е. в этом случае $c = 1$. Это означает, что все строки матрицы персонажей *A* зависимы, в точном смысле слова – коллинеарны (параллельны), т.е. матрица *A* составлена из строк, отличающихся только множителями. Это случай наибольшей однородности группы. Такая ситуация практически мало вероятна, но из нее следует ценный вывод: даже в случае *несомненной* однородности группы классический подход в терминах *расстояний* разбросает таких персонажей по *разным* кластерам. Но оценка однородности с помощью введенного выше коэффициента *c* приведет к адекватному ситуации отнесению *всей* этой группы к одному кластеру.

Противоположный *крайний* случай заключается в ситуации, при которой векторы персонажей больше всего «топорщатся» – это случай, описываемый матрицей вида

$$A = k \left\| \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \multicolumn{4}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ 0 & 0 & 0 & 1 \\ \multicolumn{4}{c}{\text{и так далее, строки из 0 и 1}} \\ \multicolumn{4}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ \multicolumn{4}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \end{array} \right\|, \; k \text{ – любое число.}$$

В этом, тоже фантастическом, но в другую сторону, случае коэффициент c будет равен $1/m$ – наименьшее возможное его значение. Все практические случаи оценки однородности родословной схемы посредством коэффициента общности c будут находиться в выявленном нами диапазоне: $1/m < c < 1$. В реальных РС, конечно, строки-персонажи, составляющие их матрицы, не будут ни коллинеарны в точном смысле этого понятия, ни совершенно несходны друг с другом, как в последнем примере. Ситуации будут расплывчатыми: найдутся группы, в которых однородность в изложенном выше смысле высока, т.е. коэффициент общности, сосчитанный по ним, достаточно велик. По всей же РС он будет мал, поскольку группы с высоким значением коэффициента общности *между собой* будут отличаться значительно.

Последние рассуждения имеют принципиальное значение для окончательной формулировки метода и алгоритма, его реализующего. По существу, представляемые ниже процедуры реализуют метод кластеризации, суть которого в новом подходе к определению топологии множества объектов, которая сближает объекты, далекие в классическом смысле расстояния.

Построение единой шкалы для сравнения множества РС, т.е. описание всего их разнообразия при помощи единственного параметра, не может, в силу сказанного выше, привести к вырази-

тельным, ясно интерпретируемым результатам. Поэтому был признан более целесообразным двухэтапный подход к решению поставленной задачи.

На первом этапе анализируются по отдельности все РС, включенные в обработку, т.е. каждая РС подвергается процедуре кластеризации. Ее алгоритмическая организация традиционна: на первом шаге кластеров столько, сколько персонажей – каждый персонаж составляет кластер. На втором шаге подбирается такая пара кластеров (пока это еще отдельные персонажи), для которой расстояние между ними равно максимуму по определенному выше коэффициенту общности *c* (напомним – чем он больше, тем однороднее кластер в описанном нами смысле). И так далее, до того момента, когда останутся два кластера, объединяющиеся в последний, уже тривиальный, кластер, включающий все персонажи, т.е. совпадающий со всей РС. Результаты каждого шага запоминаются в описании дерева кластеров, так что можно с помощью дополнительной процедуры наглядно представить получившиеся в результате разбиения на 2, 3 и более кластеров.

Ограничивая рассмотрение результатов проведенного анализа фиксированным числом кластеров (скажем, пятью), выполняем содержательное описание (интерпретацию) кластеров по каждой из рассмотренных РС, затем составляем единый список названий (меток) кластеров по всем РС. Назовем этот список базовым для данной совокупности РС. Теоретически могло случиться, что он будет большим – разные РС породят свои наборы кластеров, содержательно мало пересекающиеся с описаниями кластеров из других РС. В случае такого разнообразия мы попали бы в теоретический тупик. Практически же оказалось, что это разнообразие невелико и базовый список вполне обозрим.

Последнее обстоятельство позволяет выполнить второй этап анализа – произвести *неформальную* группировку родословных схем в соответствии с количественным составом содержащихся в них базовых кластеров. В результате каждая РС получит свою содержательную метку-название.

Обсуждение результатов: сравнение с классическими методами

Изложенный выше двухступенчатый метод анализа совокупности РС складывался, как это вообще бывает в большой работе, не в той логике, в которой он окончательно представлен. Работа началась с попыток расклассифицировать всю совокупность РС, расположив их поначалу по различным шкалам «внутриклассового разброса наблюдений» (такие, как внутриклассовая дисперсия или обобщенная внутриклассовая дисперсия, внутриклассовое расстояние между элементами и др. [3]), а затем и по предложенной нами шкале коэффициента общности *c*. Ибо ни одна из этих попыток (теперь можем сознаться – отчаянные попытки расклассифицировать большие матрицы разной размерности!) не дала и, как выше проанализировано, не могла дать удовлетворительных ясных группировок: одномерная шкала, как бы изощренно она ни была построена, – слишком бедная модель для анализа столь сложных структур, как РС.

Тогда мы обратились к многомерным представителям родословных схем, полагая, что увеличение количества шкал даст эффект. В качестве таких многомерных представителей нами рассматривались те или иные усреднения множества персонажей, входящих в родословную схему. Естественно, что поначалу мы строили вектор, состоящий из среднеарифметических значений показателей; затем перешли к их среднеквадратичным значениям, затем к средним значениям, рассчитанным по другим нормировкам. Появилась надежда, что если подвергнуть множество так или иначе построенных представителей от каждой РС какому-либо виду кластерного анализа, то мы придем к желаемому итогу. Но этот подход тоже не привел к содержательной трактовке различий родословных схем. Построенные кластеры из указанных представителей РС не позволили рационально сопоставить их с соответствующими сложными структурами РС. Не по-

могло и обращение к тем векторам – интегральным представителям РС – элементы которых рассчитывались вместе с коэффициентом однородности *c*.

Объяснение этих неудач на самом деле уже содержится в описании окончательно принятого метода: в громадных по числу персонажей схемах в принципе невозможно обнаружить их однородность, ее просто нет. Адекватный подход заключается в выявлении относительно однородных групп персонажей внутри отдельных РС.

Заметим еще, что мучительные попытки объять необъятное, т.е. описать сложные структуры РС посредством отдельных шкал, усугублялись проблемой масштабирования. Дело в том, что элементы строк-персонажей выражают в некотором масштабе специфические данные. При изменении масштабов численные значения этих элементов и характеристики, описанные выше, будут изменяться. Тем самым, все они являются в известной мере случайными, зависящими от выбора масштаба показателей. Это относится и к нашему коэффициенту однородности *c*.

Решение проблемы масштабирования

При эмпирических попытках нащупать подходящее масштабирование мы использовали следующие способы нормировки отдельного показателя: деление его значений на их евклидову норму, стандартизация показателя (приведение его значений к нулевому среднему и единичной дисперсии), деление значений на их сумму, деление на максимальное значение. Не придя к убедительному варианту, но сохраняя упорное желание расклассифицировать родословные схемы по одномерной шкале, мы углубились в проблему масштабирования теоретически.

Как избежать зависимости численных значений коэффициента *c* от масштабирования? Можно выбрать самые «плохие» масштабы в данной РС, при которых коэффициент *c* минимален, и считать истинным именно это значение. Была поставлена и решена задача минимизации коэффициента *c*, рассматриваемого как функция *m*

масштабов, т.е. множителей, на которые умножаются соответствующие столбцы матрицы *A*. Задача минимизации решалась в оболочке системы MATLAB, работающей под управлением Windows, путем обращения к процедуре fmins, в которой реализован метод Нелдера-Мида (Nelder-Mead simplex algorithm [6, p. 116–122]), – эффективный метод минимизации функции многих переменных. Наличие указанной процедуры чрезвычайно облегчило решение этой большой и трудной в вычислительном отношении задачи.

Однако ни шкала коэффициента подобия *c*, построенная по оптимальному (в указанном смысле) масштабированию, ни соответствующие векторы-представители родословных схем не дали эффекта ясно интерпретируемой классификации принятых к рассмотрению РС. Тем не менее, работа по поиску хорошего масштабирования не пропала даром: выяснилось, что разброс значений коэффициента подобия при различных способах масштабирования тем меньше, чем меньше персонажей содержит родословная схема. Это обстоятельство объясняет достаточную устойчивость результатов, полученных по предложенному нами методу, поскольку внутри него коэффициент однородности рассчитывается по относительно малым группам персонажей.

Возвращаясь к выбранной для исследований коллекции 11 РС, мы решили начать с более детального, по восьми параметрам, описания каждой генеалогии, вошедшей в экспериментальный массив. Пример такого описания представлен в следующей таблице (в данном примере при масштабировании каждый столбец поделен на его евклидову норму).

Для интерпретации кластеров пришлось использовать некоторые дополнительные процедуры. На основании таблиц, аналогичных приведенной выше, мы построили формальное описание кластеров, для чего рассчитали средние значения, минимумы, максимумы и разности между ними по всем параметрам и персонажам, вошедшим в один кластер. На рис. 1 приведен пример графического представления одного из таких кластеров, полученных по одной из РС.

Таблица 1

ФРАГМЕНТ ТАБЛИЧНОГО ПРЕДСТАВЛЕНИЯ РЕЗУЛЬТАТОВ КЛАСТЕРНОГО АНАЛИЗА ГЕНЕАЛОГИЧЕСКОГО ДЕРЕВА[1]

Уникальное имя персонажа	Число сиблингов	Число предков	Число поколений предков	Число потомков	Число детей	Число поколений потомков	Число родствен-ников отца	Число родствен-ников матери
0	0,00000	0,06637	0,03268	0,00000	0,00000	0,00000	0,04722	0,01901
6	0,00000	0,04867	0,03268	0,00000	0,00000	0,00000	0,06727	0,00000
49	0,00000	0,03540	0,03268	0,00000	0,00000	0,00000	0,06921	0,00000
22	0,00000	0,00885	0,01307	0,00000	0,00000	0,00000	0,06921	0,00000
89	0,00000	0,01770	0,02614	0,00000	0,00000	0,00000	0,06921	0,00000
90	0,00000	0,01770	0,02614	0,00000	0,00000	0,00000	0,06921	0,00000
7	0,00000	0,05310	0,03268	0,00000	0,00000	0,00000	0,06792	0,00115
12	0,00000	0,05310	0,03268	0,00000	0,00000	0,00000	0,06792	0,00115
51	0,00000	0,03540	0,03268	0,00000	0,00000	0,00000	0,06792	0,00115
73	0,00000	0,02212	0,02614	0,00000	0,00000	0,00000	0,06792	0,00115
36	0,01471	0,01327	0,01961	0,00000	0,00000	0,00000	0,06662	0,00000
38	0,01471	0,01327	0,01961	0,00221	0,00833	0,00617	0,06662	0,00000
85	0,01471	0,01327	0,01961	0,00221	0,00833	0,00617	0,06598	0,00000
87	0,01471	0,01327	0,01961	0,00221	0,00833	0,00617	0,06598	0,00000
8	0,01471	0,02655	0,01961	0,01327	0,01667	0,01235	0,00712	0,00518
43	0,01471	0,02655	0,01961	0,01327	0,01667	0,01235	0,00712	0,00518
2	0,01471	0,01770	0,01961	0,00221	0,00833	0,00617	0,00776	0,00806
70	0,01471	0,01770	0,01961	0,00221	0,00833	0,00617	0,00776	0,00806
25	0,01471	0,00442	0,00654	0,00000	0,00000	0,00000	0,00000	0,00000
94	0,07353	0,00442	0,00654	0,00000	0,00000	0,00000	0,00000	0,00000
99	0,07353	0,00442	0,00654	0,00000	0,00000	0,00000	0,00000	0,00000
101	0,07353	0,00442	0,00654	0,00000	0,00000	0,00000	0,00000	0,00000
102	0,07353	0,00442	0,00654	0,00000	0,00000	0,00000	0,00000	0,00000
61	0,04412	0,00885	0,01307	0,00000	0,00000	0,00000	0,00000	0,00115
63	0,04412	0,00885	0,01307	0,00000	0,00000	0,00000	0,00000	0,00115
103	0,01471	0,01327	0,01307	0,00221	0,00833	0,00617	0,00129	0,00115
62	0,04412	0,00885	0,01307	0,00221	0,00833	0,00617	0,00000	0,00115
91	0,07353	0,00442	0,00654	0,00664	0,00833	0,01235	0,00000	0,00000
69	0,07353	0,00442	0,00654	0,00885	0,01667	0,01235	0,00000	0,00000

[1] Кластеры отделяются один от другого пустой строкой. В таблице представлены для примера только два из пяти полученных нами кластеров. Каждая строка (запись) состоит из принадлежащих конкретному индивиду нормированных значений включенных в описание показателей.

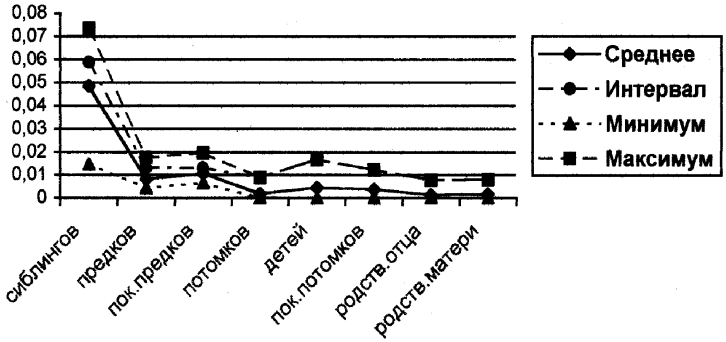

Рис. 1. Графическое представление одного из кластеров РС

Графики существенно облегчили интерпретацию. Поскольку мы остановились на выделении пяти кластеров по каждой РС, то соответственно получили пять таких графиков. А затем свели их в одну, выбрав для графического представления каждого кластера средние значения (см. рис. 2).

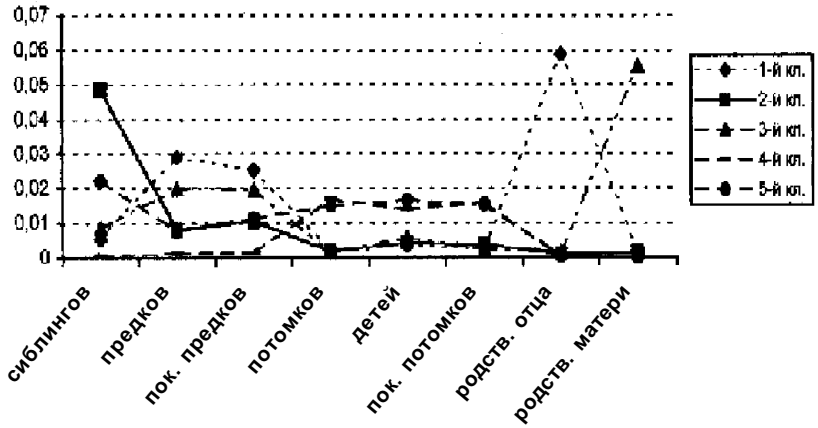

Рис. 2. Представление кластеров, полученных по одной из РС (средние значения показателей)

Видимый на графике характер колебаний значений параметров дает основание для следующей (см. табл. 2) интерпретации полученных кластеров (в данном случае – групп персонажей).

Подчеркнем, что рассматривая по отдельности каждую из 11 РС и задавая определение (вычисление) пяти кластеров (классов), в результате мы получили ***всего*** 9 кластеров, содержательно отличающихся друг от друга. Это дает основу для классификации и самих одиннадцати РС посредством количественной оценки представленности в каждой из них 9 характерных (базовых) кластеров.

Таблица 2

ИНТЕРПРЕТАЦИЯ КЛАСТЕРОВ, ПРЕДСТАВЛЕННЫХ НА РИС. 2

Номер кластера[1]	Смысл (интерпретация) кластера
1	Дети младшего и среднего возраста (самые молодые генерации) с корнями по материнской линии
2	Представители средних генераций (поколений) с родителями и детьми
3	Сиблинги средних генераций с родителями и детьми
4	Представители средних генераций (без родителей)
5	Дети младшего и среднего возраста (самые молодые генерации) с корнями по отцовской линии

На каждой из 11 генеалогий для наглядности и контроля интерпретации были выделены персонажи, образующие каждый из кластеров. В качестве примера приведем одно из реальных генеалогических деревьев, в котором различной штриховкой выделены персонажи, вошедшие в разные кластеры (см. рис. 3).

Заметим, что некоторые персонажи вообще не имеют штриховки, так как по избранным для описания и анализа параметрам они

[1]В данном случае номера кластеров даны на примере рис. 2.

имеют нулевые значения. Таким образом, они образуют еще один (можно сказать, естественный – в противовес вычисленным) кластер, который может быть назван: «супруг(а) без родителей и детей».

Для решения той же задачи (и для контроля) был использован один из традиционных методов оценки взаимосвязи между признаками – факторный анализ. Приведем для примера результат, полученный по двум генеалогическим деревьям (см. табл. 3 и 4).

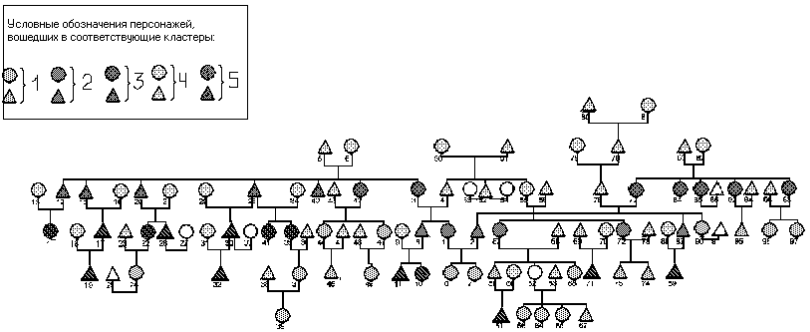

Рис. 3. Схема генеалогического дерева № 1

Таблица 3

РЕЗУЛЬТАТЫ ФАКТОРНОГО АНАЛИЗА ИСХОДНЫХ ДАННЫХ ПЕРВОГО ГЕНЕАЛОГИЧЕСКОГО ДЕРЕВА

Исходные признаки	Факторы по генеалогическому дереву № 1		
	Старшие поколения	Младшие поколения с развитой материнской линией	Сиблинги без отцовской линии
Число сиблингов		,174	**,555**
Число прямых предков	–,230	**,930**	
Число поколений предков	–,261	**,951**	
Число потомков	**,948**	–,131	
Число детей	**,911**	–,168	
Число поколений потомков	**,862**	–,329	
Число родственников отца	–,187	,315	–,861
Число родственников матери	–,160	**,710**	***,466***

Таблица 4

РЕЗУЛЬТАТЫ ФАКТОРНОГО АНАЛИЗА ИСХОДНЫХ ДАННЫХ ВТОРОГО ГЕНЕАЛОГИЧЕСКОГО ДЕРЕВА

Исходные признаки	Факторы по генеалогическому дереву № 2		
	Младшие поколения с развитой отцовской линией	Старшие поколения	Развитая материнская линия
Число сиблингов		–,329	
Число прямых предков	**,910**		*,239*
Число поколений предков	**,909**	–,169	*,301*
Число потомков	–,199	**,906**	
Число детей	–,183	**,768**	
Число поколений потомков	–,289	**,892**	–,131
Число родственников отца	**,837**	–,146	–,416
Число родственников матери	,163	–,176	**,928**

Метод: главных компонент
Метод вращения факторов: варимакс с нормализацией по **Kaiser**[1]**.**

Видим, что интерпретация латентных переменных (факторов) по каждой из РС совпадает лишь частично. Это естественно, так как различие между конкретными родословными схемами проявляется именно здесь. Однако использование кластерного анализа по предлагаемому методу в конечном итоге привело к полному описанию одиннадцати РС с помощью всего 9 различных кластеров. При обращении же к методам факторного анализа полное описание той же совокупности РС было достигнуто с помощью 13 факторов.

Значит ли это, что факторный анализ позволяет уловить более тонкие различия? Думаем, что нет. Дело скорее всего в том, что для факторного анализа (процедуры-алгоритмы статистического пакета SPSS 10.0) использовались «сырые» (не нормированные) исходные признаки. Кроме того, факторный анализ оценива-

[1]В обоих случаях вращение было реализовано в 5 итераций.

ет взаимосвязи между признаками, а не между объектами. И хотя средства SPSS позволяют вычислить индивидуальные значения факторов для каждого респондента, сама по себе эта процедура довольно громоздка, менее прозрачна и не поддается надежному контролю со стороны исследователя.

Наконец, предлагаемый нами метод – сравнение групп персонажей с помощью коэффициента общности *c*, дает возможность оценить вес (или вклад) каждого из исходных параметров в итоговую оценку. И здесь результаты, полученные с помощью разных методов, практически не противоречат друг другу.

Резюме

Специфика такого объекта социологического изучения, какими являются генеалогические деревья, подвинула нас к поиску адекватных способов и методов их анализа и оценки. И хотя эта работа еще не завершена, мы сочли возможным и даже необходимым вынести на суд коллег некоторые результаты наших поисков.

Сама по себе задача оценки сходства (различия) объектов традиционна и типична для социологического исследования. Однако далеко не всегда исследователь оказывается столь серьезно озабочен проблемой релевантности объекта и метода и, тем более, разработкой каких-либо новых подходов. Опыт сотрудничества социолога и профессиональных математиков в таком непростом деле показался нам и плодотворным, и обоюдно интересным. При этом ни математики, ни социолог не собирались переквалифицироваться. Прежде всего, мы обоюдно старались найти доходчивый и ясный язык описания своих проблем и способов, методов их решения.

Может быть, не всегда нам удавалось четко и достаточно убедительно аргументировать собственную неудовлетворенность от использования традиционных для социологии методов анализа. Но мы убеждены, что при выборе методов анализа природа объекта обязательно должна учитываться, и мы постарались это сделать.

ЛИТЕРАТУРА

1. *Божков О.Б.* Генеалогии и биографии как объект социологического анализа // Социологический журнал. 1999. № 3/4.

2. *Божков О.Б.* Биографии и генеалогии: ретроспективы социально-культурных трансформаций // Социологический журнал. 2001. № 1.

3. Прикладная статистика / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; Под ред. С.А. Айвазяна. М.: Финансы и статистика, 1989.

4. *Смирнов В.И.* Курс высшей математики. М.: Наука, 1974. Т. III. Ч. I.

5. *Гантмахер Ф.Р.* Теория матриц. М.: Физматгиз, 1966.

6. *Denis J.E. Jr., Woods D.J.* New Computing Environments: Microcomputers in Large-Scale Computing // SIAM. 1987.