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

ЭВРИСТИЧЕСКИЙ ПОТЕНЦИАЛ IBM SPSS STATISTICS И СВОБОДНОЙ СРЕДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА R-PROJECT: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

(автореферат бакалаврской работы)

студентки 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль подготовки - Прикладная информатика в социологии
Социологического факультета
Ахатчиковой Людмилы Леонидовны

Научный руководитель
кандидат социологических наук, доцент _____ Н.Ю. Кравченко
подпись, дата

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

Саратов 2017

Актуальность проблемы. Для людей, которые принимают решения, постоянно имело значение точно и своевременно сообщать информацию о будущих событиях в обществе и экономике. Поэтому прогнозирование стало очень важным процессом в планирование стратегий различных компаний и политики разных государств. Не случайно А.И. Орлов отмечает, что развитие современных экономических, социологических теорий, а также сложных компьютерных программ повлияло на подъем новых методов прогнозирования и анализа. [1]

В современных условиях, когда информационные потоки стали особенно массивными, появляется колоссальное количество данных. В результате того, что разум человека не способен непосредственно и глубоко анализировать большой массив чисел, то статистические пакеты программ помогают людям получить довольно полную информацию из данных. Сегодня насчитывается более сотни статистических пакетов прикладных программ, решающих задачи статистического анализа социологических данных. [2]

В рамках данной дипломной работы будет проведено сравнение эвристического потенциала двух программных пакетов для обработки статистических данных – IBM SPSS Statistics и свободной среды статистического анализа R-Project.

Степень изученности данной проблемы довольно низкая в силу ее специфического уклона и узкой направленности, тем не менее, существуют некоторые исследованиями данного вопроса.

Шотландский социолог Брендан О'Коннор публиковал свое исследование на тему сравнения различных пакетов программного обеспечения для обработки статистических данных, в числе которых были и рассматриваемые нами программы. Он пришел к следующим выводам:

Достоинством R стало поддержка огромного количества расширений и расширенные возможности визуализации. Достоинством для SPSS явилось

[1] Орлов А.И. Статистические пакеты – инструменты исследователя. - Журнал «Заводская лаборатория». 2008. Т.74. No.5. С.76-78.

[2]«Классификация пакетов прикладных программ» - сайт «Студопедия» Электронный ресурс http://studopedia.ru/10_206954_klassifikatsiya-paketov-prikladnih-programm.html Дата обращения 25.04.2017

легкость статистического анализа. К недостаткам R Брендан О'Коннор отнес сложность обучения, а в SPSS главным недостатком выделил высокую стоимость продукта (Приложение Б, таблица 1).[1] Так же, Роберт Мюэнкен в своей статье ограничивается сравнением данных программных пакетов по следующим категориям: количеству упоминаний в научных статьях за 2015 год (Приложение Б, рисунок 1), количеству книг, написанных по каждому из программных пакетов (Приложение Б, рисунок 2). В каждой из них SPSS выходит победителем, но по количеству упоминаний в предложениях о найме на работу (Приложение Б, рисунок 3), напротив, выигрывает R из чего можно сделать вывод о том, в западных странах реальный сектор экономики сейчас делает ставку на специалистов работающих именно с R при том, что SPSS всё еще остаётся более, популярным статистическим пакетом. Так же Роберт Мюэнкен показывает процентное изменение количества научных статей с использованием программного обеспечения за два года (2014 – 2015 г.). R находится на четвертом месте в росте (15%), в то время как пакет SPSS на 30% снизился по количеству научных статей с использованием этой статистической программы (Приложение Б, рисунок 4). Несмотря на последние годы упадка, SPSS по-прежнему чрезвычайно доминирует для научного использования[2].

Так же В.В. Величко в своей статье «Сравнительный анализ статистических пакетов программ» опубликовал результаты своего исследования: В рамках выполняемого исследования, был проведен опрос мнения 20 пользователей рассмотренных пакетов программ, по 5-ти бальной системе, усредненные результаты представлены в (Приложение Б таблице 2)[3].

[1] «Comparison of data analysis packages: R, Matlab, SciPy, Excel, SAS, SPSS, Stata» - сайт AI and Social Science – Brendan O'Connor Электронный ресурс http://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/ Дата обращения 27.04.2017

[2] «The Popularity of Data Analysis Software», автор Robert A.Muenchen, Электронный ресурс http://r4stats.com/articles/popularity/ Дата обращения 27.04.2017

[3] Международный научный журнал «Инновационная наука» №5/2016 ISSN 2410-6070 Величко В.В. «Сравнительный анализ статистических пакетов программ» С. - 32

В результате наилучшую оценку среди пользователей завоевал пакет R. Пакет SPSS получает тоже хорошую оценку за исключением стоимости программного продукта.

Преподаватель отделения интеллектуальных систем РГГУ Дмитрий Виноградов в своей статье «Среда статистических вычислений R: опыт использования в преподавании» делится своим опытом внедрения программного пакета R на замену программам SPSS и Excel во время практических занятий по его курсу «Статистический анализ данных» в 2010 году[1]. В данной статье он приходит к выводу, что если перед исследователем стоит задача изучения статистики, а также присутствует необходимость написания нестандартных процедур для статистической обработки данных, то ему крайне рекомендуется обратить свое внимание на пакет R.

Так же исследованием в этой области занимается Кандидат физико-математических наук, доцент Вадим Леонардович Аббакумов. В своем исследование на тему анализ данных на базе статистических пакетов R и SPSS он рассказывает о преимуществах этих пакетов. Так же он приводит доводы, почему язык R, все таки стоит изучать на ряду с программой SPSS.[2]

Были изучены многие аспекты программных пакетов R и SPSS, а в контексте рассмотрения обработки именно социологических данных не производилось. В результате было решено рассмотреть данные программные продукты на примере обработке социологических данных.

Целью дипломной работы ставится анализ эвристического потенциала и сравнение пакетов программного обеспечения для обработки социологических данных.

Задачи: 1.Провести классификацию функциональности статистических программных продуктов

[1] Дмитрий Виноградов «Среда статистических вычислений R: опыт использования в преподавании» Электронный ресурс https://habrahabr.ru/post/92135/ Дата обращения 29.04.2017

[2]Вадим Леонардович Аббакумов «Лекция 1. Анализ данных на R в примерах и задачах» Электронный ресурс https://www.youtube.com/watch?v=8mwJ3mEjdIg дата обращения 27.04.2017

2. Рассмотреть теоретико-методологические основы функциональности статистического программного обеспечения SPSS и R.

3. Показать возможности практического применения программных статистических пакетов SPSS и R на примере авторского исследования «Отношение молодежи г. Саратова к незарегистрированным бракам»

4. Выявить преимущества и недостатки использования статистических пакетов SPSS и R.

Объектом анализа являются программы для обработки статистических данных, а именно два сравниваемых пакета программного обеспечения – SPSS и R.

Предметом выступают актуализированные в рабочем режиме функционально – аналитические характеристики представленных пакетов программ.

Эмпирическую базу данной работы составили данные авторского исследования «Отношение молодежи города Саратова к незарегистрированному браку», статистический пакет SPSS 22. Исследование было проведено в 2015 году. Объем выборочной совокупности составил 50 человек.

Практическая значимость работы заключается в возможности использования основных положений и выводов квалификационной работы в области прикладной информатики в социологии. Также работа будет интересна социологам и бакалаврам различных направлений, применяющим количественные методы в исследованиях.

Структура работы. Диплом состоит из введения, двух разделов (раздел 1 «Программные пакеты для обработки статистических данных», раздел 2 «Сравнение программных пакетов на примере обработки массива социологических данных») заключения, списка использованных источников и приложения.

Основное содержание работы. В первом разделе приводится классификация различных программных пакетов для обработки данных и

основные сведения о них. Так же в этом же разделе рассмотрено теоретико-методологическое описание возможностей статистических пакетов SPSS и R.

На сегодняшний день на рынке представлено около тысячи компьютерных программ для статистической обработки данных. Разнообразие статистических пакетов обусловлено многоплавностью задач обработки данных с применением различных типов статистических процедур анализа для поиска ответов на вопросы из различных областей человеческой деятельности.[1]

Большая часть статистических пакетов, предоставленных на рынке, имеют гибкую модульную структуру, которая со временем пополняется и расширяется за счет пользовательских модулей, которые могут дополнительно закупаться или находится в свободном доступе в Интернете. Гибкая модульная структура дает возможность адаптировать пакет к потребностям конкретного пользователя.

Пакет для статистической обработки данных должен иметь следующий минимальный набор 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.[2]

Перед пользователями различных категорий встает вопрос выбора оптимального статистического пакета для поиска верных ответов на существующие вопросы. Очевидно, что оптимальным является вариант,

[1] Лекция 11: Статистическая обработка данных – сайт «Интуит» Электронный ресурс http://www.intuit.ru/studies/courses/3632/874/lecture/14309 Дата обращения 27.04.2017

[2] «Сравнение программных продуктов для анализа данных: R, MATLAB, SciPy, MS Excel, SAS, SPSS, Stata», Ирина Чучуева, Электронный ресурс http://www.mbureau.ru/blog/sravnenie-programmnyh-produktov-dlya-analiza-dannyh-rmatlab-scipy-ms-excel-sas-spss-stata дата обращения 03.05 2017

сочетающий в себе все необходимые параметры, это такие как: удобство пользования, количество методов, объем обрабатываемых данных, наличие методического обеспечения, высокое качество работы и умеренная цена. 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 приобретении статистического пакета пользователь должен учитывать четыре параметра:

1. Количество обрабатываемых данных;

2. Сможет ли пользователь решить поставленные задачи перед ним с помощью этого пакета;

3. Требования, которые предъявляются к знаниям пользователя в области статистики;

4. Имеет ли он в наличии персональный компьютер, и какой мощности;[1]

Все пакеты статистической обработки данных можно разделить по признаку функциональности. Они делятся на универсальные пакеты или по другому их можно назвать «популярными», профессиональные пакеты и специализированные. В таблице 1 представлена классификация статистических пакетов по функциональности.

Таблица 1- Классификация статистических пакетов по функциональности

Универсальные	STATGRAPHICS, SPSS, STATISTICA, S-PLUS, R, STADIA, STATA, Minitab
Профессиональные	SAS, BMDP
Специализированные	Большое Разнообразие

[1] Лекция 11: Статистическая обработка данных – сайт «Интуит» Электронный ресурс http://www.intuit.ru/studies/courses/3632/874/lecture/14309 Дата обращения 26.04.2017

Рассмотренные программы для обработки данных SPSS и R относятся к универсальным. Статистический пакет SPSS (StatisticalPackagefortheSocialSciences – статистический пакет для социальных наук) – универсальный пакет компании SPSSInc. Первая версия статистического пакетаSPSS вышла в 1968 году. В настоящее время, по мнению компании IBM, SPSS занимает одно из ведущих мест программного обеспечения в области статистического анализа данных. Хорошо разработанный графический интерфейс помогает пользователю в простоте и эффективности использования статистического пакета. Статистический пакет SPSS является модульной системой. В новой версии статистического пакета IBMStatistics 23 установлено шестнадцать разных модулей.Состав используемых пользователем модулей будет зависеть от варианта поставки.Поставка продуктов пакета IBMSPSS имеет несколько возможностей: IBMSPSSStatisticsStandard, IBMSPSSStatisticsPremium, IBMSPSSStatisticsforEducators (US). Командный язык Syntax - в SPSS многие задачи пользователь может выполнить с помощью мыши и диалоговых окон, однако в программе имеется мощный командный язык. Он позволяет сохранить и автоматизировать множество повторяющихся задач.

R – это мощная свободно распространяемая статистическая среда с открытым исходным кодом, которая включает в себя: программирование, интерактивную оболочку и широкие возможности по отображению графической информации. Более того, R имеет огромный набор математических и статистических функций, а также дополнительные возможности, которые предоставляются в подключаемых пакетах.

R доступен под лицензией GNU General Public License (в переводе Универсальная общественная лицензия GNU, Универсальная общедоступная лицензия GNU или Открытое лицензионное соглашение GNU) — лицензия, которая была создана в рамках проекта GNU в 1988 г. на свободное

программное обеспечение. По лицензии автор дает право на передачу программного обеспечения в общественную собственность.[1]

Система R состоит из модулей (пакетов), каждый из которых выполняет определенный круг задач.. Роберт Мюэнкен в своем исследование опубликовал график, в котором указывалось количество пакетов R в каждом году, начиная с 2002 года, когда их было около 200. К 2014 году их стало порядка 6000 (Приложение Б, рисунок 5).[2]

Одной из сильных сторон R является богатство возможностей по созданию сложных графиков и любых других форм визуального представления информации. Много чего можно сделать с помощью базовых возможностей, которые дополняются отдельными графическими пакетами.

Второй раздел «Сравнение программных пакетов на примере обработки массива социологических данных» посвящен определению массива данных для сравнения программных пакетов и практическому сравнению двух программных пакетов – SPSS и R Применительно к социологическому исследованию, данные представляются в виде таблиц, которые содержат данные двух видов: постоянную часть, как заголовок таблицы, названия строк и столбцов и переменную часть — собственно показатели таблицы (матрицы), которые являют собой непосредственно собранные в ходе социологического исследования данные. Они также могут быть введены в запоминающее устройство для обработки, в таком случае массив данных образует файл.

При сравнении программных пакетов по критерию установки и удобства загрузки данных для работы, получены следующие выводы: в целом, процесс установки R напоминает установку программного пакета SPSS, в отличии от загрузки массива данных. Так как если вы не имеете навыков программирования

---

[1] Википедия [Электронный ресурс] : свободная энциклопедия / текст доступен по лицензии Creative Commons AttriЬution-ShareAlike ; Wikimedia Foundation, Inc, некоммерческой организации. Электрон. дан. (712413 статей, 2479181 страниц, 117 104 загруженных файлов). Wikipedia®, 2001- .URL: https://ru.wikipedia.org/wiki/GNU_General_Public_License#GPL_v2 (дата обращения: 13.05.2017). Загл. с экрана. Последнее изменение страницы: 14:38, 3 апреля 2017. Яз. рус

[2] «The Popularity of Data Analysis Software», автор Robert A.Muenchen, Электронный ресурс http://r4stats.com/articles/popularity/

или знания команд для загрузки данных в R, то вы даже не сможете загрузить данные, не говоря уже об анализе. В SPSS загрузка данных происходит более интуитивно за счет интерфейса программы. Так же R имеет возможность импортировать данные за счет пакета «foreign».

При сравнении программных пакетов по критерию составления частотных таблиц, получены следующие выводы: при наличии подготовленных должным образом социологических данных оба программных пакета SPSS и R могут проводить частотный анализ и помочь исследователю сделать первые выводы о проведенном исследовании.

При сравнении программных пакетов по критерию визуализации данных на основе генерирования разных видов диаграмм, получены следующие выводы: из практического применения пакетов можно увидеть, что пакет R и SPSS могут строить разные виды диаграмм, однако диаграммы в R обладают большими возможностями персонализации и тонкой настройки в отличие от SPSS.

При сравнении программных пакетов по критерию составления таблиц зависимостей и проверке статистических гипотез, получены следующие выводы: с помощью пакетов SPSS и R можно выполнять и более сложные операции, чем выведение просты частотных таблиц и построение различных видов графиков. К более сложным операциям относятся построения таблиц сопряженности для выявления зависимости между переменными и проверка статистических гипотез при помощи различных критериев для разных видов шкал. При том, что в SPSS данные операции выполняются более интуитивно за счет интерфейса программы и является более дружественной по отношению к пользователю. В R есть тоже свои положительные стороны - зная все функции, которыми обладает программа, все лишь несколькими строчками кода команд можно выполнить нужные действия.

Заключение. Любой статистический пакет, будь то программа SPSS, имеющая пользовательский интерфейс, или R, который является языком программирования с открытым исходным кодом, или будь то программа SAS, которая представляет собой смесь языка программирования и графического

приложения – все эти программные продукты являются инструментом в руках аналитиков различных областей науки, которым приходится обрабатывать данные. При выборе инструмента для решения задачи аналитик должен учитывать множество факторов: важность и сложность задачи, которую нужно решить, сроки получения результатов, бюджет, который выделяется на покупку инструмента, а так же штат и квалификация специалистов.

Делая выводы из проведенного исследования, хочется сказать, что возможности, которыми обладают пакет SPSS и R очень схожи и для проведения анализа данных социологического исследования, можно воспользоваться любым их них. Вопрос лишь заключается в том, сколько времени вы готовы потратить для изучения азов работы, есть ли время что бы подготовить данные должным образом для анализа, сколько времени готовы потратить на поиск необходимых инструкций для разных операций и установку библиотек из репозитория, в том случае если вы решили работать в R. Или же Вам необходим программный продукт, который уже готовый, отлаженный, имеет обычный пользовательский интерфейс, в котором команды выполняются более интуитивно за счет него.

Попытаемся выделить все достоинства и недостатки каждого программного продукта, которые удалось выявить в ходе работы в них.

Далее нам хотелось бы рассказать о плюсах и минусах каждого продукта, которые удалось выявить в ходе проведенной работы.

Достоинства SPSS: развитый аппарат статистического анализа, универсальность (может быть использован для решения вопросов из различных предметных областей), большой набор статистических и графических процедур (более 50 типов диаграмм) анализа данных, а также процедур создания отчетов, детальная контекстно-ориентированная справочная система, которая позволяет неопытному пользователю с большей легкостью ориентироваться в программе, имеется значительное количество литературы по работе с пакетом, возможно свободно скачивать демонстрационную версию продукта на официальном сайте компании, имеется версия 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, который знакомится с обработкой данных и которому нужно провести небольшое исследование, будет не в состоянии потратить несколько тысяч долларов на лицензию.

У статистического пакета R, которого можно выделить следующие сильные стороны: распространение программы под GNU Public License, которое позволяет ее свободное и бесплатное использование, доступность как исходных текстов, так и бинарных модулей в обширной сети репозитариев CRAN (The Comprehensive R Archive Network), имеет возможность обмена данными с электронными таблицами и возможность сохранения всей истории вычислений для целей документирования.

К недостаткам стоит отнести: сложность обучения, так как R – мощная программа с очень большим числом доступных аналитических и графических функций. И к тому же для работы в ней необходимо владеть знанием программирования. Еще одним недостатком является то, что исходном виде отсутствует удобный графический интерфейс и очень мало русскоязычной литературы по работе с пакетом;

Как мы уже говорили, возможности SPSS схожи с возможностями R. Но пакет SPSS требует меньше времени на обучение и больших ресурсов для приобретения. R, напротив, требует самых минимальных финансовых ресурсов и самых квалифицированных специалистов или же большого количества времени на обучение. Однако, в любом случае, результат анализа данных будет зависеть не от выбранного инструмента, а от степени квалификации аналитика.