МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

СПЕЦИФИКА КЛАСТЕРИЗАЦИИ СОЦИОЛОГИЧЕСКИХ ДАННЫХ (НА ПРИМЕРЕ ИЗУЧЕНИЯ ОТНОШЕНИЯ МОЛОДЕЖИ Г. САРАТОВА К НЕЗАРЕГИСТРИРОВАННЫМ БРАКАМ)

(автореферат бакалаврской работы)

студента 5 курса 531 группы
направления 09.03.03 – Прикладная информатика
профиль Прикладная информатика в социологии
социологического факультета
Большакова Александра Викторовича

Научный руководитель
профессор, доктор социологических наук С.В. Ситникова

Заведующий кафедрой
кандидат социологических наук, доцент И.Г. Малинский

Саратов, 2020 год

ВВЕДЕНИЕ

Актуальность проблемы исследования. В современной России, как и во всем мире, социологические исследования стали привычным явлением. Опросы населения проводятся по самым различным проблемам, волнующим как представителей власти, так и рядовых граждан. Вместе с тем, разработать анкету и опросить репрезентативную выборку респондентов – задача сложная, но не единственная. Собственно доступ к собранной информации можно получить, лишь обработав и обобщив обширные разрозненные данные, а затем подвергнув их анализу с помощью математического аппарата. В докомпьютерную эпоху это была очень трудо- и времязатратная задача, решить которую могли лишь специалисты в области математических наук. В настоящее время ситуация изменилась. С распространением персональных компьютеров у исследователей-социологов появился доступ к большому числу программ, специализирующихся на математической обработке и анализе разнообразной информации. Выбор программных продуктов данного типа чрезвычайно широк: STATISTICA, STATA, R, PSPP (универсальные), SAS, BMDP (профессиональные), BioStat, DATASCOPE, DA-система (специализированные) и мн.др., однако SPSS можно назвать одной из самых популярных программ, использующейся во всем мире.

SPSS обрела свою популярность благодаря многим причинам, одной из которых является ее универсальность. К ее помощи прибегают специалисты из разных областей знаний, в том числе и социологи. Кроме того, она предлагает очень широкий спектр инструментов для работы с данными, воспользоваться которыми может человек, имеющий как серьезную математическую подготовку, так и довольно поверхностные знания в области математики.

Иерархический кластерный анализ входит в пакет возможностей SPSS. С одной стороны, он предъявляет не очень высокие требования к данным и довольно прост в реализации, с другой – его результаты довольно сложно интерпретировать. Зачастую для этого необходимо обратиться к помощи

других аналитических инструментов, что делает использование иерархического кластерного анализа не только сложной, но и интересной задачей.

Степень изученности темы исследования. Востребованность программы вызвала публикацию многочисленных учебников, пособий и руководств по работе с SPSS. Первые работы, посвященные особенностям анализа социологических данных при помощи компьютерных технологий, были переводными. Среди авторов таких работ можно назвать А. Бююля, П. Цефеля, Дж. Хили. Работы отечественных исследователей появились вскоре после публикации первых пособий зарубежных авторов по этой тематике и были довольно разнообразными. Областью статистической обработки информации заинтересовались социологи, психологи, экономисты и мн.др.

Интерес исследователей был сосредоточен не только на SPSS как одной из наиболее популярных программ для статистической обработки данных, но и на других программных продуктах. Появились попытки реализации комплексного соединения информационных технологий и социологии, а также теоретико-методологического осмысления проблем компьютерной поддержки социологического эмпирического исследования. Тем не менее, подавляющее большинство работ содержат, в основном, лишь алгоритмы проведения основных статистических процедур. Попыток более или менее системного описания возможностей программных продуктов, в том числе и SPSS, практически не представлено, что делает нашу работу весьма актуальной.

Объектом данного исследования являются виды иерархического кластерного анализа; *предметом* выступает характеристика эвристических возможностей иерархического кластерного анализа, применяемого вместе с другими аналитическими инструментами.

Цель исследования – выявить потенциал иерархического кластерного анализа в решении задачи выявления и описания отношения молодежи к незарегистрированному браку. Постановка цели определила формулирование следующих *задач* исследования:

1. Представить теоретические основы метода кластеризации статистических данных;

2. Охарактеризовать иерархический кластерный анализ как инструмент статистического анализа эмпирической информации;

3. Рассмотреть возможности иерархического кластерного анализа применительно к характеристике отношения молодежи к незарегистрированному браку;

4. Представить кластерные модели, характеризующие разные типы семейных отношений современных молодых людей.

Методологической базой исследования выступает принцип позитивистского подхода к изучению социальной реальности, впервые сформулированный основателем социологии Огюстом Контом и заключающийся в необходимости количественного представления социальных явлений и процессов, а также элементы системного анализа.

В качестве *эмпирической базы* исследования были использованы данные массового социологического опроса, проведенного автором по теме «Отношение молодежи г. Саратова к незарегистрированным бракам».

Научная новизна заключается в раскрытии аналитического потенциала иерархического кластерного анализа в социологическом исследовании, в том числе в тесном взаимодействии с другими методами анализа.

Структура работы. Данная работа состоит из введения, трех разделов (1 раздел «Иерархический кластерный анализ: общая характеристика, виды, специфика применения для социологических данных», 2 раздел «Практический пример применения иерархического кластерного анализа переменных при изучении отношения молодежи к незарегистрированным бракам», 3 раздел «Практический пример применения иерархического кластерного анализа случаев при изучении отношения молодежи к незарегистрированным бракам»), заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Иерархический кластерный анализ: общая характеристика, виды, специфика применения для социологических данных» посвящен обзору иерархического кластерного анализа, его определению, описанию механизма работы, характеристике его видов: кластерного анализа переменных и кластерного анализ случаев, уточнению их сильных и слабых сторон.

Кластерный анализ – это набор многомерных статистических методов, нацеленных на исследование структуры некоторой совокупности переменных или объектов.

Главной задачей кластерного анализа переменных заключается в переходе от первоначальной совокупности множества переменных к значительно меньшему числу кластеров.

Главным итогом иерархического кластерного анализа является дендрограмма, позволяющая определить число искомых кластеров. При её интерпретации исследователи сталкиваются проблемой отсутствия однозначных критериев выделения кластеров. Существует несколько способов ее преодоления, но лучшим считается обращение к факторному анализу.

Факторный анализ позволяет решить важную исследовательскую задачу: дать всестороннее и одновременно компактное описание объекта изучения. Для этого в ходе факторного анализа выявляются латентные переменные или факторы, которые отвечают за наличие линейных корреляционных связей между наблюдаемыми переменными.

Таким образом, оба метода, и факторный анализ, и иерархический кластерный анализ переменных, являются эффективными инструментами выявления латентных переменных через исследование взаимосвязей между наблюдаемыми переменными. Действия, выполняемые в ходе статистических операций в каждом из методов, принципиально различаются. Итоговое решение сильно зависит от того, какие меры связи между наблюдаемыми переменными будут выбраны для расчетов. Тем не менее, итоговый набор

латентных переменных (факторов и кластеров), как правило, совпадает. Поэтому с целью обеспечения более тщательного контроля над переменными исследователю целесообразно применять оба метода.

Кластерный анализ случаев выполняет задачу разбиения заданной выборки наблюдений на подмножества или кластеры таким образом, чтобы каждый отдельный кластер состоял из схожих случаев, а случаи разных кластеров существенно отличались.

Кластерный анализ является эффективным и простым методом классификации, предлагающим весьма наглядные результаты. К его основным преимуществам можно отнести отсутствие ограничений на нормальное распределение переменных; возможность классификации в случаях отсутствия априорной информации о классах; универсальность (применимость и к объектам, и к переменным). Вместе с тем, получение нескольких кластеров случаев, отличающихся друг от друга, отнюдь не означает их правильной интерпретации. Для того, чтобы дать точную характеристику вновь полученным группам респондентов, можно использовать много разных приемов статистического анализа, одним из самых распространенных является дискриминантный анализ.

Дискриминантный анализ представляет собой инструмент прогнозирования, с помощью которого можно предсказать принадлежность случаев к двум или более непересекающимся группам. Исходными данными для него выступает множество объектов, разделенных на группы таким образом, что каждый отдельный объект относится только к одной группе.

Данные, характеризующие рассматриваемые объекты, должны быть представлены в формате количественных (или условно «количественных») шкал. Данные переменные определяются как дискриминантные переменные или предикторы.

Дискриминантный анализ позволяет определить правила, которые бы позволили по значениям дискриминантных переменных (или предикторов) отнести каждый объект к одной из заданных групп и вычислить «веса» каждой

дискриминантной переменной, с помощью которой объекты разделяются на группы.

Таким образом, одновременное использование кластерного анализа случаев и дискриминантного анализа является очень эффективным инструментом статистического анализа, т.к. позволяет не только по-новому дифференцировать выборочную совокупность, но и дать точную и обоснованную характеристику новым группам.

Второй раздел «Практический пример применения иерархического кластерного анализа переменных при изучении отношения молодежи к незарегистрированным бракам» описывает пример использования метода иерархического кластерного анализа переменных.

Иерархический кластерный анализ переменных был проведен методом Уорда и корреляцией Пирсона в качестве меры измерения расстояния между переменными. Анализ данных таблицы последовательности слияния показал, что оптимальным решением задачи будет три или четыре кластера.

В результате проведения факторного анализа при помощи Обобщенного метода наименьших квадратов с вращением факторов по методу Варимакса было получено решение, состоящее из 4 факторов. Об адекватности и оптимальности полученного решения свидетельствует его высокая устойчивость (попытки рассчитать другие факторные решения давали очень похожее распределение переменных по факторам), логичность и возможность интерпретирования факторов, а также 43,5% дисперсии, охватываемых данным решением. Таким образом, в ходе проведения факторного анализа были обнаружены четыре латентные переменные, разбивающие эмпирические переменные на четыре группы, что совпало с результатом кластерного анализа. Это говорит об устойчивости выявленных связей между переменными и решения в целом. Более того, информация о факторных нагрузках позволяет нам точнее интерпретировать взаимодействия между переменными и облегчает задачу определения латентных переменных: «Ориентации на традиционные

ценности», «Стабильность семейных отношений», «Ориентация на финансовое благополучие» и «Ориентация на семейную жизнь».

Третий раздел «Практический пример применения иерархического кластерного анализа случаев при изучении отношения молодежи к незарегистрированным бракам» посвящен дифференциации совокупности респондентов на группы, придерживающиеся различных взглядов на гражданский брак.

Для проведения кластерного анализа случаев в качестве метода кластеризации был выбран метод Варда, для вычисления расстояния между объектами – квадрат расстояния Евклида. Главным результатом проведения кластерного анализа случаев является расчет таблицы последовательности слияния исследуемых случаев, которая позволяет предварительно определить число кластеров. Анализ данных таблицы последовательности показал, что резкое возрастание различий обнаруживается при переходе от шага 126 к 127. Следовательно, оптимальное количество кластеров равно 3 или 2.

Для социологической интерпретации полученного результата был использован дискриминантный анализ методом Уилкса, основанный на минимизации коэффициента Уилкса после включения в уравнение регрессии каждого нового предиктора. Тест равенства групповых средних показал, что хорошей дифференцирующей способностью, позволяющей эффективно работать в регрессионном уравнении, обладают 4 переменные из 18: «Возраст», «Средний доход в месяц», «Наличие разногласий с партнером» и «Власть». В ходе дискриминантного анализа были вычислены 2 функции, позволяющие сравнить и выявить различия между данными группами: *«Семейный доход»* и *«Наличие разногласий»*.

Результаты интерпретации трех кластеров респондентов показали, что первая группа респондентов характеризуется небольшим отрицательным значением функций «Семейный доход» и «Наличие разногласий». Таким образом, в первую группу вошли респонденты, проживающие в семьях с относительно невысоким доходом, но не сталкивающихся с разногласиями. Их

определили как *«необеспеченных, но психологически учтойчивых»*. Их доля в выборочной совокупности оказалась наибольшей - 83,9%.

Вторая группа респондентов отличается значительным положительным значением функции «Семейный доход» и небольшим, но положительным значением по переменной «Наличие разногласий», другими словами, данную группу составили респонденты, проживающие в семьях средней обеспеченности, но сталкивающиеся с разногласиями: *«средне обеспеченные, но психологически неблагополучные»* (11,5% выборочной совокупности).

Третья группа опрошенных отличается очень большим положительным значением функции «Семейный доход» и небольшим отрицательным значением по второй функции, что характеризует их как людей, проживающих в семьях очень обеспеченных и психологически благополучных: *«обеспеченные и психологически благополучные»* (4,6% выборочной совокупности).

Точность прогноза составила 98,5%, самый слабый прогноз был составлен для первой, наиболее многочисленной, группы респондентов - «необеспеченных, но психологически устойчивых» - фактическая и прогнозируемая принадлежность не совпала лишь для 1,8% представителей данной группы.

ЗАКЛЮЧЕНИЕ

Программа статистической обработки данных SPSS является мощным инструментом анализа социологической информации. Она предлагает множество методов обработки данных. К числу популярных относится кластерный анализ. В целом он является эффективным и простым методом классификации, предлагающим весьма наглядные результаты. К его основным преимуществам можно отнести отсутствие ограничений на нормальное распределение переменных; возможность классификации в случаях отсутствия априорной информации о классах; универсальность (применимость и к объектам, и к переменным).

Главная задача кластерного анализа переменных заключается в переходе от первоначальной совокупности множества переменных к значительно меньшему числу кластеров.

Вместе с тем, у данного метода есть и слабые стороны. Кластерный анализ переменных предлагает простое и визуализированное решение разбиения переменных на кластеры, которое не раскрывает особенности взаимосвязей между самими переменными, что часто затрудняет их интерпретацию. Кроме того, статистики рекомендуют обращаться к данному методу в случае небольшого числа переменных (не более 10). Выходом из данного затруднения является одновременное обращение к факторному анализу, цель которого также заключается в объединении множества переменных в небольшое число факторов. При этом факторный анализ не связан ограничением числа переменных, и исследователь имеет возможность более тонкого изучения взаимосвязей между переменными для корректной интерпретации полученных факторов. Также факторный анализ позволяет сохранить полученные факторные значения в базе данных как новые переменные, в отличие от кластерного анализа переменных.

В ходе авторского исследования использование кластерного и факторного методов анализа привело к получению практически идентичной структуры латентных переменных, не совпадающих лишь частично по «технической» причине: в факторном анализе одна переменная может присутствовать в более чем одной латентной переменной, в кластерном же анализе это невозможно. В результате была построена модель, объяснявшая 43,5% выборочной совокупности. Вместо 11 эмпирических переменных были получены 4 латентные переменные: «Ориентация на традиционные ценности», «Стабильность семейных отношений», «Ориентация на финансовое благополучие» и «Ориентация на семейную жизнь».

Для выявления групп студенческой молодежи, придерживающихся разных форм восприятия гражданского брака, мы обратились к кластерному анализу случаев.

Кластерный анализ случаев выполняет задачу разбиения заданной выборки объектов на подмножества или кластеры таким образом, чтобы каждый отдельный кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались, при этом изначально информация о числе кластеров и их составе неизвестна. Вместе с тем, получение нескольких кластеров случаев, отличающихся друг от друга, отнюдь не означает их правильной интерпретации. Для того, чтобы дать точную характеристику вновь полученным группам респондентов, целесообразно обратиться к дискриминантному анализу.

Дискриминантный анализ представляет собой инструмент прогнозирования, с помощью которого можно предсказать принадлежность случаев к двум или более непересекающимся группам. Исходными данными для него выступает множество объектов, разделенных на группы таким образом, что каждый отдельный объект относится только к одной группе, причем их принадлежность к той или иной группе является известной, что отличает его от кластерного анализа случаев. Выявленные отличия данных методов позволяет на практике использовать их в паре. Это обеспечивает не только успешную перегруппировку данных и более легкий способ характеристики вновь полученных групп, но и построение уравнения регрессии с очень высокой степенью точности прогноза.

Иллюстрируя возможности совместного применения данных методов на примере данных социологического исследования, посвященного незарегистрированному браку, с помощью кластерного анализа были выделены 3 группы респондентов, реализующих различные типы поведения в семейной жизни.

В ходе проведения дискриминантного анализа были вычислены две функции, отвечающие за прогноз принадлежности респондентов к той или иной группе, и получившие названия «Семейный доход» и «Наличие разногласий с партнером».

Анализ значений центроидов групп позволил дать характеристику выделенным группам. Первая группа получила название «Необеспеченных, но психологически устойчивых», поскольку ее отличительными признаками оказались небольшое отрицательное значение функции «Семейный доход» и незначительное влияние со стороны функции «Наличие разногласий с партнером». Вторая группа – «Средне обеспеченные, но психологически неблагополучные» - имела довольно заметное положительное значение функции «Семейный доход» и небольшое положительное значение по второй переменной. Третья группа – «Обеспеченные и психологически благополучные» - напротив, отличилась очень большим положительным значением функции «Семейный доход» и небольшим отрицательным значением по функции «Наличие разногласий с партнером».

Таким образом, согласно полученным результатам, совместное применение кластерного и дискриминантного методов анализа оказалось весьма эффективным инструментом для классификации респондентов по самым разным основаниям, в том числе весьма непривычным в рамках проведения массового обследования населения. В ходе реализации данной техники анализа была осуществлена перегруппировка респондентов по новым основаниям, заданным не одной, а нескольким переменным, дана интерпретация и подробная характеристика каждой вновь созданной группы опрошенных и вычислено дискриминантное уравнение, позволяющее прогнозировать принадлежность неизвестных объектов, в том числе из генеральной совокупности, с точностью 98,5%.

Подводя итоги, еще раз отметим, что иерархический кластерный анализ выступает мощным методом обработки социологической информации, который позволяет получить интересные результаты, что не избавляет его от ряда недостатков. Нейтрализовать эти недостатки и полностью раскрыть потенциал кластерного анализа позволяет его применение в сочетании с другими аналитическими методами, в нашем случае это были факторный и дискриминантный анализ. Программа SPSS, имеющая в своем распоряжении

широчайший набор аналитических инструментов, не случайно является одной из самых популярных программ в своем сегменте.