ОПЫТ ПРИМЕНЕНИЯ КЛАСТЕРНОГО АНАЛИЗА

М.Ф.Черныш

(Москва)

Опытным путем выявляются границы применения кластерного анализа в прикладных эмпирических исследованиях. Установлено, что кластерный анализ – это метод, не вполне пригодный для операции по классификации больших совокупностей. Отмечается, что процедуры кластерного анализа, применяемые в стандартных статистических программах, слишком чувствительны в отношении пропущенных данных, а также имеют серьезные ограничения в том, что касается количества наблюдений и переменных, включаемых в них.

Ключевые слова: идеологические установки, кластерный анализ, метод "ближайшего соседа", плотность кластера, метод "дальнего соседа", метод "увязки средних величин", метод Уорда, K-Means-кластерный анализ.

Проблемное поле

В российском обществе усилиями средств массовой информации распространилась и закрепилась мысль о том, что все граждане страны делятся на сплоченные группы сторонников и противников реформ. В число первых включаются, как правило, люди либеральных настроений, под которыми понимаются граждане, выступающие за развитие рыночных настроений, демократических институтов, а в число вторых – прежде всего те, кто противится движению страны вперед к рынку, мечтает о возврате к старым коммунистическим временам, враждебен по отношению к Западу. В последнее время эта мысль, почти целые десять лет побуждавшая российских граждан поддерживать власть

имущих, все чаще оказывается в центре идеологических противоречий. Выясняется, что сторонники рынка далеко не едины в своих оценках Запада, а адепты авторитарной власти отнюдь не единодушны в таких важных идеологических вопросах, как собственность на землю и итоги приватизации. В связи с этим возникает вопрос о том, насколько реальны группы, определяемые какой-либо одной идеологический установкой, в какой степени граждане России консолидированы внутри тех, кто поддерживает или отвергает реформы. У этого вопроса есть важная методологическая подоплека. С одной стороны, ученые, подобные Конверсу, [1] утверждали и продолжают утверждать, что общественные оценки разных явлений могут разниться в весьма значительных пределах и что каждая политическая оценка существует самостоятельно, независимо от других. С другой стороны, ученые марксистской ориентации [2] (Райт) настаивали и настаивают на правильности прямо противоположного подхода, находя тесные взаимосвязи между самыми разными идеологическими установками.

Существует и третий подход, согласно которому убеждения людей структурированы не столько по отдельным оценкам, сколько по целым направлениям или, иными словами, по жизненным мирам, создающим оценочные сферы, в пределах которых установки обнаруживают тесную взаимосвязь. Так, например, люди, одобряющие строительство рыночной экономики, будут активно поддерживать любые шаги, служащие её укреплению, а именно, приватизацию промышленных предприятий, антимоно-польное законодательство, действия, направленные на защиту частной собственности и другое. Это, однако, не означает, что те же самые люди будут столь же активно отстаивать идею строительства институтов демократического государства. Связь между этими аспектами процесса реформ может быть очевидной обществоведу, но отнюдь не всегда столь же очевидной рядовому обывателю, для которого сильное, а, может быть, и жестокое

государство может казаться единственной гарантией сохранения действительно рыночных отношений. Эмпирические исследования могут пролить свет на реальную ситуацию в обществе, выявив критерии выделения социальных групп и уровень взаимосвязи между разными установками.

Стратегия анализа

Данные настоящего исследования были получены исследовательским проектом "Социальная мобильность в переходный период" в начале 1999 года. Исследование носило общероссийский характер и базировалось на выборке, равной 2600 респондентам. В числе прочих вопросов респондентам предлагалось оценить важность целого ряда направлений развития российского общества, используя для этого пятибалльную шкалу – от "очень важно" до "совсем не важно". Полученные данные были подвергнуты процедуре кластерного анализа. Отметим, что далеко не все респонденты были включены в кластеры: если респондент не давал содержательного ответа хотя бы на один вопрос, он автоматически исключался из выборочной совокупности. В результате из выборки было изъято 1064 наблюдения.

Кластерный анализ – гибкая процедура, позволяющая варьировать способы переработки данных в весьма широких пределах. Именно на этом этапе происходит очевидное вмешательство исследователя в характер процедуры, привносящее субъективное начало в получаемый результат. Начнем с того, что методы кластеризации существенно различаются между собой. Метод "ближайшего соседа" [3] представляет собой процедуру, в рамках которой каждое новое наблюдение присоединяется к формируемому кластеру по принципу наибольшей близости. При этом близость исчисляется по следующему правилу: первые два наблюдения объединяются в том случае, если они имеют самые близкие результаты по совокупности переменных. В дальнейшем каждое новое наблюдение отыскивается по степени близости или

подобия двух точек внутри кластеров. Иными словами, в ходе названной процедуры плотность самого кластера, определяемая дисперсией признаков внутри него, не принимается во внимание. Отсюда, метод “ближайшего соседа” способен генерировать достаточно “вялые” кластеры, но при этом сама процедура агломерирования (соединения) кластеров носит постепенный, ступенчатый характер. Проиллюстрируем использование данного метода нашим примером.

Диаграмма 1.

В этом случае дистанция между различными элементами, формирующими новые кластеры, измеряется стандартизованной величиной, колеблющейся от 1 до 100. Полное слияние этих элементов происходит в точке шкалы, равной 100. Как видно из дендрограммы (Диаграмма 1), в точке 78 процедура слияния, реализованная почти полностью, привела к формированию двух кластеров, каждый из которых базируется на определенной группе

переменных. Дальнейший анализ кластеризуемых переменных (а в фокусе процедуры были переменные, а не отдельные наблюдения) затруднен по причине ступенчатости процедуры.

Примерно те же результаты дает процедура "дальнего соседа", при которой кластеры формируются по принципу отдаленности наблюдений (переменных в данном случае) друг от друга.

В нашем случае возможно применение еще одного метода, предполагающего более энергичные перепады между кластерами. Это так называемый метод "увязки средних величин" (UPGMA) [4, p. 180]. Здесь используется метод минимизации средних величин. Для соединяемых кластеров вычисляется дистанция между всевозможными парами наблюдений, входящих в них. Затем эти величины суммируются, на основании полученной суммы выявляется наименьшая из всех суммарных дистанций, а далее кластеры, разделенные наименьшей дистанцией, сливаются в один. Таким образом, при формировании кластера методом UPGMA имеет место более жесткая привязка процесса агломерации к средним величинам переменных. Эта положительная сторона метода оборачивается сложностями, когда речь идет о переменных, измеряемых по порядковой шкале, для которых средние величины - весьма сомнительная характеристика, а тем более для дихотомических переменных, которые вообще не вполне пригодны для данного метода.

Рассмотрим результаты анализа, произведенного при помощи метода UPGMA.

Диаграмма 2 подтверждает, что, как и в предыдущем случае, мы имеем постепенный процесс соединения признаков, не дающий возможности четко выделить более двух реальных кластеров: один, включающий в себя две переменные, – в верхней части дендрограммы, другой, объединяющий все остальные переменные кроме трех, – в нижней.

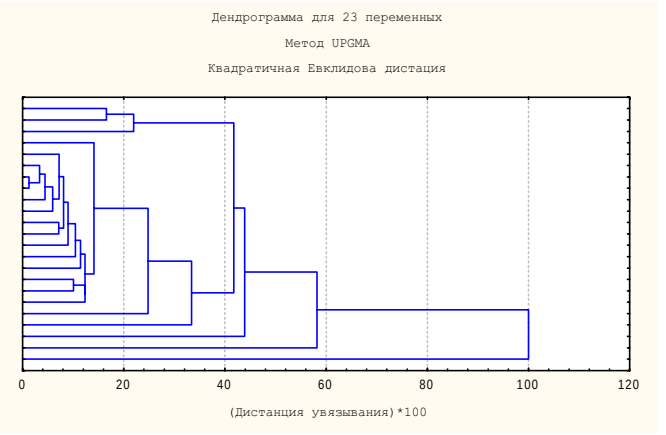

Диаграмма 2.

Метод Уорда [4, p. 180], наиболее часто применяемый в кластерном анализе, как и предыдущий, базируется на средних величинах. Для каждого кластера рассчитывается квадратичная Евклидова дистанция от средних величин переменных внутри кластера и средних величин переменных, присоединяемых к нему.

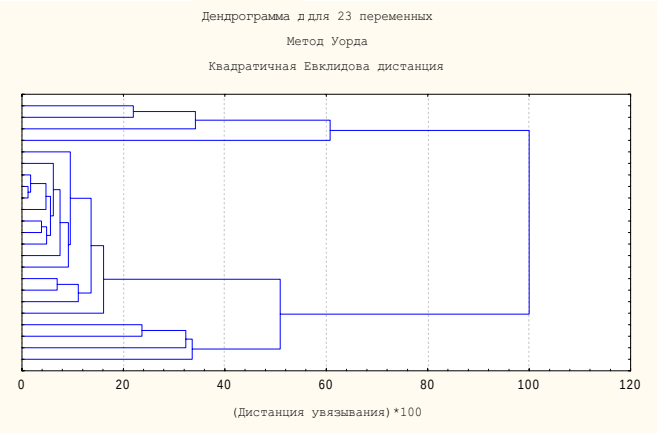

Диаграмма 3.

Метод Уорда дал результаты, схожие с предыдущими: произошло деление совокупности на два крупных кластера, в один из которых входит четыре переменных в верхней части таблицы, а остальные девятнадцать переменных - в другой. Отметим, что в данном случае большой кластер вобрал в себя и четыре переменные, не входившие ранее ни в один из кластеров. Кроме этого, при внимательном изучении полученных данных можно видеть, что большой кластер сам образован из двух возможных кластеров. Таким образом, можно предположить, что оптимальным было бы деление совокупности не на две, а на три большие группы.

В связи с вышесказанным возникает вопрос, как можно верифицировать полученный результат? По-видимому, наилучшим способом выявления надежности данных анализа было бы выявление дифференцирующего потенциала каждого из возможных вариантов. Иными словами, реальность каждого из результатов должна проверяться на конкретных примерах с использованием описательной статистики.

Еще одна проблема заключается в том, чтобы перейти от группировки переменных к группировке наблюдений. Последнее представляется довольно сложной проблемой, если учесть, что иерархический кластерный анализ работает с ограниченным числом наблюдений, как правило, не превышающим 300. Но опыт показывает, что даже эта цифра является чрезмерной, поэтому лучше свести число наблюдений к еще меньшему числу.

Кластеризация переменных – это процедура, позволяющая выявить их группировки. В этом смысле она подобна факторному анализу, делающему то же самое. Результаты кластерного анализа переменных всегда полезно соотнести с результатами факторного анализа. Возникает вопрос, в чем разница между двумя процедурами, в чем преимущества каждой из них? Следует подчеркнуть, что теоретические основания факторного анализа более детально проработаны в плане теории и базируются на надежных математических моделях. Теория кластерного анализа

менее развита и, следовательно, всякий раз используя эту процедуру, исследователь ступает по зыбкой почве субъективных интерпретаций. Существуют, однако, две сильные стороны кластеризации, заставляющие часто прибегать к ней. Во-первых, в отличие от факторов, получаемых в ходе факторного анализа, кластеры не включают в себя значения переменных с разными знаками. Кластерный анализ объединяет переменные, связанные только положительной ассоциацией. Во-вторых, кластерный анализ (с учетом первой его сильной стороны) позволяет более однозначно, чем факторный анализ, соотносить отдельные наблюдения с каждой из группировок наблюдений. Для этого в кластер включаются только наблюдения, имеющие сходную с ним конфигурацию связи между переменными. В стандартных пакетах обработки данных для этого предусмотрены специальные процедуры.

Иную альтернативу объединения наблюдений предлагает так называемый K-Means-кластерный анализ [3, p. 23], предполагающий формирование быстрых кластеров. Он, как UPGMA или метод Уорда, базируется на средних. В основе метода "быстрых кластеров" лежит центроидная стратегия, сходная с дисперсионным анализом. На предварительном этапе анализа задается некое число кластеров, а далее соотношение (F) дисперсий внутри и вне кластера увеличивается до максимальных значений. Огрубляя детали этой процедуры, можно сказать, что в данном случае выполняется одномерный дисперсионный анализ, в рамках которого размер групп неизвестен, а наблюдения включаются в кластеры так, чтобы F-величина достигала максимальных значений. "Быстрый кластер" – это процедура, в рамках которой итерации совершаются до тех пор, пока все наблюдения не отнесены к определенным кластерам. В этом одно из её основных отличий от иерархического кластерного анализа, позволяющего определить группу переменных, действительно группирующих наблюдения в кластеры.

Результаты кластеризации

Предварительная работа с данными исследования позволила нам выявить два альтернативных варианта реализации процедуры кластерного анализа. Возможно задавать в качестве условия процедуры “быстрый кластер” разбиение как на два, так и на три кластера. Протестируем сначала первый вариант разбиения – на два кластера. Исследователь определяет те переменные, которые будут определять формируемые кластеры. В нашем случае это переменные, раскрывающие возможные направления развития нашего общества.

К-КЛАСТЕР (СРЕДНИЕ ВЕЛИЧИНЫ ПО ПЯТИБАЛЛЬНОЙ ШКАЛЕ)

Таблица 1.

	Кластеры	
	Первый	**Второй**
Развитие демократии и инициативы граждан	3,70	3,58
Развитие рыночной экономики	3,94	3,55
Наведение строжайшей дисциплины во всем обществе	4,17	4,76
Приватизация госпредприятий	2,70	2,19
Вхождение в мировую экономику	3,75	3,79
Обеспечение полной занятости	4,28	4,88
Борьба с преступностью	4,66	4,95
Социальная поддержка обездоленных	4,18	4,82
Обеспечение прав каждого отдельного гражданина	4,35	4,79
Предотвращение развала страны	4,54	4,94
Защита интересов русских	3,59	4,43
Развитие духовности	4,07	4,73
Улучшение межнациональных отношений	3,73	4,69
Материальное равенство граждан	2,96	3,98
Воссоединение бывших советских республик	1,98	4,03
Борьба с бюрократией	3,81	4,77
Защита семьи и повышение рождаемости	4,06	4,77
Восстановление производства	4,71	4,97
Восстановление сельского хозяйства	4,69	4,98
Борьба с алкоголизмом и наркоманией	4,16	4,83
Борьба с коррупцией	4,38	4,91
Борьба с засильем иностранного капитала	2,86	4,44
Укрепление национальной валюты	4,45	4,80

Как видно из Таблицы 1, кластеры не слишком различаются между собой. Первый кластер (451 респондент) чуть больше значения придает развитию демократии, рыночной экономики и заметно ниже оценивает актуальность борьбы с иностранным капиталом. В этой группе сосредоточились те, кто чуть меньше жаждет возвращения к силовым методам управления страной, например, "наведения строжайшей дисциплины во всем обществе" или усиления борьбы с преступностью. У респондентов, вошедших в первый кластер, нет убежденности в том, что российская политика должна быть направлена на воссоединение СССР. Во многих случаях различия между значениями переменных в разных кластерах исчисляются десятыми долями балла (напомним, что речь идет о средней, вычисляемой на основе пятибалльной шкалы). Именно эта особенность полученных результатов не дает возможности однозначно заключить, что в первом кластере собрались "либеральные сторонники свободы и демократии", а во втором – "авторитарные сторонники силовых действий". На наш взгляд, проблема использования K-Means-кластерного анализа состоит, прежде всего, в той его особенности, о которой говорилось, а именно в обязательности распределения респондентов по разным кластерам. Это не позволяет изъять из каждой группы тех, кто очень слабо соотносится с той или иной точкой зрения. Иными словами, мы не получаем столь необходимой нам градуированной картины, на основании которой можно было бы определить группу последовательных сторонников рынка, группу колеблющихся, группу расходящихся с превалирующей в кластере точкой зрения по каким-то положениям и так далее.

Есть, однако, еще один выход, состоящий в том, чтобы увеличивать число кластеров. Это может "развести" респондентов, вынужденно соединенных в одну группу по разным кластерам.

Как видно из Таблицы 2, картина из трех кластеров выглядит более структурированной. Первый кластер, выступая за развитие

рынка и демократии, выступает одновременно за наведение строжайшей дисциплины, обеспечение полной занятости, борьбу с преступностью, восстановление отечественного производства и, в большей степени нежели другие, за приватизацию государственных предприятий. Он явно отличается от второго кластера, поддерживающего идею восстановления СССР. Что касается третьего кластера, то здесь значения всех переменных не вполне рельефны. Исключение составляет, разве что, переменная, отражающая поддержку борьбы за СССР: третий кластер испытывает наименьший энтузиазм по поводу этого направления развития страны. Можно предположить, что укрепление двух первых кластеров произошло именно за счет выведения части апатичных респондентов, не имеющих четких мнений по поводу возможного направления страны, в отдельную категорию.

Итак, резюмируя данные исследования, можно сказать, что, действительно, в обществе существует явное размежевание по вопросу будущего развития страны. Причем размежевание происходит по двум основным измерениям: первое – это поддержка силовых действий, второе – продолжение рыночных реформ. Важно также иметь ввиду, что эти измерения пересекаются между собой. Иными словами, немалая часть "рыночников" хотела бы видеть более жесткие действия государства, направленные на стабилизацию общества. В этом они вполне солидарны со своими антиподами, жаждущими вернуть утраченное прошлое.

Заключение

Пример использования кластерного анализа, приведенный выше, наглядно демонстрирует недостатки этого метода.

1. Процедура нетерпима по отношению к пропущенным данным. От исследователя требуется решить эту проблему до того, как он приступает к анализу. Любой способ решения проблемы

РЕЗУЛЬТАТЫ КЛАСТЕРНОГО АНАЛИЗА: РАЗБИЕНИЕ НА ТРИ КЛАСТЕРА (СРЕДНИЕ ЗНАЧЕНИЯ ПО ПЯТИБАЛЛЬНОЙ ШКАЛЕ): ТРИ КЛАСТЕРА

Таблица 2.

	Размер 1 кластера (516)	Размер 2 кластера (697)	Размер 3 кластера (229)
Развитие демократии и инициативы граждан	4,34	3,28	3,07
Развитие рыночной экономики	4,44	3,21	3,39
Наведение строжайшей дисциплины во всем обществе	4,65	4,74	3,79
Приватизация госпредприятий	3,26	1,73	2,28
Вхождение в мировую экономику	4,42	3,52	3,11
Обеспечение полной занятости	4,75	4,88	3,88
Борьба с преступностью	4,93	4,96	4,35
Социальная поддержка обездоленных	4,65	4,83	3,77
Обеспечение прав каждого отдельного гражданина	4,70	4,78	4,07
Предотвращение развала страны	4,89	4,93	4,21
Защита интересов русских	4,18	4,37	3,36
Развитие духовности	4,58	4,72	3,67
Улучшение межнациональных отношений	4,43	4,69	3,24
Материальное равенство граждан	3,60	3,96	2,72
Воссоединение бывших советских республик	2,54	4,36	1,98
Борьба с бюрократией	4,51	4,77	3,28
Защита семьи и повышение рождаемости	4,60	4,76	3,67
Восстановление производства	4,92	4,98	4,51
Восстановление сельского хозяйства	4,91	4,98	4,48
Борьба с алкоголизмом и наркоманией	4,69	4,83	3,72
Борьба с коррупцией	4,85	4,91	3,93
Борьба с засильем иностранного капитала	3,59	4,50	2,81
Укрепление национальной валюты	4,72	4,81	4,20

вряд ли можно признать адекватным, поскольку и математическая импутация[1] данных и любая другая подстановка неизбежно привносят субъективное начало в получаемые результаты.

[1] Импутация – замена пропущенных данных значимыми величинами. Применяется при подготовке сложных процедур, чтобы свести к минимуму изъятие наблюдений.

2. Кластерный анализ неадекватно решает проблемы классификации в том случае, когда используются порядковые или номинальные шкалы. Для номинальных шкал требуется сложная и громоздкая процедура их перекодирования в дихотомические, а затем стандартизация, что является коренным преобразованием изначальной информации и, безусловно, чревато существенными её искажениями.

3. Иерархический кластерный анализ не может работать с большим числом наблюдений. Это – очень существенное ограничение для социологов, как правило, имеющих дело с большими массивами данных.

4. Разные процедуры кластерного анализа плохо согласуются друг с другом. Например, иерархический кластерный анализ не вполне соотносится с процедурой "быстрый кластер".

5. В конечном итоге число кластеров, задаваемых исследователем, остается на его совести. Так, в приведенном выше примере, вполне допустимо продолжать увеличивать число кластеров, получая все новые и новые группы.

6. Процедура кластерного анализа не позволяет производить более дробную и четкую структуризацию совокупности, что и является основной задачей исследования.

Очевидно, что на сегодняшний день кластерный анализ в его стандартном исполнении (SPSS или Statistica) – это весьма слабое подспорье в той работе по классификации данных, которая является важной частью любого социологического исследования.

ЛИТЕРАТУРА

1. *Converse P.* The nature of belief systems in mass publics//Ideology and discontent. New York:Free Press, 1964.
2. *Wright E.O.* Classes. New York:Verso, 1985.
3. *Aldenderfer/Blashfield.* Cluster analysis. Sage Publications, 1985.
4. *Norusis M.J.* SPSS-X Advanced statistics guide. SPSS International.