УДК 004.657:004.43

ПЛАТФОРМА R-STUDIO ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ

В.С. Дроздов
Ассистент кафедры инженерной психологии и эргономики БГУИР, магистр технических наук

Д.В. Лихачевский
Декан факультета компьютерного проектирования БГУИР, кандидат техни-ческих наук

Е.А. Мельникова
Ассистент кафедры инженерной психологии и эргономики БГУИР, магистр технических наук

В.С. Осипович
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

Н.В.Щербина
Старший преподаватель кафедры инженерной психологии и эргономики БГУИР, магистр технических наук

К.Д.Яшин
Заведующий кафедрой инженерной психологии и эргономики БГУИРи, кандидат тех-нических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: kafipie@bsuir.by, e.melnikova@bsuir.by

Аннотация. Рассмотрены примеры работы в графической среде обработки данных с использованием платформы R-Studio, описаны основные технологические приемы. Представлены практические работы для обучения студентов, а также результаты их выполнения.

Ключевые слова: анализ, вектор, матрица, распределение, вероятность, большие данные, R, R-Studio.

Введение. Язык программирования R является инструментом для статистической обработки данных и работы с графикой. Это программная среда с открытым исходным кодом, развиваемая в рамках GNU-проекта. R применяется там, где нужна работа с большими данными. Это не только для статистики в узком смысле слова, но это первичный анализ (графика, таблицы сопряженности и др.). Это также продвинутое математическое моделирование. R может использоваться и там, где сейчас принято использовать аналитические программы уровня MatLab/Octave [1].

Среди существующих графических оболочек для работы с R следует отметить платформу R-Studio, которая отличается удобством и постоянно расширяющимися функциональными возможностями. Внешний вид компьютерной программы R-Studio представлен на рисунке 1.

Главные преимущества платформы R-Studio: высокая гибкость и свободный код. Гибкость позволяет создавать приложения (пакеты) практически на любые потребности. Кажется,

нет ни одного метода современного статистического анализа, который не был бы сейчас представлен в R-Studio. Свободный код – это не просто бесплатность программы, но и возможность разобраться, как именно происходит анализ, а если в коде встретилась ошибка – самостоятельно исправить ее и сделать исправление доступным для всех [1].

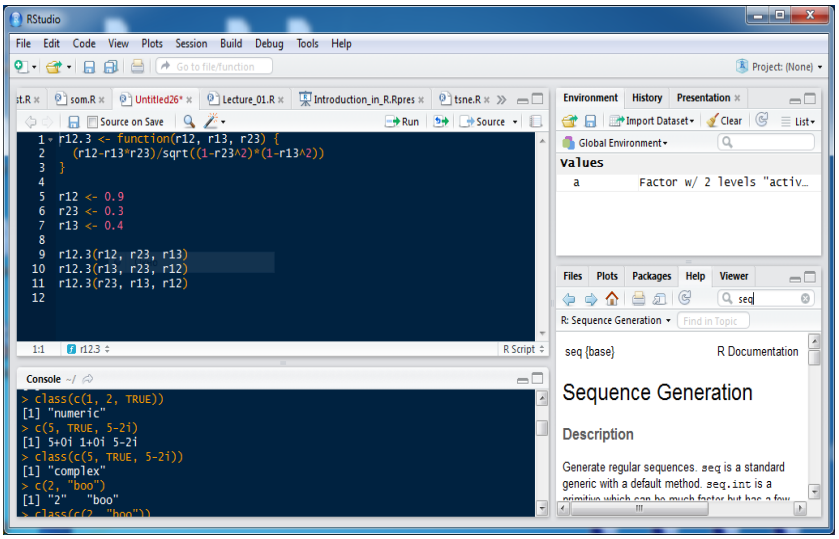

Рисунок 1. Внешний вид R-Studio [2]

Результаты работы. Широкие технические возможности графических оболочек для работы с R, в частности платформы R-Studio, послужили основанием для принятия организационного решения – ввести эту дисциплину студентам факультета компьютерного проектирования Белорусского государственного университета информатики и радиоэлектроники в качестве специального курса.

Кафедра инженерной психологии и эргономики готовит специалистов по информационным технологиям с квалификацией инженер-системотехник. Участие кафедры в организации и работе конференции BIG DATA позволило ей организовывать взаимодействие с IBM BIG DATA University (IBM BDU). Образовательная программа Ambassador от IBM BDU для нас стала доступна благодаря сотрудничеству с IBM Analyics Platform и BEZ Next (обе компании – США) [3].

В текущем учебном году для студентов 3 курса кафедра разработала и представила теоретические и практические курсы для освоения платформы R-Studio.

Цель настоящей работы – представить опыт проведения практических занятий R-Studio, осуществленных в 2017-2018 учебном году со студентами 3 курса при подготовке инженеров-системотехников.

В настоящей статье представлено несколько практических работ по изучению инструментов и технологических возможностей платформы R-Studio.

Практическая работа «Векторы и матрицы» Цель: изучить инструменты и технологию формирования векторов и матриц в пакете R-Studio. Имеем набор данных о фильмах, представленных в таблице 1 [4].

С применением R-Studio выполним следующие технологические приемы. Создадим вектор из девяти фильмов: фильм 2; фильм 1; фильм 4; фильм 8; фильм 3; фильм 6; фильм 5; фильм 3; фильм 9 фильм 7. Создадим из вектора матрицу (4x3). Выведем элемент Whiplash (через номер строки и столбца). Выведем первую строку. Выведем второй столбец. Выведем вектор матрицы. Создадим вектор длительностей фильмов. Создадим матрицу (3x3) из полученного ранее вектора. Увеличим значение каждого элемента матрицы на пять. Определим, какие фильмы можно посмотреть, потратив не более двух часов. Выведем названия фильмов.

На рисунках 2 и 3 представлены примеры компьютерных программ, формируемых в результате работы с пакетом R-Studio. Известно, что техника формирования векторов и матриц лежит в основе обработки массивов больших данных.

Таблица 1

База фильмов

Название фильма	Год выпуска	Длительность (часы)	Жанр	Рейтинг (IMDB)	Сборы (млн. USD)	Дубляж	Возраст
фильм 1	1995	81	анимация	8,3	30	0	0
фильм 2	1998	125	анимаация	8,1	10,4	1	14
фильм 3	1985	97	драма	7,9	1	0	14
фильм 4	2011	100	романтика	8	15	1	12
фильм 5	1936	87	комедия	8,6	1,5	0	10
фильм 6	1999	139	драма	8,9	6,3	0	18
фильм 7	2002	130	криминал	8,7	3,3	1	18
фильм 8	1987	119	драмма	7,9	25	0	14
фильм 9	1977	121	экшн	8,7	11	0	10
фильм 10	1999	122	драма	8,4	15	0	18
фильм 11	2015	118	драмма	8,3	13	1	18
фильм 12	1964	94	комедия	8,5	1,8	1	0
фильм 13	1998	95	хорор	7,3	1,2	1	18
фильм 14	1975	91	комедия	8,3	0,4	1	14
фильм 15	2006	98	комедия	5,2	4,2	0	16
фильм 16	2004	99	хорор	8	6,1	1	10
фильм 17	1976	113	криминал	8,3	1,3	1	18
фильм 18	1994	142	криминал	9,3	25	0	14
фильм 19	2014	169	приключенческий	8,6	165	0	18
фильм 20	1995	178	биографический	8,2	50	0	16
фильм 21	1990	145	биографический	8,2	25	0	0
фильм 22	2013	179	романтика	7,8	4,5	1	16
фильм 23	2010	108	триллер	8	13	0	16
фильм 24	1985	116	научно-фантастический	8,5	19	0	12
фильм 25	2008	107	триллер	7,6	5,5	1	14
фильм 26	2014	106	драма	8,5	3,3	1	12
фильм 27	2014	100	криминал	8,1	25,5	0	14
фильм 28	1995	104	фентези	6,9	65	0	12
фильм 29	2004	108	драма	8,3	20	0	14
фильм 30	2002	113	комедия	7,2	45	0	12

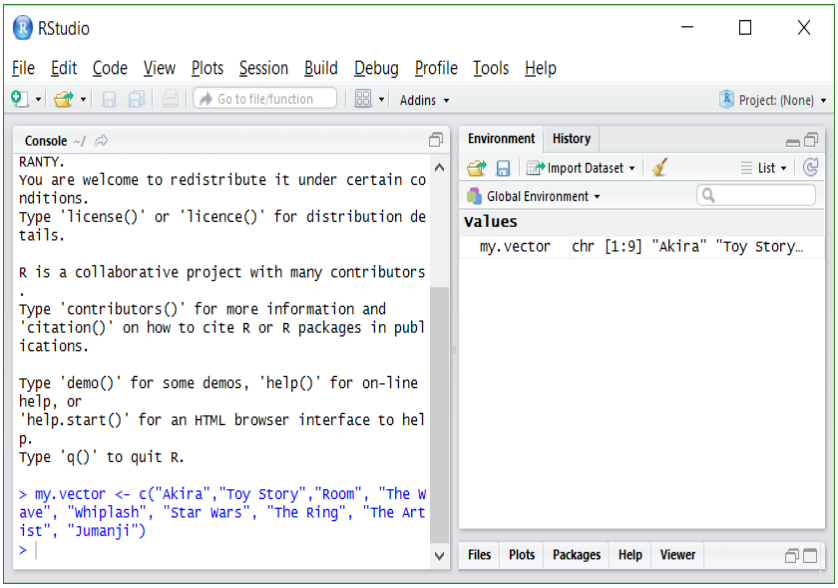

Рисунок 2. Создание вектора из девяти названий фильмов

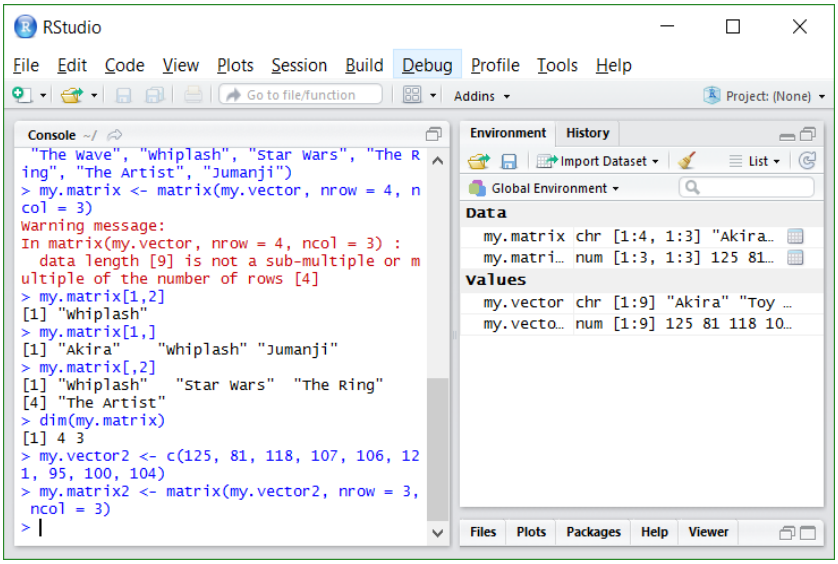

Рисунок 3. Создание матрицы (3x3) из вектора

Практическая работа «Анализ векторов». Цель: изучить технологию оценки различия между выборками данных с применением параметрического t-теста и непараметрического Wilcox-теста. При анализе математических и физических моделей часто требуется оценить, имеются ли различия между двумя выборками данных. Для этого используется параметрических t-тест и непараметрический Wilcox-тест. Эти тесты определяют различия данных только по центральным значениям. Наиболее часто используется t-тест, рассмотрим технологию применения параметрического t-теста для оценки различия между выборками данных. Создадим две нормальные выборки, плотности которых представлены на рисунке 4. Примем нулевую гипотезу H против альтернативной H1. Оценим t-статистику и число степеней свободы для нашего примера. Функция возвращает значение под кривой распределения Стьюдента. Сравним полученные результаты.

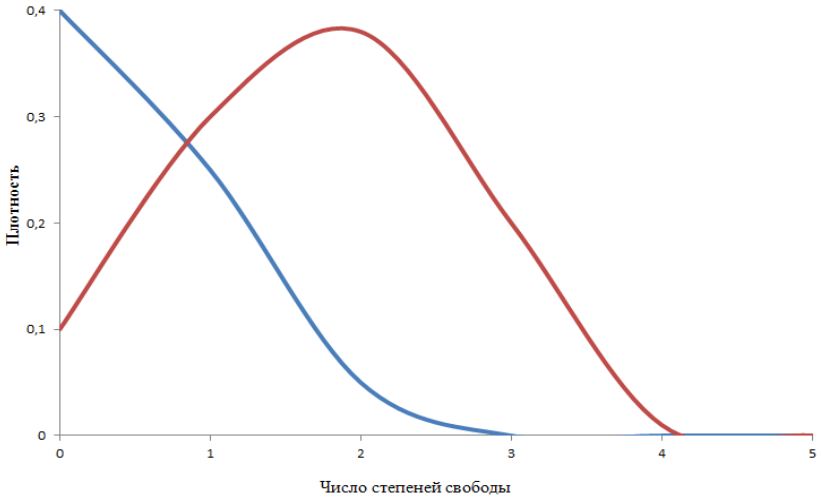

Рисунок 4. Плотности выборки

Практическая работа «Вероятность события». Цель: изучить технологические приемы построения распределения вероятности с использованием пакета R-Studio. Результаты исследований характеристик психофизиологических состояний испытуемых содержит около 400 записей. Фрагмент таблицы представлен на рисунке 5.

Испытуемый	Дата эксперимента	Дата рождения	Возраст	№ сеанса	Время, потраченное на сеанс	Количество реалаксация	Количество свыше 60 с	Количество свыше 155 с
1	19.03.2014	21.03.1970	43	1	139,7	0	0	0
1	15.05.2014		44	2	239,8	0	0	0
1	29.05.2014		44	3	1929,5	5	3	2
2	24.03.2014	24.03.1980	34	1	272,1	0	0	0
3	31.03.2014	24.03.1988	26	1	197,5	0	0	0
3	02.04.2014		26	2	413,5	0	0	0
4	21.11.2014	02.04.1980	34	1	377,7	0	0	0
5	07.04.2014	16.04.1975	39	1	1048,4	3	0	3
6	07.04.2014	26.05.1977	36	1	879,5	5	5	0
6	07.04.2014		36	2	1048,4	3	0	3
7	22.05.2014	07.08.1964	49	1	773,6	0	0	0
8	10.06.2014	21.12.1975	38	1	879,6	3	3	0
8	12.03.2015		38	2	1048,4	3	3	0
8	19.03.2015		39	3	112,3	0	0	0
8	09.04.2015		39	4	200,5	0	0	0
8	13.10.2015		39	5	680,4	0	0	0
8	14.10.2015		39	6	302,2	1	1	0
9	08.04.2014	13.12.1980	33	7	572	1	0	1
9	17.04.2014		33	1	1293,1	0	0	0
10	09.04.2014	18.02.1960	54	2	1593,9	5	1	4
10	11.04.2014		54	1	972,9	0		
10	15.05.2015		55	2	1032,7	0		
10	11.06.2015		55	3	2477,9	0		
10	06.июл		55	4	1052,5	7	7	0
10	06.07.2015		55	5	1298,6	7	7	0
10	09.07.2015		55	6	1053,4	3	3	0
10	15.07.2015		55	7	65,8	0		
10	16.07.2015		55	8	335,2	1	1	0
10	28.07.2015		55	9	70,5	0		
10	09.04.2014		55	10	893,4	5	5	0
11	09.04.2014	25.05.1989	24	1	1002,5	0		
12	09.04.2014	01.06.1976	37	1	942,8	6	6	0
12	09.12.2014		38	2	1148,6	7	6	1
13	09.04.2014	24.07.1979	34	1	1182,9	7	6	1
13	19.05.2014		34	2	2039,8	3	3	0

Рисунок 5. Фрагмент таблицы Excel

Определим и построим распределение вероятности событий от порядкового номера месяца. Для этого экспортируем данные из таблицы Excel, приводим строчные записи в формате POSIXct. Создаем объект, хранящий результат и месяц. Находим количество успешных опытов (более 30 секунд). Вызываем метод x11(), чтобы нарисовать график. Получаем распределение, представленное на рисунке 6.

С использованием тех же исходных характеристик психофизиологических состояний испытуемых осуществим попытку получить вероятность удачных результатов эксперимента в зависимости от года рождения респондентов. Применяя технологические приемы пакета R-Studio, получаем распределение, представленное на рисунке 7.

Опять же, с использованием исходных характеристик осуществим оценку вероятности удачного эксперимента в зависимости от времени суток работы испытуемых. Полученное распределение представлено на рисунке 8.

Рисунок 6. Распределение вероятности событий от порядкового номера месяца

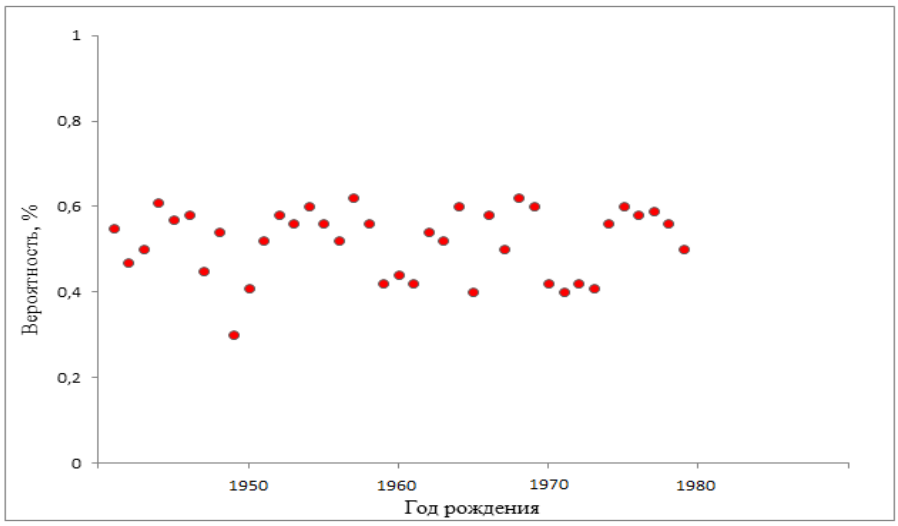

Рисунок 7. Распределение вероятности удачного эксперимента в зависимости от года рождения испытуемых

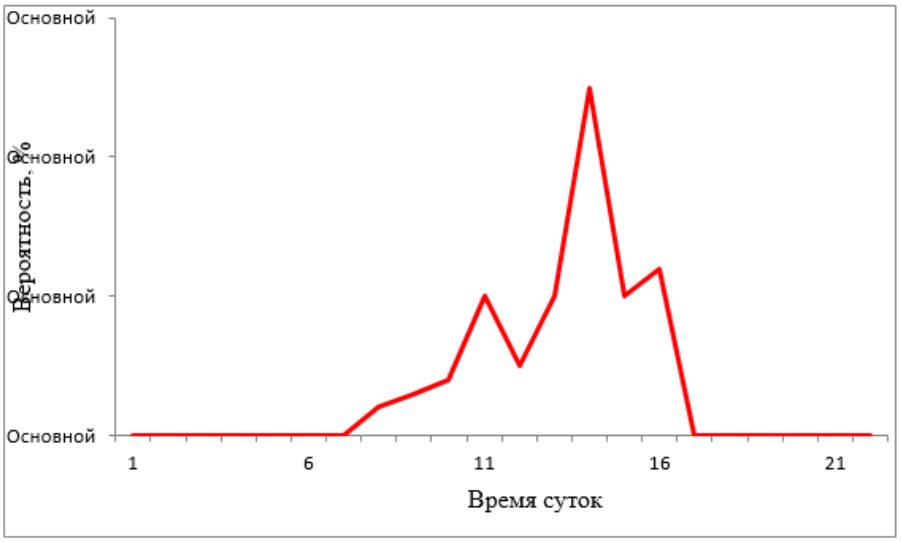

Рисунок 8. Распределение вероятности удачных экспериментов от времени суток работы испытуемых

Практическая работа «XLS-таблицы». Цель: изучить технологические приемы пакета R-Studio при поиске и обработке информации из списка по заданным параметрам. На рисунке 9 представлены фрагменты списка научно-технических журналов, индексируемых в базе данных Scopus [5]. Список представлен в формате xls и содержит несколько тысяч записей. Применяя технологические приемы работы с пакетом R-Studio, выполним следующие действия. Выведем список журналов, относящихся к тематике Computer Science. Выведем список журналов, изданных до 2016 года. Выведем список журналов, издательства Pleiades International. Построим столбчатую диаграмму, отражающую количество журналов по каждой из представленных тематик. Выведем список журналов, издаваемых с 1996 года по настоящее время. Создадим вектор из списка журналов, относящихся к тематике Chemistry. Создадим вектор из списка журналов, относящихся к тематике Energy.

Практическая работа «Графика». Цель: изучить методики построения графических изображений с использованием пакета R-Studio. Используя технологические приемы R-Studio, создадим пустой графический файл. Зададим необходимые нам масштабы по осям абсцисс, ординат или по обеим осям логарифмическим способом. Создадим подписи осей абсцисс и ординат. Создадим надписи сверху и снизу графика. Нарисуем прямую y=kx+b. Построим 2 графика, представленных на рисунке 10. Варианты: blank – нет линии; solid – сплошная линия; dashed – пунктирная линия; dotted – точечная линия; dotdash – точка-тире.

Практическая работа «Открытые ресурсы». Цель: изучить методики загрузки и чтения данных их внешних источников, технологические приемы пакета R-Studio. Компания N представляет пользователям информационные ресурсы, полученные из различных открытых источников [6]. Выполним команду download.file( ) и скачаем данные опросов, касающихся аренды и продажи жилья в 2006 году в указанной местности. Загрузим эти данные в пакет R-Studio с использованием книги кодирования, описывающей имена переменных. Скачаем Excel-таблицу из данных Nature Gas Aquisition Program. Прочитаем строки 18-23 и столбцы 7-15 в пакете R-Studio, присвоив результат переменной с именем dat. Рассчитаем значение выражения sum(dat$Zip*dat$ext,na.rm=T). Прочитаем XML данные о ресторанах указанной местности и указанного сайта. Рассчитаем количество ресторанов ,которые имеют zipcode 21231.

Практическая работа «Инструменты и пакеты». Цель работы: изучить пакеты: devtools, roxygen2, testthat, knitr. Создадим R-package: File| New Project| New Directory| R Package|. Он содержит следующие функции: create_data() и build_plot(). Заполним description. На панели

инструментов packages определим созданный пакет.

№	Название основной версии журнала	Название версии журнала в Scopus	ISSN print (Scopus)	E-ISSN (Scopus)	SourceID	Охват	Дата принятия (с 2015-)	Издательство версии журнала (Scopus)	DOI prefix (Jul 16)	Platform	Medline (см. лист Comments)	ASJC Codes
J001	Ab Imperio	Ab Imperio	2164-9731		21100316466	2013-ongoing		Ab Imperio	no	own		3312; 1202;
J002	Acarina	Acarina	0132-8077		19700173205	2009-ongoing		KMK Scientific Press	no	own		1109;
J003	Акустический журнал	Acoustical Physics	1063-7710	1562-6865	12922	1996-ongoing		Pleiades International	10.1134*	Springer		3102;
J004	Acta Naturae	Acta Naturae	2075-8251		21100256975	2013-ongoing		Park Media Ltd	no	own		1305; 1312; 1303; 1313;
J005	Актуальные проблемы теории и истории искусства	Actual Problems of Theory and History of Art	2312-2129		n/a yet	2016-ongoing	2016.дек-12	NP-Print	10.18688/	own		n/a yet
J006	Успехи геронтологии	Advances in gerontology	1561-9125		52741	2001-ongoing		Pleiades International	10.1134*	Springer		2700;
J007	Акушерство и гинекология	Akusherstvo i Ginekologiya	0300-9092	2412-5679	30069	1965-ongoing	2016.апр-08	Bionika Media Ltd.	10.18565	own		2700;
J008	Алгебра и логика	Algebra and Logic	0002-5232	1573-8302	144635	1968-ongoing		Springer	10.1007*	Springer		2603; 2609;
J009	Аналитика и контроль	Analitika i Kontrol	2073-1442	2073-1450	2,1101E+10	2016-ongoing	2016.сен-07	Ural Federal University na the first President of Russia B.N. Yeltsin	10.15826	own		n/a yet
J010	Analysis Mathematica	Analysis Mathematica	0133-3852	1588-273X	24691	1975-ongoing		Springer	10.1007*	Springer		2600;
J011	Анестезиология и реаниматология	Anesteziologiya i Reanimatologiy	0201-7563		21439	1976-ongoing		Meditsina Publishers	no	own	Medline	2703;
J012	Ангиология и сосудистая хирургия	Angiologiia i sosudistaia khirurgiia = Angiology and vascular surgery	1027-6661		20464	2003-ongoing		Infomedia Publishers	no	own	Medline	2700;
J013	Антибиотики и химиотерапия	Antibiotiki i Khimioterapiya	0235-2990		19486	1988-ongoing		OKI	no	own		2725; 2404; 2726;
J014	Прикладная биохимия и микробиология	Applied Biochemistry and Microbiology	0003-6838	1608-3024	16806	1996-ongoing,		Pleiades International	10.1134*	Springer		1303; 2402;
J015	Прикладная эконометрика	Applied Econometrics	1993-7601	2410-6445	n/a yet	2016-ongoing	2016.апр-09	CEMI RAS	no	REPEC		n/a yet
J016	Applied Magnetic Resonance	Applied Magnetic Resonance	0937-9347		27021	1990-ongoing		Springer	10.1007*	Springer		3107;
J017	Прикладная физика	Applied Physics	1996-0948	1432-0630	2,11E+10	2011-ongoing		Federal Informational-Analytical Center of the Defense Industry	no	own		3100
J018	Археология, этнография и антропология Евразии	Archaeology, Ethnology and Anthropology of Eurasia	1563-0110		4900152805	2006-ongoing		Institute of archaeology and ethnography SB RAS	10.17746	own		3316; 1204;
J019	Архив патологии	Arkhiv Patologii	0004-1955		27572	1950-ongoing		Media Sphera	10.17116	own		2734;
J020	Артериальная гипертензия	Arterial Hypertension	1607-419X	2411-8524	2,1101E+10	2017	2017.янв-27	All-Russian Public Organization "Antihypertensive League"	10.18705/	own		n/a yet
J021	Arthropoda Selecta	Arthropoda Selecta	0136-006X		20500195051	2011-ongoing		KMK Scientific Press	no	own		1105; 1109;
J022	Письма в Астрономический журнал	Astronomy Letters	1063-7737	1562-6873	26758	1996-ongoing		Pleiades International	10.1134*	Springer		3103; 1912;
J023	Астрономический журнал	Astronomy Reports	1063-7729	1562-6881	26760	1996-ongoing		Pleiades International	10.1134*	Springer		3103; 1912;
J024	Астрофизический бюллетень	Astrophysical Bulletin	1990-3413	1990-3421	1,97E+10	2009-ongoing		Pleiades International	10.1134*	Springer		3105; 3103;
J025	Оптика атмосферы и океана	Atmospheric and Oceanic Optics	1024-8560	2070-0393	21100431105	2009-ongoing	2015.авг-22	Pleiades International	10.1134*	Springer		1904; 1902; 3107; 1910;
J026	Атомная энергия	Atomic Energy	1063-4258	1573-8205	29827	1992-ongoing		Springer	10.1007*	Springer		2104;
J027	Автоматика и вычислительная техника	Automatic Control and Computer	0146-4116		24906	1973-ongoing		Allerton Press Inc.	10.3103*	Springer		1711; 1712; 2207;
J028	Автоматика и телемеханика	Automation and Remote Control	0005-1179	1608-3032	24950	1996-ongoing		Pleiades International	10.1134*	Springer		2207;

Рисунок 9. Фрагмент списка журналов, индексируемых в Scopus.

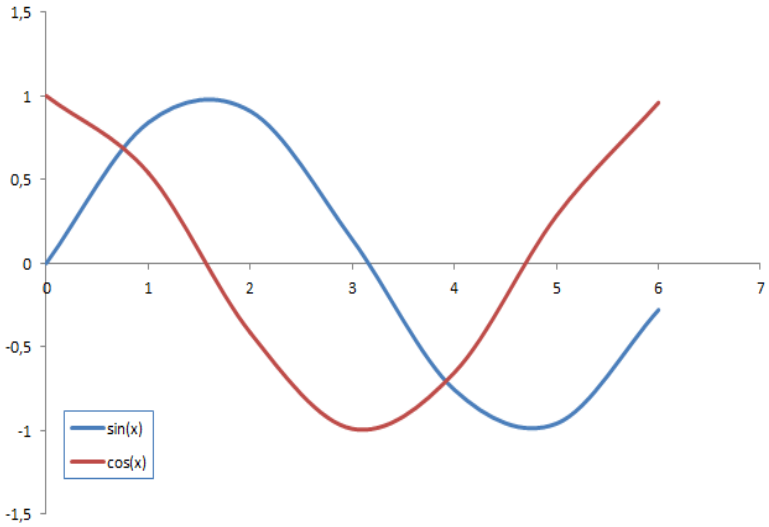

Рисунок 10. Пример построения графиков в R-Studio

Практическая работа «Функции». Цель: изучить функции создания и визуализации текстового ряда данных create_data() и build_plot(). Каждая функция содержит не менее пяти параметров. Текстовый ряд, созданный с использованием функции create_data(), может быть разных видов: временной; функция зависимости; ряд распределения и др. Визуализация ряда данных, созданная функцией build_plot(), формирует не менее трех видов графиков. Создадим описание функций в директории R с помощью команды devtools::document(). Файлы .Rd генерируются автоматически, исходя из roxygen comments в коде пакета, и сохраняются в директории man/.

Практическая работа “XLS-таблицы”. Цель: изучить технологические приемы пакета R-Studio при поиске и обработке информации из списка по заданным параметрам. Список с данными об автомобилях представлен в формате xls и содержит несколько тысяч записей. Фрагмент списка представлен на рисунке 11. С использованием инструментов пакета R-Studio выведем список автомобилей годом выпуска от 2015 и позднее. Выведем список автомобилей с мощностью двигателя менее 100 лошадиных сил.

Выведем список автомобилей марки BMW 3-й модели в серии купе. Выведем список автомобилей марки Mitsubishi с типом двигателя дизель и полным приводом. Выведем столбчатую диаграмму, содержащую соотношение мощности двигателя к его объему автомобилей марки BMW. Выведем список автомобилей марки BMW 5-й модели в серии седан, отсортированный по расходу топлива в городе на 100 км. Выведем список автомобилей марки Mitsubishi модели Lancer Evolution, отсортированный по оборотам максимального крутящего момента. Выведем столбчатую диаграмму, содержащую соотношение мощности двигателя к его объему автомобилей марки Mitsubishi модели Eclipse. Выведем кольцевую диаграмму, содержащую информацию о процентном соотношении моделей автомобилей марки Mitsubishi. Вывести информацию об автомобилях длиной больше 4500 мм, мощностью двигателя больше 200 лошадиных сил и годом выпуска ранее 2008.

208720 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118d AT (143 л.с.)
208729 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118d MT (143 л.с.)
208715 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118i AT (143 л.с.)
208722 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118i AT (143 л.с.)
208723 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118d AT (143 л.с.)
208727 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118d AT (143 л.с.)
208714 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	118i MT (143 л.с.)
208718 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120d AT (177 л.с.)
208725 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120d MT (177 л.с.)
208731 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120d MT (177 л.с.)
208716 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120i AT (170 л.с.)
208721 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120i AT (170 л.с.)
208724 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120i MT (170 л.с.)
208728 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	120i MT (170 л.с.)
208713 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	123d AT (204 л.с.)
208717 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	123d AT (204 л.с.)
208719 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	123d AT (204 л.с.)
208726 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	123d MT (204 л.с.)
208730 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	123d MT (204 л.с.)
208732 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	123d MT (204 л.с.)
208734 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	125i AT (218 л.с.)
208736 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	125i AT (218 л.с.)
208739 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	125i MT (218 л.с.)
208743 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	125i MT (218 л.с.)
208735 BMW	БМВ	1 серия	E81/E82/E87/E88 [рестайлинг]	2007	2012 Кабриолет	128i AT (233 л.с.)

Рисунок 11. Фрагмент списка данных об автомобилях

Практическая работа «Статистические модели». Цель: изучить технологию построения статистических моделей с использованием статистических тестов. Данные о температуре и напряжении солнечных панелей представлены в формате xls. Список действующих солнечных батарей и панелей, а также их технических параметров составляет сотни миллионов записей. Фрагмент списка представлен на рисунке 12. В базовом пакете R-Studio доступен широкий диапазон функций. Существует также большое количество других пакетов, которые увеличивают потенциальные возможности R. В пакете stats есть дополнительные статистические модели, в первую очередь glm – обобщенная линейная модель, позволяющая, например, моделировать логистические или логарифмические зависимости. Пакеты nlme, mgcv позволяют строить нелинейные модели. Некоторые функции позволяют пользователю отобразить полученной модели, среди которых summary() выводит определенный набор статистических параметров (статистические тесты); residuals() отображает остатки регрессии; predict() – прогнозные значения; coef() – отображает вектор с оценками параметра [7].

Определим выборочное среднее значение напряжения панелей 1-14. Определим выборочное значение температуры панелей 2-8. Определим выборочную дисперсию и выборочное среднеквадратичное отклонение напряжения панели 1, 2 и 5. Определим выборочную дисперсию и выборочное среднеквадратичное отклонение температуры панелей 3, 4 и 6. Проверим гипотезу о нормальном распределении генеральной совокупности напряжений с помощью критерия Пирсона. График зависимости среднего напряжения от времени представлен на рисунке 13. Проверим гипотезу о нормальном распределении генеральной совокупности температур с помощью критерия Пирсона. Сравним дисперсию напряжений и температур с помощью критерия Фишера. Сравним две генеральные совокупности (напряжение и температуру) с помощью критерия Стьюдента. С использованием критериев Барлетта и Кохрана сравним несколько дисперсий температуры и напряжений любых панелей. График зависимости средней температуры от времени представлен на рисунке 14.

DateTime	Module 1.6_16	Module 1.6_16	Module 1.6_1	Module 1.6_1	Module 1.6_2	Module 1.6_2	Module 1.6_3	Module 1.6_3	Module 1.6_4
11.10.2017 9:00	31,9477783	33,1317284	31,9477783	33,1317284	31,9477783	33,1317284	31,9477783	33,1317284	31,9477783
11.10.2017 9:15	13,9	30,16	23,6	29,78	11,9	29,63	13,4	27,68	16
11.10.2017 9:30	15,9	32,64	25,6	32,71	13,9	32,6	15,4	31,7	18
11.10.2017 9:45	15,9	32,8	25,6	32,81	15,4	32,7	16,9	31,91	19,5
11.10.2017 10:00	16,4	32,66	27,1	33,06	15,4	32,76	17,4	32,1	19,5
11.10.2017 10:15	16,4	32,03	25,6	32,23	15,4	32,1	16,9	31,22	19,5
11.10.2017 10:30	16,4	33,12	25,6	33,2	15,9	33,12	16,9	32,34	19,5
11.10.2017 10:45	17,4	33,2	25,6	33,3	15,9	33,11	18,9	32,37	19,9
11.10.2017 11:00	17,4	33,15	21,8	33,35	15,9	33,15	18,9	32,53	20
11.10.2017 11:15	17,4	33,86	25,6	33,98	17,4	33,9	18,9	33,3	21
11.10.2017 11:30	25,2	34,44	27,1	34,66	25,2	34,65	32,3	34,15	22
11.10.2017 11:45	20,5	35	25	35,37	20	35,3	20,5	35,07	24,1
11.10.2017 12:00	22,5	34,6	31,2	35,2	22	34,99	23,5	34,78	25,6
11.10.2017 12:15	24	34,34	31,2	34,93	23,5	34,74	25	34,49	27,6
11.10.2017 12:30	24	34,45	31,2	34,6	23,5	34,6	25,5	34,5	27,6

Рисунок 12. Фрагмент списка действующих солнечных батарей и панелей

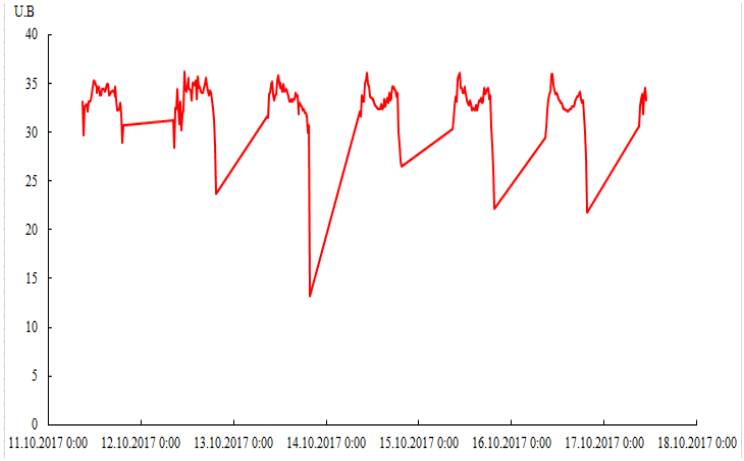

Рисунок 13. Зависимость среднего напряжения от времени

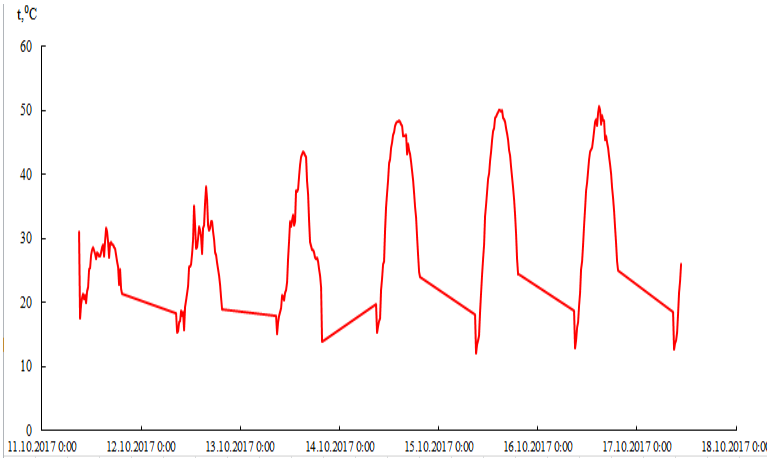

Рисунок 14. Зависимость средней температуры от времени

Заключение и перспективы развития. Разработаны и апробированы со студентами третьего курса около 20 практических работ по изучению инструментов и технологических возможностей платформы R-Studio. Часть работ представлены в этой статье. Как уже отмечалось выше, все это стало возможным в результате сотрудничества кафедры инженерной психологии и эргономики БГУИР с компанией BEZNext и IBM BDU, а также в результате освоение на кафедре образовательной программы Ambassador от IBM BDU.

При работе с большими данными и при организации практических занятий со студентами необходимо было организовать сбор и хранение больших массивов информации. Требовалось либо расширить компьютерные мощности и хранилища кафедры, либо организовать альтернативные пути. И опять, здесь пришли организационные решения благодаря сотрудничеству с упомянутыми компаниями BEZNext и IBM BDU.

Не так давно корпорация IBM вложила более миллиарда долларов США в развитие облачной платформы IBM Cloud, включая хостинг Softlayer. IBM Cloud представляет собой мощный комплекс облачных технологий и сервисов, позволяющий решать очень широкий круг задач клиентам компании. Как известно, облачные вычисления – это модель предоставления вычислительных ресурсов (от отдельных приложений до крупных центров обработки данных) через интернет с оплатой по факту использования. Центр обработки данных представляет собой очень крупное высокотехнологичное сооружение для размещения вычислительного оборудования. Изначально такие центры использовались, в основном, для собственных нужд крупных организаций. В последнее время данный термин получил широкое распространение в коммерческой сфере, ввиду возрастания интереса к услугам частных центров обработки данных и спроса на них среди клиентов. Сейчас центры обработки данных предлагают свои клиентам целый комплекс телекоммуникационных услуг, связанных с хранением и обработкой информации [8].

Таким образом, сейчас центры обработки данных – это сложные системы, которые включают в себя целый комплекс IT решений, высокотехнологичного оборудования и инженерных конструкций. Основная задача такого центра заключается в быстрой обработке любого объема данных, хранении информации и ее выдаче в стандартизированном виде пользователю. Фактически ядром центра являются мощные серверные станции, снабженные необходимым программным обеспечением, системами охлаждения и безопасности [8].

Платформа IBM Cloud представляет собой набор облачных сервисов для хранения данных. Как известно, облачные сервисы используют принцип виртуализации, то есть создание программных продуктов и развертывание серверного оборудования виртуально. Виртуализацию применяют для приложений, серверов, систем хранения данных и локальных сетей. Это достаточно эффективный способ сокращения расходов на создание ИТ-инфраструктуры кафедры. IBM Cloud обеспечивает мгновенный доступ к необходимым вычислительным ресурсам и службам для быстрого старта, непрерывного развития и успешного масштабирования. Службы для мобильных приложений, Интернета вещей, Watson и многого другого делают IBM Cloud удобной платформой для нового поколения приложений [9].

Кафедра инженерной психологии и эргономики получила доступ к облачной платформе от IBM, которая позволяет хранить и работать с большими объемами информации. Хранилище объемом 25 Гигабайт обеспечивает хранение неструктурированных данных. IBM Cloud берет на себя настройку инфраструктуры, подготовку ресурсов и хостинг.

Список литературы

[1]. Шипунов А.Б., Балдин Е.М. Анализ данных с R - Москва: ДМК Пресс, 2014 – 148 с.

[2]. Использование языка R с базами данных – Ресурс IBM для разработчиков и IT-специалистов – Режим доступа: https://www.ibm.com/developerworks/ru/library/dm-1402db2andr/index.html - Дата доступа: 26.02.2018

[3]. M.P. Batura, S.K. Dzik, B. Zibitsker, D.V. Lihachevsky, I Tsyrelchuk, K.D. Yashin. Experience in organizing educational process in BIG DATA analytics at BSUIR //Сборник материалов III международной

практической конференции «BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня», Минск, Республика Беларусь, 3-4 мая 2017 года

[4]. База фильмов – Data Science and Cognitive Computing Courses– Режим доступа: https://cognitiveclass.ai/courses/r-101// - Дата доступа: 26.02.2018

[5]. Список книг, индексируемых в Scopus – Официальный интернет-портал издательского дома Elsevier– Режим доступа: http://www.elsevierscience.ru/products/scopus// - Дата доступа: 26.02.2018

[6]. Акберова Н.И. Краткое введение в R и R-Studio – Казань: КФУ, 2014 – 33 с.

[7]. Савельев А.А., Мухарамова С.С., Пилюгин А.Г. Использование языка R для статистической обработки данных – Казань, КГУ, 2007

[8]. Центр обработки данных IBA Group – Ресурс IBM Беларусь – Режим доступа: http://iba.by/services/datacenter/ - Дата доступа: 26.02.2018

[9]. IBM Cloud – Ресурс IBM Cloud – Режим доступа: https://www.ibm.com/cloud/ - Дата доступа: 26.02.2018

R-STUDIO FOR ANALYSIS OF BIG DATA

V.S.DROZDOV
Assistant of the department of Human Engineering and Ergonomics BSUIR, Master of Technical Sciences

D.V. LIKHACHEUSKI, PhD
Dean of the Faculty of Computer Design BSUIR, Associate professor

E.A. MELNIKOVA
Assistant of the department of Human Engineering and Ergonomics BSUIR, Master of Technical Sciences

V.S. OSIPOVICH, PhD
Associate professor of department of Human Engineering and Ergonomics

N.V.SCHERBINA
Senior Lecturer of the department of Human Engineering and Ergonomics BSUIR, Master of Technical Sciences

K.D.YASHIN, PhD
Head of the Department of Engineering Psychology and Ergonomics BSUIR, Associate Professor

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: kafipie@bsuir.by, e.melnikova@bsuir.by

Abstract. The methods of working in the graphical data processing environment using the R-studio platform are considered, the main technological methods are described. The practical works for teaching students are presented, as well as the results of their implementation.

Keywords: analysis, vector, matrix, distribution, probability, big data, R, R-Studio.