CEPEL
EMPRESA DO SISTEMA ELETROBRÁS

RELATÓRIO TÉCNICO

Nº: DPP/PEN – 109/2001	Nº DE PÁGINAS: 22	ANEXOS: -

TÍTULO:	**MANUAL DE REFERÊNCIA DE ANÁLISE DE CONGLOMERADOS**

ÁREA	2000	Nº DO PROJETO:	1345

CLIENTE:	ONS – Operador Nacional do Sistema Elétrico Rua da Quitanda, 196 - Centro 20.091-000 – Rio de Janeiro – RJ
ATENÇÃO:	Vinícius Forain Rocha Simone Prado

RESUMO:	Este relatório apresenta técnicas estatísticas multivariadas usadas para organizar um conjunto de objetos em subconjuntos mutuamente exclusivos denominados clusters. Tais técnicas serão aplicadas ao modelo de geração de cenários sintéticos de vazões ou energias afim de reduzir o esforço computacional do modelo de planejamento de curto prazo.

AUTORES:
Debora Lima D. Duarte Jardim FPLF
Maria Elvira Piñeiro Maceira ACSI

PALAVRAS-CHAVE:
Análise multivariada, técnicas de agrupamento

CLASSIFICAÇÃO: Controlado

GERENTE DO PROJETO

NOME: Maria Elvira Piñeiro Maceira
Tel.: (021) 598-6454
e-mail: elvira@cepel.br

COORDENADOR

NOME: Maria Elvira Piñeiro Maceira
Tel.: (021) 598-6454
Fax.: (021) 598-6482

APROVAÇÃO:
29/01/01

DIRETOR DE PROGRAMAS DE PESQUISAS E DESENVOLVIMENTO
NOME: Dr. Luiz Alberto da Silva Pilotto

CEPEL
EMPRESA DO SISTEMA ELETROBRÁS

MANUAL DE REFERÊNCIA

DE ANÁLISE DE

CONGLOMERADOS

1 INTRODUÇÃO

Análise de conglomerados (“cluster analysis”) compreende uma grande variedade de técnicas multivariadas cujo objetivo principal é agrupar objetos baseados nas características que eles possuem. Estes procedimentos empiricamente procuram nos dados uma estrutura natural de agrupamento dos objetos ou variáveis formando assim grupos (“clusters”) homogêneos. Desta maneira se a classificação for bem sucedida, objetos dentro de um grupo estarão *próximos* e grupos diferentes estarão *afastados*.

O “cluster analysis” foi introduzido por Tryon em 1939, mas somente a partir de 1963, estimulados por Sokal e Sneath, os métodos de agrupamento começaram a se desenvolver. Em pouco tempo as publicações com aplicações dos métodos de agrupamento em todo o meio científico se multiplicaram. Existem duas razões para este rápido crescimento: (1) o desenvolvimento de computadores mais capazes e velozes que tornaram viáveis o manuseio de matrizes com dimensões elevadas e, (2) o reconhecimento da classificação como um procedimento científico de fundamental importância [1].

A análise de conglomerados, como dito anteriormente, é usado em diversas áreas do conhecimento e, por esta razão, este termo pode ser encontrado na literatura como “Q analysis”, “construção de tipologias” ,”análise de classificações”, e “taxonomia numérica”. Apesar dos diferentes nomes utilizados, todos estes métodos têm em comum a classificação dos objetos ou variáveis de acordo com as suas relações naturais.

Todos os estudos de análise de conglomerados podem ser divididos em cincos passos básicos:

(1) seleção da amostra a ser agrupada;
(2) definição do conjunto de variáveis através das quais os objetos da amostra serão medidos;
(3) cálculo da similaridade entre os objetos da amostra;
(4) utilização de um método de agrupamento para agrupar os objetos similares;

(5) validação da solução resultante.

Os métodos de agrupamento são ferramentas de análise de dados úteis em diversas situações, mas algumas considerações devem ser feitas a seu respeito:

- a maioria dos métodos de agrupamento são heurísticos. Eles são usados primordialmente como uma técnica exploratória. Eles não têm base Estatística sobre a qual obtêm-se inferências estatísticas da amostra para a população;

- as soluções não são únicas pois diferentes métodos de agrupamento podem gerar diferentes soluções para um mesmo conjunto de dados. Isto ocorre porque cada método utiliza uma regra de agrupamento diferente;

- a solução resultante dos métodos é dependente das variáveis escolhidas como base para a medida de similaridade. A adição ou exclusão de variáveis relevantes podem ter um grande impacto na solução.

2 MEDIDAS DE SIMILARIDADE [2,3]

A maioria das discussões sobre análise de conglomerados enfatiza o procedimento para criar os agrupamentos, porém a escolha de uma medida de similaridade tem um papel crucial em qualquer estudo de agrupamento.

Muitos termos foram criados para descrever importantes características com relação a estimação da similaridade. Neste manual, o termo “objeto” é usado para denotar o que será agrupado, enquanto que “variável” denota a característica dos objetos usada para avaliar a similaridade. O termo “coeficiente de similaridade” é usado para descrever qualquer tipo de medida de similaridade.

O conceito de similaridade é de fundamental importância para os métodos de agrupamentos. Apesar de sua aparente simplicidade, este conceito e, especialmente, os procedimentos usados para medir a similaridade estão longe de serem simples. A similaridade entre os objetos pode ser medida de várias maneiras, dentre as quais as duas que mais se destacam são as medidas de correlação e as medidas de distância.

A estimação quantitativa da similaridade tem sido dominada pelo conceito de métrica. Objetos são representados como pontos no espaço e a similaridade entre eles é medida através da distância entre os pontos. A dimensionalidade do espaço é determinada pelo número de variáveis usadas para descrever os objetos. Para uma medida de similaridade ser considerada uma métrica, ela tem que satisfazer as seguintes condições:

- simetria: $d(x,y) = d(y,x) \geq 0$
- desigualdade triangular: $d(x,y) \leq d(x,z) + d(y,z)$
- diferenciar objetos não idênticos: se $d(x,y) \neq 0$, então $x \neq y$
- não diferenciar objetos idênticos: se $d(x,y) = 0$, então $x = y$

2.1 MEDIDAS DE CORRELAÇÃO

Medidas de correlação representam as similaridades pela correspondência de padrões de variação ao longo das variáveis. Altas correlações indicam similaridade e baixas correlações a sua falta. Uma medida de correlação de similaridade não olha para a magnitude, mas para o padrão de variação dos valores.

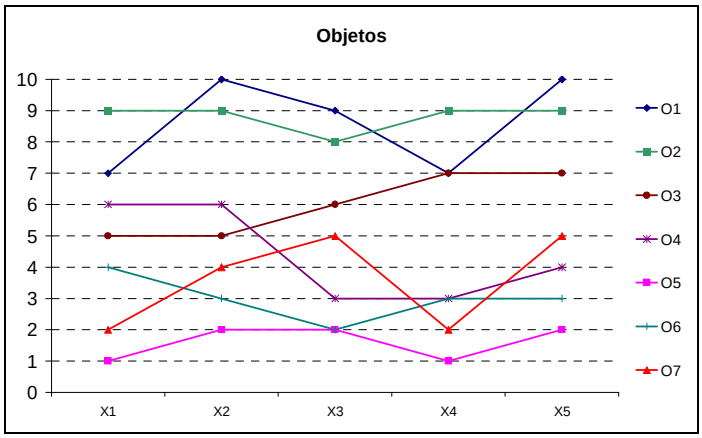

Figura 1 – Representação Gráfica de um Conjunto de Dados[2]

A tabela 1 mostra a correlação entre os sete objetos ilustrados na Figura 1. Observando-se os resultados contidos na tabela 1, verifica-se a presença de três grupos distintos.

Medida de Correlação

Objetos	1	2	3	4	5	6	7
1	1,000						
2	-0,147	1,000					
3	0,000	0,000	1,000				
4	0,087	0,516	-0,824	1,000			
5	0,963	-0,408	0,000	-0,060	1,000		
6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

Tabela 1- Cálculo da similaridade utilizando medida de correlação

O primeiro contendo os objetos 1, 5 e 7 que têm alta correlação positiva entre si e padrão semelhante. Da mesma maneira, os objetos 2, 4 e 6 também têm alta correlação positiva entre si e correlação negativa com relação aos outros objetos. O objeto 3 tem correlação pequena ou negativa com todos os outros objetos, formando deste modo um grupo por si só. Os grupos obtidos são mostrados nas Figruas 2 a 4.

Medidas de correlação são raramente usadas, pois a grande maioria das aplicações dos métodos de agrupamento visam analisar a magnitude dos objetos e não o padrão de variação destes.

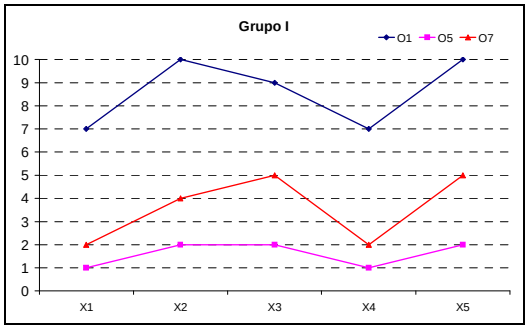

Figura 2 – Grupo I

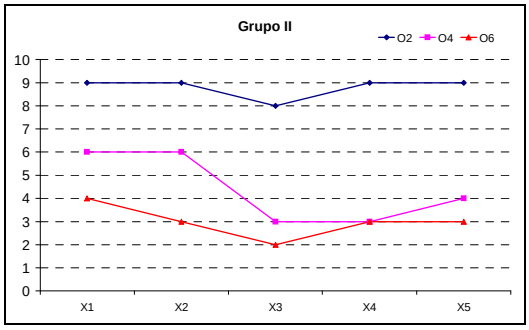

Figura 3 –Grupo II

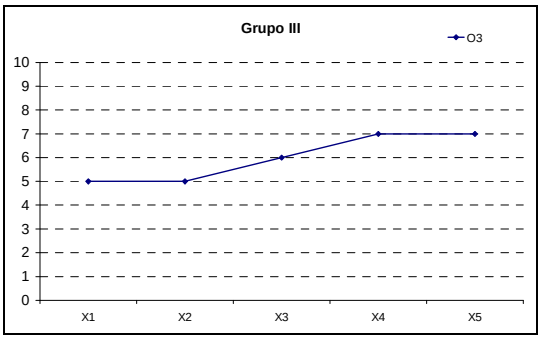

Figura 4 – Grupo III

2.2 MEDIDAS DE DISTÂNCIA

Representam a similaridade como a proximidade de um objeto a outro através de suas variáveis. Como é a medida mais intuitiva, as medidas de distâncias se tornaram as mais difundidas e utilizadas. Medidas de distância representam uma medida de não similaridade, pois quanto maior a distância entre dois objetos maior a diferença entre eles. A distância é então convertida em uma medida de similaridade através do uso de uma relação inversa.

A diferença entre as medidas de correlação e distância pode ser vista referenciando-se novamente a Figura 1. Distância foca a magnitude dos valores e retrata como objetos similares aqueles que estão próximos e, não necessariamente, com o mesmo padrão de variação. A tabela 2 mostra a *similaridade* dos objetos em questão através de uma medida de distância.

Medida de Distância

Objetos	1	2	3	4	5	6	7
1	0,000						
2	3,320	0,000					
3	6,860	6,630	0,000				
4	10,240	10,200	6,000	0,000			
5	15,780	16,190	10,100	7,070	0,000		
6	13,110	13,000	7,280	3,870	3,870	0,000	
7	11,270	12,160	6,320	5,100	4,900	4,360	0,000

Tabela 2 - Cálculo da similaridade utilizando medida de distância

Pode-se observar que os objetos 1 e 2 formam um grupo que tem como característica em comum a presença de valores elevados. Os objetos 4, 5, 6 e 7 formam outro grupo, agora com a presença de valores baixos. Novamente o grupo 3 forma um grupo exclusivo. Os grupos obtidos são mostrados nas Figuras 5 a 7. A configuração final dos grupos formados foi diferente para cada uma das medidas de similaridade utilizadas.

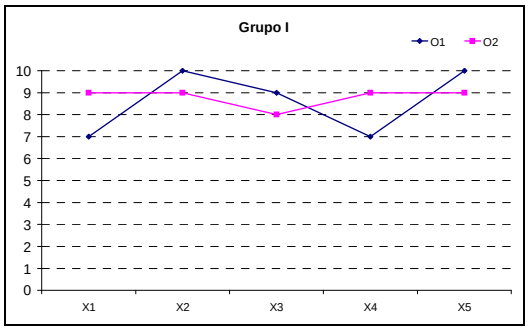

Figura 5 – Grupo I

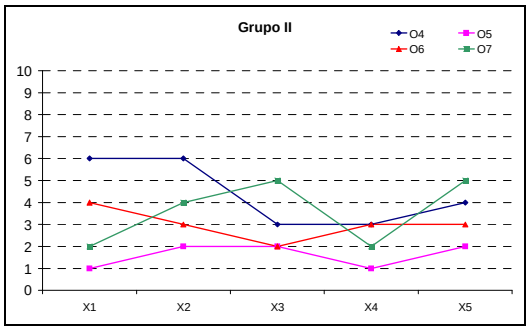

Figura 6 – Grupo II

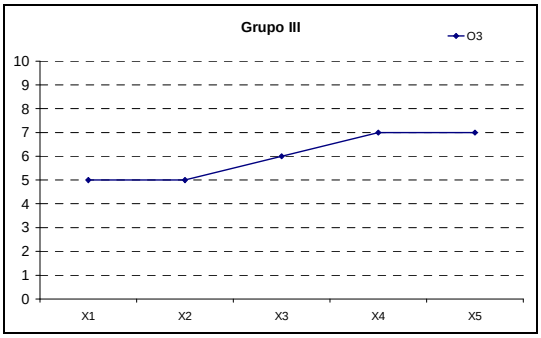

Figura 7 – Grupo III

A escolha de uma ou outra medida requer uma interpretação diferente dos resultados obtidos. Agrupamentos baseados na correlação não têm valores similares e sim padrões de variação semelhantes, enquanto que os baseados em uma medida de distância têm valores similares através das variáveis, mas os padrões de variação podem ser bem diferentes.

2.2.1 Tipos de medidas de distância

Muitas medidas de distância estão disponíveis na literatura. Entre as mais usuais está a distância Euclidiana, definida como:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

A Figura 8 é um exemplo da distância Euclidiana entre dois objetos com duas variáveis X e Y.

$$d_{xy} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

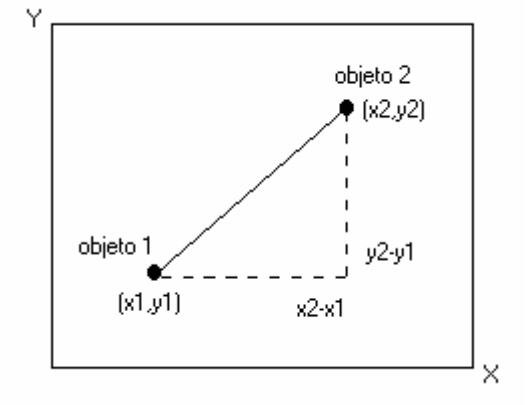

Figura 8 – Um exemplo da distância Euclidiana

Para evitar o uso da raiz quadrada, utiliza-se a distância Euclidiana quadrada, que nada mais é do que o valor da distância Euclidiana elevado ao quadrado. A distância quadrada é recomendada como medida de distância para os métodos de agrupamento Ward e para o método Centróide, que serão vistos mais tarde.

Outras opções não baseadas na distância Euclidiana também estão disponíveis. Uma das mais usadas é a distância Manhattan, também conhecida como City-Block. Esta medida tem a seguinte definição:

$$d_{ij} = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$$

Outra medida de distância é a métrica Minkowski:

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} | x_{ik} - x_{jk} |^{r}}$$

para r=1, d_{ij} é a distância City-Block entre dois objetos com dimensão p. Para r=2, d_{ij} transforma-se na distância Euclidiana.

As seguintes medidas de distância são definidas somente para variáveis não negativas. São elas a métrica Canberra e o coeficiente Czekanowski, definidas logo abaixo:

$$d_{ij} = \sum_{k=1}^{p} \frac{| x_{ik} - x_{jk} |}{x_{ik} + x_{jk}} \qquad \text{Canberra}$$

$$d_{ij} = 1 - \frac{2\sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} x_{ik} + x_{jk}} \qquad \text{Czekanowski}$$

Um problema visto por todas as medidas de distância que usam dados não padronizados é a variação da solução encontrada pelo método de agrupamento quando a escala das variáveis analisadas são diferentes. No sentido de amenizar este problema, antes do cálculo da distância, as variáveis são padronizadas.

Uma medida de distância que já incorpora um procedimento de padronização é a distância Mahalanobis (D^2), definida por:

$$d_{ij} = (x_i - x_j)' \sum^{-1} (x_i - x_j)$$

onde Σ é a matriz de covariância dos objetos. Quando a correlação entre as variáveis é zero, a distância Mahalanobis é equivalente a distância Euclidiana quadrada.

Diferentes medidas de distância ou mudanças na escala das variáveis podem gerar soluções distintas. Por isso é aconselhável que se experimente várias medidas, e que se compare os resultados obtidos com valores teóricos ou conhecidos previamente.

2.3 PADRONIZAÇÃO DOS DADOS

Muitas medidas de distância são sensíveis a variações na escala ou na magnitude entre as variáveis. A forma mais comum de padronização é a conversão de cada variável em valores padrão pelo decréscimo do valor médio e a divisão pelo seu respectivo desvio padrão. Esta transformação elimina a influência introduzida pelo uso de diferentes escalas nas variáveis usadas na análise. Não existe diferença nos valores padrão quando a escala é alterada.

3 MÉTODOS DE AGRUPAMENTO [3]

O principal objetivo quando se usa a análise de conglomerados é encontrar grupos de objetos similares em um conjunto de dados de tal forma que as variâncias entre os grupos seja máxima, e dentro deles, mínima. Estes agrupamentos também são conhecidos por “clusters”.

3.1 MÉTODOS DE AGRUPAMENTO HIERÁRQUICO

Existem basicamente dois tipos de procedimentos no método hierárquico – aglomerativo e divisivo.

O procedimento aglomerativo é baseado em uma série de fusões. Inicialmente cada objeto está em um agrupamento exclusivo, isto é, existem tantos grupos quantos forem os objetos. Estes agrupamentos são fundidos de acordo com a sua similaridade. Grupos similares são fundidos e formam, desta maneira, um novo agrupamento. Este processo vai se repetindo até que no final todos os objetos estejam em um único agrupamento. A Figura 9 ilustra este processo.

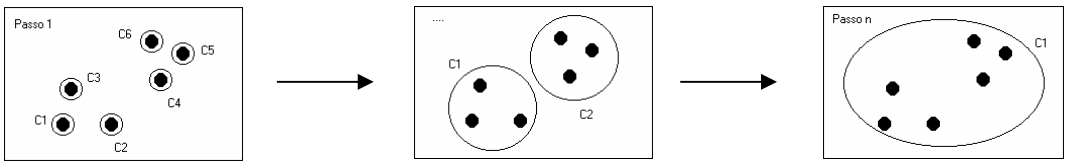

Figura 9 – Processo Aglomerativo

No procedimento divisivo o processo acontece de maneira contrária, isto é, todos os objetos estão inicialmente em um mesmo grupo então estes começam a ser divididos em subgrupos. Os objetos em um subgrupo são diferentes dos objetos de um outro subgrupo. O processo continua até que cada objeto esteja sozinho em um grupo. A Figura 10 ilustra o processo divisivo.

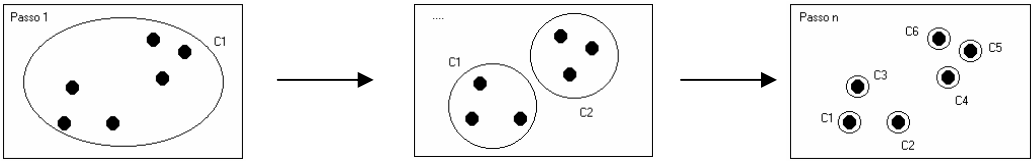

Figura 10 – Processo Divisivo

Dentre os métodos de agrupamento hierárquicos, o método hierárquico aglomerativo é largamente utilizado. A seguir estão descritos os passos para se agrupar N objetos através do método hierárquico aglomerativo:

❶ Comece com N agrupamentos cada um com um único objeto e construa a matriz de similaridade NxN
   ➔ matriz de similaridade é aquela cujos elementos a_{ij} são medidas de similaridade entre os objetos i e j

❷ Ache na matriz a maior similaridade entre dois agrupamentos
   ➔ suponha U e V os grupos mais similares

❸ Forme o agrupamento UV (U e V). Atualize a matriz de similaridade
   ➔ retire linhas e colunas pertinentes aos grupos U e V e adicione uma linha e uma coluna com a similaridade entre o grupo UV e os demais

❹ Repita os passos 2 e 3 N-1 vezes
   ➔ no final do algoritmo todos os objetos estarão em um único grupo

Este algoritmo é válido para todos métodos hierárquicos aglomerativos. O algoritmo pode ser interrompido quando um número pré determinado de agrupamentos estiver realizado ou quando a distância entre os grupos atingir um valor máximo (ou mínimo no método divisivo) definido pelo usuário.

Os resultados do método aglomerativo podem ser mostrados em um diagrama bidimensional conhecido como *dendograma*. O dendograma ilustra as fusões que vão sendo feitas sucessivamente a cada passo. A Figura 11 permite observar que o método aglomerativo se move de baixo para cima (grupos se fundindo).

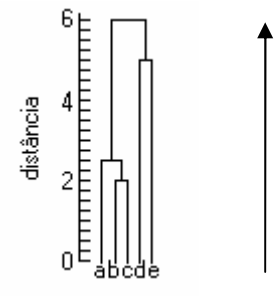

Figura 11 - Dendograma

A seguir estão descritos os métodos hierárquicos aglomerativos mais utilizados. A diferença entre eles está na maneira de se calcular a similaridade entre os agrupamentos.

3.1.1 Método de Encadeamento (Linkage Method)

- Encadeamento Simples (Single Linkage)

As entradas deste algoritmo podem ser distâncias ou similaridades entre pares de objetos. Grupos serão formados através da fusão dos objetos mais próximos (menor distância). A distância entre dois grupos quaisquer é a menor distância entre um objeto de um grupo com outro pertencente ao outro grupo, como mostra a Figura 12. Neste caso a menor distância é entre o objeto 2 (grupo 1) e o objeto 4 (grupo 2).

$$d = d_{24}$$

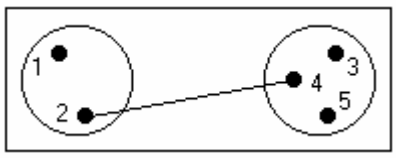

Figura 12 – Distância entre agrupamentos através do método Encadeamento Simples

- Encadeamento Completo (Complete Linkage)

O procedimento seguido pelo método Encadeamento Completo é similar ao utilizado pelo Encadeamento Simples, exceto que o critério é baseado na máxima distância. A cada estágio a distância entre os agrupamentos é definida como sendo a maior distância que existe entre um objeto de um grupo com os demais de outro grupo, conforme mostra a Figura 13.

$$d = d_{13}$$

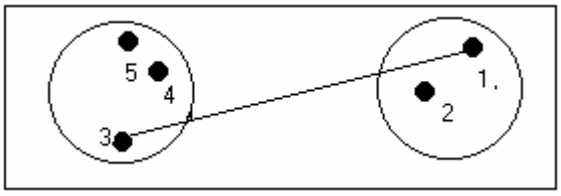

Figura 13 - Distância entre agrupamentos através do método Encadeamento Completo

- Encadeamento Médio (Average Linkage)

Este trata a distância entre dois grupos como a distância média entre todos os objetos de um grupo e todos os objetos do outro grupo. Este método não depende de valores extremos, e a fusão é baseada em todos os membros do grupo e não apenas em um par de objetos como nos métodos de Encadeamento Simples e Completo. A Figura 14 ilustra este método.

$$d = \frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$

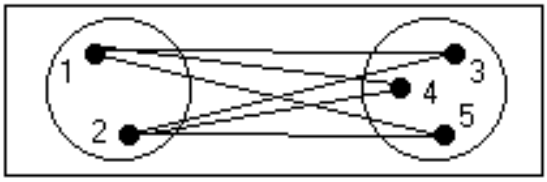

Figura 14 - Distância entre agrupamentos através do método Encadeamento Médio

3.1.2 Método Centróide

No método Centróide a distância entre dois grupos é a distância entre os seus centróides, como mostrado na Figura 15. O centróide é uma representação do valor médio dos objetos em um grupo. Neste método, cada vez que um novo objeto é aglomerado, o centróide daquele agrupamento deve ser recalculado. A vantagem é que este método é menos afetado pela presença de objetos espúrios do que os métodos hierárquicos que utilizam o encadeamento simples e completo.

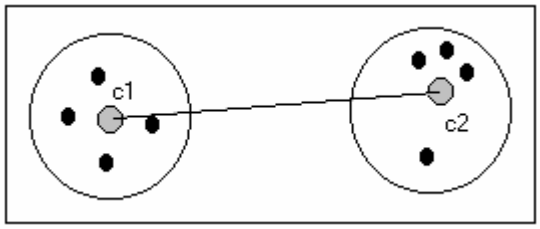

Figura 15 - Distância entre agrupamentos através do método Centróide

3.1.3 Método Ward

Este método se baseia na minimização da "perda de informação" quando se funde dois grupos. Perda de informação se refere à soma dos quadrados dos desvios (ESS) dos grupos. Primeiramente, é testada todas as possibilidades de junção dos grupos, mas somente irá se efetivar aquela que produzir o menor acréscimo na ESS.

$$ESS = \sum_{i=1}^{NC} \sum_{j=1}^{NOi} (x_j - \bar{x}_i)'(x_j - \bar{x}_i)$$

onde NC é o número de grupos e NO_i é o número de objetos no grupo i e $\bar{x}_i$ é o centróide do grupo i.

Este método, que é o percursor hierárquico dos métodos não hierárquicos. Para ilustrar este método será utilizada a Figura 16.

1. Inicialmente cada objeto forma um grupo, logo a soma dos quadrados dos desvios (ESS) será nula.

2. Todas as possibilidades de agrupamento, fusão de grupos 2 a 2, são testadas e para cada uma é calculada o acréscimo na ESS. O agrupamento que será efetivamente realizado será aquele que gerar o menor acréscimo na ESS.

3. O passo 2 será repetido até que todos os grupos sejam fundidos em um único grupo.

No exemplo da Figura 16, suponha que a fusão dos grupos 1 e 2 tenha gerado o menor acréscimo em ESS, logo os grupos 1 e 2 serão fundidos pois esta configuração foi a que gerou menor perda de informação.

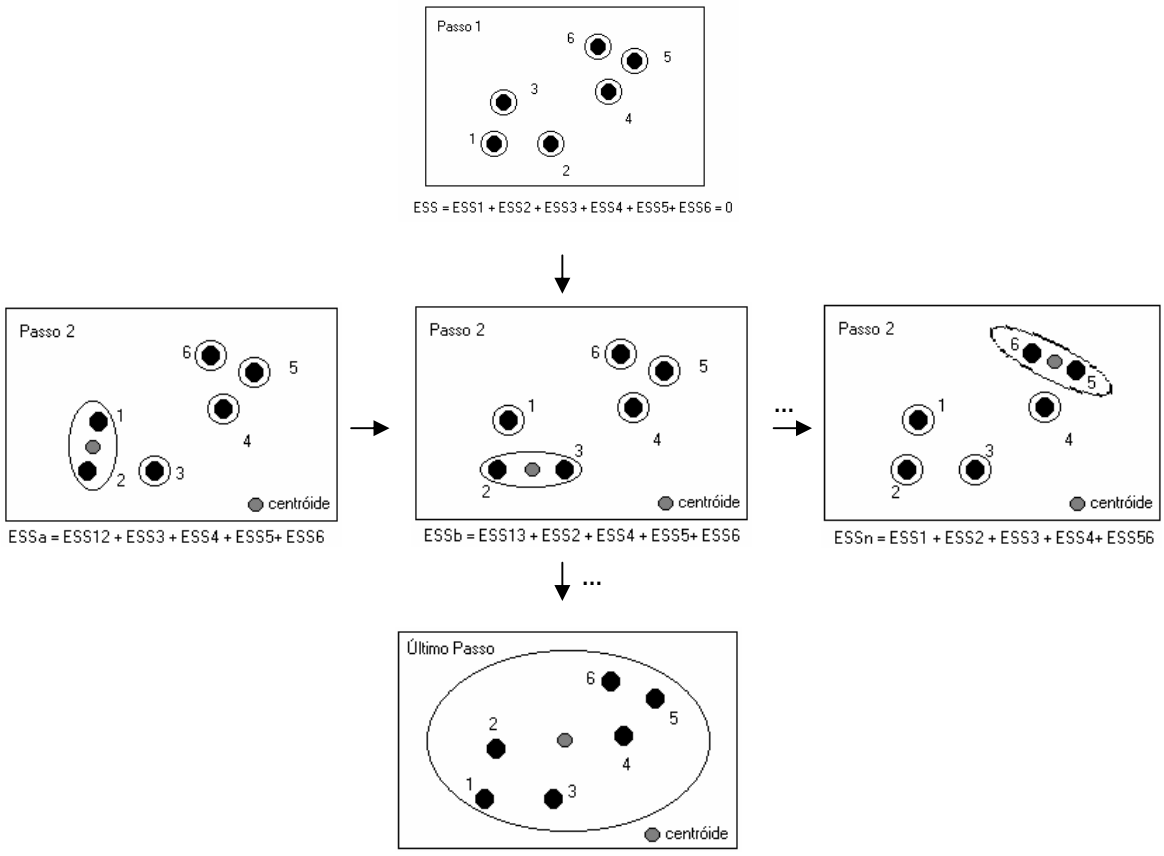

Figura 16 – Método Ward

Observações Finais

Nos métodos hierárquicos não há realocação dos objetos. Se algum objeto for incorretamente agrupado em um estágio anterior não há possibilidade de realocá-lo em um estágio posterior. Uma outra desvantagem destes métodos é a necessidade da construção da matriz de similaridade, o que torna estes métodos inviáveis para grandes conjuntos de dados.

3.2 MÉTODO DE AGRUPAMENTO NÃO HIERÁRQUICO

Métodos de agrupamento não hierárquico são utilizados para agrupar objetos em um número específico de grupos. Como a matriz de similaridade não precisa ser calculada, estes métodos podem ser aplicados a grandes conjuntos de dados. Dentre os métodos não hierárquicos o mais conhecido é o K-Means.

O primeiro passo deste método é formar uma partição inicial aleatória no conjunto de dados. O número de grupos deve ser estabelecido previamente. O próximo passo é o cálculo dos centróides destes grupos. Então, as distâncias (geralmente distância Euclidiana) entre todos os objetos e os centróides dos grupos são calculadas. Os objetos são realocados para o grupo que tiver o centróide mais próximo. Este último passo é repetido até que não haja mais realocações de objetos. Vale a pena lembrar que toda vez que um objeto for realocado os centróides devem ser recalculados.

Ao invés de inicializar o processo com uma partição do conjunto de dados, pode-se definir pontos aleatoriamente que servirão como centróides para os agrupamentos. Estes pontos sorteados podem ser quaisquer, ou pontos pertencentes ao conjunto de dados a ser agrupado.

Uma versão simplificada deste método é apresentada abaixo:

❶ Divida os N objetos em K agrupamentos
   ➔ partição inicial ou especificação de K centróides iniciais
❷ Realoque um objeto para o grupo cujo centróide é o mais próximo deste objeto.

- ➔ recalcule o centroíde do grupo que recebeu e que perdeu o objeto

❸ Repita o passo 2 até que não haja mais realocações

Observações Finais

O maior problema deste método é como selecionar inicialmente os núcleos dos agrupamentos. Uma vez que este processo é aleatório, a cada novo sorteio o método pode resolver o problema de uma maneira diferente.

4 MODELO DE GERAÇÃO DE SÉRIES SINTÉTICAS DE VAZÕES PARA SISTEMAS HIDROELÉTRICOS USANDO ANÁLISE DE CONGLOMERADOS

4.1 Aspectos Gerais

Um problema encontrado no projeto e na operação de sistemas hidrotérmicos é a consideração apropriada da variabilidade hidrológica. As técnicas quantitativas usadas para lidar com a variabilidade hidrológica geralmente requerem um modelo estocástico de afluências naturais para cada usina hidráulica. Este modelo estocástico pode ser usado para criar um grande número de sequências de afluências equiprováveis. O uso de séries sintéticas se dá devido ao fato de os registros históricos geralmente não serem suficientemente longos para fornecer uma boa estimativa dos riscos envolvidos na operação de um reservatório de um dado projeto, podendo ainda não incluir os casos mais extremos de cheias e secas. Sequências hidrológicas sintéticas são amplamente usadas em conjunto com modelos de simulação para avaliar projetos propostos e estratégias de operação para reservatórios. O problema chave em todos os usos de sequências hidrológicas sintéticas é escolher um modelo de séries temporais estocásticas que garanta uma proximidade entre as sequências de afluências histórica e sintética em termos das estatísticas as quais poderão influenciar o projeto ou a operação do reservatório em questão.

Um modelo de geração de vazões mensais foi desenvolvido para o sistema hidroelétrico Brasileiro baseado em uma modelagem auto-regressiva periódica das séries hidrológicas (modelo GEVAZP). O modelo pode ser aplicado diretamente às afluências naturais evitando o uso de transformações especiais tal como a transformação Box-Cox. A geração de afluências negativas pode também ser evitada, e uma metodologia para assegurar a adequação do modelo foi desenvolvida [4].

Um modelo foi desenvolvido para calcular estratégias ótimas de operação a curto prazo em um sistema hidroelétrico (modelo DECOMP) com diversos reservatórios, baseado na programação dinâmica dual estocástica [5]. Em geral, a evolução dos

reservatórios pode ser representada por uma estrutura de árvore, Figura 17, onde cada bifurcação corresponde a um vetor de afluências alternativas.

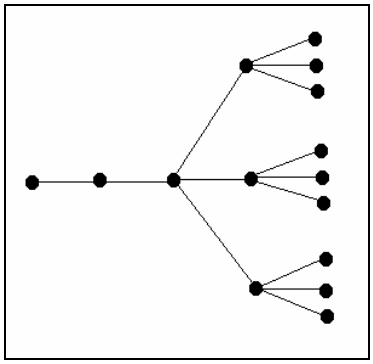

Figura 17 - Estrutura de árvore

Devido à restrições de tempo computacional é interessante trabalhar com o menor número de cenários hidrológicos possível e esta limitação não é ideal para o esquema de geração de Monte Carlo. Um número reduzido de séries não é suficiente para caracterizar bem um processo estocástico.

Aplicadas à um grande número de cenários hidrológicos gerados, as técnicas de aglomeração proporcionam a escolha de um conjunto representativo de cenários. Este conjunto representativo de cenários hidrológicos irá conter toda a informação necessária para representar o processo estocástico de vazões. Os cenários que fazem parte deste conjunto representativo são obtidos através de agrupamento de cenários semelhantes e possuem características similares aos demais componentes do grupo em que estão localizados. Desta forma, estes cenários resultantes são não equiprováveis.

As técnicas de aglomeração foram incorporadas no programa de geração de séries sintéticas de energias e vazões (GEVAZP/CEPEL) utilizado pelos modelos de planejamento da operação de médio e curto prazo do sistema hidroelétrico Brasileiro, permitindo a redução do número de cenários hidrológicos gerados através de agrupamento de cenários similares. Os resultados obtidos mostram que os cenários gerados pelo modelo de geração em conjunto com as técnicas de aglomeração são

representativos e desta forma, conseguem caracterizar bem o processo estocástico de vazões.

4.2 Método de Agrupamento Utilizado

O método escolhido foi o método de agrupamento não hierárquico K-Means, pois o problema em questão é agrupar um grande número de cenários hidrológicos. Como já visto no item 3.2, este método não hierárquico é ideal para trabalhar com grandes conjuntos de dados pois não requer o cálculo da matriz de similaridade.

A cada estágio do horizonte, o processo de aglomeração é incorporado à geração de cenários hidrológicos sintéticos.

Supondo-se, por exemplo, que a cada estágio um nó dê origem à três novos nós na estrutura da árvore, como ilustrado na Figura 17, e que para todos os meses o processo estocástico de vazões seja representado por um modelo auto-regressivo de ordem três. No primeiro estágio serão geradas 1000 cenários equiprováveis de vazões multivariadas a partir de uma única sequência de vazões multivariadas conhecidas (A_3, A_2, A_1). A seguir, o processo de aglomeração é aplicado resultando em três cenários não equiprováveis de vazões multivariadas (A_4^1, A_4^2, A_4^3), Figura 18. No segundo estágio, para a sequência (A_4^1, A_3, A_2), são gerados 1000 cenários equiprováveis de vazões multivariadas. Do processo de aglomeração três cenários não equiprováveis de vazões multivariadas são produzidos ($A_{5,1}^1$, $A_{5,1}^2$, $A_{5,1}^3$), Figura 18a. Ainda neste estágio, a partir da sequência (A_4^2, A_3, A_2) são gerados 1000 cenários equiprováveis de vazões multivariadas. Da mesma forma, o processo de aglomeração é aplicado e três cenários não equiprováveis de vazões multivariadas são obtidos ($A_{5,2}^1$, $A_{5,2}^2$, $A_{5,2}^3$), Figura 18b. Por último, a partir da sequência (A_4^3, A_3, A_2), são gerados 1000 cenários de vazões multivariadas. O processo de aglomeração é novamente aplicado resultando em três cenários não equiprováveis de vazões multivariadas ($A_{5,3}^1$, $A_{5,3}^2$, $A_{5,3}^3$), Figura 18c. Este procedimento é repetido para cada um dos estágios do horizonte, Figura 19.

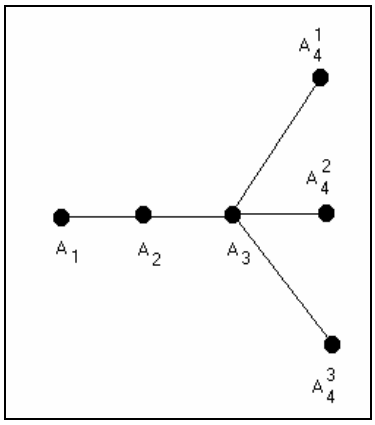

Figura 18 – Caso Exemplo (1º per. Estocástico)

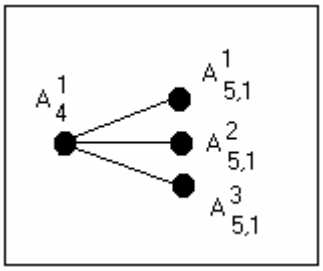

Figura 18a - Caso Exemplo (2º per. Estocástico)

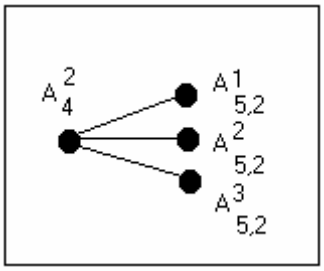

Figura 18b - Caso Exemplo (2º per. Estocástico)

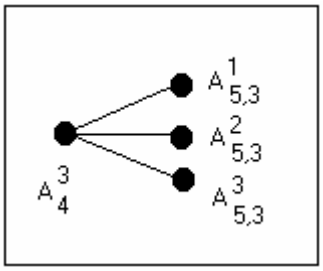

Figura 18c - Caso Exemplo (2º per. Estocástico)

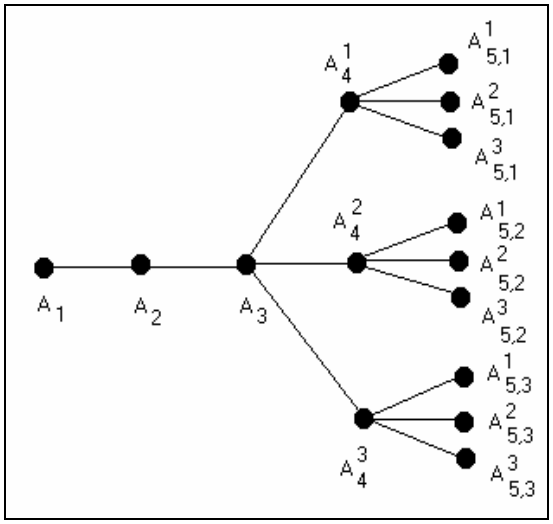

Figura 19 – Árvore Completa

Caso Exemplo

No processo de aglomeração, dos 1000 cenários equiprováveis de vazões multivariadas apenas alguns cenários (número este selecionado previamente pelo usuário) permanecerão formando o conjunto de cenários não equiprováveis de vazões multivariadas representativos do conjunto gerado originalmente. Deste processo não resultam novos valores de vazões multivariadas. Sendo assim, a correlação temporal ou serial não é alterada pois a cada estágio (mês) é gerado novamente todo o conjunto de cenários necessários para descrever o processo estocástico (neste caso 1000 cenários). Note que a geração deste conjunto de cenários é realizada para cada nó do estágio corrente, e isto é feito levando-se em conta o passado destes nós.

A Figura 20 ilustrado o processo completo de aglomeração incorporado ao modelo de geração de séries sintéticas.

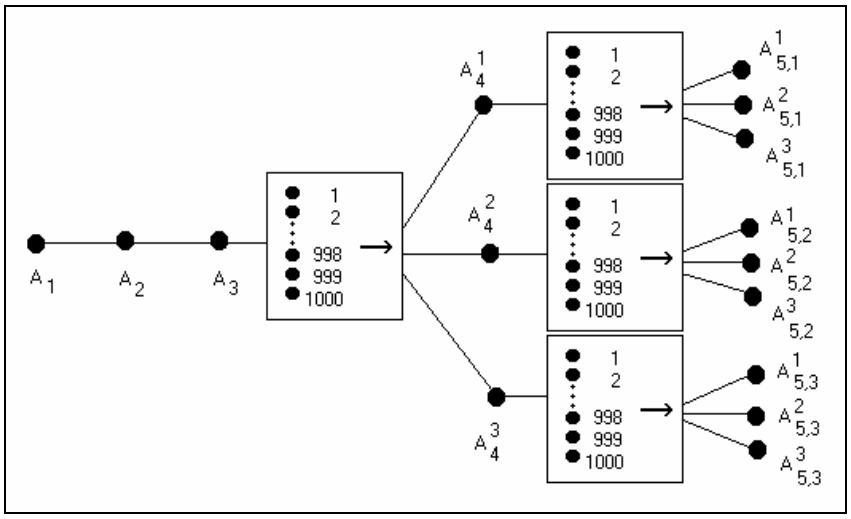

Figura 20 – Geração de Séries Sintéticas com a Utilização de Técnicas de Aglomeração

4.3 Medida de Distância

A medida de distância utilizada para calcular a similaridade entre os objetos (cenários hidrológicos multivariados) e os grupos foi a distância Euclidiana, a mais usual dentre todas as descritas anteriormente.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

onde d_{ij} é a distância entre os cenários i e j. A variável x_{ik} é a vazão da usina hidráulica k no cenário i e a variável x_{jk} é a vazão da usina hidráulica k no cenário j. O número de usinas hidráulicas no vetor de vazões multivariadas x é p.

Os dados de entrada (vazões afluentes) foram padronizados pela média e desvio padrão mensal, esta transformação homogeniza as escalas nas variáveis usadas.

$$x'_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$$

onde x'_{ik} é o vazão padronizada do cenário i da usina hidráulica k, $\mu_k = \frac{1}{n}\sqrt{\sum_{i=1}^{n} x_{ik}}$ é a média mensal da usina k, n é o número total de cenários hidrológicos gerado sinteticamente e $\sigma_k = \frac{1}{n}\sqrt{\sum_{i=1}^{n} (x_{ik} - \mu_k)^2}$ é o desvio padrão mensal desta mesma usina.

4.4 Determinação dos Cenários Representativos

O processo de agrupamento foi inicializado através do sorteio de sementes aleatórias para representar os centróides dos grupos. Estes pontos aleatórios foram sorteados diretamente do conjunto de entrada. Logo são cenários hidrológicos do conjunto original gerado sinteticamente. Desta maneira, pode-se garantir que nenhum grupo ficará vazio.

Nos passos seguintes até a convergência do processo de aglomeração, o centróide dos grupos (c_k) será o ponto médio destes grupos, isto é :

$$c_k = \frac{1}{no_k} \sum_{i=1}^{no_k} x_i$$

onde no_k é o número de cenários hidrológicos pertencentes ao grupo k. Os cenários hidrológicos x são vetores de vazões multivariadas.

Após a convergência do processo, o centróide dos grupos será o objeto mais próximo do ponto médio destes grupos. A Figura 21 ilustra como é escolhido o representante de cada grupo formado.

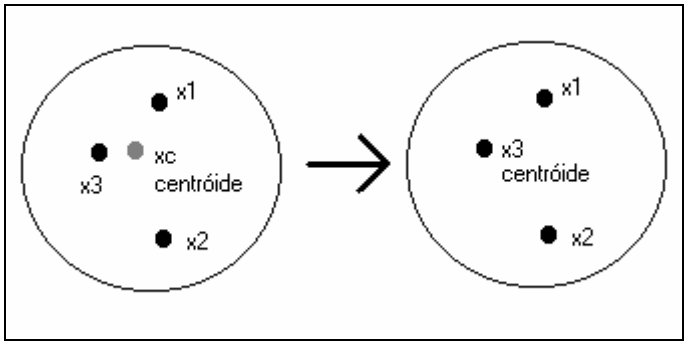

Figura 21 – Escolha do Cenário Representativo

Tomando a Figura 21 como exemplo, supondo que após a convergência do processo de aglomeração os cenários hidrológicos x_1, x_2 e x_3, cada um composto por vazões afluentes a todas as usinas hidráulicas, formem um grupo cujo ponto médio é representado por pelo vetor x_c. Observe que o cenário x_3 é o mais próximo do ponto médio real do grupo, logo x_3 será o cenário hidrológico escolhido para representar o grupo.

4.5 Correlação Espacial

Cada cenário hidrológico considerado no processo de agrupamento, é composto por vazões afluentes a cada usina hidráulica considerada, que foram geradas de forma a preservar a correlação espacial histórica, conforme descrito no Manual de Referência

do Modelo GEVAZP, capítulo 2. Assim, os cenários hidrológicos selecionados como representativos da amostra original são alguns dos 1000 cenários gerados originalmente, não havendo troca de vazões entre cenários e usinas hidráulicas. Portanto, a correlação espacial entre as vazões às usinas hidráulicas não sofre alteração.

5 CONCLUSÃO

Um processo estocástico é totalmente descrito pelo conjunto de todas as séries temporais que o compõe e pode ser bem caracterizado através de um grande número de séries sintéticas.

Neste relatório, foram descritas técnicas de agregação para selecionar um número reduzido de cenários sintéticos de vazões para o modelo de otimização do planejamento da operação a curto prazo (DECOMP), que representem adequadamente o processo estocástico de vazões, diminuindo assim o esforço computacional do modelo.

6 REFERÊNCIAS

(1) Aldenderfer, M. ;Blashfield, R. – “Cluster Analysis”, Sage Publication, 1984.

(2) Hair, J; Anderson, R; Tathan, R; Black, W. – “Multivariate Data Analysis”, Ed. Prentice Hall, 1998.

(3) Johnson, R.; Wichern, D. – “Applied Multivariate Statistical Analysis”, Ed. Prentice Hall, 1998.

(4) Maceira, M.E.P.; Bezerra,C.M.B. – “Modelo de Geração de Séries Sintéticas de Energias e Vazões (Gevazp) - Manual de Metodologia”, Relatório Técnico CEPEL nº DPP/PEN – 083/2000.

(5) Costa,J.P.; Prado,S.; Binato,S. – “Modelo DECOMP - Manual de Metodologia”, Relatório Técnico CEPEL nº DPP/PEL– 639/99.