МИНОБРНАУКИ РОССИИ

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

Кафедра социальной информатики

ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ СОЦИОЛОГИЧЕСКОГО ИССЛЕДОВАНИЯ

(автореферат бакалаврской работы)

студентки 5 курса 531 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Лаур Татьяны Леонидовны

Научный руководитель кандидат
физико-математических наук, доцент _____ Л.Б. Тяпаев
подпись, дата

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

Саратов 2022

ВВЕДЕНИЕ

Актуальность проблемы. В нынешнее время информация, растущая в значительных объёмах, выявляет потребность в обработке больших объёмов данных. В этом направлении существенная роль отведена интеллектуальному анализу данных. Данное направление включает в себя методы, отличающиеся от классического анализа. Они основаны на моделировании, вероятностных, и решающие задачи обобщения, ассоциирования и отыскания закономерностей. Развитию данной дисциплины поспособствовало проникновение искусственного интеллекта в сферу анализа данных идей.

В данной работе будет рассмотрена частная задача данного анализа, а конкретно, задачу иерархического кластерного анализа. Анализ и интерпретация полученных данных будет выполнена по данным социологического исследования.

Главной задачей кластерного анализа является выделение необходимого числа групп объектов схожих между собой внутри группы и максимально отличных от экземпляров других классов. Подобный анализ широко применяется в информационных системах для отыскания закономерностей в данных.

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. Получить доступ к собранной данным можно лишь структурировав и сгруппировав обширные разрозненную информацию, а потом проанализировав их с помощью математического программного обеспечения. 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 появлением ПК у социологов, проводящих исследования, появился возможность использовать различные программы,

специализирующиеся 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. [1], 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 содержит в себе очень большой спектр инструментов для работы с данными, работать с которыми может человек, знающий, как серьезную математическую базу, так и общие 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. Поэтому данный метод используют вместе с другими аналитическими инструментами, что делает использование иерархического кластерного анализа не только сложной, но и интересной задачей.

Целью выпускной квалификационной работы является описание и анализ иерархического кластерного анализа данных социологического исследования.

Сформулированная цель достигается решением, следующим **задач**:

1) Выявить особенности иерархического кластерного анализа
2) Изучить выполнение иерархического кластерного анализа с помощью программы SPSS.
3) Рассмотреть возможности иерархического кластерного анализа применительно к характеристике отношения молодежи к незарегистрированному браку.

[1] Статистическая обработка данных // Национальный открытый университет «Интуит» [Электронный ресурс]: сайт. URL: http://www.intuit.ru/studies/courses/3632/874/lecture/14309 (дата обращения: 26.04.2022).Загл. с экрана. Яз. рус.

4) Выполнить интерпретацию полученных данных социологического исследования.

Предметом выступают построение и интерпретация иерархического кластерного анализа в программе SPSS.

Объектом исследования является иерархический кластерный анализ.

В качестве **эмпирической базы** исследования были использованы данные 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, двух разделов заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Иерархический кластерный анализ» рассматриваются понятие и разновидности кластерного анализа, в частности определение иерархического кластерного анализа, описание его механизма работы, разновидности его возможностей: кластерный анализ наблюдений и выявление его сильных и слабых сторон.

Кластерный анализ или кластеризация – это задача группировки набора объектов таким образом, что объекты в одной и той же группе (называемой кластером) были более схожи (в том или ином смысле) друг с другом, чем с другими группами (кластерами).

Основой иерархического кластерного анализа является последовательное объединение кластеров для формирования больших кластеров или разделение больших кластеров на маленькие. Данная кластеризация достаточно наглядна только лишь при небольших объёмах исходных данных.

Метода иерархические кластеризации отличаются по правилам образования кластеров. По этим правилам определяются схожи ли объекты между собой и можно ли их отнести в один кластер. Методы, объединяющие

объекты в кластеры называются агломеративными, а разделяющие – дивизимными.

− агломеративные методы, в которых каждый объект первоначально находится в отдельном кластере. Затем объединяются два «ближайших» (наиболее похожих) кластера, процесс повторяется до тех пор, пока все объекты не будут в одном кластере. В конечном итоге оптимальное количество кластеров выбирается из всех кластерных решений;

− дивизивные методы, в которых все объекты первоначально находятся в одном кластере, и приведенная выше стратегия применяется в обратном порядке, пока каждый объект не перейдет в отдельный кластер.

Главной задачей кластерного анализа заключается в распределения исходной выборки данных на определенное количество кластеров.

Результатом иерархического кластерного анализа является дендрограмма, позволяющая определить число необходимых кластеров. При её интерпретации исследователи сталкиваются проблемой с такой проблемой, как отсутствие однозначных критериев выделения кластеров. Существует несколько способов ее преодоления:

Анализ таблицы последовательности слияния с целью найти резкое увеличение различий по шагам кластеризации и найти шаг, который отличает резкое возрастание различий. Адекватному числу кластеров будет соответствовать разность между числом случаев и порядковым номером шага, на котором был зафиксирован резкий скачок различий (в случае с кластерным анализом объектов).

— Визуальный анализ дендрограммы (особенно в случае небольшого объема выборки), поскольку данный вид диаграммы весь процесс кластеризации представляет графически в форме древовидной структуры. С помощью дендрограммы можно не только перейти к любому случаю на любом уровне кластеризации, но и можно увидеть, на каком расстоянии друг от друга находятся кластеры или случае на определенном уровне и, соответственно,

сделать вывод о том, на каком этапе остановить процесс кластеризации (в случае с обоими видами кластерного анализа).

— Сравнение результатов кластеризации, выполненной различными методами, в том числе с результатами факторного анализа (в случае с кластерным анализом переменных).

Кластерный анализ наблюдений 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. В этой работе будут рассмотрены анализ средних и таблицы сопряженности.

Создание таблиц сопряженности с кластерной переменной и переменными, используемыми ранее для разбиения выборки на кластеры. В рамках данного метода придется провести целый ряд отдельный итераций, затем сопоставить полученные результаты с целью получения интегральной характеристики всех имеющихся кластеров. Это делает данный способ очень результативным

Таким образом, иерархический кластерный анализ наблюдений является эффективным и простым методом классификации, что является его основным преимуществом. Вместе с тем, рекомендуют его использовать в случаях, когда

число переменных не превышает десяти, иначе интерпретировать полученные результаты будет очень сложно.

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 «Применения иерархического кластерного анализа с помощью программы SPSS при изучении отношения молодежи города Саратова к незарегистрированным бракам» рассматривает два примера использования метода иерархического кластерного анализа наблюдений.

Для первого анализа был выбран вопрос о мотивах вступления респондентов в незарегистрированный брак. В качестве переменных были предложены 6 вариантов, из которых респондент отмечал подходящие для себя:

1) переменная № 1 – Проверка чувств;
2) переменная № 2 – Возможность материальной независимости друг от друга;
3) переменная № 3 – Свобода сексуальных отношений;
4) переменная № 4 – Проверка бытовой совместимости;
5) переменная № 5 – Отсутствие процедуры развода;
6) переменная № 6 – Проверка схожести взглядов на дальнейшую жизнь.

Иерархический кластерный анализ наблюдений был проведен методом Уорда и квадрат Евклидовой в качестве меры измерения расстояния между наблюдениями. Анализ данных дендрограммы и таблицы шагов алгомирации показал, что оптимальным решением задачи будет два кластера.

Для описания полученных кластеров обратимся к анализу средних. В результате построения определили, что что в первой кластер попали респонденты, для которых мотивами вступления в незарегистрированный брак являются проверка чувств, проверка бытовой совместимости, проверка схожести взглядов на дальнейшую жизнь (у этих переменных высокое значение

среднее). Следовательно, данному кластеру можно дать название «репетиция семейной жизни». Для респондентов из этого кластера незарегистрированный брак является «промежуточным» этапом перед регистрацией брака.

Во второй кластер попали переменные возможность материальной независимости друг от друга, свобода сексуальных отношений, отсутствие процедуры развода. По таблице выше видно, что у данного кластера противоположное отношение на данный вид брака, в связи с этим дадим ему название «незарегистрированный брак, как постоянная форма сожительства». Респонденты данного кластера рассматривают такую форму брака приемлемой для себя и готовы прожить так всю жизнь.

Для дальнейшей интерпретации результатов построили таблицы сопряженности с переменными: пол, образование, доход в месяц.

Таким образом, в ходе проведения иерархического кластерного анализа были определены два кластера, распределяющие респондентов на две группы «репетиция семейной жизни» и «незарегистрированный брак, как постоянная форма сожительства». В первом кластере преобладают респонденты, получившие образование, имеющие постоянную работу и малый и средний заработок. Они уже имеют представления о дальнейшей жизни, планах, знают какие качества хотят видеть в партнере, поэтому вступают в такой вид брака для проверки бытовой совместимости, для обеспечения материальной стабильности, тем самым для них незарегистрированный брак является промежуточным этапом, для того чтобы «встать на ноги» перед официальные регистрации брака. Ко второму кластеру относятся студенты, на данный момент получающие образование, не имеющие постоянного заработка, находящиеся в таком возрасте, что могут меняться ценности, приоритеты и качества, которые хотят видеть в партнере, поэтому на данном этапе для них незарегистрированный брак является приемлемым, без дальнейшего оформления.

Во втором анализе рассмотрели вопрос отношения респондентов к незарегистрированному браку. В качестве переменных были предложены 9

вариантов, каждую из которых респонденты оценивали по 5-бальной шкале, где 1 – совсем не согласен, 5 – полностью согласен:

1) переменная № 1 – Вступая в незарегистрированный брак, можно проверить серьезность отношений;

2) переменная № 2 – Я отрицательно отношусь к незарегистрированному браку, так как в нем нет постоянных обязательств;

3) переменная № 3 – В незарегистрированном браке мне нравится то, что можно проверить свою совместимость с партнером;

4) переменная № 4 – Я считаю достоинством незарегистрированного брака отсутствие юридической ответственности;

5) переменная № 5 – Мне нравится незарегистрированный брак тем, что его легко можно разорвать;

6) переменная № 6 – Считаю, что на незарегистрированный брак партнеры соглашаются легче, чем на зарегистрированный;

7) переменная № 7 – Незарегистрированный брак подрывает выстраивание долгосрочных планов;

8) переменная № 8 – Долгое проживание в незарегистрированном браке способствует развитию разных планов на его продолжение;

9) переменная № 9 – В незарегистрированном браке нет крепости отношений.

Анализ был проведен также методом Уорда и квадрат Евклидовой в качестве меры измерения расстояния между наблюдениями. Анализ данных дендрограммы и таблицы шагов алгомирации показал, что оптимальным решением задачи будет два или 3 кластера.

Для определения точного количества кластеров обратились к анализу средних. 2-кластерное решение отличается не совсем равномерным распределением переменных по кластерам, также представляется довольно неоднозначным для интерпретации, так как 1 кластер включает в себя противоречивые по смыслу переменные: «Вступая в незарегистрированный брак, можно проверить серьезность отношений» и «Мне нравится

незарегистрированный брак тем, что его легко можно разорвать», «В незарегистрированном браке мне нравится то, что можно проверить свою совместимость с партнером» и «Я считаю достоинством незарегистрированного брака отсутствие юридической ответственности».

В свою очередь, 3-кластерное решение является более сбалансированным по распределению переменных между кластерами. При таком распределении кластеры хорошо поддают социологической интерпретации, поэтому остановимся на 3-кластерном распределение наблюдений.

Анализ средних 3-кластерого решения показал, в первой кластер попали респонденты, которые выбрали переменные «Проверка совместимости», «Позволяет проверить серьезность отношений», «Долгое проживание в незарегистрированном браке способствует развитию разных планов на его продолжение» (у этих переменных высокое значение среднее). Следовательно, данному кластеру можно дать название «репетиция семейной жизни». Для респондентов из этого кластера незарегистрированный брак является «промежуточным» этапом перед регистрацией брака.

Во второй кластер попали переменные «Достоинство незарегистрированного брака - отсутствие юридической ответственности», «незарегистрированный брак легко можно разорвать», «на незарегистрированный брак партнеры соглашаются легче, чем на зарегистрированный». Из таблицы выше видно, что у данного кластера отношение к данный виду брака также положительное, но смысл его для них совсем другой. В связи с этим дадим ему название «незарегистрированный брак, как постоянная форма сожительства». Респонденты данного кластера относятся к незарегистрированному браку положительно и рассматривают возможность прожить так всю жизнь.

Третий кластер можно охарактеризовать как негативное отношение к незарегистрированному браку. Респонденты из этого кластера выбирали переменные «Незарегистрированный брак подрывает выстраивание долгосрочных планов», «Я отрицательно отношусь к незарегистрированному

браку, так как в нем нет постоянных обязательств», «В незарегистрированном браке нет крепости отношений», что говорит об их негативном отношении к данному виду брака, так как возможно был неудачный опыт или есть друзья/знакомые с таким опытом. В связи с этим данному кластеру можно дать название «противники незарегистрированного брака».

Далее рассмотрели данные кластеры относительно переменных пол, образование и доход. В данном случае результат немного схож с результатом анализа мотивации респондентов при вступлении в незарегистрированный брак. В первом кластере преобладают респонденты обоих полов практически в равном соотношении, получившие образование, имеющие постоянную работу и малый и средний заработок. Они уже имеют представления о дальнейшей жизни, планах, знают какие качества хотят видеть в своем партнере, поэтому вступают в такой вид брака для проверки бытовой совместимости, для обеспечения материальной стабильности, тем самым для них незарегистрированный брак является промежуточным этапом, для того чтобы «встать на ноги» перед официальные регистрации брака.

Ко второму кластеру относятся студенты, на данный момент получающие образование, не имеющие постоянного заработка, находящиеся в таком возрасте, что могут меняться ценности, приоритеты и качества, которые хотят видеть в партнере, и респонденты уже, получившие образование, но пока не уверенные в своих намерениях и планах, поэтому на данном этапе для них незарегистрированный брак является приемлемым, без дальнейшего оформления. Так же если провести аналогию с предыдущем анализом мотивов респондентов, то видно, что в данном случае вместе со студентами в кластере есть молодые люди с окончившим образованием – это может говорить о том, что определенный процент респондентов изменили свои мнение и намерения относительно незарегистрированного брака, возможно из-за неудачного опыта или развода, могли понять, что не нет особый преимуществ и разницы между зарегистрированным и незарегистрированным браком, поэтому изменили свое мнение на этот счет.

В третий кластер вошел наименьший процент респондентов из всей выборки. В нем практически в равном соотношении присутствуют и парни, и девушки, имеют ежемесячный доход выше среднего и получившие высшее образование – это говорит о том, что данные респонденты уже успели социально адаптироваться, определились со своими планами и целями, имеют хороший доход и могут позволить себе сыграть пышную свадьбу, в данном возрасте уже могут задумываться о детях – поэтому рассматривают для себя сразу зарегистрированный брак. Что касается негативной отношения к незарегистрированному браку – это может быть воспитание родителей, которые не приемлют для себя данный вид брака, негативный опыт, личный или друзей/знакомых, по религиозным причинам и т.д. Но так как данный кластер является наименьший среди трех кластеров – это позволяет сделать вывод, что общее отношение молодежи города Саратова к данному виду брака положительное и наблюдается тенденция его распространения и популяризации.

ЗАКЛЮЧЕНИЕ

Программа статистической обработки данных SPSS является мощным инструментом анализа социологической информации. Она предлагает множество методов обработки данных. К числу популярных относится иерархический кластерный анализ. К его основным достоинствам относятся: отсутствие ограничений на нормальное классификацию переменных; возможность распределения в случаях отсутствия априорных сведений о классах; возможность применения к наблюдениям и к переменным.

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

Вместе с тем, у данного метода есть и слабые стороны. Иерархический кластерный анализ предоставляется простое и наглядное решение классификации переменных на кластеры, которое не показывает особенности взаимосвязей между самими переменными, из-за чего сложно поддаются

интерпретации. Кроме того, статистики рекомендуют обращаться к данному методу в случае небольшого числа переменных. Выходом из данного затруднения является одновременное обращение к другим функциям, как в нашем случае построение таблиц анализа средних и таблиц сопряженности.

В ходе авторского исследования были проведены два иерархических кластерных анализа мотивов респондентов при вступлении в незарегистрированный брак и их отношение к данному виду брака. Для того что определить группы молодежи города Саратова, придерживающихся разных мнений касаемо незарегистрированного брака, обратимся к иерархическому кластерному анализу наблюдений. Иерархический кластерный анализ наблюдений выполняет функцию распределения предоставленной выборки данных на группы или кластеры таким образом, чтобы каждый выбранный кластер состоял из похожих объектов, а объекты отличающихся кластеров имели существенные отличия, при этом перед началом кластеризации информация о количестве кластеров и их содержании неизвестна.

Иллюстрируя возможности SPSS на примере данных социологического исследования, посвященного незарегистрированному браку, с помощью иерархического кластерного анализа были выделены 2 группы респондентов, придерживавшихся разных мотивов при вступлении в незарегистрированный брак и 3 группы респондентов с различным отношением к данному виду брака.

При анализе мотивов респондентов была сформировано два кластера, получившие название «репетиция семейной жизни» и «незарегистрированный брак, как постоянная форма сожительства». Различиями данных кластеров является цели, которые преследуют респонденты в каждом из кластеров. Так в первом кластере молодые люди рассматривают незарегистрированный брак как промежуточный этап перед регистрацией брака, для проверки бытовой совместимости и получения материальной стабильности в том числе для проведения свадьбы. Во втором кластере преобладают в основном студенты, не имеющий пока для четких планов, стабильного заработка, поэтому для них

незарегистрированный брак на текущий момент конечная точка, о дальнейших планах на регистрацию брака они пока не думают.

Результатом анализа отношения молодежи к незарегистрированному браку получились 3 кластера «репетиция семейной жизни», «незарегистрированный брак, как постоянная форма сожительства», «противники незарегистрированного брака». Первый и второй кластер существенно превосходят по количеству респондентов относительно третьего кластера – это говорит о том, что тенденция на незарегистрированный брак растет, данный вид брака становится все более популярным. Единственное, что молодежь, как и в первом анализе, закладывает в незарегистрированный брак разные понятие и цели, так в первом кластере люди рассматривают незарегистрированный как репетицию и промежуточный этап, а во втором кластере – рассматривают его как постоянную форму сожительства, готовы прожить так всю жизнь, либо не задумывают пока об официальной регистрации.

Следовательно, как было описано выше, иерархический кластерный анализ является мощным методом анализа социологической информации, который дает возможность получить интересные результаты. Обойти указанные недостатки, а также полностью показать возможности иерархического кластерного анализа позволяет его применение в сочетании с другими аналитическими методами, в нашем случае это были анализ средний и таблицы сопряженности. Программное обеспечение SPSS, содержащая в своем составе широчайший спектр аналитических инструментов, заслуженно занимает свое месте в ТОПе программ в своей отрасли.