

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

**ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ОБРАБОТКИ
СОЦИОЛОГИЧЕСКОЙ СТАТИСТИКИ:
СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПАКЕТОВ SPSS И SAS**

(автореферат бакалаврской работы)

студентки 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Лейкман Юлии Александровны

Научный руководитель

кандидат философских наук, доцент _____ А.И. Завгородный
подпись, дата

Зав. кафедрой

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

Саратов 2018

Введение. Актуальность проблемы. Точная и своевременная информация о том, что может произойти в экономике и обществе в будущем, всегда была важна для лиц, принимающих решения. Прогнозирование стало важной частью процесса планирования стратегии любой компании и любой государственной политики. Развитие современных экономических, социологических теорий, а также сложных компьютерных программ повлияло на появление новых методов прогнозирования и анализа.

В современных условиях, когда информационные потоки стали особенно массивными, существует огромный объем данных. Кроме того, из-за того, что скорость всех процессов в обществе возрастает, необходимо определять быстрые ответы на возникающие вопросы, чтобы искать и проводить огромный объем исследований и сбор данных во всех сферах жизни.

Сегодня в нашей стране более широко распространены конкретные социологические исследования, которые являются источником социальной информации, необходимой для успешного решения важных социально-экономических проблем, для научного управления социальными процессами и явлениями. Эффективный и четкий анализ полученной информации фактически невозможно для человека из-за его объема и сложности и требует новых методов их решения. В этих критериях автоматизация различных областей человеческой жизни посредством внедрения компьютерных технологий повлияла на процесс классификации. Основой этого процесса является идея использования математических методов для сбора, сортировки и классификации объектов.

Актуальность выбора соответствующих инструментов для обработки социологических данных обусловлена тем, что проведение любых исследований, которые включают кластерный анализ данных, будь то психологические, медицинские исследования, научные исследования, а также социологические исследования и исследования, выбор инструментов

для их обработки, диагностики и прогнозирования очень важен для принятия управленческих решений, на первый план выходит набор скорости, точности и удобства при анализе результатов. Сравнительному анализу предполагается подвергнуть два пакета программ. SPSS - это модульная программа для обработки статистических данных. Его основой является базовый модуль базы SPSS Base –он дает нам возможность управлять данными и содержит наиболее общий статистический анализ данных: ведение описательной статистики; построение линейных и нелинейных моделей; реализация; конверсия данных; фактор, кластер, дисперсионный анализ; расчет корреляций; диаграммы; подготовка отчетов и другие функции. Дополнительные модули пакета могут быть установлены для проведения расширенного и углубленного анализа данных.

В качестве альтернативы был выбран программный продукт SAS (разработанный SAS Institute Inc.). Это интегрированный компонент системы SAS, разработанный специально для идентификации информации, которая необходима для принятия решений в огромных наборах данных.

Предназначен для поиска и анализа глубоко скрытых шаблонов в данных SAS. Пакет SAS (система статистического анализа) представляет собой профессиональный статистический пакет, разработанный SAS Institute Inc. Первая версия пакета была выпущена в 1976 году. Последняя версия SAS Statistics позиционируется как статистический пакет для аналитической работы на современных предприятиях.

Пакет работает эффективно с различными типами данных: маркетинговые базы данных, результаты клинических испытаний, обзоры состояния здоровья, исследования, предпочтения потребителей, исследования рынка, ценные бумаги и т. Д. Встроенные инструменты статистического анализа могут применяться для решения широкого круга вопросов. Пакет позволяет работать со следующими данными: ввод,

восстановление, изменение и управление данными; написание отчетов и подготовка графиков; статистический анализ; бизнес-планирование, прогнозирование и поддержка принятия решений; исследования операций и управления проектами; улучшение качества; разработка приложения. **Степень изученности** этой проблемы довольно низкая из-за её специфики и узкой направленности.

Целью данной работы является сравнение эвристического потенциала специальных программных пакетов для обработки социологических данных SPSS и SAS с точки зрения возможностей кластерного анализа.

Цель работы, сформулированная подобным образом предполагает решение следующих задач:

1. Рассмотреть специальные пакеты обработки статистических данных.
2. Провести кластерный анализ на примере авторского исследования в программе SPSS.
3. Провести кластерный анализ на примере авторского исследования в программе SAS.
4. Сравнить полученные результаты.

Объектом анализа являются программы обработки статистических данных, а именно два сравниваемых пакета программного обеспечения - SPSS и SAS.

Предметом являются функциональные и аналитические характеристики представленных программных пакетов, обновленных в рабочем режиме.

В качестве основы социологических данных для сравнения эвристического потенциала двух программных пакетов - SPSS и SAS - было принято решение использовать данные, собранные в ходе моего исследования - изучения особенностей принятия независимых решений современными молодыми людьми в 2016 году. На основании темы, выбранной для исследования, объектом исследования являлась молодежь

города Саратова в возрасте от 14 до 29 лет, предметом исследования были особенности принятия независимых решений.

Эмпирической основой исследования являются: пакеты программ, базы данных.

Методологической базой исследования является: сравнительный анализ, структурно функциональный анализ.

Практическая значимость исследования заключается в возможности использования основных положений и выводов данной работы для дальнейшего изучения выбранной проблематики.

Структура выпускной бакалаврской работы представлена введением, двумя главами, заключением, библиографией и приложением.

Основное содержание работы. В первой главе «Современные программные пакеты для обработки статистических данных» приводится классификация

различных программных пакетов для обработки данных и основные сведения о них. Отрасль разработки подобного программного обеспечения развивается стремительными темпами. На сегодняшний день на рынке представлено около тысячи компьютерных программ для статистической обработки данных (далее – статистические пакеты). Разнообразие статистических пакетов обусловлено многоплановостью задач обработки данных с применением различных типов статистических процедур анализа для поиска ответов на вопросы из различных областей человеческой деятельности.

Большинство представленных на рынке статистических пакетов обладают гибкой модульной структурой, которая может пополняться и расширяться за счет пользовательских модулей, дополнительнокупаемых или находящихся в свободном доступе в Интернете. Подобная гибкость позволяет адаптировать большинство пакетов к потребностям конкретного

пользователя.

Статистический пакет должен удовлетворять следующему минимальному набору требований:

- ❖ Модульность;
- ❖ Ассистирование при выборе способа обработки данных;
- ❖ Использование простого проблемно-ориентированного языка для формулировки задания пользователя;
- ❖ Автоматическая организация процесса обработки данных;
- ❖ Ведение банка данных пользователя и составление отчета о результатах проделанного анализа;
- ❖ Диалоговый режим работы пользователя с пакетом;
- ❖ Совместимость с другим программным обеспечением.

Перед пользователями различных категорий встает вопрос выбора оптимального статистического пакета для поиска верных ответов на вопросы существующие. Очевидно, что оптимальным является вариант, сочетающий в себе необходимые функциональные возможности, высокое качество и умеренную цену работы. Однако, создатели всех программных статистических пакетов заявляют, что их продукт превосходит аналоги. Отсутствие у большинства исследователей времени для освоения нескольких программ, а также недостаток хорошо структурированной и легкодоступной обывателю информации делает непростым ее выбор.

При выборе пакета, как правило, учитываются следующие параметры:

- ❖ соответствие характеру решаемых задач;
- ❖ объем обрабатываемых данных;
- ❖ требования, предъявляемые к квалификации пользователя (уровень знаний в области статистики);
- ❖ имеющееся в наличии компьютерное оборудование.

Все программы статистической обработки данных по признаку

функциональности можно разделить на профессиональные, универсальные - назовем их «популярными» - и специализированные. Статистические программы относятся к наукоемкому программному обеспечению, цена их часто недоступна индивидуальному пользователю. Профессиональные пакеты имеют большое количество методов анализа, популярные пакеты - количество функций, достаточное для универсального применения. Специализированные же пакеты ориентированы на какую-либо узкую область анализа данных.

Вторая глава «Сравнение программных пакетов на примере социологического исследования» посвящена определению набора данных для сравнения программных пакетов и практическим сравнение двух программных пакетов-SPSS и SAS Применительно к социологическим исследованиям данные представлены в таблицах, содержащих два типа данных: постоянная часть, как название таблицы, имена строк и столбцов и переменная часть - это фактически индикаторы таблицы (матрицы), которые непосредственно собираются в ходе социологического данные исследований. При сравнении пакетов по критерию загрузки, удобство данные работы, получены следующие результаты:SPSS очень позволяет удобно импортировать данные, подготовка данных для работы в SAS-одна из самых больших проблем для новичка. Сама обработка данных в SAS он подробно описана в различных руководствах

При сравнении программных пакетов, были получены следующие результаты:

Пакет SPSS

1. SPSS-это программное обеспечение, которое используется в статистическом анализе. Это аббревиатура для статистического пакета для социальных наук. Он также используется исследователями рынка, исследователями здоровья, компаниями обзора, правительством,

исследователями образования, организациями маркетинга и т.д. SPSS был разработан для социальных наук и был первым языком статистического программирования для ПК. Он был разработан в 1968 году в Стэнфордском университете и спустя восемь лет компания SPSS Inc. была основана, в дальнейшем она была куплена IBM в 2009 году.

2. SPSS-это удивительный инструмент для людей, которые не являются статистиками. Он имеет удобный интерфейс и имеет простое в использовании меню. Главным образом SPSS играет важную роль в социальных науках.
3. SPSS может использоваться, когда данные меньше, чем 100 Мб. Он будет предоставлять данные надлежащим образом.
4. SPSS имеет простой интерфейс, в котором пользователю не нужно учиться программировать. Он имеет функцию paste, которая создает синтаксис для шагов, выполняемых в пользовательском интерфейсе.
5. SPSS используется во многих университетах. Он имеет очень хороший интерфейс с полной документацией. Он обеспечивает функциональность, с помощью кнопки "Вставить". Кроме того, он имеет очень хорошую официальную поддержку для любых вопросов, с которыми столкнется пользователь.
6. SPSS является дорогостоящим и имеет различные лицензии для различных функций. Его синтаксис ограничен, и он сравнительно медленно внедряет новые технологии. Кроме того, он не может обрабатывать большие наборы данных.

Пакет SAS:

1. SAS является языком программирования и имеет набор, разработанный для расширенной аналитики, бизнес-аналитики, управления данными и прогнозного анализа. Он разработан для статистической системы анализа. SAS был создан в университете

штата Северная Каролина и в первую очередь был разработан, для того чтобы анализировать большие объемы данных сельского хозяйства.

2. Предполагается, что SAS имеет большое количество высококачественного производственного кода для различных целей. Он считается ведущим в коммерческом аналитическом пространстве. SAS имеет сильные регулируемые возможности и свои обновления программного обеспечения в контролируемой окружающей среде.
3. Когда данные присутствуют в огромном объеме, то SAS является более мощным программным продуктом и предоставляет различные средства, такие как сортировка данных.
4. SAS использует PROC SQL, который делает кодирование намного проще, легко работать в SAS тем, кто знает SQL. SAS можно легко обучить с нуля.
5. SAS, активно используется в современном мире. Он имеет хороший интерфейс, и способен обрабатывать большие наборы данных.
6. SAS также является дорогостоящим продуктом. Для опций, которые не задействованы в пользовательском интерфейсе, вам нужно будет написать код и соответственно настроить. В SAS существуют различные программы для различных визуализаций или интеллектуального анализа данных

Заключение. Любой программный продукт, будь то программный продукт с SPSS или SAS - это инструмент в руках аналитика, социолога, экономиста, словом любого специалиста, которому приходится иметь дело с обработкой данных. Выбирая инструмент для решения задачи, ему необходимо учитывать множество факторов, таких как сложность и важность задачи, сроки получения результатов, штат и квалификацию специалистов, бюджет, выделенный на покупку инструмента.

Исходя из проведенного исследования, можно утверждать, что

возможности двух программных пакетов очень схожи и для нужд небольшого социологического исследования можно выбрать любой из них. Вопрос лишь в том, какое время вы готовы потратить на изучение азов работы, позволяют ли сроки вашего исследования подготовить данные должным образом, тратить время на поиск необходимых инструкций для различных вычислений и установку библиотек из репозитория в случае, если вы выбрали SAS. Или же вам необходимо готовое, отлаженное, производительное решение «из коробки», для работы с которым вы всегда можете найти подготовленные кадры или без труда отдать работу «фрилансерам» на аутсорсинг. Постараемся же выделить те плюсы и минусы каждого продукта, которые удалось выявить в ходе проведенной работы.

Достоинства SPSS:

- ❖ Развитый аппарат статистического анализа;
- ❖ Универсальность (может быть использован для решения широкого круга вопросов из различных предметных областей, требующих проведения статистического анализа данных);
- ❖ Широкий набор статистических и графических процедур (более 50 типов диаграмм) анализа данных, а также процедур создания отчетов;
- ❖ Детальная контекстно-ориентированная справочная система, позволяющая неопытному пользователю с большей легкостью ориентироваться в программе, Наличие значительного количества литературы по работе с пакетом.
- ❖ Возможность свободного скачивания демонстрационной версии продукта на официальном сайте компании, наличие версий продукта на различных языках;

Недостатки SPSS:

- ❖ Высокая цена по сравнению со статистическими пакетами

аналогичного уровня.

Да, значительный недостаток у SPSS всего один, но, как и в случае с, например, операционными системами для персональных компьютеров, это может иметь большое значение, как для индивидуального пользователя, так и для организации. Ведь рядовой студент, знакомящийся с обработкой данных или, например, индивидуальный пользователь, которому нужно провести небольшое исследование вряд ли будем готов потратить несколько тысяч долларов на лицензию. Также и в случае с организациями, которым необходимо установить программу на целый парк компьютеров, сталкивается с тем, что стоимость обучения сотрудников работе с бесплатным ПО с открытым исходным кодом может оказаться выгоднее, чем приобретение лицензий на 10-20 рабочих станций.

И тогда, им как раз подойдет SAS, у которого можно выделить следующие сильные стороны:

- ❖ Простой синтаксис, быстрое обучение “с нуля”;
- ❖ Отладка кода проходит значительно проще, чем на R;
- ❖ Интеграция с БД (Oracle/Teradata);
- ❖ Удобный формат выходных данных (особенно таблиц);
- ❖ Мощная поддержка со стороны компании SAS;
- ❖ Многолетний успешный опыт эксплуатации компаниями разной величины, с разными задачами и разным объёмом входных данных. В частности, России SAS используют ОАО “РЖД”, МТС, ЦБ РФ, а также ведущие банки, среди которых Сбербанк, Альфабанк, Тинькофф и многие другие.

Минусы:

- ❖ Профессиональное использование языка предполагает покупку программного продукта;
- ❖ Исходники многих исполняемых алгоритмов SAS не являются

публичными, следовательно изучение работы языка сильно ограничено;

- ❖ SAS значительно уступает в производительности;
- ❖ С точки зрения объема кода SAS также зачастую сильно проигрывает (иногда в несколько раз)..

Резюмируя, SPSS требует больших финансовых ресурсов и меньшего времени на обучение, однако является фактическим стандартом в отрасли, что подтверждается, например, наличием обучающих курсов по работе именно с данным программным обеспечением в различных вузах.

SAS сложнее узнать, что точка и нажмите интерфейс SPSS. SPSS легче узнать, поскольку он обеспечивает функциональность вставки. Все таблицы, диаграммы могут быть легко переданы одним нажатием кнопки. SAS, с другой стороны, не имеет такой функциональности. Трудно настроить SAS, так как нужно иметь знания кодирования для создания вещей в соответствии с настроенными требованиями. SPSS также предоставляет интерфейс, который облегчает обучение. Документация SPSS намного лучше и дает лучшую ясность в алгоритмах, используемых для статистических процедур. Моделирование проще в SPSS, но SAS может обеспечить больше контроля благодаря интерфейсу командной строки и расширенному кодированию редактора. В SAS Enterprise не так хорош визуальный интерфейс, как SPSS.

Обработка данных в SAS выполняется быстрее, чем в SPSS. SPSS обрабатывает данные быстро, но только тогда, когда они невелики. Когда данные становятся больше, трудно обрабатывать их через SPSS. SAS может легко обрабатывать большие объемы данных. Он предоставляет различные функции, такие как сортировка и разделение данных, что упрощает для SAS обработку больших кусков. SPSS, с другой стороны, предоставляет бессрочную лицензию, в то время как SAS предоставляет ежегодную лицензию, которая делает ее дороже, чем SPSS. SAS почти в 1,75 раза

дороже по стоимости одной установки, чем SPSS. Графические возможности SAS и SPSS являются чисто функциональными; хотя можно внести незначительные изменения в графики, полностью настроить графики и визуализации в SAS и SPSS может быть очень громоздким или даже невозможным.