МИНОБРНАУКИ РОССИИ

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

Кафедра социальной информатики

СТРАТЕГИИ МИГРАЦИИ ЖИТЕЛЕЙ МАЛОГО ГОРОДА: ОПЫТ ПРИМЕНЕНИЯ ИЕРАРХИЧЕСКОГО КЛАСТЕРНОГО АНАЛИЗА

(автореферат бакалаврской работы)

студента 5 курса 531 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Череваткина Максима Александровича

Научный руководитель
доцент, кандидат социологических наук,
_____ С.В. Курганова
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2019

ВВЕДЕНИЕ

Актуальность проблемы. Согласно истории развития эмпирической социологии, в середине XX столетия наступил расцвет массовых социологических исследований. Конечно, достижение таких успехов эмпирической социологией было обусловлено всей предшествовавшей историей ее развития, разработкой теоретико-методологических оснований социологических исследований в целом, шлифовкой и совершенствованием методов сбора информации, накоплением практического опыта организации исследований и т.д. Однако был еще один довольно редко упоминающийся фактор, который внес свой вклад в расцвет эмпирических исследований того времени, а именно появление и бурное развитие специализированных компьютерных программ, позволивших резко повысить качество сбора, хранения и обработки результатов проведенных исследований.

Уже в 1965-м году американские студенты Норманн Най и Дейл Вент, обучавшиеся политологии в Стэнфордском университете, США, попытались найти компьютерную программу, с помощью которой можно было бы проанализировать статистическую информацию. Они перебрали все имевшиеся на тот момент программы, но ни одна из них не показалась им более или менее пригодной: они были либо неудачно построенными, либо не обеспечивали наглядность представления обработанной информации. Тогда студенты решили разработать собственную программу со своей единой концепцией и единым синтаксисом. Через год была готова первая версия программы, а еще через год – версия, которая смогла работать на IBM 360. Так была создана программа SPSS. С тех пор она стала одной из популярнейших программ статистической обработки данных, позволяющей обрабатывать данные из областей социологии, маркетинга, биологии, психологии и медицины. За это время кроме SPSS на рынке появились и другие программы, также позволяющие решить проблему анализа статистических данных, например, STATISTICA,STATA, R, PSPP(универсальные), SAS, BMDP (профессиональные), BioStat, DATASCOPE,

DA-система (специализированные) и мн.др. Но SPSS до сих пор является одной из самых популярных программ, использующейся во всем мире.

Степень научной разработанности. 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 сравнительный анализ эвристических возможностей разных видов иерархического кластерного анализа при работе с социологическими данными.

Цель исследования – выявить потенциал иерархического кластерного анализа в решении задачи выявления и описания поведенческих стратегий разных групп населения. Постановка цели определила формулирование следующих **задач** исследования:

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

2. Рассмотреть возможности иерархического кластерного анализа применительно к характеристике поведенческих стратегий разных групп населения.

В качестве **эмпирической базы** исследования были использованы данные массового социологического опроса, проведенного студентами социологического факультета СГУ по теме «Специфика миграционного поведения населения малого города (на примере г. Петровска)»[1].

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, двух разделов (1 раздел «Иерархический кластерный анализ: общая характеристика, виды, возможности реализации в программе SPSS», 2 раздел «Практический пример применения иерархического кластерного анализа при изучении стратегий миграции жителей малого города»), заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Иерархический кластерный анализ: общая характеристика, виды, возможности реализации в программе SPSS» 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.

[1] Данное социологическое исследование было проведено в 2012 году студентами социологического факультета СГУ. Метод сбора информации – стандартизированное анкетирование. Выборка строилась по принципу стратифицированной (объектом исследования выступали жители города Петровска в возрасте от 18 лет и старше). Всего было опрошено 150 респондентов.

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 при изучении стратегий миграции жителей малого города» описывается пример использования методов иерархического кластерного анализа с целью выявить стратегии трудовой миграции жителей малого города Красноармейска.

Кластерный анализ переменных мы использовали для объединения нескольких переменных-ценностей в более крупные латентные переменные. Были получены два решения: 3-кластерное и 4-кластерное, но оба имели недостатки. 3-кластерное решение оказалось сложным для интерпретации, т.к. 1 и 2 кластеры содержали противоречивые по смыслу переменные. Тем не менее, с некоторой натяжкой их можно определить как «Рыночные ценности», «Традиционные / семейные ценности» и «Индивидуалистические ценности». Что же касается второго решения, то оно было непригодно уже с математической точки зрения, поскольку четвертый кластер содержал лишь 1 переменную.

Факторный анализ позволил обнаружить две латентные переменные «Традиционные ценности» и «Либеральные ценности», что несколько отличалось от кластерных решений, но только частично. В итоге это решение было выбрано в качестве рабочего.

К кластерному анализу случаев мы обратились для выявления групп респондентов, реализующих разные стратегии трудовой миграции. В анализе было использовано 14 переменных с количественными или условно количественными шкалами.

Анализ таблицы последовательности слияния исследуемых случаев показал, что оптимальное число кластеров в данном случае – 4. Теперь надо было охарактеризовать данные группы, для чего обратились к дискриминантному анализу.

Поскольку в результате проведенного кластерного анализа случаев были получены 4 группы респондентов, реализующих свои стратегии трудовой миграции, то в ходе дискриминантного анализа были вычислены 3 функции, позволяющие сравнить и выявить различия между данными группами.

В первой функции, названной нами *«Оседлость»*, доминировали переменные «Традиционные ценности», «Поменять место жительства», «Считаю ее средством достижения более высокого социально-экономического статуса» и «Уехать работать в другой город». Вторая функция *«Трудовая миграция»* продемонстрировала сильную корреляционную связь с переменными «Уехать работать в другой город», «Вовлеченность в трудовую миграцию», «Искать другое место работы с более высокой оплатой в своем городе» и «Традиционные ценности». Третья функция *«Территориальная миграция»* характеризуется корреляционной связью с переменными «Считаю ее причиной разрушения семьи», «Искать другое место работы с более высокой оплатой в своем городе», «Уехать работать в другой город».

Для первой группы отличительными признаками являются большое положительное значение функции «Трудовая миграция», большое отрицательное значение функции «Оседлость» и незначительное влияние со стороны функции «Территориальная миграция». Они получили название *«отходников»* (вовлеченных в отходничество). Их доля в выборочной совокупности составила 18,5%.

Вторая группа респондентов имела большие отрицательные значения функции «Территориальная миграция» и небольшими значениями по другим переменным - это *«инертные»*. Всего в выборке их оказалось 20%.

Третья группа опрошенных отличалась большим положительным значением функции «Оседлость» и небольшими значениями по другим функциям - *«укорененные»* – 45,4%.

Четвертая группа получила высокие отрицательные значения всех функций «Оседлость» и «Трудовая миграция» и среднее положительное – функции «Территориальная миграция» - *«потенциальные мигранты»*. В выборке они оказались самыми малочисленными – всего 16,2%.

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 выявлению различной структуры латентных переменных-ценностей, но совпадающих в основных моментах. И поскольку кластерный анализ переменных не позволяет сохранить решение в базе данных,

то в итоге была выбрана модель, построенная с помощью факторного анализа и включающая два фактора: «Традиционные ценности» и «Либеральные ценности».

Для выявления групп жителей г. Красноармейска, придерживающихся разных стратегий относительно трудовой миграции, 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 трудовой миграции, с помощью кластерного анализа были выделены 4 группы респондентов, реализующих различные стратегии.

В ходе проведения дискриминантного анализа были вычислены три функции, отвечающие за прогноз принадлежности респондентов к той или иной

группе, и получившие названия «Оседлость», «Трудовая миграция» и «Территориальная миграция».

Анализ значений центроидов групп позволил дать характеристику выделенным группам. Первая группа получила название «Отходников», поскольку ее отличительными признаками оказались большое положительное значение функции «Трудовая миграция», большое отрицательное значение функции «Оседлость» и незначительное влияние со стороны функции «Территориальная миграция». Таким образом, в первую группу вошли респонденты, ориентированные на поиск заработка за пределами своего города, но продолжающие в нем проживать. Вторая группа – «Инертные» - имела большое отрицательное значение функции «Территориальная миграция» и небольшие значения по другим переменным, т.е. данную группу составили респонденты, не сделавшие явного выбора в пользу хоть какой-нибудь стратегии. Третья группа – «Укорененные» - напротив, отличилась большим положительным значением функции «Оседлость» и небольшими значениями по другим функциям, что характеризует их как людей, которых не привлекают какие-либо формы миграции, они предпочитают не покидать свой город. Наконец, четвертую группу – «Потенциальных мигрантов» - отличают высокие отрицательные значения функций «Оседлость» и «Трудовая миграция» и среднее положительное – функции «Территориальная миграция», что указывает на их стремление покинуть свой город совсем.

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 14 переменными, 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, превышающей 93%.

Применение таблиц сопряженности и критерия независимости Хи-квадрат Пирсона позволило выявить ряд социально-демографических закономерностей для выявленных групп. Отходники – это, как правило, молодые мужчины без профессионального образования, холостые и не имеющие детей, укорененные же являются их антиподом – женщины более зрелого возраста, с профессиональным образованием, замужние и имеющие детей. Для двух других групп – инертных и потенциальных мигрантов, таких профилей выявить не удалось, что объясняется их социально-демографическим разнообразием.

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.