Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра социальной информатики

ВОЗМОЖНОСТИ ЭМПИРИЧЕСКОЙ ДИФФЕРЕНЦИАЦИИ СОВОКУПНОСТИ РЕСПОНДЕНТОВ (НА БАЗЕ ПРОГРАММЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ SPSS)

(автореферат бакалаврской работы)

Студента 5 курса 531 группы
направления 090303 – Прикладная информатика,
профиль – Прикладная информатика в социологии
социологического факультета
Богова Алексея Александровича

Научный руководитель
Кандидат философских наук, доцент А.И. Завгородный

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

Саратов, 2017 год

Введение. Согласно истории развития эмпирической социологии, в середине XX столетия наступил расцвет массовых социологических исследований. Отечественный социолог Н.И. Лапин отмечает, что первые десятилетия после Второй мировой войны являются периодом развитых эмпирических 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.

Объектом данного исследования являются методы кластерного анализа случаев и дискриминантного анализа; *предметом* выступает эмпирическая дифференциация выборочной совокупности.

Цель исследования – выявить потенциал кластерного анализа случаев и дискриминантного анализа в решении задачи эмпирической дифференциации выборочной совокупности. Постановка цели определила формулирование следующих *задач* исследования:

1. Охарактеризовать процесс формирования выборочной совокупности в массовом опросе;
2. Определить кластерный анализ случаев и дискриминантный анализ в качестве инструментов статистического анализа эмпирической информации;
3. Рассмотреть возможности кластерного анализа случаев и дискриминантного анализа применительно к эмпирической дифференциации однородной выборочной совокупности.

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.

В качестве *эмпирической базы* исследования были использованы данные массового социологического опроса, проведенного студентами социологического факультета СГУ по теме «Специфика досуга в малом городе (на примере г. Петровска)»[1].

[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 14 лет и старше). Всего было опрошено 150 респондентов.

Структура диплома. Данная работа состоит из введения, трех разделов (1 раздел «Особенности формирования выборочной совокупности в массовом опросе», 2 раздел «Кластерный анализ случаев и дискриминантный анализ как методы дифференциации выборочной совокупности», 3 раздел «Эмпирический пример применения кластерного анализа случаев и дискриминантного анализа»), заключения, списка использованных источников и приложения.

Основное содержание работы. В первом разделе «Особенности формирования выборочной совокупности в массовом опросе» рассматриваются основные принципы и технологии применения выборочного метода с точки зрения особенностей структурирования генеральной и выборочной совокупностей, разбиения их на группы в соответствии с определенными признаками.

История развития массовых статистических обследований населения насчитывает около трехсот лет. Однако расцвет опросного метода, использующего принцип выборки, наступил лишь в 30-40 гг. XX века в США, и связан он был с именем Дж. Гэллапа. Именно тогда были сформулированы основные критерии массового опроса, среди которых были обозначены такие как случайный или квотный характер выборки и сбор индивидуальных данных, когда каждое наблюдение может быть соотнесено с конкретным индивидуумом в выборке.

Идея выборочного метода состоит в том, чтобы в процессе обследования охватить только часть элементов генеральной совокупности таким образом, чтобы характеристики людей, попавших в выборочную совокупность, максимально полно повторяли характеристики людей в генеральной совокупности. Наиболее последовательно данный принцип реализован собственно случайной или равновероятностной выборкой, однако данный тип выборки на практике встречается крайне редко. Это объясняется рядом трудностей, с которыми сталкивается исследователь при попытке использовать случайную выборку, которые часто являются труднопреодолимыми. В таком случае исследователи обращаются к другим типам выборок: с введением

элементов неслучайности, лишь частично решающих недостатки случайной выборки, и метод неслучайного отбора, который полностью решает эти проблемы, и этим можно объяснить его распространенность. Его главным достоинством является экономия ресурсов в самом широком смысле слова.

При использовании *квотной выборки* отбирают один или несколько признаков, по которым будет контролироваться выборка. Количество единиц в выборке, обладающих определенными характеристиками, должно быть пропорционально количеству таких единиц в генеральной совокупности. Выбранные признаки должны быть независимыми и тесно связанными с изучаемыми характеристиками. Чаще всего в качестве таких признаков исследователи используют социально-демографические характеристики, так как они часто носят ключевой характер, и информация о распределении их в генеральной совокупности является доступной. Как правило, используют не более трех – четырех признаков, так как при увеличении их числа растет число ограничений и, соответственно, растут затраты на поиск респондентов.

Таким образом, проведение массовых опросов по стандартизированной анкете, особенно выборочным методом по принципу неслучайности, структурирует выборочную совокупность согласно социально-демографическим признакам, достаточно традиционным для статистики. Что оставляет открытым вопрос дифференцирования опрошенных согласно менее привычным, латентным переменным.

Второй раздел «Кластерный анализ случаев и дискриминантный анализ как методы дифференциации выборочной совокупности» посвящен обзору основных методов статистического анализа, позволяющих дифференцировать выборочную совокупность с помощью статистико-математических методов.

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

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.

В третьем разделе «Эмпирический пример применения кластерного анализа случаев и дискриминантного анализа» описывается пример использования методов кластерный анализ случаев и дискриминантный анализ с целью дифференцировать совокупность респондентов по достаточно нестандартному для массовых социологических опросов признаку – особенностям организации и проведения досуга.

В анализе было использовано 9 переменных, каждая из которых имела одинаковую дихотомическую шкалу, содержащую варианты ответа «да» и «нет», закодированных кодами 0 и 1. Такая кодировка позволяет рассматривать данные переменные как обладающие характеристиками условно «количественных» шкал, что соответствует требованиям, предъявляемым к переменным со стороны кластерного и дискриминантного анализа.

Для проведения кластерного анализа случаев в качестве метода кластеризации был выбран метод Варда, а 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. Для этого требуется оценить динамику увеличения различий по шагам кластеризации и выявить шаг, на котором отмечается резкое возрастание различий. В рассматриваемом примере число кластеров оказалось равно 4.

Для проведения дискриминантного анализа был выбран метод пошагового введения независимых переменных в уравнение, из пошаговых методов был выбран метод Уилкса, основанный на минимизации коэффициента Уилкса после включения в уравнение регрессии каждого нового предиктора.

Поскольку в результате проведенного кластерного анализа случаев были получены 4 группы респондентов, предпочитающих свой специфический способ организации досуга, то в ходе дискриминантного анализа были вычислены 3 функции, позволяющие сравнить и выявить различия между данными группами.

В первой функции, названной нами *«Домашний досуг»*, доминируют переменные «Наличие детей» (связанная с функцией обратной корреляционной связью), «Наличие интернета», «Занятие домашними делами» и «Наличие хобби» (связанные прямой корреляцией). Вторая функция *«Активный досуг»* продемонстрировала сильную корреляционную связь с переменными «Наличие хобби», «Занятие спортом», «Наличие детей» (связаны прямой корреляционной связью) и «Отдых дома» (обратная корреляция). Третья функция *«Коммуникативный досуг»* характеризуется прямой корреляционной связью с переменными «Отдых дома», «Посвящение свободного времени близким, друзьям, родным», «Посещение библиотеки» и «Выезд на природу».

Для первой группы респондентов отличительными признаками являются большое положительное значение функции «Домашний досуг», большое, но отрицательное значение функции «Активный досуг» и незначительное влияние со стороны функции «Коммуникативный досуг», поэтому она была определена как *«домоседы»*. Вторая группа респондентов была охарактеризована большими положительными значениями функций «Активный досуг» и «Домашний досуг» и также положительным, но небольшим значением функции «Коммуникативный досуг» - *«активисты»*. Третья группа опрошенных, напротив, отличилась большими отрицательными значениям функций «Домашний досуг» и «Коммуникативный досуг» - *«инертные»*. Четвертую группу отличают высокие значения всех трех функций, причем функции «Домашний досуг» и «Активный досуг» получили отрицательные значения, а функция «Коммуникативный досуг» - положительное - «*коммуникаторы»*.

Совместное использование двух методов, кластерного анализа случаев и дискриминантного анализа, позволило разработать эффективную модель, дающую точность прогноза в 93%.

Заключение. Для современных обществ проведение массовых обследований населения, дающих объективную информацию о состоянии социума, является привычным и хорошо знакомым явлением. Этому способствует широкое использование выборочного метода, обеспечивающего репрезентативность получаемых данных. В распоряжении исследователя находится много методов формирования выборки, как основанных на принципе равновероятностного отбора, так и относящихся к неслучайным методам. Тем не менее, все рассмотренные методы носят априорный характер, т.е. выборка формируется и далее в ходе анализа дифференцируется по признакам, легко фиксируемым визуально и, как правило, отслеживаемым органами государственной статистики. Однако исследователь нередко сталкивается с ситуацией, когда ему надо структурировать выборку по иным, более трудно фиксируемым признакам. Задача многократно усложняется, когда надо использовать не один-два, а множество признаков одновременно. Здесь на помощь исследователю приходят специализированные программы статистической обработки информации, обладающие большим арсеналом сложных математических методов работы с информацией, способных справиться с поставленной задачей.

Одними из наиболее эффективных методов, направленных на дифференциацию выборочной совокупности, являются кластерный анализ случаев и дискриминантный анализ. Не смотря на то, что решают они одну задачу, принципы их работы серьезно отличаются друг от друга. Так, кластерный анализ случаев является набором многомерных статистических методов, нацеленных на исследование структуры некоторой совокупности объектов, и 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. При этом изначально информация о числе кластеров и их составе неизвестна.

В целом 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, 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 особенностям организации досуга в малом городе, с помощью кластерного анализа были выделены 4 группы респондентов, реализующих различные практики организации своего досуга.

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

группе, и получившие названия «Домашний досуг», «Активный досуг» и «Коммуникативный досуг».

Анализ значений центроидов групп позволил дать характеристику выделенным группам. Первая группа опрошенных получила название «домоседов» (для них оказалось свойственным большое положительное значение функции «Домашний досуг» и большое отрицательное значение функции «Активный досуг»). Вторая группа была названа «активистами» (здесь были отмечены большие положительные значения функций «Активный досуг» и «Домашний досуг»). Третью группу опрошенных или «инертных» отличали большие отрицательные значения функций «Домашний досуг» и «Коммуникативный досуг». Наконец, для четвертой группы или «коммуникаторов» было характерно высокое положительное значение функции «Коммуникативный досуг» и также высокие, но отрицательные значения двух других функций.

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 девятью 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%.