ANÁLISE DE AGRUPAMENTOS DE AUTOCORRELAÇÕES DE VAZÕES MÉDIAS DIÁRIAS CENSURADAS

Eloy Kaviski[1] e Ingrid Illich Müller[1]

Resumo – Este artigo apresenta um método para definição das tipologias de curvas de autocorrelações de vazões médias diárias censuradas. O método proposto é baseado em técnicas de análise de agrupamentos. O método é aplicado para classificar as curvas de autocorrelação censuradas usando-se os dados de 200 estações fluviométricas localizadas na região Sul do Brasil.

Abstract – This paper presents a method for definition of the serial correlation curves from censored mean daily discharges. The method is based in the cluster analysis technic. The method is applied to classify the censored serial correlation curves from 200 hydrometric stations located in the South Brazilian Region.

Palavras-Chave – coeficiente de correlação diário, censura de vazões diárias, análise de agrupamentos.

[1] Engenheiro Pesquisador do Centro de Hidráulica e Hidrologia Prof. Parigot de Souza – CEHPAR
COPEL - Companhia Paranaense de Energia, UFPR – Universidade Federal do Paraná
Caixa Postal 1309, CEP 81504-000, Curitiba - PR
Fone: (041) 267-1754, fax: (041) 266-2935, e-mail: ingrid@cch.copel.br

INTRODUÇÃO

Em estudos que tratam do planejamento e construção de pequenas centrais hidrelétricas é notadamente interessante que se conheça, além da energia gerada pela usina quando esta opera isolada, também o valor do ganho incremental de energia quando a usina opera integrada em um sistema interligado de porte.

Fill (1989) propôs uma expressão analítica, baseada na teoria estocástica dos reservatórios, para cálculo da energia incremental no caso de pequenas centrais hidrelétricas, quando estas operam integradas. Porém, a operacionalização da fórmula proposta por Fill, exige o conhecimento do comportamento das estatísticas que servem como entrada à fórmula, basicamente a média e desvio padrão das energias médias afluentes à usina, e o coeficiente de correlação dessas afluências com as do sistema interligado, quando considerada à *censura* imposta às vazões naturais devida a motorização. Nagayama (1995) propôs uma forma de avaliação dos efeitos da *censura* sobre as três estatísticas mencionadas, e recomenda, que se obtenha métodos mais precisos de avaliação destes parâmetros, especialmente do coeficiente de correlação censurado.

Este trabalho tem por objetivo apresentar um método de análise e composição das tipologias de curvas de autocorrelações de vazões médias diárias censuradas. O método proposto é baseado em técnicas de análise de agrupamentos e é aplicado para classificar as curvas de autocorrelação das séries de vazões diárias censuradas de 200 estações fluviométricas localizadas na região Sul do Brasil.

ANÁLISE DE AGRUPAMENTOS

Define-se análise de agrupamentos como a exploração e a organização dos dados para identificação de objetos, de forma que exista similaridade dentro dos grupos e dissimilaridade entre os grupos. O objetivo do uso da análise de agrupamentos é detectar inerentes grupos, com o propósito de produzir a indicação de novos sistemas e revelar novos atributos gerais.

Genericamente, o problema de agrupamento pode ser escrito como: seja o conjunto $\underline{X} = \{\underline{x}_i\}_{i=1}^{N}$, representando os dados de *N* vetores amostrais, $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^T$; a análise de agrupamentos deve solucionar o problema de decompor o conjunto $\underline{X}$, em *K* sub-conjuntos $\{G^{(k)}\}_{k=1}^{K} : \bigcup_{k=1}^{K} G^{(k)} = G$, de maneira que o sub-conjunto $G^{(k)} (1 \le k \le K)$, contenha vetores que possuam alguma característica similar, uns em relação aos outros; em geral, $\left\{G^{(k)}\right\}_{k=1}^{K}$ pode ser considerado como a forma de particionar o conjunto $G : G^{(j)} \bigcap G^{(k)} = \Phi \; \forall j,k \in \{1,2,...,K\}$ e $j \neq k$. Deve ser considerado também que um grupo genérico $G^{(k)}$ é composto por $L^{(k)}$ objetos, cujos

vetores incluídos em $G^{(k)}$ são representados por: $\underline{z}_i^{(k)} = \left(z_{i1}^{(k)}, z_{i1}^{(k)}, ..., z_{in}^{(k)}\right)^T$ e $\underline{\mu}_k$ é o vetor médio (centróide) do grupo $G^{(k)}$.

O maior problema na aplicação de técnicas baseadas em análise por agrupamentos está em como definir os grupos, ou seja, definir critérios para caracterizar a similaridade dentro dos grupos e a dissimilaridade entre os grupos.

São muitos os métodos utilizados para detectar e medir a similaridade, podendo os mesmos serem classificados em várias categorias. Neste trabalho, emprega-se a classificação apresentada por Kaviski (1992), onde todos os métodos são incluídos em duas grandes classes: *técnicas hierárquicas e não-hierárquicas*.

Pelas *técnicas hierárquicas*, as medidas de similaridade são usadas para construir uma matriz de representação da similaridade, apresentando todos os pares de associações obtidas por meio de alguma análise criteriosa das amostras. As técnicas operam sobre a matriz de similaridades, permitindo construir uma árvore que caracteriza a relação entre as amostras. O processo é iniciado com *N* objetos e *N* grupos. Os dois grupos mais similares (os mais próximos) formam um novo agrupamento, reduzindo o número de grupos para *(N-1)*. Repetindo-se este procedimento, todos os objetos formam um grupo único no final do processo. Para todos os agrupamentos hierárquicos pode-se construir um diagrama de árvore, também chamado de *dendograma*. Pelo diagrama pode ser visto o agrupamento por objetos através de todos os passos do processo. Para seleção do número de grupos *(K)* pode-se usar a fórmula de Sturges:

$$K = 1 + 1{,}43 \ln N \qquad (1)$$

O critério que define a proximidade entre dois grupos depende do método que estiver sendo utilizado.

Como *técnicas não-hierárquicas* são considerados todos os métodos que não necessitam do cálculo e armazenamento da matriz de similaridades. Existe uma grande variedade de métodos que podem ser classificados nesta categoria, mas todos possuem uma regra comum, que consiste na identificação inicial dos objetos para cada agrupamento. Os passos seguintes do algoritmo verificam se é necessário que os membros dos agrupamentos sejam modificados, desde que sejam respeitadas as instruções específicas de cada método. Os vários algoritmos têm sido definidos com critérios que procuram atingir os propósitos de definição do melhor particionamento ou para definir as melhores partições. Kaviski (1992) apresenta uma sub-divisão das técnicas não-hierárquicas nas seguintes sub-classes:

(i) *Técnicas de particionamento ótimo*, nos quais os objetos são agrupados em agrupamentos mutuamente exclusivos, segundo um critério de otimização.

(ii) *Técnicas de pesquisa de densidade*, nas quais os objetos são agrupados em diferentes sub-grupos, por meio de pesquisa para as regiões que tenham uma alta densidade de probabilidade relativa ("density-search").

(iii) *Técnicas de acumulação* (“clumping”), nas quais as classes (“clumps”) são sobrepostas, e uma classe e seus complementos são tratados como diferentes tipos de classes.

(iv) *Outras técnicas*, as quais não são classificadas num dos tipos previamente definidos, ou aquelas que são do tipo mistas, combinação de duas ou mais técnicas.

MÉTODO PROPOSTO

O método fundamenta-se na aplicação sucessiva de duas técnicas de análise de agrupamentos. Inicialmente aplica-se o algoritmo “K-means” determinando-se um conjunto de grupos de curvas de autocorrelação denominado de *formas-fortes* e em seguida completa-se a análise classificando-se as *formas-fortes* por meio da técnica hierárquica “complete linkage”.

O algoritmo “K-means”, devido a MacQueen (Park, Chen e Simons, 1979) utiliza como critério de otimização minimizar o erro devido ao particionamento de N objetos em K agrupamentos [$e(N,K)$], o qual é obtido por:

$$e(N,K) = \sum_{k=1}^{K} \sum_{l=1}^{L^{(k)}} d(\underline{z}_l^{(k)}, \underline{\mu}_k)^2 \tag{2}$$

Para determinação do valor correto de K geralmente considera-se a seguinte relação:

$$F_1 = \left[\frac{e(N,K)}{e(N,K+1)} - 1 \right] (N - K + 1) \tag{3}$$

A expressão (3), sob a hipótese de normalidade da distribuição dos dados, mede a redução da variância quando se passa de K para $K+1$ agrupamentos. Como regra empírica, se o valor de F_1 resultar maior do que 10, é justificado passar de K para $K+1$ agrupamentos.

O método “average linkage” (Mardia, Kent e Bibby, 1982) usa como critério de similaridade a maximização da seguinte função:

$$\overline{\phi}\left(G^{(j)}, G^{(k)}\right) = \frac{1}{L^{(j)} L^{(k)}} \sum_{\underline{x} \in G^{(j)}, \underline{z} \in G^{(k)}} \phi(\underline{x}, \underline{z}) \tag{4}$$

onde $\phi(\underline{x}, \underline{z})$ é a medida de similaridade, no caso, o coeficiente de correlação entre os objetos incluídos nos grupos $G^{(j)}$ e $G^{(k)}$. Para cada par de grupos analisados, a média de todos os possíveis pares com o novo grupo é calculada, e aqueles que maximizam a função $\overline{\phi}(G^{(j)}, G^{(k)})$ são escolhidos.

Pode-se usar como critério a minimização da função descrita em (4), usando-se como medida de similaridade, a própria função distância Euclidiana, ou seja: $\phi(\underline{x}, \underline{z}) = d(\underline{x}, \underline{z})$. Restringe-se num limite superior o valor da distância entre dois objetos quaisquer para que estejam num mesmo grupo.

Os métodos "single linkage" e "complete linkage" são semelhantes ao método "average linkage", onde o critério de similaridade é simplesmente a distância entre dois grupos, e com a condição de que sejam menores que um valor pré-fixado. Pelo método "single linkage", a distância entre dois agrupamentos é definida como a distância entre os objetos de cada grupo de forma que se minimize a equação (5); e pelo método "complete linkage", de forma que se maximize a expressão (6):

$$\phi(G^{(j)}, G^{(k)}) = \min \left[d(\underline{x}_j, \underline{x}_k) \right], \underline{x}_j \in G^{(j)} \text{ e } \underline{x}_k \in G^{(k)} \tag{5}$$

$$\phi(G^{(j)}, G^{(k)}) = \max \left[d(\underline{x}_j, \underline{x}_k) \right], \underline{x}_j \in G^{(j)} \text{ e } \underline{x}_k \in G^{(k)} \tag{6}$$

Os parâmetros necessários para utilização do método proposto, são:

i)número de curvas de autocorrelação a serem classificadas (*N*);
ii) número de ordenadas das curvas de autocorrelação usadas como parâmetros classificadores (*no*);
iii) número de inicializações dos grupos (*nig*);
iv) número de iterações máximo usado pelo método não-hierárquico (*niter*);
v) número mínimo de curvas de autocorrelação por grupo (*lmin*);
vi) número de grupos a serem classificados pela técnica não-herárquica "K-means" (*K*).

Usando-se a técnica "K-means" *nig* vezes (com inicializações diferentes) determina-se *nig* conjuntos de curvas de autocorrelação classificadas em *K* grupos. As inicializações são realizadas por meio da classificação das *N* curvas de autocorrelação em *K* grupos, usando-se cada uma das ordenadas das curvas de autocorrelação separadamente (uma em cada tentativa de classificação). As ordenadas das curvas são classificadas em ordem de importância em função do valor de sua variância (um método mais preciso pode ser obtido usando-se análise dos componentes principais). O algoritmo usado para aplicação da técnica "K-means" encontra-se descrito em Kaviski (1996).

Com os *nig* resultados obtidos identifica-se os grupos de curvas de autocorrelação cujas curvas estão agrupadas *nig* vezes, definindo-se então as chamadas *formas-fortes*. As *formas-fortes* são obtidas verificando-se os *nig* resultados determinados através da técnica "K-means" e identificando-se quais as curvas de autocorrelação que permanecem agrupadas nos mesmos sub-conjuntos. Desta maneira o número de *formas-fortes* (*nff*) que se obtêm resulta entre *K* e *N*.

Usando-se a técnica hierárquica "complete linkage" para agrupar os *nff* sub-conjuntos obtidos, pode-se reduzi-los até o número de grupos desejado (no mínimo igual a 1 e no máximo igual a *nff*), definindo-se as tipologias existentes. O algoritmo usado para aplicação da técnica "complete linkage" encontra-se descrito em Kaviski (1996).

Para definir o número de grupos pode-se avaliar o valor da função objetivo (*Fob*), expressa pela equação (2), e também determinar a razão entre a variação entre agrupamentos (*Vb*) e a variação dentro dos agrupamentos (*Vw*) das ordenadas das curvas de autocorrelação. As expressões usadas para determinar *Vb* e *Vw* são (Kaviski, 1992):

$$Vb = \sum_{j=1}^{k}\sum_{i=1}^{no}(\bar{y}_{ij} - \bar{y}_{i})^2 /(k-1) \tag{7}$$

$$Vw = \sum_{j=1}^{k}\sum_{i=1}^{no}\sum_{l=1}^{Lk}(y_{ijl} - \bar{y}_{ij})^2 /(n-k) \tag{8}$$

sendo que:

$$\bar{y}_{ij} = \sum_{l=1}^{L_j} y_{ijl} / L_j \tag{9}$$

$$\bar{y}_{i} = \sum_{j=1}^{K}\sum_{l=1}^{Lj} y_{ijl} / n \tag{10}$$

sendo y_{ijl} (*i = 1,...,no*) o valor das ordenadas da curva de autocorrelação *l* (*l = 1,...,Lj*) pertencente ao grupo *j* (*j = 1,...,K*).

Como regra geral pode-se considerar que os grupos constituídos que resultarem no menor valor possível para a função objetivo (*Fob*) e no maior valor da relação *Vb/Vw* é o de melhor qualidade.

Para utilização do método proposto foi desenvolvido um programa escrito na linguagem Turbo-Pascal (Kaviski, 1996).

ESTUDO DE CASO

O método proposto foi aplicado aos dados de 200 estações fluviométricas localizadas na região Sul do Brasil. Para cada local são considerados 100 valores de autocorrelações de vazões médias diárias censuradas, correspondentes às censuras realizadas desde 0,1 a 10 vezes o valor da vazão média de longo período.

Para determinação das *formas-fortes* por meio da aplicação da técnica "K-means" adotaram-se os seguintes parâmetros:

i) número de curvas de autocorrelações, $N = 200$;
ii) número de grupos, $K = 4$;
iii) número de inicializações, $nig = 10$.

Desta maneira obteve-se 17 *formas-fortes* (*nff*). A partir das quais, aplicando-se a técnica "complete linkage", obteve-se os demais agrupamentos. Os valores de *Fob* e de *Vb/Vw* correspondentes a cada valor de *K* são apresentados na Tabela 1. A Tabela 2 apresenta os grupos determinados para $K = 6$, que correspondem ao maior valor da relação *Vb/Vw*.

Tabela 1 - Função objetivo *Fob* e relação *Vb/Vw*

K	**Fob**	**Vb**	**Vw**	**Vb/Vw**
8	2964	1.578	0.2069	7.629
7	3113	1.833	0.2131	8.603
6	3172	2.108	0.2139	9.856
5	3184	1.934	0.2134	9.063
4	4987	2.516	0.2593	9.705
3	7008	2.517	0.3080	8.172
2	8855	0.953	0.3531	2.699

Tabela 2 - Grupos obtidos para $K = 6$

Grupo L	Códigos (ANEEL) das estações por grupo						
1 83	64215083	64270083	64332083	64440000	64442800	64447000	64447011
	64447500	64452000	64453000	64465000	64475000	64476000	64477600
	64480000	64480001	64481200	64490100	64491011	64498511	64498550
	64501000	64501000	64501000	64501011	64506000	64507000	64507001
	64507011	64516083	64535083	64535084	64640000	64670012	64670015
	65015400	65020000	65024000	65025000	65028000	65035000	65208000
	65774411	65774999	65808000	65815000	65815050	65815051	65815100
	65815999	65816000	65816001	65816005	65825000	65825001	65825999
	65826399	65826799	65826800	65829000	65883000	65883050	65883055
	65894990	65895999	65925444	65925666	65927000	65927001	65927010
	65962000	65963102	65963103	65963105	65973501	65985000	65986000
	65990000	65993000	65993500	65993501	65993555	65994000	
2 68	64359001	64360000	64362000	64370000	64380000	64390000	64460000
	64496998	64496999	64497002	64497005	64497012	64497013	64497015
	64571083	64619950	64619999	64620000	64620001	64620002	64620999
	64622001	64625000	64645000	64655000	64655002	64660000	64660500
	64670011	64671000	64675002	64795000	64797000	64799500	64800000
	65385000	65415000	65415010	65774412	65774415	65776000	65776010
	65802000	65803002	65803005	65811000	65816004	65817000	65817001
	65830000	65855000	65855001	65883052	65883053	65894992	65894993
	65894995	65895003	65925000	65925001	65945000	65955000	65960000
	65975000	65975001	65981500	65987000	65988000		
3 34	64382000	64491260	64497001	64497003	64497014	64652000	64652001
	64659000	64670013	64671001	64765000	64767000	64771500	64773000
	64775000	65260000	65365000	65370000	65383500	65770000	65775900
	65809000	65810000	65816002	65948000	65969990	65970000	65970500
	65979000	81021000	81301001	81301999	82160000	82230815	
4 6	65883054	82170000	82195002	82230001	82230002	82230011	
5 4	64670016	81301002	81301005	82230920			
6 5	82230012	82230013	82230015	82230811	82230812		

CONCLUSÕES E RECOMENDAÇÕES

Neste trabalho não foi realizada uma análise de sensibilidade nos resultados dos agrupamentos considerando-se a não utilização das 100 autocorrelações. Intuitivamente, nos parece que reduzindo-se o número de valores, desde que seja considerado um critério bem elaborado para realizar esta simplificação, torna-se possível melhorar a qualidade dos grupos formados. Acredita-se que com o uso das ordenadas que realmente representem as curvas de autocorrelações, sem o uso das demais, que podem ser pouco representativas, consegue-se definir uma classificação de melhor qualidade.

Uma das técnicas recomendadas na literatura, que pode ser usada para identificar as autocorrelações representativas, é o chamado método dos componentes principais (Mardia, Kent e Bibby, 1982; Kaviski, 1992). Um problema de degenerescência nos resultados da análise de agrupamentos pode acontecer quando usa-se as 100 autocorrelações para classificar um pequeno número de estações (casos em que $N<100$).

Os estudos realizados podem ser complementados investigando-se os seguintes pontos:

(i) verificação da possibilidade da existência de degenerescência nos resultados obtidos, causado pelo uso das 100 autocorrelações na realização da classificação de 200 estações;

(ii) identificação das autocorrelações mais significativas, usando-se o método de análise dos componentes principais;

(iii) comparação da classificação das autocorrelações obtida por análise de agrupamentos em relação a outras classificações determinadas por meio de métodos alternativos, como por exemplo, usando análise dos componentes principais ou usando redes neurais artificiais (Faghri e Hua, 1995).

REFERÊNCIAS BIBLIOGRÁFICAS

FILL, Heinz Dieter. 1989. Avaliação analítica da energia garantida incremental de uma usina hidrelétrica. In: VIII Simpósio Brasileiro de Recursos Hídricos, 8., Foz do Iguaçu. *Anais*. Rio de Janeiro, v.1, p. 122-129.

FAGHRI, A.; HUA, J. 1995. Roadway seasonal classification using neural networks. *Journal of Computing in Civil Engineering.* Vol. 9. No.3, July. p.209-215.

KAVISKI, E. 1992. *Métodos de regionalização de eventos e parâmetros hidrológicos.* Dissertação de mestrado. Curitiba: CEHPAR, UFPR, Paraná.

KAVISKI, E. 1996. *Projeto HG-83 - Consultorias de Pequena Duração - Serviço N.6. Relatório n.3: Análise de Agrupamentos de Curvas de Carga.* Centro de Hidráulica e Hidrologia Prof. Parigot de Souza - CEHPAR. Curitiba. 46p.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. 1982. *Multivariate analysis.* Academic Press, Inc., USA.

NAGAYAMA, Marcílio Ulysses. 1995. *Parâmetros anuais de afluências censuradas na escala diária, utilizados no cálculo da energia garantida de pequenas centrais hidrelétricas.* Dissertação de mestrado. Curitiba: CEHPAR, UFPR, Paraná.

PARK, J.K.; CHEN, Y.H.; SIMONS, D.B. 1979. Cluster analysis based on density estimates and its application to Landsat imagery. *Hidrology Papers.* Colorado State University, No. 98, september.