ISSN: 2317-0840

# O uso da análise de cluster no estudo de características pluviométricas

*Éder Comunello*[1,2†], *Lucio B. Araújo*[3], *Paulo Cesar Sentelhas*[4], *Mirian F. C. Araújo*[3], *Carlos Tadeu S. Dias*[4], *Carlos Ricardo Fietz*[2].

[1] *Programa de Pós-Graduação em Engenharia de Sistemas Agrícolas (USP/Esalq/PPGESA)*
[2] *Embrapa Agropecuária Oeste*
[3] *Universidade Federal de Uberlândia, Faculdade de Matemática (FAMAT/UFU)*
[4] *Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ/USP)*

**Resumo:** *O objetivo deste trabalho foi identificar e caracterizar ambientes homogêneos com base na precipitação observada em Mato Grosso do Sul. Para esse fim, utilizaram-se dados provenientes de estações e postos pluviométricos localizados em Mato Grosso do Sul e entorno. Por meio de análise de agrupamento, empregando um método hierárquico e de aglomeração, baseado na distância euclidiana como medida de dissimilaridade e soma dos quadrados como critério de fusão (método de Ward), foram identificados três ambientes homogêneos em Mato Grosso do Sul. Esses ambientes homogêneos foram caracterizados avaliando-se suas estatísticas básicas e por meio de nova análise de agrupamento, desta vez avaliando o comportamento decendial da precipitação em cada grupo isoladamente.*

**Palavras-chave:** *precipitação; análise multivariada; estatística climatológica.*

**Abstract:** *The aim of this study was identify and characterize homogeneous environments based on precipitation observed in Mato Grosso do Sul State, Brazil. For this, data from different climatic stations located in Mato Grosso do Sul and its surroundings were used. By cluster analysis, using a hierarchical method based on Euclidean distance as dissimilarity measure and sum of squares as a criterion for fusion (Ward´s method) were identified three homogeneous environments in Mato Grosso do Sul. These homogeneous environments were characterized evaluating their basic statistics and using new cluster analysis, this time evaluating the behavior of the decendial rainfall in each group separately.*

**Keywords:** *Precipitation; Multivariate analysis; Climatological statistics.*

## Introdução

Mato Grosso do Sul é uma importante região produtora de grãos no Brasil. Com uma área agricultável ao redor de 2,6 milhões de hectares, esse estado foi responsável por aproximadamente 8%, 7% e 6% da produção brasileira de milho, soja e cana na safra de 2011/2012 (IBGE, 2011). Contudo, a despeito de sua importância e participação no cenário agrícola, essa região é marcada por uma alta variabilidade na produtividade agrícola decorrente, sobretudo, de fatores limitantes ambientais, destacando-se entre eles, o déficit hídrico nas fases críticas da produção. Uma vez que o uso de irrigação ainda é bastante incipiente no estado, há uma forte dependência da água fornecida pelas chuvas.

Ao considerar a agricultura brasileira como um todo, a precipitação pluvial constitui-se no principal fator de risco climático para a atividade agrícola, com estiagens e chuvas excessivas respondendo pela maioria dos sinistros agrícolas (GÖPFERT et al., 1993). Mesmo durante a estação chuvosa, podem ocorrer déficits de precipitação pluvial pontuais e consequentemente, redução na produtividade.

Os danos provocados pelo déficit hídrico podem ser intensificados pela associação com outros fatores climáticos tais como temperatura do ar elevada e baixa umidade do ar. Além disso, as características e condição local do solo interferem nesse agravamento, principalmente em função da capacidade de retenção de água e do potencial erosivo (MISHRA; SINGH, 2010; SALAS et al., 2005).

[†]Autor correspondente: ecomunel@gmail.com.

Sendo assim, há de se desenvolver estratégias de convivência com os períodos de estiagem, sendo que o primeiro passo consiste na análise e determinação de padrões espaço-temporais. A divisão de um território em regiões homogêneas considera quesitos espaço-temporais e contribui diretamente para atenuar o efeito da precipitação pluvial na produção agrícola.

Keller Filho et al. (2005) discorrem sobre a importância da definição de regiões homogêneas para o processo de Zoneamento Agrícola do Brasil, demonstrando como este procedimento auxilia na escolha das culturas de menor risco e para o estabelecimento de datas de plantio mais favoráveis. Outro estudo conduzido por Fernandes et al. (2012) utiliza o conhecimento acerca das regiões homogêneas na definição de estratégias para o desenvolvimento de genótipos adaptados aos diferentes ambientes encontrados. Segundo esses mesmos autores, um modo de melhor quantificar os riscos climáticos e também minimizar as interações genótipo $\times$ ambiente complexas é a identificação e caracterização de ambientes climáticos homogêneos.

Desse modo, pode-se inferir que um estudo dessa natureza seria de grande validade para o Estado de Mato Grosso Sul, motivando a execução deste trabalho que visa identificar e caracterizar ambientes homogêneos com base na precipitação observada, auxiliando no entendimento e formulação de estratégias para atenuação do risco inerente à produção agrícola.

## Material e métodos

A região de estudo compreende o Estado de Mato Grosso do Sul (Figura 1), localizado entre as coordenadas 17°S, 50°W e 25°S,59°W, e com área aproximada de 358.000 Km$^2$, ou 4,2% do território brasileiro. Na maior parte do território do estado predomina o clima do tipo tropical ou tropical de altitude, com chuvas de verão e inverno seco, caracterizado por médias termométricas que variam entre 20°C 25°C. No extremo meridional, com maior latitude e relevo de planalto, ocorre o clima subtropical.

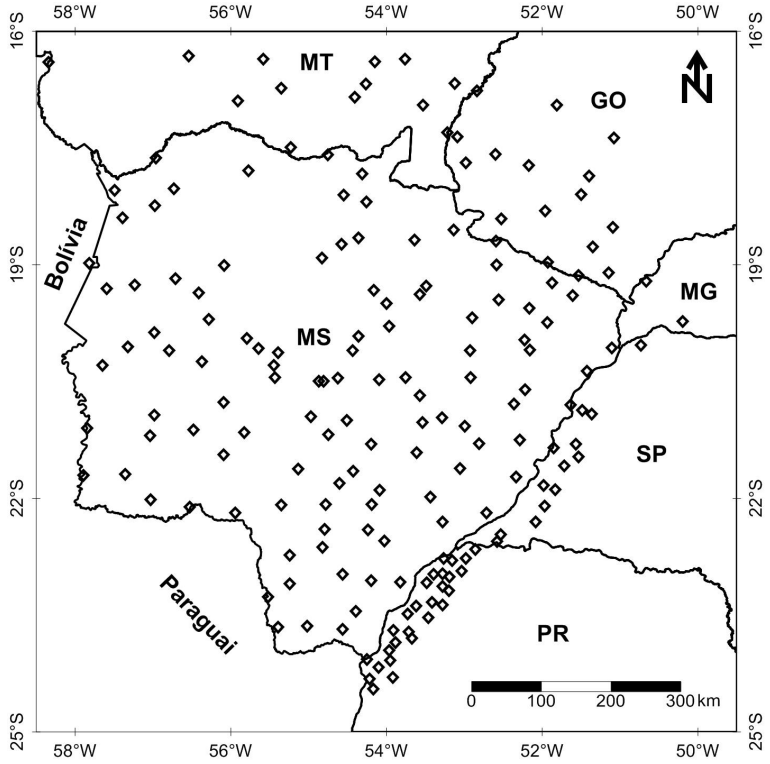


Figura 1: Mapa da área de estudo com estações e postos pluviométricos utilizados (pontos).

A definição de regiões homogêneas implica em uma técnica de classificação de objetos em categorias similares e um dos métodos mais utilizados para este fim é a análise de agrupamentos (cluster). Esse

método estatístico considera um conjunto inicial de objetos, aos quais são associadas variáveis classificatórias, que são medidas utilizadas para se obter grupos de objetos assemelhados em relação aos valores assumidos por essas variáveis (EVERITT; DUNN, 2011).

A análise de agrupamentos já é uma técnica de aplicação comum na climatologia para a definição de regiões climáticas homogêneas, tendo sido empregada em diversos estudos (FERNANDES et al.; 2012; KELLER FILHO et al., 2005; MUNÕZ-DIAZ; RODRIGO, 2003; UNAL et al., 2003; UVO, 2003).

Um ponto chave para a aplicação do método é a escolha das variáveis classificatórias, a qual irá refletir o julgamento do investigador sobre a relevância dos fatores climáticos para os propósitos da pesquisa. Embora a precipitação pluvial seja influenciada por fatores físico-geográficos, tais como posição, extensão latitudinal e relevo, para efeitos desse estudo considera-se apenas a precipitação pluvial observada, organizada sob a forma de decêndios (períodos de dez dias) e assim buscando explorar características pertinentes à intensidade de precipitação destes períodos, bem como seu comportamento temporal.

O agrupamento dos dados em decêndios é uma medida comum em estudos climáticos, e além de suavizar o efeito de erros nas observações, pode ser encarada também como uma estratégia de redução da dimensionalidade. Os dados anuais, inclusive aqueles oriundos de anos bissextos, são padronizados em 36 períodos, cada qual considerado como uma variável classificatória.

Para a composição destas variáveis foram utilizados dados da precipitação pluvial diária obtidos da Agência Nacional de Águas (ANA), inicialmente referentes a 238 postos pluviométricos. Além de postos localizados no território sul-matogrossense, incluíram-se postos do entorno, visando assegurar continuidade espacial e conexão com estudos realizados em áreas subjacentes. Inicialmente, os dados foram avaliados quanto à atualidade, ausência de erros, completude e consistência. Desta análise inicial, foram selecionados 181 postos pluviométricos com série mínima de 20 anos, atualizações a partir de 2002, e proporção de falhas inferior a 1/3 do período de observação. As séries históricas de cada posto selecionado foram utilizadas para o cálculo de decêndios, atribuindo-se o valor nulo à medida no caso de dados faltantes. Considerando apenas decêndios completos, obteve então a média decendial para cada estação. Por fim, os dados foram organizados de forma a compor uma matriz com dimensão de 181 linhas (estações) por 36 colunas (decêndios). Adicionalmente foram ainda organizados os identificadores e informações gerais e geográficas de cada observação (postos).

A essa matriz aplicou-se o método de agrupamento hierárquico e de aglomeração, com base na distância euclidiana como medida de dissimilaridade e soma dos quadrados como critério de fusão (EVERITT; DUNN, 2011; WARD, 1963). De acordo com Lund et al. (2009) este método possui ótimo desempenho para uma grande quantidade de dados climáticos, a despeito de sua relativa simplicidade. Além disso, diversos autores utilizaram e recomendaram esse método para determinar ambientes climatologicamente homogêneos. (FERNANDES et al.; 2012; KELLER FILHO et al., 2005; MUNÕZ-DIAZ; RODRIGO, 2003; UNAL et al., 2003; UVO, 2003; BALDO et al., 2000).

A definição da medida de similaridade ou de distância a ser empregada é fundamental na análise de agrupamentos. A distância Euclidiana tem propriedades métricas é a mais utilizada para variáveis classificatórias reais e medidas em uma escala de intervalo (EVERITT; DUNN, 2011). Para variáveis classificatórias tomadas com unidades distintas recomenda-se a padronização. Uma vez que as variáveis utilizadas neste estudo são tomadas na mesma unidade, considerou-se desnecessário o procedimento.

O método de agrupamento hierárquico difere dos não-hierárquicos, por não produzir um número fixo de agrupamentos, mas sim formá-los por meio de uma sequência crescente de partições ou junções sucessivas de grupos (abordagem aglomerativa). O método hierárquico aglomerativo é o mais utilizado na construção de agrupamentos (KAUFMAN; ROUSSEAU, 1990).

Além do método de Ward, às vezes referido como método da variância mínima, as técnicas de agrupamento hierárquico mais utilizadas são a ligação simples (*single linkage method*) e a ligação completa (*complete linkage method*). Tal como realizado neste estudo, a opção pela técnica é, de certo modo, subjetiva e realizada com base em vários estudos empíricos (KELLER FILHO et al., 2005).

Ainda segundo Keller Filho et al. (2005), no método de Ward, formação dos agrupamentos em cada estágio da hierarquia é avaliada pela soma dos quadrados dos desvios em relação ao centro de gravidade dos grupos, geralmente indicada por R2. O critério para a fusão de cada par de grupos é o de que seja

obtido o menor acréscimo possível no valor de R2.

Formados os grupos pela aplicação da técnica de agrupamento, o passo seguinte implica na definição de grupos. Não existe um método inteiramente satisfatório para a determinação do número ideal de grupos (HARTIGAN, 1985) e sua definição é, via de regra empírica, valendo-se, sobretudo, da experiência e habilidade do pesquisador em vislumbrar os agrupamentos e explica-los. Como sugestão de partida pode-se iniciar a análise tomando os agrupamentos formados na metade da maior distância do dendrograma (FERNANDES et al., 2012).

## Resultados e discussão

A etapa de preparação dos dados envolveu a avaliação de cada um dos 238 postos pluviométricos inicialmente disponíveis. Cada posto foi avaliado quanto à atualidade, ausência de erros, completude e consistência dos dados. Ao final desta etapa, foram selecionados 181 postos pluviométricos com série mínima de 20 anos, atualizações a partir de 2002, e proporção de falhas inferior a 1/3 do período de observação.

A matriz de dados submetida à análise foi formada por 181 observações avaliadas (postos pluviométricos) por 36 variáveis classificatórias, compreendidas pela média histórica decendial de cada posto. Inicialmente a matriz de dados foi analisada utilizando a técnica de componentes principais, sendo possível explicar 80,3% da variação dos dados com base nos quatro primeiros componentes principais. A técnica da ACP facilitou o entendimento do conjunto de dados e demonstrou a relação entre variáveis. Na etapa seguinte aplicou-se a técnica de análise de agrupamentos, utilizando a métrica Euclidiana no procedimento de Ward e cujo resultado da análise de agrupamentos pode ser observado na Figura 2.

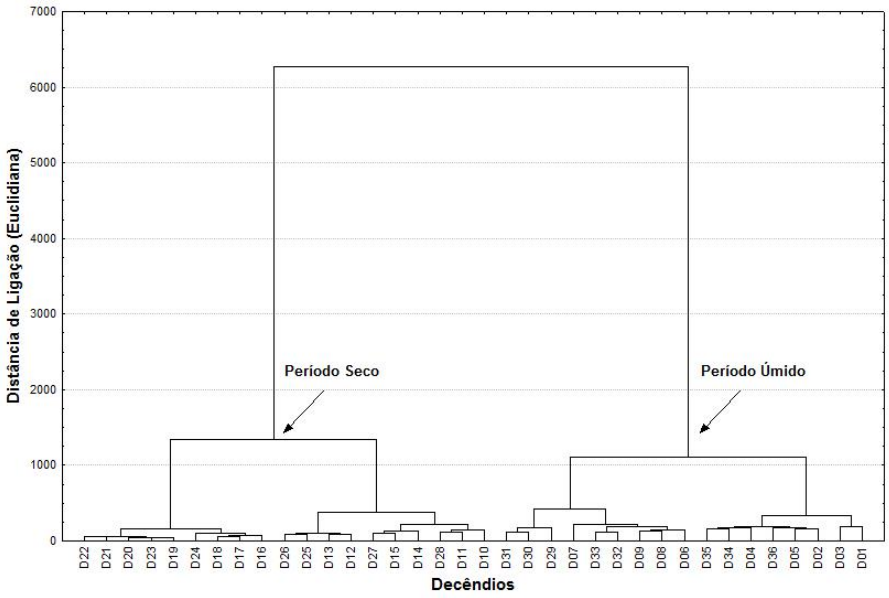


Figura 2: Dendrograma avaliando 36 decêndios, resultante do método de agrupamento hierárquico e de aglomeração, com base na distância euclidiana e soma dos quadrados como critério de fusão.

De acordo com a Figura 2, fica nítida a distinção de dois períodos, podendo ser interpretados como duas épocas homogêneas, relativas a um período seco e um período úmido ou chuvoso. Esta análise leva em consideração os montantes pluviométricos em cada decêndio, bem como sua distribuição temporal, considerando todas as observações. Todavia deve-se verificar se esta interpretação pode ser generalizada para todo o conjunto de observações. O resultado de uma nova análise de agrupamentos, buscando avaliar as observações no lugar das variáveis é demostrada na Figura 3.

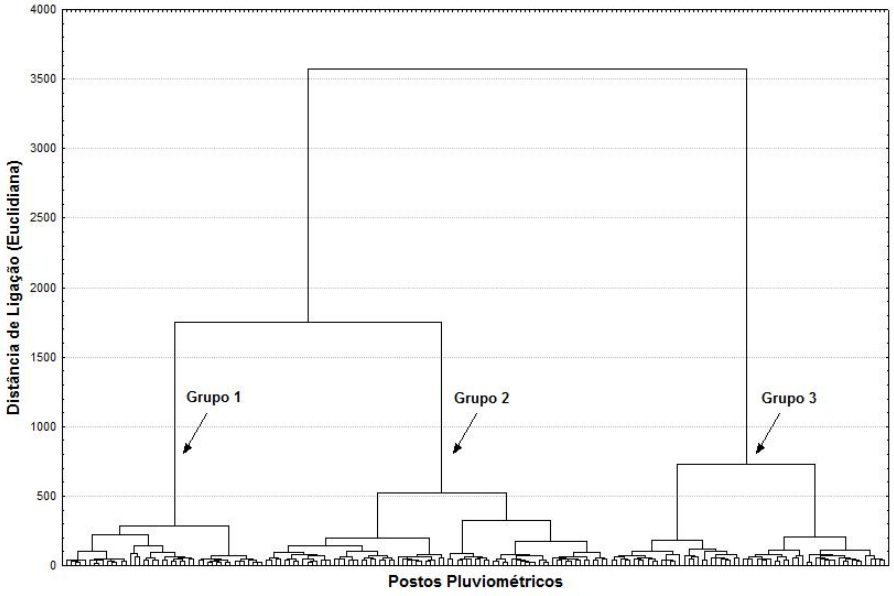


Figura 3: Dendrograma avaliando os 181 postos pluviométricos, resultante do método de agrupamento hierárquico e de aglomeração, com base na distância euclidiana e soma dos quadrados como critério de fusão

Embora a generalização de duas épocas distintas de precipitação proporcionadas pela primeira análise, a segunda análise revela a existência de pelo menos três ambientes distintos na área de estudo. O critério empírico de se utilizar como limiar de corte a metade da maior distância é pouco útil neste caso, coincidindo com o limiar para a formação de três ou quatro grupos. Optou-se pela formação de três grupos cuja consistência foi examinada verificando a coerência espacial da divisão. Sendo a precipitação uma variável contínua no espaço, é esperado que ela se apresente agrupada tal como obtido na Figura 4.

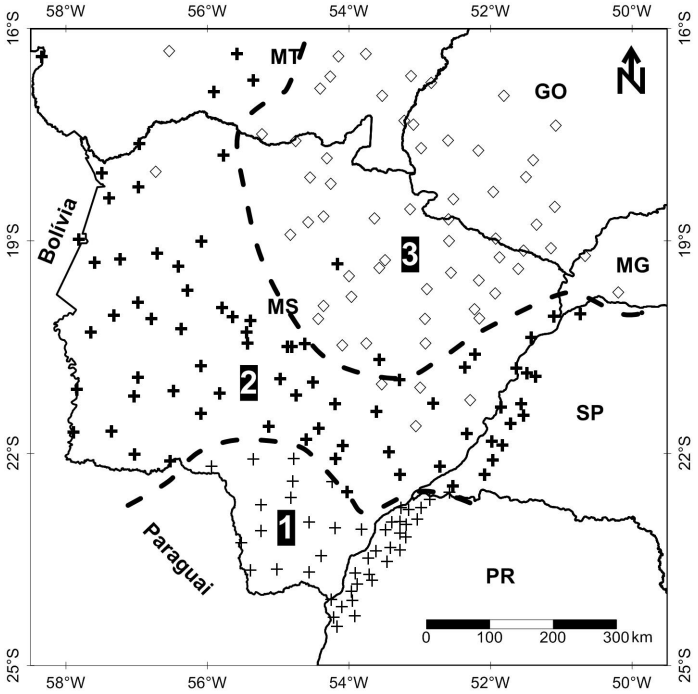


Figura 4: Representação espacial dos grupos homogêneos formados pela análise de cluster.

Uma vez que a representação dos resultados do método hierárquico da variância mínima em três grupos foi considerada coerente, passou-se a designar as três áreas como Regiões Homogêneas (RH) RH1 - Sul de MS, RH2 - Corredor Central de MS e RH3 - Norte de MS, cada uma com, respectivamente, 44, 76 e 61 observações. A matriz de dados foi subdividida, de modo a calcular as estatísticas básicas representativas de cada grupo (Tabela 1).

Tabela 1: Médias decendiais e totais anuais para cada região homogênea determinada.

| **Dec** | **RH1** | **RH2** | **RH3** | **Dec** | **RH1** | **RH2** | **RH3** |
|---|---|---|---|---|---|---|---|
| | ***n=44*** | ***n=76*** | ***n=61*** | | ***n=44*** | ***n=76*** | ***n=61*** |
| D01 | 49.10 | 65.15 | 93.85 | D19 | 19.36 | 9.08 | 3.85 |
| D02 | 61.98 | 63.74 | 81.14 | D20 | 19.92 | 8.68 | 4.67 |
| D03 | 57.83 | 66.69 | 98.04 | D21 | 17.53 | 10.09 | 7.14 |
| D04 | 50.88 | 53.71 | 77.55 | D22 | 20.43 | 10.58 | 4.96 |
| D05 | 54.05 | 57.92 | 79.23 | D23 | 20.87 | 8.06 | 4.62 |
| D06 | 43.22 | 44.61 | 59.75 | D24 | 22.28 | 13.38 | 13.27 |
| D07 | 33.26 | 48.88 | 72.75 | D25 | 36.10 | 20.91 | 19.45 |
| D08 | 39.21 | 44.79 | 66.84 | D26 | 31.70 | 18.52 | 17.69 |
| D09 | 40.05 | 41.44 | 61.04 | D27 | 46.65 | 30.08 | 28.77 |
| D10 | 34.44 | 30.45 | 39.55 | D28 | 46.68 | 31.88 | 32.07 |
| D11 | 44.21 | 30.97 | 33.10 | D29 | 60.76 | 37.86 | 41.73 |
| D12 | 34.80 | 22.43 | 18.71 | D30 | 55.53 | 44.89 | 51.49 |
| D13 | 35.89 | 20.97 | 16.13 | D31 | 57.11 | 46.01 | 53.16 |
| D14 | 46.86 | 24.56 | 17.60 | D32 | 46.73 | 49.91 | 60.97 |
| D15 | 43.08 | 30.21 | 23.88 | D33 | 48.43 | 44.81 | 62.09 |
| D16 | 30.35 | 15.83 | 8.66 | D34 | 55.36 | 58.88 | 74.54 |
| D17 | 25.29 | 13.18 | 7.25 | D35 | 56.38 | 53.78 | 73.91 |
| D18 | 26.97 | 10.18 | 5.39 | D36 | 56.27 | 62.88 | 82.28 |
| **Total** | | | | | **1469.53** | **1245.98** | **1497.15** |

Uma etapa auxiliar, útil para caracterização dos grupos foi realizada empregando-se a análise de agrupamentos considerando os decêndios em cada grupo separadamente. Os resultados são apresentados nas Figuras 5, 6 e 7.

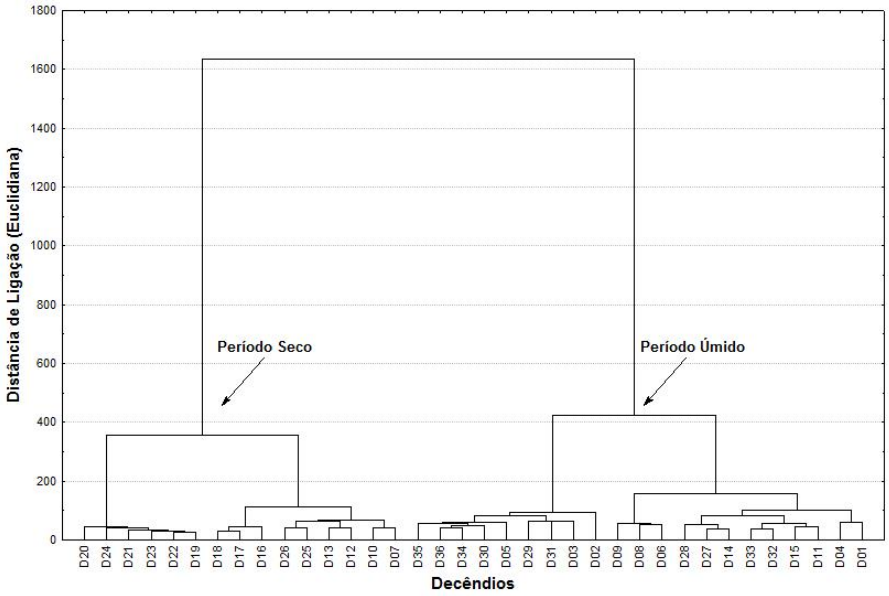


Figura 5: Dendrograma avaliando 36 decêndios, referente à RH1 – Sul de MS.

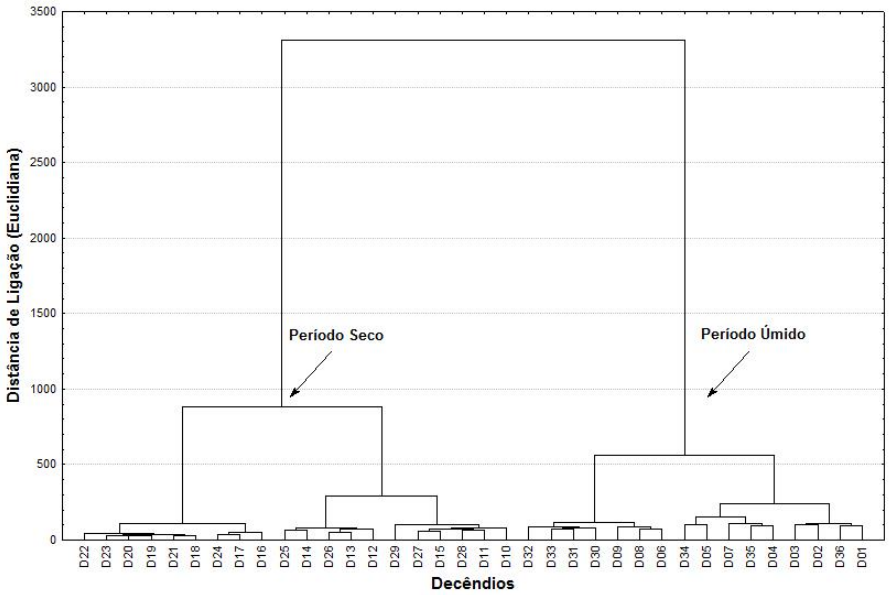


Figura 6: Dendrograma avaliando 36 decêndios, referente à RH2 – Corredor Central de MS.

Apesar de todos dendrogramas apresentarem nítida distinção entre os períodos seco e úmido, o agrupamento dos decêndios difere, sendo úteis para distinção das regiões entre si. A RH1 tem período úmido maior que as demais áreas, agrupando nesta fase dos decêndios que seguem desde D27 até D11, ou seja, distribuição da precipitação ao longo de 21 decêndios. As duas outras áreas tem comportamento similar, com distribuição das chuvas durantes apenas 17 meses (D30 até D9).

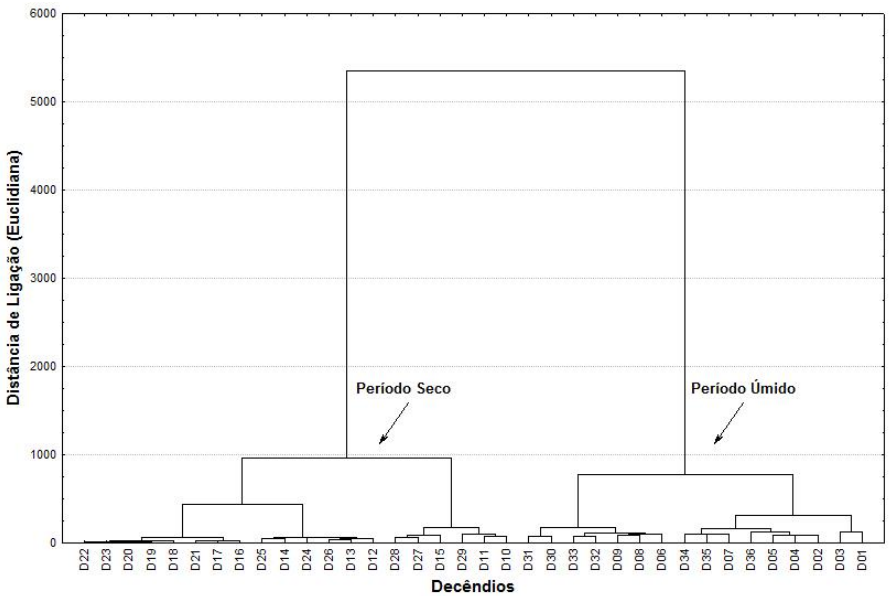


Figura 7: Dendrograma avaliando 36 decêndios, referente à RH3 – Norte de MS.

Avaliando os dendrogramas das Figuras 5 a 7 e os dados da Tabela 1, expressos na Figura 8, pode-se caracterizar as três regiões. Inicialmente temos que a RH2 é a que tem menor disponibilidade hídrica anual (1245,98 mm), embora seu período úmido (D30-D9) seja o mesmo da RH3. A RH3, por sua vez, tem disponibilidade hídrica (1497,15 mm) muito similar ao da RH1 (1469,53 mm), mas se diferencia

muito em relação à distribuição desta chuva. A RH1 apresenta o maior período úmido (D27-D11), sem implicar, no entanto, em maiores montantes pluviométricos.

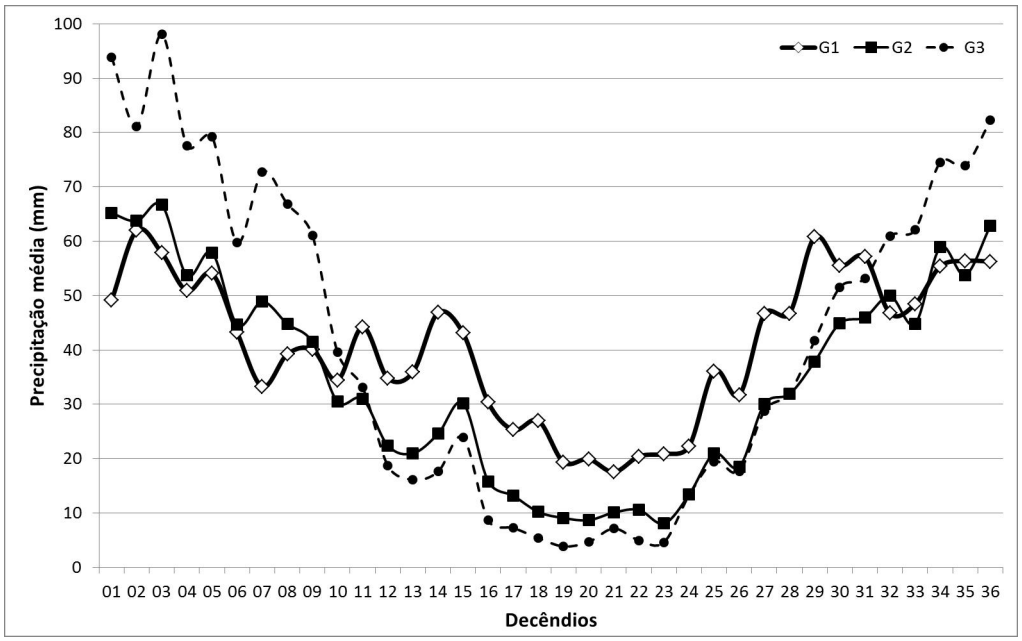


Figura 8: Perfil médio das regiões homogêneas RH1, RH2 e RH3.

A RH3 é potencialmente a de menor risco climático, por concentrar os maiores montantes pluviométricos ao longo de um prazo menor do que observado na RH1. A agricultura praticada na RH1 deve avaliar precisamente os momentos críticos do cultivo de interesse, buscando ajustá-los com a oferta hídrica disponível.

## Conclusões

A análise de agrupamento hierárquica é um instrumento adequado na identificação de zonas homogêneas tomando por base dados pluviométricos diários. A utilização de variáveis classificatórias definidas por meio da formação de decêndios permite, de forma eficiente, formar grupos similares quanto ao regime de chuva. A análise hierárquica de agrupamento permitiu três regiões pluviometricamente homogêneas localizadas em território sul-matogrossense.

## Referências

BALDO, M.C.; ANDRADE, A.R.; MARTINS, M.L.O.F; NERY, J.T. Análise de precipitação pluvial do Estado de Santa Catarina associada com a anomalia da temperatura da superfície do oceano Pacífico. *Revista Brasileira de Agrometeorologia*, v.8, p.283-293, 2000.

EVERITT, B. S.; DUNN, G. *Applied Multivariate Analisys*. 2rd ed. London: Wiley, 2011, 342p.

FERNANDES, D. S.; Kruger, L. F.; HeinemannI,A. B.; Rocha, R. P. Identificação e caracterização de ambientes homogêneos de eventos de seca/umidade com base em simulações climáticas regionais. *Bragantia*, Campinas, v.71, p.290-298 2012.

GÖPFERT, H.; ROSSETTI, L. A.; SOUZA, J. *Eventos generalizados e seguridade agrícola. Brasília: IPEA*, 1993, 65p.

HARTIGAN, J.A. Statistical theory in clustering. *Journal of Classification*, v.2, p.63-76, 1985.

IBGE. Sistema IBGE de recuperação automática - SIDRA: Banco de Dados Agregados: Dados de previsão de safra: produção - hectare - Brasil e Mato Grosso do Sul - setembro de 2011. [Rio de Janeiro, 2011]. URL: http://www.sidra.ibge.gov.br/bda/prevsaf/default.asp?t=3&z=t&o....

KAUFMAN, L.; ROUSSEAU, W. *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley & Son, 1990, 342 p.

KELLER-FILHO, T.; ASSAD, E. D.; LIMA, P. R. S. R. Regiões pluviometricamente homogêneas no Brasil. *Pesqui. Agropecu. Bras*, Brasília, v.40, p.311-322, 2005.

LUND, R.; LI, B. Revisiting climate region definitions via clustering. *Journal of Climate*, v.22, p.1787-1800, 2009.

MISHRA, A. K.; SINGH, V. P. A review of drought concepts. *Journal of Hydrology*, v.391, p.202-216, 2010.

MUÑOZ-DIAZ, D.; RODRIGO, F.S. The North Atlantic oscillation and winter rainfall over the Siberian Peninsula as captured by cluster analysis. *Geophysical Research Abstracts*, v.5, p.865-885, 2003.

SALAS, J. D.; FU, C.; CANCELLIERE, A.; DUSTIN, D.; BODE, D.; PINEDA, A.; VINCENT, E. Characterizing the severity and risk of drought in the Poudre River. *Journal of Water Resources Planning and Management*, Colorado, v.131, p.383-393, 2005.

UNAL, Y.; KINDAP, T.; KARACA, M. Redefining the climate zones of Turkey using cluster analysis. *International Journal of Climatology*, v.23, p.1045-1055, 2003.

UVO, B.C. Analysis and regionalization of the Northern European winter precipitation based on its relationship with the North Atlantic oscillation. *International Journal of Climatology*, v.23, p.1185-1194, 2003.

WARD, J.H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, v.58, p.236-244, 1963.