ANÁLISE DE AGRUPAMENTO DE DADOS HIDROLÓGICOS: UMA REVISÃO DA LITERATURA

Kássia Regina Bazzo [1] & Júlio Gomes [2]*

Resumo – O presente artigo apresenta uma revisão de literatura sobre a utilização da análise de agrupamentos em estudos nas áreas de hidrologia e recursos hídricos. Um grande problema nos estudos hidrológicos é a escassez de dados provenientes de monitoramento. Nesse sentido, a mineração de dados apresenta-se como uma alternativa promissora para extrair maiores informações sobre os dados existentes para subsidiar decisões relativas ao planejamento e gestão dos recursos hídricos. A análise de agrupamento foi usada por diversos autores, principalmente na escala espacial, visando à obtenção de regiões hidrologicamente homogêneas. A análise dos estudos anteriores permitiu concluir que a performance dos diferentes métodos de agrupamento depende largamente das características da região a ser estudada e da definição dos parâmetros de entrada: número de agrupamentos (c), índice de difusividade (m) e grau de tolerância (δ). Verificou-se também que os métodos difusos foram utilizados e avaliados em um maior número de vezes e apresentaram-se mais adequados para a aplicação em recursos hídricos. Finalmente, constatou-se que é fundamental a avaliação da homogeneidade dos grupos para garantir a qualidade dos resultados obtidos.

Palavras-Chave – mineração de dados; análise de agrupamentos; *fuzzy c-means*.

CLUSTER ANALYSIS OF HYDROLOGICAL DATA: A LITERATURE REVIEW

Abstract - This paper presents a literature review on the use of cluster analysis in hydrology and water resources fields. A major problem in hydrological studies is the scarcity of data from monitoring. In this sense, data mining presents a promising alternative to extract more information about existing data to support decisions related to water resources planning and management. Several authors have been used cluster analysis, mainly in the spatial scale, to define hydrologically homogeneous regions. The analysis of the previous studies allowed us to conclude that the performance of different clustering methods depends largely on both the characteristics of the studied region and on the definition of input parameters: number of clusters (c), diffusivity index (m) and degree of tolerance (δ). Also, it was possible to conclude that diffuse methods were used and evaluated in a greater number of times, and they were indicated as more suitable for application in water resources. Finally, it was verified that evaluation of homogeneity of the groups is fundamental to guarantee the quality of the achieved results.

Keywords – Data mining; cluster analysis; fuzzy c-means.

[1] Eng. Sanitarista e Ambiental. Mestranda, Programa de Pós-graduação em Engenharia de Recursos Hídricos e Ambiental – PPGERHA/UFPR
[2] Eng. Civil, Dr., Professor, Programa de Pós-graduação em Engenharia de Recursos Hídricos e Ambiental – PPGERHA/UFPR

INTRODUÇÃO

A escassez de dados de monitoramento hidrológico é um fato conhecido no Brasil. Enquanto essa realidade não é mudada, cabe aos pesquisadores a aplicação e a análise de técnicas que possibilitem explorar ao máximo os dados *in loco* existentes, visando extrair maiores informações e expandir o leque de conhecimento relacionado aos eventos hidrológicos. As técnicas de mineração de dados e da estatística multivariada por meio da análise de agrupamentos vêm sendo amplamente utilizadas com este objetivo e se destacam por viabilizar a análise conjunta das diversas variáveis, o que as tornam promissoras neste ramo do conhecimento.

A análise de agrupamentos refere-se ao processo de categorizar uma série de dados dispersos em grupos (ou conjuntos) com características que se assemelham entre si e que os diferem dos demais grupos (Chen e Wang, 2012). Em hidrologia, na escala espacial, diversos autores utilizaram esta técnica para: identificação de regiões hidrologicamente homogêneas de precipitações (Boschi *et al*., 2011; Uda *et al*., 2015); obtenção de informações em áreas sem monitoramento por meio da regionalização de vazões (Gibertoni *et al*., 2004; Elesbon, 2012) e de curvas de permanência (Pessoa, 2015); e análise regional de frequência (Rao e Srinivas, 2006a, 2006b; Carvalho, 2007; Basu e Srinivas, 2016, 2014). Na escala temporal, a técnica é usada no agrupamento de eventos extremos com características altamente homogêneas internamente (Wang *et al*., 2015).

Em qualquer área da pesquisa científica, previamente à aplicação direta das técnicas apresentadas na literatura, compete aos pesquisadores a revisão bibliográfica e a análise dos resultados obtidos por outros autores que utilizaram a mesma técnica de interesse. Tal procedimento serve como auxílio na definição dos caminhos mais adequados a serem adotados para a tomada de decisão em relação à hipótese inicialmente formulada, ou seja, na definição do método da pesquisa.

O presente trabalho tem por objetivo apresentar uma revisão da literatura sobre as técnicas de análise de agrupamentos mais utilizadas no âmbito da hidrologia, assim como avaliar o potencial da sua utilização. Estes objetivos são alcançados por meio da discussão da temática e do levantamento dos estudos já realizados que visaram identificar agrupamentos homogêneos, tanto na escala temporal (categorização de eventos extremos), quanto na escala espacial (identificação de regiões hidrologicamente homogêneas).

ANÁLISE DE AGRUPAMENTO

A mineração de dados (*data mining*) é um processo no qual são aplicados métodos inteligentes para extrair padrões a partir dos dados observados (Han *et al*., 2012). Segundo os mesmos autores, a análise de agrupamentos (*cluster analysis*), um dos métodos usados na mineração de dados, visa realizar o agrupamento de dados, promovendo a maximização da similaridade dentro do mesmo grupo e a minimização da similaridade entre grupos diferentes.

Dentre as diversas classificações das técnicas de agrupamentos, no âmbito deste estudo, cabe destacar as classes hierárquicas e não hierárquicas ou particionais (Han *et al.*, 2012; Tan *et al.*, 2005). As técnicas hierárquicas fazem sucessivos agrupamentos, diminuindo ou aumentando o número de grupos (*clusters*) ao longo da análise. Está técnica é caracterizada por utilizar dendogramas para a definição do número de grupos de maneira subjetiva. Os principais métodos são: *single linkage; complete linkage;* e de *Ward*. Já as técnicas não hierárquicas supõem que o número de grupos é conhecido, sendo, portanto, um parâmetro de entrada. Os elementos são alocados ao grupo com a similaridade mais próxima de maneira iterativa e os principais métodos são: *k-means; c-means;* e *fuzzy c-means*, considerado uma generalização do método *c-means* (Carvalho, 2007).

Boschi *et al.* (2011) utilizaram o algoritmo particional *k-means* para a definição de áreas pluviométricas homogêneas em 2 decênios diferentes no Rio Grande do Sul. Os autores estabeleceram que o número ideal de grupos estaria contido em um intervalo pré-definido e compararam os resultados entre si para a obtenção do valor ótimo. Após a sobreposição das áreas homogêneas identificadas para os 2 períodos, foram identificadas 6 regiões comuns aos decênios e verificaram incrementos significativos na precipitação anual em 5 destas 6 regiões.

Elesbon (2012) avaliou 3 técnicas de agrupamento hierárquico de bacias hidrográficas como suporte à regionalização de vazões mínimas, médias e máximas na bacia do Rio Doce. O autor utilizou os métodos do vizinho mais próximo (*single linkage*), vizinho mais distante (c*omplete linkage*) e o método de *Ward* para agrupar vazões de 61 estações fluviométricas na bacia em estudo, e propôs um método para a identificação do número de regiões homogêneas baseado na análise gráfica. Para o método *single linkage*, os resultados foram descartados por apresentarem agrupamento irregular. Para o método c*omplete linkage*, foram obtidas 4 regiões homogêneas, com 45 estações sendo agrupadas em uma única região. Para o método de *Ward*, o número de agrupamentos não foi idêntico para as diferentes vazões avaliadas (mínimas, médias e máximas). O autor não identificou o motivo deste comportamento hidrológico diferenciado, descartando o método. Considerando os resultados obtidos, sugeriu-se o método c*omplete linkage* como o método mais adequado para o agrupamento de regiões hidrologicamente homogêneas. Cabe ressaltar que não foram realizados testes estatísticos para a comparação dos resultados de homogeneidade obtidos em cada um dos métodos, inviabilizando a inferência estatística sobre qual a melhor técnica.

Pessoa (2015) comparou os métodos de análise de agrupamento hierárquico *Ward* e não hierárquico *fuzzy c-means* na obtenção de regiões homogêneas na Amazônia legal para a regionalização de curvas de permanência de vazões. O índice de desempenho usado para avaliar os métodos foi o Coeficiente de Nash-Sutcliffe. Os resultados obtidos indicaram o método *fuzzy c-means* como o mais indicado para a região em estudo.

Segundo Andrade (2004), a qualidade das técnicas hierárquicas pode ser prejudicada devido à impossibilidade de se realizar ajustes de fusão e divisão de grupos, além de não permitir a troca de elementos entre os grupos. Para a obtenção de melhores resultados na segmentação, é sugerida a integração de técnicas hierárquicas e não hierárquicas, cujo processamento é iterativo. Com esse intuito, alguns estudos utilizaram a combinação de dois métodos para obter resultados mais satisfatórios (Rao e Srinivas, 2006a; Uda *et al.*, 2015)

Rao e Srinivas (2006a) aplicaram uma técnica de agrupamento particional para refinar os resultados obtidos por técnicas de agrupamento hierárquicas na identificação de regiões homogêneas em bacias hidrográficas de Indiana, EUA, através da utilização de três algoritmos híbridos. Os três métodos hierárquicos utilizados foram *single linkage*, *complete linkage* e *Ward*, enquanto o método não hierárquico foi o *k-means*. O algoritmo que apresentou melhores resultados foi a combinação dos métodos *Ward* e *k-means*, sendo recomendado para estudos de regionalização na região.

Uda *et al.* (2015) usaram os métodos hierárquico *complete linkage* e não hierárquico *k-means* para a análise de agrupamento espaço-temporal da precipitação na bacia hidrográfica do Rio Iguaçu. O primeiro método foi utilizado para a identificação dos *outliers* e do número ideal de grupos através da análise subjetiva do dendograma resultante. Escolheu-se um elemento de cada grupo, designado como *semente*, para a aplicação do método não hierárquico, usado para refinar o agrupamento inicial. Segundo os autores, a utilização conjunta dos dois métodos apresentou bons resultados.

TEORIA DIFUSA NA ANÁLISE DE AGRUPAMENTO

Na lógica difusa, os conceitos não são precisos, mas aproximados, sendo, portanto, adequados na área de recursos hídricos devido às incertezas associadas aos processos envolvidos, à falta de dados observados e aos conflitos existentes nas tomadas de decisão (Gibertoni *et al.*, 2004). Segundo Basu e Srinivas (2014), um agrupamento é dito difuso quando cada elemento pode pertencer a mais de um grupo simultaneamente, corroborando com o cenário real.

Dentre os principais métodos utilizados tem-se o *fuzzy c-means* e as suas variações. As primeiras aplicações da teoria difusa para análise de agrupamento por meio do método *fuzzy c-means* foram realizadas por Ruspini (1969). Os trabalhos de Bedzek (1974, 1981) e Dunn (1974) tornaram-se um marco, nos quais o reconhecimento de padrões tornou-se claramente estabelecido. No conjunto difuso de dados, a função analisada é a *função de pertinência*, cujos valores assumidos (graus de pertinência) não são binários, mas sim contidos no intervalo [0, 1] (Gibertoni, *et al.*, 2004; Han *et al.*, 2012), sendo que a soma das pertinências de conjuntos adjacentes é igual a uma unidade.

O algoritmo *fuzzy c-means* atua de forma iterativa, baseado na minimização de uma função objetivo. Esta função relaciona o grau de pertinência do elemento a um determinado grupo, a distância da similaridade entre o elemento e o centro do grupo e o índice de difusividade, que é um parâmetro com valor maior que 1 que controla a difusividade no processo de classificação. A cada iteração, o grau de pertinência e o centro do agrupamento são ajustados, e deve-se definir de uma tolerância para estabelecer o fim das iterações. Resumidamente, para o uso do método, tem-se como parâmetros de entrada: número de agrupamentos (c), o índice de difusividade (m) e o grau de tolerância (δ).

Recentemente, Wang *et al.* (2015) utilizaram o método *fuzzy c-means* para o agrupamento de inundações anuais observadas em uma estação fluviométrica no Rio Wuijiang, China, a partir da definição de indicadores de severidade. Os parâmetros de entrada c, m e δ foram definidos como iguais a 4, 2 e 10^{-5}, respectivamente. O algoritmo usado gerou uma matriz que associava a cada ano da série de dados analisada o grau de pertinência aos 4 grupos. Cada ano foi então considerado como pertencente ao grupo para o qual apresentava o maior grau de pertinência. A partir dos resultados obtidos, os autores categorizaram as cheias como: catastróficas, altas, normais e pequenas. Embora, o algoritmo tenha sido efetivo para avaliar a intensidade das inundações, não foram discutidas as incertezas relacionadas aos agrupamentos realizados, nem à definição dos parâmetros de entrada.

Na escala espacial, Basu e Srinivas (2016) aplicaram uma variação da técnica difusa de agrupamento não hierárquico de bacias hidrográficas baseada na entropia para a regionalização da análise de frequência de cheias nos quatro maiores rios da Índia. Os resultados foram comparados às técnicas de agrupamento *K-means clustering* (GKM) e Região de Influência (ROI). A partir da avaliação do tamanho e da homogeneidade dos grupos formados, foram definidos valores ótimos para os parâmetros de entrada. Verificou-se que o agrupamento de bacias hidrográficas baseado na entropia produziu resultados mais precisos em comparação às abordagens GKM e ROI e as técnicas utilizadas permitiram a obtenção de quantis de cheias confiáveis para as bacias hidrográficas sem monitoramento.

Basu e Srinivas (2014) propuseram uma abordagem alternativa de agrupamento difuso baseada nas funções núcleo (*Kernel functions*) para a análise da frequência de cheias regionais e obtenção de quantis de cheia em bacias hidrográficas sem monitoramento. As funções núcleo foram utilizadas por permitirem mapear os dados em um espaço multidimensional relativamente maior que o original e propiciar a separação linear dos dados devido aos limites dos agrupamentos possuírem uma forma regular, enquanto, no espaço original, os limites dos agrupamentos são formados arbitrariamente. As abordagens propostas, tradicional e alternativa, foram aplicadas a bacias hidrográficas no estado de

Ohio, Estados Unidos, e verificou-se uma melhoria na estimativa dos quantis de cheias na abordagem alternativa em comparação ao f*uzzy c-means* convencional. A homogeneidade dos grupos formados utilizando-se o *kernel fuzzy c-means* e o convencional *fuzzy c-means* foi comparada por meio da medida de Heterogeneidade e da precisão na estimativa de quantis de cheias para locais sem monitoramento. Os resultados mostram menores valores de heterogeneidade e maior precisão na obtenção dos quantis de cheia na abordagem alternativa (*kernel fuzzy c-means*) em comparação à abordagem tradicional (*fuzzy c-means*), evidenciando, portanto, um melhor desempenho da abordagem alternativa.

No âmbito nacional, Gibertoni *et al.* (2004) propuseram uma análise difusa por meio do uso do método *fuzzy c-means*, juntamente com a regressão por metas difusas, para a regionalização de vazões médias diárias e vazões máximas anuais em bacias hidrográficas de pequeno e médio porte no Paraná. Para a definição do número de grupos (*c*) ideal, os autores avaliaram a consequência hidrológica de aumentar o número de agrupamentos de *c* para (*c* + 1), representada pela variação dos parâmetros hidrológicos a serem regionalizados dentro dos agrupamentos. O valor do índice de difusividade (*m*) foi definido como igual a 2, valor máximo indicado por Ross (1995), citado por Hall e Minn (1999), representando um alto nível de incerteza a respeito da ambiguidade dos dados. Os resultados demonstraram que a análise difusa proposta (*fuzzy c-means* + regressão por metas difusa) apresentou melhores estimativas da vazão média em relação aos métodos *K-means* e *fuzzy c-means*. Embora a utilização da regressão difusa tenha melhorado a precisão na estimativa das vazões, segundo os autores, a sua aplicação não apresentou bom desempenho em locais que não foram utilizados na construção do modelo, tornando-se uma limitação no âmbito da regionalização, cujo objetivo é a obtenção de dados em locais sem monitoramento. Os autores sugerem a atribuição de valores menores do coeficiente *m*, caso a região estudada apresente características fisiográficas, meteorológicas e fluviais bastante distintas entre si, ou características fluviais que possuam forte dependência das características fisiográficas.

Índices de validação

A obtenção do número ótimo de grupos sem a aplicação de métodos hierárquicos pode ser feita por meio do uso de índices de validação, cujos procedimentos são vastamente utilizados na análise de agrupamentos convencional (Basu e Srinivas, 2014). Dentre os diversos índices conhecidos, tem-se aqueles baseados em: (i) características da matriz de pertinência dos elementos aos grupos, como o coeficiente de partição e entropia de partição (Bezdek, 1974, 1981); (ii) distância entre os grupos, como os índices de Dunn (Dunn, 1974) e Davies-Bouldin (Davies e Bouldin, 1979); (iii) grau de pertinência e características dos vetores, como os índices de Fukuyama-Sugeno (Fukuyama e Sugeno, 1989), Xie-Beni (Xie e Beni, 1999) e PBM (Pakhira *et al.*, 2004). Para maior conhecimento, alguns índices encontram-se sintetizados e discutidos em Boutin e Hascoet (2004).

Rao e Srinivas (2006b) na regionalização de bacias hidrográficas através do método *fuzzy c-means* testaram cinco índices de validação para diferentes números de grupos (*c*) e coeficiente de difusividade (*m*) para avaliar a sua adequação na obtenção de regiões homogêneas. Os índices avaliados foram: coeficiente de partição; entropia de partição; extensão do Xie-Beni; Fukyyama-Sugeno; e Kwon. Os valores de *c* e *m* que produziram grupos mais homogêneos foram 7 e 1,4, respectivamente, sendo que os índices Kwon e extensão de Xie-Beni indicaram corretamente o número de agrupamentos. Ao testar outros quatro índices de validação por meio de um método híbrido de agrupamento, Rao e Srinivas (2006a) concluíram que os índices de Dunn e Davies-Bouldin apresentaram melhor efetividade na identificação do número ótimo de partições por resultarem em agrupamentos mais homogêneos

Basu e Srinivas (2014) testaram seis índices de validação de agrupamento difuso, dentre eles, o coeficiente de partição e entropia de partição. Os seis índices avaliados falharam na identificação do número ideal de partições e, portanto, foi proposto um novo procedimento envolvendo a investigação do grau de pertinência, o tamanho do agrupamento e a homogeneidade dos diferentes agrupamentos obtidos com a variação dos parâmetros de entrada. Embora a proposição de um novo método para a identificação do número ideal de partições tenha sido satisfatória, ainda assim os grupos finalmente formados apresentaram-se somente próximos a homogeneidade, demonstrando a influência dos demais parâmetros de entrada na medida da heterogeneidade avaliada.

No âmbito nacional, Carvalho (2007) utilizou o método *fuzzy c-means* para o agrupamento de regiões homogêneas de precipitação mensal e anual no Sul do Brasil, visando a regionalização da análise de frequência. Para a identificação do número de agrupamentos ideal (*c*), foram avaliados os comportamentos de seis diferentes índices de validação: partição difusa, entropia da partição; Fukuyama-Sugeno; Xie-Beni e; Kwon. Foi verificada uma variabilidade no número de partições ideais obtidos pelos índices avaliados, indicando que não é razoável a utilização de um único índice, porém recomendada a utilização conjunta de diversos.

Homogeneidade dos agrupamentos

Segundo Rao e Srinivas (2006b), no geral, as regiões fornecidas pelos métodos de agrupamento não são estatisticamente homogêneas. Para garantir a efetividade dos métodos na identificação de grupos homogêneos, recomendam a realização de testes de homogeneidade. Hosking e Wallis (1993) propõem uma forma de medir o grau de heterogeneidade de um grupo através da comparação da variação dos momentos L amostrais entre os locais para as regiões ditas homogêneas e sugerem intervalos arbitrários para a aceitação ou não da homogeneidade dos grupos.

Dentre os estudos levantados que utilizaram a análise de agrupamento, poucos realizaram a avaliação da homogeneidade dos resultados obtidos e, destes, a minoria apresentou resultados satisfatórios. Como consequência, são utilizados ajustes visando diminuir a heterogeneidade dos grupos, sendo tradicionalmente utilizadas as diretrizes fornecidas por Hosking e Wallis (1997). Dentre estas, destacam-se as 3 principais diretrizes: (i) eliminar elementos do conjunto de dados; (ii) transferir elementos entre regiões e; (iii) dividir uma região em mais regiões.

Rao e Srinivas (2006a) utilizaram os ajustes descritos por Hosking e Wallis (1997) para aprimorar a homogeneidade das regiões agrupadas, dado que os 3 testes de homogeneidade utilizados para avaliar os agrupamentos formados, descritos por Hosking e Wallis (1993), indicaram que estes apresentavam heterogeneidade. A aplicação destas diretrizes resultou na obtenção de 7 agrupamentos, sendo 6 considerados homogêneos, demonstrando melhoria nos resultados obtidos. Rao e Srinivas (2006b) também obtiveram melhoria na formação de grupos homogêneos com a utilização de ajustes.

CONSIDERAÇÕES FINAIS

Como pode ser visto ao longo do trabalho, a análise de agrupamentos de dados hidrológicos é utilizada nacional e internacionalmente, principalmente na identificação de regiões hidrologicamente homogêneas. A teoria difusa se mostra adequada a esta área de estudo por permitir a avaliação das incertezas envolvidas nos agrupamentos, como, por exemplo, por meio da matriz de pertinência.

Nesse sentido, constatou-se que o método difuso mais utilizado e avaliado é o método *fuzzy c-means* e suas variações. Segundo os estudos analisados, este algoritmo é capaz de produzir resultados satisfatórios, quando da definição ótima dos parâmetros de entrada necessários. Para isto, indica-se o uso de diferentes índices de validação para a definição do número ideal de agrupamentos, além da avaliação de diferentes valores de coeficiente de difusividade e grau de tolerância. É importante

destacar que os resultados também são influenciados pelas características dos elementos a serem agrupados (áreas de estudo, por exemplo).

Adicionalmente, embora a maior parte dos estudos levantados não realize tal verificação, é necessária a avaliação da efetividade dos agrupamentos na formação de áreas homogêneas através de testes de homogeneidade. Quando a homogeneidade não é atingida, recomenda-se a realização de ajustes entre os agrupamentos formados para eliminar e heterogeneidade dos grupos. Em resumo, embora haja uma ampla utilização de métodos de análise de agrupamento, a aplicação direta dos métodos sem a definição dos parâmetros ideais e a utilização dos resultados obtidos sem a devida avaliação da efetividade em produzir regiões homogêneas podem trazer resultados inadequados para os estudos na área de recursos hídricos.

REFERÊNCIAS

ANDRADE, L. P. de. (2004). *Procedimento Interativo de Agrupamento de dados.* Dissertação (mestrado). Universidade Federal do Rio de Janeiro, Rio de Janeiro, 193 p.

BASU, B.; SRINIVAS, V. V. (2014). Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resources Research*, 50, 4, pp. 3295–3316.

BASU, B.; SRINIVAS, V. V. (2016). Regional Flood Frequency Analysis Using Entropy-Based Clustering Approach. *Journal of Hydrologic Engineering*, pp. 1–12.

BEZDEK, J. C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3, 3, pp. 58–73.

BEZDEK, J. C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms.* Spring Street, New York 267 p.

BOSCHI, R. S.; OLIVEIRA, S. R. DE M.; ASSAD, E. D. (2011). Técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul. *Engenharia Agrícola*, 31, 6, pp. 1189–1201.

BOUTIN, F.; HASCOET, M. (2004). Cluster validity indices for graph partitioning. Proceedings. *Eighth International Conference on Information Visualisation*, pp. 376–381.

CARVALHO, T. L. L. de. (2007). *Análise regional de frequências aplicadas à precipitação pluvial*. Dissertação (mestrado). Universidade Federal do Rio Grande do Sul, Porto Alegre, 111 p.

CHEN, X.; WANG, L. (2012). Flood feature identification and clustering in Wujiang River, south China. In *Anais* XIX International Conference on Water Resources. Univerity of Illinois. 8 p.

DAVIES, D. L. BOULDIN, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 4, pp. 224–227.

DUNN, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 1, pp. 95–104.

ELESBON, A. A. A. (2012). Gestão de Recursos hídricos: análise estatística multivariada em suporte a regionalização de vazões e proposta metodológica para avaliação, rearranjo e otimização de redes de monitoramento hidrométrico. Tese (Doutorado). Universidade Federal de Viçosa – Viçosa – MG, p. 148.

FUKUYAMA, Y., SUGENO, M. (1989). A new method of choosing the number of clusters for the fuzzy c-means method. In *Anais* Fifth Fuzzy Systems Symposium, Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT), Kobe, Japan.

GIBERTONI, R. de F. C.; KAVISKI, E.; MINE, M. R. M. (2004). Regionalização de Parâmetros Hidrológicos Utilizando Análise Difusa. *Revista Brasileira de Recursos Hídricos*, 9, 3, pp. 69–79.

GOMES, E. P.; BLANCO, C. J. C.; PESSOA, F. C. L. (2016). Formação de regiões homogêneas de precipitação por agrupamento fuzzy c-means. In *Anais* XIII Congresso Nacional de meio ambiente de Poço de Caldas. 8 p.

HALL, M. J. & MINNS, A. W. (1999). The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal*, 44(5), pp.693- 704.

HAN, J.; KAMBER, M.; PEI, J. (2012). *Data Mining: Concepts and Techniques*. 3rd ed. Elsevier, Waltham, USA, 740 p.

HOSKING, J. R. M.; WALLIS, J. R. (1993). Some statistics useful in Regional Frequency Analysis. *Water Resources Research*, 29, 9, pp. 271-281.

HOSKING, J.R.M., WALLIS, J.R. (1997). Regional Frequency Analysis: An approach based on L-Moments. 1ª ed. New York: Cambridge University press, Cambridge, 224 p.

PAKHIRA M. K.; BANDYOPADHYAY S.; MAULIK K. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37, 3, pp. 481-501.

PESSOA, F. C. L. (2015). Regionalização de curvas de permanência de vazões na Amazônia legal. Tese (doutorado). Universidade Federal do Pará. 237 p.

RAO, A. R.; SRINIVAS, V. V. (2006a) Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology*, 318, 1–4, pp. 37–56.

RAO, A. R.; SRINIVAS, V. V. (2006b). Regionalization of watersheds by fuzzy cluster analysis. *Journal of Hydrology*, 318, 1–4, pp. 57–79.

ROSS, T. J. (1995). Fuzzy Logic with Engineering Applications. McGraw –Hill, New York.

RUSPINI, E. H. (1969). A new approach to clustering. *Information and Control*, 15, 1, p. 22–32.

TAN, P. N.; STEINBACH, M.; KUMAR, V. (2005). Cluster Analysis: Basic Concepts and Algorithms. In: I*ntroduction to Data Mining*, 82 p.

UDA, P. K.; FRANCO, A. C. L.; QUEEN, G.; BONUMÁ, N. B.; KOBIYAMA, M. (2015). Análise de cluster da precipitação na bacia do Rio Iguaçu, região Sul do Brasil. In *Anais* XXI Simpósio Brasileiro de Recursos Hídricos, Brasília - DF, nov.2015. pp. 1–7.

WANG, L.; CHEN, X.; SHAO, Q.; LI, Y. (2015). Flood indicators and their clustering features in Wujiang River, South China. *Ecological Engineering*, 76, pp. 66–74.

XIE, X.L.; BENI, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13,8, pp. 841–847.