MÉTODOS DE CLUSTERIZAÇÃO PARA DEFINIÇÃO DE REGIÕES HOMOGÊNEAS DE PRECIPITAÇÕES INTENSAS

Daniele Feitoza Silva[1]; Dirceu Silveira Reis Júnior[2] & Carlos Henrique Ribeiro Lima[3]

Palavras-Chave – Regionalização de chuvas intensas; Clusterização; Curvas IDF.

INTRODUÇÃO

A estimativa de precipitações intensas é um elemento central no dimensionamento e avaliação de obras hidráulicas, sendo usualmente realizada a partir de curvas Intensidade–Duração–Frequência derivadas de séries históricas pluviométricas. Contudo, a disponibilidade de registros consistentes, longos e representativos é limitada, o que aumenta a incerteza dos quantis estimados e induz decisões pouco robustas de projeto. Uma estratégia para mitigar esse problema é a análise de frequência regional (Hosking, Wallis, 1997; Hailegeorgis, Thorolfsson, Alfredsen, 2013; Haruna, Blanchet, Favre, 2022), na qual informações de diferentes estações são combinadas sob a hipótese de que pertencem a regiões homogêneas, isto é, conjuntos de séries que podem ser tratadas como realizações de um mesmo processo gerador de extremos, diferindo apenas por um fator local de escala.

Historicamente, formular regiões homogêneas para precipitações máximas baseava-se em critérios empíricos (Dalrymple, 1960 *apud* Dawdy, Griffis, Gupta, 2012). A partir da década de 1990, a Análise de Frequência Regional por Momentos-L forneceu um arcabouço estatístico, interpretando regiões homogêneas como populações cujas séries podem ser tratadas como realizações de uma mesma distribuição adimensional, escalonada por um índice local, recomendando o uso de análise de *clusterização* para identificar conjuntos de estações com comportamento semelhante (Hosking, Wallis, 1997). Neste contexto, a *clusterização* surge como ferramenta para propor regiões homogêneas candidatas, sobre as quais testes de homogeneidade e modelos regionais de frequência podem ser aplicados de forma consistente (Jain, 2010; Torres, 2025). Este trabalho revisa, de forma sintética, as principais famílias de métodos de *clusterização* aplicáveis à definição de regiões homogêneas de precipitações intensas, com vistas à posterior regionalização de curvas IDF.

METODOLOGIA

O estudo baseou-se em revisão bibliográfica direcionada sobre *clusterização* espacial de precipitações intensas, regionalização de chuvas extremas, curvas IDF, teoria de valores extremos e dependência espacial. Foram consultadas bases como *Google Scholar*, *Scopus* e *Web of Science*, além de livros e relatórios técnicos de referência, resultando em aproximadamente 80 trabalhos analisados.

MÉTODOS DE *CLUSTERIZAÇÃO*

Os métodos de *clusterização* aplicados à definição de regiões homogêneas podem ser organizados em diferentes famílias, segundo o princípio estatístico para medir similaridade. Entre os métodos mais utilizados estão os métodos particionais (*hard* ou *soft*), os métodos hierárquicos, e os métodos baseados em modelos. Em todos os casos, o objetivo é identificar conjuntos de estações com comportamento semelhante, seja a partir de estatísticas de precipitação extrema, seja com base em covariáveis físico-climáticas ou em parâmetros de modelos probabilísticos.

Os métodos particionais e hierárquicos são baseados em medidas de distância ou dissimilaridade (e.g., norma Euclidiana, F-madograma, correlação etc.). Algoritmos como K-means e K-medoids/PAM (P*artitioning Around Medoids*) dividem os dados em k grupos previamente

[1] Universidade de Brasília. E-mail: pariconha@gmail.com
[2] Departamento de Engenharia Civil e Ambiental, Universidade de Brasília. E-mail: dirceureis@unb.br
[3] Universidade de Brasília. E-mail: chrlima@unb.br

definidos, buscando minimizar a variabilidade interna e maximizar a separação entre grupos. O primeiro utiliza centróides e o segundo medoids, o que o torna mais robusto e permite empregar distâncias não euclidianas em aplicações com extremos. Já os métodos hierárquicos, como o de Ward, produzem dendrogramas e permitem visualizar relações de similaridade em múltiplas escalas, sendo amplamente empregados em estudos de regionalização por sua simplicidade e interpretabilidade. As abordagens fuzzy diferem por admitir que uma mesma estação possua graus de pertinência em mais de um grupo, útil em regiões de transição climática (Chavoshi *et al*., 2012). Esses métodos são usados para identificar padrões ou regimes de variabilidade, e partem, em geral, de uma métrica euclidiana e de uma decomposição da variância intra/intergrupos, comum quando os dados podem ser aproximados por distribuições quase gaussianas. Essa estratégia tende a funcionar para médias mensais, em que o teorema do limite central dá respaldo à aproximação normal em cada regime, mas é menos adequado para descrever diretamente o comportamento de extremos (Bernard *et al*., 2013).

Em eventos extremos de precipitação, a distribuição dos dados tende a ser assimétrica: dois postos podem ter médias parecidas, mas comportamentos distintos nos extremos. Essa lacuna motivou o desenvolvimento de metodologias associadas à dependência espacial. Bernard *et al*. (2013) propuseram a clusterização baseado no F-madograma, uma medida de dependência entre extremos construída a partir das distribuições de probabilidade em cada ponto. Essa medida é então combinada com o algoritmo K-medoids, que busca partições coerentes do campo de extremos. No K-medoids, cada grupo é representado por um medoide, e não por uma média. Em outras palavras, a representatividade do cluster está ancorada em um registro, e não em um centróide não observável.

Saunders, Stephenson, Karoly (2020) retomam essa abordagem comparando o K-medoids com o método hierárquico. Eles mostram que o K-medoids pode ser sensível à distribuição espacial das estações: tende a concentrar medoides em áreas com maior densidade de postos. Já a clusterização hierárquica fornece uma sequência ordenada de partições em que é possível interpretar a passagem de uma dependência extremal à medida que o nível de corte sobe. Para avaliar essas regiões, os autores ajustaram, em cada *cluster*, processos max-estáveis e analisaram as elipses de dependência resultantes. Essas elipses descrevem o alcance e a direção preferencial da dependência extrema, permitindo verificar se a regionalização reproduz de forma adequada a escala espacial e a anisotropia da dependência entre extremos dentro e entre as regiões definidas.

Os métodos baseados em modelos assumem que as observações provêm de uma mistura de distribuições (e.g., gaussianas ou de extremos), com cada componente correspondendo a um cluster com parâmetros próprios (Fraley, Raftery, 2002). Essa família é particularmente promissora para extremos, pois possibilita clusterizar parâmetros de modelos de extremos em vez de variáveis.

Nesse contexto, a abordagem bayesiana hierárquica proposta por Sun *et al*. (2015) combina uma mistura de clusters, modelos GEV não estacionários e *pooling* parcial, permitindo identificar regiões com comportamento semelhante de tendência e reduzir incertezas na estimação dos parâmetros. Na formulação, cada cluster é associado a um conjunto de parâmetros regionais e a coeficientes de regressão que ligam os parâmetros da GEV ao tempo, enquanto coeficientes locais variam entre estações sob um modelo de *pooling* parcial. Essa estrutura permite que a mesma cadeia de MCMC (Cadeia de Markov via Simulação Monte Carlo) faça, simultaneamente (i) a classificação das estações em sub-regiões, (ii) a estimação de parâmetros locais e regionais e (iii) a quantificação completa da incerteza, evitando o viés de esquemas em dois estágios (clusterização *a priori* seguida de ajuste de extremos).

Trabalhos mais recentes estendem essa família para incluir explicitamente covariáveis climáticas. Zeng *et al.* (2024) formularam um modelo hierárquico bayesiano, em que os coeficientes de regressão associam os parâmetros da distribuição à índices climáticos de larga escala com *pooling* parcial. Os autores mostraram que o modelo regional reduziu os intervalos de confiança em

comparação com modelos locais, reforçando que modelos baseados em misturas e estruturas hierárquicas são particularmente adequados quando se deseja: (i) incorporar não-estacionariedade; (ii) compartilhar informação; e (iii) que as regiões homogêneas emerjam da dinâmica dos extremos.

RESULTADOS

A Tabela 1 sintetiza os métodos de *clusterização* revisados, contrastando a natureza da *clusterização*, a medida de dissimilaridade adotada e o tratamento (ou não) da dependência extremal entre estações.

Tabela 1 – Síntese comparativa das famílias de métodos de clusterização revisadas.

Método	Natureza	Medida de dissimilaridade	Apropriada para extremos?
Hierárquico (Ward)/ K-means	Rígida	Distância euclidiana	Não, assume estrutura aproximadamente gaussiana
Fuzzy c-means	Flexível	Distância euclidiana	Não; adequada a regimes médios
Modelos de mistura (mistura gaussiana)	Probabilística	Verossimilhança	Limitada; útil para regime central, não para caudas pesadas sem modelagem específica
F-madograma + K-medoids	Rígida	F-madograma (dependência extremal)	Sim; mas K-medoids é sensível à densidade de postos
Hierárquico bayesiano (mistura de GEV)	Probabilística	Verossimilhança de extremos + *prior* hierárquica; covariáveis	Sim; parâmetros de extremos no espaço do modelo

Em síntese, a literatura indica que os métodos baseados em distância Euclidiana permanecem úteis como ferramentas exploratórias e para variáveis quase gaussianas (e.g., médias mensais). Abordagens ancoradas em teoria de valores extremos permitem regionalizar diretamente padrões de máximos e quantis de projeto, reduzindo a incerteza. No contexto de curvas IDF, isso se traduz em estimativas mais consistentes. Já os métodos clássicos permanecem atrativos por sua simplicidade computacional, ampla disponibilidade em pacotes estatísticos, inclusive em Python ou R, e facilidade de interpretação, o que explica sua predominância na literatura aplicada.

Um resultado relevante da revisão é a limitação, na literatura nacional, de trabalhos que apliquem clusterização de natureza probabilística voltada a eventos intensos de precipitação. A maioria dos trabalhos consultados emprega o K-means ou Ward (Oliveira, 2023; Torres, 2025) ou não são direcionados à *clusterização* de eventos intensos.

CONCLUSÕES

A revisão evidenciou que a literatura de *clusterização* para definição de regiões homogêneas de precipitações intensas evoluiu de critérios empíricos para abordagens estatísticos explícitos a partir dos Momentos-L e, mais recentemente, métodos que tratam diretamente a dependência extremal e *clusterização* hierárquica bayesiana no espaço dos parâmetros da distribuição.

Como principais conclusões, destacam-se: (i) a predominância do K-means na literatura, inclusive como comparação para métodos mais recentes; (ii) métodos clássicos de clusterização baseados em distância euclidiana são úteis como etapa exploratória, mas são limitados para definir regiões homogêneas de precipitações extremas; e (iii) métodos desenvolvidos com base na teoria de valores extremos são mais adequadas para regionalização voltada a quantis de projeto e curvas IDF, por representarem explicitamente a dependência extremal e aspectos de não estacionariedade.

REFERÊNCIAS

BERNARD, E.; NAVEAU, P.; VRAC, M.; MESTRE, O. (2013). *"Clustering of maxima: Spatial dependencies among heavy rainfall in France"*. Journal of Climate, v.26, n.20, p.7929–7937. DOI: 10.1175/JCLI-D-12-00836.1.

CHAVOSHI, S.; AZMIN SULAIMAN, W. N.; SAGHAFIAN, B.; SULAIMAN, M. D. N. B.; LATIFAH, A. M. (2012). *"Soft and hard clustering methods for delineation of hydrological homogeneous regions in the southern strip of the Caspian Sea Watershed"*. Journal of Flood Risk Management, v.5, n.4, p.282–294. DOI: 10.1111/j.1753-318X.2012.01149.x.

DAWDY, D. R.; GRIFFIS, V. W.; GUPTA, V. K. (2012). *"Regional flood-frequency analysis: How we got here and where we are going"*. Journal of Hydrologic Engineering, v.17, n.9, p.953–959. DOI: 10.1061/(ASCE)HE.1943-5584.0000584.

FRALEY C.; RAFTERY, A. E. (2002). *"Model-based clustering, discriminant analysis, and density estimation"*. Journal of the American Statistical Association, 97(458), 611–631. https://doi.org/10.1198/016214502760047131.

HAILEGEORGIS, T. T.; THOROLFSSON, S. T.; ALFREDSEN, K. (2013). *"Regional frequency analysis of extreme precipitation with consideration of uncertainties to update IDF curves for the city of Trondheim"*. Journal of Hydrology, 498, 305-318. https://doi.org/10.1016/j.jhydrol.2013.06.019.

HARUNA, A.; BLANCHET, J.; FAVRE, A-C. (2022). *"Performance-based comparison of regionalization methods to improve the at-site estimates of daily precipitation"*. Hydrology and Earth System Sciences, 26, 2797-2811. https://doi.org/10.5194/hess-26-2797-2022.

HOSKING, J. R. M.; WALLIS, J. R. (1997). "*Regional Frequency Analysis: An Approach Based on L-Moments*". Cambridge University Press, Cambridge/UK, 224 p.

JAIN, A. K. (2010). *"Data clustering: 50 years beyond k-means"*. Pattern Recognition Letters, v.31, n.8, p.651–666. DOI: 10.1016/j.patrec.2009.09.011.

OLIVEIRA, L. B. (2023). *"Application of clustering methods for hydrological regionalization using CAMELS-BR"*. Dissertação (Mestrado em Engenharia Civil), Universidade Federal do Ceará, Fortaleza/CE, 77 p.

SAUNDERS, K. R.; STEPHENSON, A. G.; KAROLY, D. (2020). *"A regionalisation approach for rainfall based on extremal dependence"*. Extremes, 23, 1–26. DOI: 10.1007/s10687-019-00373-2.

SUN, X.; THYER, M.; RENARD, B.; LANG, M.; KUCZERA, G. (2015). *"Hierarchical Bayesian clustering for nonstationary flood frequency analysis: Application to trends of annual maximum flow in Germany"*. Water Resources Research, v.51, n.8, p.6759–6779. DOI: 10.1002/2015WR017117.

TORRES, I. P. (2025). *Uso de regiões homogêneas para obtenção de equações de chuvas intensas pelo método "Transformação da Máxima Precipitação Diária"*. Dissertação (Mestrado em Meteorologia Agrícola), Universidade Federal de Viçosa, Viçosa/MG, 150 p.

ZENG, H.; ZHOU, Y.; LIU, P.; LI, X.; HUANG, J.; ZHOU, H.; YU, W. (2024). *"Climate-informed clustering based nonstationary regional extreme flood events spatio-temporal evolution using hierarchical Bayesian modeling"*. Journal of Hydrology: Regional Studies, 56, 102066. https://doi.org/10.1016/j.ejrh.2024.102066.

AGRADECIMENTOS

Os autores agradecem à Agência Nacional de Águas e Saneamento Básico (ANA) e à Fundação de Empreendimentos Científicos e Tecnológicos (FINATEC) pelo apoio institucional e financeiro que possibilitou o desenvolvimento deste trabalho (Processo SEI 23106.017474/2024-93).