IV SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R
R & PYTHON E AS TENDÊNCIAS DE COLABORAÇÃO
NITERÓI, 21 A 23 DE MAIO DE 2019

APLICAÇÃO DA TÉCNICA DE CLUSTERIZAÇÃO COM R: UMA ABORDAGEM PARA O SETOR DA SILVICULTURA NO BRASIL

Robert Armando Espejo[1]

Rildo Vieira de Araújo[2]

Reginaldo Brito da Costa [3]

Michel Constantino [4]

Resumo

O presente artigo teve como campo empírico de estudo o setor da silvicultura desenvolvido no Brasil, destacando-se como ramo de atividade próspero no meio nacional e internacional, ocasionada principalmente pelas exportações de celulose. O objetivo deste artigo foi apresentar as ferramentas que o R oferece para trabalhar com *Clusters* e fazer uma análise de como se realiza o agrupamento das empresas do setor de Papel e Celulose, mediante as técnicas *Ward, Single Link* e *Centroid*. Metodologicamente fizeram parte dos procedimentos o cálculo da eficiência utilizando o modelo de Análise Envoltória de Dados e posteriormente utilizou-se da técnica de Clusters para a devida classificação dos elementos. Os resultados demonstraram que empresas com maior quantidade de vendas e maior número de funcionários tiveram maior destaque na separação dos grupos.

Palavras-chave: Grupo, Papel e Celulose, software "*R*".

Abstract

This article has as an empirical field of study the forestry sector developed in Brazil, standing out as a branch of prosperous activity in the national and international environment, caused mainly by pulp exports. The objective of this article was to present the tools that the R offers to work with Clusters and to make an analysis of how the grouping of companies in the Pulp and Paper sector, through the techniques Ward, Single Link and Centroid. Methodologically, procedures were part of the calculation of efficiency using the model of Data Envelopment Analysis and later the Clusters technique was used for the proper classification of the elements. The results showed that companies with greater sales volume and higher number of employees were more prominent in the separation of groups.

Keywords: Cluster, Paper and Pulp, software "R".

[1] Universidade Católica Dom Bosco (UCDB), (UFMS) robert.espejo@ufms.br

[2] Universidade Católica Dom Bosco (UCDB), (IFMT) ifmt.rildo@gmail.com

[3] Universidade Católica Dom Bosco (UCDB), reg.brito.costa@gmail.com

[4] Universidade Católica Dom Bosco (UCDB), michel@ucdb.br

Introdução

A silvicultura brasileira está densamente relacionada ao incremento da indústria nacional de base florestal. Na década de 1950, o empenho de idealização do Estado brasileiro para agenciar o desenvolvimento econômico se consolidou no Plano de Metas, que nomeava cinco áreas importantes para destinação de investimentos e implantava finalidades para serem abrangidos em cinco anos. No período entre 2005 e 2015, a produção brasileira de celulose cresceu a uma taxa de 5,9% a.a. (SILVA et al. 2016).

Com o crescimento obtido na produção de papel e celulose o Brasil tem demostrado ser eficiente na sua produção, obtendo alcance não apenas nacional como também atingindo o mercado externo com volume considerável de exportações. Sua elevada competitividade é proveniente de condições climáticas adequadas e de um extenso histórico de aquisição em pesquisa e desenvolvimento florestal, efetivado em tão alto grau pelas principais empresas do setor, como por entidades de pesquisas (HORA, 2015).

A ampliação do setor florestal no Brasil chama atenção para realização de estudos e pesquisas visando compreender melhor o seu enquadramento no segmento, podendo confirmar por meio de técnicas e ferramentas que auxiliam no planejamento de grupos empresariais. Dentro da estatística multivariada é possível utilizar da técnica de cluster como um tipo de instrumento que contribui na classificação de cada grupo, complementando a análise de cada caso.

Objetivo

Este artigo irá apresentar as ferramentas que o R oferece para trabalhar com *Clusters* e fazer uma análise de como se realiza o agrupamento das empresas do setor de Papel e Celulose do ano de 2016, mediante o uso das técnicas *Ward, Single Link* e *Centroid*.

Material e Método

A abordagem escolhida para este estudo foi a analítica, envolvendo uma avaliação de informações disponíveis em empresas ligadas na área de silvicultura (THOMAS, et al. 1996).

Os dados utilizados foram obtidos da B3 – Brasil, Bolsa, Balcão e empresas do segmento de Papel e Celulose. Conforme a Tabela 1, foram trabalhadas as seguintes variáveis: a) a quantidade de funcionários (*INPUT*); b) tempo de vida das empresas; e c) o

resultado das vendas líquidas (*OUTPUT*). O número de empresas que apresentaram dados resumiu-se a 12 companhias, que apresentaram elementos para as três variáveis.

Tabela 1 – Empresas de Papel e Celulose.

Empresas	Vendas	func	Tempo
A	7.160,20	13.833	120
B	9.331,50	7.747	94
C	934,40	1.897	36
D	2.698,90	4.851	9
E	1.339,80	1.740	67
F	5.144,20	4.493	10
G	1.777,80	1.212	10
H	1.890,90	4.735	46
I	468,00	1.035	71
J	589,20	952	63
K	786,40	2.472	76
L	504,80	829	71

Fonte: BM&FBOVESPA e empresas do setor, 2019

Técnicas de Agrupamentos (Cluster)

Fadel et al (2014) exemplificam que o problema clássico de agrupamento é determinado como: dado um conjunto formado por "*n*" objetos $X = \{x_1, x_2, ..., x_i, ..., x_n\}$ em que com cada objeto $x_i \in X$ possui "*p*" atributos, carece arquitetar "*k*" grupos C_j ($j=1,...,k$) a partir de X, de forma a garantir que os componentes de cada grupo sejam homogêneos conforme a uma determinada medida de similitude conforme mostra a Figura 01 . Uma solução (ou partição) pode ser representada como $\pi = \{ C_1, C_2, ..., C_k \}$ (HAN et al., 2012).

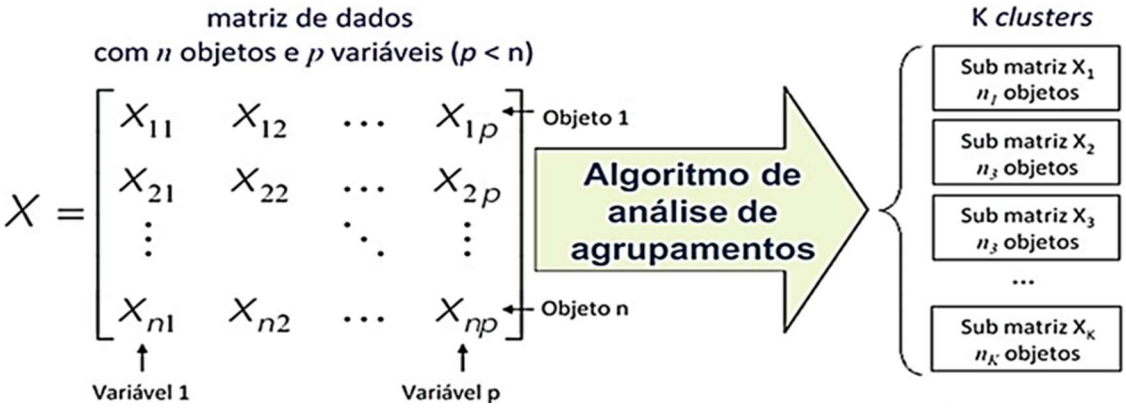

Figura 1 - Análise de agrupamentos

Fonte: Adaptado de Pessanha (2017)

Existem duas técnicas principais para a construção de clusters, podendo ser classificadas como hierárquicas e não-hierárquicas. Mingotti (2005) destaca que os métodos não-hierárquicos utilizam algoritmos interativos e, em comparação com métodos hierárquicos, apresentam maior capacidade de trabalhar com um volume grande de informações.

À medida que os objetos se afastam, os que decompõem posições muito conexas instituem clusters determinados por um método para agrupá-los hierarquicamente. Isto constitui na regra de amalgamação ou ligação (CÂMARA, 2009).

Tabela 2: Tipos de distância usadas em análise de cluster para variáveis intervalares e categóricas

Tipo de distância	**Fórmula**	**Observações**
Euclideana	$D(x, y)=\{\Sigma_i(x_i-y_i)^2\}^{1/2}$	Usa dados brutos. Não é afetada por adição, mas por mudança de escala.
Euclideana quadrada	$D(x, y)=\Sigma_{i(}(x_i-y_i)^2$	Quando se quer colocar maior peso nos objetos que estão mais separados.
City-block (Manhattan)	$D(x, y)=\Sigma_i\|x_i-y_i\|$	Semelhante à distância Euclidiana simples.
Percentagem de discordância	D(x, y) =número de $\Sigma_i(x_i\neq y_i)/i$	Útil se os dados incluídos na análise são categóricos (ou nominais) por natureza.

Fonte: CÂMARA (2009).

Conforme demonstrado na Tabela 2 foi utilizada a distância euclidiana para o desenvolvimento da pesquisa, onde a utilização dos dados brutos não é afetada por adição, mas por mudança de escala. Ponderando o vetor "**x**" de coordenadas reais (x_1, x_2, ..., x_p) como descritor dos objetos que investigarão a dissemelhança entre os grupos. A medida mais notória para indicar a adjacência entre os objetos A e B é a distância euclidiana d (A, B):

$$d(A,B) = [\sum_{i=1}^{p}(x_i(A) - x_i(B))^2]^{1/2}$$

ou em linguagem matricial:

$$d(A,B) = [(\mathbf{x}(A) - \mathbf{x}(B))'(\mathbf{x}(A) - \mathbf{x}(B))]^{1/2}$$

Nesse sentido, dentre diversos métodos de agrupamento, foram trabalhados três técnicas, conforme citados a seguir:

O Método de Ward, segundo Hair et al. (1998), é um algoritmo de agrupamento que principia com todos os objetos em um único grupo, sendo calculado como a soma de quadrados entre os grupos somados sobre todas as variáveis, tendendo a resultar em

agrupamentos de tamanhos quase iguais, devido a sua minimização de variação interna. Este é o método mais utilizado em estudos de cluster (BEM et al. 2015).

O método hierárquico aglomerativo do *single-link* emprega uma norma de mínima distância, que inicia juntando os dois indivíduos mais próximos, que estabelecerão o primeiro agrupamento. Na fase seguinte, pode acontecer que um terceiro indivíduo se unirá ao agrupamento já formado, ou outros dois indivíduos que agora estão mais próximos se unirão para formar um segundo agrupamento. A decisão é determinada pelo discernimento da distância mínima entre os elementos do grupo. A metodologia repete-se até que todos os indivíduos pertençam a um só grupo (VALLI, 2002).

A técnica do Centroide, pondera que a distância entre dois conglomerados é a distância entre seus centroides, que nada mais é que a média para todas as variáveis. A cada agrupamento novo de objetos, deve-se calcular um novo centroide. A maior dificuldade dessa técnica é de ter de recalcular os centros dos grupos a cada junção realizada. (HAIR et al. 2010).

Análise Envoltória de Dados

A Análise Envoltória de Dados (DEA) trabalha com modelos não-paramétricos, um dos modelos é conhecido como CCR, sendo representado em homenagem aos seus autores, porém, também é conhecido como modelo de retornos constantes à escala (*Constant Returns to Scale* - CRS); o modelo é linearizado, tornando-se um problema de programação linear, cuja fórmula é apresentada na Figura 2 (GOMES et al. 2003).

$$\max \quad h_o = \frac{\sum_{j=1}^{s} u_j y_{jo}}{\sum_{i=1}^{r} v_i x_{io}}$$

$$\text{sujeito a}$$

$$\frac{\sum_{j=1}^{s} u_j y_{jk}}{\sum_{i=1}^{r} v_i x_{ik}} \leq 1, \quad k = 1,\ldots,n$$

$$u_j, v_i \geq 0 \quad \forall i, j$$

Figura 2 – Fronteira Eficiente Modelo CCR

Fonte: GOMES et al. (2003)

Em sua formulação matemática, analisa-se que cada DMU *k*, *k* = 1, ..., *n*, é uma unidade de produção que utiliza r *INPUTS* x_{ik}, *i* =1, …, *r*, para produzir *s* *OUTPUTS* y_{jk}, *j* =1, …, *s*. O modelo CCR, mencionado, maximiza o quociente entre a combinação linear dos *OUTPUTS* e

a combinação linear dos *INPUTS*, com a restrição de que, para qualquer DMU, esse quociente não pode ser maior que 1. Assim, para uma DMU o h_o é a eficiência; x_{io} e y_{jo} são os *INPUTS* e *OUTPUTS* da DMU o v_i e u_j são os pesos calculados pelo modelo para *INPUTS* e *OUTPUTS*, respectivamente.

No primeiro momento foi utilizada a técnica de Análise Envoltória de Dados(DEA) para encontrar os resultados da eficiência gerados no modelo CCR e em seguida incorporados como uma nova variável, conforme a Tabela 3 e acrescentados aos dados originais para uma demonstração da técnica de *Cluster*, buscando identificar uma classificação de grupos que tenham maior similaridade ou dissimilaridades.

Tabela 3 – Empresas de Papel e Celulose com o resultado da eficiência.

Empresas	vendas	func	tempo	eff_ccr
A	7160,2	13833	120	0,35
B	9331,5	7747	94	0,82
C	934,4	1897	36	0,34
D	2698,9	4851	9	0,58
E	1339,8	1740	67	0,52
F	5144,2	4493	10	1,00
G	1777,8	1212	10	1,00
H	1890,9	4735	46	0,27
I	468	1035	71	0,31
J	589,2	952	63	0,42
K	786,4	2472	76	0,22
L	504,8	829	71	0,42

Fonte: ESPEJO et al., 2019

Para o processamento dos dados, foi utilizado o software R por meio da *package stats,* que realizou o cálculo estatístico para geração do *cluster* entre as empresas, visando identificar como se procedeu a sua classificação.

Resultados e Discussão

De acordo com a tabela 3, observa-se que o resultado de eficiência foi atingido pelas empresas “F e G” que obtiveram um resultado satisfatório atingindo a pontuação máxima relativa de 1,00 ou 100%, onde buscou-se a minimização dos insumos e a maximização do produto.

A partir da Figura 3 é possível destacar algumas estatísticas básicas que identificam o comportamento amostral das variáveis. Em relação às vendas, o valor mínimo foi de R$

468.000,00 e atingiu seu valor máximo em R$ 9.331.500,00, sendo que as vendas médias no período foram de R$ 2.718.800,00. Na questão dos funcionários, a menor quantidade foi de 829 empregados e chegou a uma escala máxima de 13.833 funcionários, com uma média de 3.816 empregados. Em relação ao tempo de vida (em anos) desde a data de fundação das empresas, a mais nova tinha 9 anos e a mais antiga 120 anos, chegando a um período médio de 56 anos. A última variável envolvendo o cálculo da eficiência por meio do método de DEA, evidenciou que o valor mínimo de eficácia atingido dentre as empresas foi de 0,21 e o valor máximo de 1,00 atingindo a sua eficiência; quanto à média obtida entre as DMUs o resultado foi de 0,52.

vendas	func	tempo	eff_ccr
Min. : 468.0	Min. : 829	Min. : 9.00	Min. : 0.2169
1st Qu. : 737.1	1st Qu.: 1168	1st Qu.: 29.50	1st Qu.: 0.3289
Median : 1558.8	Median : 2184	Median : 65.00	Median :0.4185
Mean : 2718.8	Mean : 3816	Mean : 56.08	Mean : 0.5210
3rd Qu.: 3310.2	3rd Qu.: 4764	3rd Qu.: 72.25	3rd Qu.: 0.6425
Max. : 9331.5	Max. :13833	Max. : 120.00	Max. : 1.0000

Figura 3 – Estatística descritiva

Fonte: ESPEJO et al. 2019

A Figura 4 demonstra a medida de distância utilizada no trabalho para auxiliar na visualização dos valores por empresa e assim facilitar a organização posterior do dendrograma. As marcações destacadas por coluna representam as menores distâncias entre as empresas, para uma posterior classificação das unidades.

	A	B	C	D	E	F	G	H	I	J	K
B	6461.7812										
C	13462.3823	10234.1171									
D	10029.5485	7237.7775	3440.9745								
E	13420.8984	9997.6021	435.8431	3395.4142							
F	9555.7290	5303.6830	4945.9386	2471.3675	4696.3515						
G	13721.2234	9988.5767	1086.8418	3753.7644	688.3874	4700.8095					
H	10514.0193	8027.2646	2994.8683	817.1224	3045.3534	3262.4870	3524.9989				
I	14442.1863	11118.1431	980.7130	4420.7030	1121.1941	5816.2129	1323.1124	3964.2489			
J	14460.3406	11072.5243	1006.4378	4433.5033	1088.2832	5769.7067	1217.8585	4000.7250	147.1137		
K	13026.8793	10042.1454	595.0874	3053.1601	917.6909	4804.0835	1604.6279	2518.3306	1471.8602	1532.7938	
L	14608.2431	11214.7174	1151.6966	4581.9650	1235.7840	5912.0766	1330.7664	4144.7237	209.2612	149.3866	1666.9651

Figura 4 – Distância "*Euclidean*"

Fonte: ESPEJO et al. 2019

Como o método *Single Link* é um dos algoritmos de agrupamento hierárquico mais simples por utilizar a técnica de vizinho mais próximo para classificar os grupos, ao analisar a Figura 5, foi definido um corte que demonstrou a quantidade de grupos formada por meio da distância euclidiana. Os resultados demonstraram que as empresas que apresentaram maiores vendas foram classificadas separadamente, representadas por: "A", "B" e "F".

Nota-se que as empresas que tiveram maiores resultados entre as variáveis vendas, quantidade de funcionários e tempo de fundação foram decisivos para se distanciarem das

demais empresas, destacando-se as empresas “A” e “B”. As empresas “D” e “H” ficaram juntas no grupo, pois apresentaram depois do primeiro agrupamento as que tiveram maiores vendas e quantidade de funcionários. Como as demais empresas tiveram um resultado menor entre as empresas analisadas, favoreceu para que mantivessem em um novo agrupamento, levando em consideração as variáveis vendas e número de funcionários.

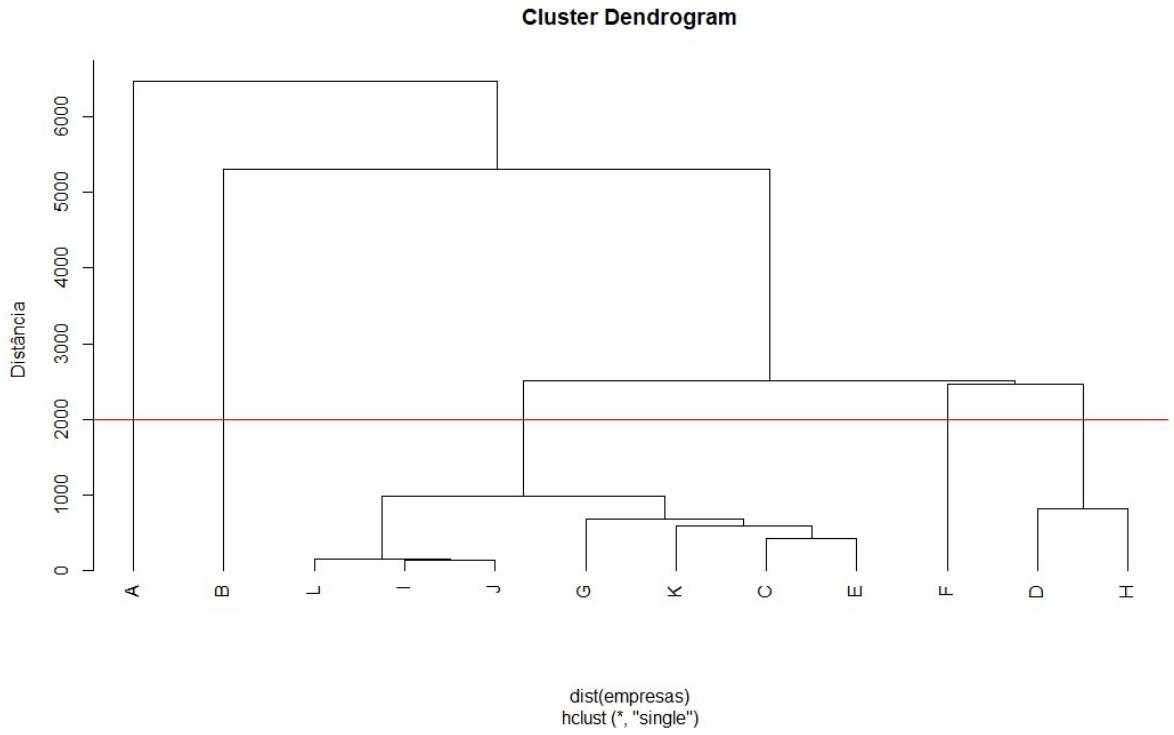

Figura 5 – Método “*Single Link*”

Fonte: ESPEJO et al. 2019

Diferentemente do *Single Link*, o método *Ward* é um dos algoritmos de agrupamento hierárquico que se baseia na perda de informação ocasionada pelo agrupamento dos elementos e medida por meio da soma dos quadrados dos desvios. Desta forma, o software *R* realiza um novo cálculo de distância que modificam a forma de agrupamento, sendo demonstrados na Figura 6.

O resultado *Ward* favoreceu a formação mais clara de pares de grupos, onde inicialmente ficou evidente dois grandes grupos, mas optou-se em um corte melhor que apresentasse a distribuição em três grupos, explicados da seguinte forma: As empresas que apresentaram maiores vendas, maior quantidade de funcionários e também maior tempo de mercado, foram classificadas em grupos conjuntamente através das unidades “A” e “B”, diferentemente no método *Single Link* onde as empresas foram classificadas de forma independente. Já as empresas “D”, “F” e “H” ficaram juntas no grupo, pois apresentaram,

depois do primeiro agrupamento, as que tiveram maiores vendas e uma quantidade maior de funcionários. A última classificação foi composta pelo restante das sete empresas, por apresentarem um menor resultado em relação as vendas e também um número menor funcionários.

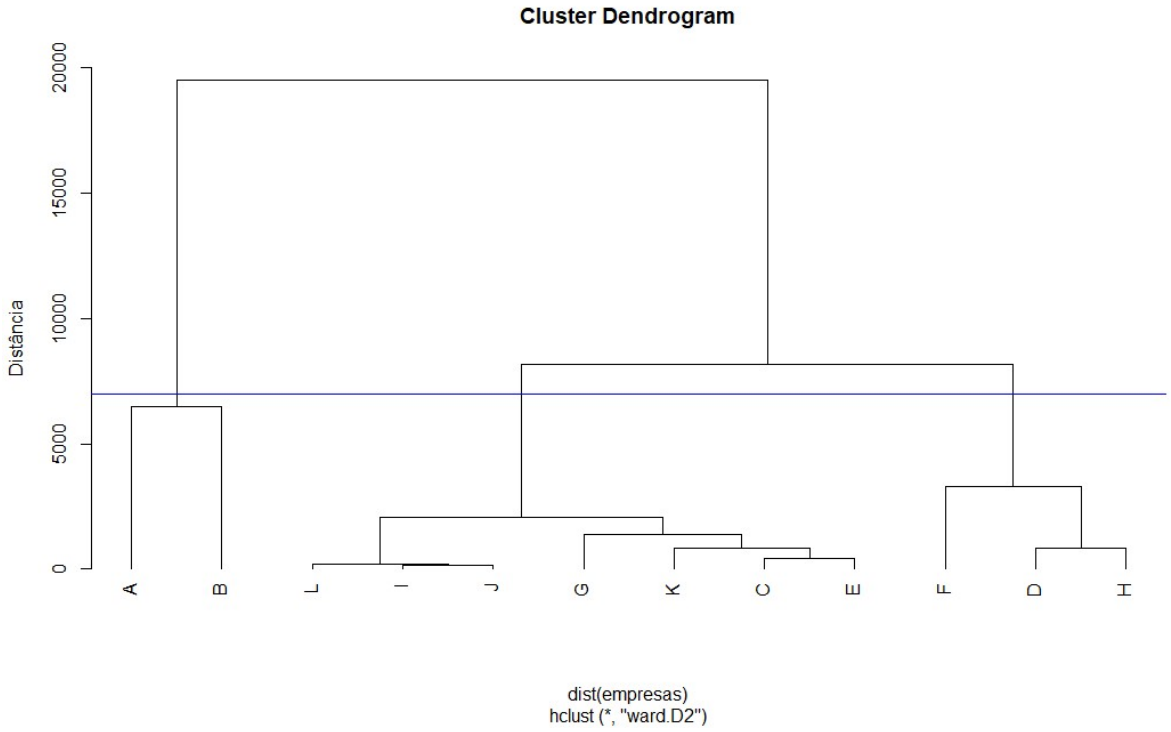

Figura 6 – Método "*Ward*"

Fonte: ESPEJO et al. 2019

Comparando-se com os métodos *Single Link* e *Ward*, a Figura 7 apresenta o resultado pelo método *Centroid* e pode ser visto como representando a observação média inclusa a um agrupamento entre todas as variáveis na análise. Dentre a amostra de 15 empresas, foi comparado as distâncias para ver quão diferentes os grupos estão um do outro.

Coincidentemente com o método *Single Link*, este dendrograma foi criado usando uma partição final de 5 grupos, e ocorrem em um nível de similaridade de aproximadamente 2000, onde o primeiro agrupamento à esquerda envolve sete empresas observadas nas linhas (K, I, J, G, K, C e E). O segundo corresponde a uma única empresa denominada "F". O terceiro agrupamento foi entre as empresas "D e H". E por fim, as unidades "A e B" foram classificados separadamente, pois verificou-se que apresentam maior porte, ao observar as vendas, número de empregados e tempo de fundação.

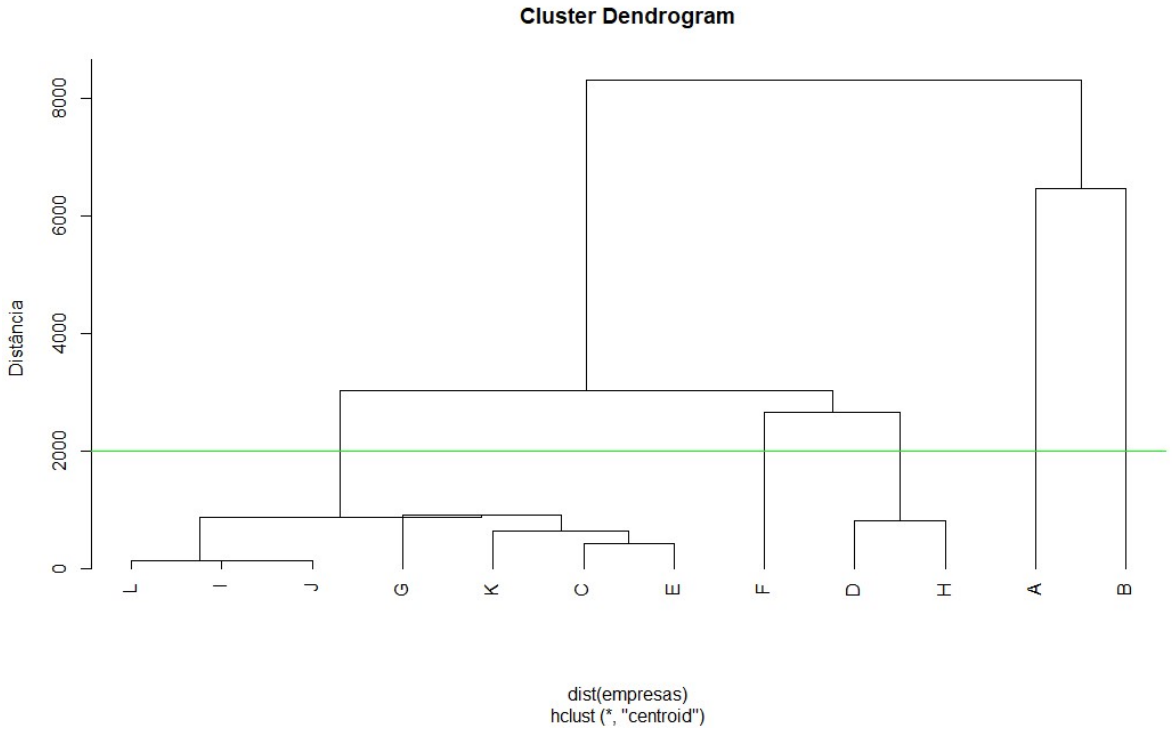

Figura 7 – Método "*Centroid*"

Fonte: ESPEJO et al. 2019

Conclusão

Neste trabalho foi apresentada uma proposta metodológica para avaliar a quantidade de agrupamentos entre as empresas de Papel e Celulose, sendo utilizado técnicas de mineração de dados com a clusterização dos algoritmos hierárquicos.

Dentre os métodos utilizados, cabe ressaltar que a experiência do pesquisador com os dados favorecerá na escolha de um ou de outro método para explicar o agrupamento das unidades que tenham certa similaridade ou dissimilaridade. Ao se avaliar como se realiza o agrupamento de acordo com os métodos, conclui-se que as empresas com maior quantidade de vendas e maior número de funcionários tiveram maior destaque na hora de realizar a separação dos grupos, denotando que as empresas de maior porte tendem a se unir com empresas com a mesma similaridade e em alguns casos a quantidade de tempo em atividade também influenciou nos resultados.

Em relação à variável eficiência introduzida nos dados, nota-se pouca influência na formação dos *clusters*, devido às grandezas numéricas das demais variáveis; sugere-se para futuras pesquisas a sua padronização, podendo obter novos resultados com outros modelos.

Dentre os três métodos apresentados, o método *Ward* identificou um agrupamento com menor número de clusters, facilitando na explicação do conjunto de empresas utilizado na amostra e levando principalmente em consideração as características comuns dos dados observados.

Referências

CÂMARA, Fernando. **Coluna psiquiatria contemporânea.** Psiquiatria e estatística. Parte II: Fundamentos da análise de clusters (classificação numérica). IMPPG-UFRJ. Janeiro de 2009 - Vol.14 - Nº 1. Disponível em:< http://www.polbr.med.br/ano09/cpc0109.php>. Acesso em 19/03/2019.

CHARNES, A.; Cooper, W. W.; RHODES, E. **Measuring The E-ciency Of Decision Making Units**. European Journal Of Operational Research, Piotrowo, V. 2, N. 3, P. 429- 444, 1978.

FADEL, Augusto C.; SEMAAN, Gustavo da S.; BRITO, José A. de M. **Um estudo da aplicação de técnicas de combinação de agrupamentos**. XVII Simpósio de Pesquisa Operacional e Logística da Marinha. Agosto: 2014.

GOMES, E.G.; MELLO, J.C.C.B. S.; ASSIS, A.S.; et al. **Uma medida de eficiência em segurança pública**. Niterói: Relatórios de Pesquisa em Engenharia de Produção, v. 3, n. 7, p. 1-15, 2003. Disponível em < www.producao.uff.br/conteudo/rpep/volume32003/relpesq_303_07.doc>. Acesso em 21/03/2018

HAIR, J. F. et al. (EDS.). **Multivariate data analysis**. 7th. ed. Upper Saddle River, NJ: Prentice Hall, 2010.

HAN, J.; Kamber, M.; Pei, J. **Data Mining**: Concepts and Techniques. 3ª. ed. Morgan Kaufmann Publishers, 2012.

HORA. André B. da. **PANORAMAS SETORIAIS 2030 PAPEL E CELULOSE. 2015.**

Bem.Judite Sanson,Giacomini Nelci Maria Richter, Waismann Moisés. **Utilização da técnica da análise de clusters ao emprego da indústria criativa entre 2000 e 2010:** estudo da Região do Consinos, RS. INTERAÇÕES, Campo Grande, v. 16, n. 1, p. 27-41, jan./jun. 2015.

MALHOTRA, N. **Pesquisa de Marketing: uma orientação aplicada.** Porto Alegre: Bookman, 2001.

MINGOTI, Sueli A. **Análise de dados através de métodos de estatística multivariada:** Uma abordagem aplicada. 2ª impressão. Belo Horizonte: Editora UFMG, 2005.

PESSANHA. José Francisco Moreira. **Análise de Agrupamentos algoritmos e aplicações**. Universidade do Estado do Rio de Janeiro (Uerj). Outubro, 2017.

R Core Team and contributors worldwide. **stats**: R statistical functions. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/

SILVA. Carlos A. F.; BUENO, Jefferson M.; NEVES, Manoel R. **A INDÚSTRIA DE CELULOSE E PAPEL NO BRASIL**. 2016.

THOMAS, Jerry R. e NELSON, Jack K. **Research methods in physical activity**. 3.ed. Champaign: Human Kinetics, 1996.

VALLI, M. **Análise de Cluster**. Augusto Guzzo Revista Acadêmica (São Paulo), v. 04, p. 77, 2002.

Anexo

```
empresas<-read.csv2(file='dendoser.csv')
output_cluster<-hclust(dist(empresas),method='ward.D2')
plot(output_cluster, labels= NULL, hang = -0.1, ylab= "Distância")
abline(h=7000, col="blue")
dendograma_output_cluster<-plclust(output_cluster,labels=objetos,ylab='distancia', hang=-1, main = "Cluster Dendrogram")
output_cluster<-hclust(dist(empresas),method='single')
plot(output_cluster, labels= NULL, hang = -0.1, ylab= "Distância")
abline(h=3500, col="red")
dendograma_output_cluster<-plclust(output_cluster,labels=objetos,ylab='distancia', hang=-1, main = "Cluster Dendrogram")
output_cluster<-hclust(dist(empresas),method='centroid')
plot(output_cluster, labels= NULL, hang = -0.1, ylab= "Distância")
abline(h=2000, col="green")
dendograma_output_cluster<-plclust(output_cluster,labels=objetos,ylab='distancia', hang=-1, main = "Cluster Dendrogram")
```