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

СОСТОЯНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ОБРАБОТКИ СОЦИОЛОГИЧЕСКОЙ СТАТИСТИКИ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПАКЕТОВ

АВТОРЕФЕРАТ ДИПЛОМНОЙ РАБОТЫ

студента 6 курса 631 группа
специальности 080801.65 - Прикладная информатика в социологии
Социологического факультета
Емельянова Антона Владимировича

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

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

Саратов 2016

Введение. Точная и своевременная информация о том, что может произойти в экономике и обществе в будущем, всегда имела значение для тех, кто принимает решения. Прогнозирование стало важной частью процесса планирования стратегии любой компании и политики любого государства. Развитие современных экономических, социологических теорий, а также сложных компьютерных программ повлияло на подъем новых методов прогнозирования и анализа.

В современных условиях, когда информационные потоки стали особенно массивными, появляется колоссальное количество данных. Также, ввиду того, что скорость всех процессов в обществе возрастает, увеличивается и потребность в выявлении быстрых ответов на возникающие вопросы, с целью поиска которых и проводится огромное количество исследований и сбор данных во всех сферах жизни.

Сегодня рынок статистического программного обеспечения впечатляет своим многообразием, несмотря на его специфичность. Существует более тысячи разнообразных программ решающих задачи статистического анализа социологических данных. Однако даже на рынке такого специфического программного обеспечения существует конкуренция.

Актуальность выбора соответствующего инструментария для обработки социологических данных обусловлена тем, что при проведении любых исследований, включающих в себя анализ массивов данных, будь то психологические, медицинские исследования, научные изыскания, а также социологические опросы и исследования, выбор инструментария для их обработки, диагностики и прогнозирования является очень важным для принятия управленческих решений, а на первый план выходят совокупность быстроты, точности и удобства в процессе анализа полученных результатов. 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. Стоит отметить, что, к сожалению,

последняя совершенно неизвестна на просторах нашей родины.

Степень изученности данной проблемы довольно низкая в силу ее специфического уклона и узкой направленности, тем не менее, существуют некоторые исследованиями данного вопроса. Например, Роберт Мюэнкен в своей статье ограничивается сравнением данных программных пакетов по следующим категориям: количеству упоминаний в академических статьях за 2014 год (прил.1, рис.1), количеству книг, написанных по каждому из программных пакетов (прил.1, рис.2) и количеству ссылок на основной сайт в интернете (прил.1, рис.3). В каждой из них SPSS выходит победителем, но по количеству упоминаний в предложениях о найме на работу[1] (прил.1, рис.4), напротив, выигрывает R, из чего можно сделать вывод о том, что реальный сектор экономики сейчас делает ставку на специалистов работающих именно с R при том, что SPSS всё еще остаётся более, если можно так выразиться, «на слуху» [2].

Шотландский социолог Брендан О'Коннор также публиковал свое исследование на тему сравнения различных пакетов программного обеспечения для обработки статистических данных, в числе которых были и рассматриваемые нами программы[3]. Он пришел к следующим выводам:

	R	**SPSS**
Преимущества	Поддержка огромного количества расширений, расширенные возможности визуализации	Легкость статистического анализа
Недостатки	Сложность обучения	Высокая стоимость
Рекомендуемое	Финансовая сфера,	Объемные и

[1] Для сбора статистики о предложениях о найме на работу использовались следующие ключевые слова: R, статистический анализ, интеллектуальный анализ данных (data mining), аналитика данных, машинное обучение, количественный анализ, бизнес анализ, статистическое программное обеспечение, предсказательное моделирование.

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

[3] http://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/

применение	обработка статистических данных	комплексные научные изыскания и исследования

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.

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

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.

Сравнительному анализу предполагается подвергнуть два пакета программ. SPSS является модульной программой для обработки статистических данных. Ее основу составляет базовый модуль – SPSS Base – позволяющий осуществлять управление данными и содержащий наиболее распространенные методы статистического анализа данных: проведение описательной статистики; построение линейных и нелинейных моделей; осуществление преобразования данных; проведение факторного, кластерного, дисперсионного анализов; вычисление корреляций; построение

[1] https://habrahabr.ru/post/92135/

графиков; подготовка отчетов и прочие функции. Для проведения расширенного и углубленного анализа данных могут быть установлены дополнительные модули пакета.

В качестве альтернативы была выбрана система R. Эта система начала разрабатываться усилиями Роберта Джентльмена и Росса Ихака на факультете статистики университета Мельбурна в 1995 году. Первые буквы имен авторов определили ее название. R широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ. R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей применения. В базовую поставку R включен основной набор пакетов, а всего по состоянию на 2013 год доступно более 4000 пакетов. Ещё одной особенностью R являются графические возможности, заключающиеся в возможности создания качественной графики, которая может включать математические символы.

В 2010 году R вошёл в список победителей конкурса журнала InfoWorld в номинации на лучшее открытое программное обеспечение для разработки приложений.

В качестве массива социологических данных для сравнения эвристического потенциала двух пакетов программного обеспечения – SPSS и R – было решено использовать данные, собранные в ходе проведенного мной исследования вторичной занятости студентов в 2012 году. На основании темы, выбранной для исследования, объектом исследования являлись студенты 1-5 курсов дневного отделения Социологического факультета Саратовского Государственного Университета, а предметом исследования являлась вторичная занятость студентов дневного отделения Социологического факультета Саратовского Государственного Университета.

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 «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, дополнительно закупаемых 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.[1]

Перед пользователями различных категорий встает вопрос выбора оптимального статистического пакета для поиска верных ответов на существующие вопросы. Очевидно, что оптимальным является вариант, сочетающий в себе необходимые функциональные возможности, высокое качество работы и умеренную цену. 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.[2]

---

[1] «Сравнение программных продуктов для анализа данных: R, MATLAB, SciPy, MS Excel, SAS, SPSS, Stata», Ирина Чучуева, http://www.mbureau.ru/blog/sravnenie-programmnyh-produktov-dlya-analiza-dannyh-r-matlab-scipy-ms-excel-sas-spss-stata

[2] Ш. Ф. Фарахутдинов, А. С. Бушуев Обработка И Анализ Данных Социологических Исследований В Пакете Spss 17.0: Курс Лекций, Тюмень Тюмгнгу, 2014. – 219 с.

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 исследлвания данные. Они также могут быть введены в запоминающее устройство для обработки, в таком случае массив данных образует файл.

При сравнении програмных пакетов по критерию удобства загрузки данных для работы, получены следующие результаты: SPSS позволяет очень удобно импортировать данные, подготовка данных к работе в R — это одна из самых больших проблем для новичка R. Сама по себе обработка данных подробно описана в разных руководствах и пособиях, а вот информация как добиться того, чтобы R прочитал приготовленные в другой программе данные, как правило, опускается.

При сравнении програмных пакетов по критерию составления частотных таблиц, получены следующие результаты: оба программных пакета – SPSS и R – при наличии подготовленным должным образом данных и некоторой первичной подготовке, как то: корректное озаглавливание переменных и именование ответов на вопросы, позволяют провести частотный анализ и сделать первые выводы о проведенном социологическом исследовании. По крайней мере, в той его части, которая касается социального портрета респондентов.

При сравнении програмных пакетов по критерию визуализации данных на основе генерирования круговых диаграм, получены следующие результаты: возможности построения круговых диаграмм в R схожи с SPSS, однако обладают большими возможностями персонализации и тонкой

настройки.

При сравнении проргаммных пакетов по критерию составления таблиц зависимостей, получены следующие результаты: при помощи обоих программных пакетов – SPSS и R – можно довольно интуитивно выполнять и более сложные операции, чем построение простых частотных таблиц, а именно построение таблиц сопряженности для нахождения зависимостей между переменными в массиве социологических данных, что является основным инструментом для подтверждения одной из выдвинутых в ходе данного конкретного исследования гипотезы. При том, в SPSS эта операция является более дружественной по отношению к пользователю и выполняется более интуитивно.

Заключение. Любой программный продукт, будь то программный продукт с пользовательским интерфейсом (SPSS) или язык программирования (R), или смесь графического приложения и языка программирования (MATLAB, 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 R. 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 R, у которого можно выделить следующие сильные стороны:

- Распространение программы под GNU Public License, позволяющее ее свободное и бесплатное использование;
- Доступность как исходных текстов, так и бинарных модулей в обширной сети репозитариев CRAN (The Comprehensive R Archive Network).
- Возможность обмена данным с электронными таблицами;
- Возможность сохранения всей истории вычислений для целей документирования.

К недостаткам стоит отнести:

- Сложность обучения;
- В исходном виде отсутствует удобный графический интерфейс. Операции выполняются в командной строке.

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. Язык R предоставляет большую гибкость, но требует самой высокой квалификации специалистов. Баланс достичь несложно, однако результат, в любом случае, будет зависеть от квалификации аналитика, а не от выбранного инструмента.