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 №3:

По курсу «Макростатический анализ и прогнозирование»

Работу выполнил: студент группы С18-712: Луканов А.В.

Проверила: Домашова Д. В.

Москва 2021

# 1. Постановка задачи

Субъект РФ характеризуются следующими признаками:

X1 - ОБЩИЕ КОЭФФИЦИЕНТЫ СМЕРТНОСТИ (число умерших на 1000 человек населения);

Х2 – СРЕДНЕДУШЕВЫЕ ДЕНЕЖНЫЕ ДОХОДЫ НАСЕЛЕНИЯ (в месяц; рублей);

Х3 – ОБЩАЯ ПЛОЩАДЬ ЖИЛЫХ ПОМЕЩЕНИЙ, ПРИХОДЯЩАЯСЯ В СРЕДНЕМ НА ОДНОГО ЖИТЕЛЯ (на конец года; квадратных метров);

Х4 – Уровень занятости для людей в возрасте 15-72 лет(процентов)

Х5 – коэффициенты миграционного прироста на 10 000 человек населения;

Х6 – зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1000 человек);

Х7 – Зарегистрировано преступлений экономической направленности на 1000 человек;

Х8 – ЧИСЛО ПРЕДПРИЯТИЙ И ОРГАНИЗАЦИЙ (на конец года) на 1000 человек;

Х9 – РАСХОДЫ НА ОХРАНУ ОКРУЖАЮЩЕЙ СРЕДЫ(млн руб) на 1000 чел;

Х10 - ВЫБРОСЫ ЗАГРЯЗНЯЮЩИХ ВЕЩЕСТВ В АТМОСФЕРНЫЙ ВОЗДУХ,   
ОТХОДЯЩИХ ОТ СТАЦИОНАРНЫХ ИСТОЧНИКОВ тысяч тонн;

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%.

# 2. Ход работы.

Первым делом стандартизируем исходные данные, чтобы привести их к одному масштабу.

Если исходные признаки, по которым производится классификация объектов, имеют разные единицы измерения, то необходимо перейти к стандартизованным переменным одним из следующих способов:

; ; ; ; ; ,

где  - исходное значение j-го признака на *i*-ом объекте наблюдения;

 - нормированное значение исходного *j*-го признака на *i*-ом объекте наблюдения;

 - среднее значение *j*-го признака;

 - выборочное среднеквадратическое отклонение *j*-го признака;

 - максимальное значение *j*-го признака;

 - минимальное значение *j*-го признака.

Была рассчитана матрица корреляций для исходных данных. Матрицы представлены на Рисунках 1.1 и 1.2.

Изображение выглядит как текст, кроссворд, табло

Автоматически созданное описание

Рисунок 1.1 – Результат расчета корреляционной матрицы (Statistica).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1.2 – Результат расчета корреляционной матрицы (Python).

Теперь проверим гипотезу о незначимости матрицы корреляций. Для этого посчитаем собственные числа для данной матрицы. Результаты представлены на Рисунках 2.1, 2.2.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 2.1 – Результаты расчета оценок собственных чисел корреляционной матрицы (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 3.1 – Результаты расчета оценок собственных чисел корреляционной матрицы (Python)

Затем были посчитаны наблюдаемое и критическое значение статистики Хи-квадрат, в результате гипотеза о незначимости корреляционной матрицы была отвергнута. Результаты представлены на Рисунках 3.1 и 3.2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hi2набл | Hi2крит1 | Hi2крит2 | alpha | k | n | det(R) = |
| 310,9174 | 65,64662 | 118,1359 | 0,05 | 10 | 85 | 0,021356 |

Рисунок 3.1– Проверка гипотезы о незначимости корреляционной матрицы в Excel



Рисунок 3.2– Проверка гипотезы о незначимости корреляционной матрицы в Python

Дальше были посчитаны доверительные интервалы для собственных значений корреляционной матрицы, результаты представлены на рисунках 4.1 и 4.2.

|  |  |  |
| --- | --- | --- |
| Предел 1 | Собственные числа | Предел 2 |
| 2,064305 | 2,688611 | 3,854247 |
| 1,573852 | 2,049830 | 2,938525 |
| 1,054556 | 1,373485 | 1,968954 |
| 0,971825 | 1,265732 | 1,814486 |
| 0,638714 | 0,831880 | 1,192538 |
| 0,474461 | 0,617952 | 0,885862 |
| 0,435703 | 0,567472 | 0,813497 |
| 0,195551 | 0,254692 | 0,365112 |
| 0,154579 | 0,201329 | 0,288614 |
| 0,114415 | 0,149018 | 0,213623 |

Рисунок 4.1 – Доверительные интервалы собственных чисел, посчитанные в Excel

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 4.2 – Доверительные интервалы собственных чисел, посчитанные в Python

Посчитаем информативность (Рисунок 5.1 и 5.2).

|  |  |
| --- | --- |
| Оценка информативности | |
| I\_1(z(x)) = | 0,268861 |
| I\_2(z(x)) = | 0,473844 |
| I\_3(z(x)) = | 0,611193 |
| I\_4(z(x)) = | 0,737766 |

Рисунок 5.1 – Оценка информативности (Excel)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 5.2 – Оценка информативности (Python)

Информативность превышает 70% при использовании 4 главных компонент. Проверим гипотезу о достаточности 4 главных компонент с помощью критерия Кайзера (Рисунок 6.1 и 6.2).



Рисунок 6.1 – Критерий Кайзера (Excel)

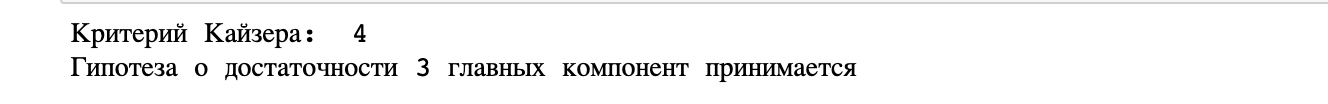


Рисунок 6.2 – Критерий Кайзера (Python)

Построим график «осыпи». Результаты представлены на рисунках 7.1 и 7.2.

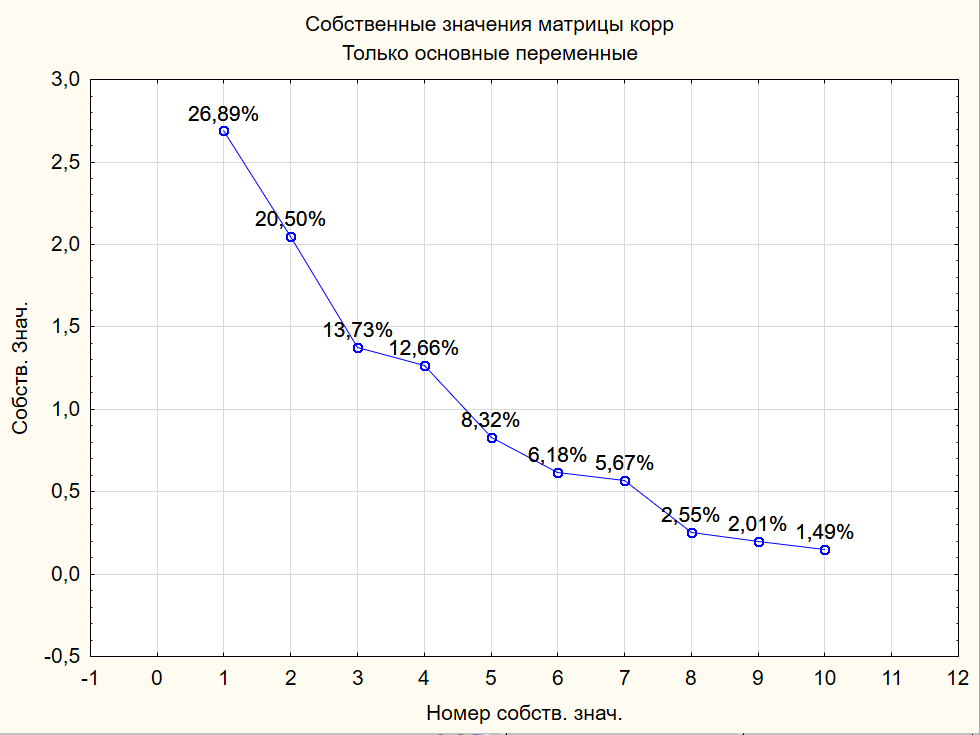


Рисунок 7.2 – График «осыпи» (Statistica)

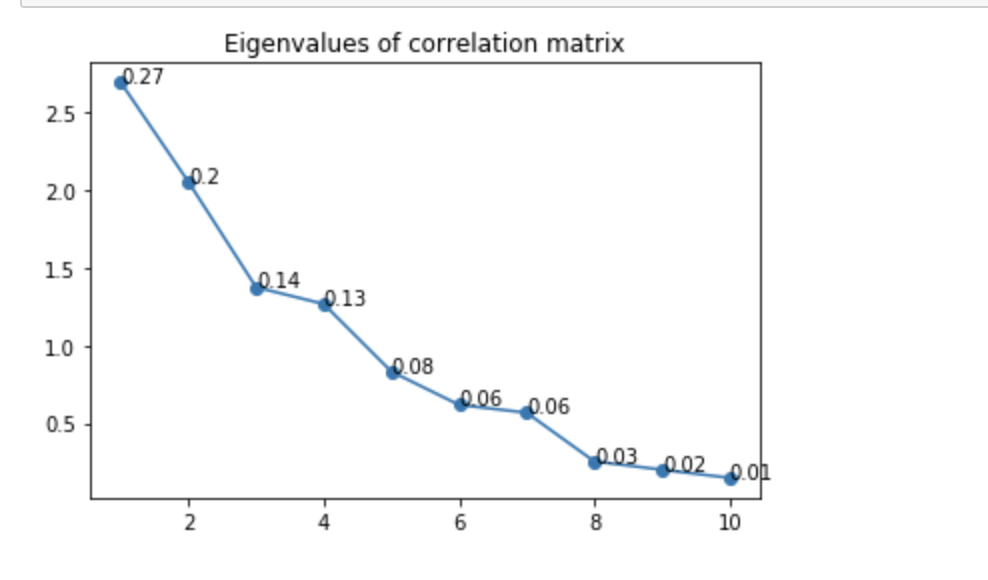


Рисунок 7.2 – График «осыпи» (Python)

Определим вклад каждой главной компоненты в суммарную дисперсию исходных признаков (Рисунок 8.1 и 8.2).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 8.1 – Вклады главных компонент в суммарную дисперсию исходных признаков (Statictica)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 8.2 – Вклады главных компонент в суммарную дисперсию исходных признаков (Python)

Рассчитаем собственные вектора корреляционной матрицы. Результаты представлены на рисунке 9.

Изображение выглядит как текст, кроссворд, табло

Автоматически созданное описание

Рисунок 9 – Результаты расчета собственных векторов корреляционной матрицы (Statistica)

После этого посчитаем коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Рисунок 10).

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Рисунок 10.1 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Statistica)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 10.2 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Python)

Рассчитаем матрицу нагрузок . Результаты представлены на рисунках 11.1 и 11.2.

Изображение выглядит как текст, кроссворд, окно

Автоматически созданное описание

Рисунок 11.1 – Результаты расчета элементов матрицы нагрузок (Statistica)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 11.2 – Результаты расчета элементов матрицы нагрузок (Python)

Проведем анализ 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 8 главными компонентами не наблюдается тесной связи (не имеется значений >0,7) Это подтверждает правильность выделения только 3 первых главных компонент.

Первая главная компонента тесно отрицательно связана (модуль коэффициента корреляции больше 0,7) с двумя исходными признаками: x2 (Среднедушевые доходы населения), x4 (Уровень занятости для людей в возрасте 15-72 лет), поэтому первую главную компоненту можно интерпретировать «Уровень доходов населения».

Вторая главная компонента тесно связана с признаком x1 (Коэффициент смертности) и с признаком х3 (Общая площадь жилых помещений, приходящихся на одного человека). Вторую главную компоненту можно назвать «Показатель смертности в регионе»

Третья главная компонента тесно связана с х6 (зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1000 человек)) и х7 (Зарегистрировано преступлений экономической направленности на 1000 человек) и отрицательно связана с х5(коэффициенты миграционного прироста на 10 000 человек населения). Третью главную компоненту можно назвать «Уровень преступности в регионе»

Х6 – зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1000 человек);

Рассчитаем матрицу индивидуальных значений центрировано - нормированных главных компонент. Результаты представлены на рисунках 12.1 и 12.2.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 12.1 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Statistica)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 12.2 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Python)

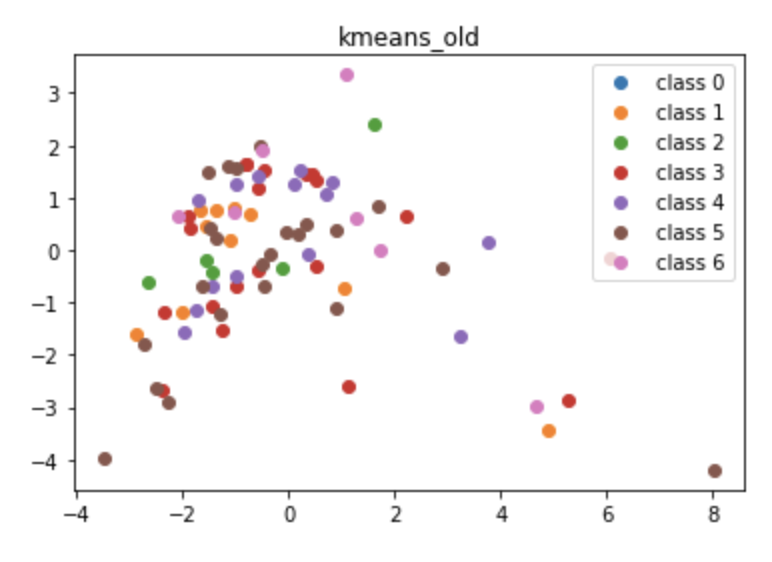
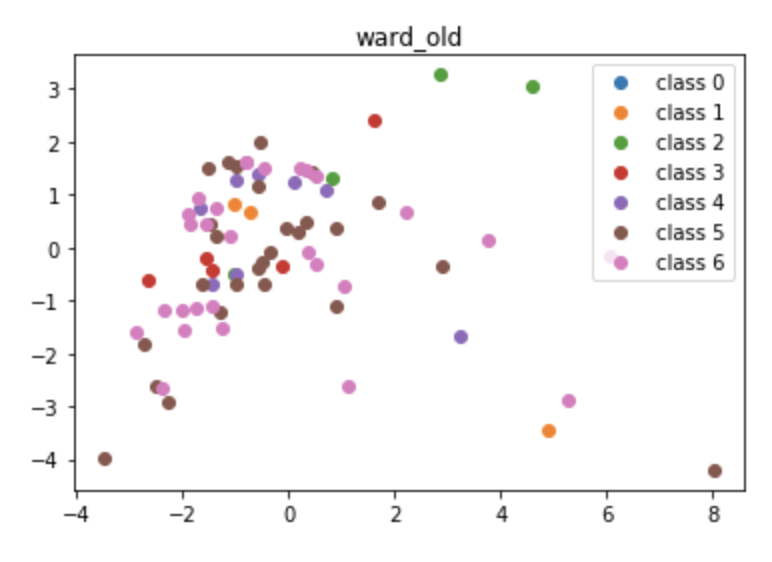
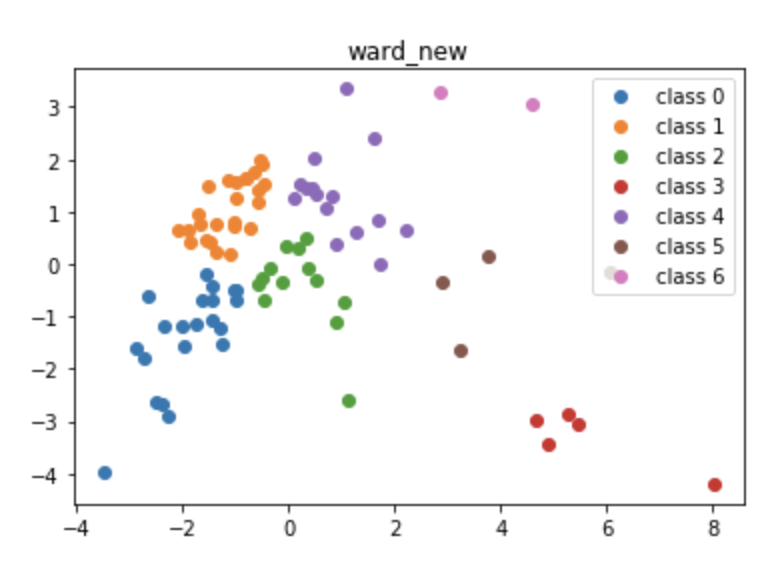


Рисунок 13.1, 13.2 – точечные графики старых кластеров в новой системе координат (2 главных компоненты)



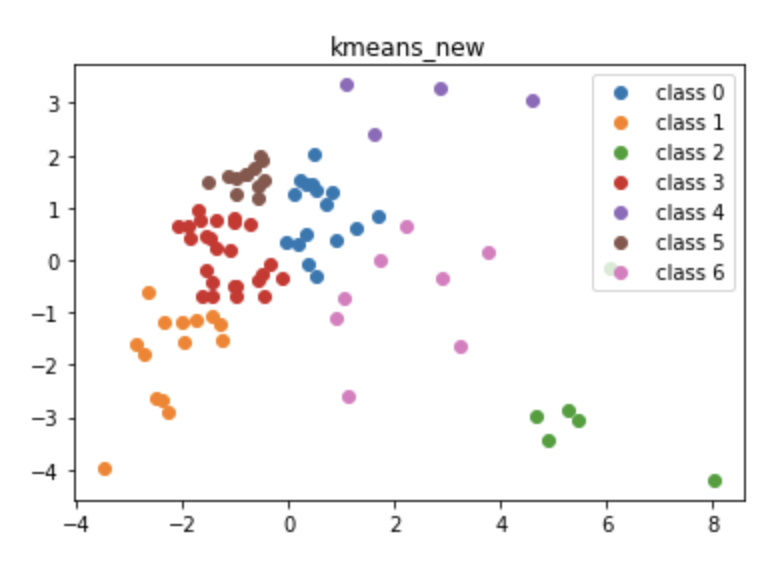
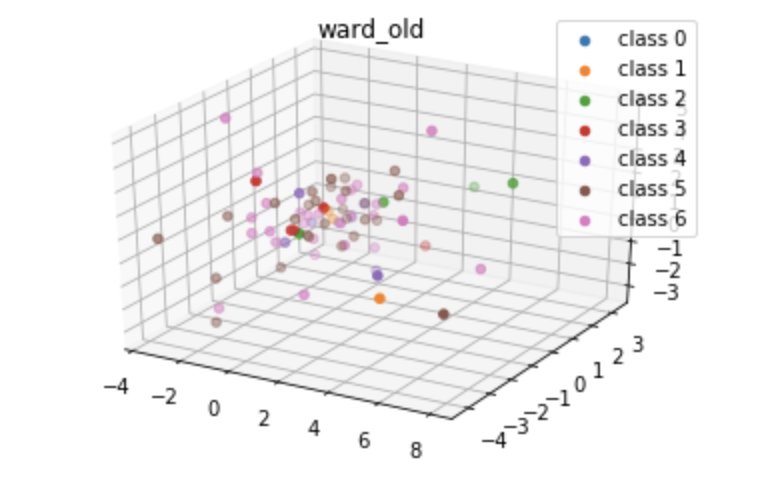


Рисунок 14.1, 14.2 – точечные графики новых кластеров в новой системе координат (2 главных компоненты)



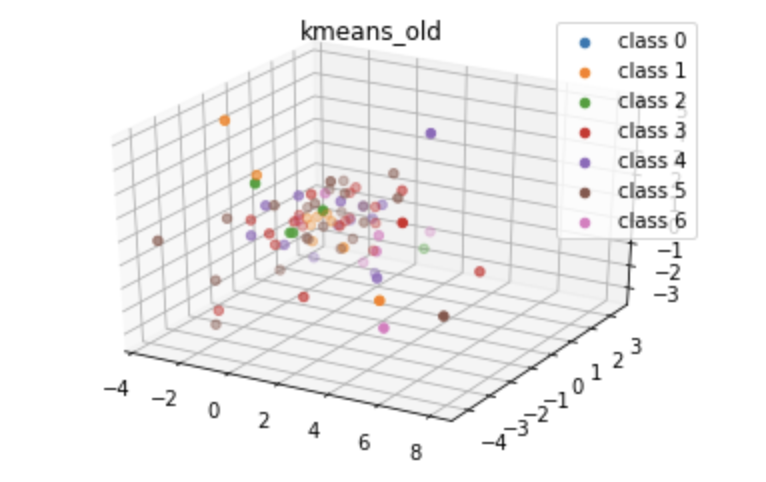
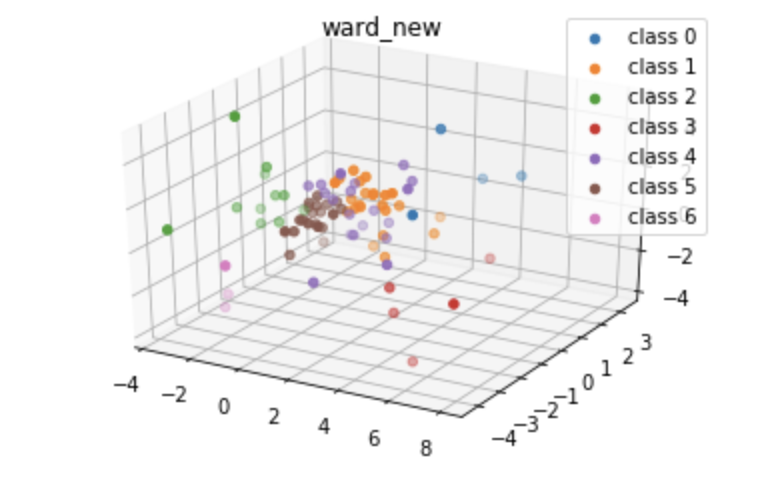


Рисунок 15.1, 15.2 – точечные графики старых кластеров в новой системе координат (3 главных компоненты)



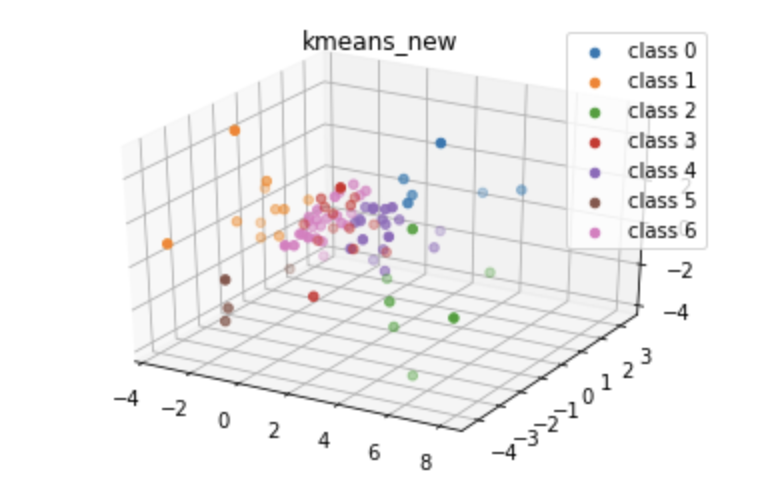


Рисунок 16.1, 16.2 – точечные графики новых кластеров в новой системе координат (3 главных компоненты)

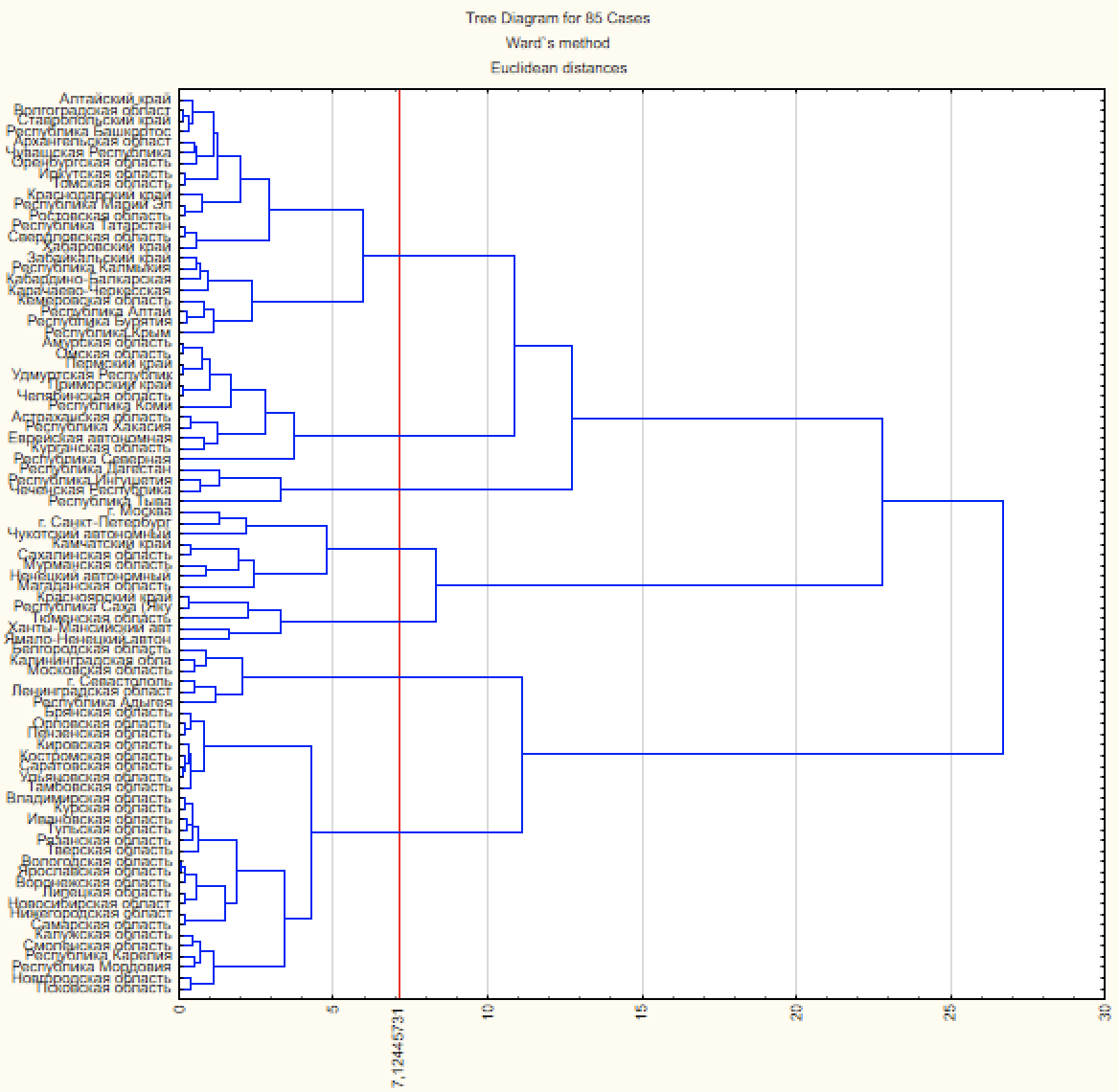


Рисунок 17 – Дендрограмма для кластеризации методом Уорда, полученная по методу главных компонент