UNIVERSIDADE FEDERAL DE MINAS GERAIS

Programa de Pós-Graduação em Estatística com Ênfase em Indústria e Mercado

Monografia

“Análise de Agrupamentos com Uso do Excel”

Autor: Davidson Marcos de Oliveira

Orientador: Prof. Roberto da Costa Quinino

2015

Davidson Marcos de Oliveira

“Análise de Agrupamentos com Uso do Excel”

Monografia para Especialização apresentada ao Programa de Pós-Graduação em Estatística com Ênfase em Indústria e Mercado do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do Grau de Especialista em Estatística

Área de Concentração: Estatística Multivariada
Orientador: Prof. Roberto da Costa Quinino

Belo Horizonte
Universidade Federal de Minas Gerais
Instituto de Ciências Exatas ICEx

Agosto/2015

Dedico esta conquista aos meus pais Nelito e Neuza, minha esposa Elenir e minha filha Letícia.

Vocês são os meus pilares e a fonte de inspiração para que eu busque sempre os melhores resultados.

“A escada do conhecimento não conhece o último degrau!”

AGRADECIMENTOS

Agradeço à Deus por possibilitar a realização deste projeto, colocando em minhas mãos preciosos recursos.

Agradeço ao meu orientador Roberto Quinino (UFMG) pelo tempo disponibilizado a me orientar.

Agradeço a todos os professores do curso pelas lições valiosas.

Agradeço a minha esposa Elenir e minha filha Letícia pelo apoio, orações e abnegação para que este projeto fosse concluído.

Agradeço meu Tio José Raimundo e a minha Tia Conceição pela acolhida durante a realização do curso.

Agradeço ao colega Tobias pelo apoio durante o projeto.

SUMÁRIO

SUMÁRIO DE FIGURAS

RESUMO

Neste trabalho mostramos como realizar uma análise de agrupamentos com uso do Excel por meio de um modelamento via programação matemática. Todas as etapas são descritas de tal forma que possa ser usado como elemento didático em cursos de Estatística Multivariada que contenha o tópico análise de agrupamentos.

Palavras-chave: Análise de Agrupamentos, Excel, Programação Matemática.

ABSTRACT

In this study we show how to perform cluster analysis using Excel by means of a mathematical modeling program. All steps are described in a way that it can be used as a didactic element in Multivariate Statistics courses that contain the topic cluster analysis.

Key words: *Cluster Analysis, Excel, Mathematical Programing.*

1. INTRODUÇÃO

Considere que exista uma amostra de *n* objetos, cada um dos quais tem um escore em *p* variáveis. A ideia de uma análise de agrupamentos é usar os valores das variáveis para planejar um esquema para agrupar os objetos em classes, de modo que objetos similares estejam em uma mesma classe.

Muitos algoritmos têm sido propostos para análise de agrupamento. Neste trabalho consideraremos aqueles que começam com o cálculo das distâncias de cada objeto a todos os outros objetos.

Segundo Mingoti [1], *a Estatística Multivariada consiste em um conjunto de métodos estatísticos utilizados em situações nas quais muitas variáveis são medidas simultaneamente, em cada elemento amostral*.

O objetivo mais comum da análise de agrupamentos é tratar a heterogeneidade nos dados. O resultado é um pequeno número administrável de grupos, cada um consistindo em um número de objetos relativamente homogêneos com uma variação dentro do grupo consideravelmente menor do que o total de variação no conjunto completo de dados.

Várias são as aplicações potenciais da análise de agrupamentos. Por exemplo, em biologia evolucionária e ecológica podemos querer identificar e discriminar diferentes espécies e subespécies de plantas e animais de acordo com a similaridade relativa de suas características físicas. Em campanhas publicitárias é necessário segmentar, para um melhor desempenho da campanha, os indivíduos no mercado alvo com respeito às suas necessidades e suas reações comportamentais. O objetivo é dividir o mercado alvo em grupos menores que são mais homogêneos e, portanto, mais facilmente servidos por um tipo específico de produto ou uma campanha promocional específica. Na indústria, variáveis obtidas em processos podem ser agrupadas de maneira a encontrar aquelas que podem estar ou não interferindo no resultado de um processo.

Segundo Triola [2], *se dados amostrais não forem coletados de maneira apropriada, eles podem ser de tal modo inúteis que nenhuma manipulação estatística poderá salvá-los*, portanto, é fundamental que sejam empregadas técnicas confiáveis para explorar as informações e sobre elas tomar decisões. A abordagem quanto à quantidade de variáveis sendo analisadas e confrontadas podem ser decomposta em Análise Univariada, Bivariada ou Multivariada, sendo a última o objeto de nosso estudo.

O problema aqui discutido não é simples e aumenta a sua complexidade com o aumento dos objetos. Por exemplo, se considerarmos 20 objetos (uma quantidade pequena em casos reais) e desejarmos criar quatro grupos diferentes de igual tamanho já teríamos cerca de 488 milhões possibilidades [20!/ (5!5!5!5!4!)].

Neste trabalho apresentaremos uma solução com uso da programação matemática. Usaremos como ferramenta computacional o Excel, o que pode facilitar a utilização da técnica em empresas que não possuem softwares estatísticos específicos. Em geral existem vários algoritmos para análise de agrupamentos e o leitor interessado em métodos alternativos ao apresentado neste trabalho pode consultar Johnson & Wichen [3].

2. ESTUDO DE CASO

Análise de Agrupamento é um recurso que permite avaliar se determinados elementos podem ser agrupados com base na similaridade de características, valores ou comportamento. Ela tem como objetivo *dividir os elementos da amostra ou população em grupos de forma que os elementos pertencentes a um esmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas*, (Mingoti 2005).

O desafio das técnicas de agrupamento é a definição de critérios que irão determinar o quão distante elementos de um subgrupo são considerados semelhantes ou não.

Segundo Mingoti [1], *é necessário considerar medidas que descrevam a similaridade entre elementos amostrais de acordo com as características que nelas foram medidas*. Considerando o fato de estar lidando com múltiplas variáveis, as análises resultantes deverão ser agrupadas em um vetor, e este vetor deverá ser comparado entre os elementos do grupo.

Neste trabalho, para melhor entendimento, usaremos um exemplo para explicar a metodologia. Esta abordagem permitirá ao leitor usar a metodologia em outros trabalhos sem grandes dificuldades. Recomendamos que o leitor tenha disponível a planilha em Excel desenvolvida pelos autores para um melhor entendimento. Esta pode ser obtida diretamente com os autores por meio do endereço eletrônico.

Considere 49 cidades dos Estados Unidos (n = 49) e delas extraídas informações e dado um código conforme a Tabela 1. O objetivo é agrupar as cidades de acordo com similaridades das variáveis p1 a p6 descritas na Tabela 1.

Tabela 1: Codificação das variáveis de estudo

Código	Variável
p1	Percentual da população afro descendente
p2	Percentual da população de origem hispânica
p3	Percentual da população de origem asiática
p4	Idade média da população
p5	Taxa de desemprego
p6	Renda per capta

Os dados da pesquisa estão contidos na Tabela 2.

Tabela 2 : Dados da Pesquisa

Ordem	Cidade	p1	p2	p3	p4	p5	p6
1	Albuquerque	3	35	2	32	5	18
2	Atlanta	67	2	1	31	5	22
3	Austin	12	23	3	29	3	19
4	Baltimore	59	1	1	33	11	22
5	Boston	26	11	5	30	5	24
6	Charlotte	32	1	2	32	3	20
7	Chicago	39	20	4	31	9	24
8	Cincinnati	38	1	1	31	8	21
9	Cleveland	47	5	1	32	13	22
10	Columbus	23	1	2	29	3	13
11	Dallas	30	21	2	30	9	22
12	Denver	13	23	2	34	7	23
13	Detroit	76	3	1	31	9	21
14	El Paso	3	69	1	29	11	13
15	Fort Worth	22	20	2	30	9	20
16	Fresno	9	30	13	28	13	16
17	Honolulu	1	5	71	37	5	24
18	Houston	28	28	4	30	7	22
19	Indianapolis	22	1	1	32	5	21
20	Jacksonville	25	3	2	32	7	19
21	Kansas City	30	4	1	33	6	21
22	Las Vegas	11	13	4	33	5	20
23	Long Beach	14	24	14	30	8	21
24	Los Angeles	14	40	10	31	11	21
25	Memphis	55	1	1	32	9	20
26	Miami	27	63	1	36	12	17
27	Milwaukee	31	6	2	30	5	22
28	Minneapolis	13	2	4	32	5	23
29	Nashville	23	1	1	33	3	24
30	New Orleans	62	4	2	32	7	18
31	NY	29	24	7	34	11	27
32	Oakland	44	14	15	33	10	24
33	Oklahoma City	16	5	2	32	6	17
34	Omaha	13	3	1	32	5	20
35	Philadelphia	40	6	3	33	9	23
36	Phoenix	5	20	2	31	4	19
37	Pittsburgh	26	1	2	35	7	21
38	Portland	8	3	5	35	7	20
39	Sacramento	15	16	15	32	8	20
40	St. Louis	48	1	1	33	8	23
41	San Antonio	7	56	1	30	5	17
42	San Diego	9	21	12	31	8	20
43	San Francisco	11	14	29	36	6	31
44	San Jose	5	27	20	30	8	26
45	Seattle	10	4	12	35	5	28
46	Toledo	20	4	1	32	6	19
47	Tucson	4	29	2	31	3	19
48	Tulsa	14	3	1	33	4	20
49	Virginia Beach	14	3	4	29	6	18

3. ETAPAS DA ANÁLISE DE AGRUPAMENTOS POR MEIO DA PROGRAMAÇÃO MATEMÁTICA E PLANILHA EXCEL

3.1 Etapa 1: Padronização da Variáveis

Quando as variáveis estão em unidades diferentes e você deseja minimizar o efeito das diferenças de escala uma sugestão é padronizar as variáveis. Assim, nesta etapa converteremos todas as variáveis para uma escala comum, subtraindo cada variável da sua média e dividindo pelo seu respectivo desvio padrão. Os dados padronizados estão na Tabela 3.

Tabela 3 : Dados Padronizados

Ordem	Cidade	p1	p2	p3	p4	p5	p6
1	Albuquerque	-1,17872	1,238954	-0,36257	0,061342	-0,75146	-0,87523
2	Atlanta	2,355188	-0,76443	-0,4523	-0,43962	-0,75146	0,324386
3	Austin	-0,68177	0,510449	-0,27285	-1,44154	-1,49534	-0,57533
4	Baltimore	1,91345	-0,82514	-0,4523	0,562301	1,480155	0,324386
5	Boston	0,091278	-0,21806	-0,09339	-0,94058	-0,75146	0,924195
6	Charlotte	0,422582	-0,82514	-0,36257	0,061342	-1,49534	-0,27542
7	Chicago	0,809103	0,328323	-0,18312	-0,43962	0,736282	0,924195
8	Cincinnati	0,753886	-0,82514	-0,4523	-0,43962	0,364346	0,024482
9	Cleveland	1,250842	-0,58231	-0,4523	0,061342	2,224028	0,324386
10	Columbus	-0,07437	-0,82514	-0,36257	-1,44154	-1,49534	-2,37475
11	Dallas	0,312147	0,389031	-0,36257	-0,94058	0,736282	0,324386
12	Denver	-0,62655	0,510449	-0,36257	1,063261	-0,00759	0,624291
13	Detroit	2,852145	-0,70373	-0,4523	-0,43962	0,736282	0,024482
14	El Paso	-1,17872	3,30305	-0,4523	-1,44154	1,480155	-2,37475
15	Fort Worth	-0,12959	0,328323	-0,36257	-0,94058	0,736282	-0,27542
16	Fresno	-0,84742	0,93541	0,624433	-1,9425	2,224028	-1,47504
17	Honolulu	-1,28916	-0,58231	5,828653	2,566139	-0,75146	0,924195
18	Houston	0,201712	0,813993	-0,18312	-0,94058	-0,00759	0,324386
19	Indianapolis	-0,12959	-0,82514	-0,4523	0,061342	-0,75146	0,024482
20	Jacksonville	0,03606	-0,70373	-0,36257	0,061342	-0,00759	-0,57533
21	Kansas City	0,312147	-0,64302	-0,4523	0,562301	-0,37953	0,024482
22	Las Vegas	-0,73698	-0,09664	-0,18312	0,562301	-0,75146	-0,27542
23	Long Beach	-0,57133	0,571158	0,714161	-0,94058	0,364346	0,024482
24	Los Angeles	-0,57133	1,542497	0,355249	-0,43962	1,480155	0,024482
25	Memphis	1,69258	-0,82514	-0,4523	0,061342	0,736282	-0,27542
26	Miami	0,146495	2,938798	-0,4523	2,06518	1,852091	-1,17514
27	Milwaukee	0,367364	-0,5216	-0,36257	-0,94058	-0,75146	0,324386
28	Minneapolis	-0,62655	-0,76443	-0,18312	0,061342	-0,75146	0,624291
29	Nashville	-0,07437	-0,82514	-0,4523	0,562301	-1,49534	0,924195
30	New Orleans	2,079102	-0,64302	-0,36257	0,061342	-0,00759	-0,87523
31	NY	0,25693	0,571158	0,086066	1,063261	1,480155	1,823908
32	Oakland	1,08519	-0,03593	0,803889	0,562301	1,108219	0,924195
33	Oklahoma City	-0,4609	-0,58231	-0,36257	0,061342	-0,37953	-1,17514
34	Omaha	-0,62655	-0,70373	-0,4523	0,061342	-0,75146	-0,27542
35	Philadelphia	0,86432	-0,5216	-0,27285	0,562301	0,736282	0,624291
36	Phoenix	-1,06829	0,328323	-0,36257	-0,43962	-1,1234	-0,57533
37	Pittsburgh	0,091278	-0,82514	-0,36257	1,56422	-0,00759	0,024482
38	Portland	-0,90263	-0,70373	-0,09339	1,56422	-0,00759	-0,27542
39	Sacramento	-0,51611	0,085488	0,803889	0,061342	0,364346	-0,27542
40	St. Louis	1,306059	-0,82514	-0,4523	0,562301	0,364346	0,624291
41	San Antonio	-0,95785	2,513837	-0,4523	-0,94058	-0,75146	-1,17514
42	San Diego	-0,84742	0,389031	0,534705	-0,43962	0,364346	-0,27542
43	San Francisco	-0,73698	-0,03593	2,06008	2,06518	-0,37953	3,023526
44	San Jose	-1,06829	0,753284	1,252529	-0,94058	0,364346	1,524004
45	Seattle	-0,7922	-0,64302	0,534705	1,56422	-0,75146	2,123813
46	Toledo	-0,24003	-0,64302	-0,4523	0,061342	-0,37953	-0,57533
47	Tucson	-1,1235	0,874701	-0,36257	-0,43962	-1,49534	-0,57533
48	Tulsa	-0,57133	-0,70373	-0,4523	0,562301	-1,1234	-0,27542
49	Virginia Beach	-0,57133	-0,70373	-0,18312	-1,44154	-0,37953	-0,87523

3.2 Etapa 2: Agrupamento Inicial

Inicialmente devemos decidir uma quantidade de agrupamentos a serem utilizados na análise. Segundo Mingoti [1], *uma questão de grande importância é de como se deve proceder para escolher o número final g de grupos que define a partição do conjunto de dados analisado, ou de outra forma, em qual passo k o algoritmo de agrupamentos deve ser interrompido. Não existe uma resposta exata para esta pergunta. Entretanto, existem alguns critérios que podem auxiliar na decisão final.*

Foram escolhidos inicialmente 4 agrupamentos para realizar a análise e depois uma avaliação da necessidade de mais ou menos. Devemos considerar que uma quantidade grande de agrupamentos (>8) deve ser evitado uma vez que em termos práticos a sua administração torna-se complicada e consequentemente de pouca utilidade prática. Considere então que adotaremos nesta etapa inicial quatro agrupamentos. De forma aleatória devemos escolher um elemento para cada um dos quatro agrupamentos. Considere que as cidades Albuquerque, Atlanta, Austin e Baltimore foram alocadas respectivamente aos agrupamentos de 1 a 4. A Tabela 4 ilustra os procedimentos descritos e respectivos dados padronizados para as variáveis p1 a p6.

Tabela 4: Agrupamentos Iniciais

Nome	Ordem	p1	p2	p3	p4	p5	p6
Albuquerque	1	-1,17872	1,238954	-0,36257	0,061342	-0,75146	-0,87523
Atlanta	2	2,355188	-0,76443	-0,4523	-0,43962	-0,75146	0,324386
Austin	3	-0,68177	0,510449	-0,27285	-1,44154	-1,49534	-0,57533
Baltimore	4	1,91345	-0,82514	-0,4523	0,562301	1,480155	0,324386

O resultado das variáveis p1 a p6 deve ser obtido da Tabela 3 e neste caso sugerimos o uso da função Excel *procv*. Veja a Tabela 5 para um melhor entendimento. Nela destacamos a função usada para obter o resultado da variável p1 para cidade Albuquerque.

Tabela 5: Uso da função procv

Nome	Ordem	p1	p2	p3	p4	p5	p6
=PROCV($R4;$B$4:$O$52;9;0)						-0,75146	-0,87523
Atlanta	2	2,355188	-0,76443	-0,4523	-0,43962	-0,75146	0,324386
Austin	3	-0,68177	0,510449	-0,27285	-1,44154	-1,49534	-0,57533
Baltimore	4	1,91345	-0,82514	-0,4523	0,562301	1,480155	0,324386

3.3 Etapa 3: Cálculo da Distância de cada Objeto para o Agrupamento

Existem várias medidas de similaridade vetorial entre elementos, sendo uma delas a Distância Euclidiana. Segundo Mingoti [1], *a distância Euclidiana entre dois elementos* X_l *e* X_k*, sendo* $l \neq k$*, é definida por*:

$$d(X_l, X_k) = [(X_l - X_k)'(X_l - X_k)]^{1/2}$$

$$d(X_l, X_k) = \left[\sum_{i=l}^{p} [(X_{il} - X_{ik})^2]^{1/2}\right]$$

Ou seja, dois elementos amostrais são comparados em cada variável pertencente ao vetor de observações.

A Tabela 6 contém a distância de cada cidade para o agrupamento inicial. Esta distância é a Euclidiana. Por exemplo, a distância entre a cidade 1 e o agrupamento 2 é dado por:

$$D_{1;2} = \sqrt{(p1_1 - p1_2)^2 + \ldots + (p5_1 - p5_2)^2 + (p6_1 - p6_2)^2}$$

$$D_{1;2} = \sqrt{(-1{,}179 - 2{,}355)^2 + \ldots + (-0{,}751 + 0{,}751)^2 + (-0{,}875 - 0{,}324)^2} = 4{,}266$$

Tabela 6: Distâncias para os Grupos				
Cidade	Distância p/ 1	Distância p/2	Distância p/3	Distância p/4
1	0,000	4,266	1,920	4,528
2	4,266	0,000	3,640	2,487
3	1,920	3,640	0,000	4,715
4	4,528	2,487	4,715	0,000
5	2,837	2,483	2,053	3,379
6	2,782	2,216	2,315	3,420
7	3,243	2,496	3,240	2,131
8	3,211	1,976	2,949	1,919
9	4,418	3,218	4,676	1,141
10	3,247	3,841	2,324	4,911
11	2,757	2,824	2,655	2,619
12	2,154	3,664	3,152	3,286
13	4,827	1,598	4,510	1,595
14	3,709	6,504	4,491	6,157
15	2,347	3,194	2,381	2,947
16	3,795	5,353	3,988	4,685
17	7,153	7,884	7,593	7,689
18	2,262	2,829	2,038	3,188
19	2,486	2,553	2,301	3,081
20	2,428	2,646	2,544	2,612
21	2,640	2,328	2,823	2,479
22	1,619	3,383	2,245	3,601
23	2,243	3,646	2,249	3,616
24	2,649	4,428	3,419	3,678
25	3,884	1,807	3,845	1,101
26	3,942	5,822	5,521	4,683
27	2,817	2,066	1,946	3,119
28	2,569	3,050	2,426	3,442
29	3,087	2,797	2,907	3,628
30	3,835	1,530	3,677	1,993
31	3,997	3,995	4,679	2,736
32	3,881	2,923	4,198	1,840
33	2,015	3,257	2,262	3,415
34	2,109	3,083	2,100	3,472
35	3,463	2,371	3,727	1,367
36	1,149	3,724	1,154	4,338
37	3,081	3,130	3,741	2,576
38	2,664	3,959	3,591	3,413
39	2,176	3,521	2,679	3,188
40	3,767	1,856	3,831	1,305
41	1,666	4,922	2,299	5,374
42	1,869	3,764	2,291	3,561
43	5,200	5,487	5,678	5,193
44	3,296	4,457	3,264	4,385
45	3,969	4,260	4,344	4,189
46	2,159	2,819	2,250	3,031
47	1,015	4,020	1,157	4,774
48	2,214	3,173	2,400	3,651
49	2,564	3,350	1,682	3,895

Observe na Tabela 7 que a distância Euclidiana pode ser calculada com o uso da função SOMAXMY2 do Excel.

Tabela 7: Uso da FunçãoSOMAXMY2

Cidade	Distância p/ 1	Distância p/2	Distância p/3	Distância p/4
=(SOMAXMY2(J4:O4;$S$4:$X$4))^0,5			1,920	4,528
2	4,266	0,000	3,640	2,487
3	1,920	3,640	0,000	4,715
4	4,528	2,487	4,715	0,000
5	2,837	2,483	2,053	3,379
6	2,782	2,216	2,315	3,420
7	3,243	2,496	3,240	2,131

3.4 Etapa 4: Decidindo em qual grupo será alocado a cada objeto/cidade

Cada linha da Tabela 6 representa uma cidade e as colunas as distâncias. A menor distância em cada linha definirá o grupo que a cidade será alocada. Por exemplo, a cidade 5 será alocada ao Grupo 3 que corresponde a distância 2,053 que é a menor. No Excel a menor distância é calculada usando a função MÍNIMO. Todas as distâncias mínimas estão representadas no Tabela 9.

Tabela 9: Calculando as Distâncias Mínimas

Cidade	Distância p/ 1	Distância p/2	Distância p/3	Distância p/4	Distância Mínima
1	0,000	4,266	1,920	4,528	0,000
2	4,266	0,000	3,640	2,487	0,000
3	1,920	3,640	0,000	4,715	0,000
4	4,528	2,487	4,715	0,000	0,000
5	2,837	2,483	2,053	3,379	2,053
6	2,782	2,216	2,315	3,420	2,216
7	3,243	2,496	3,240	2,131	2,131
8	3,211	1,976	2,949	1,919	1,919
9	4,418	3,218	4,676	1,141	1,141
10	3,247	3,841	2,324	4,911	2,324
11	2,757	2,824	2,655	2,619	2,619
12	2,154	3,664	3,152	3,286	2,154
13	4,827	1,598	4,510	1,595	1,595
14	3,709	6,504	4,491	6,157	3,709
15	2,347	3,194	2,381	2,947	2,347
16	3,795	5,353	3,988	4,685	3,795
17	7,153	7,884	7,593	7,689	7,153
18	2,262	2,829	2,038	3,188	2,038
19	2,486	2,553	2,301	3,081	2,301
20	2,428	2,646	2,544	2,612	2,428
21	2,640	2,328	2,823	2,479	2,328
22	1,619	3,383	2,245	3,601	1,619
23	2,243	3,646	2,249	3,616	2,243
24	2,649	4,428	3,419	3,678	2,649
25	3,884	1,807	3,845	1,101	1,101
26	3,942	5,822	5,521	4,683	3,942
27	2,817	2,066	1,946	3,119	1,946
28	2,569	3,050	2,426	3,442	2,426
29	3,087	2,797	2,907	3,628	2,797
30	3,835	1,530	3,677	1,993	1,530
31	3,997	3,995	4,679	2,736	2,736
32	3,881	2,923	4,198	1,840	1,840
33	2,015	3,257	2,262	3,415	2,015
34	2,109	3,083	2,100	3,472	2,100
35	3,463	2,371	3,727	1,367	1,367
36	1,149	3,724	1,154	4,338	1,149
37	3,081	3,130	3,741	2,576	2,576
38	2,664	3,959	3,591	3,413	2,664
39	2,176	3,521	2,679	3,188	2,176
40	3,767	1,856	3,831	1,305	1,305
41	1,666	4,922	2,299	5,374	1,666
42	1,869	3,764	2,291	3,561	1,869
43	5,200	5,487	5,678	5,193	5,193
44	3,296	4,457	3,264	4,385	3,264
45	3,969	4,260	4,344	4,189	3,969
46	2,159	2,819	2,250	3,031	2,159
47	1,015	4,020	1,157	4,774	1,015
48	2,214	3,173	2,400	3,651	2,214
49	2,564	3,350	1,682	3,895	1,682

3.5 Etapa 5: Alocando cada objeto/cidade ao seu Grupo

Nesta etapa cada cidade é alocada a cada grupo e o resultado é registrado na coluna Grupo. No Excel a função CORRESP pode realizar esta tarefa uma vez que dado o valor mínimo encontrado, a função indica a qual coluna ela corresponde.

Tabela 10: Alocação das Cidades aos Grupos

Cidade	Distância p/ 1	Distância p/2	Distância p/3	Distância p/4	Distância Mínima	Grupo
1	0,000	4,266	1,920	4,528	0,000	1
2	4,266	0,000	3,640	2,487	0,000	2
3	1,920	3,640	0,000	4,715	0,000	3
4	4,528	2,487	4,715	0,000	0,000	4
5	2,837	2,483	2,053	3,379	2,053	3
6	2,782	2,216	2,315	3,420	2,216	2
7	3,243	2,496	3,240	2,131	2,131	4
8	3,211	1,976	2,949	1,919	1,919	4
9	4,418	3,218	4,676	1,141	1,141	4
10	3,247	3,841	2,324	4,911	2,324	3
11	2,757	2,824	2,655	2,619	2,619	4
12	2,154	3,664	3,152	3,286	2,154	1
13	4,827	1,598	4,510	1,595	1,595	4
14	3,709	6,504	4,491	6,157	3,709	1
15	2,347	3,194	2,381	2,947	2,347	1
16	3,795	5,353	3,988	4,685	3,795	1
17	7,153	7,884	7,593	7,689	7,153	1
18	2,262	2,829	2,038	3,188	2,038	3
19	2,486	2,553	2,301	3,081	2,301	3
20	2,428	2,646	2,544	2,612	2,428	1
21	2,640	2,328	2,823	2,479	2,328	2
22	1,619	3,383	2,245	3,601	1,619	1
23	2,243	3,646	2,249	3,616	2,243	1
24	2,649	4,428	3,419	3,678	2,649	1
25	3,884	1,807	3,845	1,101	1,101	4
26	3,942	5,822	5,521	4,683	3,942	1
27	2,817	2,066	1,946	3,119	1,946	3
28	2,569	3,050	2,426	3,442	2,426	3
29	3,087	2,797	2,907	3,628	2,797	2
30	3,835	1,530	3,677	1,993	1,530	2
31	3,997	3,995	4,679	2,736	2,736	4
32	3,881	2,923	4,198	1,840	1,840	4
33	2,015	3,257	2,262	3,415	2,015	1
34	2,109	3,083	2,100	3,472	2,100	3
35	3,463	2,371	3,727	1,367	1,367	4
36	1,149	3,724	1,154	4,338	1,149	1
37	3,081	3,130	3,741	2,576	2,576	4
38	2,664	3,959	3,591	3,413	2,664	1
39	2,176	3,521	2,679	3,188	2,176	1
40	3,767	1,856	3,831	1,305	1,305	4
41	1,666	4,922	2,299	5,374	1,666	1
42	1,869	3,764	2,291	3,561	1,869	1
43	5,200	5,487	5,678	5,193	5,193	4
44	3,296	4,457	3,264	4,385	3,264	3
45	3,969	4,260	4,344	4,189	3,969	1
46	2,159	2,819	2,250	3,031	2,159	1
47	1,015	4,020	1,157	4,774	1,015	1
48	2,214	3,173	2,400	3,651	2,214	1
49	2,564	3,350	1,682	3,895	1,682	3

Tabela 11: Função Corresp

Cidade	Distância p/ 1	Distância p/2	Distância p/3	Distância p/4	Distância Mínima	Grupo
1	0,000	4,266	1,920	4,528	=CORRESP(AG5;AC5:AF5;0)	
2	4,266	0,000	3,640	2,487	0,000	2
3	1,920	3,640	0,000	4,715	0,000	3
4	4,528	2,487	4,715	0,000	0,000	4
5	2,837	2,483	2,053	3,379	2,053	3
6	2,782	2,216	2,315	3,420	2,216	2
7	3,243	2,496	3,240	2,131	2,131	4
8	3,211	1,976	2,949	1,919	1,919	4
9	4,418	3,218	4,676	1,141	1,141	4
10	3,247	3,841	2,324	4,911	2,324	3
11	2,757	2,824	2,655	2,619	2,619	4
12	2,154	3,664	3,152	3,286	2,154	1
13	4,827	1,598	4,510	1,595	1,595	4
14	3,709	6,504	4,491	6,157	3,709	1

3.6 Etapa 6: Decidindo se o Agrupamento Obtido é o Melhor

O desempenho obtido com os agrupamentos utilizados é dado pela soma das distâncias mínimas da Tabela 10. No exemplo em questão a distância mínima obtida foi de 107, 462. Esta distância é válida para o Agrupamento inicial utilizado. Cada agrupamento inicial geraria uma soma de distâncias mínimas. Para termos a solução ótima precisaríamos avaliar todas as possibilidades de agrupamentos iniciais e utilizar a combinação que gerasse a menor soma de distâncias mínimas. O problema é que temos 211.876 possibilidades o que torna o processo complicado. Definindo SDM (i,j,k,l) como a soma das distâncias mínimas e N o número de objetos o problema pode ser modelado como um problema de programação matemática dado por:

$$
\begin{aligned}
&\textit{Minimizar}\ \text{SDM(i;j;k;l)} \\
&s.a \\
&1 \le i, j, k, l \le N \qquad (1) \\
&i, j, k, l \ \text{são inteiros} \\
&i \ne j \ne k \ne l
\end{aligned}
$$

O problema de programação matemática pode ser resolvido por meio do Solver do Excel com uso do método evolucionário.

Um algoritmo genético ou evolucionário aplica os princípios da evolução presentes na natureza para a necessidade de encontrar uma solução ótima para um problema. Trata-se de uma heurística famosa pela sua eficiência e motivação de grande interesse. São procedimentos interativos que mantém um grupo de soluções melhoradas a cada

interação. O método funciona tentando imitar o processo de seleção natural oriundo da biologia: ele elimina as soluções menos aptas e gera descendentes das soluções mais aptas.

As soluções iniciais são geradas aleatoriamente (mas em geral viáveis), formando a população inicial de indivíduos. Para cada uma dessas soluções calcula-se o valor da função objetivo ou aptidão do indivíduo. Os melhores indivíduos são escolhidos para reproduzir na geração corrente. Os indivíduos com piores desempenho são eliminados (morrem), mantendo a população em um nível (tamanho) desejável. Esse processo continua até que o critério de parada seja atendido. As restrições ao problema são tratadas de forma similar ao procedimento de multiplicadores de Lagrange sendo assim introduzidas na função objetivo, transformando o problema com restrições num problema irrestrito.

O problema formulado na expressão (1) pode ser usado no nosso caso. A figura 1 ilustra a formulação no Excel/Solver.

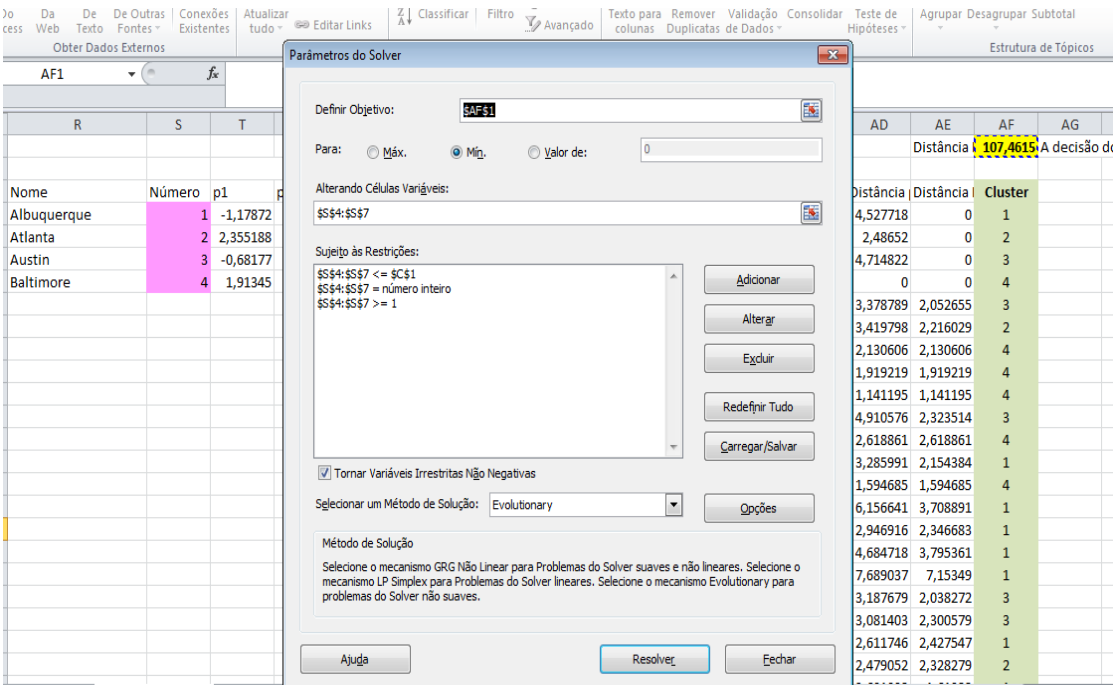

Figura 1: Formulação da Análise de Agrupamentos das Cidades com Excel/Solver

A solução ótima encontrada pelo Solver foi utilizar como agrupamento inicial as cidades San Francisco (43), Philadelphia (35), Omaha (34) e Long Beach (23) resultando em uma soma de distâncias mínimas igual a 77,58. O resultado final do agrupamento é descrito na Tabela 12.

Tabela 12: Agrupamento final com 4 Grupos

Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**
Honolulu	Atlanta	Albuquerque	Dallas
San Francisco	Baltimore	Austin	El Paso
Seattle	Chicago	Boston	Fort Worth
	Cincinnati	Charlotte	Fresno
	Cleveland	Columbus	Houston
	Detroit	Denver	Long Beach
	Memphis	Indianapolis	Los Angeles
	Miami	Jacksonville	Sacramento
	New Orleans	Kansas City	San Antonio
	NY	Las Vegas	San Diego
	Oakland	Milwaukee	San Jose
	Philadelphia	Minneapolis	
	Pittsburgh	Nashville	
	St. Louis	Oklahoma City	
		Omaha	
		Phoenix	
		Portland	
		Toledo	
		Tucson	
		Tulsa	
		Virginia Beach	

Existem várias técnicas para avaliarmos se a quantidade de grupos escolhido é razoável. Uma delas é o testar quantidades diferentes da que foi utilizada e registrar o valor da soma das distâncias mínimas. Uma das técnicas apresentada por Mingoti (2013) é o Coeficiente de Correlação semiparcial (Método de Ward). Segundo Mingoti [1], *para cada passo do agrupamento, calcula-se o coeficiente de correlação* semiparcial, *traçando-se então, um gráfico do passo versus o valor do coeficiente de correlação parcial observado. Busca-se no gráfico, o ponto da curva no qual ocorre um salto consideravelmente maior que os restantes*. Analogamente, optamos por calcular o

desempenho total obtido variando a quantidade de agrupamentos e calculando o salto no valor encontrado sempre que se aumenta um agrupamento. O Gráfico 1 apresenta os valores encontrados quando se varia o número de *clusters* de 1 a 6 e os respectivos saltos (subtração do valor anterior pelo novo valor encontrado. Com quatro grupos obtivemos 77,58. Para três e cinco grupos obtêm-se respectivamente 88,84 e 71,74.

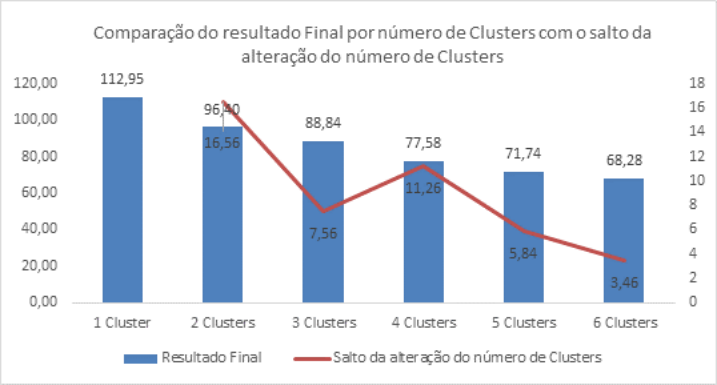

Gráfico 1 – Comparação do resultado Final por número de clusters com o salto da alteração do número de clusters.

Assim, o ganho que seria obtido aumentando para cinco grupos é relativamente pequeno em função do aumento da complexidade de trabalhar com um grupo a mais. Por outro lado, a diminuição para três grupos aumentou um pouco mais a soma das distâncias. Como quatro grupos não é difícil administrar, entendemos que seja melhor permanecer com quatro grupos. A decisão do número de grupos em geral não é objetiva e deve seguir o princípio da parcimônia.

Destacamos também que neste trabalho utilizamos a Distância Euclidiana para avaliar a similaridade entre os objetos (cidade). Com o uso do Excel é tarefa simples utilizar outras medidas como a Distância Euclidiana ao quadrado, Métrica de *Minkowski* e Distância de *Mahalanobis*. Pesquisas poderiam ser feitas formulando métricas originais e avaliando o seu desempenho.

Portanto, sem elaborar julgamento sobre a coerência das variáveis utilizadas, foi possível fazer um ajuste que distribuíssem as cidades em quatro grupos a partir da Distância Euclidiana.

4. CONCLUSÕES

Neste trabalho apresentamos a análise de agrupamentos como um problema de programação matemática. Entendemos que o trabalho permite uma compreensão didática do objetivo e mecanismos para realização da análise de agrupamentos. Com o uso do Excel entendemos que o aprendizado ficou mais motivador permitindo inclusive avaliar alternativas de métricas para a similaridade entre objetos. Entendemos que o Excel seria também uma boa alternativa para o aprendizado de outras técnicas multivariadas como a Análise Fatorial, Análise de Discriminante, Escalonamento Multidimensional, etc.

REFERÊNCIAS BIBLIOGRÁFICAS

MINGOTE, S. A. Análise de dados através de Estatística Multivariada: uma abordagem aplicada. Belo Horizonte; Editora UFMG, 2005.

TRIOLA, Mario F. Introdução à Estatística; tradução Vera Regina Lima de Farias e Flores, - 10ª ed. – rio de Janeiro: Editora LTC, 2008LATTIN, J; Crorroll, J. D.; Green, P. E. Análise de dados Multivaiados. São Paulo. Editora Cengage, 2011.

JOHSON, R. A.; Wichen, D. W. Applied Multivariate Statistical Analysis, Prentice Hall, 1988.

MANLY, B. J. F. Métodos Estatísticos Multivariados: Uma Introdução. Porto Alegre. Editora Bookman, 2008.

REIS, Elizabeth. Estatística Multivariada Aplicada. Lisboa. Editora Sílabo 2001.