ТИМОВКИН С. Н., ЕГОРОВА Д. К.

РЕАЛИЗАЦИЯ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ В ВЫЧИСЛИТЕЛЬНОЙ СРЕДЕ R

Аннотация. В работе рассмотрена реализация метода главных компонент в вычислительной среде R. В качестве исходных данных использовались рейтинговые показатели вузов.

Ключевые слова: вычислительная среда R, анализ, визуализация, многомерные данные, снижение размерности, метод главных компонент.

TIMOVKIN S. N., EGOROVA D. K.

IMPLEMENTATION OF PRINCIPLE COMPONENT ANALYSIS IN COMPUTING ENVIRONMENT R

Abstract. The article considers the principal component analysis in the computing environment R. The initial data are the ratings of the universities.

Keywords: computing environment R, analysis, visualization, multidimensional data, dimensional reduction, principal component analysis.

Среда R является программным средством, предназначенным для анализа многомерных данных и их визуализации, с мощными графическими возможностями. Среди множества других видов программного обеспечения ее выделяет то, что это среда с открытым исходным кодом, развиваемая в рамках проекта по разработке свободного программного обеспечения (GNU), причем функциональность ее базовой версии может быть значительно увеличена путем применения расширенной системы команд или пакетов, которые так же находятся в свободном доступе. В данной среде используется интерфейс командной строки, но для удобства работы созданы и интегрированные среды разработки (IDE), например, RStudio, JGR, RKWard, SciViews-R, R Commander. В данной работе рассмотрена реализация метода главных компонент в RStudio, которая написана на C++ и использует фреймворк Qt для графического интерфейса пользователя [1].

Рассмотрим следующую задачу: требуется провести анализ и визуализацию рейтинговых показателей вузов. В качестве исходных данных были выбраны двадцать семь вузов Приволжского федерального округа, городов Москва и Санкт-Петербург. Данные взяты с интернет-сайта Рейтинг вузов России [2]. Перечислим их: КФУ, КГМУ, КНИТУ, КНИТУ-КАИ им. А.Н. Туполева, СамГТУ, СамГМУ, СамГУ, САУ им. ак. С.П. Королёва, МГУ им. Н.П. Огарёва, НГТУ им. Р. Е. Алексеева, ННГУ им. Н.И. Лобачевского, ИжГТУ им. М. Т. Калашникова, СГУ им. Н. Г. Чернышевского, СГМУ им. В. И. Разумовского,

ПНИПУ, ПГНИУ, МИ «РЕАВИЗ», НИЯУ, НИУ ВШЭ, ТПУ, ТГУ, НГУ, ИТМО, МИФИ, НИТУ «МИСиС», НИУ «МЭИ», МАИ, НИУ «МИЭТ». Каждый из вузов оценивается по девяти рейтинговым показателям, подразделенными на три основные группы и итоговым рейтинговым функционалом.

Условия для получения качественного образования

1. Международная интеграция.
2. Востребованность среди абитуриентов.
3. Уровень преподавания.
4. Ресурсное обеспечение.

Востребованность у работодателя

5. Сотрудничество с работодателями.
6. Качество карьеры выпускников.

Научно исследовательская активность

7. Инновационная активность.
8. Научные достижения.
9. Инновационная инфраструктура.

Визуализация рейтинговых показателей возможна, если будет произведено снижение размерности исходного признакового пространства с девяти показателей к двум или трем, тогда каждый вуз будет характеризоваться двумерной или трехмерной точкой в декартовой прямоугольной системе координат. В качестве метода снижения размерности был выбран метод главных компонент [3]. Согласно данному методу, переход к новому набору показателей осуществляется таким образом, чтобы выполнялось условие $I_{p'}\left(\tilde{Z}(X)\right) = \max\limits_{Z \epsilon F}\{I_{p'}(Z(X))\}$, где $I_{p'}$ – мера информативности, $Z(X)$ – p-мерная вектор-функция исходных переменных, $\tilde{Z}(X)$ – p'-мерная вектор-функция новых переменных и $p' \ll p$. В качестве класса допустимых преобразований F берут всевозможные линейные ортогональные нормированные комбинации исходных показателей и переходят к p' первым главным компонентам по формуле

$$z^{(1)}(X) = c_{j1}(x^{(1)} - \mu^{(1)}) + \cdots + c_{jp}(x^{(p)} - \mu^{(p)}).$$

Здесь на коэффициенты накладываются условия нормировки $\sum_{\nu=1}^{p} c_{j\nu}^2 = 1, = 1,2, \dots, p'$, $\sum_{\nu=1}^{p} c_{j\nu} * c_{k\nu} = 0,\ k = 1,2, \dots, p, j \neq k$, где $\mu^{(\nu)}$ – математическое ожидание $x^{(\nu)}$. Критерием оптимальности p'-мерной системы новых показателей является выражение $I_{p'}(Z(X)) = \frac{Dz^{(1)} + \cdots + Dz^{(p')}}{Dx^{(1)} \dots + Dx^{(p')}}$, где D – дисперсия.

Для реализации метода главных компонент в среде R к ее базовой версии могут быть подключены пакеты *ade4* и *vegan* (далее в работе будут приведены отрывки кода R с соответствующими комментариями) [4]. Следующая команда R ставит в соответствие каждому из двадцати семи вузов вектора с рейтинговыми показателями, затем все они записываются в массив:

```
VUZ<-rbind(KazanPrFedUn,SamGosAeroUn,KazGosMedUnZdr,NizhGosUnLob,SamGosMedUnZdr,
PermNazIslPolUn,SamGosTechUn,KazNazIslTechUn, KazNazIslUnTupol, SamGosUn,
MordGosOGAREVA,NizhGosTechAleks,IzhGosTechKalash, SarGosUnChernish,
SarGosMedRazumovsk, PermGosNazIslUn, MedInstReaviz, MIFI,VShEco, TomsPolTec,
Novosib,TomsGos,MISiS,MEI, SPBInfTech, MoskAvia,MIET)
```

Далее каждому рейтинговому показателю (компоненте) присвоены имена:

```
colnames(VUZ) <- c("MezhdIntegr", "VostrebAbitur", "UrovenPrepod", "ResursObesp",
"SotrudnSRabotodmi", "KachestvoKareriVip", "InnovatzAktiv", "NauchnDostig",
"InnovationInfrastr")
```

Затем была произведена оценка взаимосвязи исходных данных путем составления корреляционной таблицы, из которой можно отметить достаточно высокую взаимосвязь:

- «Международной интеграции» с «Уровнем преподавания», «Ресурсным обеспечением», «Востребованностью абитуриентами» (входят в группу «Условия для получения качественного образования»);
- «Востребованности абитуриентами» с «Ресурсным обеспечением»;
- «Уровня преподавания» с «Ресурсным обеспечением» и «Качеством карьеры выпускников»;
- «Ресурсного обеспечения» и «Качества карьеры выпускников»;
- «Инновационной активности» и «Инновационной инфраструктуры».

Полное отсутствие корреляции наблюдается у «Сотрудничества с работодателями» со всеми другими критериями.

Следующие команды R создают объект VUZ.pca и формируют главные компоненты:

```
VUZ.pca <-prcomp(VUZ)
pc <-VUZ.pca$x
round(pc,4)
```

Результат выполнения команд приведен на рисунке 1.

```
##                      PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## KazanPrFedUn      -3.1334  0.0279 -0.6446 -1.0625 -0.4141  0.0439 -0.1118  0.3969  0.2303
## SamGosAeroUn      -0.2790  0.3641 -0.3254  1.9837 -0.1741  0.1249 -0.1059 -0.0433  0.1072
## KazGosMedUnZdr     1.1261 -1.1972  1.8275  0.1365  0.5045 -0.7698  0.2700 -0.4384  0.3910
## NizhGosUnLob      -0.1768 -0.5181 -1.0181  0.4141  0.8561  0.1161 -0.1419  0.3750  0.0242
## SamGosMedUnZdr     2.0767  0.3743  1.1172  0.1166  0.0784  0.0760  0.1759  0.7126  0.2544
## PermNazIslPolUn    1.4958  1.4356 -0.9270  0.5311  0.0104 -0.0639 -0.2668  0.0643 -0.0757
## SamGosTechUn       1.9513  0.9483 -0.0364  0.5204 -0.7608  0.1829  0.9649  0.2349 -0.1226
## KazNazIslTechUn    1.7175 -0.0582 -0.6398 -0.4909  0.3439 -0.0977  0.6258  0.0529  0.2390
## KazNazIslUnTupol   1.5016 -0.3023 -0.6120 -0.5417 -0.8341  0.1751 -0.7153 -0.1706  0.0314
## SamGosUn           1.6339 -1.6237 -1.1333  0.3075 -0.9972 -0.7228  0.7535 -0.4096 -0.1440
## MordGosOGAREVA     2.0398  1.2830 -0.5847 -0.5343  0.3181  0.7254  0.2646 -0.5878  0.1331
## NizhGosTechAleks   2.6496  0.9065  0.3103  0.4377 -0.0255 -0.0771  0.0672 -0.0828 -0.0775
## IzhGosTechKalash   2.3457  1.0751  0.1649 -0.4581  1.3018 -0.1682  0.1481  0.2080 -0.7001
## SarGosUnChernish   1.7140 -0.3955 -2.1058 -0.2580  0.2547 -0.7889 -0.5482 -0.1475  0.2451
## SarGosMedRazumovsk 3.1626 -0.7705  1.4617 -0.2735 -0.1353  0.1855 -0.1671  0.0274  0.0741
## PermGosNazIslUn    2.4908 -0.8453 -0.6263 -0.0153 -0.6259  0.4602 -0.0974  0.3952 -0.2118
## MedInstReaviz      3.5688 -0.5861  1.2731 -1.0084 -0.2158 -0.6027 -0.2380 -0.0183 -0.2238
## MIFI              -4.6318  0.5311  0.2319  0.4380  0.4475 -1.1256  0.0346  0.0502  0.0590
## VShEco            -4.1826  0.0411  0.8583  0.9165 -0.5500  0.3969  0.1132 -0.2666 -0.1983
## TomsPolTec        -4.5735 -0.0704 -0.3439 -0.6626 -0.0798 -0.2433  0.1234  0.0173 -0.2283
## Novosib           -3.9651 -0.2649  0.5397 -0.0130 -0.7503 -0.5753 -0.3352  0.4370 -0.2086
## TomsGos           -3.0387  0.4365 -0.2848 -0.3757  0.6477  0.2017  0.5780 -0.1619  0.2007
## MISiS             -2.9984 -1.2168  0.0758 -0.6255  0.1696  1.0558  0.3805  0.3010  0.2040
## MEI               -2.3315  0.9768  0.5406 -0.7264 -0.4947  0.5540 -0.2490 -0.7005 -0.1793
## SPBInfTech        -1.7680 -1.4494 -0.0191  0.4729  1.1425  0.2637 -0.4130 -0.2738 -0.2362
## MoskAvia           0.1359  2.2645  0.6430  0.0243 -0.2412 -0.1620 -0.5934  0.0372  0.3617
## MIET               1.4687 -1.3663  0.2573  0.7468  0.2237  0.8353 -0.5168 -0.0088  0.0510
```

Рис. 1. Таблица главных компонент.

Выясним, как первая главная компонента связана с рейтинговым функционалом в таблице. Следующие команды создают вектор с рейтинговыми данными и вычисляют коэффициент корреляции главной компоненты и рейтингового функционала:

```
M<-c(3.4193,2.7737,2.5191,2.5113,2.1431,2.0928,2.0859,1.9914,1.9839,1.96,1.8972,
1.8894,1.8868, 1.7874,1.7649,1.6854,1.5942,4.1514,4.0377,3.9215,3.883,3.5104,
3.4247, 3.3442,3.0739, 2.7339,  2.1233)
cor(M,pc[,1])
## [1] -0.9870379
```

Высокое значение коэффициента корреляции свидетельствует о том, что рейтинговый функционал и первая главная компонента сильно взаимосвязаны. Первая главная компонента вбирает в себя максимум отличий между вузами. Таким образом, не зная способа подсчета рейтингового функционала (представленного в [2]), были восстановлены позиции вузов в таблице рейтинга.

Следующая команда визуализирует корреляцию первой компоненты с итоговым местом вуза в рейтинге (рис.2):

```
plot(Mesto,pc[,1])
```

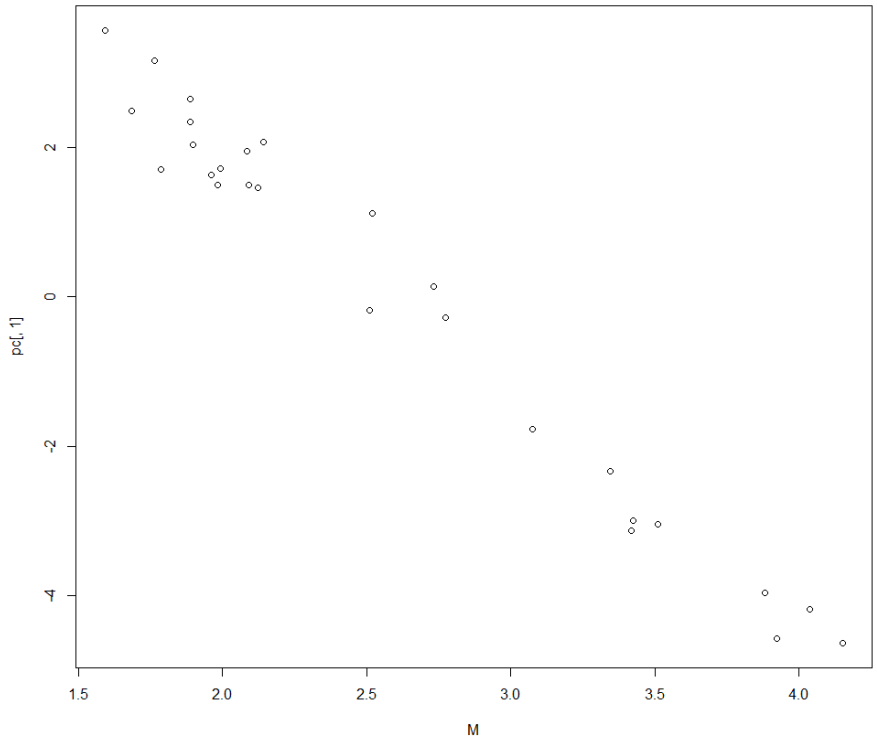

Рис. 2. Корреляционное поле.

На графике корреляционного поля между первой главной компонентой и значением рейтингового функционала наблюдается сильно выраженная прямая отрицательная корреляционная зависимость. Также можно отметить, что с увеличением рейтингового функционала значение главной компоненты убывает.

Следующая команда R позволяет найти основные статистические характеристики компонент. Результат приведен на рисунке 3.

```
summary(VUZ.pca)
```

```
## Importance of components:
##                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## Standard deviation     2.65 0.98407 0.88107 0.66817 0.59094 0.52228 0.41989 0.32713 0.24007
## Proportion of Variance 0.69 0.09518 0.07629 0.04388 0.03432 0.02681 0.01733 0.01052 0.00566
## Cumulative Proportion  0.69 0.78519 0.86148 0.90536 0.93968 0.96649 0.98382 0.99434 1.00000
```

Рис. 3. Основные статистические характеристики компонент.

Первая строка «Standard deviation» показывает среднеквадратичное отклонение главных компонент. С большим числом компонент отклонение уменьшается, что и должно быть. Вторая строка «Proportion of Variance» объясняет долю отклонений компонент. Из строки «Cumulative Proportion» следует, что первая главная компонента PC1 описывает 69%

дисперсии исходного набора данных, первая и вторая главные компоненты PC1 и PC2 объясняют приблизительно 78,5% исходных данных. Далее видно, что, например, пять главных компонент PC1-PC5 описывают около 94% набора исходных данных и т.д.

С помощью следующих команд извлечем в матрицу *v1* веса, с которыми старые переменные, результаты отдельных критериев вузов, входят в новые переменные (рис.4):

```
v1<-VUZ.pca$rotation
round(v1,4)
```

```
##                       PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## MezhdIntegr        -0.3935 -0.1671  0.3863 -0.4027  0.4748 -0.0001  0.2137 -0.0932 -0.4754
## VostrebAbitur      -0.2892 -0.1989  0.2497  0.6056  0.1579  0.2747 -0.4568 -0.3716  0.0244
## UrovenPrepod       -0.3487 -0.0331  0.1664 -0.3790 -0.2558 -0.5431 -0.5002 -0.0928  0.2945
## ResursObesp        -0.4139 -0.0866  0.1452 -0.1745 -0.2174  0.5977  0.0513  0.4909  0.3485
## SotrudnSRabotodmi  -0.0166  0.4682  0.5161  0.2358  0.2324 -0.2223  0.3704  0.0032  0.4668
## KachestvoKareriVip -0.3685  0.2241  0.1065  0.3046 -0.6358 -0.1483  0.2561  0.0094 -0.4669
## InnovatzAktiv      -0.2884  0.3055 -0.3192  0.2394  0.4051 -0.2187 -0.2844  0.5856 -0.1688
## NauchnDostig       -0.3960 -0.4685 -0.4369  0.1881  0.0995 -0.2629  0.4573 -0.1221  0.3071
## InnovationInfrastr -0.3041  0.5845 -0.4113 -0.2364  0.0684  0.2898  0.0052 -0.4957  0.0802
```

Рис. 4. Матрица весов.

Матрица *v1* показывает, как новые оси смещены относительно старых. Далее каждая компонента вычисляется как сумма произведений значений весов на соответствующий критерий. Например, первая главная компонента будет вычислена так:

$$PC1 = \text{MezhdIntegr} \times (-0.3935) + \text{VostrebAbitur} \times (-0.2892) + \text{UrovenPrepod} \times (-0.3487) +$$
$$+\text{ResursObesp} \times (-0.4139) + \text{SotrudnSRabotodmi} \times (-0.0166) + \text{KachestvoKareriVip} \times (-0.3685) +$$
$$+\text{InnovatzAktiv} \times (-0.2884) + \text{NauchnDostig} \times (-0.3960) + \text{InnovationInfrastr} \times (-0.3041).$$

В первом столбце матрицы *v1* показано, что критерий «Международная интеграция» входит в главную компоненту PC1 с весом -0.3935, «Востребованность абитуриентами» с весом -0.2892 и так далее. По таблице можно сказать, что наиболее существенные переменные – это «Международная интеграция», «Ресурсное обеспечение» и «Научные достижения» с весами -0.3935, -0.4139 и -0.3960 соответственно; критерий «Сотрудничество с работодателями» входит в первую главную компоненту с очень маленьким весом -0.0166.

Перед построением графика главных компонент визуализируем долю объясняемой дисперсии разными компонентами. Для этого используем следующую команду:

```
plot(VUZ.pca)
```

Результат приведен на рисунке 5.

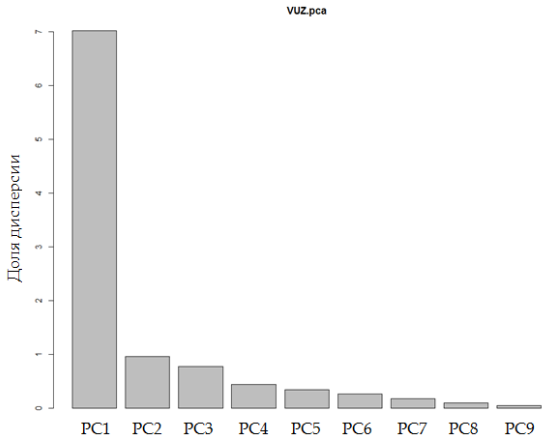

Рис. 5. График относительных вкладов компонент.

График показывает относительные вклады каждой компоненты в общий разброс данных. Хорошо видно, что компонент девять, как и признаков, но, в отличие от первоначальных признаков, наибольший вклад вносит первая главная компонента. Аналогичные результаты, но в текстовом виде, получили в «Результатах оценивания».

Визуализируем наблюдения в осях первой и второй главных компонент командой:

```
biplot(VUZ.pca, xlim = c(-0.8, 0.8), ylim = c(-0.6, 0.6))
```

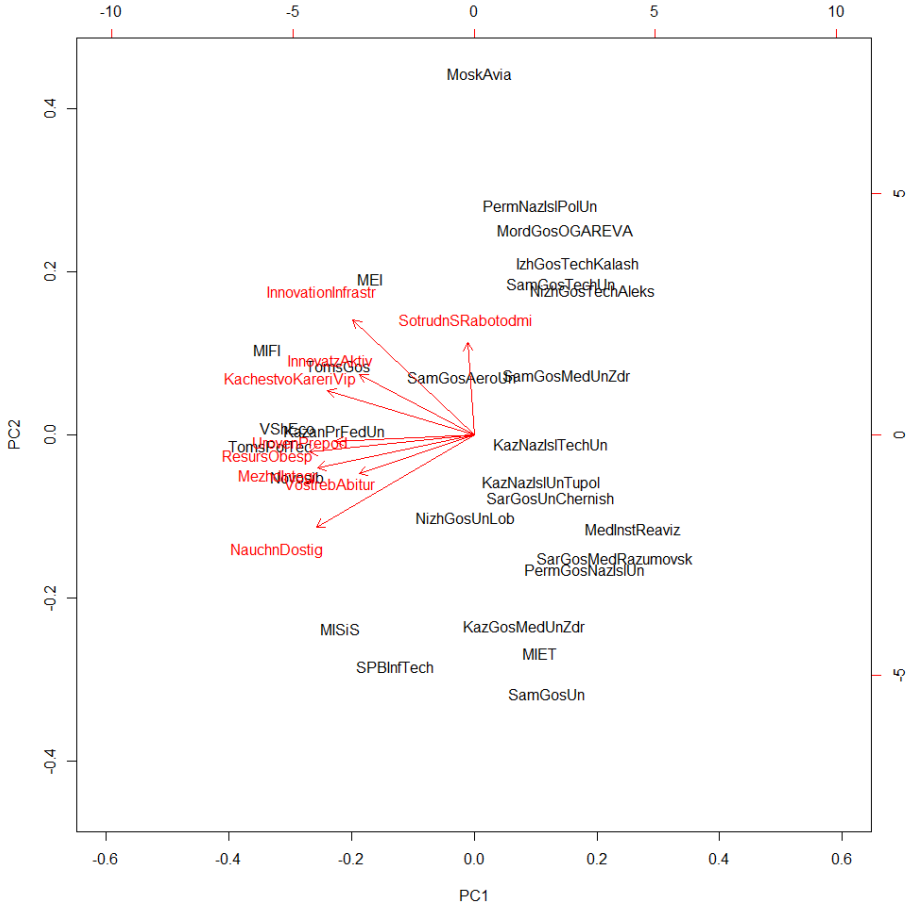

Рис. 6. Визуализация первой и второй главных компонент.

На рисунке 6 приведен график величины вклада каждого из девяти исходных признаков в первые две компоненты.

Красные векторы – это проекции единичных векторов исходных координат на плоскость первых двух главных компонент. Векторы, соответствующие исходным координатам, направлены влево. Это подтверждает то, что коэффициенты в первом столбце матрицы вращения отрицательные. При этом стрелки *SotrudnSRabotodmi, KachestvoKareriVip, InnovatzAktiv, InnovationInfrastr* направлены вверх и соответствующий коэффициент (вторая главная компонента) во втором столбце положительный, остальные – вниз и соответственно коэффициент отрицательный. Проекция вектора *SotrudnSRabotodmi* на ось первой главной компоненты почти нулевая, а на ось второй компоненты – положительная и заметно больше нуля, что также подтверждается таблицей.

Признаки *UrovenPrepod, ResursObesp, MezhdIntegr* и *VostrebAbitur* вносят очень маленький вклад во вторую главную компоненту, они входят в группу «Условия для получения качественного образования». Их корреляционная зависимость была найдена ранее. Можно предположить, что они дают информацию об общем статусе вуза: высшие учебные заведения занявшие самые высокие позиции в итоговой таблице (МИФИ, НИУ ВШЭ, НГУ, ТПУ) расположились именно в этой области, но мало информации о том, как это влияет на «Научно-исследовательскую активность» и «Востребованность работодателями».

ЛИТЕРАТУРА

1. Кабаков Р. И. R в действии. Анализ и визуализация данных в программе R. – М.: ДМК Пресс, 2016. – 588 с.

2. Рейтинг вузов России [Электронный ресурс]. – Режим доступа: https://top100universities.ru.

3. Айвазян А.С., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 609 с.

4. Шипунов А. Б., Балдин Е. М., Волкова П. А. и др. Наглядная статистика. Используем R [Электронный ресурс]. – Режим доступа: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf.