Н.П. Третьяков
(Российский государственный социальный университет;
e-mail: trn12@smtp.ru)

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА К МИРОВОЙ СТАТИСТИКЕ ПОЖАРОВ

В статье рассмотрена группировка 48 стран по совокупности таких показателей, как среднее число пожаров (абсолютное и в расчёте на 1000 чел.) и среднее число жертв (абсолютное, в расчёте на 100000 чел. и в расчёте на 100 пожаров) в год. Таким образом, решалась задача выявления групп (кластеров) стран, близких между собой по общей пожарной обстановке. Использованы статистические данные Центра Пожарной Статистики Международной Ассоциации Пожарно-спасательных Служб (CTIF). Мелкомасштабное разбиение содержит 11 кластеров (Россия входит в один кластер с Украиной и Индией), крупномасштабное – 4 кластера (Россия вошла в группу с упомянутыми странами, а также с рядом стран СНГ).

Ключевые слова: статистика пожаров, кластерный анализ, группировка, многомерная классификация.

N.P. Tretiyakov

APPLICATION OF CLUSTER ANALYSIS TO WORLD FIRE STATISTICS

The article is devoted to the grouping of 48 countries according to such parameters as mean number of fires (total and per 1000 inh.) and mean number of fire deaths (total, per 100000 inh. and per 100 fires). The definition of clusters of countries that are similar in general fire situation is performed. Statistical data provided by Center of Fire Statistics (International Association of Fire and Rescue Services - CTIF) is used. The small scale partition contains 11 clusters, while the large scale partition contains 4 clusters.

Key word: fire statistics, cluster analysis, grouping, multidimensional classification.

Введение

Несмотря на то, что мировая пожарная статистика в последние два десятилетия существенно продвинулась вперёд [1-4], в основном благодаря деятельности Международной Ассоциации Пожарно-спасательных Служб [4], математические методы, используемые для обработки данных [5], весьма просты. Между тем, только исследование многомерных статистических распределений позволяет выявить скрытые закономерности и провести классификацию по совокупности показателей. Проведенное автором исследование является попыткой применения кластерного анализа для классификации стран по общей пожарной опасности. Обработаны 5 исходных показателей пожарной опасности: среднее число жертв (абсолютное, в расчёте на 100000 чел. и в расчёте на 100 пожаров) и среднее число пожаров (абсолютное и в расчёте на 1000 чел.) в год по данным CTIF [4]. Таким образом, решалась задача выявления групп (кластеров) стран, близких между собой по общей пожарной опасности.

Метод

Кластерный анализ – один из методов многомерного анализа, предназначенный для группировки (кластеризации) совокупности элементов, которые характеризуются многими факторами, и получения однородных групп (кластеров). Разбиение на кластеры происходит с помощью некоторой метрики, например, евклидова расстояния. Задача кластерного анализа, как и компонентного, состоит в представлении исходной информации об элементах в сжатом виде без ее существенной потери. Однако, в отличие от метода главных компонент, сжатие исходной информации производится иначе. Кластерный анализ не позволяет произвести ранжирование исходных объектов, но дает возможность сгруппировать их по степени близости по совокупности используемых показателей. Полученные группы объектов (кластеры) могут быть в той или иной степени ранжированы, например, по величине расстояния их средних точек (центров тяжести) до начала координат многомерного пространства.

В качестве основных методов анализа программный пакет STATISTICA предлагает Joining (tree clustering) – группу иерархических методов, которые используются в том случае, если число кластеров заранее не известно, и K-Means Clustering (метод К-средних), в котором пользователь заранее определяет количество кластеров.

Следует отметить, что в случае ранговых переменных применение евклидовой метрики является некорректным ввиду того, что для них не определены алгебраические операции сложения и умножения. Однако для ранговых переменных определена операция сравнения. Существует другая часто используемая метрика, представляющая собой сумму модулей разностей переменных двух объектов:

$$\rho(x,y) = \sum_i |x_i - y_i| .$$

Такая метрика в западной литературе получила название "City-block (Manhattan) distance". Поскольку арифметические разности между ранговыми переменными, выраженными натуральными числами, действительно отражают меры их близости, а последние уже можно складывать друг с другом, использование такой метрики в этом случае является возможным.

Можно указать на еще одно достоинство упомянутой метрики, заключающееся в том, что в ней сглаживаются (демпфируются) эффекты слишком большого различия отдельных координат. Это свойство представляется ценным в данном исследовании группировки стран по совокупности показателей, характеризующих пожарную обстановку.

Кроме метрики, в кластерном анализе необходимо задать другую опцию: расстояние между кластерами. Обычно эта опция определяется экспериментально, по ходу самого анализа. В данном исследовании оказалось, что оптимальным является расчет расстояния по принципу Ворда (Ward method).

Результаты исследования

Для исследования многомерных распределений показателей пожарной обстановки 48 стран были отобраны 5 исходных показателей: 1) среднее число пожаров в год; 2) среднее число пожаров в год в расчёте на 1000 чел.; 3) среднее число жертв в год; 4) среднее число жертв в год в расчёте на 100000 чел.; 5) среднее число жертв в год в расчёте на 100 пожаров. Таким образом, первый и третий показатели являются экстенсивными, т.е., вообще говоря, зависящими от величины страны (численности населения), а остальные показатели интенсивными, от величины страны не зависящими. Именно они качественно отражают степень экономического развития региона (города). Величины указанных показателей приведены в табл. 1.

Таблица 1

Исходные показатели пожарной опасности

1	2	3	4	5	6	7	8
№	Страна	Население, тыс. чел.	Среднее число жертв			Среднее число пожаров в год	Среднее число пожаров на 1000 чел. в год
			в год	на 100 тыс. чел.	на 100 пожаров		
1	Китай	1321852	2206	0,17	0,18	251786	0,19
2	Индия	1129866	8500	0,75	4,25	200000	0,18
3	США	301140	3625	1,20	1,20	1613400	5,36
4	Россия	141378	18759	13,27	13,16	236698	1,67
5	Филиппины	91077	249	0,27	2,50	9877	0,11
6	Вьетнам	85262	90	0,11	0,11	2154	0,03
7	Германия	82401	479	0,58	0,26	184485	2,24
8	Турция	71159	340	0,48	0,48	59618	0,84
9	Франция	63714	455	0,71	0,13	357654	5,61
10	Великобритания	60776	532	0,87	0,11	489942	8,06
11	Италия	58148	112	0,19	0,05	211504	3,64
12	Украина	46300	3909	8,44	7,30	53546	1,16
13	Южная Африка	42880	1817	4,24	3,52	51620	43831,00
14	Польша	38518	503	1,30	0,28	179815	4,67
15	Узбекистан	27780	184	0,66	1,20	15295	0,55
16	Малайзия	24821	62	0,25	0,23	27012	1,09
17	Тайвань	22859	169	0,74	2,23	7590	0,33
18	Румыния	21537	217	1,01	1,88	11957	0,56
19	Австралия	20434	122	0,60	0,60	113442	5,55
20	Казахстан	15285	125	3,43	3,03	17340	1,13
21	Греция	10706	70	0,66	0,26	27391	2,56
22	Португалия	10643	93	0,87	0,14	64560	6,07
23	Чехия	10229	102	1,00	0,53	19369	1,89
24	Сербия	10150	14	0,14	0,09	16334	1,61

Продолжение табл. 1

1	2	3	4	5	6	7	8
25	Венгрия	9956	154	1,55	0,62	24897	2,50
26	Беларусь	9725	105	12,57	10,25	11916	1,23
27	Швеция	9031	42	1,16	0,39	26772	2,96
28	Австрия	8200	34	0,51	0,13	32204	3,93
29	Швейцария	7555	103	0,45	0,22	15126	2,00
30	Болгария	7323	34	1,41	0,46	12585	3,08
31	Таджикистан	7077	8	0,48	2,67	1272	0,18
32	Лаос	6522	37	0,12	6,78	118	0,02
33	Иордания	6053	82	0,62	0,37	9867	1,63
34	Дания	5468	52	1,50	0,49	16626	3,04
35	Словакия	5448	106	0,96	0,43	11978	2,20
36	Финляндия	5239	59	2,01	0,72	14757	2,82
37	Норвегия	4628	2	1,28	0,46	12826	2,77
38	Сингапур	4553	37	0,05	0,04	4828	1,06
39	Хорватия	4493	200	0,82	0,49	8039	1,79
40	Молдова	4321	34	4,23	7,62	2623	0,61
41	Новая Зеландия	4116	39	0,84	0,15	22524	5,47
42	Ирландия	4109	13	0,96	0,13	31051	7,56
43	Албания	3601	265	0,37	0,71	1827	0,51
44	Литва	3575	51	7,42	1,40	18983	5,31
45	Монголия	2952	6	1,72	2,41	2112	0,72
46	Кувейт	2506	239	0,24	0,13	4775	1,91
47	Латвия	2260	16	10,58	1,98	12050	5,33
48	Словения	2009	2206	0,81	0,24	6571	3,27

Исследование проводилось с использованием программного пакета Statistica и состояло из следующих этапов:

А) Ранжирование объектов (стран) отдельно по каждому из 5 показателей пожарной опасности и получение таким образом 5 новых (ранговых) переменных (ранжирование проведено по убыванию, т.е. ранг 1 получает страна с наибольшим значением показателя).

Б) Проведение кластерного анализа полученных ранговых переменных с использованием иерархических алгоритмов, поскольку число кластеров заранее неизвестно.

В) Использование метода К-средних для определения состава кластеров и их характеристик.

Результат анализа п. Б) в виде иерархического дерева приведен на рис. 1. На графике чётко проявляются четыре крупные группы (кластера) и 11 мелкомасштабных групп. Отметим, что приведенная нумерация их не имеет отношения к какому-либо ранжированию и является случайной (соответствует той, что получилась на следующем этапе анализа).

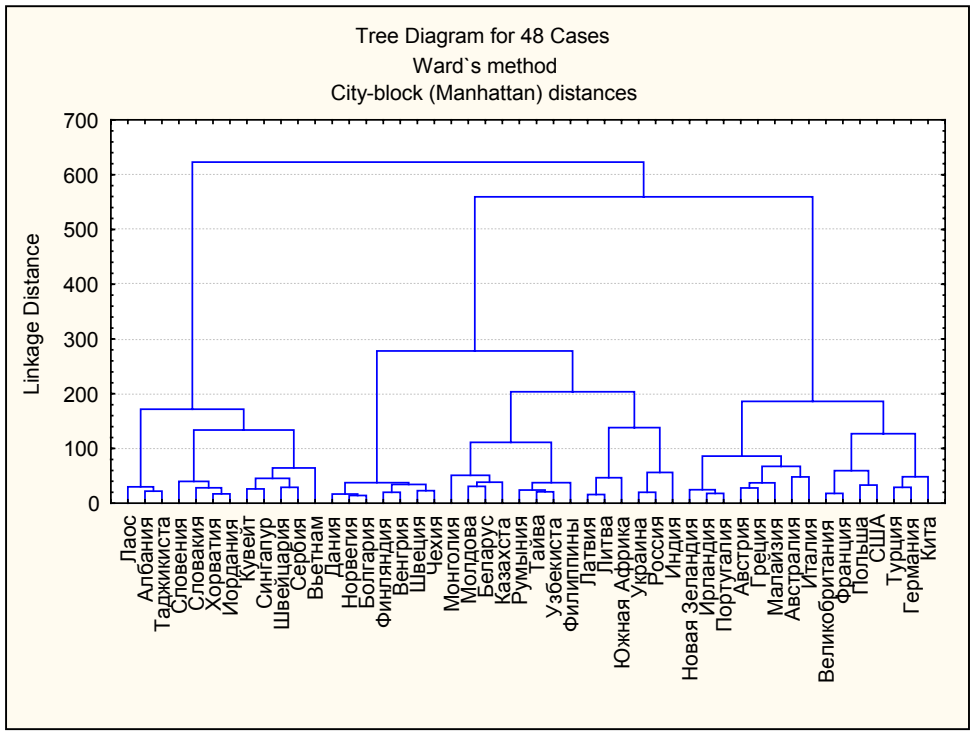

Рис. 1. Иерархическая диаграмма результатов кластерного анализа

На этапе В) были определены составы каждого из кластеров. Для крупномасштабного разбиения они приведены в табл. 2.

Таблица 2

Состав кластеров 1 и 2

Кластер 1		Кластер 2		Кластер 3		Кластер 4	
Страна	Расстояние от центра	Страна	Расстояние от центра	Страна	Расстояние от центра	Страна	Расстояние от центра
Южная Африка	14,0	Индия	12,3	Китай	15,1	Вьетнам	11,1
Чехия	6,0	Россия	14,5	США	12,4	Малайзия	10,0
Венгрия	4,7	Филиппины	12,0	Германия	4,7	Сербия	9,3
Швеция	4,0	Украина	10,3	Турция	10,8	Швейцария	6,5
Болгария	3,3	Узбекистан	7,6	Франция	6,4	Таджикистан	11,7
Дания	2,3	Тайвань	8,2	Великобритания	9,5	Лаос	14,6
Словакия	7,6	Румыния	4,1	Италия	9,0	Иордания	3,8
Финляндия	4,2	Казахстан	5,4	Польша	8,2	Сингапур	9,5
Норвегия	5,0	Беларусь	8,5	Австралия	7,6	Хорватия	7,7
Новая Зеландия	11,0	Молдова	8,2	Греция	7,9	Албания	7,5
Ирландия	11,4	Монголия	12,5	Португалия	8,4	Кувейт	6,6
Литва	7,9			Австрия	10,0	Словения	10,5
Латвия	9,1						

В таблице, помимо названий стран, приведены также их расстояния до центра (средней точки) данного кластера (в тех же единицах, что обрабатываемые ранги показателей). Эти расстояния позволяют судить, насколько та или иная страна по своим показателям близка к значениям, характерным для кластера, в состав которого она входит. Отметим, что эти расстояния никак не определяют рейтинги или ранжирования стран в составе группы; кластерный анализ не в состоянии ранжировать объекты, а лишь группирует их.

На рис. 2 приведена диаграмма средних значений рангов всех шести показателей для каждого кластера.

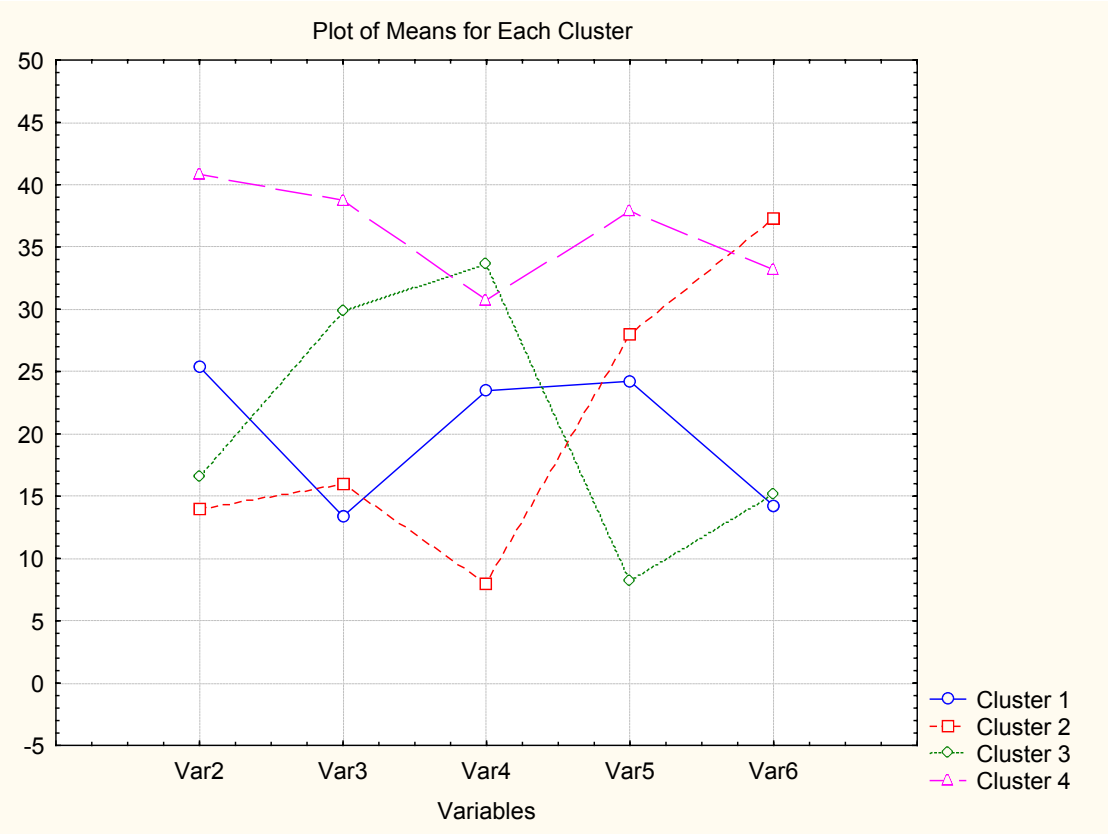

Рис. 2. Средние значения рангов показателей для каждого кластера

Из рис. 2 можно сделать вывод, что кластер 4 – это группа стран, наиболее благополучных по общей пожарной опасности (высокие ранги означают относительно малые значения показателей – ранжирование проводилось по убыванию). Кластер 2, в который вошла Россия, – это группа стран, неблагополучных по первым трём показателям (абсолютное и удельные количества жертв), однако относительно благополучных в отношении 4-го и 5-го показателей (по сравнению с кластерами 1 и 3).

Литература

1. Brushlinsky N.N. Magnetometric method of investigating fire sites. "Fire Technology", vol. 33, № 3, 1997.

2. Брушлинский Н.Н. Экономическая оценка борьбы с пожарами в современном мире. - М.: ВНИИПО, 1998.

3. Брушлинский Н.Н. Мировая пожарная статистика в конце XX века. - М.: Академия ГПС, 2000.

4. Fire statistics. CTIF Report, № 11. Moscow-Berlin, June 2006.

5. Fire Data Analysis Handbook. Second Edition FA-266 / January 2004. US Department of Homeland Security, FEMA, 2004.

Статья поступила в редакцию Интернет-журнала 15 мая 2009 г.