UTILIZAÇÃO DE MÉTODOS HIERÁRQUICOS DE AGRUPAMENTO EM CISTERNAS DE PLACAS NO PAJEÚ PERNAMBUCANO

Manoel Rivelino Gomes de OLIVEIRA[1]
Moacyr Cunha FILHO[1]
David Venancio da CRUZ[1]
José Ramon Barros CANTALICE[1]
Ana Patrícia Siqueira Tavares FALCÃO[2]

- Resumo: A qualidade da água é considerada como uma série de características físicas, químicas e biológicas. Estas qualidades devem ser estudadas de forma minuciosa, levando-se em consideração que podem conter substâncias nocivas a saúde humana. Assim, o programa um milhão de cisternas do Governo Federal surgiu como uma alternativa que tem se consagrado em toda região semiárida do Nordeste Brasileiro por sua rapidez na construção, o baixo custo de execução da obra e pala a maneira fácil de captar água das chuvas que são bastante reduzidas no semiárido Brasileiro. Este estudo objetivou agrupar estas cisternas de acordo com suas similaridades ou dissimilaridades segundo as variáveis mensuradas, utilizando os métodos estatísticos multivariado de agrupamentos hierárquicos. A solução obtida pelo método de Ward apresentou a melhor distribuição das cisternas entre os grupos além de ter sido a de maior variabilidade média entre os grupos. Os métodos do vizinho mais distante a da média das distâncias obtiveram as menores variâncias internas dos grupos, enquanto que o método da média das distâncias obteve a maior correlação cofenética. Portanto, as soluções obtidas pelos métodos hierárquicos aglomerativos sugeriram bons agrupamento das cisternas segundo a qualidade da água no Pajeú.

- PALAVRAS-CHAVE: Agrupamentos hierárquicos; captação de água; semiárido brasileiro.

1 Introdução

A estiagem é um fenômeno natural conhecido no mundo inteiro pela ausência de precipitação. Modelos climáticos globais prevê aumento na frequência e intensidade de eventos climáticos extremos, incluindo secas severas (CHERWIN e KNAPP, 2012). Este fenômeno tem um poder catastrófico principalmente por causar danos aos recursos hídricos ou mesmo pela causa da falta direta de água, em regiões assoladas pela seca ou estiagem.

[1]Universidade Federal Rural de Pernambuco, UFRPE, Programa de Pós-Graduação em Biometria e Estatística Aplicada, CEP: 52171-900, Recife, PE, Brasil. E-mail: rivelino_gomes@hotmail.com; moacyr@deinfo.ufrpe.br; davidvenacio@hotmail.com; cantalic@terra.com.br
[2] Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco, IFPE, Pró-Reitora de Pesquisa e Inovação, CEP: 55602-970,Vitória de Santo Antão, PE, Brasil, E-mail: apstfalcao@hotmail.com.

As regiões mais afetadas pelo fenômeno da estiagem são aquelas com clima semiárido. As zonas áridas são aquelas cujos índices pluviométricos em geral não são elevados. A precipitação média anual nas regiões semiáridas varia de 300 a 800 mm. Assim a evapotranspiração é uma componente chave do equilíbrio da água (BUNTING *et al*., 2014).

A Caatinga compreende os estados do Nordeste e ainda o norte de Minas Gerais em uma área de 73.683.649 ha. A ocorrência de secas periódicas e regimes pluviométricos sazonais estabelecem regimes intermitentes aos rios e vegetação sem folhas em boa parte do ano, isto por que o semiárido do nordeste Brasileiro é propenso a secas, experimentando uma curta estação chuvosa praticamente em todos os anos (HASTENRATH, 2012).

Segundo Gleick (2010), a compreensão do ciclo hidrológico melhorou a capacidade de gestão dos recursos hídricos e outros recursos naturais que são limitados para uma população crescente. O principal problema decorrente da escassez de chuvas é a falta de água em regiões onde o clima predominante é o semiárido como no Nordeste brasileiro.

Condicionada pelas chuvas escassas, a falta de água está aliada ainda a sua forma de armazenamento, isto é, quando a água está em determinados reservatórios pode ocasionar em grandes percas por evapotranspiração e infiltração assim como problemas relacionados a poluição. A evapotranspiração acarreta um déficit de água subterrânea e reduções das vazões (CAPDEVILA *et al*., 2011).

Dentre as diversas formas de armazenar água no semiárido destacam-se açudes, barragens, barreiros, barragens subterrâneas, tanques de pedra, poços artesianos, cisternas de placas, entre outras. Cisternas de placas são reservatórios de captação da água da chuva, cuja finalidade é armazenar água da chuva para o consumo doméstico. Apesar de mais de dois milhões de pessoas que vivem em regiões semiáridas do Brasil consumir água da chuva armazenada em cisternas, pouca informação está disponível sobre a qualidade da água (ALVES *et al*., 2014).

Esta forma de armazenar água é amplamente utilizada nas famílias rurais residentes na região semiárida durante o período de estiagem ou quando não há disponibilidade de água para o consumo residencial. O monitoramento da qualidade da água é uma das ferramentas para o desenvolvimento sustentável no fornecimento de informações importantes para a gestão da água (JALALI, 2009).

Em muitas partes do mundo, a contaminação das águas vem sendo um risco generalizado a saúde humana, especialmente em países economicamente menos desenvolvidos (NNANE *et al*., 2011). Dentre os parâmetros da qualidade da água os abordados para estudo neste trabalho foram: Cátions, Ânions, Potencial hidrogeniônico (pH), Sólidos dissolvidos totais (S.D.T .), Coliformes fecais e Coliformes totais. Estas seis variáveis de qualidade da água foram selecionadas e medidas para posteriores análises.

Este trabalho objetivou agrupar estas cisternas considerando variáveis ou parâmetros de qualidade da água de acordo com suas similaridades ou dissimilaridades utilizando os métodos estatísticos multivariado de agrupamentos hierárquicos, onde os agrupamentos formados darão suporte para observar as cisternas que tem águas similares em pontos distintos da área de estudo, e ainda podem-se detectar alguns reservatórios que poderiam ter água imprópria para o consumo humano se estes formam grupos muito diferentes dos demais.

Análise de agrupamento tem aplicação em várias áreas, como se pode observar em Xu e Furey (2007), para classificar solventes farmacêuticos para estudo de cristalização. Bellis *et al* (2012), utilizaram técnicas de agrupamento para identificar localidades com múltiplos problemas de saúde e necessidades sociais. Martinent *et al* (2013) também utilizaram analise de agrupamento para identificar perfis efetivos de atletas antes e durante a competição para examinar as diferenças entre esses perfis no enfretamento e cumprimento de metas desportivas em uma amostra de 360 atletas. A análise de agrupamento também foi utilizada por Passarino *et al* (2007) para definir fenótipos de envelhecimento humano. Já Clifford *et al* (2011), comparam métodos de agrupamento para investigação de dados de metilação do genoma. Kassomenos *et al* (2010) utilizaram técnicas de comparação de agrupamento estatísticos para a classificação das trajetórias atmosféricas modeladas. A análise de agrupamento funcional ainda foi utilizada no estudo dos dados de qualidade de água na Escócia (HAGGARTY *et al*., 2012).

2 Material e métodos

A área de estudo está localizada nas comunidades de Serra Grande: Poço do Serrote, Poldrinho, Catolé e Três Irmãos. A área é próxima a Serra da Lagartixa, entre as cidades de Serra Talhada (Figura 1) e Floresta, pertencente a região do Alto Sertão do Pajeú, que apresenta ambientes semiáridos do Estado de Pernambuco, de coordenadas geográficas de 38°23'55.51" longitude Oeste e 8°07'06.72" latitude Sul.

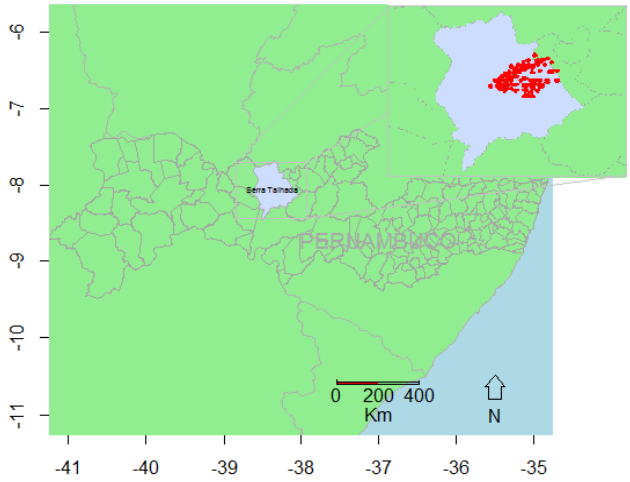

Figura 1 - Localização Geográfica do município de Serra Talhada – PE.

O clima da região, de acordo com a classificação de Köppen, se enquadra no tipo Bwh, denominado semiárido, quente e seco, com chuvas de verão e outono com pluviosidade média anual de 647 mm, para o período de 1912 a 1991 (SUDENE, 1990) e temperatura média anual superior a 25 ^{o}C.

Uma mudança proposta por uma rede de organizações sociais do semiárido chamada Articulação do Semiárido (ASA), na forma de abordar a questão da seca, visa a convivência com o clima semiárido e não mais o combate à seca. A primeira grande proposta da ASA para o semiárido é o programa de construção de um milhão de cisternas para captação de água da chuva na zona rural (FERREIRA, 2009).

As cisternas de placas são reservatórios de captação de água da chuva, para o consumo doméstico nas comunidades rurais do semiárido nordestino, com a finalidade de suprir o consumo doméstico em períodos de longas estiagens. As cisternas mais comumente construídas no sertão nordestino têm capacidade para 16.000 litros. Quantidade suficiente para suprir as necessidades domésticas de uma família com quatro pessoas no período de cinco meses se considerarmos um consumo diário de 25 litros d'água por pessoa.

As cisternas de placas comparados a outros sistemas de captação de água, apresenta um baixo custo, um curto período de construção por sua simplicidade e praticidade das obras (GNADLINGER, 2008). Este sistema deixa a população menos dependente de carros-pipa, e ainda traz como benefício a redução das verminoses proporcionando uma água de boa qualidade para as famílias beneficiadas.

O nome cisterna de placa surgiu por causa do material utilizado em sua construção, isto é, placas de cimentos pré-moldadas, confeccionadas a parte, antes da montagem da estrutura da cisterna. A cisterna apresenta forma cilíndrica (Figura 2), com 3,40 m de diâmetro e 1,80 m de profundidade, devendo ficar 1,30 m enterrada no solo para segurança de sua estrutura.

Figura 2 - Cisternas de placas instaladas em residência na comunidade de Serra Grande, em Serra Talhada – PE. Fonte: Lima, 2014.

As amostras da água destinadas para a análise foram realizadas em 100 cisternas de novembro de 2014 a março de 2015, período esse considerado de maior incidência de eventos pluviométricos nas regiões semiáridas, em especial no sertão pernambucano. Estas precipitações nas regiões semiáridas em geral, são convectivas apresentando altas variações temporais e espaciais (AUGUSTINE, 2010).

A coleta foi feita em garrafas pet de 1000 ml, lavadas com solução de limpeza e esterilizadas antes do procedimento de coleta. As garrafas tiveram suas tampas retiradas no local da coleta e nestas foi colocada água da cisterna, tampadas e armazenadas posteriormente. Esse procedimento foi realizado em cada uma das cisternas de placas que armazenam água de chuva destinada ao consumo doméstico nas comunidades rurais de Serra Grande nos Assentamentos Poço do Serrote, Assentamento Poldrinho, Assentamento Catolé e Assentamento Três Irmãos.

Uma vez que as coletas foram concluídas, as amostras de água foram devidamente armazenadas em local apropriado e em temperatura adequada, em seguida levadas para análise de qualidade no laboratório de análise química do Instituto Federal de Educação, Ciência e Tecnologia (IFPE), Campus de Vitória de Santo Antão, PE, no Agreste Pernambucano.

Foram feitas análises químicas para determinar o pH com a utilização do instrumento ph metro de bancada TECNAL modelo TEC-3MP, os Sólidos Dissolvidos Totais, seguindo o método Gravimétrico de secagem à 180° e a condutividade elétrica com um condutivímetro de bancada de marca Digimed modelo DM-31. Ainda foram realizadas análises microbiológicas onde foram detectadas coliformes fecais e totais utilizando-se a técnica de Tubos Múltiplos de acordo com o Standard Methods (APHA *et al*., 1995).

Ainda com relação às análises químicas foram determinados os cátions e os ânions responsáveis pela dureza da água. Os cátions incluindo sódio (Na^{+}) e potássio (K^{+}) e os ânions tais como sulfato (SO_4^{-2}) e cloretos naturais (Cl^{-}) existem naturalmente na água, e segundo (DEVIC *et al*., 2014), são determinantes em testes de qualidade da água.

Os métodos hierárquicos de agrupamento foram utilizados para classificar as cisternas em distintos grupos de acordo com suas similaridades ou dissimilaridades com relação às variáveis de qualidade de água ora analisadas.

Segundo Evaritt (1980), a análise de agrupamento é uma técnica estatística multivariada projetada para encontrar a estrutura dentro de um conjunto de dados. A abordagem envolve a divisão de um conjunto de dados para um número de grupos que sejam tão diferentes o quanto seja possível. É um método de classificação de objetos, uma vez que existem técnicas e critérios, examinando a minimização da distância dentro de um grupo e a maximização da distância entre grupos para encontrar um número ideal de grupos. Além dos métodos hierárquicos de agrupamento, existem também os métodos não hierárquicos como k – means (ARMSTRONG *et al*., 2012).

Existe uma variedade de métodos aglomerativos que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos. Entretanto, aqui se utilizaram: o método da média, o método do vizinho mais próximo, o método do vizinho mais distante e o método da variância de Ward, todos inseridos no conceito de agrupamento aglomerativo (ANDERBERG, 1973).

De acordo com Barroso e Artes (2003), os métodos aglomerativos são diferenciados pela regra de redefinição da matriz de dissimilaridade com presença de cada união de pares de objeto. Os métodos hierárquicos aglomerativos foram apresentados considerando G_1 e G_2, dois grupos de objetos, com g_1 e g_2 objetos, respectivamente com $d[G_1, G_2]$ sendo a distância entre eles, os métodos:

Método da ligação média (*average linkage*). Este método baseia-se na média da distância entre todos os objetos de G_1 e os de G_2, isto é.

$$[G_1, G_2] = \sum_{i \in G_1} \sum_{j \in G_2} \frac{d_{ij}}{g_1 g_2} \quad (1)$$

Os grupos são fundidos com base na distância entre todos os objetos ou cisternas dos dois grupos, em contraste com o método da ligação simples, os dois pontos da matriz de dados mais próximos não podem fundir os grupos se os outros pontos da matriz de dados são suficientemente semelhantes.

Método do vizinho mais próximo (*single linkage*). Este método baseia-se na distância mínima entre um objeto do grupo G_1 e um objeto do grupo G_2, ou seja.

$$d[G_1, G_2] = \underset{i \in G_1}{\underset{j \in G_2}{min}} (d_{ij}) \quad (2)$$

Este método usa apenas os dois pontos mais próximos da ligação simples por vez, e não consegue recuperar agrupamentos compactos. Entretanto, o método tem a capacidade de isolar valores atípicos como agrupamentos únicos no dendograma, uma vez que estes pontos, na matriz de dados, estarão longe dos seus mais próximos vizinhos. Segundo Hartigan (1984), o método de agrupamento da ligação simples é consistente em certo sentido, para grupos de alta densidade em uma dimensão.

Método do vizinho mais distante (*complete linkage*). Este método, também denominado de ligação completa, baseia-se na distância máxima entre um objeto do grupo G_1 e um objeto do grupo G_2, ou seja.

$$[G_1, G_2] = \underset{i \in G_1}{\underset{j \in G_2}{max}} (d_{ij}) \quad (3)$$

Este método é realizado da mesma maneira que o vizinho mais próximo, diferindo apenas por que a distância tomada entre grupos é a distância máxima entre os objetos de dois grupos.

Método de Ward. Neste método a alocação de um elemento a um grupo é feita maximizando a homogeneidade dentro dos grupos ou minimizando o total das somas de quadrados dentro dos grupos também conhecida como soma de quadrado dos erros (ESS), que é calculada por:

$$ESS = \sum_{i=1}^{n} (y_i - \bar{y})'(y_i - \bar{y}) \quad (4)$$

em que y_i é o vetor multivariado de medidas associado ao i-ésimo objeto e $\bar{y}$ é a média de todos os itens. De acordo com Milligan (1980), o dendograma gerado pelo método de

Ward tende a ser influenciado por valores extremos assim como produzir grupos com aproximadamente o mesmo número de observações.

O resultado do agrupamento hierárquico é, geralmente, afetado pela escolha da medida de distância entre os objetos do grupo. Esta escolha irá influenciar nos agrupamentos uma vez que alguns pontos da matriz de dados poderiam estar mais próximos um dos outros, se uma medida de distância é utilizada, e mais distante para outra medida. Existem muitas distâncias para medida entre dois pontos, mas neste trabalho será utilizada a distância euclidiana com a seguinte expressão:

$$d(ij) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \qquad (5)$$

em que x_{ij} é a *i*-ésima observação do objeto g_i e a *j*-ésima observação do objeto g_j com relação a variável *p*.

A validade da informação gerada pela função ligação é compará-la com os dados originais da distância. Se o agrupamento é válido, a ligação dos objetos no agrupamento tem uma forte correlação com as distâncias entre objetos no vetor de distâncias.

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij} - \bar{c})(d_{ij} - \bar{d})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij} - \bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(d_{ij} - \bar{d})^2}} \qquad (6)$$

em que: c_{ij} é o valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética, e d_{ij} é o valor de dissimilaridade entre indivíduos i e j, obtidos a partir da matriz de dissimilaridade.

3 Resultados e discussão

Observou-se no Q-Q plot (Figura 3) um *outlier* que se confirmou pelo *Boxplot* dos cátions, dos ânions, dos coliformes fecais e totais da (Figura 4), onde ainda foi possível verificar uma considerável assimetria nos coliformes fecais, e uma moderada assimetria nos cátions, nos ânions, e nos coliformes totais. Existe uma relação linear entre os cátions e os ânions que foi claramente observada no *Scatterplot* apresentados na (Figura 3), o que não se verificou entre outras variáveis que tiveram relação dispersa.

A utilização do Q-Q plot (Figura 3) é importante para se verificar a normalidade da matriz de dados das variáveis de qualidade da água, e observa-se a existência de normalidade uma vez que todos os pontos estão próximos a reta. Neste gráfico também foi possível verificar a existência de um *outlier* que possivelmente deve ser cisternas com qualidade de água fora dos padrões do consumo doméstico. O *Scatterplot* (Figura 3) é utilizado com o objetivo de observar a relação existente entre as variáveis estudadas. Neste caso, observa-se que existe relação apenas entre os cátions e os ânions, as demais várias não são influenciadas pelas outras variáveis em estudo, mas sim por fatores

externos. Ainda é possível observar no *Scatterplot* que as cisternas apresentam taxas coliformes fecais que variam, isto é, baixas, moderadas e elevadas.

As análises gráficas mostram a existência de *outlier* multivariado. O *Q-Q plot* e o *Scatterplot* são observados na (Figura 3). Os *boxplots* observados na (Figura 4) mostram que uma padronização das variáveis foi necessária uma vez que as variáveis aqui utilizadas: cátions, ânions, pH, sólidos dissolvidos totais (S.D.T.), coliformes fecais e coliformes totais estão aqui apresentadas em diferentes escalas.

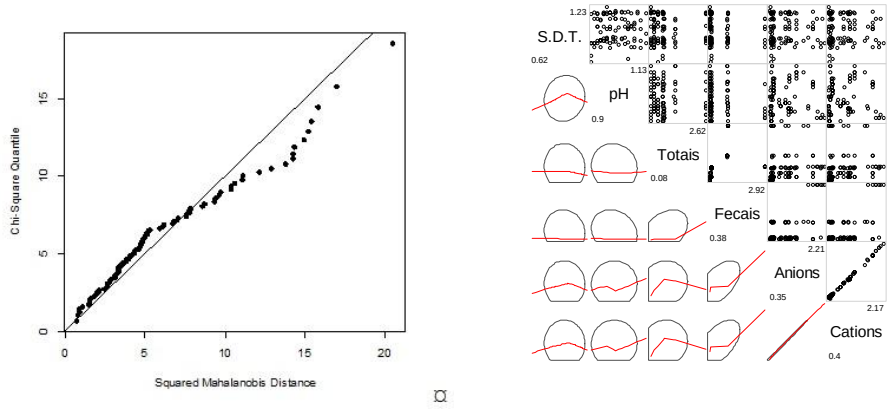

Figura 3 - *Q-Q plot* e *Scatterplot* das variáveis: cátions, ânions pH, sólidos dissolvidos totais, coliformes fecais e totais.

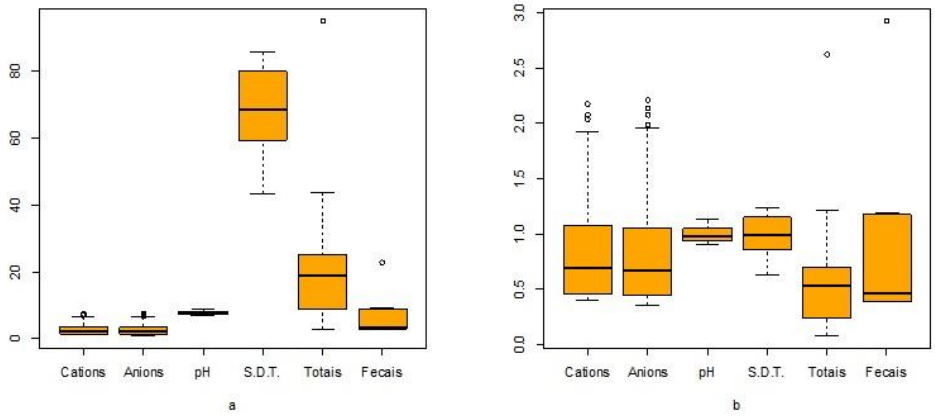

Figura 4 - *Boxplots* das variáveis com dados originais (a) e com os dados padronizados (b).

A partir da matriz de dissimilaridade das seis variáveis de qualidade de água, foram utilizados os seguintes métodos hierárquicos aglomerativos, isto é, o método do vizinho mais próximo, o método do vizinho mais distante, o método da média das distâncias, e o método de Ward. Observando-se os dendogramas dos métodos hierárquicos da (Figura 5),

considerando a presença de seis grupos em cada figura, o método do vizinho mais próximo (A), do vizinho mais distante (B) indicaram a presença de um grupo apenas com duas cisternas, o que não se verifica no método da média das distâncias (C).

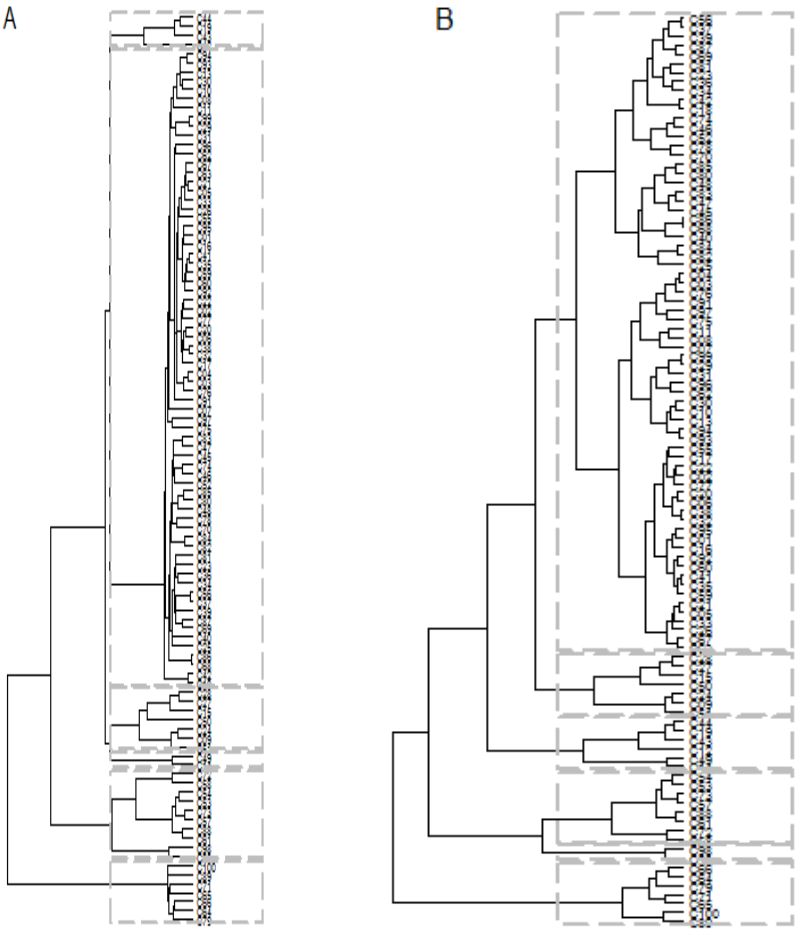

Figura 5 - Dendogramas da distância das variáveis: cátions, ânions, pH, sólidos dissolvidos totais, coliformes fecais e totais obtidos pelos métodos: (A) Vizinho mais próximo e (B) Vizinho mais distante.

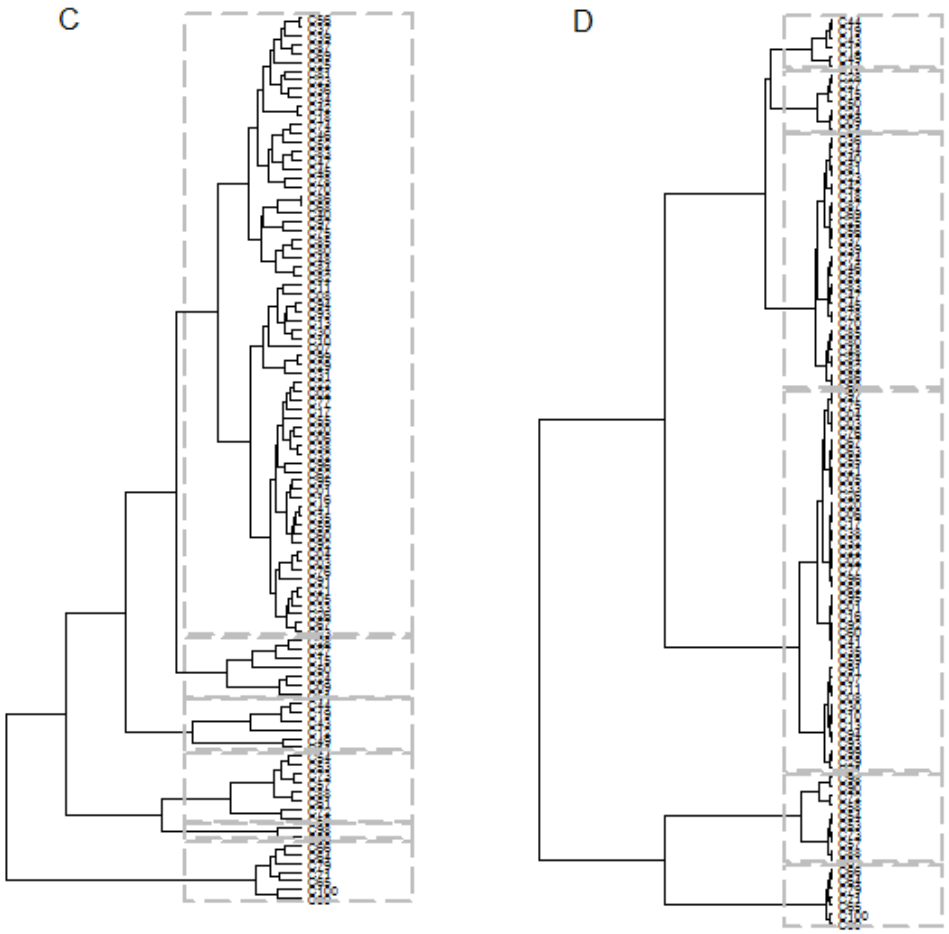

Figura 5 (continuação) - Dendogramas da distância das variáveis: cátions, ânions, pH, sólidos dissolvidos totais, coliformes fecais e totais obtidos pelos métodos: (C) Média das distâncias e (D) variância de Ward.

De acordo com os dendogramas apresentados na Figura 5, é possível visualizar que cada um dos quatro métodos de agrupamento hierárquicos utilizados selecionam as cisternas de forma bastante similar em cada grupo, ou seja, as cisternas são alocadas praticamente nos mesmos grupos por pelo menos três métodos de agrupamento, o que apresenta diferenças em sua maneira de agrupar é o método de Ward (Figura 5. D). Desta forma, as cisternas alocadas em cada um dos seis grupos ora classificados pelos métodos de agrupamento apresentam qualidade de água bastante similar.

Com base nas correlações cofenéticas dos agrupamentos hierárquicos apresentados na Tabela 1, verificou-se que o método das médias das distâncias proporcionou o melhor agrupamento das cisternas, apresentando a maior correlação cofenética 0,9562. A correlação cofenética mede a qualidade dos agrupamentos formados por cada método e é uma média que está no intervalo de zero a um [0; 1], e quanto mais próximo de 1, melhor a qualidade do agrupamento gerado por um determinado método de agrupamento.

Tabela 1 - Correlações cofenéticas dos agrupamentos hierárquicos

Método Hierárquico	Correlação Cofenética
Vizinho mais próximo	0,9381
Vizinho mais distante	0,9514
Média das distâncias	0,9562
Ward	0,8320

Na Tabela 2 que apresenta o tamanho, a média e a variância interna de cada grupo, pode-se perceber que as menores variâncias estão no método do vizinho mais distante e no método da média das distâncias.

Tabela 2 - Número de cisternas nos grupos (n), média e variâncias internas (var) das soluções obtidas pelo método do vizinho mais próximo (M1), vizinho mais distante (M2), média das distâncias (M3) e Ward (M4)

Grupos	Método											
	M1			M2			M3			M4		
	n	média	var	n	média	var	n	média	var	n	média	var
Grupo 1	70	4,12	0,25	70	4,13	0,25	70	4,13	0,25	42	3,59	0,09
Grupo 2	7	5,90	0,21	7	5,91	0,21	7	5,90	0,21	7	5,9	0,21
Grupo 3	4	6,62	0,08	6	7,06	0,38	6	7,07	0,38	6	7,07	0,38
Grupo 4	2	7,26	0,71	8	6,96	0,14	8	6,96	0,14	28	4,93	0,10
Grupo 5	10	7,35	0,48	7	9,43	0,08	7	9,43	0,08	10	7,35	0,48
Grupo 6	7	9,43	0,08	2	8,89	0,03	2	8,89	0,03	7	9,43	0,08

Observou-se ainda que a variância interna do grupo 6 foi 0,03 para as soluções obtidas pelo o método do vizinho mais distante e método da média das distâncias que teve duas cisternas para as duas soluções, denotando que as qualidades da água nestas duas cisternas são semelhantes estatisticamente.

O método de Ward teve um bom desempenho, onde apresentou uma baixa variância interna para o grupo 1 que agrupou 42 cisternas, ou seja, 42% das observações. Uma característica interessante é que as soluções apresentadas pelos métodos do vizinho mais próximo, do vizinho mais distante e da média das distâncias, tiveram as maiores

variâncias internas para o grupo 1, uma vez que este grupo foi o maior nestes três métodos de agrupamento, sugerindo que a maioria das cisternas apresenta qualidade da água similar segundo as variáveis estudadas.

A partir da matriz de distância das seis variáveis de qualidade da água obtiveram-se as medidas apresentadas na Tabela 2, isto é, o número de observações, a média e a variância interna grupos obtidos por cada um dos quatro métodos de agrupamento aqui analisados. A média interna de cada grupo é obtida a partir da matriz de distância original dos dados, ou seja, uma vez que os grupos foram obtidos por cada um dos métodos de agrupamento, esta média foi calculada com relação às seis variáveis de qualidade de água simultaneamente utilizando apenas as cisternas alocadas no grupo para o qual a média está sendo calculada. De forma similar foram obtidas as variâncias internas de cada grupo, a partir da matriz distância dos dados originais. Os termos "média" e "variância interna" são utilizados devido a estas medidas serem obtidas com as observações ou cisternas de cada grupo individualmente, e não com todas as observações.

O número de grupos que foram 6 em cada método de agrupamento, foram obtidos a partir da medida de correção cofenética, onde com seis grupos esta média apresentou os melhores resultados para cada método. Além da medida de correlação cofenética a escolha do número de grupos teve o auxílio do ponto de corte em cada dendograma, onde este ponto de corte foi realizado em pelo menos 40% da distância total, que foi a distância euclidiana.

Avaliando as médias totais dos grupos apresentadas na Tabela 2, observou-se que as soluções obtidas pela aplicação dos diferentes métodos da Análise de Agrupamento foram satisfatórias, uma vez que apresentaram médias distintas, ou seja, quanto mais distintas as médias dos grupos, melhor será a garantia de que as características de um determinado grupo diferem dos demais grupos. Denotando, desta forma, que os grupos estão bem classificados do ponto de vista de similaridade dentro dos grupos e dissimilaridade entre os grupos. Os métodos dos vizinhos mais distantes e da média das distâncias apresentaram exatamente as mesmas médias para todos os grupos e o mesmo tamanho de grupo, proporcionando dessa forma um resultado confiável no que diz respeito à alocação das cisternas que supostamente têm qualidade de água similar segundo as variáveis medidas.

Os métodos do vizinho mais próximo e Ward apresentaram as maiores dispersões entre as médias de cada grupo, sugerindo que os grupos obtidos por estes métodos sejam mais heterogêneos entre si, em relação às outras técnicas envolvidas. Assim estas duas técnicas nos levam a crer que os grupos por elas formados, foram bem selecionados, isto é, as cisternas foram bem selecionadas, não precisando ser removida para outro grupo diferente do que se encontra.

Os métodos de agrupamento do vizinho mais próximo e Ward por apresentarem grupos heterogêneos, sugerem que as cisternas pertencentes a um mesmo grupo têm qualidade da água similar. Por outro lado, apresentou qualidade de água diferente das cisternas pertencentes a outros grupos.

A solução proposta pelo método de Ward mostrou-se eficiente na montagem de grupos heterogêneos. Sugere-se utilizar os agrupamentos obtidos pelos métodos do vizinho mais distante e da média das distâncias se o interesse estiver na variância interna dos grupos, e o método de Ward, se o interesse estiver na dispersão da média interna dos grupos.

Conclusões

Neste trabalho, investigou-se a sensibilidade da classificação das cisternas de placas utilizando quatro métodos de agrupamentos hierárquicos. A correlação cofenética e o corte no dendograma foram utilizados para definir o número de grupos no estudo de qualidade da água no sertão do Pajeú.

O uso de técnicas de análise de agrupamento pode auxiliar bastante na construção de grupos de cisternas, baseando-se em informações de mais de uma característica ou variável. Na decisão pela melhor solução, recomenda-se avaliar a qualidade dos agrupamentos obtidos, comparando-se as variâncias internas dos grupos. Ainda devem-se utilizar diferentes métodos de agrupamentos em um conjunto de cisternas, visto que as soluções resultantes nem sempre apresentam grupos do mesmo tamanho.

Nas condições em que as soluções dos métodos de agrupamentos foram obtidas, observou-se que os métodos do vizinho mais distante e da média das distâncias apresentaram a menor variância interna entre os seis grupos de cisternas, proporcionando, desta maneira, grupos mais homogêneos, isto é, devem ter cisternas com qualidade de água bastante semelhante.

O método de Ward foi o que apresentou a melhor distribuição das cisternas entre os grupos, e ainda a maior variabilidade entre as médias dos grupos de cisternas. Esse resultado sugere que os grupos são heterogêneos, ou seja, as cisternas que estão em um determinado grupo têm uma baixa probabilidade de ser alocado em outro grupo, garantindo uma maior similaridade entre a qualidade da água das cisternas que estão em um mesmo grupo.

O método hierárquico de Ward foi o mais sensível em detectar a qualidade da água para todos os grupos, por apresentar as maiores variâncias internas. Enquanto que o método do vizinho mais distante e da média das distâncias parecem serem menos sensíveis a qualidade da água das cisternas uma vez que a variabilidade interna de cada grupo formado por estes métodos foi pequena.

O método da média das distâncias foi o que apresentou melhor qualidade de agrupamento das cisternas de acordo com o coeficiente de correlação cofenética, isto é, foi o maior entre os quatro métodos estudados. O método do vizinho mais distante, apesar de inferior, apresentou pequena diferença em relação ao método da média das distâncias.

As técnicas de agrupamentos hierárquicos que não foram utilizadas neste trabalho podem ser, preferencialmente, utilizadas assim como os métodos de agrupamentos não hierárquicos como o k – means e o k – medoid para a investigação de qualidade de água das cisternas de placas da região do Pajeú.

O uso de análise de agrupamento das cisternas de placas segundo as variáveis de qualidade da água em regime captação pode desempenhar um papel fundamental no estudo da poluição da água nos assentamentos Poldrinho, Poço do Serrote, Catolé e Três Irmãos na região do Pajeú Pernambucano. As técnicas de análise de agrupamento também podem contribuir na quantificação do teor de cada característica ou variável de qualidade da água em estudo, podendo detectar cisternas com qualidade de água imprópria para o consumo doméstico, as quais seriam estatisticamente chamadas de *outlier*.

Agradecimentos

Os autores agradecem à Universidade Federal Rural de Pernambuco pelo apoio e ao editor e revisores pelos comentários e sugestões.

OLIVEIRA, M. R. G; CUNHA FILHO, M.; CRUZ, D. V.; CANTALICE, J. R. B.; FALCÃO, A. P. S. T. Use of hierarchical clustering methods of plaque cisterns in the Pajeú Pernambucano. *Rev. Bras. Biom.* Lavras, v.35, n.2, p.226-241, 2017.

- *ABSTRACT: Water quality is considered as a series of physical, chemical and biological characteristics. These qualities should be studied in detail, taking into account that may contain substances harmful to human health. So, the program "one million cisterns", sponsored by the Brazilian government, has emerged as an alternative that has been established throughout the semi-arid northeastern of Brazil, for its speed of construction, the low cost of the work execution and visor easy way to capture rainwater that are pretty slim in the Brazilian semiarid. This study aimed to apply multivariate statistical methods of hierarchical clustering in order to group these cisterns according to their similarities or dissimilarities according to the measured variables. The results show that the Ward method presented the best distribution of tanks between the groups, in addition providing the highest average variability among groups. The centroid method had the lowest variance of internal groups, while the method of the average distance had the highest cophenetic correlation. Therefore, the solutions obtained by agglomerative hierarchical method suggested good grouping of tanks according to their water quality in Pajeú.*

- *KEYWORDS: Hierarchical clustering; water catchment; Brazilian semi-arid.*

Referências

ALVES, F.; KÖCHLING, T.; LUZ, J.; SANTOS, S, M.; GAVAZZA, S. Water quality and microbial diversity in cisterns from semiarid areas in Brazil. *Journal of Water and Health*, v.12, n.3, p.513–525, 2014.

ANDERBERG, M. R. *Cluster analysis for applications*. New York: Academic Press, 1973.

APHA; AWWA; WEF. *Standard Methods for the Examination of Waterand Wastewater*, twenty firsted. America Public Health Association, Washington, DC, 2005.

ARMSTRONG, J. J,; ZHU, M.; HIRDES, J. P.; STOLEE, P. K-Means Cluster Analysis of Rehabilitation Service User in the Home Health Care System of Ontario: *Examining the Heterogeneity of a Complex Geriatric Population*. v.93, n.21, p.98 – 205, 2012.

AUGUSTINE, D. J. Spatial versus temporal variation in precipitation in a semiarid ecosystem. *Landscape Ecology*, v.25, n.6, p.913–925, 2010.

BARROSO, L. P.; ARTES, R. Análise multivariada. In: SEAGRO: Simpósio de Estatística Aplicada a Experimentação Agronômica, 10., RBRAS – *Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria*, 48., 2003, Lavras. Minicurso. Lavras: UFLA 2003. 156p.

BELLIS, M. A.; JARMAN, I.; DOWNING, J.; PERKINS, C.; BEYNON, C.; HUGHES, K.; LISBOA, P. Using clustering techniques to identify localities with multiple health and social needs. *Health & Place*, v.18, p.138–143, 2012.

BUNTING, D. P.; KURC, S. A.; GLENN, E. P.; NAGLER, P. L.; SCOTT, R. L. Insights for empirically modeling evapotranspiration influenced by riparian and upland vegetation in semiarid regions. *Journal of Arid Environments*, v.111, p.42–52, 2014.

CAPDEVILA, A. S.; SCOTT, R. L.; SHUTTLEWORTH, W. J.; VALDÉS, J. B. Estimating evapotranspiration under warmer climates: Insights from a semi-arid riparian system. *Journal of Hydrology*, p.1–11, 2011.

CHERWIN, K.; KNAPP, A. Unexpected patterns of sensitivity to drought in three semi-arid grasslands. *Global Change Ecology*, v.169, p.845–852, 2012.

CLIFFORD, H.; WESSELY, F.; PENDURTHI, S.; EMES, R. D. Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in Genetics*, v.2, n.88, p.1–11, 2011.

DEVIC, G.; DJORDJEVIC, D.; SAKAN.; S. Natural and anthropogenic factors affecting the groundwater quality in Serbia. *Science of the total environment*, v.468–469, p.933–942, 2014.

EVERITT, B. *Cluster analysis*. Halstead: New York, 1980.

FERREIRA, I. A. R. *Água e Política no Sertão: Desafios ao Programa um Milhão de Cisternas*. Brasília: CDS. 2009, 141f. (Dissertação de Mestrado.) - Universidade de Brasília, 2009.

GLEICK, P. H. *Roadmap for sustainable waterre sources in South western North America*. Pacific Institute, Oakland, v.107, n.50, p.21300–21305, 2010.

GNADLINGER, J. *Técnica de diferentes tipos de cisternas, construídas em comunidades rurais do Semiárido brasileiro*. Juazeiro, BA: IRPAA, 2008.

HAGGARTY, R. A.; MILLER, C. A.; SCOTT, E. M.; WYLLIE, F.; SMITH, M.Functional clustering of water quality data in Scotland. *Environmetrics*, v.23, p.685–695, 2012.

HARTIGAN, J. A. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, v.76, p.388–394, 1981.

HASTENRATH, S. Exploring the climate problems of Brazil's Northeastern: a review. *Climatic Change*, v.112, p.243–251, 2012.

JALALI, M. Groundwater geochemistry in the Alisadr, Hamadan. *Environmental Monitoring Assess*, Western Iran, 2009.

KASSOMENOS, P.; VARDOULAKIS, S.; BORGE, R.; LUMBRERAS, J.; PAPALOUKAS, C.; KARAKITSIOS, S. Comparison of statistical clustering techniques for the classification of modelled atmospheric trajectories. *Theoretical and Applied Climatology*, v.102, p.1–12, 2010.

LIMA, I. P. A. *Aplicação do controle estatístico de qualidade de água em cisternas instaladas em comunidades na região do Sertão do Pajeú – Semiárido Pernambucano*. 2014. 68f. Dissertação (Mestrado em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco, Recife, 2014.

MARTINENT, G.; NICOLAS, M.; GAUDREAU, P.; CAMPO, M. A Cluster Analysis of Affective States Before and During Competition. *Journal of Sport & Exercise Psychology*, v.35, p.600-611, 2013.

MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms, *Psychometrika*, v.45, p.325–342, 1980.

NNANE, D. E.; EBDON, J, E.; TAYLOR, H. D. Integrated analysis of water quality parameters for cost-effective faecal pollution management in river catchments. *Science Direct*, v.45, n.6, 2011.

PASSARINO, G.; MONTESANTO, A.; RANGO, F. D.; GARASTO, S. BERARDELLI, M.; DOMMA, F.; MARI, V.; FERACO, F.; FRANCESCHI, C.; BENEDICTIS, G. D. A cluster analysis to define human aging phenotypes, *Biogerontology*, v.8 p.283–290, 2007.

SUDENE - Superintendência do Desenvolvimento do Nordeste. *Dados pluviométricos mensais do Nordeste*. Recife: SUDENE, 1990.

XU, D.; FUREY, N. R. Statistical cluster analysis of pharmaceutical solvents. *International Journal of Pharmaceutics*, v.339, p.175–188, 2007.

Recebido em 23.12.2015

Aprovado após revisão em 23.09.2016