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

ОПЕРАЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ ПРОГРАММЫ SPSS ПО ОБРАБОТКЕ СОЦИОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

АВТОРЕФЕРАТ ДИПЛОМНОЙ РАБОТЫ

студентки 6 курса 631 группы
специальности 080801.65 - Прикладная информатика в социологии
Социологического факультета
Хабаровой Марии Александровны

Научный руководитель
Кандидат социологических наук, доцент Н.Ю. Кравченко

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

Саратов, 2016 год

Введение. Процесс работы с первичной социологической информацией предполагает несколько этапов ее трансформации. На первом этапе исследователь сосредоточен на так называемом исследовательском (или программном) вопросе, который непосредственно указывает на исследуемую проблему. Задачей исследователя на этом этапе является «перевод» программного вопроса на язык анкеты, который должен быть, с одной стороны, понятен и удобен респонденту, а с другой – обеспечивать точной и объективной информацией. Второй этап включает в себя «обратный перевод» языка анкеты на язык исследователя, который реализуется посредством анализа и интерпретации полученных данных. Однако между ними находится еще один, не менее важный этап, о котором практически не упоминается в социологической литературе, но который также предполагает значительную трансформацию социологических данных – работа с информацией в рамках электронной базы. Содержание данной работы состоит в изменении параметров информации таким образом, чтобы она была максимально адаптирована для математической обработки на следующем этапе социологического исследования.

Например, для исследователя может представлять интерес среднее значение или сумма баллов по нескольким переменным для каждого респондента. Иногда желательно упорядочить данные файла по какому-либо признаку. Нередко возникает необходимость трансформации шкалы переменной из одного типа в другой. Также может потребоваться обработка не всех данных файла, а только их части, выделенной по каким-либо параметрам (например, полу, возрасту, успеваемости и т.д.) Далеко не полный список перечисленных проблем указывает на то, что для регулярной аналитической работы недостаточно умения вводить данные и применять к ним статистические процедуры. Возникает задача эффективного управления данными.

Вместе с тем, как один из целого ряда этапов проведения социологического исследования, этап трансформации данных также вносит

свою лепту в приращение, сохранение или потерю качества данных. С одной стороны, активная трансформация собранных данных позволяет исследователю находить глубокие и невидимые для поверхностного взгляда связи и закономерности. При этом возможности математической переконфигурации данных практически безграничны. С другой же стороны, возникает вопрос: насколько исследователь может быть уверен в том, что те данные, которые он получил после трансформации, продолжают оставаться репрезентативными по отношению к изучаемой социальной реальности? Подобная постановка вопроса позволяет рассматривать выбранную тему исследования как актуальную не только с точки зрения компьютерных наук, но и социальных.

Объектом нашего исследования является программа SPSS, **предметом** – операции модификации и отбора социологических данных.

Цель работы – продемонстрировать возможности использования программы SPSS для оптимизации социологических баз данных, полученных в ходе сбора первичной социологической информации.

Выбор объекта, предмета и цели исследования определил формулирование следующих **задач**:

1. описать процедуры модификации и отбора данных, предлагаемых программой SPSS;

2. представить примеры возможных модификаций и отбора социологических данных.

Методологическая база исследования. Работа выстроена в традициях позитивизма О. Конта, ориентированного на экспериментальное познание социальной жизни. Также используются принципы сравнительного анализа, элементы системного подхода.

Эмпирическая база. В работе использовались база данных, созданная в процессе проведения авторского исследования[1].

[1] База данных была создана в ходе авторского исследования, проведенного методом анкетирования в 2014 году. Объектом исследования выступали предприниматели г. Энгельса. Выборочная совокупность формировалась по принципу целевой и составила 105 респондентов (текст анкеты см. в приложении А).

Структура диплома. Дипломная работа состоит из введения, первого раздела «Потенциал статистического пакета SPSS в модификации и отборе данных» и второго раздела «Операции модификации и отбора социологических данных», заключения, списка использованных источников и приложения.

Основное содержание работы. Первый раздел «Потенциал статистического пакета SPSS в модификации и отборе данных» посвящен инструментам трансформации данных, встроенным в программу SPSS. Часть инструментов связаны с создаем новых переменных через модификацию уже имеющихся. В процессе модификации получаются агрегированные данные. Другая часть инструментов ориентирована на работу с объектами и подразумевает разбиение всей совокупности объектов на отдельные подвыборки.

К числу важнейших инструментов относятся: вычисление новых переменных путем; подсчет частоты появлений определенных значений; перекодирование значений; агрегирование данных; ранговые преобразования; упорядочение объектов в соответствии с определенными критериями; расщепление файла; выбор подмножества объектов для дальнейшего анализа и вычисление весов наблюдений.

Путем арифметических вычислений в SPSS можно образовать новые переменные и добавить их в файл данных. Как и прочие переменные, их можно применять как для аналитических целей, так и для дальнейшей трансформации данных.

Для построения численных выражений применяются такие арифметические операторы, как сложение, вычитание, умножение, деление, возведение в степень. Также в арифметических выражениях могут участвовать не только переменные и константы, но и встроенные в SPSS функции.

В SPSS есть возможность подсчитать число появлений одного и того же значения или значений для определенной переменной. Например, если взять множественный вопрос об имеющихся источниках дохода и во всех

наблюдениях подсчитать число появлений значения 1 (= да), то для каждого респондента мы получим количество источников дохода, которые у него есть.

Первоначально собранные данные можно перекодировать. Перекодирование численных данных необходимо, например, когда первоначальное разнообразие исходных данных не нужно для последующего анализа. В этом случае перекодирование означает уменьшение объема обрабатываемой информации. Перекодирование данных можно выполнить вручную или автоматически. Кроме того, перекодированные значения можно хранить в той же переменной или перенести их в другую переменную.

На базе значений одной или нескольких группирующих переменных (переменных разбиения) можно объединить наблюдения в группы (агрегировать) и создать новый файл данных, содержащий по одному наблюдению для каждой группы разбиения. Для этого SPSS предоставляет большое количество функций агрегирования. По умолчанию в качестве функции агрегирования принято среднее значение. Вместо него можно выбрать и другие функции агрегирования (всего в SPSS встроено 16 функций).

В SPSS существует возможность задавать ранги для измеренных значений переменной, проводить оценки Сэвиджа, вычислять процентные ранги и формировать процентильные группы, добавляя в файл данных соответствующие переменные. Программа SPSS включает также процедуры ручного ранжирования, результаты которого часто представляют самостоятельную ценность для исследователя.

С помощью сортировки объектов можно расположить данные файла в таком порядке, который наиболее удобен исследователю на данном этапе исследования. При определении параметров сортировки можно пользоваться сразу несколькими переменными.

Использование процедуры расщепления файла дает возможность анализировать данные, собранные в электронной базе, раздельно по группам.

Выбор объектов для анализа позволяет исследователю отобрать для аналитической работы не все данные файла, а лишь его часть, удовлетворяющую определенным условиям.

- Наконец, SPSS предоставляет возможность определения веса данных. При этом данным, относящимся к разным наблюдениям, присваиваются различные весовые коэффициенты посредством так называемой переменной взвешивания. Эта процедура является очень полезной, когда выборка не является репрезентативной, то есть частотные характеристики выборки по важным для обеспечения репрезентативности переменным не соответствуют частотным характеристикам генеральной совокупности, а также для анализа данных, уже представленных в виде частотных таблиц.

Во втором разделе «Операции модификации и отбора социологических данных» представлены примеры практической реализации инструментов трансформации данных, также анализируются последствия данной трансформации для результатов исследования. Всего рассматривается пять процедур: вычисление новой переменной, перекодировка данных, расщепление файла, выбор объектов и вычисление весов наблюдений.

Для вычисления новой переменной «Степень выраженности либеральности взглядов предпринимателей» через вычисление среднего арифметического были выбраны переменные, характеризующие либеральность взглядов опрошенных предпринимателей: «Мое материальное положение в настоящем и в будущем зависит, прежде всего, от меня», «Для достижения успеха в жизни надо рисковать, это дает шанс», «Материальных успехов люди должны добиваться сами, а те, кто этого не хочет, пусть живут бедно – это справедливо», «Свобода – то, без чего жизнь человека теряет смысл» и «Современный мир жесток, чтобы выжить и преуспеть, необходимо бороться за свое место в нем, а то и переступить через некоторые нормы морали». Варианты ответов в данном вопросе были построены по принципу 4-балльной ранговой шкалы, где 1 балл соответствует полному согласию, а 4 балла – полному несогласию.

В результате было выявлено, что уровень дохода не влияет на степень выраженности либеральных ценностей и установок у предпринимателей. Кроме того, создание новой агрегированной переменной не изменило ранее полученные результаты.

С помощью команды перекодировки была трансформирована переменная «Стаж предпринимательской деятельности»: шкалу отношений преобразовали в номинальную шкалу. Однако результаты корреляционного анализа, проведенного с переменной «Стаж предпринимательской деятельности» (в обоих форматах) и переменными, измеряющих степень выраженности либеральных ценностей, показал, что степень выраженности либеральных настроений не связан с длительностью занятия предпринимательством. Данный вывод был подтвержден и корреляцией с агрегированной переменной «Степень выраженности либеральных ценностей предпринимателей».

Процедура взвешивания весов наблюдений была применена для корректировки параметров выборочной совокупности по такому параметру, как уровень образования. В ходе авторского исследования было опрошено 44,8% респондентов без высшего образования и 55,2% - с высшим образованием. Однако данная пропорция не выдерживает принцип репрезентативности, поскольку, согласно статистическим данным, в 2014 году среди российских предпринимателей 73,8% не имели высшего образования, тогда как 26,2% были с высшим образованием. Представленные данные послужили исходным материалом для задания пропорций для расчета соответствующих весовых коэффициентов. В итоге процедура взвешивания наблюдений и ремонта выборки отразилась на результатах анализа взаимосвязи между уровнем образования респондентов и степенью выраженности либеральных ценностей методом Хи-квадрата и привела к менее выраженной статистической взаимосвязи между рассматриваемыми признаками.

Таким образом, рассмотрение ряда ключевых переменных привело нас к выводу о том, что формирование либеральных ценностей и установок не коррелирует с развитием предпринимательства. Данный вывод подтолкнул нас

к выдвижению новой гипотезы о том, что российское предпринимательство является крайне разнородной социальной группой, поэтому линейной связи между успешностью предпринимательской деятельности и степенью выраженности либеральных установок и не выявляется. Для продолжения исследования взаимосвязей, теперь уже нелинейного характера, между интересующими нас переменными были привлечены другие способы трансформации данных – расщепление файла и выбор объектов.

Процедура расщепления файла позволила выявить группы предпринимателей, являющиеся наиболее выраженными носителями либеральных ценностей и установок. Корреляционные матрицы с максимальным числом значимых связей (со значимостью не ниже 0,1), охватывающих практически все переменные-индикаторы либеральных ценностей, были получены для групп предпринимателей в возрасте от 31 до 40 лет, с высшим образованием, стажем предпринимательской деятельности не более 10 лет и уровнем личного дохода не выше 50 000 рублей, но при этом с потребительскими возможностями не ниже среднего.

Процедура выбора объекта позволила нам выделить искомую подвыборку с помощью переменных «Стаж предпринимательской деятельности» и «Личный ежемесячный доход». Всего в данную группу попали 39 опрошенных или 37% от выборочной совокупности.

Использование фильтра в качестве переменной, разделяющей выборочную совокупность на либерально ориентированных и либерально неориентированных предпринимателей позволило нам сформировать социальный портрет либеральной части предпринимательства г. Энгельса: это женщины и молодежь с высшим образованием, еще не создавшая собственной семьи, стажем предпринимательской деятельности до 10 лет и уровнем личного дохода не более 50 000 рублей в месяц, но при этом с потребительскими возможностями среднего уровня.

Заключение. В результате краткого обзора процедур модификации данных, предусмотренных в программе SPSS, мы смогли убедиться в широте ее

математических возможностей. Вместе с тем, остаются некоторые пробелы, которые ограничивают возможности исследователя. В первую очередь, они связаны с типом шкалы модифицируемой переменной. Например, вычисления новых переменных и подсчет частоты появления определенных значений имеют серьезные ограничения в работе со строковыми переменными. При помощи механизма перекодировки данных можно двигаться только по направлению укрупнения делений шкалы, но не наоборот и так далее.

Однако еще более острой является проблема социологического характера. Весь цикл проведения социологического исследования ориентирован не только на сбор первичной социологической информации, но и на обеспечение ее качества. Механизмы модификации исходных эмпирических данных, являясь одним из этапов математической обработки информации, изменяют ее характеристики и тем самым порождают дополнительные риски утраты достоверности собранных данных, что для исследователей является неприемлемым.

В ходе нашей работы мы рассмотрели несколько примеров модификации и отбора социологических данных как способы изменения параметров взаимосвязи между исследуемыми переменными: вычисление новой переменной, перекодировка данных, вычисление весов наблюдений, расщепление файла и выбор объектов.

Сравнение результатов корреляционного анализа с исходными переменными выраженности либеральных ценностей и установок и агрегированной переменной показало, что модификация переменных к значительным изменениям в структуре данных не привела. И в первом, и во втором случае мы фиксируем, что уровень личного дохода не оказывает на степень выраженности либеральных ценностей статистически значимого влияния. Таким образом, мы получили обобщенный и доказательный вывод о том, что с ростом уровня личного дохода степень выраженности либеральных настроений в российском предпринимательстве не увеличивается.

Во втором случае модификации данных в виде перекодировки шкал структура социологических данных также не изменилась. Так, анализ взаимосвязи стажа предпринимательской деятельности, измеренного шкалой отношений, показал, что с его увеличением степень актуализации либеральных ценностей и установок не увеличивается. Анализ взаимосвязи стажа предпринимательства, измеренного при помощи дихотомической шкалы, привел нас к аналогичным выводам.

Третий случай модификации данных в виде процедуры взвешивания наблюдений продемонстрировал возможности программы SPSS выявлять и исправлять некоторые ошибки, допущенные на предыдущих этапах социологического исследования, в частности – ошибки репрезентативности выборки. Так, нерепрезентативное распределение респондентов по уровню образования указывало на тенденцию роста либеральных настроений с увеличением уровня образования. В результате же ремонта выборки при помощи математической процедуры взвешивания мы обнаружили ослабление зависимости между уровнем образования предпринимателей и степенью выраженности у них либеральных ценностей и установок.

Если использование инструментов модификации переменных по позволило нам выстроить линейную зависимость между предпринимательством и либеральными ценностями и установками, то обращение к инструментам отбора данных помогло обнаружить связи нелинейного характера. Так, с помощью команды расщепления мы разбили всю выборочную совокупность предпринимателей на отдельные подвыборки и обнаружили среди них собственно носителей либеральных ценностей и установок. Исследование этой части респондентов показало, что среди них преобладают женщины и молодежь, люди несемейные и с высшим образованием, стажем предпринимательской деятельности до 10 лет и уровнем личного дохода не более 50 000 рублей в месяц, но при этом с потребительскими возможностями среднего уровня. Интересно отметить, что по переменным, характеризующим собственно предпринимательскую

деятельность, значимые различия между либерально ориентированным и либерально неориентированным предпринимательством выявлены не были.

Таким образом, статистический пакет SPSS является сложным инструментом анализа социологических данных, неосторожное использование которого с целью трансформации данных может привести к риску появления дополнительных ошибок и, как следствие, к ухудшению качества данных. С другой стороны, программа является эффективным инструментом поиска глубоких взаимосвязей между исследуемыми переменными, а также обеспечения контроля качества данных и исправления некоторых видов ошибок, допущенных на предыдущих этапах проведения социологического исследования, причем делает это с минимальными затратами ресурсов.

Для того, чтобы снизить риск потери качества данных в процессе их модификации, мы можем рекомендовать искать подтверждения выявленной взаимосвязи переменных, обращаясь сразу к нескольким методам анализа, предоставляемых SPSS. Особую бдительность в проверке сохранения качества социологических данных надо проявлять, когда в ходе модификации были совершены кардинальные изменения.