

Министерство науки и высшего образования Российской Федерации
Севастопольский государственный университет
Кафедра ИС

Отчет
по лабораторной работе №4
«Кластерный анализ. Основные этапы и задачи кластерного анализа данных»
по дисциплине
«ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

Выполнил студент группы ИС/б-17-2-о
Горбенко К. Н.
Проверил
Сырых О.А.

Севастополь
2020

1 ЦЕЛЬ РАБОТЫ

- закрепить теоретические знания и приобрести практические навыки в проведении кластерного анализа по экспериментальным данным;
- исследовать возможности языка R для проведения кластерного анализа.

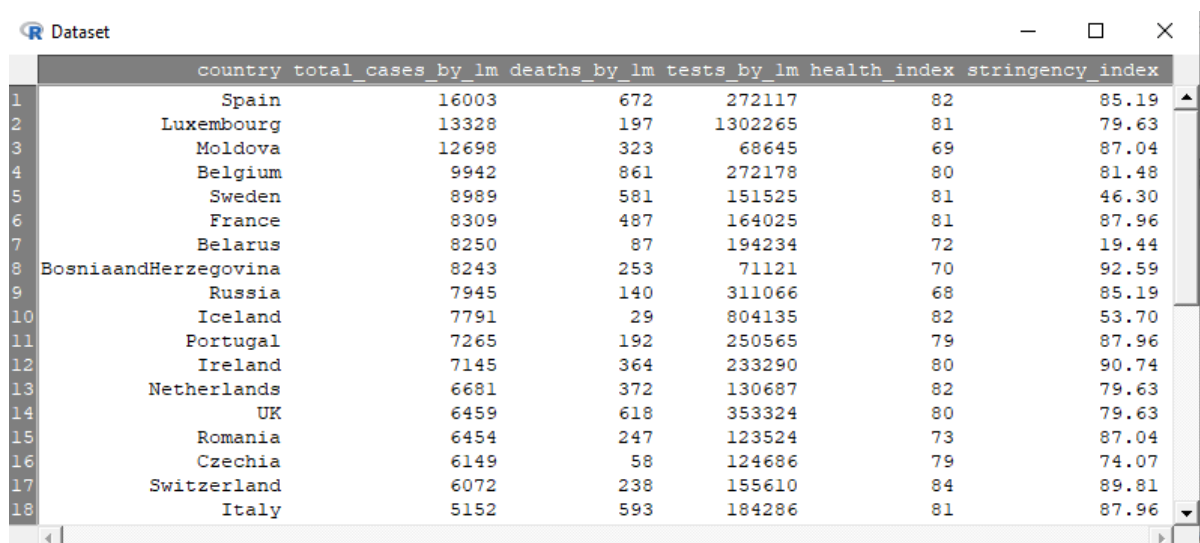
2 ЗАДАНИЕ НА РАБОТУ

1. Создать файл с исходными данными.
2. Провести кластерный анализ экспериментальных данных.
3. Проведя процедуру кластеризации несколько раз при различных значениях числа кластеров (от 2-х до 10 кластеров), необходимо выбрать лучшую группировку в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний. Полученные результаты оформите в виде таблицы. Изобразить графически значения данного показателя качества классификации. Для этого построить диаграмму, на которой по оси X – количество кластеров, по оси Y – значения показателя J.

3 ХОД РАБОТЫ

3.1 Кластерный анализ

Выполним иерархический кластерный анализ. Загрузим данные о количестве заражений коронавирусом 1.



	country	total_cases_by_lm	deaths_by_lm	tests_by_lm	health_index	stringency_index
1	Spain	16003	672	272117	82	85.19
2	Luxembourg	13328	197	1302265	81	79.63
3	Moldova	12698	323	68645	69	87.04
4	Belgium	9942	861	272178	80	81.48
5	Sweden	8989	581	151525	81	46.30
6	France	8309	487	164025	81	87.96
7	Belarus	8250	87	194234	72	19.44
8	BosniaandHerzegovina	8243	253	71121	70	92.59
9	Russia	7945	140	311066	68	85.19
10	Iceland	7791	29	804135	82	53.70
11	Portugal	7265	192	250565	79	87.96
12	Ireland	7145	364	233290	80	90.74
13	Netherlands	6681	372	130687	82	79.63
14	UK	6459	618	353324	80	79.63
15	Romania	6454	247	123524	73	87.04
16	Czechia	6149	58	124686	79	74.07
17	Switzerland	6072	238	155610	84	89.81
18	Italy	5152	593	184286	81	87.96

Рисунок 1 – Загруженные данные

- total_cases_by_1m - количество заражений на 1 миллион населения;
- total_death_by_1m - количество смертей на 1 миллион населения;
- tests_by_1m - количество тестов на 1 миллион населения;
- health_index - индекс развитости системы здравоохранения;
- stringency_index - индекс реакции властей на пандемию.

Функция кластерного анализа в R:

```
kmeans(x, centers, iter.max=10, nstart=1,
algorithm=c("Hartigan-Wong" "Lloyd" "Forgy" "MacQueen"))
```

Выполним разбиение на два кластера по переменным «Индекс развитости системы здравоохранения» и «Индекс реакции властей на пандемию». Результат разбиения изображен на рисунке 2.

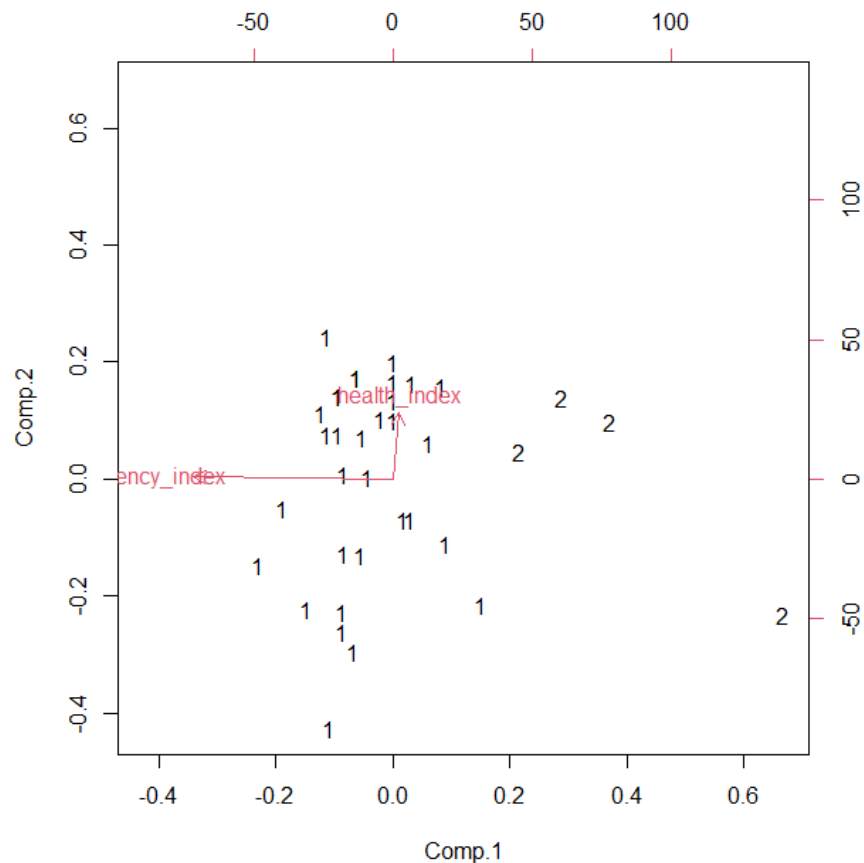


Рисунок 2 – Разбиение данных на 2 кластера

Результаты разбиения представлены на рисунке 3.

```
Rcmdr> .cluster <- KMeans(model.matrix(~-1 + health_index + stringency_index,
Rcmdr+   Dataset), centers = 2, iter.max = 10, num.seeds = 10)

Rcmdr> .cluster$size # cluster sizes
[1] 32 4

Rcmdr> .cluster$centers # cluster centroids
      new.x.health_index new.x.stringency_index
1          76.84375          83.82531
2          78.50000          44.90750

Rcmdr> .cluster$withinss # within cluster sum of squares
[1] 2487.778 1022.395

Rcmdr> .cluster$tot.withinss # Total within sum of squares
[1] 3510.174

Rcmdr> .cluster$betweenss # Between cluster sum of squares
[1] 5394.984

Rcmdr> biplot(princomp(model.matrix(~-1 + health_index + stringency_index,
Rcmdr+   Dataset)), xlabs = as.character(.cluster$cluster))

Rcmdr> remove(.cluster)
```

Рисунок 3 – Результаты разбиения

- первый кластер содержит 32 элемента, второй – 4;
- сумма квадратов расстояний внутри кластера: 1 – 2487.778, 2 – 1022.395;
- общая сумма квадратов расстояний внутри кластеров: 3510.174;
- сумма квадратов расстояний между кластерами – 5394.984.

Для выбора лучшей группировки в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний было проведено деление на 2 – 10 кластеров и заполнена таблица в MS Excel.

m	Dii	Dij	J1	J2	J
2	3510.174	5394.984	1755.087	2697.492	0.650637
3	2265.939	6639.219	755.313	2213.073	0.341296
4	1418.837	7486.32	354.70925	1871.58	0.189524
5	922.531	7982.627	184.5062	1596.5254	0.115567
6	690.357	8214.801	115.0595	1369.1335	0.084038
7	508.062	8397.096	72.58028571	1199.58514	0.060504
8	411.28	8493.878	51.41	1061.73475	0.048421
9	329.784	8575.373	36.64266667	952.819222	0.038457
10	306.693	8598.465	30.6693	859.8465	0.035668

Рисунок 4 – Расчет численного показателя меры качества классификации

Значения данного показателя качества классификации представлено графически на рис 4. Для этого построена диаграмма, на которой по оси X – количество кластеров, по оси Y – значения показателя J.

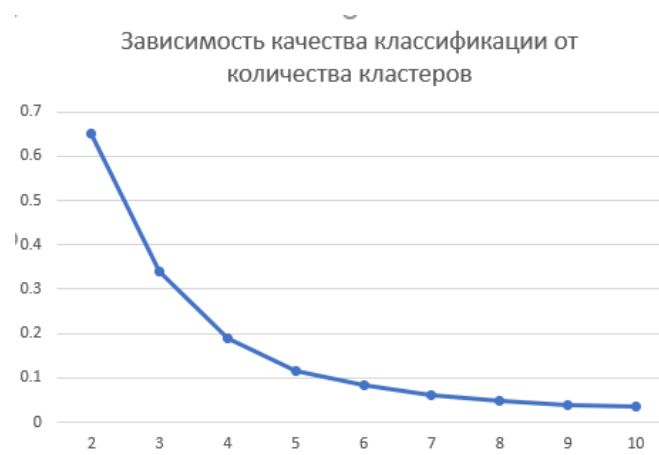


Рисунок 5 – Диаграмма численной меры качества классификации

В соответствии с этим критерием оптимальным разбиением экспериментальных данных является разбиение на 3 кластера.

3.2 Иерархический анализ

Был проведен иерархический анализ методом Уорда:

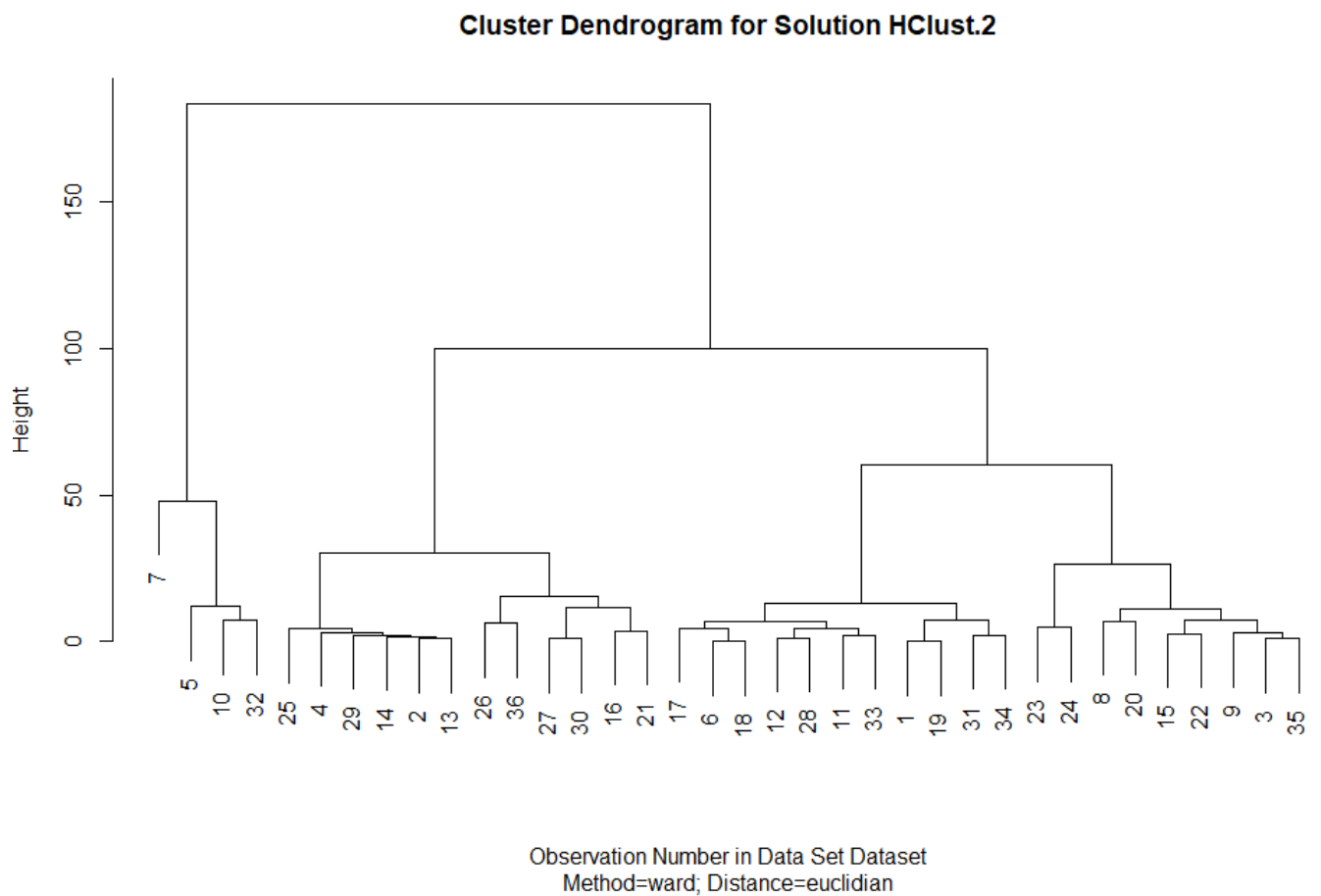


Рисунок 6 – Иерархический анализ методом Уорда

Результатом анализа является количество элементов в каждом из кластеров:

```
1 Rcmdr> summary(as.factor(cutree(HClust.3, k = 3))) # Cluster Sizes
2 1 2 3
3 20 12 4
```

Был проведен иерархический анализ методом простой связи:

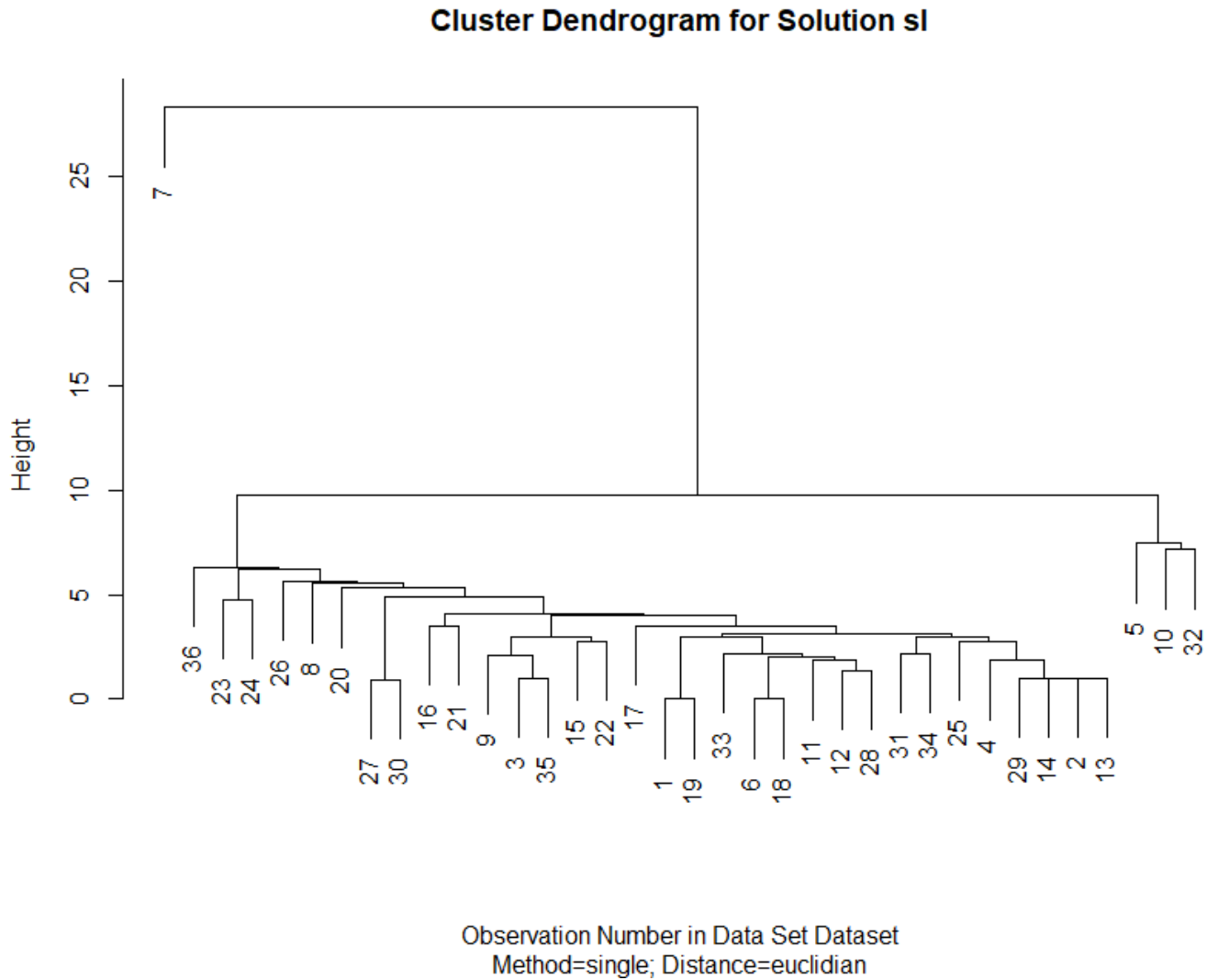


Рисунок 7 – Иерархический анализ методом простой связи

Результатом анализа является количество элементов в каждом из кластеров:

```
1 Rcmdr> summary(as.factor(cutree(sl, k = 3))) # Cluster Sizes
2 1 2 3
3 32 3 1
```

ВЫВОДЫ

В ходе лабораторной работы провели многомерный анализ данных. Для этого использовали кластеризацию, которая предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек". Целью кластеризации является поиск существующих структур. Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной. Провели иерархический анализ с использованием k-средних. Данный алгоритм является наиболее распространённым.