IDENTIFICAÇÃO DE GRUPOS REGIONAIS DE CHUVA POR ANÁLISE DE AGRUPAMENTOS

Frederico Ivanchechen de Mattos[1], Mino Viana Sorribas[2], Einara Zahn[2] e José Eduardo Gonçalves[2]

[1]Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental (PPGERHA), Universidade Federal do Paraná (UFPR), Brasil.
[2]Sistema Metereológico do Paraná (SIMEPAR), Brasil.
E-mail: fivanche@gmail.com, mino.sorribas@simepar.br, einara.zahn@gmail.com, jose.eduardo@simepar.br

Introdução

Medições manuais e automáticas de precipitação concedem informações cruciais para o gerenciamento de recursos hídricos e projetos de obras hidráulicas. Entretanto, problemas na telemetria, entupimento dos pluviômetros e avarias dos instrumentos de medição durante tempestades podem causar falhas e/ou erros na série de dados. Consequentemente, a Agência Nacional de Águas (ANA) exige a consistência desses dados em antemão a sua disponibilização. Nesse sentido, métodos de análise de frequência regional são utilizados para identificar erros e, na escala temporal adequada, preencher falhas. O requerimento primário para essa análise é a identificação de áreas pluviometricamente homogêneas. Nesse estudo de caso explora-se a análise de agrupamentos pelo método de Ward utilizando a distância Euclidiana como medida de dissimilaridade e parâmetros de validação de agrupamentos para determinar o número ótimo de grupos. Apesar da forte semelhança entre as observações de todos os postos, a análise de agrupamento isolou sete grupos regionais, reduzindo a arbitrariedade na tomada de decisões.

Palavras-chave: Análise de grupos, Método de Ward, Hidrologia estatística.

Metodologia

Dados de entrada

Séries de precipitação acumulada mensal foram levantadas entre 2000 e 2015 para 30 estações pluviométricas monitoradas pela CTG-Brasil (Chinese Three Gorges Corporation) e mais uma adicional pertencente à rede do SIMEPAR (Sistema Meteorológico do Paraná). Todas as estações se encontram na bacia hidrográfica do Paranapanema, um rio de 929 km de extensão que divide os estados de São Paulo e Paraná, drenando cerca de 100800 km² (ANA, 2016). Para cada posto $\mathbf{x_i}$ foram acumulados os registros diários de precipitação no mês $\boldsymbol{j}$, obtendo-se o valor $\mathbf{z_{ij}}$. Como a análise de agrupamentos exige que haja observações no mesmo mês para todas as estações, 5% dos dados foram invalidados devido a falhas no histórico.

Análise de agrupamentos

Conquanto existam diversos métodos de identificação de regiões homogêneas (Satyanarayana e Srinivas, 2008), a análise de agrupamentos é amplamente empregada para dados de precipitação em estudos no mundo inteiro (Pelczer e Cisneros-Iturbe, 2008; Santos, Lucio e Silva, 2015; Shirin e Thomas, 2016; Terassi e Galvani, 2017). Há duas vertentes da análise, ambas com o mesmo objetivo: formar grupos cujos membros pertencentes sejam os mais semelhantes possíveis entre si e os mais distintos dos demais. Métodos hierárquicos, como o de Ward, iniciam com cada estação em um grupo (de tamanho um, denominado singleton) e, à medida que se itera, o método funde os grupos mais semelhantes. Essa união pode ser baseada em uma gama de princípios, como mínima distância entre grupos (single linkage) e menor distância máxima entre grupos (complete linkage). A grande vantagem do método de Ward é que essa junção se baseia no mínimo incremento de variância, portanto, procurando minimizar a distância média entre as séries históricas dos postos. Foi adotado a fórmula de Lance-Williams para o algoritmo de Ward, conforme a equação [1]:

$$\boldsymbol{d}(\boldsymbol{C_i} \cup \boldsymbol{C_j}, \boldsymbol{C_k}) = \sqrt{\frac{(|\boldsymbol{C_k}| + |\boldsymbol{C_i}|)\boldsymbol{d}(\boldsymbol{C_k}, \boldsymbol{C_i})^2 + (|\boldsymbol{C_k}| + |\boldsymbol{C_j}|)\boldsymbol{d}(\boldsymbol{C_k}, \boldsymbol{C_j})^2 + |\boldsymbol{C_k}|\boldsymbol{d}(\boldsymbol{C_i}, \boldsymbol{C_i})^2}{|\boldsymbol{C_i}| + |\boldsymbol{C_j}| + |\boldsymbol{C_k}|}} \quad [1]$$

Em que $\mathbf{d}(\mathbf{C_i} \cup \mathbf{C_j}, \mathbf{C_k})$ representa o custo de fusão do grupo $\mathbf{C_i}$ ao $\mathbf{C_j}$ em relação ao grupo $\mathbf{C_k}$, que não pertence a nenhum dos outros dois. $|\mathbf{C}_*|$Representa o número de membros no grupo, enquanto que $\mathbf{d}(\mathbf{C_a}, \mathbf{C_b})$, a distância entre os grupos $\mathbf{C_a}$ e $\mathbf{C_b}$.

Medida de Dissimilaridade

A distância, ou dissimilaridade, entre grupos pode ser medida por métricas como distância Euclidiana (Santos, Lucio e Silva, 2015), distância Euclidiana Padronizada (Satyanarayana e Srinivas, 2008), distância Mahalanobis (Detzel, Bessa e Mine, 2013) e outras. Optou-se pela distância Euclidiana pelo fato dela poder ser entendida como uma medida em um espaço multidimensional (Terassi e Galvani, 2017) e ter sido amplamente utilizada em estudos similares. Sua fórmula é expressa em [2].

$$\boldsymbol{d}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n_j}} (\boldsymbol{z_{ik}} - \boldsymbol{z_{jk}})^2} \quad [2]$$

Onde x_i representa o vetor de observações $(z_{i1},\ z_{i2}, \dots, z_{in_j})$.

Determinação do número ótimo de grupos

A inspeção visual de dendrogramas, como o da figura 1, é o método mais simples para determinar o número de grupos. Ele é um gráfico em formato de árvore cujas folhas representam os postos pluviométricos e a união dos galhos, os pontos de mesclagem dos grupos. Espera-se que a distância vertical entre os nódulos seja pequena na base do gráfico e que haja um salto logo após atingir-se o número ótimo de agrupamentos. Para reduzir a incerteza, recorreu-se a três parâmetros de desempenho de agrupamentos sugeridos pela literatura técnica (Pelczer e Cisneros-Iturbe, 2008; Satyanarayana e Srinivas, 2008), todos partindo da premissa que bons agrupamentos são compactos e distantes entre si. O índice Dunn (1974) é a razão entre a menor distância entre grupos e a maior distância entre membros de um mesmo grupo, sendo calculado pelas equações [3], [4] e [5]. Quanto maior o índice de Dunn (DI), melhor o agrupamento. O índice de silhueta (SI) também relaciona distância intergrupos e intragrupos, embora com as equações [6], [7] e [8]. *SI* Próximos da unidade apontam bons agrupamentos, os próximos de zero indicam grupos sobrepostos e valores negativos são sinais de alocação errônea de postos nos grupos. Satyanarayana e Srinivas (2008) usam o índice Davies-Bouldin (DBI), uma razão entre desvios das observações dos membros em relação ao centróide de seu grupo e a distância entre grupos. Verificando sua formulação nas equações [9] e

[10] é possível entender que grupos compactos e bem separados têm IDB próximos de zero.

$$DI = \min \left[\frac{\min_{1\le i\le j\le n_c}[d(C_i,C_j)]}{\max\ [\Delta(C_k)]}\right] \quad [3]$$

$$d(C_i,C_j) = \max_{x_i,x_j\in C_i}\left[d(x_i,x_j)\right] \quad [4]$$

$$\Delta C_k = \max_{x_i\in C_i,x_j\in C_j}\left[d(x_i,x_j)\right] \quad [5]$$

$$SI = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{b(x_i)-a(x_i)}{\max\,[b(x_i),a(x_i)]} \quad [6]$$

$$a(x_i) = \frac{1}{n_c}\sum_{j=1,j\neq i}^{n_c} d(x_i,x_j)\ \mid x_i,x_j\in C_i \quad [7]$$

$$b(x_i) = \frac{1}{n_c}\sum_{j=1}^{n_c} d(x_i,x_j)\ \mid x_i\in C_i, x_j\in C_{j*} \quad [8]$$

$$DBI = \frac{1}{n_c}\sum_{i=1}^{n_c}\max_{i\neq j}\left[\frac{S_i+S_j}{d(\overline{C_i},\overline{C_j})}\right] \quad [9]$$

$$S_i = \frac{1}{n_i}\sum_{x_i\in C_i} d(x_i\,,\overline{C_i}) \quad [10]$$

Onde n_c é o número total de grupos, n_ié o número de membros do grupo C_i, C_{j*} é o grupos mais próximo de C_i e $\overline{C_i}$ é o centróide do grupo C_i.

Resultados

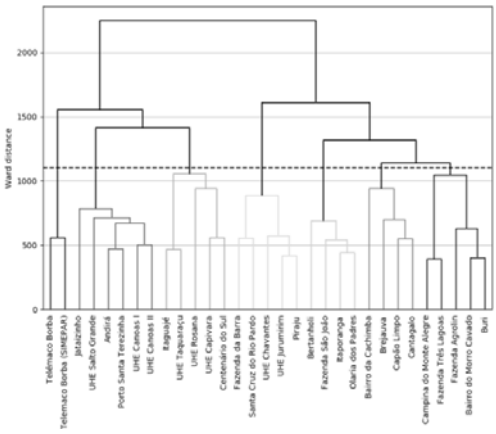

Figura 1.-Dendrograma com linha de corte em 1100.

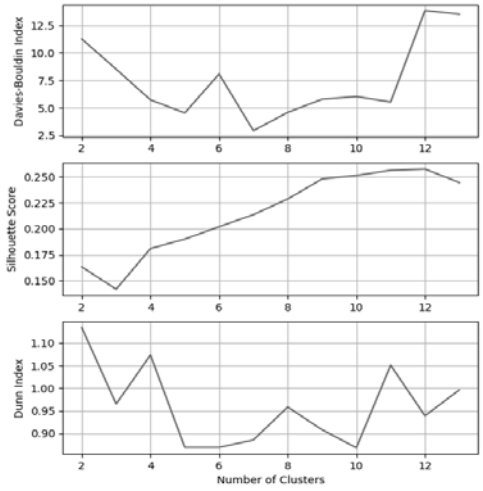

Figura 2.-Índices de desempenho de agrupamentos.

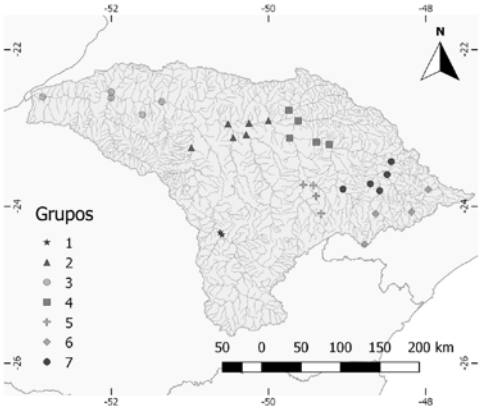

Figura 3.- Disposição geográfica dos grupos.

Embora DI e SI apontem para 11 grupos, esse resultado gera duas estações em singletons e três grupos de tamanho 2, o que não é interessante. O DBI sugere sete como número ótimo de grupos, o que representa um agrupamento muito coerente, visto que mesmo sem entrar com as posições geográficas dos postos no algoritmo, estações próximas geograficamente ficaram nos mesmos grupos. Esse fato concorda com a premissa que agentes influenciadores da precipitação, como clima e relevo, não variam significativamente dentro de um pequeno espaço. Contudo, os grupos são muito similares, pois o SI é próximo de zero. Tal semelhança torna-se mais evidente quando se plota as séries de precipitação acumulada mensal média dos grupos umas contra as outras. Portanto, é provável que um ou mais postos pertençam a mais de um agrupamento.

Referências bibliográficas

ANA (2016). *Relatório Executivo do PIRH Paranapanema.* Agência Nacional de Águas, Brasil.

Detzel, D.H.M.; Bessa, M.R. e Mine, M.R.M. (2013). "Amostragem de Séries Sintéticas Hidrológicas". *Revista Brasileira de Recursos Hdricos* ,Vol. 18, No. 4, pp. 173-184.

Dunn, J.C. (1974). "A fuzzy relative of the ISODATA process and its use in detecting compact well separeted clusters". *Journal of Cybernetics*, Vol. 3, no. 3, pp. 32-57.

Pelczer, I.J. e Cisneros-Iturbe, H.L. (2008). "Identification of rainfall patterns over the Valley of Mexico". *11th International Conference on Urban Drainage*. Edinburgo, pp. 1-9.

Santos, E.B.; Lucio, P.S. and Silva, C.M.S. (2008). "Precipitation regionalization of the Brazilian Amazon". *Atmospheric Science Letters*, Vol. 16, pp. 185-192.

Satyanarayana, P. and Srinivas, V.V. (2008) "Regional frequency analysis of precipitation using large-scale atmospheric variables". *Journal of Geophysical Research*, Vol. 113, D24110, pp. 1-16.

Shirin, S.A.H., and Thomas, R. (2016). "Regionalization of Rainfall in Kerala State". *Procedia Tech*, Vol. 24, pp. 15-22.

Terassi, P. M. B.; Galvani, E. (2017). "Identification of Homogeneous Rainfall Regions in the Eastern Watersheds of the State of Paraná, Brazil". *Climate*, Vol. 5, pp. 53-66.