|  |
| --- |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ |
| Федеральное государственное автономное образовательное  учреждение высшего образования |
| **НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»** |



***ИНСТИТУТ ФИНАНСОВЫХ ТЕХНОЛОГИЙ И ЭКОНОМИЧЕСКОЙ БЕЗОПАСНОСТИ***

КАФЕДРА «ФИНАНСОВЫЙ МОНИТОРИНГ»

ОТЧЕТ

по лабораторной работе №3

по курсу «Макростатистический анализ и прогнозирование»

Выполнил студент группы С18-712:

Степаненкова Маргарита

Преподаватель: Домашова Д. В.

Москва, 2021

# Постановка задачи

Исходные данные:

* X1 - валовой региональный продукт (млн руб на 1 000 человек населения)
* X2 - инвестиции в основной капитал (в фактически действовавших ценах; млн руб на 1 000 человек населения)
* X3 - число предприятий и организаций (на конец года на 1 000 человек населения)
* X4 - уровень занятости для людей в возрасте 15-72 лет (процентов)
* X5 - коэффициенты миграционного прироста на 10 000 человек населения
* X6 - зарегистрировано преступлений особой тяжести (на 1 000 человек населения)
* X7 - зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1 000 человек населения)
* X8 - общие коэффициенты смертности (число умерших на 1000 человек населения)
* X9 - мощность амбулаторно-поликлинических организаций всего, тыс. посещений в смену (на конец года, на 1 000 человек населения)
* X10 - численность врачей всех специальностей, всего, тыс. человек (на конец года, на 1 000 человек населения)
* X11 - численность пенсионеров, тыс. человек (на 1 000 человек населения)

Ставится задача на основании статистических данных по показателям снизить размерность признакового пространства методом главных компонент, обеспечив уровень информативности новой системы признаков не ниже 55%.

# Выполнение работы

Исходные данные для анализа представлены в виде матрицы. Фрагмент таблицы с исходными данными в пакете Statistica представлен на Рисунке 1.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1 – Исходные данные для анализа

Для приведения исходных переменных к стандартизованному виду можно воспользоваться операцией центрирования и нормирования данных. Результаты преобразования данных представлены на Рисунке 2.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 2 – Центрировано-нормированные значения признаков

Рассчитаем выборочные значения коэффициентов корреляции для каждой пары признаков (Рисунок 3.1 и Рисунок 3.2).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 3.1 - Результаты расчета корреляционной матрицы (Statistica)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 3.2 - Результаты расчета корреляционной матрицы (Python)

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 (Рисунок 4).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 4 - Результаты расчета оценок собственных чисел корреляционной матрицы в пакете Statistica

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 4.1 - Результаты расчета оценок собственных чисел корреляционной матрицы в Python

Расчет наблюдаемого значения проводится в Excel на основании полученных оценок собственных значений, которые приводятся на Рисунке 5.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 5 - Проверка гипотезы о незначимости корреляционной матрицы в Excel

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 5.1 - Проверка гипотезы о незначимости корреляционной матрицы в Excel

Наблюдаемое значение рассчитывается по формуле:

χ2 = ,

где |R ̂x| - определитель матрицы R ̂\_x, равный произведению оценок собственных чисел матрицы;

k – число факторов;

n – объем выборки.

Наблюдаемое значение составило χ2 набл = 575,20. Критические значения и определяются из уравнений:

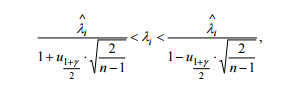


Для решения этих уравнений необходимо воспользоваться функцией ХИ2ОБР (вероятность, v) пакета Excel. Вероятность рассчитывается как 100\*()% для χкр1 и 100\*( )% для χкр2. Число степеней свободы . Критические точки принимают следующие значения:

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 Рисунках 6.1 и 6.2.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 6.1 - Доверительные интервалы собственных чисел, посчитанные в Excel

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 6.2 - Доверительные интервалы собственных чисел, посчитанные в Python

Второе собственное значение попадает в доверительный интервал первого. Поэтому проверим гипотезу о кратности собственных чисел.

Гипотеза проверяется с помощью статистики:

Наблюдаемое значение составило 304,03. Число степеней свободы

Критические точки принимают следующие значения:

Так как , то гипотеза о кратности собственных чисел отвергается.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 7.1 – Проверка гипотезы о кратности собственных чисел (Excel)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 7.2 – Проверка гипотезы о кратности собственных чисел (Python)

Оценим уровень информативности главных компонент.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 8 – Результаты оценки уровня информативности (Excel)

Изображение выглядит как текст, внутренний, снимок экрана

Автоматически созданное описание

Рисунок 8 – Результаты оценки уровня информативности (Python)

Рассчитаем критерий Кайзера.



Рисунок 9 – Значение критерия Кайзера (Python)



Рисунок 9.1 – Значение критерия Кайзера (Excel)

Рассчитаем Критерий Бартлетта.

Изображение выглядит как текст, оранжевый, снимок экрана, закрыть

Автоматически созданное описание

Рисунок 9.2 – Значение критерия Бартлетта

Так же можно использовать Критерий каменистой осыпи, для этого в Statistica и с помощью Python строится графическое представление собственных чисел (Рисунок 10.1 и Рисунок 10.2).

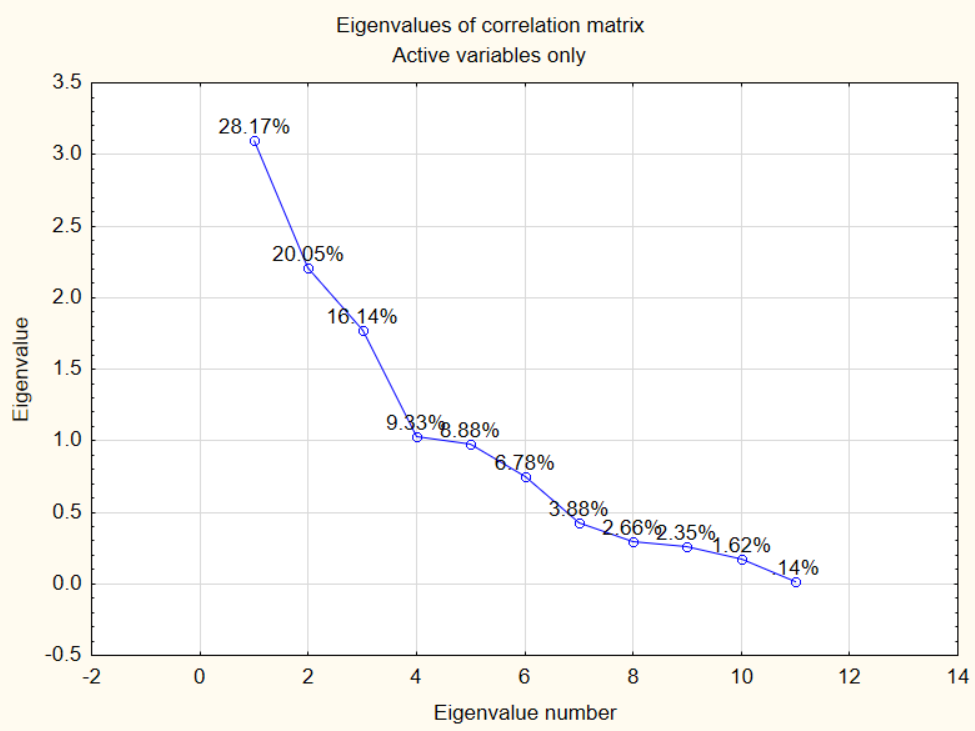


Рисунок 10.1 - График собственных чисел (Statistica)

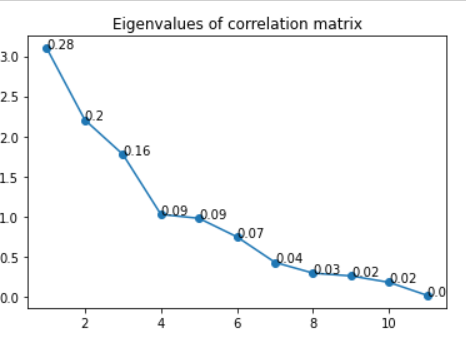


Рисунок 10.2 - График собственных чисел (Python)

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 4 главные компоненты.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 11.1 - 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.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 11.2 - Вклады главных компонент в суммарную дисперсию исходных признаков, рассчитанные в пакете Statistica

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 12 - Результаты расчета собственных векторов корреляционной матрицы в пакете Statistica

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 13.1 - Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Statistica)

Изображение выглядит как текст, окно

Автоматически созданное описание

Рисунок 13.2 - Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Python)

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

Statistica:

z1 = -0,509x1\*+0.493x2\*+0.347x3\*+0.440x4\*+0.048x5\*+0.10x6\*-0.05x7\*-0.23 x8\*+ 0.06 x9\*+0.33 x10\*-0.002 x11\*,

z2 = 0.039x1\*+0.072x2\*-0.018x3\*-0.097x4\*+0.248x5\*-0.456x6\*-0.329x7\*-0.418x8\*-0.445 x9\*-0.177 x10\*-0.449 x11\*,

z3 = 0.075x1\*+0.090x2\*-0.362x3\*-0.162x4\*-0.315x5\*+0.324x6\*+0.540x7\*-0.367x8\*-0.136 x9\*+0.131 x10\*-0.403 x11\*,

Python:

Изображение выглядит как текст

Автоматически созданное описание

Для интерпретации новых признаков необходимо провести анализ матрицы нагрузок.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 14.2 - Результаты расчета элементов матрицы нагрузок в Python

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 14.1 - Результаты расчета элементов матрицы нагрузок в пакете Statistica

Так как расчеты проводятся на основании корреляционной матрицы, то элементы матрицы нагрузок являются коэффициентами корреляции исходных признаков и главных компонент. Матрица нагрузок имеет вид:

|  |  |  |  |
| --- | --- | --- | --- |
|  | F1 | F2 | F3 |
| X1: ВРП (млн руб на 1 000 человек населения) | 0.897439 | -0.057401 | -0.100554 |
| X2: инвестиции в основной капитал (в фактически действовавших ценах; млн руб на 1 000 человек населения) | 0.868408 | -0.106477 | -0.120513 |
| X3: число предприятий и организаций (на конец года на 1 000 человек населения) | 0.611102 | 0.027074 | 0.482674 |
| X4: уровень занятости для людей в возрасте 15-72 лет (процентов) | 0.774899 | 0.144469 | 0.215321 |
| X5: коэффициенты миграционного прироста на 10 000 человек населения | 0.083940 | -0.368767 | 0.420091 |
| X6: зарегистрировано преступлений особой тяжести (на 1 000 человек населения) | 0.175227 | 0.616497 | -0.431364 |
| X7: зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1 000 человек населения) | -0.081084 | 0.488559 | -0.720164 |
| X8: общие коэффициенты смертности (число умерших на 1000 человек населения( | -0.410148 | 0.620903 | 0.488684 |
| X9: мощность амбулаторно-поликлинических организаций всего, тыс. посещений в смену (на конец года, на 1 000 человек населения) | 0.107832 | 0.661182 | 0.181818 |
| X10: численность врачей всех специальностей, всего, тыс. человек (на конец года, на 1 000 человек населения) | 0.584273 | 0.263171 | -0.174305 |
| X11: численность пенсионеров, тыс. человек (на 1 000 человек населения) | -0.004324 | 0.667177 | 0.536632 |

Первая главная компонента тесно положительно связана (коэффициент корреляции больше 0,7) с тремя исходными признаками: Х1 - валовой региональный продукт (млн руб на 1 000 человек населения), Х2 – инвестиции в основной капитал (в фактически действовавших ценах; млн руб на 1 000 человек населения), Х4 - уровень занятости для людей в возрасте 15-72 лет (процентов). Поэтому первую главную компоненту можно интерпретировать как «Показатель экономической активности». Вторая главная компонента тесно положительно связана с Х9 - мощность амбулаторно-поликлинических организаций всего, тыс. посещений в смену (на конец года, на 1 000 человек населения) и Х11 - численность пенсионеров, тыс. человек (на 1 000 человек населения). Ее можно интерпретировать как «Влияние уровня медицины на возрастной состав населения». Третья главная компонента тесно отрицательно связана с X7 - зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1 000 человек населения). Ее можно интерпретировать как «Уровень преступности».

Центрировано-нормированные исходные признаки связаны с центрировано-нормированными главными компонентами f1, f2, f3 следующими выражениями:

Изображение выглядит как текст, газета

Автоматически созданное описание

Рисунок 15.1 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Statistica)



Рисунок 15.2 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Python)

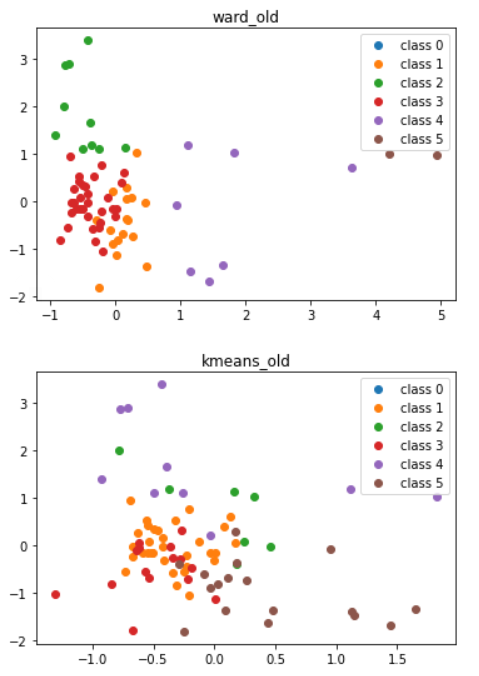


Рисунок 16.1, 16.2 – точечные графики старых кластеров в новой системе координат (2 главных компоненты)

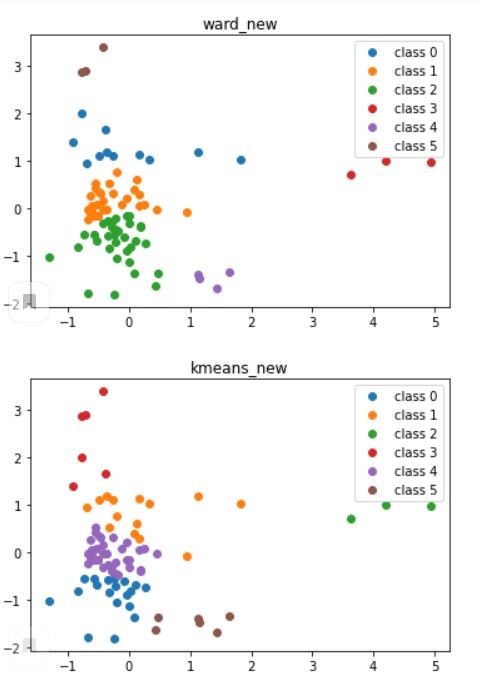


Рисунок 16.3, 16.4 – точечные графики новых кластеров в новой системе координат (2 главных компоненты)

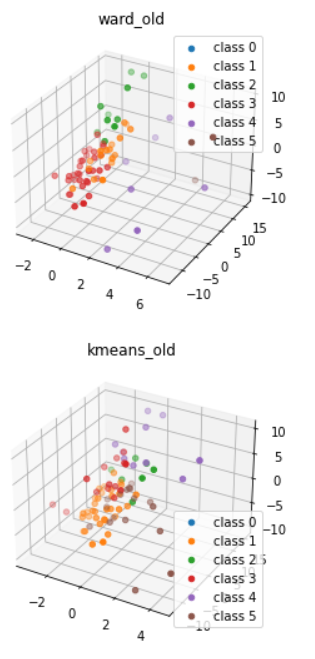


Рисунок 17.1, 17.2 – точечные графики старых кластеров в новой системе координат (3 главных компоненты)

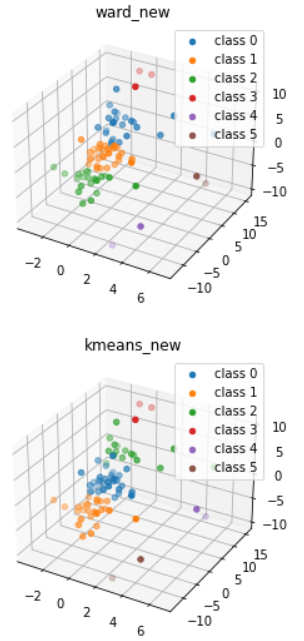


Рисунок 17.3, 17.4 – точечные графики новых кластеров в новой системе координат (3 главных компоненты)

# 

Рисунок 18.1 – дендрограмма для классификации методом Уорда, полученной по методу главных компонент (Statistica)