

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра теории вероятностей и математической статистики

ОТЧЕТ по лабораторной З
«АДЭФС. Лабораторная работа №3»

Зборовского Артёма Николаевича
студента 3 курса, 8 группы
Преподаватель
Полузеров Тимофей Дмитриевич

Минск 2025

1. ВВЕДЕНИЕ

Данный отчет представляет результаты комплексного анализа данных с применением методов факторного анализа и кластеризации. Цель исследования - выявление скрытых структур в данных и группировка наблюдений в однородные кластера. Анализ проведен на данных за период 2005-2011 годов.

2. ИСХОДНЫЕ ДАННЫЕ

2.1. Характеристики набора данных

-Источник данных: файл `Annual 2005-2011.csv`

- Объем данных: 2695 наблюдений, 22 переменные

- Период: 2005-2011 годы

- Исходные переменные: 20 анализируемых переменных (k1-k20)

2.2. Предобработка данных

1. Загрузка и индексация: 2695 наблюдений успешно загружены

2. Выбор переменных: Отобраны 20 переменных k1-k20 для анализа

3. Обработка пропусков: Пропущенные значения отсутствуют

4. Стандартизация: Данные стандартизированы (среднее = 0, стандартное отклонение = 1)

3. ФАКТОРНЫЙ АНАЛИЗ

3.1. Методология

- Метод: Метод главных компонент (PCA) с последующим вращением

- Количество факторов: 5

- Метод вращения: Квартимакс (Quartimax)

3.2. Результаты факторного анализа

Таблица 1: Характеристики факторов

Результаты факторного анализа после вращения:

Фактор	Собственное значение	Доля дисперсии %	Накопленная доля %
F1	1	3.564	17.814
F2	2	3.504	17.514
F3	3	1.558	7.786
F4	4	1.505	7.522
F5	5	2.745	13.720

4. КЛАСТЕРНЫЙ АНАЛИЗ

4.1. Методология

- Алгоритм: K-means
- Количество кластеров: 4
- Инициализация: k-means++
- Количество запусков: 10
- Сходимость: достигнута за 32 итерации
- Inertia: 38485.92

4.2. Характеристики кластеров

Таблица 2: Распределение наблюдений по кластерам

Cluster	Распределение по кластерам:	
	Кол-во наблюдений	Процент
1	779	28.9
2	225	8.3
3	670	24.9
4	1021	37.9

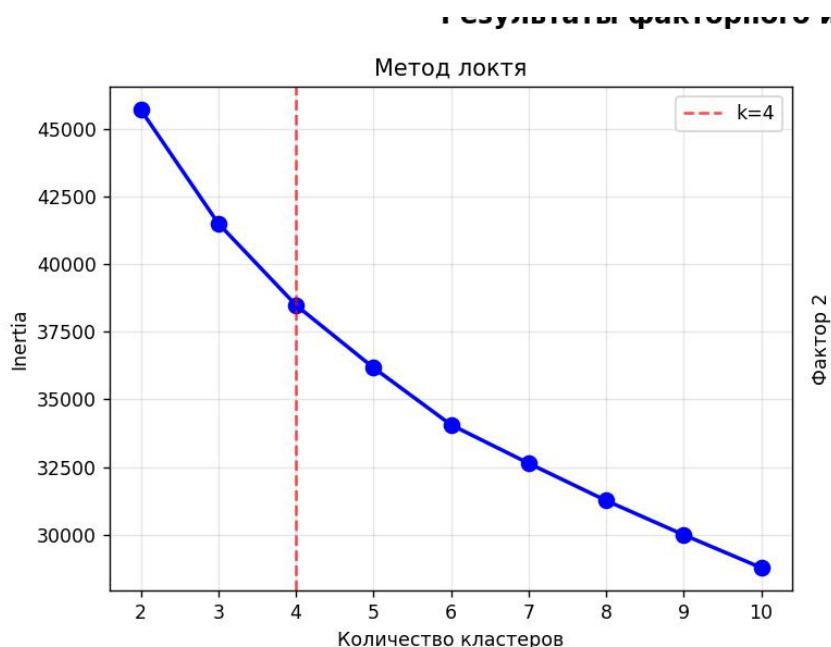
Таблица 3: Описательная статистика интегрального показателя по кластерам

	Количество	Среднее	Медиана	...	Максимум	Q1	Q3
Cluster				...			
1	779	-0.4134	-0.3659	...	1.3185	-0.6622	-0.1036
2	225	0.5566	0.4965	...	2.7785	0.2446	0.8446
3	670	0.5127	0.4601	...	2.4337	0.2473	0.6989
4	1021	-0.1437	-0.1292	...	1.5689	-0.3636	0.0802

5. ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

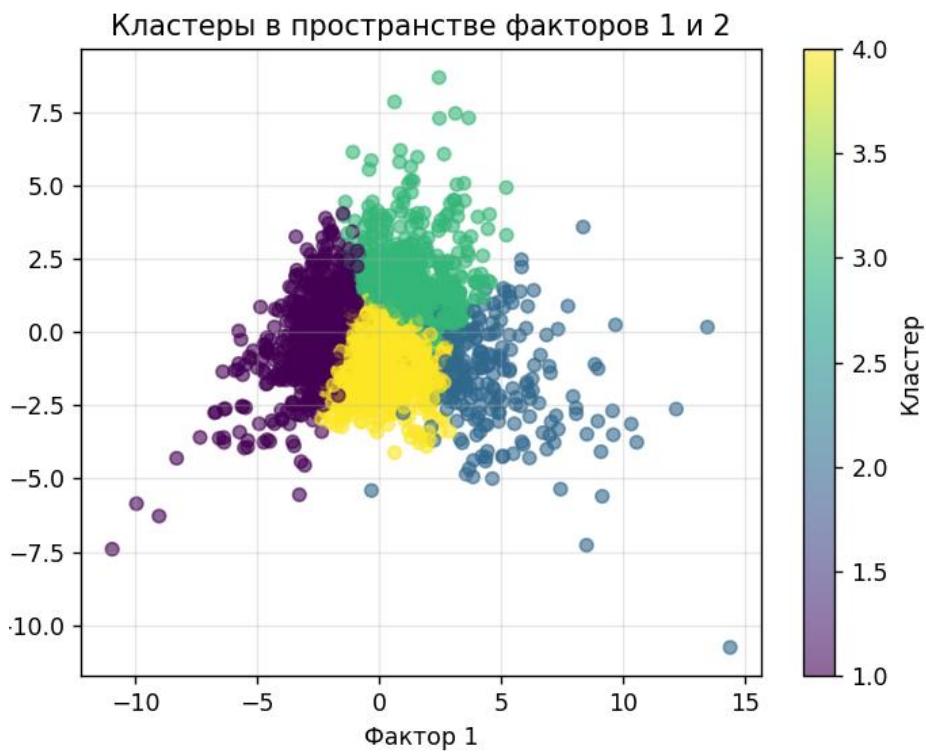
5.1. График 1: Метод локтя

График показывает зависимость инерции от количества кластеров. Точка изгиба при $k=4$ указывает на оптимальное количество кластеров.



5.2. График 2: Кластеры в пространстве факторов

Диаграмма рассеяния показывает распределение наблюдений в пространстве первых двух факторов (объясняют 35.3% дисперсии).



5.3 Полный вывод результатов

```
Размер данных: (2695, 22)

Первые 5 строк:
   empl_num      k1      k2      k3 ...     k18      k19      k20    year
1   6095.0  0.942380  0.060563  0.678302 ...  0.076738  0.055049  0.034904  5.0
2   255.0   1.980494  0.274382  0.916775 ...  0.116068  0.059740  0.025647  5.0
3   114.0   0.374160  0.001494  0.085138 ... -0.584879  0.010563  0.000000  5.0
4   365.0   7.859079  0.831978  2.449864 ...  0.171731  0.496295  0.312415  5.0
5   168.0   1.779376  0.005596  0.883293 ...  0.064809  0.025726  0.011839  5.0

[5 rows x 22 columns]

Переменные для анализа: ['k1', 'k2', 'k3', 'k4', 'k5', 'k6', 'k7', 'k8', 'k9', 'k10', 'k11', 'k12', 'k13', 'k14', 'k15', 'k16', 'k17', 'k18', 'k19', 'k20']
Размер данных: (2695, 20)
      k1      k2      k3 ...     k18      k19      k20
1  0.942380  0.060563  0.678302 ...  0.076738  0.055049  0.034904
2  1.980494  0.274382  0.916775 ...  0.116068  0.059740  0.025647
3  0.374160  0.001494  0.085138 ... -0.584879  0.010563  0.000000
4  7.859079  0.831978  2.449864 ...  0.171731  0.496295  0.312415
5  1.779376  0.005596  0.883293 ...  0.064809  0.025726  0.011839

[5 rows x 20 columns]

Размер данных после удаления пропущенных значений: (2695, 20)
```

Результаты факторного анализа после вращения:

Фактор	Собственное значение	Доля дисперсии %	Накопленная доля %
F1	1	3.564	17.814
F2	2	3.504	35.328
F3	3	1.558	43.114
F4	4	1.505	50.636
F5	5	2.745	64.356

Всего объяснено дисперсии: 64.4%

Факторные значения для первых 10 наблюдений:

	Фактор1	Фактор2	Фактор3	Фактор4	Фактор5
1	-0.715497	-1.019521	0.174932	0.651189	-1.325775
2	1.373659	-1.618779	0.688106	0.547963	-0.919859
3	-4.741825	-3.761297	0.155155	-0.049490	0.136615
4	7.734329	0.894109	0.651998	-1.876357	0.743991
5	-0.463468	-2.460454	-0.249057	-0.153288	-2.961011
6	3.816360	-1.413269	-0.274427	0.108969	0.217120
7	-1.790639	0.267790	-0.092734	-0.304781	-0.350713
8	3.402507	-3.776316	-0.115279	0.405018	0.537360
9	1.214154	0.028369	0.886277	-0.852726	-0.557507
10	-0.210609	0.275879	0.386015	-0.333168	-0.382206

Финальная таблица результатов (первые 10 строк):

	I	k1	k2	k3	...	Фактор2	Фактор3	Фактор4	Фактор5
1	-0.425	0.942	0.061	0.678	...	-1.020	0.175	0.651	-1.326
2	-0.070	1.980	0.274	0.917	...	-1.619	0.688	0.548	-0.920
3	-1.476	0.374	0.001	0.085	...	-3.761	0.155	-0.049	0.137
4	1.546	7.859	0.832	2.450	...	0.894	0.652	-1.876	0.744
5	-0.951	1.779	0.006	0.883	...	-2.460	-0.249	-0.153	-2.961
6	0.449	4.200	1.310	2.356	...	-1.413	-0.274	0.109	0.217
7	-0.350	0.836	0.008	0.182	...	0.268	-0.093	-0.305	-0.351
8	0.040	4.097	1.194	1.824	...	-3.776	-0.115	0.405	0.537
9	0.150	2.194	0.112	0.741	...	0.028	0.886	-0.853	-0.558
10	-0.037	1.216	0.063	0.414	...	0.276	0.386	-0.333	-0.382

Кластеризация завершена:

Количество итераций: 32

Inertia: 38485.92

Центроиды кластеров (в исходных единицах измерения):

	k1	k2	k3	k4	...	k17	k18	k19	k20
1	1.1620	0.0538	0.4055	-0.5094	...	70.1405	-0.0116	0.0140	-0.0182
2	6.1023	1.4358	3.1045	0.6747	...	188.6455	0.1130	0.1482	0.1156
3	1.9438	0.2199	0.8368	0.2432	...	383.0280	0.1188	0.1774	0.1956
4	1.7777	0.1265	0.6352	0.1810	...	184.8714	0.0511	0.0493	0.0321

[4 rows x 20 columns]

Распределение по кластерам:

Cluster	Кол-во наблюдений	Процент
1	779	28.9
2	225	8.3
3	670	24.9
4	1021	37.9

Описательная статистика по кластерам:

Cluster	Количество	Среднее	Медиана	...	Максимум	Q1	Q3
1	779	-0.4134	-0.3659	...	1.3185	-0.6622	-0.1036
2	225	0.5566	0.4965	...	2.7785	0.2446	0.8446
3	670	0.5127	0.4601	...	2.4337	0.2473	0.6989
4	1021	-0.1437	-0.1292	...	1.5689	-0.3636	0.0802