Отчет о летней производственной практике

на тему:

«Применение факторного анализа для кластеризации геологических проб»

Место выполнения: ИПММ, СПбПУ Петра Великого

Студент группы 3630102/70301\_\_\_\_\_<*подпись*>\_\_\_\_\_\_\_\_\_ Королевская К.Д.

Оценка научного руководителя:\_<*оценка*>\_\_ <*подпись*>\_\_ Баженов А.Н.

к.ф.-м.н., доцент высшей школы

прикладной математики и вычислительной физики

ИПММ, СПбПУ Петра Великого

**Оглавление**

[Введение 3](#_Toc46258356)

[Основная часть. 4](#_Toc46258357)

[Выводы 18](#_Toc46258358)

[Резюме 19](#_Toc46258359)

[Список литературы 19](#_Toc46258360)

Введение

Цель работы: проверить отличаются ли показатели, взятые с русского Севера и центральной Африки, исследовать по каким компонентам они различаются.

Задачи: изучить метод главных компонент, поработать с пакетами, необходимыми для обработки данных, представленных в виде эмиссионной матрицы.

Основная часть.

У нас есть набор 2D данных – следы жизни в геологических объектах. Образцы взяты с двух разных регионов: русский Север и центральная Африка. Известна область для каждой аминокислотых[1].

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Тип компонента | Буквенное обозначение |
| 320-350 | 420-480 | Humic-like | C |
| 250-260 | 380-480 | Humic-like | A |
| 310-320 | 380-420 | Marine Humic-like | M |
| 270-280 | 300-320 | Tyrosine-like, Protein-like | B |
| 270-280 | 320-350 | Tryptophane-like, Protein-like or phenol-like | T |

Таблица 1. Основные флуоресцентные компоненты

Ранее уже было получено разделение по критерию (отношение сложной и простой органики): для данных Севера ,а для Африки

.

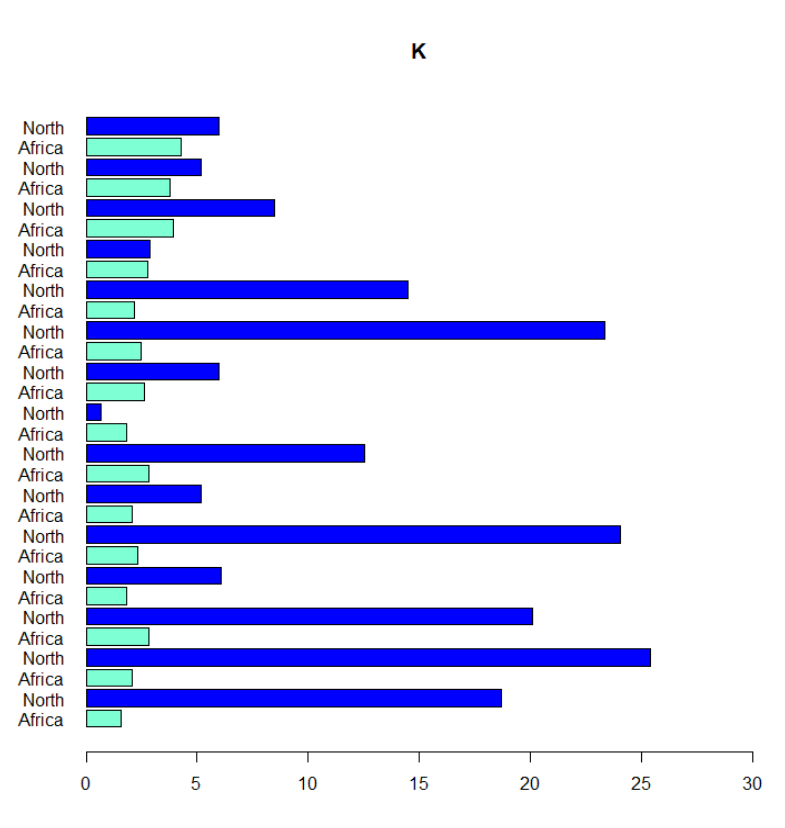


Рисунок 1. Коэффициент К.

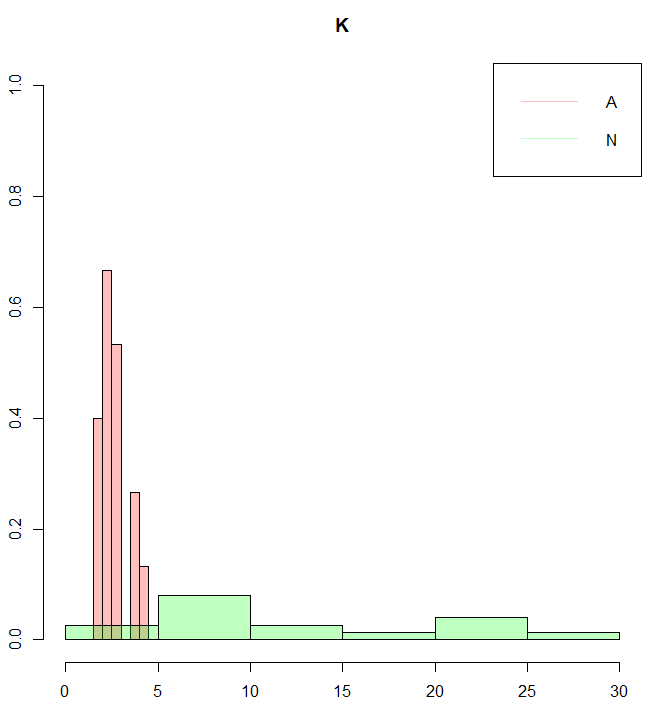


Рисунок . Гистограмма параметра К.

Также для данных была получена таблица:

|  |  |  |
| --- | --- | --- |
|  | Гуминовые компоненты (A + C) | Белковоподобные компоненты (T + B) |
| Африка |  |  |
| Север |  |  |

Таблица 2. Разделение данных по компонентам.

В ходе этой работы был изучен Метод главных компонент(МГК) [2].

МГК применяется к данным, записанным в виде матрицы. Перед применением метода данные необходимо отцентрировать и отнормировать. Цель этого метода – извлечение из этих данных нужной информации. В результате этого метода мы приходим от большого количества переменных к новому представлению, размерность которого значительно меньше.

С помощью этого метода мы раскладываем нашу матрицу X размерности на две:

где

Матрица T называется матрицей счетов, а матрица Р – матрицей нагрузок.

Алгоритм обработки каждого набора данных:

1. Считывание и визуализация данных:

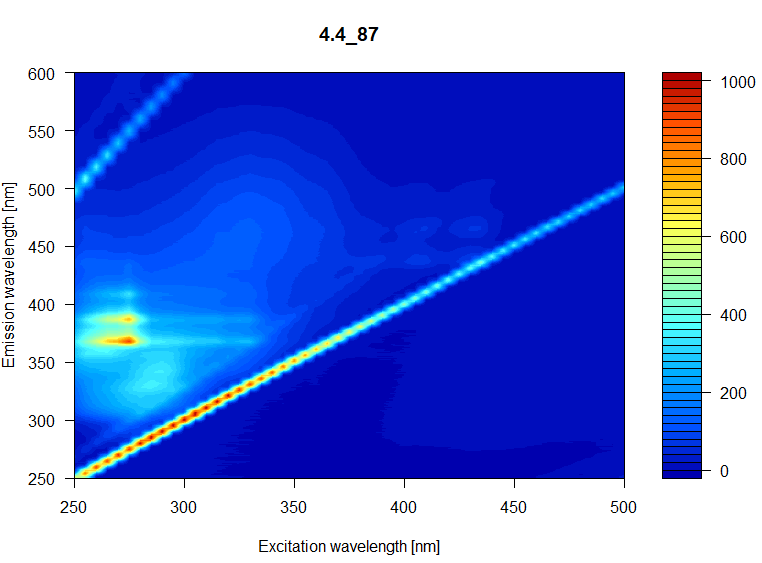


Рисунок 3. Визуализация файла 4.4\_87(Африка)

1. Обрезка графика

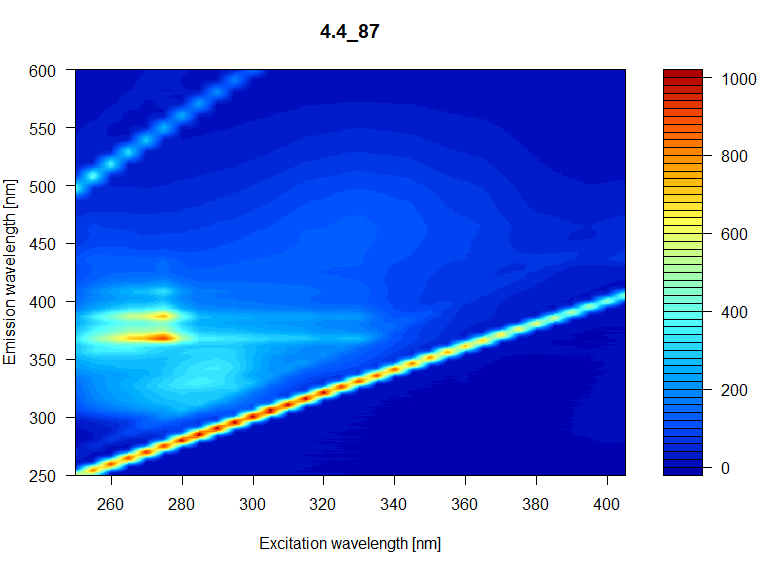


Рисунок 4. Обрезка графика (400:600)

1. Удаление лучей рэлеевского рассеяния

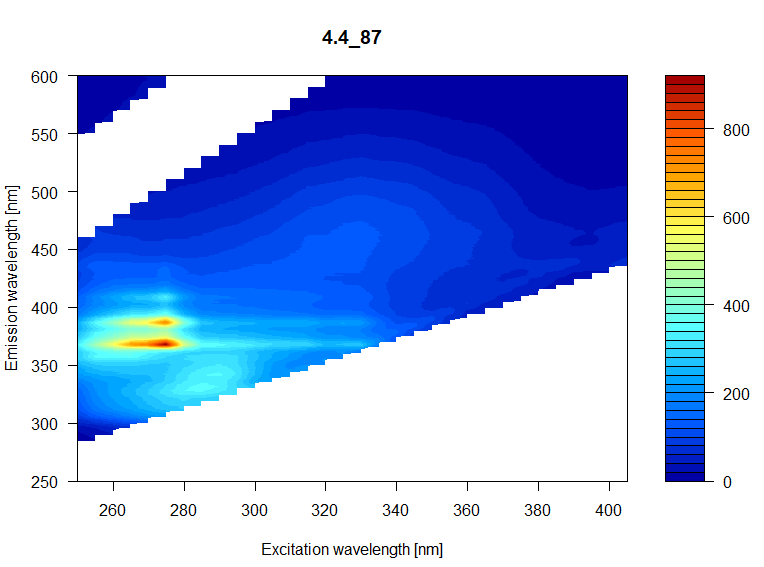


Рисунок 5.Удаление у пробы лучей рэлеевского рассеяния

1. Отображение областей пиков и вычисление интегралов интенсивности

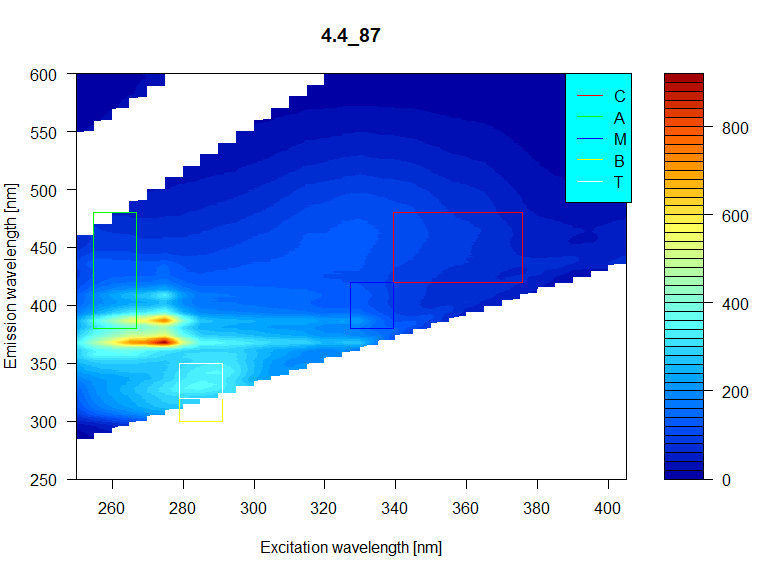


Рисунок 6. Области пиков

Полученные результаты для всех файлов:

|  | **NameFile** | **Region** | **C** | **A** | **M** | **B** | **T** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1.1\_70. | Africa | 27637.371 | 15901.663 | 8280.282 | 8847.592 | 18994.630 |
| **2** | 1701 | North | 254510.671 | 106219.655 | 53529.277 | 7156.039 | 12150.725 |
| **3** | 1.2\_21 | Africa | 44299.522 | 17860.389 | 10944.548 | 9787.680 | 20385.952 |
| **4** | 1702 | North | 208747.601 | 98056.478 | 47482.860 | 4114.232 | 7973.801 |
| **5** | 1.3\_68 | Africa | 84971.269 | 33205.100 | 22349.206 | 12679.743 | 28993.264 |
| **6** | 1704 | North | 62079.794 | 44620.021 | 15616.355 | 2051.332 | 3262.508 |
| **7** | 1.4\_114 | Africa | 7808.704 | 10276.516 | 3666.342 | 3011.712 | 6803.685 |
| **8** | 1706 | North | 113082.273 | 34739.732 | 22186.164 | 11108.461 | 13179.679 |
| **9** | 1.5\_11 | Africa | 46284.663 | 15830.898 | 10869.338 | 8912.177 | 18020.170 |
| **10** | 1708\_1to10 | North | 28110.263 | 9396.959 | 3809.299 | 257.763 | 1300.336 |
| **11** | 1.6\_37 | Africa | 43116.107 | 17584.536 | 10654.783 | 9562.391 | 19427.253 |
| **12** | 1708\_1to20 | North | 13160.902 | 5361.934 | 2825.748 | 1395.107 | 2172.467 |
| **13** | 2.3\_5 (400) | Africa | 30790.999 | 30622.243 | 15370.053 | 6702.295 | 15023.209 |
| **14** | 1711 | North | 99851.295 | 57454.808 | 25296.511 | 5179.702 | 7384.226 |
| **15** | 2.3\_5 (600) | Africa | 7808.704 | 10276.516 | 3666.342 | 3011.712 | 6803.685 |
| **16** | 1712 | North | 3166.849 | 2580.343 | 1132.392 | 4156.429 | 4213.396 |
| **17** | 2.3\_5 | Africa | 64928.502 | 56813.466 | 32160.834 | 14245.489 | 32194.158 |
| **18** | 1727 | North | 35150.555 | 23946.585 | 9209.576 | 4543.336 | 5343.927 |
| **19** | 2.4\_7 | Africa | 30765.041 | 12190.613 | 7726.828 | 5742.752 | 11548.541 |
| **20** | 1728 | North | 172170.658 | 92432.433 | 43010.285 | 3327.836 | 8005.972 |
| **21** | 3.1\_14 | Africa | 76705.161 | 27022.439 | 19764.457 | 15118.346 | 32090.572 |
| **22** | 1729 | North | 99726.835 | 58533.628 | 23495.943 | 3900.187 | 7019.745 |
| **23** | 3.2\_69 | Africa | 53790.643 | 21756.001 | 14319.347 | 8589.816 | 18565.242 |
| **24** | 1730 | North | 13317.395 | 8646.659 | 3295.380 | 3140.594 | 4555.072 |
| **25** | 3.3\_15 (600) | Africa | 9059.810 | 12920.868 | 6602.240 | 1905.545 | 3655.235 |
| **26** | 1732 | North | 70629.692 | 39587.362 | 16915.751 | 5747.826 | 7208.943 |
| **27** | 3.4\_20(800) | Africa | 6439.392 | 14502.318 | 7533.037 | 1811.117 | 3753.948 |
| **28** | 1733 | North | 44532.668 | 25364.819 | 11509.320 | 5101.810 | 8345.626 |
| **29** | 3.4\_20 | Africa | 24513.053 | 54038.659 | 28561.785 | 5597.266 | 12751.716 |
| **30** | 1734 | North | 98481.255 | 45260.690 | 23024.502 | 11070.989 | 12906.910 |

Таблица 3. Интегралы интенсивности.

Полученная матрица главных компонент (матрица счетов):

|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| --- | --- | --- | --- | --- | --- |
| **1** | -0.59703999 | 1.3772987 | 0.02073497 | 0.1016102120 | -0.1184636385 |
| **2** | 4.64186936 | -1.6561731 | -0.41513814 | 0.1965302124 | 0.0755029989 |
| **3** | -0.19813137 | 1.5298249 | -0.10988824 | 0.1224230731 | -0.0581926550 |
| **4** | 3.50331435 | -2.1641073 | -0.03866739 | 0.1990901959 | 0.0540661630 |
| **5** | 1.37307403 | 2.2412172 | -0.00445231 | 0.3107850644 | 0.0393746072 |
| **6** | -0.38640990 | -1.4469204 | 0.15827815 | -0.0324479287 | -0.2037877787 |
| **7** | -1.83288181 | -0.3417775 | 0.08471708 | 0.1148776255 | -0.0327191241 |
| **8** | 1.06577447 | 0.6470918 | -0.81764493 | -0.3882855388 | 0.1630166965 |
| **9** | -0.35310996 | 1.2177831 | -0.18201522 | 0.1243042203 | 0.0185354639 |
| **10** | -2.02063471 | -1.2830670 | -0.14274340 | 0.3020115844 | 0.0614077285 |
| **11** | -0.26956487 | 1.4290323 | -0.12573306 | 0.0869377849 | -0.0535084022 |
| **12** | -2.16313885 | -0.9245899 | -0.11219942 | 0.1279232730 | 0.0998939056 |
| **13** | -0.26450944 | 0.4792438 | 0.38951037 | -0.0168728840 | -0.0232154134 |
| **14** | 0.92527703 | -0.9888513 | 0.01906274 | -0.1820087423 | -0.0872994606 |
| **15** | -1.83288181 | -0.3417775 | 0.08471708 | 0.1148776255 | -0.0327191241 |
| **16** | -2.12019801 | -0.2321538 | -0.23511338 | -0.1903410956 | 0.0390307847 |
| **17** | 2.25401928 | 2.4673647 | 0.79042912 | -0.1441649714 | -0.0181017334 |
| **18** | -1.04255741 | -0.5016055 | -0.10490429 | -0.2183860673 | -0.0765739287 |
| **19** | -1.09870121 | 0.3256610 | -0.10897084 | 0.1211476713 | 0.0469670072 |
| **20** | 2.83917952 | -2.0532648 | 0.27496157 | 0.1698443406 | -0.0163467488 |
| **21** | 1.33838831 | 3.0008169 | -0.14582650 | 0.1686466231 | -0.0004407408 |
| **22** | 0.77041671 | -1.2165705 | 0.09008003 | 0.0009717715 | -0.1982571781 |
| **23** | -0.03943392 | 1.0818566 | -0.05354057 | 0.1802301347 | 0.0546730041 |
| **24** | -1.88820590 | -0.4935954 | -0.11398611 | -0.0104827488 | 0.0060738366 |
| **25** | -1.82057788 | -0.8306454 | 0.14808895 | 0.0230696658 | 0.1087380340 |
| **26** | 0.02033099 | -0.5308865 | -0.14730405 | -0.2459973492 | -0.0931067472 |
| **27** | -1.77874713 | -0.8556152 | 0.23581516 | 0.0046212882 | 0.1239401796 |
| **28** | -0.71168764 | -0.2515187 | -0.06772634 | -0.0637030679 | -0.0324133552 |
| **29** | 0.52488080 | -0.2401421 | 1.11541845 | -0.3765021136 | 0.1799245350 |
| **30** | 1.16188697 | 0.5560709 | -0.48595949 | -0.6007098590 | -0.0259989160 |

Таблица 4. Матрица главных компонент.

Данная матрица(табл.4) дает проекции исходных данных на подпространство главных компонент. Ее строки – это координаты данных в новой СК, а

столбцы – проекции на новую координатную ось. Графикам, основанным именно на этой матрице, уделяют особое внимание.

Полученная матрица нагрузок:

|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| --- | --- | --- | --- | --- | --- |
| **C** | 0.5275275 | -0.2146383 | -0.7106661 | 0.4090055 | -0.05756213 |
| **A** | 0.5239462 | -0.2786959 | 0.3874170 | -0.2421326 | -0.66263781 |
| **M** | 0.5552962 | -0.1519265 | 0.3248485 | -0.1274301 | 0.73945891 |
| **B** | 0.2749165 | 0.6447891 | -0.3466574 | -0.6226209 | -0.02897929 |
| **T** | 0.2515198 | 0.6613806 | 0.3452012 | 0.6084334 | -0.09979207 |

Таблица 5. Матрица нагрузок.

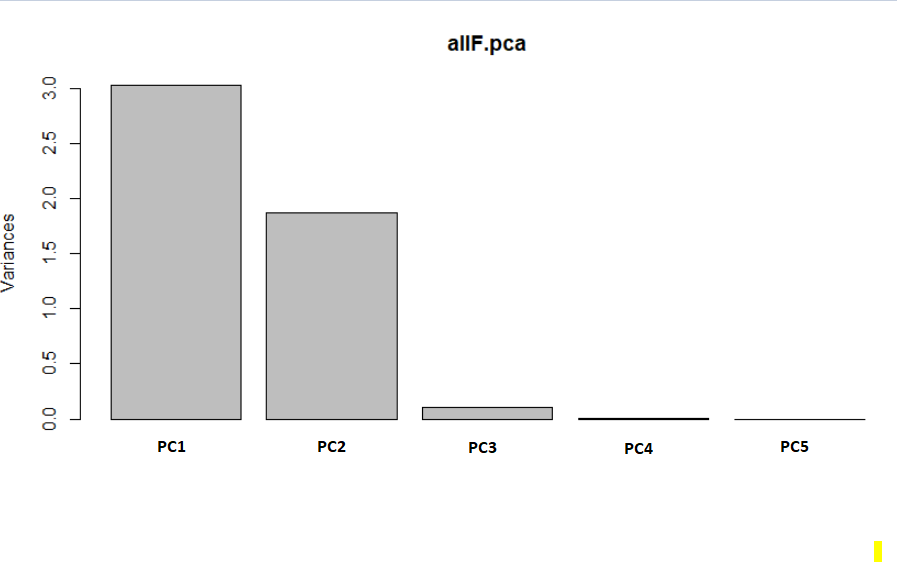
Это матрица (табл.5) перехода из исходного пространства в пространство главных компонент. Каждый столбец матрицы – это проекция соответствующей начальной переменной на новую СК, а каждая строка – это коэффициенты, связывающие исходные и конечные переменные. 

Рисунок 7. Метод главных компонент. Компоненты

Как видно из графика (рисунок 5) наибольший вклад вносят компоненты 1 и 2.

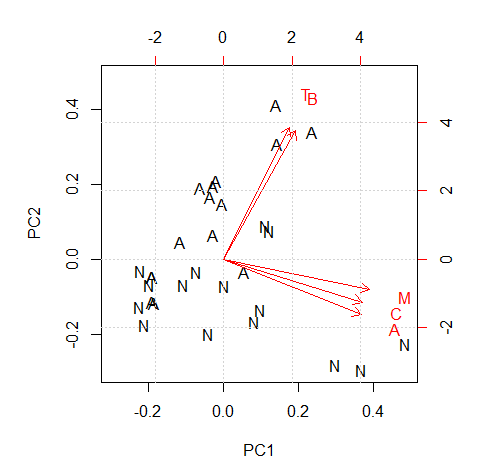


Рисунок 8. Метод главных компонент. Визуализация.

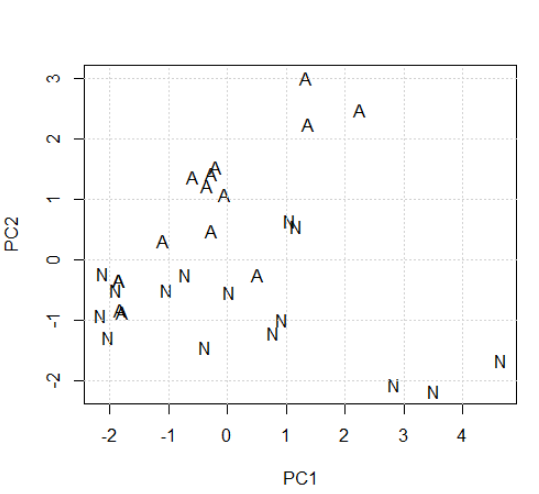


Рисунок 9. МГК. РС1 - РС2

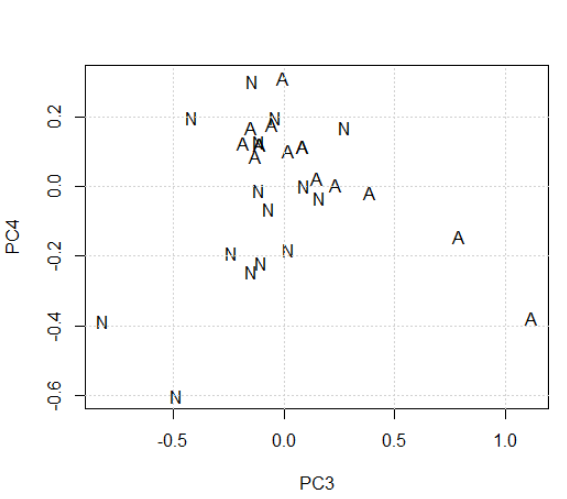


Рисунок 10. МГК. РС3 - РС4

На графике РС1-РС2 (рис. 7) отчетливо видно разделение на две группы – Север и Африку, чего не скажешь о графике РС3-РС4 (рис. 8).

Первая компонента разделяет данные по С, М и А, а вторая компонента по В и Т.

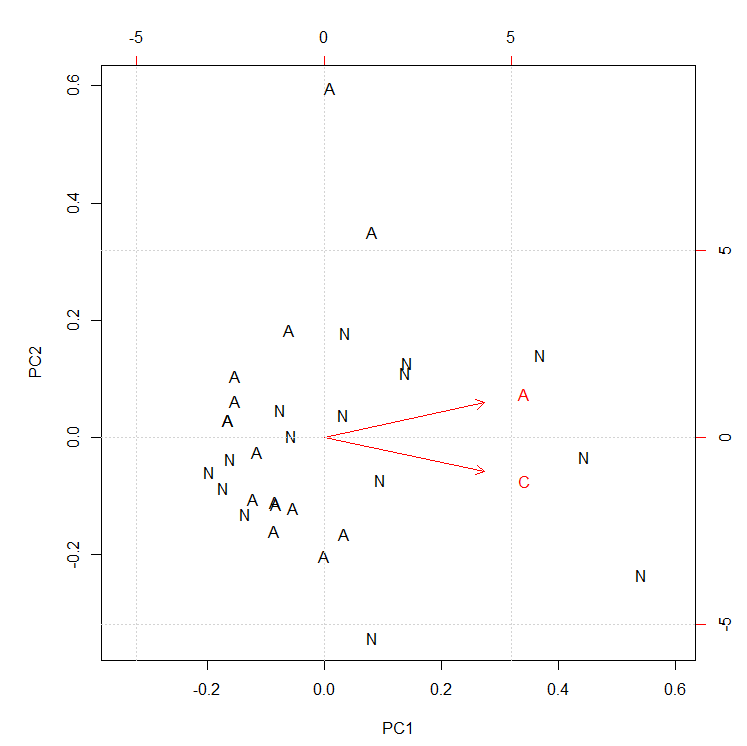


Рисунок 11. Разделение по А и С.

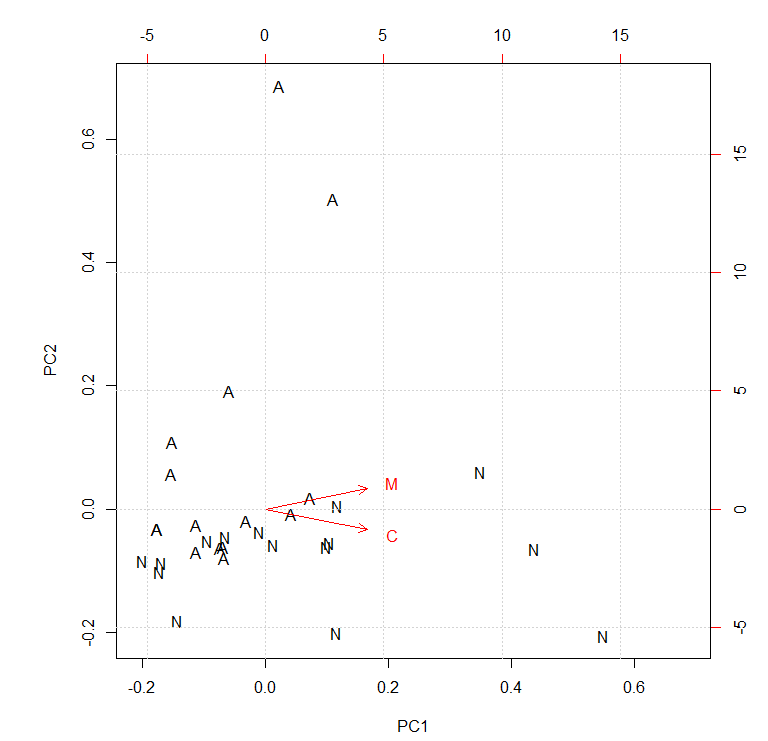


Рисунок 12. Разделение по М и С.

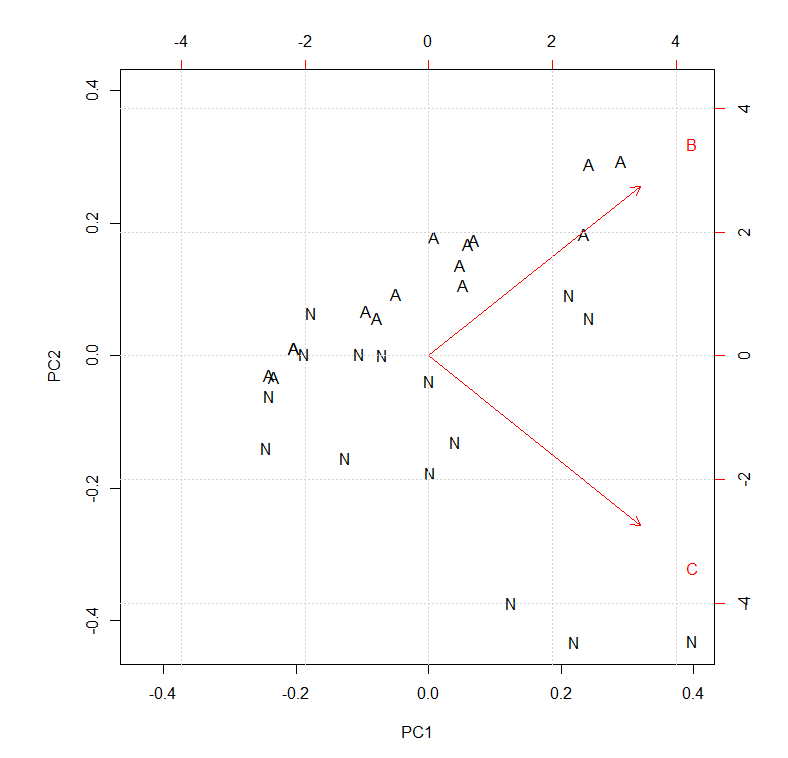


Рисунок 13. Разделение по В и С.

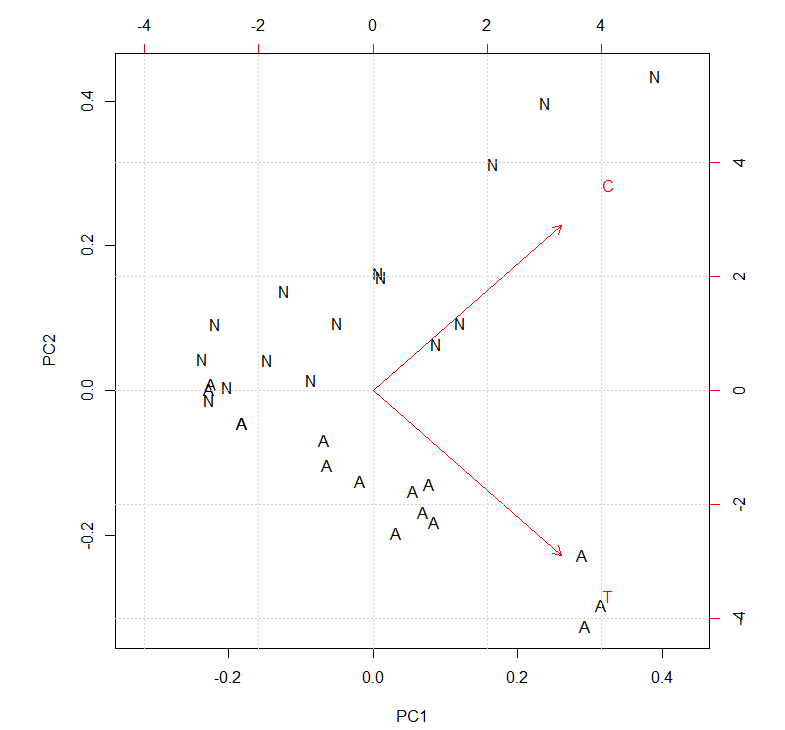


Рисунок 14. Разделение по С и Т.

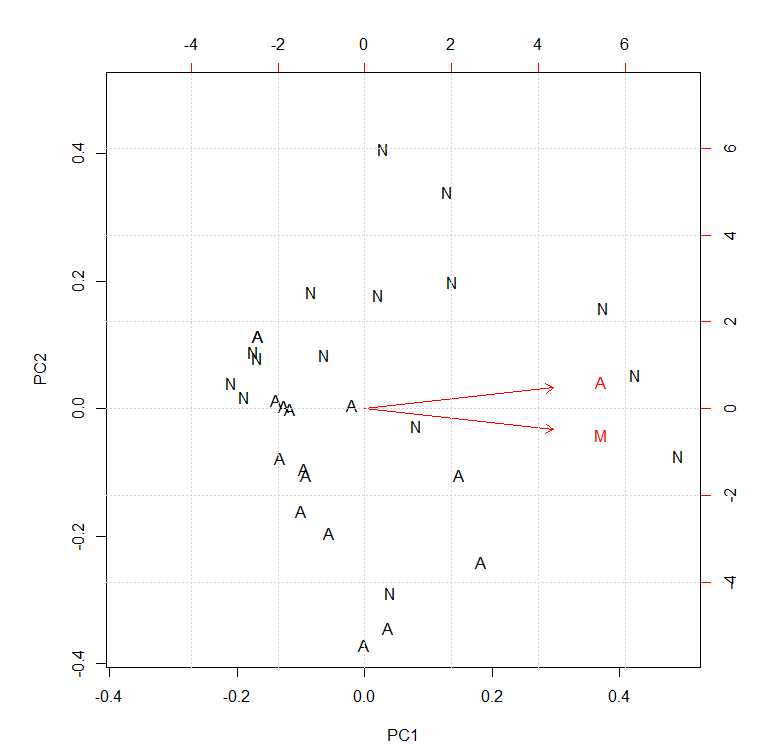


Рисунок 15. Разделение по А и М.

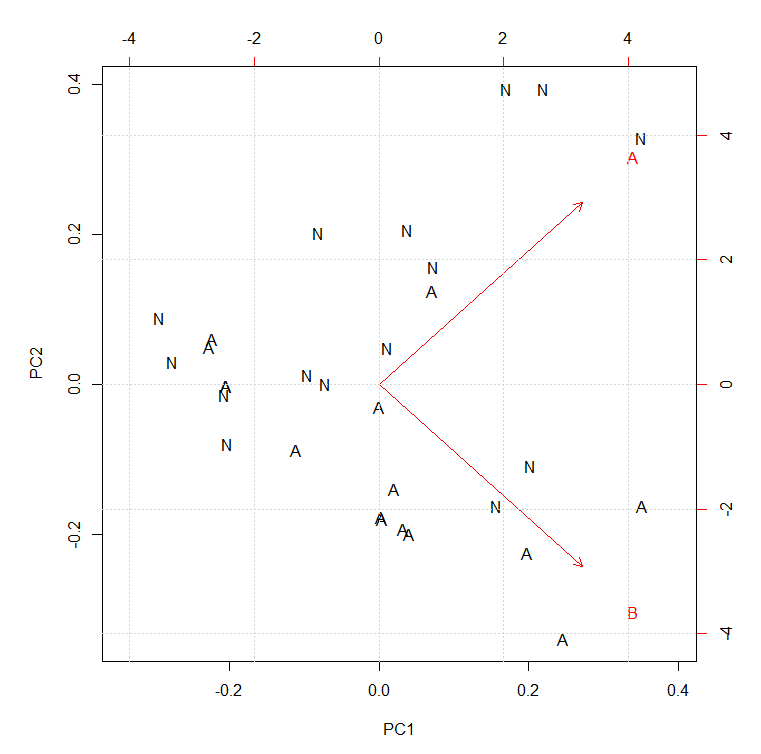


Рисунок 16. Разделение по А и В.

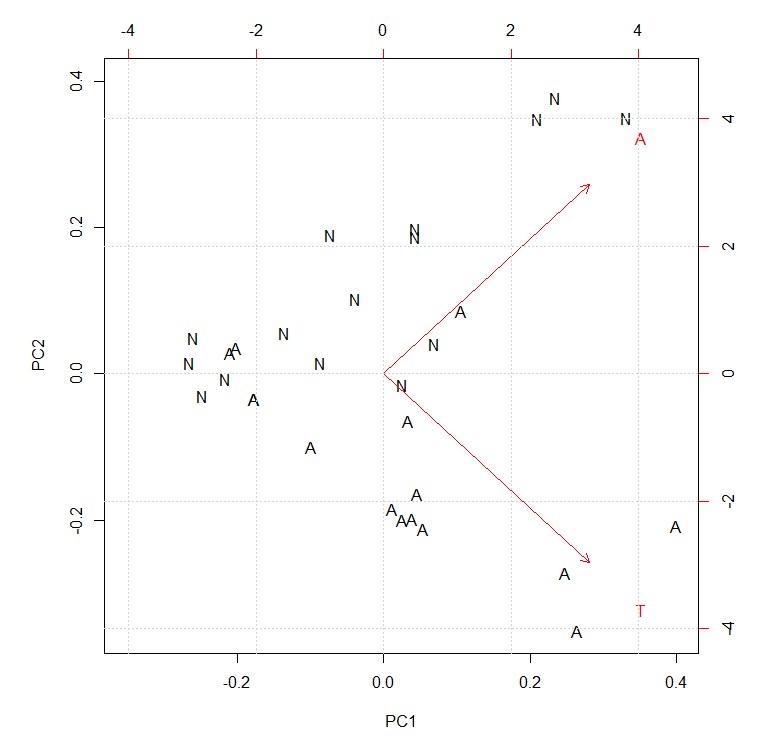


Рисунок 17. Разделение по А и Т.

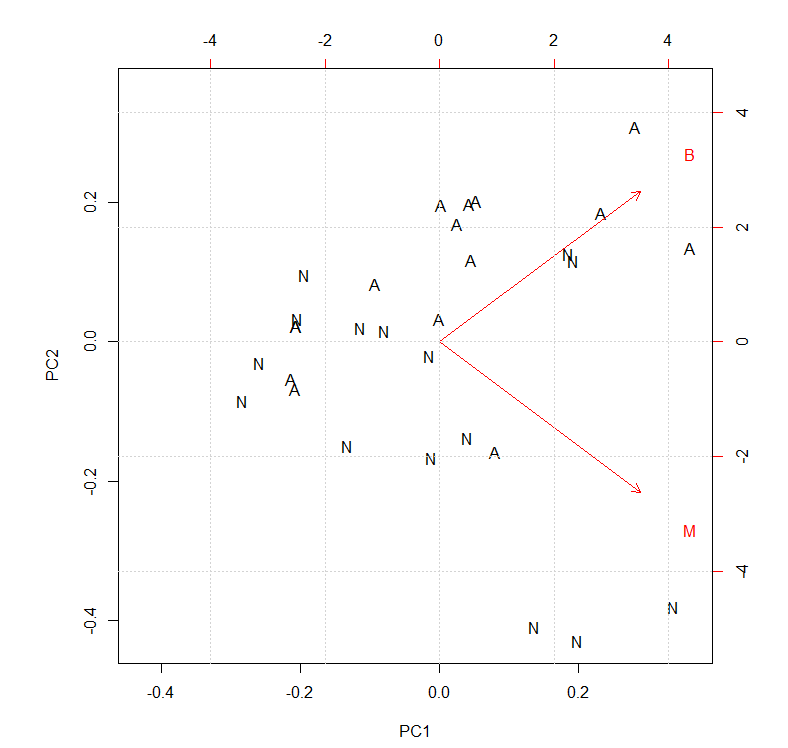


Рисунок 18.Разделеление по В и М.

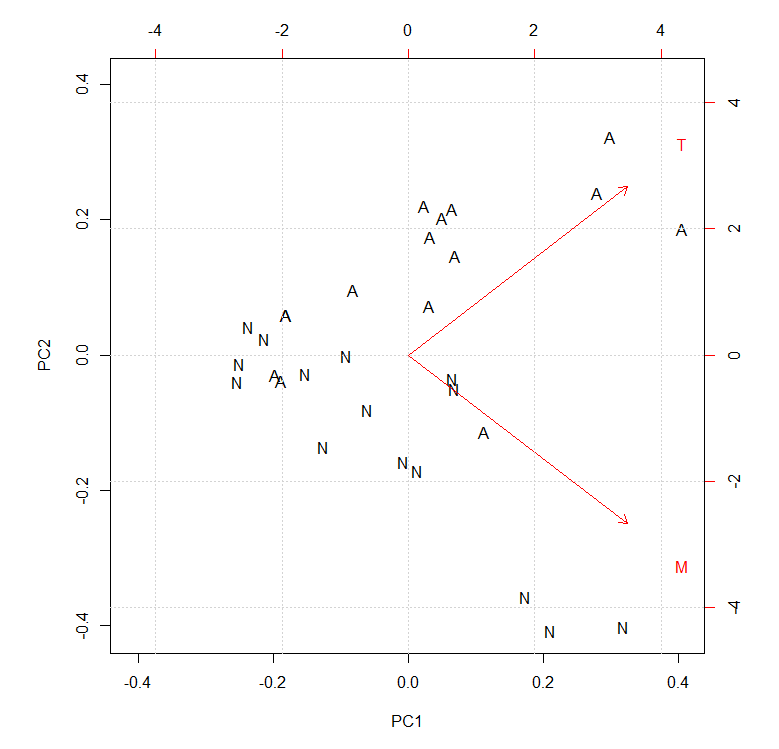


Рисунок 19. Разделение по Т и М.

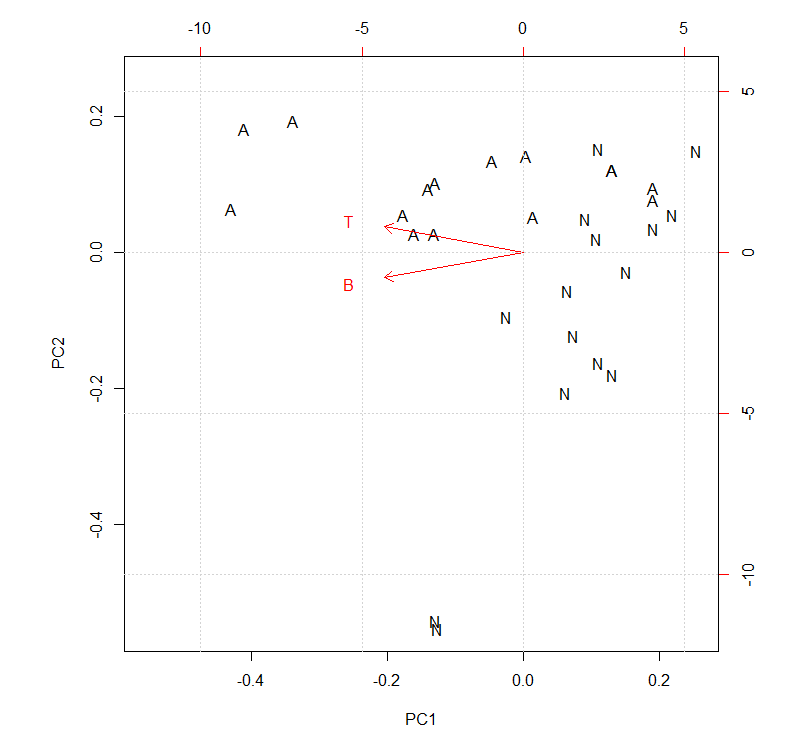


Рисунок 20. Разделение по Т и В.

Из графиков ([рис. 9 – 18](#рис11)) также видно, что пробы могут быть разделены по всем компонентам, взятым по раздельности, за исключением компоненты М.

Выводы

Полученные результаты подтвердили то, что пробы Африки и Севера могут быть разделены, но также подтвердилось и то, что есть исключения: например, при разделении по компоненте С выбиваются файлы 3.3\_15(600) и 3.4\_20(800) Африки, по компоненте А 1712 и 1730 Севера и 3.4\_20 Африки и т.д. При «суммарном» же разделении от остальных отличаются файлы Севера: 1706 и 1734, а Африки: 1.4\_114, 2.3\_5(600) и те же 3.3\_15(600), 3.4\_20(800).

При помощи МГК нам удалось уменьшить размерность до двумерной.

Из рассмотрения Таблицы 4 видно, что основной вклад в данные вносят главные компоненты PC1, PC2. Например, рассмотрим 2 первых строчки данных в таблице 4:

|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| --- | --- | --- | --- | --- | --- |
| **1** | -0.59703999 | 1.3772987 | 0.02073497 | 0.1016102120 | -0.1184636385 |
| **2** | 4.64186936 | -1.6561731 | -0.41513814 | 0.1965302124 | 0.0755029989 |

Таким образом установлено, что пробы, полученные с разных регионов, могут быть разделены с помощью метода главных компонент.

Резюме

В ходе практической работы был изучен метод главных компонент. С помощью данного метода были проанализированы данные, полученные с русского Севера и Африки. Было получено, что пробы из разных регионов могут быть разделены с помощью МГК. Также были освоены пакеты EEM и stats языка программирования R в среде разработки RStudio.

Список литературы

1. <https://cran.r-project.org/web/packages/eemR/vignettes/introduction.html>
2. <https://www.chemometrics.ru/ru/books/metod-glavnykh-komponent/lyudi-i-strany/>
3. <https://github.com/korolevskaya-kd/MathStatistics/tree/master/Practic>