

$$\min \{ h : \sum_{k=0}^{h_{\text{opt}}} \frac{(T\lambda_1)^k}{k!} > e^{T\lambda_1} (1 - \alpha) \},$$

$$\beta = \sum_{k=0}^{h_{\text{opt}}} \frac{(T\lambda_2)^k}{k!} e^{-T\lambda_2}.$$

4. ЗАКЛЮЧЕНИЕ

Используя результаты теоремы можно вычислить оптимальный порог и оценить качество алгоритма для заданных параметров распределений, которые выбираются исходя из конкретной задачи.

Также, для экспоненциального распределения и распределения Эрланга легко получить, если имеется в том необходимость, зависимость какого-либо параметра алгоритма от другого (например T от α) при заданном качестве обнаружения без дополнительного имитационного моделирования. Для чего можно использовать, например, Microsoft Excel, где реализованы функции вычисления гамма-функции обратной к гамма-функции.

ЛИТЕРАТУРА

- [1]. Медведовский И.Д., Семьянов П.В., Платонов В.В. Атака через Internet. Санкт-Петербург: НПО "Мир и семья-95", 1997.

- [2]. Xiaobing Zhang, S.Felix Wu, Zhi Fu, Tsung-Li Wu. Malicious Packet Dropping: How It Might Impact the TCP Performance and How We Can Detect It. IEEE Network Protocols, Proceeding International Conference, 2000, p.263-272.
- [3]. Башарин Г.П., Бочаров П.П., Коган Я.А. Анализ очередей в вычислительных сетях. М.: Наука, 1989.
- [4]. Бертсекас Д., Галагер Р. Сети передачи данных. М.: Мир, 1989.
- [5]. Page E.S. Continuous inspection schemes. Biometrika, 1954, v.41, N2, p.100-114.
- [6]. Page E.S. A test for a change in parameter occurring at an unknown point. Biometrika, 1955, v.42, N4, p.523-527.
- [7]. Page E.S. On problems in which a change in parameter occurs at an unknown point. Biometrika, 1957, v.44, N2, p.248-252.
- [8]. Вальд А. Последовательный анализ. М.: Физматгиз, 1960, с. 328.
- [9]. Rodionov A.S., Shakhov V.V. About point-of-change problem in the information systems with exponential distribution. Proceeding of Digital Computer Communication Network Conference, Tel-Aviv (Israel), 1999, p.141 - 145.
- [10]. Никифоров И.В. Последовательное обнаружение изменения свойств временных рядов. М.: Наука, 1983.
- [11]. Клигене Н., Телькснис Л. Методы обнаружения моментов изменения свойств случайных процессов. АИТ, 1983, н.10, с.5-56.
- [12]. Shakhov V.V. Simulation optimisation problem. Bull. NCC, Special Issue, 1999.

МНОГОМЕРНАЯ КЛАССИФИКАЦИЯ НА ОСНОВЕ МЕТРИКИ КОЛМОГорова — СМирнова

А.А.Кулешов¹, В.Н.Тамашевич²

¹ - Кафедра уравнений математической физики, механико-математический факультет, Белорусский государственный университет, просп. Ф. Скорины, 4, Минск, 220050, БЕЛАРУСЬ, тел. 209-50-47. kuleshov@cit.bsu.unibel.by.

² - Научно-исследовательский институт статистики при Министерстве статистики и анализа Республики Беларусь, Партизанский просп., 12-а, Минск, 220070, БЕЛАРУСЬ, тел. 249-34-13. niistat@postmaster.co.uk.

В настоящей публикации обсуждаются возможности применения и основные принципы работы программы кластерного анализа, созданной на базе компьютерной технической системы *Mathematica* и MS Excel. Математические алгоритмы, положенные в основу программы, в республике реализованы впервые. Описываемая программа не имеет аналогов ни в одном из известных коммерческих статистических пакетов.

Большинство объектов, изучаемых в социальных и экономических науках, может быть охарактеризовано кибернетическим понятием «сложная система». Одна из трудностей исследований в этих областях состоит в том, что почти не существует объектов, которые можно было бы рассматривать как отдельные (внесистемные) элементы. Так, экономика объединяет огромное число элементов, отличается многообразием внутренних связей и связей с другими системами (природной средой, экономической

деятельностью других субъектов, социальными отношениями и т.д.). В промышленности и сельском хозяйстве взаимодействуют природные, технологические, социальные процессы, объективные и субъективные факторы.

Одними из наиболее важных и перспективных в сфере экономики стали методы *многомерной классификации*, в частности методы *кластерного анализа* [1]. Это эффективное средство статистической обработки данных, позволяющее решать следующие задачи:

- изучение структуры совокупности с целью выделения группы объектов, схожих между собой по нескольким признакам;
- проверка выдвигаемых предположений о наличии структуры в изучаемой совокупности, а также построение новых классификаций;
- выявление групп схожих ситуаций, в которых может приниматься одно и то же решение;
- сопоставление ситуации принятия решения с ранее встречавшимися ситуациями.

В многомерном анализе образовались самостоятельные разделы, которые, однако, не изолированы, а проникают и переходят из одного в другой. Это кластерный анализ, таксономия, распознавание образов, метод главных компонент, факторный анализ. Наиболее ярко отражает черты многомерного анализа в классификации объектов кластерный анализ, а в исследовании связей – факторный анализ. В кластерном анализе группировочные признаки подвергаются объединению с помощью некоторой метрики – евклидова расстояния или иной. Чаще всего в кластерном анализе используются такие меры расстояния между объектами: евклидово расстояние, взвешенное евклидово расстояние, расстояние city-block, расстояние Минковского, расстояние Махаланобиса.

В упомянутой выше программе использован иной подход к объединению группировочных признаков. Он основан на интерпретации признаков, описывающих исследуемый объект, как статистической выборки. В этом случае объект погружается в бесконечномерное пространство функций распределения. Сами функции распределения строятся следующим образом. По статистической выборке восстанавливается плотность распределения, затем функция распределения ставится в соответствие наблюдаемому объекту. После этого для сравнения объектов можно использовать то или иное расстояние между функциями. Алгоритм программы использует метрику Колмогорова —

Смирнова, т.е. максимальное значение модуля разности (с точностью до постоянного множителя, о котором будет сказано ниже) функций распределения.

Приведем характерные графики плотности распределения и соответствующей функции распределения для объектов экономической природы.

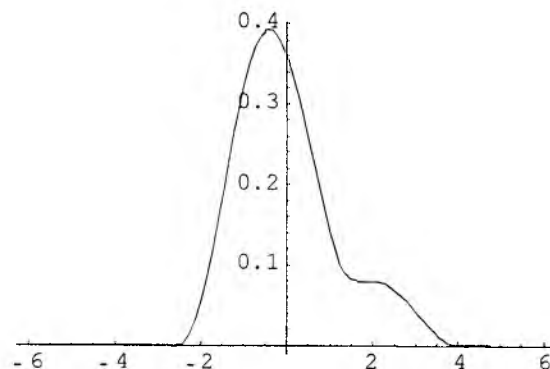


Рисунок 1. График плотности распределения

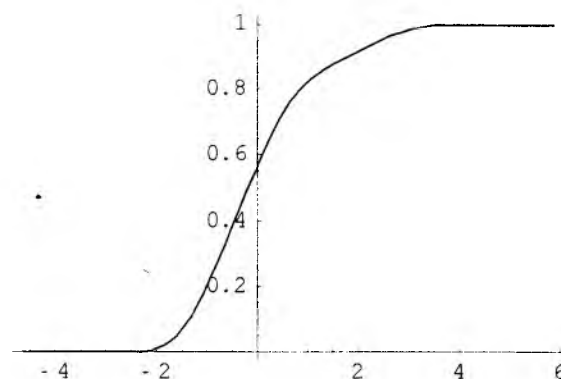


Рисунок 2. График функции распределения

В данной работе не обсуждаются преимущества использования той или иной метрики, поскольку это задача для экспертов предметной области. Тем не менее, представляется, что выбор бесконечномерного пространства для исследования свойств объектов, является преимуществом метрики Колмогорова — Смирнова перед евклидовой метрикой, поскольку в этом случае количество получаемых характеристик объектов существенно больше. Имеется и еще одно значительное отличие расстояния Колмогорова — Смирнова от евклидова. В теории вероятностей доказывается существование так называемого предельного распределения Колмогорова, которое позволяет проверить гипотезу о совпадении двух объектов с заданным

уровнем значимости и определить вероятность того, что расстояние между объектами найдено правильно. Ниже при рассмотрении примеров мы не останавливаемся на проверке гипотез. Заинтересованный читатель может найти подробное изложение техники проверки указанной выше гипотезы в [2].

К особенностям использованного алгоритма можно также отнести и то, что само количество измеряемых признаков у сравниваемых объектов может быть различным.

В качестве первой иллюстрации возможностей предложенного алгоритма рассмотрим пример.

Предположим, что нам даны шесть объектов, характеризующихся следующим набором числовых признаков.

Объект 1: {4.39, 3.6714, 1.50311, 3.98583, 1.98281, 5.96176, 2.74316, 6.02542, 3.68021, 4.63527}

Объект 2: {-1.61102, -1.05754, 0.612499, -1.29972, 0.243031, -2.8216, -0.342599, -2.87063, -1.06433, -1.79992}

Объект 3: {-5.21148, -3.89478, 0.0781538, -4.47092, -0.800791, -8.0914, -2.19398, -8.20804, -3.91094, -5.66088}

Объект 4: {-3.10114, -2.23179, 0.39135, -2.61218, -0.188974, -5.00261, -1.10883, -5.07962, -2.24245, -3.39785}

Объект 5: {2.77977, 2.4025, 1.26414, 2.56758, 1.51598, 3.60496, 1.91517, 3.63838, 2.40713, 2.90854}

Объект 6: {7.02437, 5.74734, 1.89408, 6.30612, 2.74654, 9.81753, 4.09776, 9.93066, 5.763, 7.46023}

Вычисляем расстояние между предложенными объектами по следующему алгоритму.

1. Нормируем исходные данные, т. е.

$$\text{объект}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \text{ где } \bar{x}_j \text{ — среднее}$$

арифметическое числовых признаков j -го объекта, а σ_j — дисперсия.

2. Рассчитываем матрицу расстояний, используя метрику Колмогорова — Смирнова между функциями распределения выборочных векторов (*объект*₁, *объект*₂, *объект*₃, *объект*₄, *объект*₅, *объект*₆), т. е. расстояние

$$(\text{объект}_i, \text{объект}_j) =$$

$$= \sqrt{\frac{m_i \cdot m_j}{m_i + m_j}} \max_x |F_i(x) - F_j(x)|, \quad (1)$$

где длина выборки m_i — количество признаков i -го объекта, $F_i(x)$ — функция распределения, построенная по выборке признаков i -го объекта (например, Рис.2).

Получаем следующие результаты.

$$\begin{aligned} d(1, 3) &= 0.03382155002031068 \\ d(1, 5) &= 2.482534153247273\text{E-}16 \\ d(1, 6) &= 7.447602459741819\text{E-}16 \\ d(2, 3) &= 7.447602459741819\text{E-}16 \\ d(3, 1) &= 0.03382155002031068 \\ d(3, 2) &= 7.447602459741819\text{E-}16 \\ d(3, 4) &= 2.482534153247273\text{E-}16 \\ d(4, 3) &= 2.482534153247273\text{E-}16 \\ d(5, 1) &= 2.482534153247273\text{E-}16 \\ d(6, 1) &= 7.447602459741819\text{E-}16 \end{aligned}$$

Интерпретировать полученные результаты

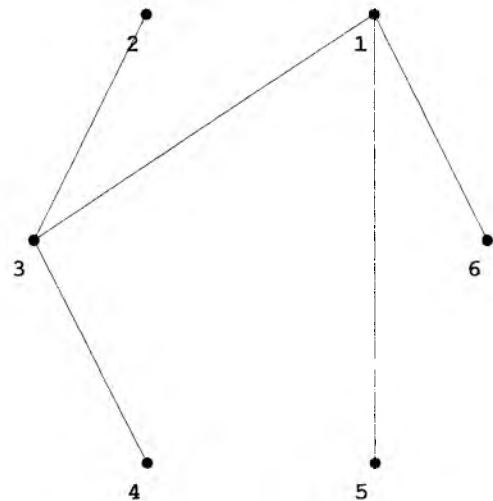


Рисунок 3. Граф

удобно с помощью графа (Рис.3), вершинами которого являются рассматриваемые объекты, а длинами ребер — вычисленные расстояния.

Самым длинным ребром графа является ребро (3; 1). Удаляя это ребро, получаем «грубую» классификацию исследуемых объектов: в одну группу попадают объекты с номерами 1, 6, 5, а в другую — 2, 3, 4.

Процедура кластеризации этих же данных, реализованная в известном пакете Statistica приводит к иному результату:

Объекты	Принадлежность кластеру	Расстояние до центра кластера
1	1	1,000
2	1	0,000
3	1	1,000
4	2	1,000
5	2	0,000
6	2	1,000

Рассмотрим пример из области экономики.

Предположим, что перед некоторой фирмой, для определения перспективных направлений инвестирования, стоит задача отбора группы стран с устойчивым экономическим положением и достаточно высоким уровнем развития.

Для решения этой задачи выделена совокупность из 17 стран, по каждой из которых имеются статистические данные по пяти признакам: X_1 — валовой внутренний продукт, в расчете на душу населения (дол. США); X_2 — среднемесячный уровень заработной платы (дол. США); X_3 — среднегодовой индекс потребительских цен (%); X_4 — уровень безработицы (в % к активному населению); X_5 — размер инвестиций в расчете на душу населения (дол. США):

	X_1	X_2	X_3	X_4	X_5
Германия	28870	2521,0	116,5	8,8	4167,5
Бельгия	26440	2233,6	111,6	9,7	3652,6
Дания	32100	3066,2	109,9	6,9	3496,8
Финляндия	23240	2251,1	108,0	15,4	2829,3
Франция	26270	1695,3	110,0	12,4	3282,2
Греция	11460	914,7	173,2	9,4	2606,2
Великобритания	19600	2026,9	114,4	8,2	2876,6
Ирландия	17110	1944,0	111,5	11,6	3251,6
Италия	19880	1649,0	124,6	12,0	3225,9
Люксембург	45360	2597,8	112,8	3,3	6250,0
Нидерланды	25940	2189,2	113,2	6,3	3812,5
Австрия	28110	2058,7	115,6	4,3	4862,9
Португалия	10160	545,6	131,1	7,3	2900,9
Швеция	25710	2930,3	112,7	10,0	2590,2
Испания	14350	1443,0	125,7	22,1	2844,4
США	28020	2484,4	115,2	5,4	4650,0
Япония	40940	3686,6	103,7	3,4	6476,0

По данным за 1996г., источник: Statistisches Jahrbuch 1998 für das Ausland.- Statistisches Bundesamt, Wiesbaden 1998

Группировка стран, одновременно по нескольким признакам, в нашем случае — пяти, возможна только с применением многомерных статистических методов. Обработку исходных данных произведем так же, как и в первом примере.

Используя метрику расстояния (1) получим матрицу расстояний 17×17 , в которой содержатся оценки попарных сравнений стран. Так как полная матрица расстояний — большого размера, воспроизведем ее отдельный фрагмент (Табл. 1):

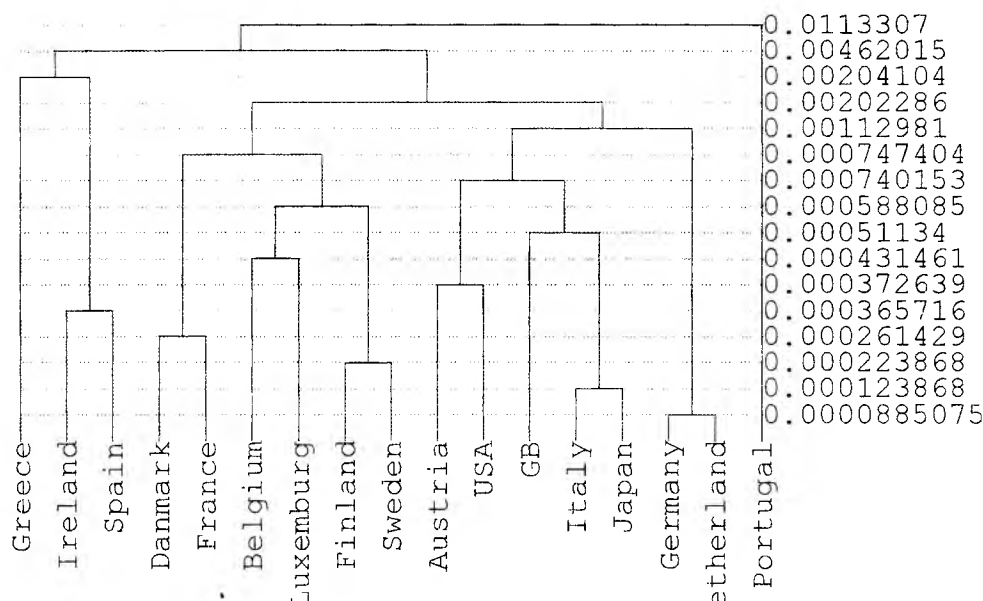
Таблица 1. Фрагмент матрицы расстояний, исчисленной с использованием метрики Колмогорова — Смирнова

	Германия	Бельгия	Дания	Финляндия	Франция	Греция
Германия	0	0,000474	0,001673	0,001055	0,001669	0,005711
Бельгия	0,000474	0	0,001199	0,000582	0,001195	0,006185
Дания	0,001673	0,001199	0	0,000618	0,000261	0,007383
Финляндия	0,001055	0,000582	0,000618	0	0,000619	0,006767
Франция	0,001669	0,001195	0,000261	0,000619	0	0,007376
Греция	0,005711	0,006185	0,007383	0,006767	0,007376	0

На основе полученной матрицы расстояний методом иерархического кластер-анализа с использованием K -расстояния (по Колмогорову) при $\tau = 1$:

$$\rho_{\tau}^{(K)}(S_l, S_m) = \sqrt{\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d^{\tau}(X_i, X_j)},$$

получим дендрограмму, демонстрирующую взаимосвязи стран и расстояния между ними в пяти-признаковом пространстве:



Дендрограмма позволяет легко выделить четыре группы стран:

I группа	II группа	III группа	IV группа
Бельгия	Великобритания	Греция	Португалия
Дания	Япония	Ирландия	
Финляндия	Австрия	Испания	
Франция	США		
Швеция	Италия		
Люксембург	Германия		
	Нидерланды		

Состав групп, даже в простом перечневом виде, позволяет видеть их однородность по уровню экономического развития. Португалия занимает особое место, имея специфические значения характерных признаков.

Построим аналитическую таблицу, где сведем среднегрупповые признаковые данные по каждой из четырех подсовкупностей стран (Табл. 2.):

Таблица 2. Многомерная группировка стран с использованием метрики Колмогорова — Смирнова

Кластер	Кол-во стран	Средне групповые значения признаков (X_{ij})				
		ВВП в расчете на душу населения, дол. США	Средне-саяная заработная плата, дол. США	Индекс потребительских цен, %	Уровень безработицы, %	Инвестиции в расчете на душу населения, дол. США
I	6	29853,3	2462,4	110,8	9,6	3683,5
II	7	27337,1	2373,7	114,7	6,9	4295,9
III	3	14306,7	1433,9	136,8	14,4	2900,7
IV	1	10160,0	545,6	131,1	7,3	2900,9
В среднем по совокупности стран	17	24915,3	2131,6	118,2	9,2	3751,5

Обобщенные данные, приведенные в Табл. 2, подтверждают ранее сделанные теоретические предположения, о разделении стран на четыре группы. При этом две первые группы стран имеют близкие характеристики и при снижении порогового уровня кластеризации, в данном случае выбираемого исследователем, они могли бы быть объединены.

В то же время, полученное распределение показывает, что различия среди стран первой и второй группы все-таки имеются. Во второй группе, очевидно более низкий средний уровень безработицы и более высокий поток инвестиций в расчете на душу населения. Следует полагать, что страны этой группы в большей мере нацелены на динамичное развитие в будущем.

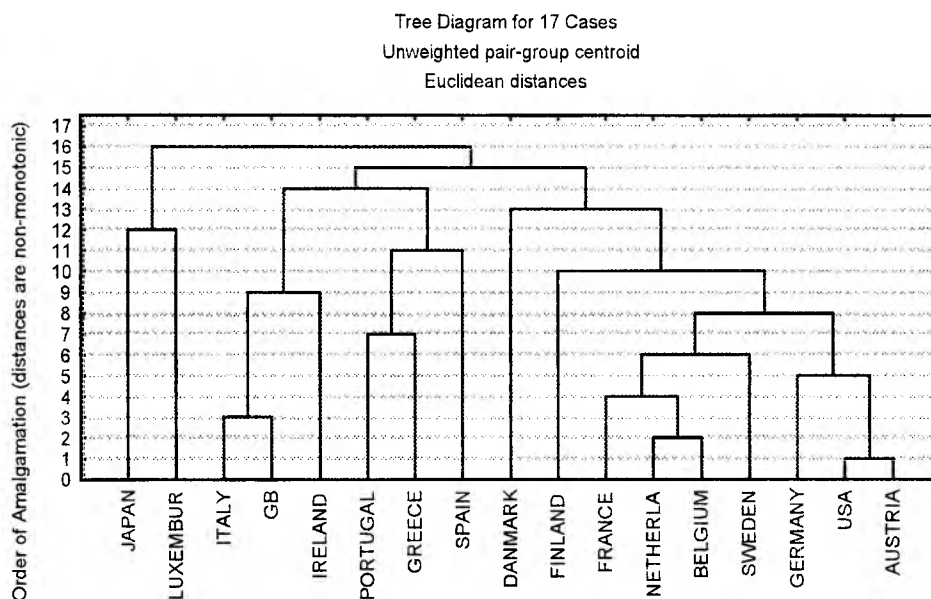
Страны первой и второй групп, в свою очередь, существенно отличаются от стран третьей и четвертой групп. Страны, стоящие на низшей ступени иерархии (третьей и четвертой группы) по своим характеристикам примерно в два раза уступают передовым странам. При этом Португалия выделяется особенно низкими уровнями ВВП в расчете на душу населения и среднемесячной оплаты труда.

Таким образом, при принятии управляющих решений следует учитывать, что инвестиции в странах с динамично развивающейся экономикой (первой и второй группы) могут встретить конкурирующие инвестиционные потоки, но здесь приложение капитала будет наиболее эф-

фективным и перспективным, в том числе и с точки зрения его приложения. Более того, стабильная экономическая обстановка в этих странах является своеобразным гарантом сохранения в определенном периоде времени благоприятных условий для использования инвестиций.

Инвестиции, направленные в страны с более низким уровнем экономического развития (третьей и четвертой групп), будут использоваться в менее благоприятной экономической среде. Во-первых, здесь, при более высоких уровнях инфляции, они подвергаются повышенному риску обесценивания. Во-вторых, потенциально их уровень доходности будет ниже из-за более низкой производительности труда и относительно более низкого выхода ВВП. Наконец, в-третьих, низкие уровни оплаты труда и высокий уровень безработицы могут привести к недовольству некоторой части населения, забастовкам и т. п., обуславливая тем самым дестабилизацию экономического положения.

Заметим что, результаты кластеризации основанные на предложенном подходе существенно отличаются от результатов, полученных традиционными методами. Так были проведены опытные расчеты с использованием метрики Евклида, реализованные в пакете Statistica. Сравнительный анализ полученных результатов предоставляется читателю.



I группа	II группа	III группа	IV группа
Люксембург	Австрия	Великобритания	Греция
Япония	Бельгия	Ирландия	Испания
	Германия	Италия	Португалия
	Дания		
	Нидерланды		
	США		
	Финляндия		
	Франция		
	Швеция		

Таблица 3. Многомерная группировка стран с использованием тривиальной метрики "евклидова расстояния"

Кластер	Кол-во стран	Среднегрупповые значения признаков (X_{ij})				
		ВВП в расчете на душу населения, дол. США	Средне-сячная заработная плата, дол. США	Индекс потребительских цен, %	Уровень безработицы, %	Инвестиции в расчете на душу населения, дол. США
I	2	43150,0	3142,2	108,3	3,4	6363,0
II	9	27188,9	2381,1	112,5	8,8	3704,9
III	3	18863,3	1873,3	116,8	10,6	3118,0
IV	3	11990,0	967,8	143,3	12,9	2783,8
В среднем по совокупности стран	17	24915,3	2131,6	118,2	9,2	3751,5

Как видно по данным Табл. 3, использование тривиального подхода в оценке расстояний между наблюдаемыми объектами, приводит в конечном счете к появлению искажающих эффектов. В данном случае Португалия, по своим признаковым характеристикам значительно удаленная от основной совокупности стран, вошла в группу, объединяющую Грецию и Испанию, хотя две последние страны имеют значения признаков отличающиеся более чем в 2-2,5 раза. Великобритания объединена в одной группе с Ирландией и Италией, также имеющими более чем двукратные отклонения признаков значений и т.д. Подобные группиров-

ки часто ставят исследователя перед проблемой идентификации образа и решения задачи по устранению "конфликта" между формальными аналитическими результатами и логикой.

ЛИТЕРАТУРА

- [1]. Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М. Многомерный статистический анализ в экономике. – М.: ЮНИТИ, 1999. – 599с.
Кулешов А.А., Позняк Ю.В., Метелица О.Н., Земсков С.В. Классификация экономических показателей. //Банковский вестник, №8, 1999. С. 25-29