МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский ядерный университет «МИФИ» (НИЯУ МИФИ)

Институт Финансовых Технологий и Экономической Безопасности Кафедра Финансового мониторинга

**Лабораторная работа №4 по курсу**

**«Макростатический анализ и прогнозирование»**

|  |  |
| --- | --- |
| **Выполнил студент группы С21-702:** | Нашивочников А.В. |
| **Проверил:** | Домашова Д.В. |

**Оглавление**

[Задание к лабораторной работе 3](#_Toc87857400)

[Порядок выполнения лабораторной работы 3](#_Toc87857401)

[Матрица нагрузок 9](#_Toc87857402)

[Вывод 25](#_Toc87857403)

## **Задание к лабораторной работе**

Субъекты РФ характеризуются социально-экономическими показателями, обозначение и наименование которых приведены в таблице А.1. Значения показателей для 84 субъектов приведены в таблице А.2. Ставится задача на основании статистических данных по показателям, соответствующим нужному варианту, снизить размерность признакового пространства методом главных компонент, обеспечив уровень информативности новой системы признаков не ниже 70%.

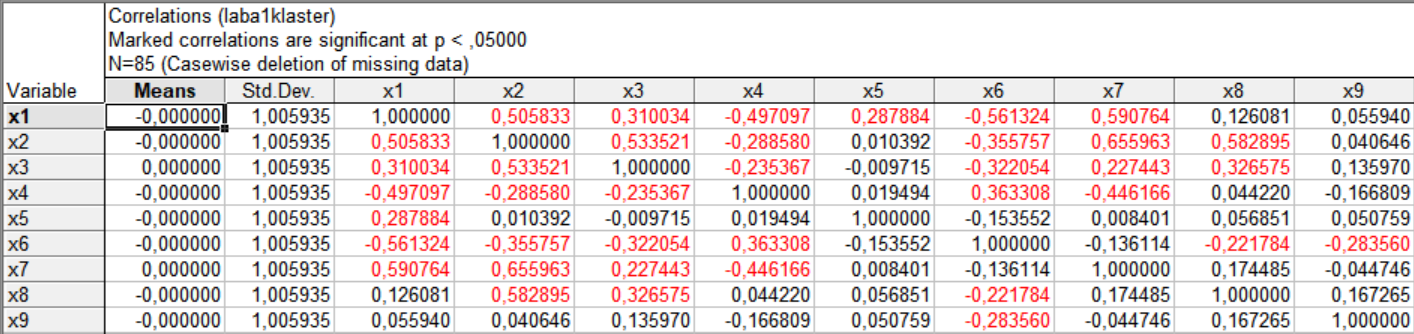
## **Порядок выполнения лабораторной работы**

Порядок выполнения лабораторной работы включает следующие показатели для анализа:

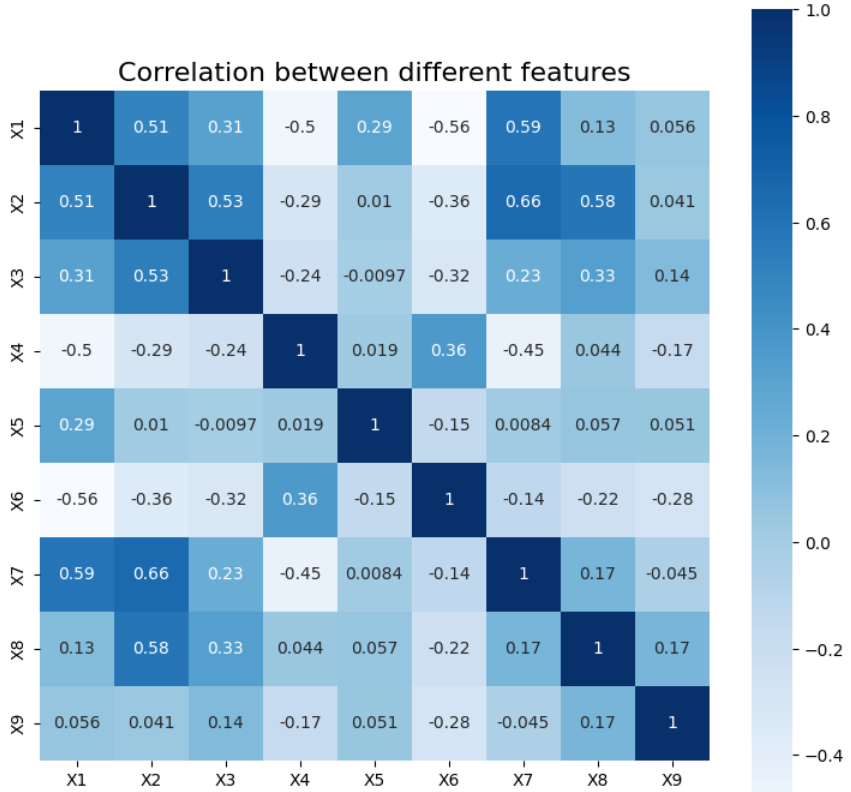
1. X1 – брачность (число браков на 1000 человек)
2. X2 - средний размер назначенных пенсий
3. X3 – заболеваемость на 1000 человек населения
4. X4 - процент безработных мужчин
5. X5 - число посещений музея на 1000 человек населения
6. X6 - число населения на одну больничную койку
7. X7 – потребительские расходы в среднем на душу населения тыс. рублей
8. X8 – количество врачей на 1000 человек населения
9. X9 – число спортивных сооружений на 1000 человек

Поскольку исходные признаки отличаются масштабом измерения, то будем рассматривать вектор центрировано-нормированных признаков x\* = (x1\*, x2\*,…, xk\*)T  и на основе исходной матрицы данных X рассчитаем оценку корреляционной матрицы.

Таблица 1 – Результаты расчета корреляционной матрицы





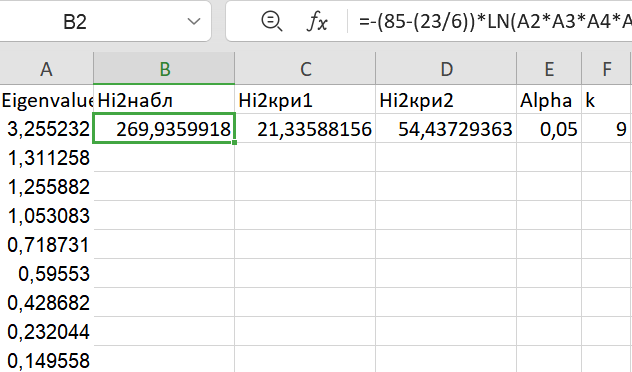


Далее согласно алгоритму, предполагая, что выборка извлечена из нормально распределенной генеральной совокупности, на уровне значимости α = 0,05 проверим гипотезу о незначимости корреляционной матрицы. **H0: Rx = Е;**

**H1: Rx ≠ E.**

Для проверки гипотезы потребуются оценки собственных чисел корреляционной матрицы.

Расчет наблюдаемого значения проводится в Excel на основании полученных оценок собственных значений, которые приводятся на рисунке 1.



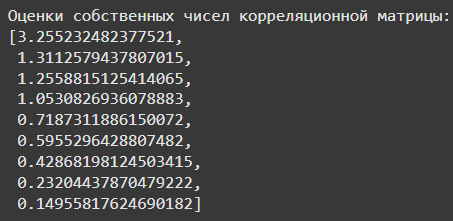


Рисунок 1 – Проверка гипотезы о незначимости корреляционной матрицы

Наблюдаемое значение рассчитывается по формуле:

χ2 = -(n - (2k+5))ln||,

где || - определитель матрицы , равный произведению оценок собственных чисел матрицы;

k – число факторов;

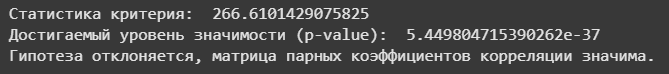
n – объем выборки.

Наблюдаемое значение составило χ 2набл =269,94. Критические значения χкр1 и 2 χкр2 определяются из уравнений:



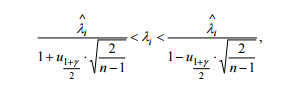
Для решения этих уравнений необходимо воспользоваться функцией ХИ2ОБР(вероятность, ) пакета Excel. Вероятность рассчитывается как 100\*(1-)% для χкр1 и 100\*()% для χкр2 . Число степеней свободы ν = . Критические точки принимают следующие значения:

Так как , то гипотеза  отвергается, матрица парных коэффициентов корреляции значима.



С вероятностью γ = 0,95 построим доверительные интервалы для собственных чисел.

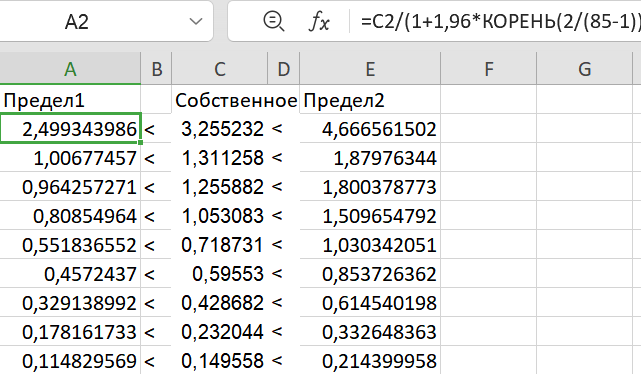
Доверительный интервал для i-ого собственного числа λi при большом объеме выборки имеет вид:

,

где – квантиль уровня стандартного нормального распределения;

n – объем выборки.

Квантиль уровня q можно найти с помощью функции НОРМСТОБР(q) пакета Excel. Получаем: u0,975 = 1,96. Доверительные интервалы для собственных чисел имеют вид, представленный на рисунке 2.



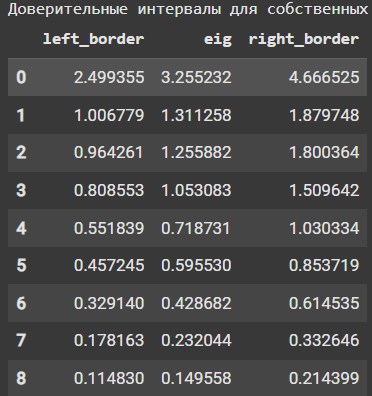
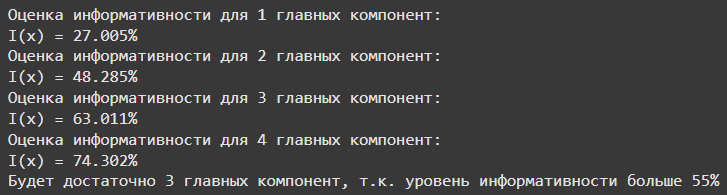


Рисунок 2 – Доверительные интервалы собственных чисел

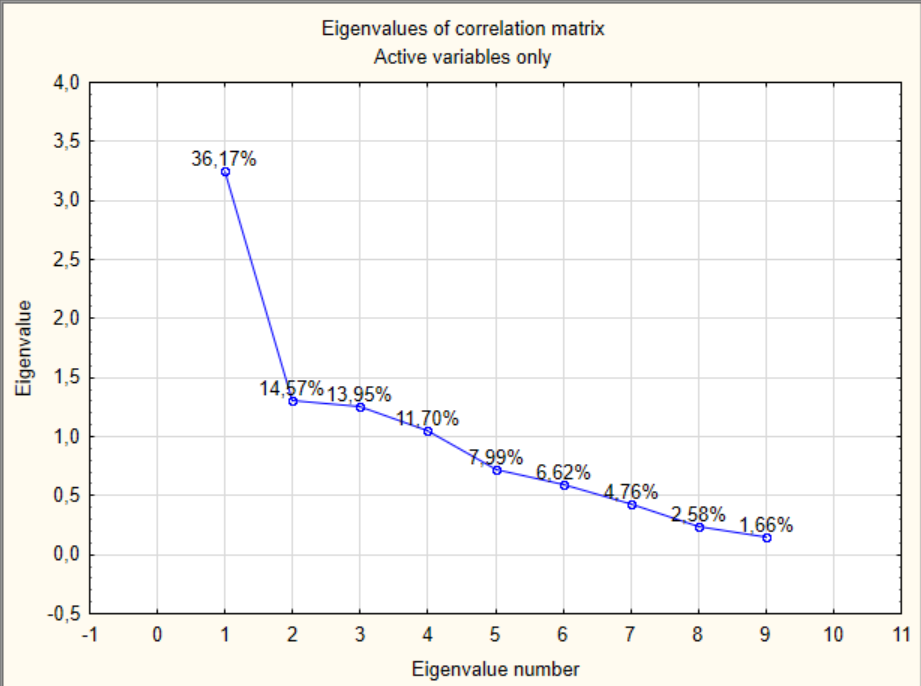
Третье собственное число попадает в доверительный интервал второго, значит можно заподозрить кратность главных компонент.

Количество главных компонент можно посчитать с помощью Критерия Кайзера (отбираются факторы с собственными числами > 1, в данном примере их получается 3).

Так как собственные числа корреляционной матрицы являются дисперсиями главных компонент, то оценка уровня информативности первых трёх главных компонент составляет I3(z(x)) = , что превышает требуемый уровень 55%. I2(z(x)) =



Так же можно использовать Критерий каменистой осыпи, для этого в Statistica строится графическое представление собственных чисел («График осыпи» в русской Statistica).



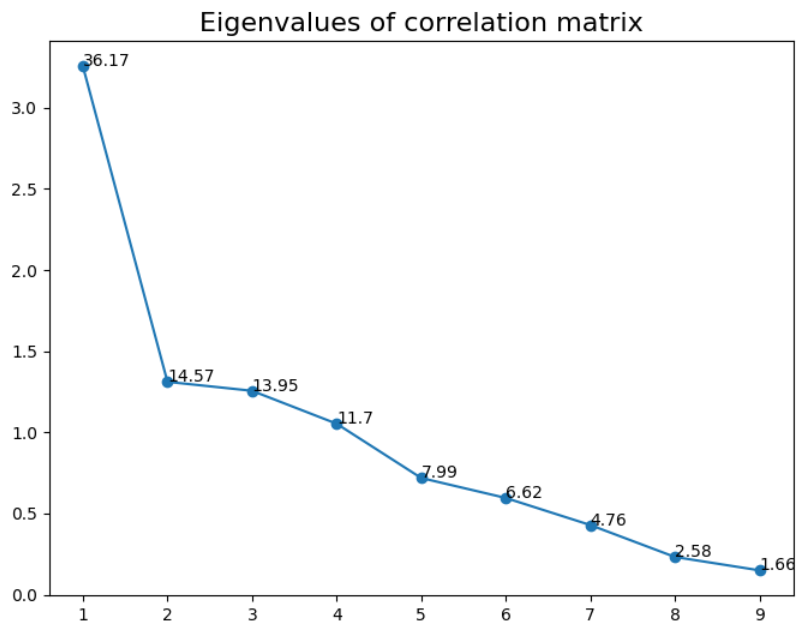
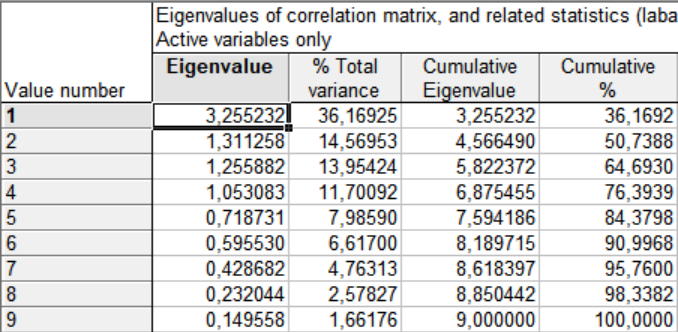
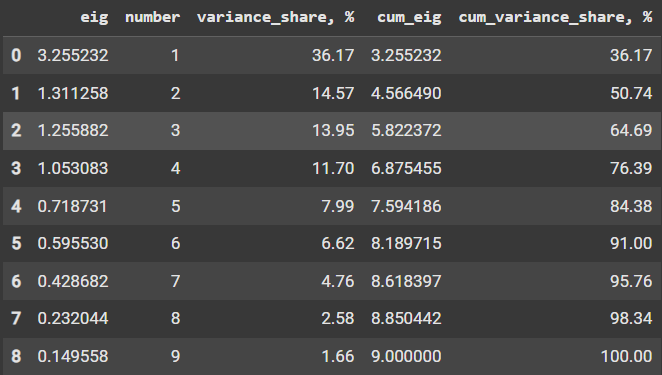


Рисунок 3 – График собственных чисел

Следует найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь". В соответствии с этим критерием можно оставить в этом примере 2 главные компоненты.

Таблица 2 – Вклады главных компонент в суммарную дисперсию исходных признаков, рассчитанные в пакете Statistica





В первом столбце таблицы приведены оценки собственных чисел, в третьем столбце – накопленные значения собственных чисел, во втором и в четвертом столбцах – относительный вклад каждой главной компоненты в суммарную дисперсию и накопленный относительный вклад соответственно.

Таблица 3 – Результаты расчета собственных векторов корреляционной матрицы

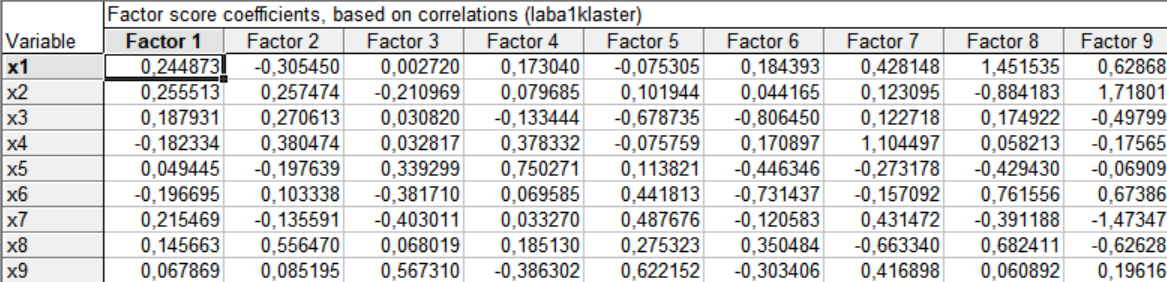
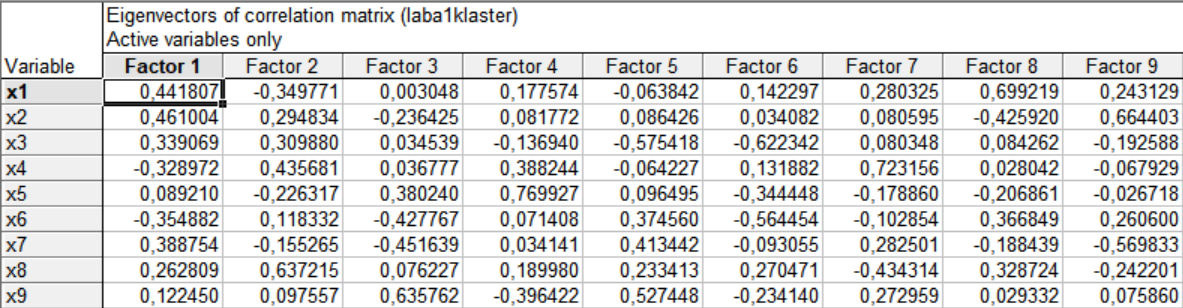
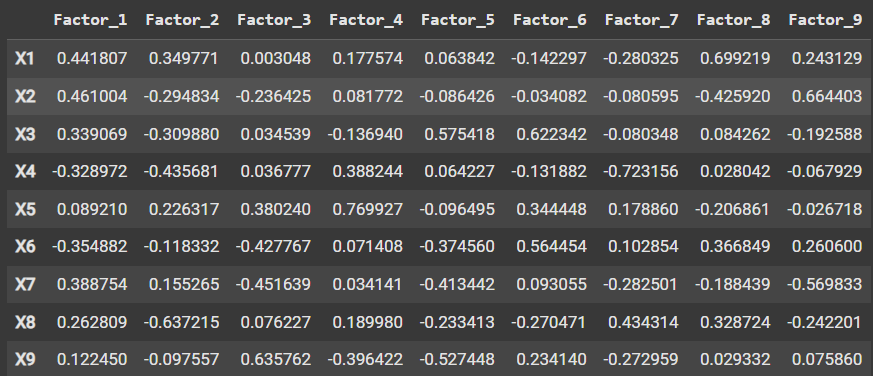


Таблица 4 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков





При снижении размерности признакового пространства до трех главных компонент следует рассматривать только три первых столбца матрицы U.

Главные компоненты связаны с центрировано-нормированными исходными признаками следующими линейными комбинациями:

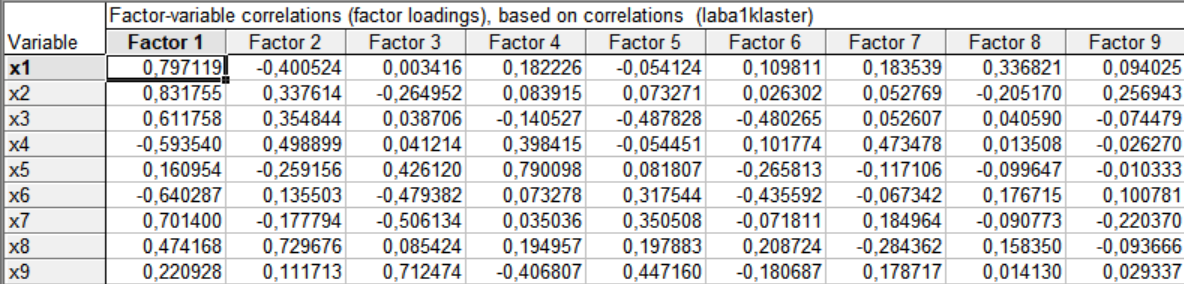
*z1* = 0,44x1\*+0,46x2\*+0,33x3\*-0,32x4\*+0,09x5\*-0,35х6\*+0,39x7\*+0,26x8\*+0,12х9\*

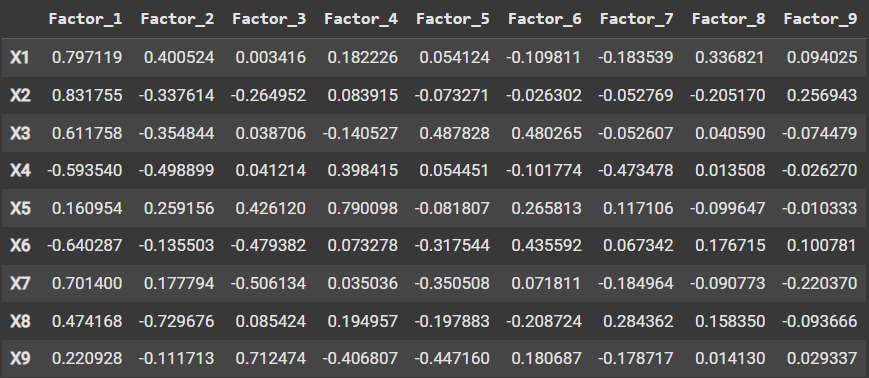
*z2* = -0,35x1\*+0,29x2\*+0,31x3\*+0,44x4\*-0,23x5\*+0,11х6\*-0,15x7\*-0,64x8\*-0,09х9\*

*z3* = 0,003x1\*-0,24x2\*+0,03x3\*+0,04x4\*+0,38x5\*-0,43х6\*-0,45x7\*+0,08x8\*+0,64х9\*

# **Матрица нагрузок**

Таблица 5 – Результаты расчета элементов матрицы нагрузок





Так как расчеты проводятся на основании корреляционной матрицы, то элементы матрицы нагрузок являются коэффициентами корреляции исходных признаков и главных компонент. Как видно из таблицы, между исходными признаками и последними пятью главными компонентами не наблюдается тесной связи (не имеется значений >0,55). Это подтверждает правильность выделения только трех первых главных компонент. Матрица нагрузок имеет вид:

Первая главная компонента тесно связана (коэффициент корреляции больше 0,55) с 4 исходными признаками *x1* – брачность, *х2 -* средний размер назначенных пенсий, *x3 -* заболеваемость на 1000 человек населения, *x7 -* отребительские расходы в среднем на душу населения тыс. Рублей. И тесно отрицательно с *х6 -* число населения на одну больничную койку. Поэтому первую главную компоненту можно интерпретировать как «уровень здравоохранения и брачности».

Вторая главная компонента тесно положительно связана с признаками *x8* – количество врачей на 1000 человек населения. Поэтому вторую главную компоненту можно интерпретировать как «обеспеченность врачами».

К третьей главной компоненте можно отнести признаки *х9* – число спортивных сооружений на 1000 человек и охарактеризовать как «обеспеченность спортивными сооружениями». *х5* – число посещений музея на 1000 человек населения относим к четвертой компоненте, охарактеризуем как «посещаемость музея». Центрировано-нормированные исходные признаки связаны с центрировано- нормированными главными компонентами f1, f2, f3 следующими выражениями:

*x1\* =* 0,79*f1  f2+*0,003*f3;*

*x2\* =*0,83 *f1 +* 34*f2*26*f3;*

*x3\* =* 0,61*f1 f2+04f3;*

*x4\* = f1 + f2+f3;*

*x5\* =* 0,16*f1  f2+*0,43*f3;*

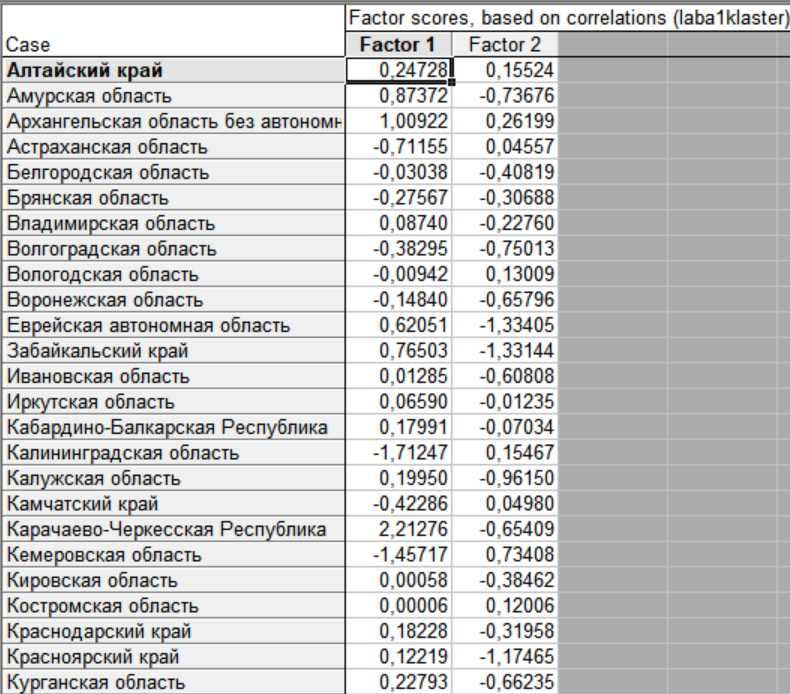
*x6\* = f1 + f2-*0,48*f3;*

*x7\* =* 0,70*f1 f2*0,51*f3;*

*x8\* =* 0,47*f1* 73*f2+*0,08*f3;*

*x9\* =* 0,22*f1* 11*f2+*0,71*f3;*

Рисунок 4 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент





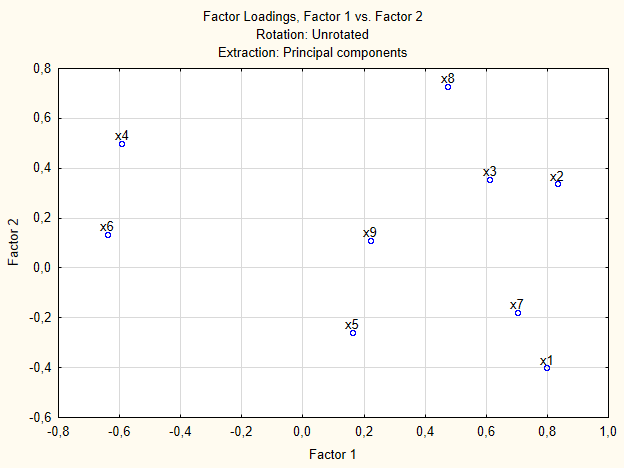


Рисунок 7 – Распределение признаков в пространстве первых двух главных компонент

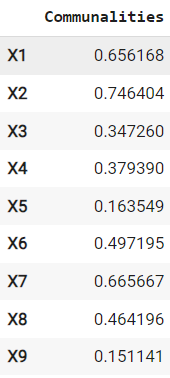
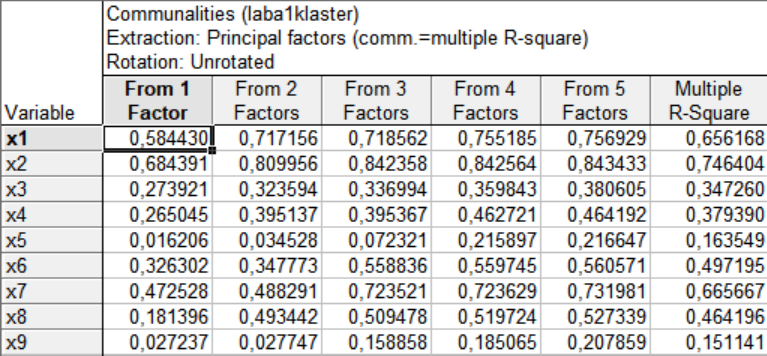
# **Метод главных факторов**

Оценками общностей в данном алгоритме будут служить квадраты оценок множественных коэффициентов корреляции , т.е.:

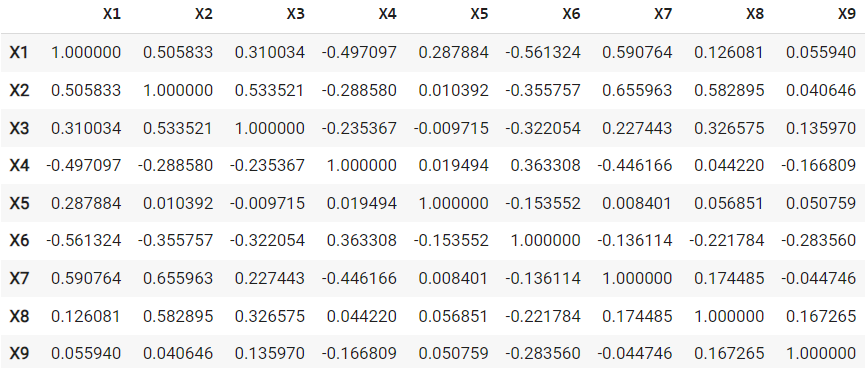
, .

Оценки общности представлены ниже:

Таблица 6 – Результаты расчета общностей для двух главных факторов



В первом, втором столбцах таблицы содержатся вклады одного и двух факторов в дисперсию признаков. Оценки общностей приведены в третьем столбце таблицы. На основе матрицы парных коэффициентов корреляции и оценок общностей можно составить оценки редуцированной матрицы



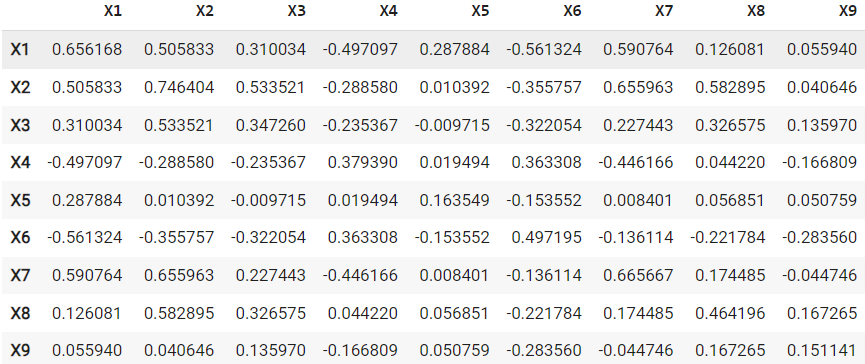
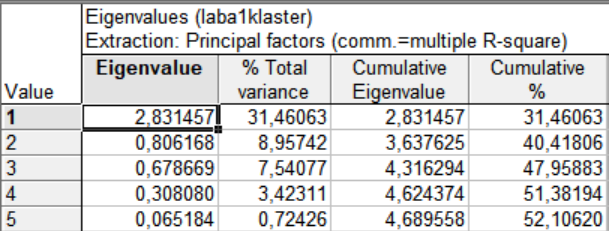
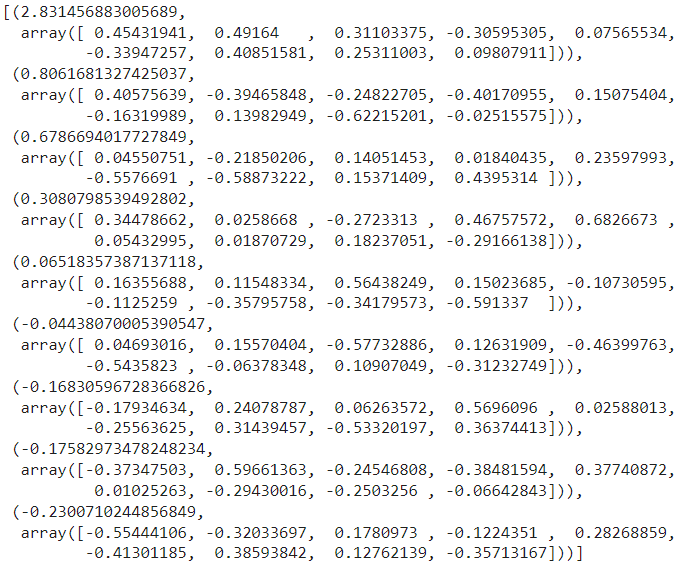
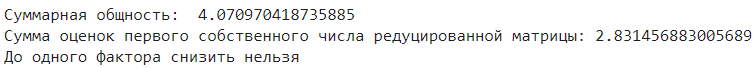


Таблица 7 – Вклад двух главных факторов в суммарную дисперсию исходных признаков







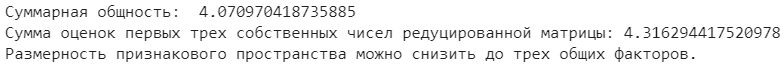


Таблица 8 – Результаты расчета общностей для трех главных факторов

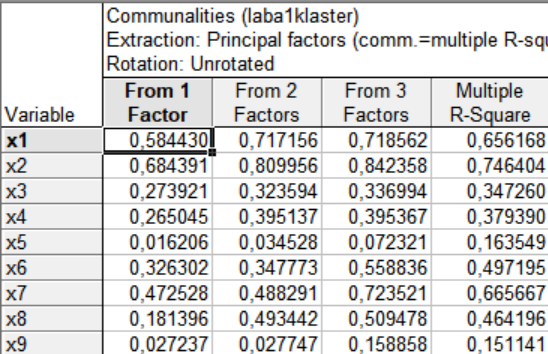
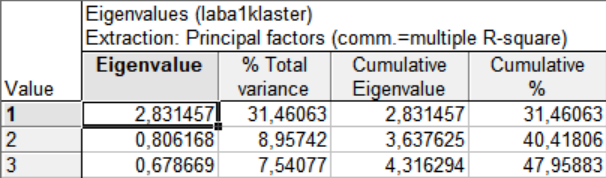


Таблица 9 – Вклад трех главных факторов в суммарную дисперсию исходных признаков



Сумма оценок трех собственных чисел редуцированной матрицы будет в этом случае больше суммарной общности (

Вклад трех главных факторов в суммарную дисперсию исходных признаков (в дисперсию процесса) составляет 47,96%.

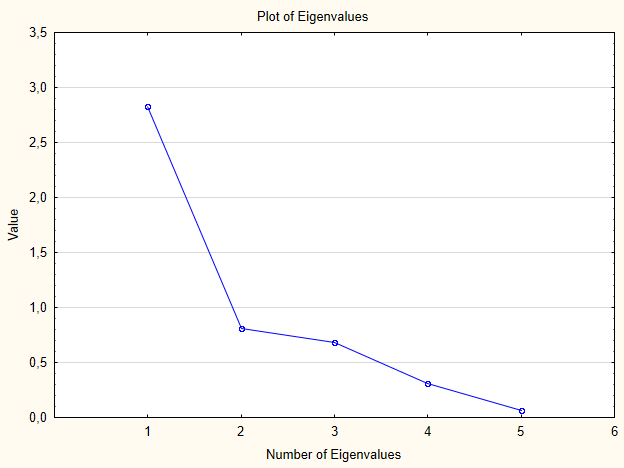
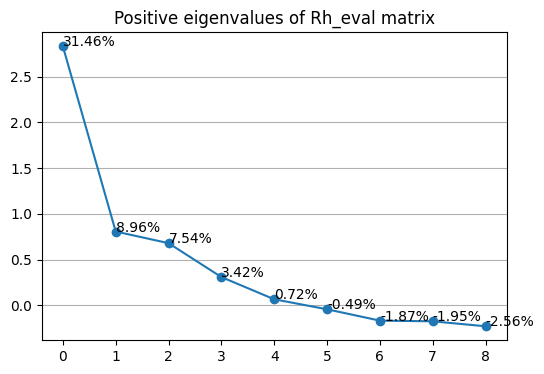


Рисунок 8 – График собственных значений

Таблица 10 – Весовые коэффициенты при общих факторах

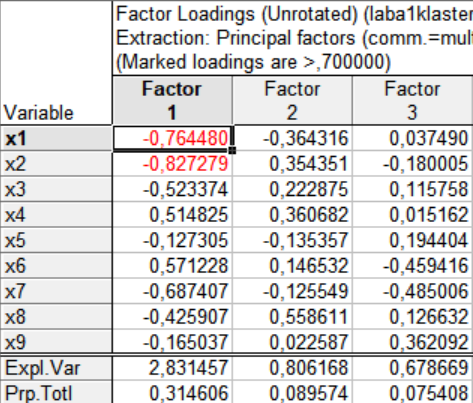
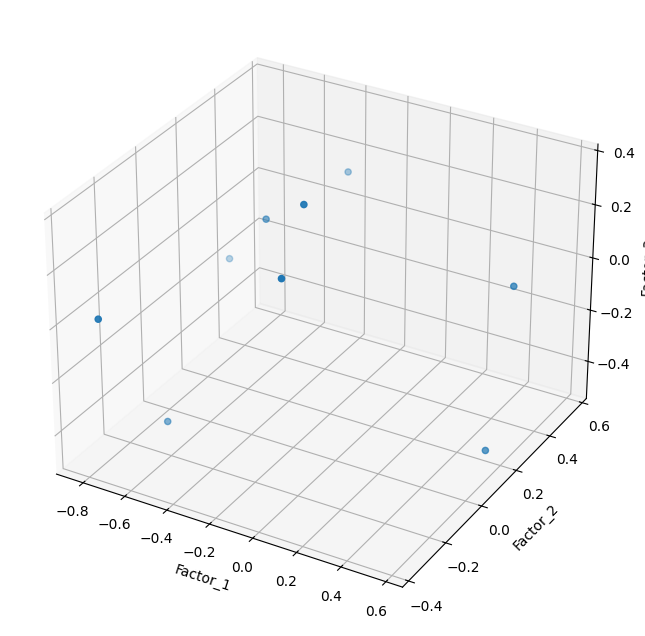
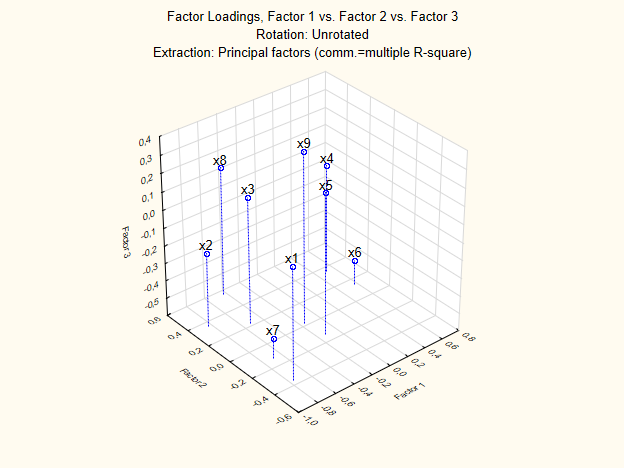


Рисунок 9 – Расположение исходных признаков на плоскости, образованной главными факторами



Попробуем упростить структуру главных факторов с помощью вращения. Для оценки структуры обобщённых факторов выберем критерий «Квартимакс», т.к. он показал наилучший результат среди других критериев.

Таблица 11 – Весовые коэффициенты факторов после вращения

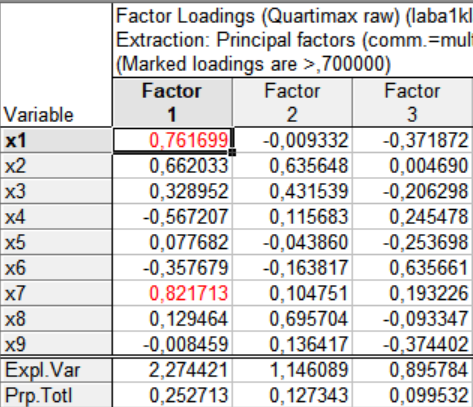
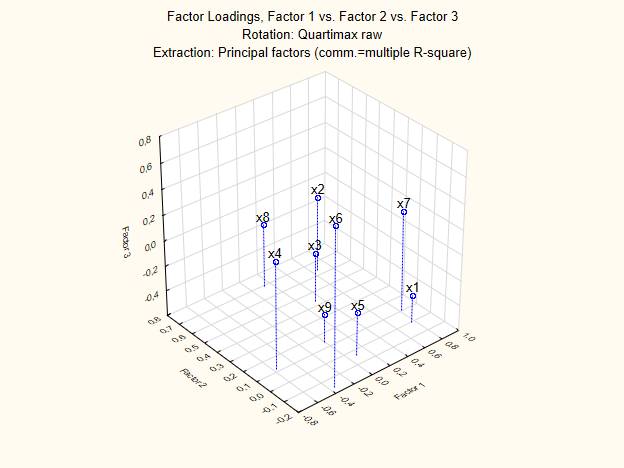
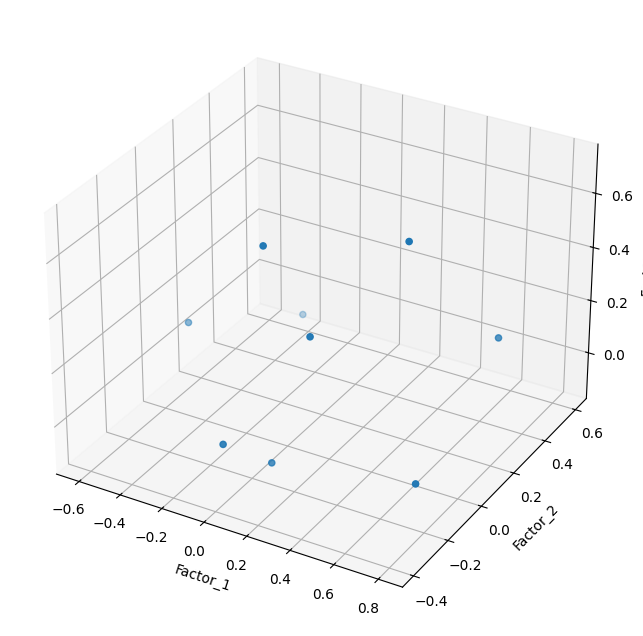


Рисунок 10 – Расположение исходных признаков на плоскости, образованной главными факторами после вращения



Матрица факторных нагрузок после вращения имеет вид:

|  |  |  |
| --- | --- | --- |
| 0,761699 | -0,009332 | -0,371872 |
| 0,662033 | 0,635648 | 0,004690 |
| 0,328952 | 0,431539 | -0,206298 |
| -0,567207 | 0,115683 | 0,245478 |
| 0,077682 | -0,043860 | -0,253698 |
| -0,357679 | -0,163817 | 0,635661 |
| 0,821713 | 0,104751 | 0,193226 |
| 0,129464 | 0,695704 | -0,093347 |
| -0,008459 | 0,136417 | -0,374402 |

|  |  |  |
| --- | --- | --- |
| 0,584430 | 0,717156 | 0,718562 |
| 0,684391 | 0,809956 | 0,842358 |
| 0,273921 | 0,323594 | 0,336994 |
| 0,265045 | 0,395137 | 0,395367 |
| 0,016206 | 0,034528 | 0,072321 |
| 0,326302 | 0,347773 | 0,558836 |
| 0,472528 | 0,488291 | 0,723521 |
| 0,181396 | 0,493442 | 0,509478 |
| 0,027237 | 0,027747 | 0,158858 |

X1 – брачность (число браков на 1000 человек)

X2 - средний размер назначенных пенсий

X3 – заболеваемость на 1000 человек населения

X4 - процент безработных мужчин

X5 - число посещений музея на 1000 человек населения

X6 - число населения на одну больничную койку

X7 – потребительские расходы в среднем на душу населения тыс. рублей

X8 – количество врачей на 1000 человек населения

X9 – число спортивных сооружений на 1000 человек

Вывод:

Сравнивая матрицы А и В можно сделать вывод, что структура обобщённых факторов после вращения значительно улучшилась. Первый фактор тесно связан (коэффициент корреляции > 0,7) с несколькими исходным признаком: X1 брачностью и X7 потребительскими расходами. Следовательно, первый главных фактор можно характеризировать как «причина брачности и расходов».

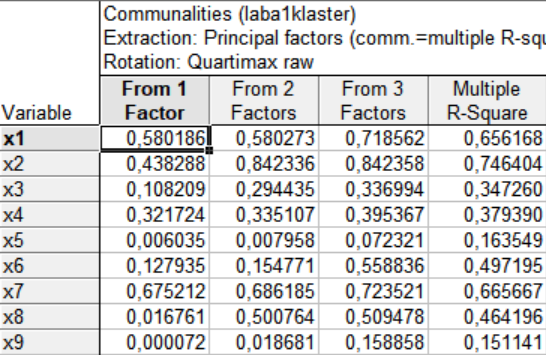
Второй фактор связан с исходными признаками: X8 количество врачей на 1000 человек населения. Следовательно, второй главных фактор можно характеризировать как «Обеспеченность врачами».

Третий фактор также связан с признаками: X6 число населения на одну больничную койку.

Далее представлена таблица вкладов факторов в дисперсию исходных признаков. Элементы первого столбца таблицы равны квадратам соответствующих элементов первого столбца матрицы В, элементы второго столбца – сумме квадратов соответствующих элементов первого и второго столбцов матрицы В, элементы третьего столбца – сумме квадратов элементов первого, второго и третьего столбцов матрицы В. Таким образом, в третьем столбце представлены оценки общностей, рассчитанные по матрице В, а в четвертом столбце – оценки общностей, рассчитанные по формуле:

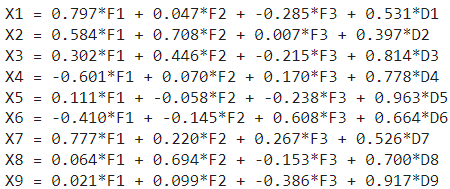
, 

Таблица 12 – Вклады трех факторов в дисперсию признаков



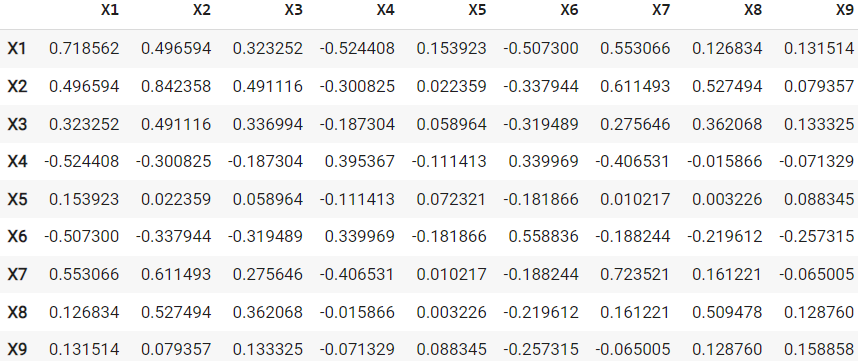
Т.к. исходные признаки процентированы и пронормированы, а главные факторы не коррелированы между собой, то оценки характерностей можно рассчитать следующим образом:

Центрировано-нормированные исходные признаки связаны с главными и характерными факторами следующими выражениями: ():



Оценка редуцированной матрицы парных коэффициентов корреляции, рассчитанная по матрице нагрузок В, и оценка остаточной матрицы парных коэффициентов корреляции представлены ниже:

Таблица 13 – Оценка редуцированной матрицы



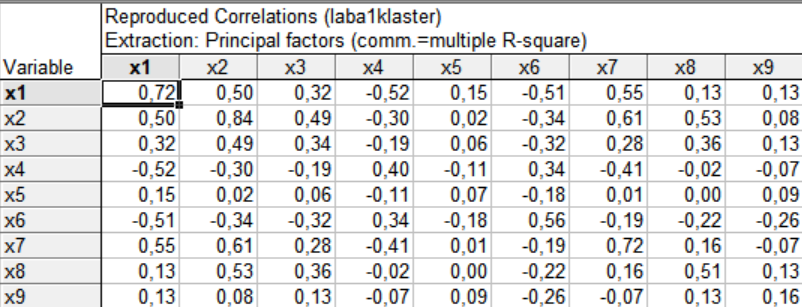
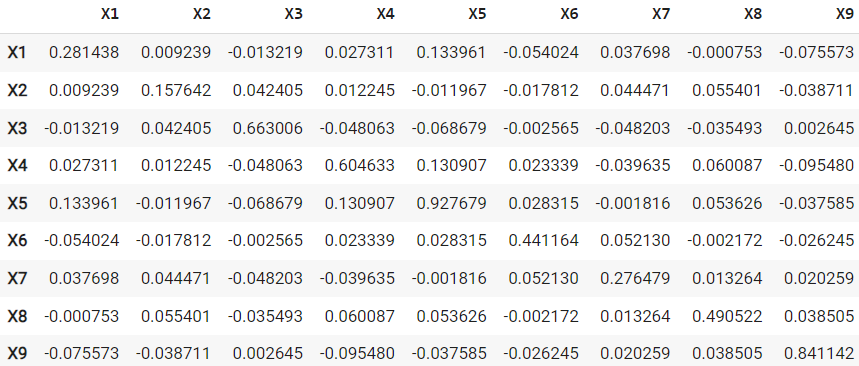
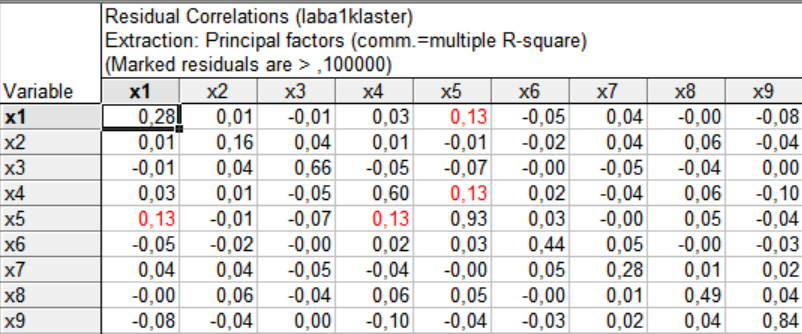


Таблица 14 – Оценка остаточной матрицы парных коэффициентов корреляции





На главной диагонали матрицы, представленной в табл.12, расположены оценки характерностей .

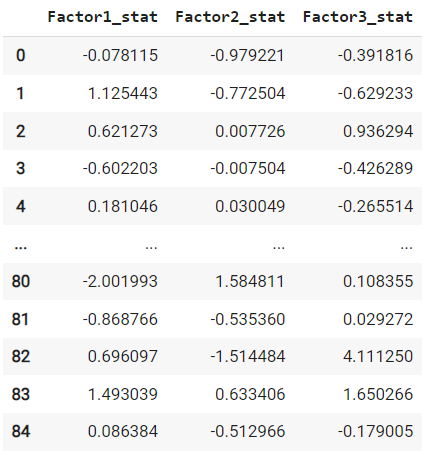
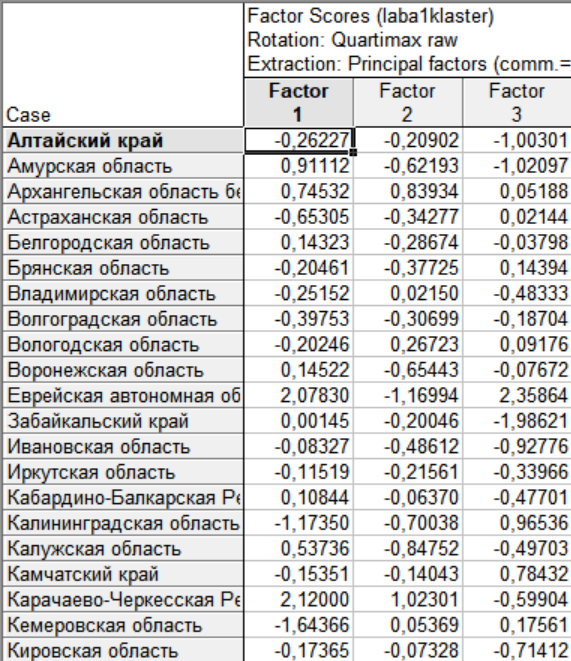




Рисунок 11 – Индивидуальные значения общих факторов

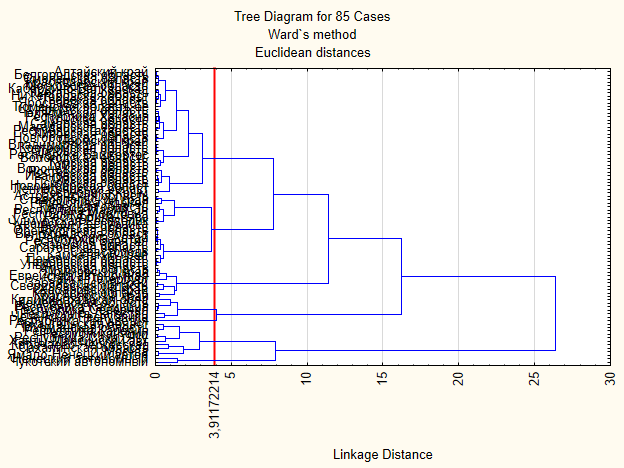


Рисунок 5 – Дендрограмма по Уорду по оставшимся двум главным компонентам

Таблица 6 – Результаты классификации субъектов РФ методом Уорда

|  |  |  |
| --- | --- | --- |
| **Номер**  **кластера** | **Кол-во**  **объектов** | **Состав класса** |
| кластер 1  {S11} | 1 | Республика Ингушетия |
| кластер 2  {S12} | 9 | Амурская область Еврейская автономная область Забайкальский край Калужская область Красноярский край Приморский край Санкт-Петербург Свердловская область Хабаровский край |
| кластер 3  {S13} | 6 | Калининградская область Кемеровская область Республика Дагестан Республика Калмыкия Республика Северная Осетия – Алания Чеченская Республика |
| кластер 4 {S14} | 2 | Ненецкий автономный округ Чукотский автономный округ |
| кластер 5 {S15} | 10 | Архангельская область без автономного округа Карачаево-Черкесская Республика Москва Мурманская область Республика Карелия Республика Коми Республика Саха (Якутия) Сахалинская область Ханты-Мансийский автономный округ – Югра Ямало-Ненецкий автономный округ |
| кластер 6 {S16} | 36 | Алтайский край Белгородская область Владимирская область Вологодская область Воронежская область Ивановская область Иркутская область Кабардино-Балкарская Республика Кировская область Костромская область Краснодарский край Курганская область Курская область Ленинградская область Магаданская область Московская область Нижегородская область Новгородская область Новосибирская область Омская область Орловская область Пермский край Псковская область Республика Башкортостан Республика Крым Республика Татарстан Республика Хакасия Ростовская область Самарская область Смоленская область Тверская область Томская область Тульская область Тюменская область без автономных округов Челябинская область Ярославская область |
| кластер 7 {S17} | 21 | Астраханская область Брянская область Волгоградская область Камчатский край Липецкая область Оренбургская область Пензенская область Республика Адыгея Республика Алтай Республика Бурятия Республика Марий Эл Республика Мордовия Республика Тыва Рязанская область Саратовская область Севастополь Ставропольский край Тамбовская область Удмуртская Республика Ульяновская область Чувашская Республика |

Диаграмма 1 – График средних значений кластеров по методу Уорда

Рисунок 6 - График рассеяния объектов по методу Уорда для 2 ГК(New)

Рисунок 6 - График рассеяния объектов по методу Уорда для 2 ГК(Old)

Таблица 7 – Результаты классификации субъектов РФ методом k-средних (2)

|  |  |  |
| --- | --- | --- |
| **Номер**  **кластера** | **Кол-во**  **объектов** | **Состав класса** |
| кластер 1  {S11} | 10 | Астраханская область Камчатский край Липецкая область Республика Адыгея Республика Марий Эл Республика Мордовия Республика Тыва Ставропольский край Удмуртская Республика Чувашская Республика |
| кластер 2  {S12} | 20 | Алтайский край Белгородская область Воронежская область Ивановская область Кабардино-Балкарская Республика Кировская область Краснодарский край Курганская область Московская область Нижегородская область Новгородская область Новосибирская область Омская область Псковская область Ростовская область Свердловская область Смоленская область Тверская область Тюменская область без автономных округов Ярославская область |
| кластер 3  {S13} | 28 | Брянская область Владимирская область Волгоградская область Вологодская область Иркутская область Костромская область Курская область Ленинградская область Магаданская область Оренбургская область Орловская область Пензенская область Пермский край Республика Алтай Республика Башкортостан Республика Бурятия Республика Крым Республика Татарстан Республика Хакасия Рязанская область Самарская область Саратовская область Севастополь Тамбовская область Томская область Тульская область Ульяновская область Челябинская область |
| кластер 4 {S14} | 6 | Калининградская область Кемеровская область Республика Дагестан Республика Калмыкия Республика Северная Осетия – Алания Чеченская Республика |
| кластер 5 {S15} | 9 | Амурская область Еврейская автономная область Забайкальский край Калужская область Красноярский край Приморский край Санкт-Петербург Сахалинская область Хабаровский край |
| кластер 6 {S16} | 11 | Архангельская область без автономного округа Карачаево-Черкесская Республика Москва Мурманская область Ненецкий автономный округ Республика Карелия Республика Коми Республика Саха (Якутия) Ханты-Мансийский автономный округ – Югра Чукотский автономный округ Ямало-Ненецкий автономный округ |
| кластер 7 {S17} | 1 | Республика Ингушетия |



Диаграмма 2 - средние значения метода к-средних для 2 ГК

Рисунок 7 - График рассеяния объектов по методу К-средних для 2 ГК(New)