МИНОБРНАУКИ РОССИИ

РГУ НЕФТИ И ГАЗА (НИУ) ИМЕНИ И.М. ГУБКИНА

Факультет **Автоматики и вычислительной техники**
Кафедра **Прикладной математики**

ОТЧЕТ

по дисциплине Big Data

на тему Кластеризация

ПРОВЕРИЛА:

Косова Ксения Олеговна
(фамилия, имя, отчество)

(подпись)

(дата)

ВЫПОЛНИЛ:
Студент группы **АМ-17-06**
(номер группы)

Хуснутдинов Эдуард Тимурович
(фамилия, имя, отчество)

(подпись)

(дата)

Москва, 20 19

ПОСТАНОВКА ЗАДАЧИ

В ходе выполнения данной работы будет проведен анализ социально-экономических показателей субъектов Российской Федерации, а затем, кластеризация последних с помощью агломеративных (ближайшего соседа, средней связи, дальнего соседа) и эталонных (k-средних) методов, с помощью вычисления критерия Т и анализа дендрограмм будет определено оптимальное количество кластеров и метод кластеризации. Целью кластеризации является выявлении динамики экономического и социального развития субъектов РФ.

ИСХОДНЫЕ ДАННЫЕ

В качестве исходных данных были взяты следующие социально-экономические показатели РФ за 2017 год: среднедушевые денежные доходы населения, удельный вес аварийного жилищного фонда в общей площади всего жилищного фонда, коэффициенты естественного прироста населения на 1000 человек населения, уровень безработицы, использование сети интернет населением. Для проведения работы мною были выбраны три федеральных округа Российской Федерации, а именно: Центральный федеральный округ (субъекты 1-18), Северо-Кавказский федеральный округ (субъекты 19-25) и Приволжский федеральный округ (субъекты 26-39).

Таблица 1 – Исходные данные

№	Субъект РФ	Среднедушевые денежные доходы	Удельный вес аварийного жилищного фонда	Коэффициенты естественного прироста населения	Уровень безработицы	Использование сети
1	Белгородская область	30074	0,002	-0,2	3,9	76,6
2	Брянская область	26402	0,003	-0,8	4,4	75,3
3	Владимирская область	23988	0,006	-0,8	4,8	76,5
4	Воронежская область	29327	0,003	-0,1	4,3	78,7
5	Ивановская область	24760	0,005	-0,8	4,7	81,7
6	Калужская область	28108	0,006	-0,2	4	73,6
7	Костромская область	24745	0,01	-0,7	5,3	76,5
8	Курская область	26425	0,001	-0,7	4,1	77,4
9	Липецкая область	29294	0,006	-0,5	3,9	79,9
10	Московская область	41286	0,003	1,1	3,2	90,4
11	Орловская область	24122	0,006	-1	6,5	68,5
12	Рязанская область	24789	0,006	-0,5	4,1	69,7
13	Смоленская область	25398	0,003	-0,4	5,7	78,7
14	Тамбовская область	25938	0,003	-0,7	4,4	74,3
15	Тверская область	24077	0,006	-1	4,5	78,3
16	Тульская область	27774	0,015	-0,5	3,9	82,4
17	Ярославская область	27625	0,008	-0,4	6,6	75,7
18	г. Москва	62532	0	1	1,4	83,1
19	Республика Дагестан	29206	0,002	0,7	12	78,4
20	Республика Ингушетия	15131	0,007	1,6	27	73,7
21	Кабардино-Балкарская Республика	20385	0,002	0,2	10,5	82
22	Карачаево-Черкесская Республика	17142	0,002	0,03	13,5	76,9
23	Республика Северная Осетия – Алания	22773	0,005	-0,2	11,8	82
24	Чеченская Республика	22202	0,004	1,6	14	63
25	Ставропольский край	23403	0,001	-0,1	5,2	80,8
26	Республика Башкортостан	28442	0,003	-0,1	5,6	82
27	Республика Марий Эл	19017	0,011	0,3	6,1	79,6
28	Республика Мордовия	18065	0,005	-0,4	4,2	70,3
29	Республика Татарстан	31719	0,002	0,2	3,5	91,2
30	Удмуртская Республика	23925	0,007	-0,2	4,8	76,1
31	Чувашская Республика	17892	0,006	-0,4	5,1	69
32	Пермский край	28655	0,018	-0,3	6,1	74
33	Кировская область	21560	0,008	-0,7	5,3	73,3
34	Нижегородская область	30742	0,006	-0,4	4,2	75,9
35	Оренбургская область	22689	0,007	-0,6	4,6	77,5
36	Пензенская область	21611	0,005	-0,7	4,5	75,7
37	Самарская область	26988	0,009	-0,3	4,2	81,8
38	Саратовская область	19825	0,007	-0,7	4,8	78,9
39	Ульяновская область	23133	0,003	-0,5	4,4	67,2

Проведем первичный анализ данных:

1. Среднедушевые денежные доходы населения

Рисунок 1 – Среднедушевые доходы населения

Из диаграммы видно, что наибольшие среднедушевые доходы населения достигаются в Центральном федеральном округе, в г. Москва и в Московской области, самые низкие же доходы в Республике Ингушетия.

2. Удельный вес аварийного жилищного фонда в общей площади всего жилищного фонда

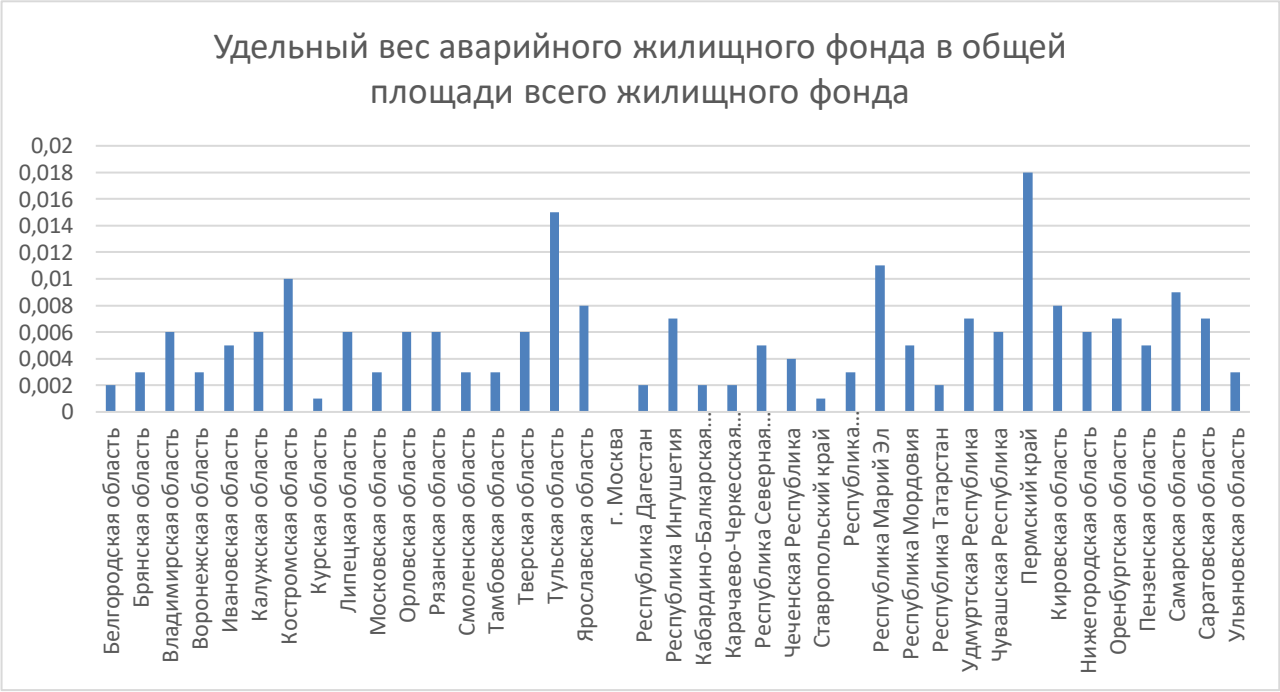

Рисунок 2 – Удельный вес аварийного жилищного фонда в общей площади всего жилья

Самый низкий показатель достигается в г. Москва, он стремится к нулю, так же, очень низкий показатель в Курской области и в Ставропольском крае, в

Пермском крае и в Тульской области, напротив, показатель достигает невиданных высот.

3. Коэффициенты естественного прироста населения на 1000 человек населения

Рисунок 3 – Коэффициенты естественного прироста населения на 1000 человек населения

Наибольший прирост населения наблюдается в Северо-Кавказском федеральном округе и в двух региона Центрального ФО (г. Москва и Московская область).

4. Уровень безработицы

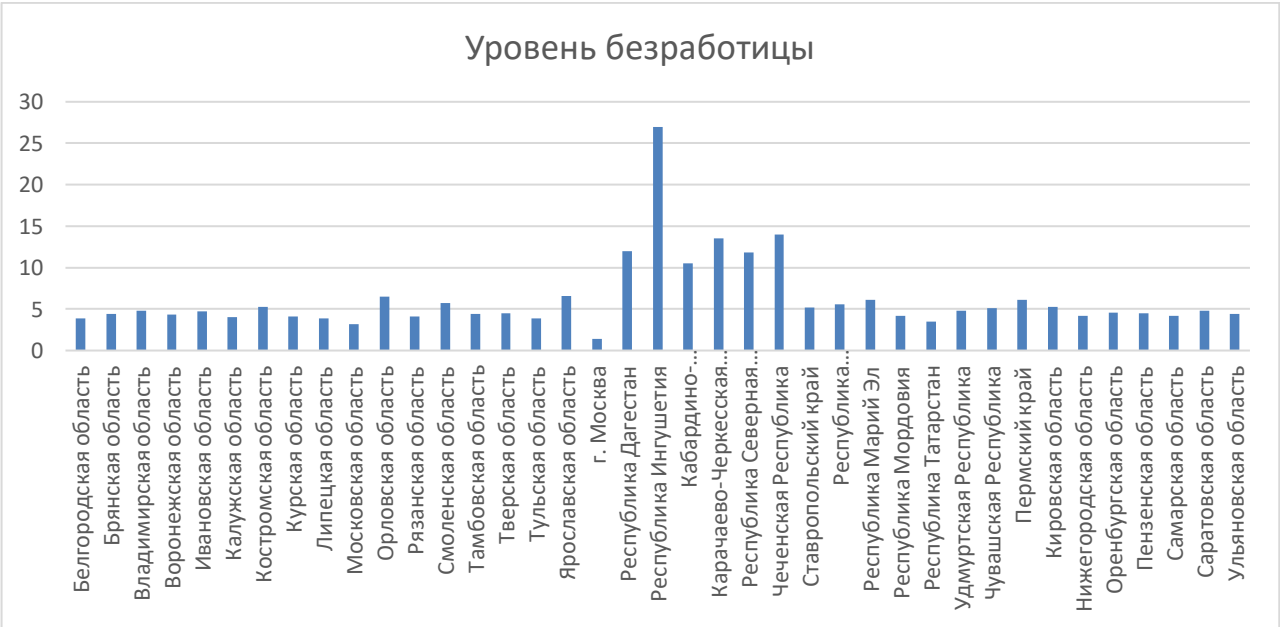

Рисунок 4 – Уровень безработицы

По уровню безработицы наибольший показатель установлен в Северо-Кавказском ФО, особенно в Республике Ингушетия, самый низкий же достигается в г. Москва.

5. Использование сети интернет населением

Рисунок 5 – Использование сети интернет населением

По данному показателю мы имеем в целом равные значения по регионам, стоит отметить лишь, что самый низкий показатель достигнут в Чеченской Республике, самые высокие в Московской области и Республике Татарстан.

Таким образом, можно выдвинуть гипотезу о маргинальности следующих регионов: г. Москва (№ 18), Московская область (№ 10), Республика Ингушетия (№ 20), Чеченская Республика (№ 24) .

Нормировка данных:

Для дальнейшей работы используем нормировку сравнения, данные будут лежать в диапазоне от -∞ до 1. Так как для показателя №3 (коэффициенты естественного прироста населения) максимум будет положительным, то при нормировке данные будут лежать от -1 до 1.

Листинг программы с нормировкой:

```
for i=1:5
     maxx=max(X(:,i));
     X(:,i)=X(:,i)/maxx;
end;
```

Таблица 2 – Нормированные данные

№	Субъект РФ	Среднедушевые денежные	Удельный вес аварийного жилищного фонда	Коэффициенты естественного прироста	Уровень безработицы	Использование сети
1	Белгородская область	0,48093776	0,111111111	-0,125	0,144444444	0,839912281
2	Брянская область	0,422215825	0,166666667	-0,5	0,162962963	0,825657895
3	Владимирская область	0,383611591	0,333333333	-0,5	0,177777778	0,838815789
4	Воронежская область	0,468991876	0,166666667	-0,0625	0,159259259	0,862938596
5	Ивановская область	0,39595727	0,277777778	-0,5	0,174074074	0,895833333
6	Калужская область	0,449497857	0,333333333	-0,125	0,148148148	0,807017544
7	Костромская область	0,395717393	0,555555556	-0,4375	0,196296296	0,838815789
8	Курская область	0,422583637	0,055555556	-0,4375	0,151851852	0,848684211
9	Липецкая область	0,468464146	0,333333333	-0,3125	0,144444444	0,876096491
10	Московская область	0,660237958	0,166666667	0,6875	0,118518519	0,99122807
11	Орловская область	0,385754494	0,333333333	-0,625	0,240740741	0,751096491
12	Рязанская область	0,396421032	0,333333333	-0,3125	0,151851852	0,764254386
13	Смоленская область	0,406160046	0,166666667	-0,25	0,211111111	0,862938596
14	Тамбовская область	0,414795625	0,166666667	-0,4375	0,162962963	0,814692982
15	Тверская область	0,385034862	0,333333333	-0,625	0,166666667	0,858552632
16	Тульская область	0,444156592	0,833333333	-0,3125	0,144444444	0,903508772
17	Ярославская область	0,441773812	0,444444444	-0,25	0,244444444	0,83004386
18	г. Москва	1	0	0,625	0,051851852	0,911184211
19	Республика Дагестан	0,467056867	0,111111111	0,4375	0,444444444	0,859649123
20	Республика Ингушетия	0,24197211	0,388888889	1	1	0,808114035
21	Кабардино-Балкарская Республика	0,325993092	0,111111111	0,125	0,388888889	0,899122807
22	Карачаево-Черкесская Республика	0,274131645	0,111111111	0,01875	0,5	0,843201754
23	Республика Северная Осетия – Алания	0,364181539	0,277777778	-0,125	0,437037037	0,899122807
24	Чеченская Республика	0,355050214	0,222222222	1	0,518518519	0,690789474
25	Ставропольский край	0,374256381	0,055555556	-0,0625	0,192592593	0,885964912
26	Республика Башкортостан	0,454839122	0,166666667	-0,0625	0,207407407	0,899122807
27	Республика Марий Эл	0,304116292	0,611111111	0,1875	0,225925926	0,872807018
28	Республика Мордовия	0,288892087	0,277777778	-0,25	0,155555556	0,770833333
29	Республика Татарстан	0,507244291	0,111111111	0,125	0,12962963	1
30	Удмуртская Республика	0,382604107	0,388888889	-0,125	0,177777778	0,834429825
31	Чувашская Республика	0,286125504	0,333333333	-0,25	0,188888889	0,756578947
32	Пермский край	0,458245378	1	-0,1875	0,225925926	0,811403509
33	Кировская область	0,344783471	0,444444444	-0,4375	0,196296296	0,80372807
34	Нижегородская область	0,49162029	0,333333333	-0,25	0,155555556	0,832236842
35	Оренбургская область	0,362838227	0,388888889	-0,375	0,17037037	0,849780702
36	Пензенская область	0,345599053	0,277777778	-0,4375	0,166666667	0,83004386
37	Самарская область	0,431587027	0,5	-0,1875	0,155555556	0,896929825
38	Саратовская область	0,317037677	0,388888889	-0,4375	0,177777778	0,865131579
39	Ульяновская область	0,369938591	0,166666667	-0,3125	0,162962963	0,736842105

ХОД РАБОТЫ

Итак, после нормировки, запускаем цикл от 1 до 39 с шагом 1, это необходимо для поиска оптимального количества кластеров, внутри цикла кластеризуем данные сначала эталонным методом, который определяется функцией kmeans(), на вход которой поступают исходные данные и количество кластеров, на которые их необходимо разбить (количеством кластеров является текущий счетчик цикла). Далее проводим кластеризацию агломеративными методами (ближнего соседа, средней связи и полной связи). Общее описание работы агломеративных методов: в начале работы алгоритма каждый элемент выборки считается отдельным кластером. После чего кластеры последовательно объединяются, пока все элементы не попадут в один кластер. На каждом шаге объединяются два кластера, расстояние между которыми минимально, минимальное же расстояние определяется конкретным методом, так, в методе ближнего соседа минимальным расстоянием между кластерами будет являться расстояние до ближайшего члена кластера, а, например, в методе полной связи минимальным расстоянием между кластерами будет являться максимальное расстояние до члена кластера.

Реализация агломеративных методов происходит похожим образом, сначала определяем квадратичную форму матрицы попарных расстояние, также используем метрику Махаланобиса, так как между двумя из показателей наблюдается корреляция 0.58, что больше необходимых нам 0.4 :

```
p=squareform(pdist(X,'mahalanobis'));
```

Далее используется функция linkage(), она строит иерархическое дерево бинарных кластеров, на вход ей подается найденная раннее матрица квадратичной формы попарных расстояний и метод (‘single’, ‘average’ или ‘complete’) для метода ближнего соседа, средней связи и полной связи соответственно. В переменную Ts записываем результат кластеризации с помощью функции cluster(), на вход которой поступает иерархическое дерево бинарных кластеров и количество кластеров (номер счетчика цикла).

Рассмотрим более подробно специфику работы алгоритмов и сравним их:

1. Метод ближнего соседа

Расстояние между классами рассчитывается по формуле:

$$\rho_{min}(S_l, S_p) = \min d(x_i, x_j), \quad \text{где } x_i \in S_l, x_j \in S_p,$$

Листинг:

```
pb=squareform(pdist(X,'mahalanobis'));
Zb = linkage(pb, 'single');
Tb = cluster(Zb,'maxclust',cl);
```

Дендрограмма:

```
H1 = dendrogram(Zb)
```

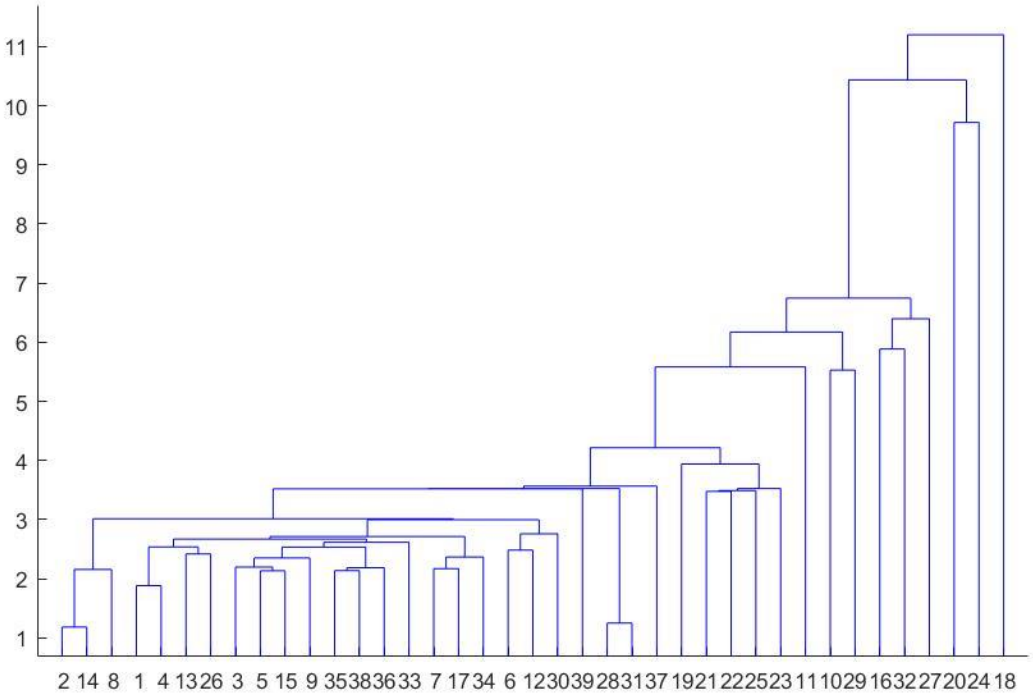

Рисунок 6 – Дендрограмма метода ближнего соседа

По данной дендрограмме видно, что регионы 18, 20 и 24 плохо кластеризуются, это г. Москва, Республика Ингушетия и Чеченская Республика, ранее мы выдвинули гипотезу о маргинальности данных регионов.

2. Метод средней связи

Расстояние между классами рассчитывается по формуле: $\rho_{mean}(S_l, S_p) = \frac{1}{n_l}\frac{1}{n_p}\sum_{x_i \in S_l}\sum_{x_j \in S_p} d(x_i, x_j)$, где n – количество объектов в кластере

Листинг:

```
psr=squareform(pdist(X,'mahalanobis'));
Zsr = linkage(psr, 'average');
Tsr = cluster(Zsr,'maxclust',cl);
```

Дендрограмма:

```
H2 = dendrogram(Zsr)
```

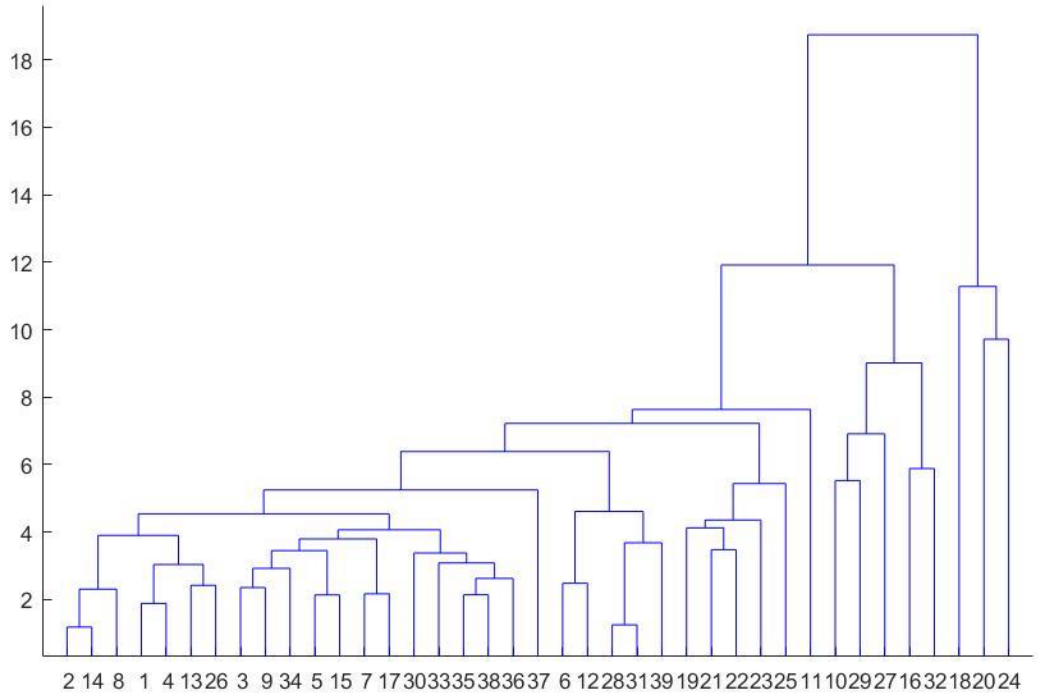

Рисунок 7 – Дендрограмма метода средней соседа

Здесь мы видим, что при делении на 4 кластера, г. Москва заносится в отдельный кластер, так же как и Чеченская Республика и Республика Ингушетия образуют собой один кластер, все эти регионы являются по нашей гипотезе маргинальными.

3. Метод полной связи

Расстояние между кластерами рассчитывается по формуле

$$\rho_{max}(S_l, S_p) = \max d(x_i, x_j), \text{ где } x_i \in S_l,\ x_j \in S_p,$$

где S_l и S_p – разные кластеры.

Листинг:

```
pp=squareform(pdist(X,'mahalanobis'));
Zp = linkage(pp, 'complete');
Tp = cluster(Zp,'maxclust',cl);
```

Дендрограмма:

```
H3 = dendrogram(Zp)
```

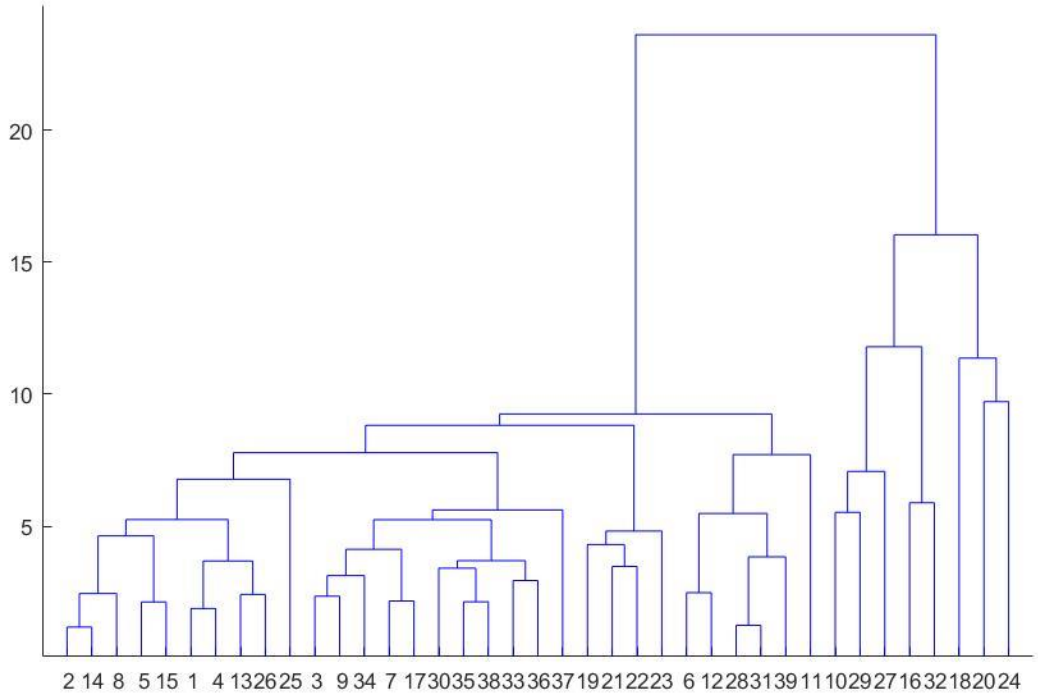

Рисунок 8 – Дендрограмма метода полной соседа

Результат в целом совпадает с предыдущим методом.

4. Метод k-средних

На первом шаге работы алгоритма исследуемое множество объектов случайным образом делится на заданное число k кластеров, в каждом из которых вычисляется центр. Объект относится к тому кластеру, к центру которого он ближе в выбранной метрике. Основная идея заключается в том, что на каждой итерации центры вычисляются заново для каждого кластера, полученного на предыдущем шаге, затем элементы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Листинг:

```
[idkm,C1] = kmeans(X,cl);
```

Вычисление оптимального количества кластеров:

Для определения оптимального количества кластеров помимо анализа исходных данных в виде дендрограмм, вычисляют также критерий T (величина, являющаяся аналогом коэффициента детерминации в статистике).

Критерий T рассчитывается по формуле:

T = 1 - $\frac{W}{V}$, где W – внутриклассовый разброс, V – общее рассеивание.

Листинг программы с вычислением внутриклассового разброса, общего рассеивания и критерия T:

```
for metod=1:4
if metod==1
    TU=idkm;
end;
if metod==2
    TU=Tsr;
end;
if metod==3
    TU=Tb;
end;
if metod==4
    TU=Tp;
end;
BBs=[];
srvb=[];
for i=1:length(X)
    srvb=[srvb sum(X(i,:))];
end;
srvb=sum(srvb)/length(X);
for i=1:cl
    n=find(TU==i);
    sr=[];
        for j=1:length(n)
            sr=[sr sum(X(n(j),:))];
        end;
        sr=sum(sr);
        sr=sr/length(n);
        s=[sr;srvb];
    BBs=[BBs length(n)*(pdist(s))^2];
end;
BB=sum(BBs);
L=[];
for i=1:length(X)
    ss=[sum(X(i,:));srvb];
    L=[L pdist(ss)^2];
end;
V=sum(L);
B=BB;
W=V-B;
T=1-(W/V);
Coef=[Coef T];
end;
```

Задача сводится к тому, чтобы достичь максимального T при наименьшем числе кластеров, данное количество кластеров будет являться потенциально оптимальным количеством кластеров.

Визуализируем результаты кластеризации эталонным методом k-средних:

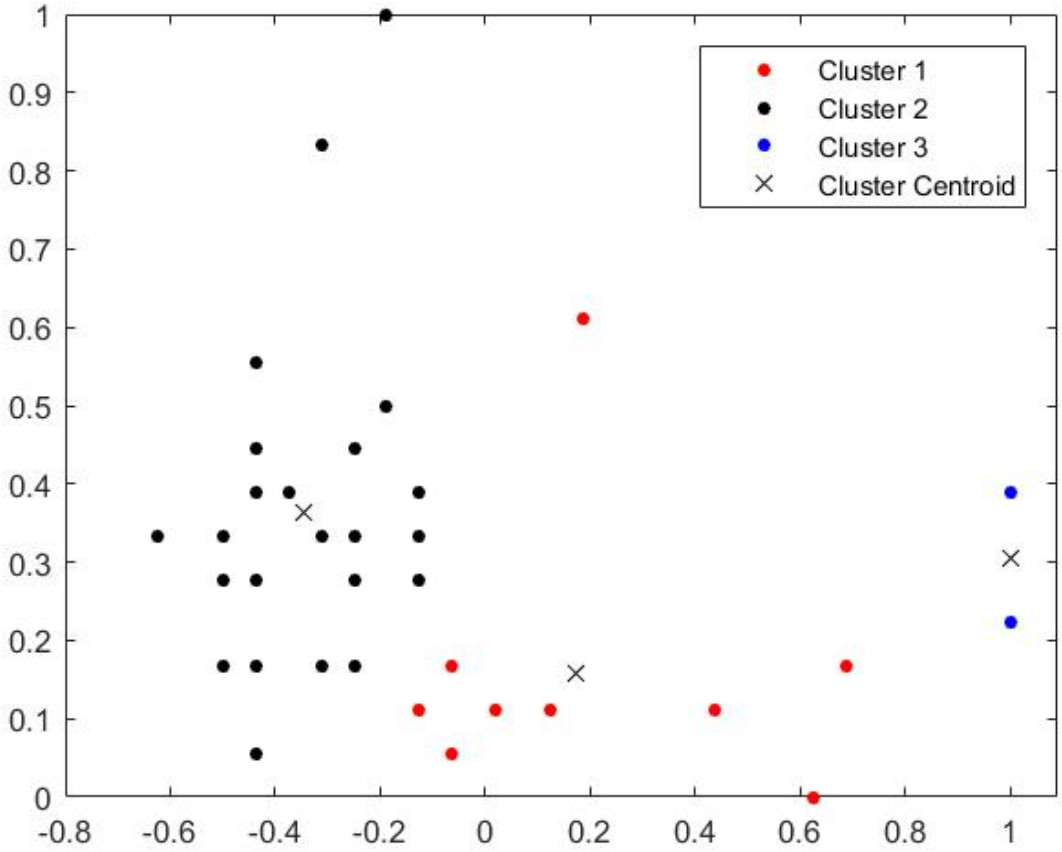

Рисунок 9 – Метод k-средних для 3 кластеров

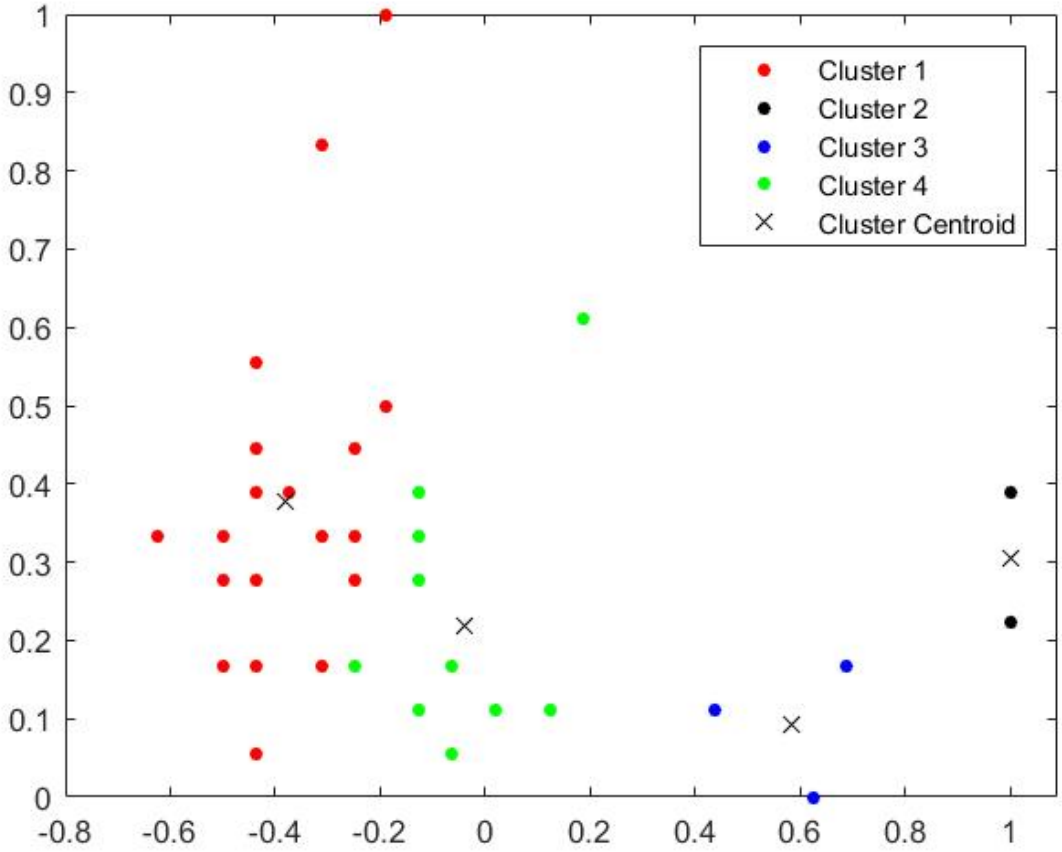

Рисунок 10 – Метод k-средних для 4 кластеров

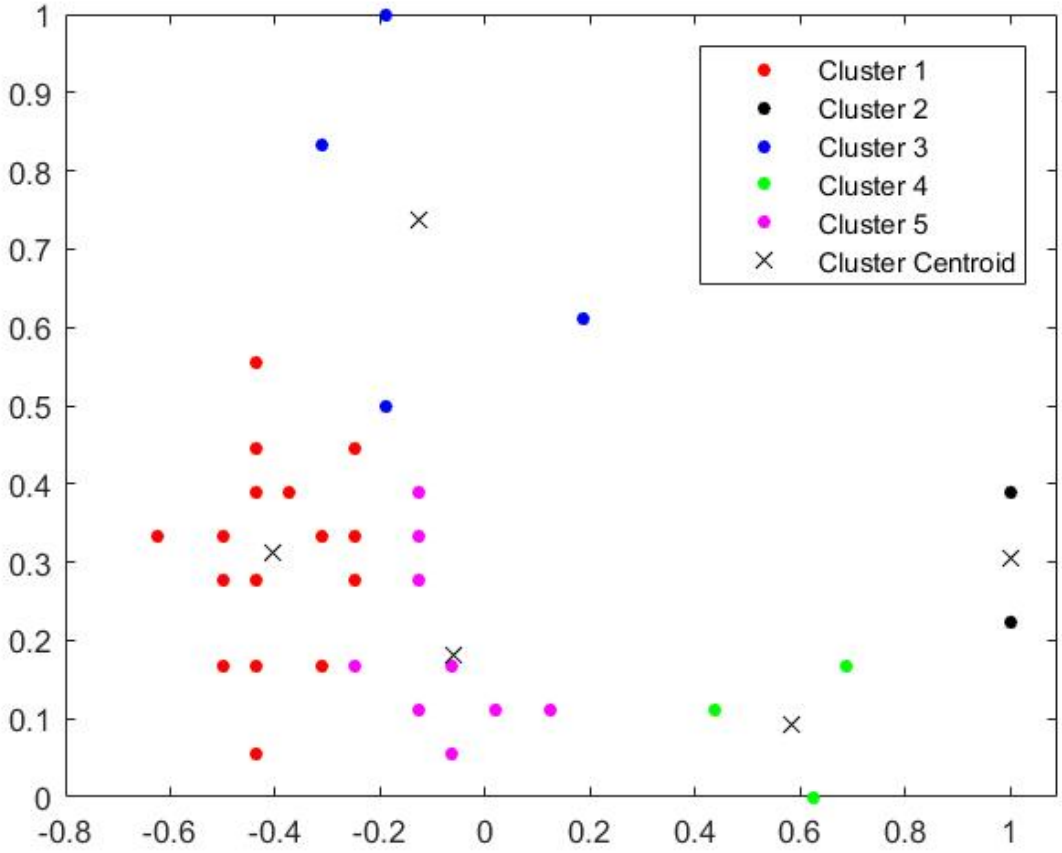

Рисунок 11 – Метод k-средних для 5 кластеров

Исходя из рисунков 9, 10, 11 видно, что начиная уже с k=3, г. Москва (№ 18) выделяется в отдельный кластер, а при k=4 Московская область также идет в отдельный кластер.

Уберем маргинальные регионы: г. Москва, Московская область, Республика Ингушетия, Чеченская Республика и проведем анализ и кластеризацию еще раз.

Результаты:

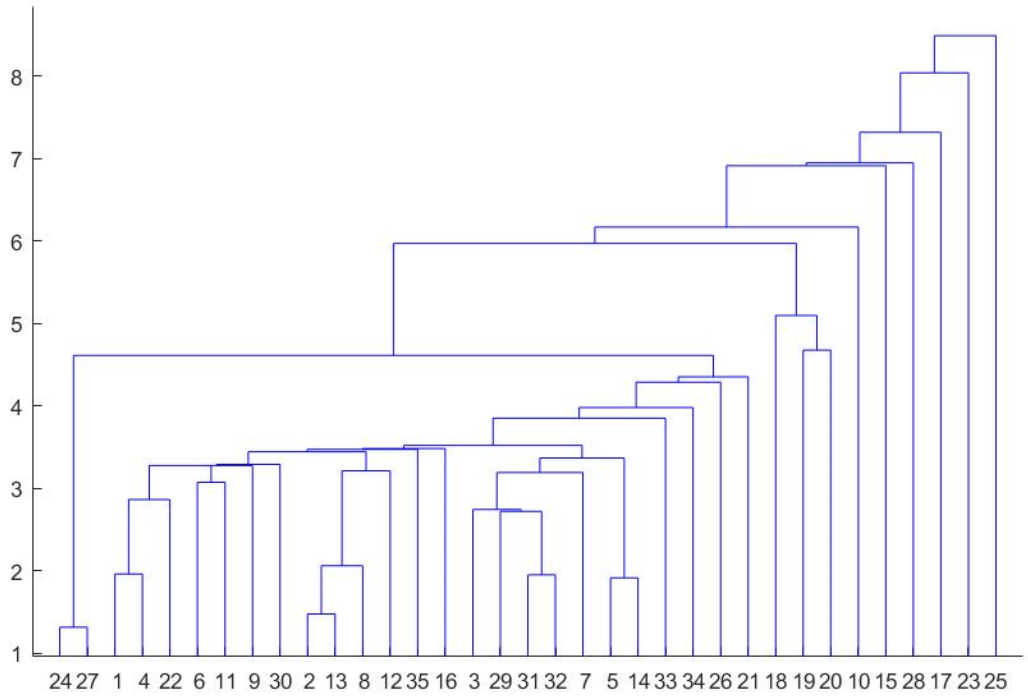

Рисунок 12 – Дендрограмма метода ближнего соседа (без маргинальных регионов)

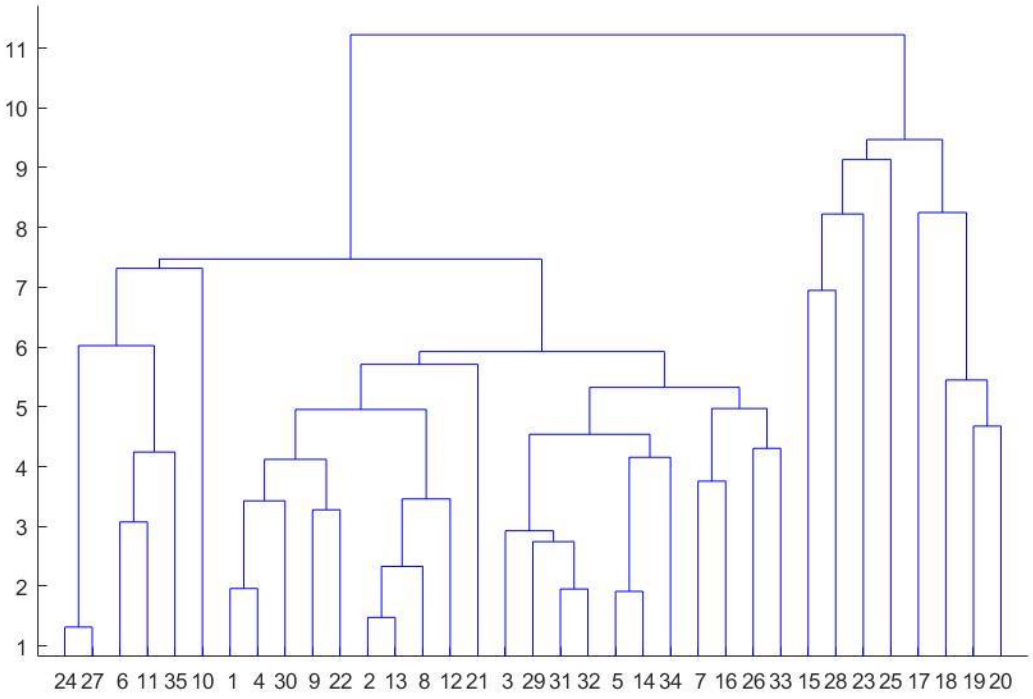

Рисунок 13 – Дендрограмма метода средней связи (без маргинальных регионов)

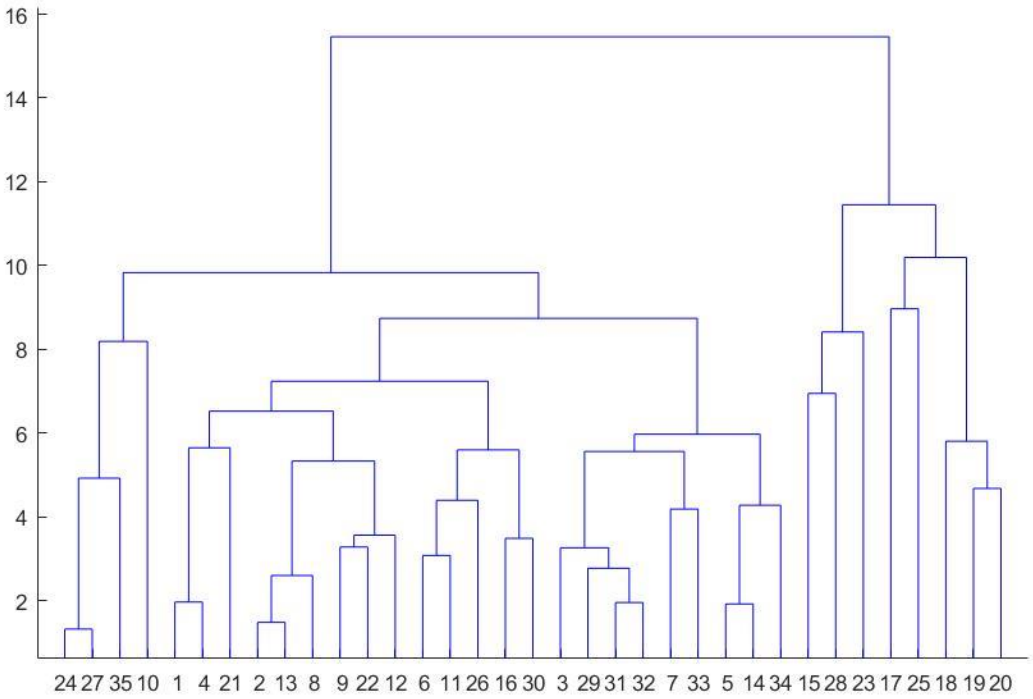

Рисунок 14 – Дендрограмма метода полной связи (без маргинальных регионов)

Из рисунка 12 видно, что метод ближнего соседа работает плохо, 25, 23, 17, 28 регионы относятся каждый к отдельному кластеру. Судя по 13 рисунку неплохой будет кластеризация на 2 или 3 кластера, но если сравнить

получаемые результаты с исходными данными, то понятно, что как минимум, Республика Дагестан и Карачаево-Черкесская Республика не могут оказаться в одном кластере, так как по исходным данным они имеют огромные различия по первому и третьему показателю, потому этот метод нас тоже не устраивает. Метод полной связи похож на метод средней связи, так же можно было бы поделить регионы по данной дендрограмме на 2-3 кластера, но возникают те же проблемы, что и в методе средней связи, в один кластер попадают совершенно различные по показателям регионы.

Приведем результаты вычисления критерия Т для всех методов при различном числе кластеров.

Таблица 3 – критерий Т

Кол-во кластеров	Метод k-средних	Метод ближнего соседа	Метод средней связи	Метод полной связи
1	0	0	0	0
2	0,483675861	0,047404395	0,598834317	0,598834317
3	0,740779692	0,139308764	0,60637988	0,603866656
4	0,786284736	0,411762493	0,60641325	0,635157123
5	0,842175027	0,485204307	0,67549844	0,659452485
6	0,834307392	0,504354097	0,686191593	0,699793801
7	0,903116941	0,539773485	0,692570543	0,756196395
8	0,91291578	0,721937342	0,712740707	0,766889548
9	0,933633989	0,723742776	0,722429766	0,7768596
10	0,931546045	0,723837153	0,724755916	0,790631417
11	0,957441531	0,72654246	0,738233358	0,800320476
12	0,972056854	0,735224855	0,742995741	0,83463252
13	0,964322715	0,752075566	0,744801174	0,852098607
14	0,956676458	0,757601122	0,84119804	0,85390404
15	0,965827045	0,781291321	0,845985914	0,854283409
16	0,978610489	0,833987118	0,927046524	0,855343456
17	0,978667079	0,845483291	0,927140901	0,885297886
18	0,980005334	0,855362492	0,93038308	0,949112255
19	0,987048699	0,933613423	0,930602268	0,950265387
20	0,976111819	0,942383386	0,941597817	0,950359764
21	0,988443545	0,942465965	0,943790016	0,954053427
22	0,991552983	0,945707365	0,943790362	0,956245626
23	0,993920694	0,954357474	0,951634373	0,96636872
24	0,991211024	0,972946863	0,970223762	0,969944409
25	0,99448916	0,97790704	0,97107744	0,97107744
26	0,997654605	0,988223164	0,979935448	0,979935448
27	0,996350101	0,990251572	0,990251572	0,98104725
28	0,997211433	0,991363374	0,991363374	0,991363374
29	0,998113705	0,991433004	0,991433004	0,991433004
30	0,998526388	0,991460362	0,991460362	0,991460362
31	0,999193317	0,99311228	0,99311228	0,99311228
32	0,998151636	0,996280747	0,996280747	0,996280747
33	0,999100583	0,999225815	0,999225815	0,999225815
34	0,999650982	0,999664633	0,999664633	0,999664633
35	1	1	1	1

В качестве оптимального метода кластеризации будем использовать эталонный метод k-средних, так как данный метод дает наилучшие результаты при вычислении критерия Т, и в целом, как было проанализировано ранее по рисункам 12, 13, 14, остальные методы работают плохо и постоянно выделяют субъекты в отдельные единичные кластеры.

Результаты кластеризации методом k-средних при 3, 4, 5 кластерах.

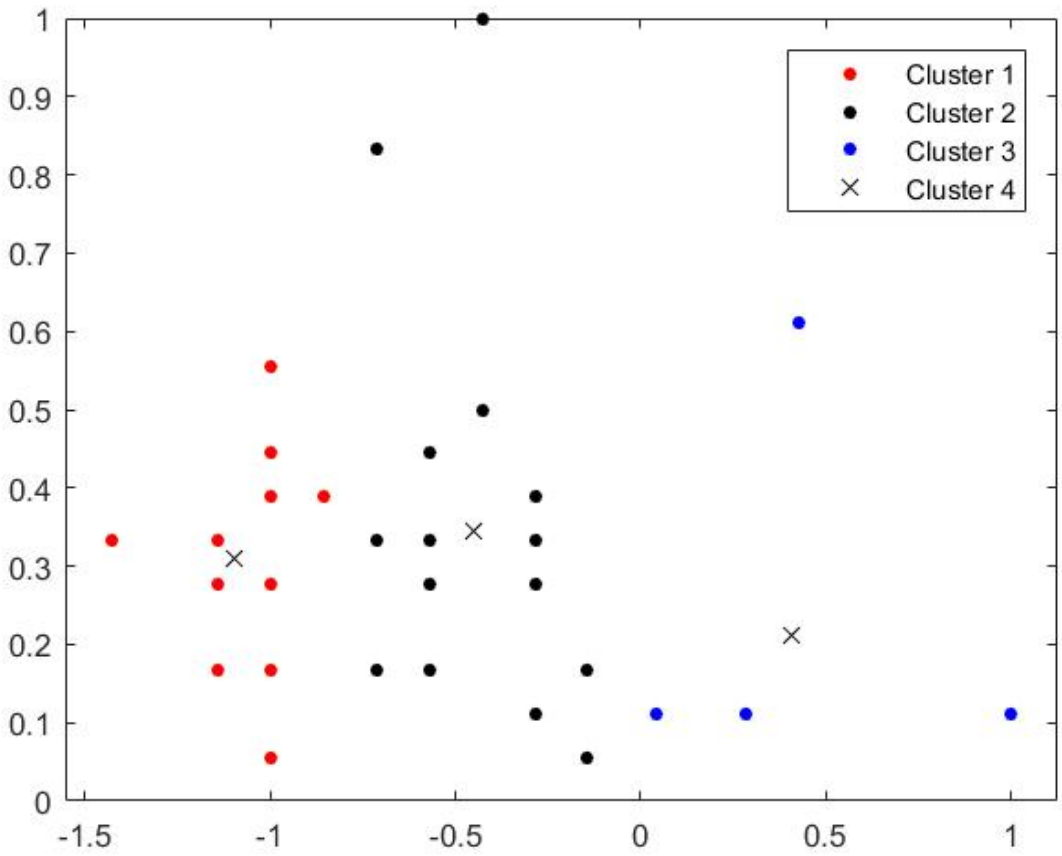

Рисунок 15 – метод k-средних (3 кластера)

Кластеризация при 3 кластерах:

Кластер 1 – Брянская область, Владимирская область, Ивановская область, Костромская область, Курская область, Орловская область, Тамбовская область, Тверская область, Кировская область, Оренбургская область, Пензенская область, Саратовская область.

Кластер 2 - Белгородская область, Воронежская область, Калужская область, Липецкая область, Рязанская область, Смоленская область, Тульская область, Ярославская область, Республика Северная Осетия – Алания, Ставропольский край, Республика Башкортостан, Республика Мордовия, Удмуртская Республика, Чувашская Республика, Пермский край, Нижегородская область, Самарская область, Ульяновская область.

Кластер 3 - Республика Дагестан, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Марий Эл, Республика Татарстан.

Разбиение на 3 кластера я считаю не очень удачным, как минимум по тому, что по исходным данным Республика Татарстан лидирует по многим показателям, в отличии от, например, Карачаево-Черкесской Республики, а метод k-средних при 3 кластерах относит их к одному, в остальном же кластеризация в целом приемлима.

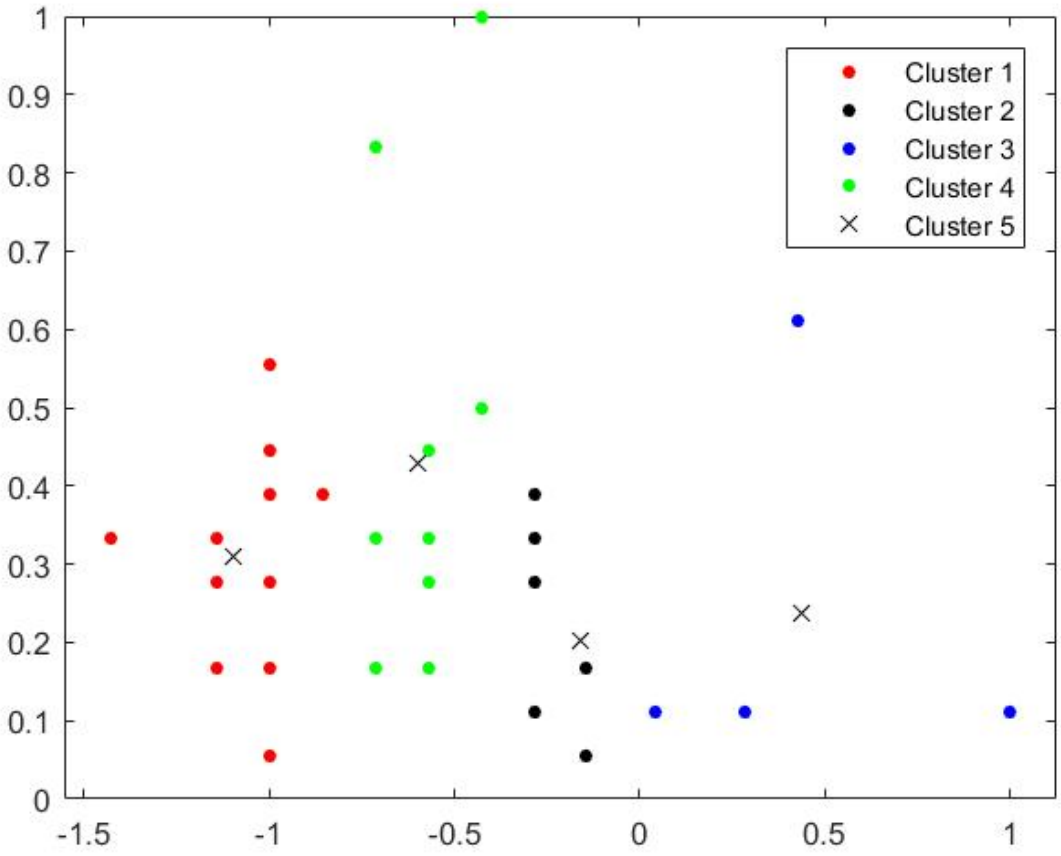

Рисунок 16 – метод k-средних (4 кластера)

Кластер 1 – Липецкая область, Рязанская область, Смоленская область, Тульская область, Ярославская область, Республика Мордовия, Чувашская Республика, Пермский край, Нижегородская область, Самарская область, Ульяновская область.

Кластер 2 - Белгородская область, Воронежская область, Калужская область, Республика Северная Осетия – Алания, Ставропольский Край, Республика Башкортостан, Республика Татарстан, Удмуртская Республика.

Кластер 3 - Республика Дагестан, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Марий Эл.

Кластер 4 - Брянская область, Владимирская область, Ивановская область, Костромская область, Курская область, Орловская область, Тамбовская область, Тверская область, Кировская область, Оренбургская область, Пензенская область, Саратовская область.

Данная кластеризация по моему мнению является наиболее оптимальной, так как во многом ее результаты согласовываются с проведенным анализом исходных данных.

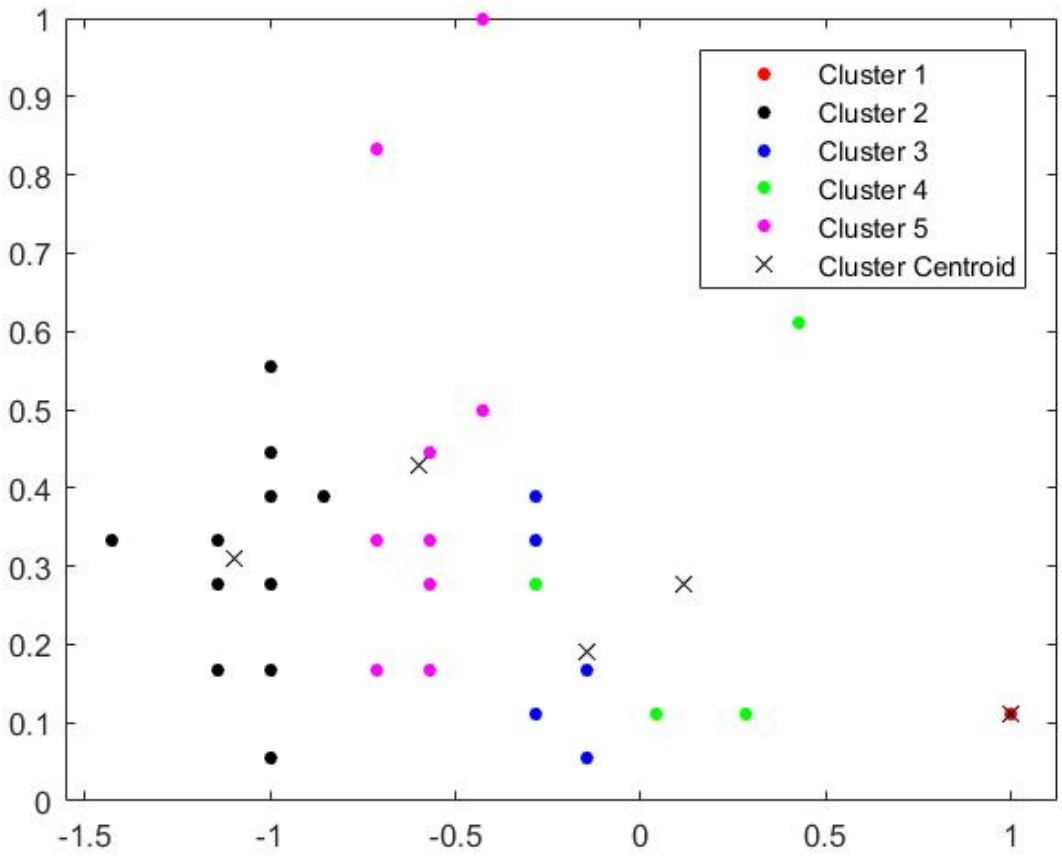

Рисунок 17 – метод k-средних (5 кластера)

Кластер 1 – Республика Дагестан.

Кластер 2 – Брянская область, Владимирская область, Ивановская область, Костромская область, Курская область, Орловская область, Тамбовская область, Тверская область, Кировская область, Оренбургская область, Пензенская область, Саратовская область.

Кластер 3 – Белгородская область, Воронежская область, Калужская область, Ставропольский край, Республика Башкортостан, Республика Татарстан, Удмуртская Республика.

Кластер 4 – Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Северная Осетия – Алания, Республика Марий Эл.

Кластер 5 – Липецкая область, Рязанская область, Смоленская область, Тульская область, Ярославская область, Республика Мордовия, Чувашская Республика, Пермский край, Нижегородская область, Самарская область, Ульяновская область.

Оптимальным количеством кластеров будем считать 4 кластера, так как при разбиение на 5 кластеров, один субъект выделяется в отдельный класс. Анализ

исходных данных также показывает, что такая кластеризация уместна. При взгляде на рисунки 15, 16, 17 можно увидеть, что разбиение на 4 кластера является оптимальным (1 кластер – с Орловской области до Республики Марий Эл включительно, 2 кластер – от Ивановской области до Тамбовской области включительно, 3 кластер – от Смоленской области до Ставропольского края включительно и 4 кластер – от Республики Башкортостан до Республики Татарстан). Вертикальная ось на рисунке 15 представляет собой квадрат суммы нормированных показателей по каждому региону (квадрат для того, чтобы проще было увидеть разницу между регионами).

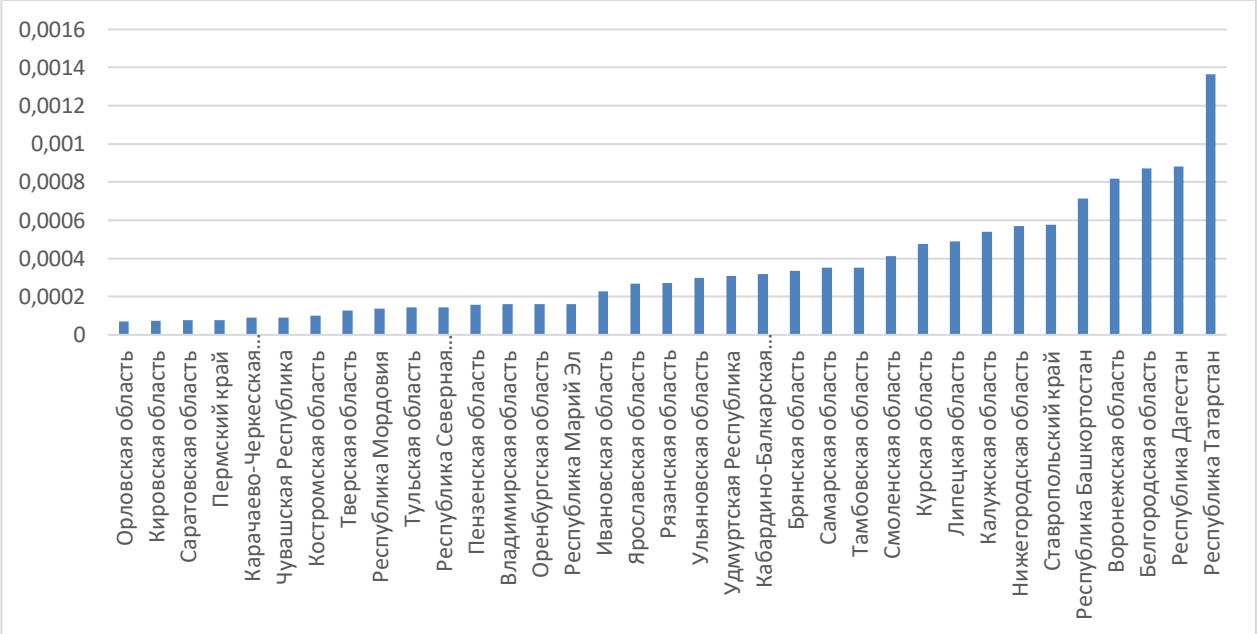

Рисунок 18 – диаграмма по суммированному показателю

Итог кластеризации методом k-средних:

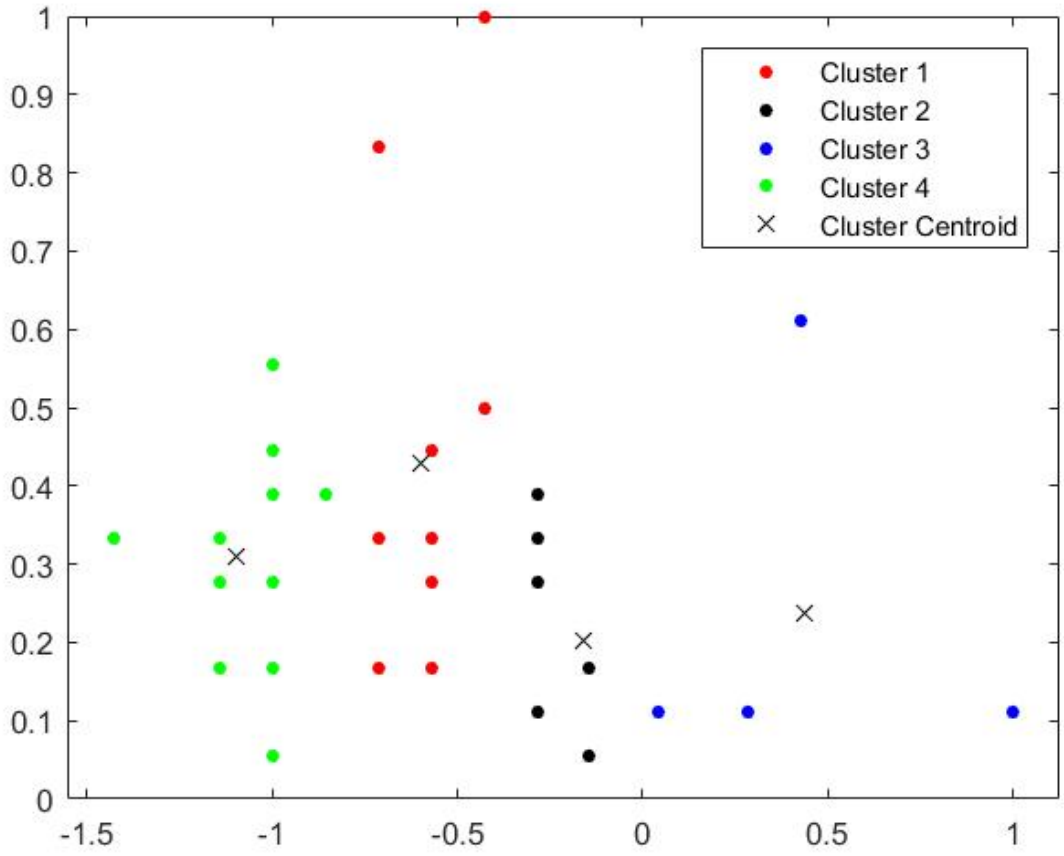

Рисунок 19 – Итоговая кластеризация

Кластер 1 – Липецкая область, Рязанская область, Смоленская область, Тульская область, Ярославская область, Республика Мордовия, Чувашская Республика, Пермский край, Нижегородская область, Самарская область, Ульяновская область.

Кластер 2 - Белгородская область, Воронежская область, Калужская область, Республика Северная Осетия – Алания, Ставропольский Край, Республика Башкортостан, Республика Татарстан, Удмуртская Республика.

Кластер 3 - Республика Дагестан, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Марий Эл.

Кластер 4 - Брянская область, Владимирская область, Ивановская область, Костромская область, Курская область, Орловская область, Тамбовская область, Тверская область, Кировская область, Оренбургская область, Пензенская область, Саратовская область.

ВЫВОД

В результате проделанной работы были выполнены следующие действия: была поставлена задача, подобраны данные, влияющие на динамику экономического и социального развития регионов, данные были выбраны с корреляцией <0.6, был проведен анализ исходных данных, дендрограмм, вычислен критерий T и совокупностью эмпирических и практических наблюдений и вычислений был определен наилучший метод кластеризации и оптимальное число кластеров. Подводя итоги можно сделать вывод о результатах кластеризации:

Кластер 1 – Липецкая область, Рязанская область, Смоленская область, Тульская область, Ярославская область, Республика Мордовия, Чувашская Республика, Пермский край, Нижегородская область, Самарская область, Ульяновская область. В этом кластере динамика развития регионов выше среднего.

Кластер 2 - Белгородская область, Воронежская область, Калужская область, Республика Северная Осетия – Алания, Ставропольский Край, Республика Башкортостан, Республика Татарстан, Удмуртская Республика. В данном кластере регионы лидируют по динамике экономического развития.

Кластер 3 - Республика Дагестан, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Марий Эл. В этом кластере регионы имеют довольно низкую динамику развития.

Кластер 4 - Брянская область, Владимирская область, Ивановская область, Костромская область, Курская область, Орловская область, Тамбовская область, Тверская область, Кировская область, Оренбургская область, Пензенская область, Саратовская область. В этом кластере можно говорить, что динамика развития здесь ниже среднего.