MANOEL RIVELINO GOMES DE OLIVEIRA

MÉTODOS MULTIVARIADOS APLICADOS NO MONITORAMENTO DA QUALIDADE DA ÁGUA DE CISTERNAS DE PLACAS NA REGIÃO DO PAJEÚ – PE

RECIFE-PE AGOSTO/2016

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

MÉTODOS MULTIVARIADOS APLICADOS NO MONITORAMENTO DA QUALIDADE DA ÁGUA DE CISTERNAS DE PLACAS NA REGIÃO DO PAJEÚ – PE

Tese apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência à obtenção do título de Doutor.

Área de Concentração: Biometria e Estatística Aplicada.

Orientador: Prof.Dr. Moacyr Cunha Filho

Co-orientadora:Profª.Drª.: Ana Patrícia Siqueira Tavares de Falcão

RECIFE-PE AGOSTO/2016

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

O48 Oliveira, Manoel Rivelino Gomes de
Métodos multivariados aplicados no monitoramento da qualidade da água de cisternas de placas na região do Pajeú - PE / Manoel Rivelino Gomes de Oliveira.
– 2016.
136 f. : il.

Orientador: Moacyr Cunha Filho.
Coorientador: Ana Patrícia Siqueira Tavares de Falcão.
. Tese (Doutorado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife, BR-PE, 2016.
Inclui referências.

1. Métodos multivariados 2. Análise de agrupamento 3. Agrupamento fuzzy 4. Estatística da silhueta 5. Cisterna de placa 6. Qualidade de água I. Cunha Filho, Moacyr, orient. II. Falcão, Ana Patrícia Siqueira Tavares de, coorient. III. Título

CDD 574.018

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

MÉTODOS MULTIVARIADOS APLICADOS NO MONITORAMENTO DA QUALIDADE DA ÁGUA DE CISTERNAS DE PLACAS NA REGIÃO DO PAJEÚ – PE

MANOEL RIVELINO GOMES DE OLIVEIRA

Tese julgada adequada para obtenção do título de Doutorem Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 26/08/2016 pela Banca Examinadora.

Orientador:

Prof. Dr. Moacyr Cunha Filho
Universidade Federal Rural de Pernambuco

Banca Examinadora:

Prof(a). Dr(a). Tatijana Stosic
Universidade Federal Rural de Pernambuco - UFRPE

Prof. Dr. Victor Cassimiro Piscoya
Universidade Federal Rural de Pernambuco - UFRPE

Prof(a). Dr(a). Ana Patrícia Siqueira Tavares Falcão
Instituto Federal de Educação, Ciência e Tecnologia – IFPE

Prof. Dr. Lazaro de Souto Araújo
Universidade Federal da Paraíba - UFPB

Dedicatória

Aos meus pais Eraldo e Rosilene, aos meus irmãos Rivaldo e Ronildo, à minha irmã Raiane, DEDICO ESTA TESE COM MUITO AMOR E CARINHO.

Agradecimentos

Agradeço

- A **Deus,** em primeiro lugar,por mais essa conquista alcançada.
- Aos meus pais, **Eraldo** e **Rosilene**, pela dedicação e amor que sempre tiveram por mim.
- Aos meus irmãos,**Rivaldo** e **Ronildo,** pelo apoio e força que têm me dado ao longo dessa luta.
- À minha irmã, **Raiane**, com seu amor incondicional, sempre querendo meu bem.
- À **Ingridy**, por todo seu apoio nesse meu projeto conquistado.
- À **Bruna Ferraz,** pelo apoio e pelo carinho.
- À **Tatiana Simões** pelo apoio e colaboração com a correção gramatical.
- À minha tia **Lindomar**, a seu esposo **Lino** e aos meus primos **Geovane** e **Guilherme**, pela presença ao meu lado,pelo carinho, pelo apoio e pela força que me deram ao longo dessa batalha.
- Ao meu orientador, Prof. **Moacyr Cunha**, pela oportunidade de trabalhar num projeto tão interessante, pela orientação, pela disponibilidade, pela boas conversas, pelos ensinamentos de vida e paciência que sempre teve comigo.
- À minha co-orientadora, Prof. **Ana Patrícia**, pela cooperação na realização deste trabalho.
- À professora **Tatijana Stosic**, pelo apoio, pela ajuda e atenção. Ao professor e amigo **Lázaro,** pelo apoio que sempre me deu e pelas ajudas que foram muito pertinentes. Ao professor **Victor Piscoya,** pela amizade, pelo carisma, pelo companheirismo e apoio em todos os momentos. Agradeço também aos três professores citados, por fazerem parte da minha banca e contribuírem de forma significativa, melhorando a qualidade deste trabalho.
- Aos meus colegas **Denis, Leandro Lucena, Alvino, Neilson, Neto, Djalma, Pedro, Carlão, Luiz Henrique**, **Silvio e Thaíze**,pelo apoio, pelo companheirismo e pelos bons momentos em que compartilhamos e aprendemos juntos.
- Aos amigos **Samuel, Diego, Augusto, Milton** e **Rodrigo Silva**, pelas ajudas que foram muito pertinentes e pela amizade que firmamos ao longo desses anos.

- Ao grande amigo **David Venâncio**, pela sua amizade e pela colaboração científica em nossos trabalhos.
- Ao amigo **Neilson,**de forma especial, pelas boas conversas, pelas palavras de conforto e conselhos que sempre me deu nos momentos difíceis da nossa jornada e também pelas muitas caronas ofertadas.
- Ao amigo **Henrique,** pelas boas conversas, pela risadas, pelas caronas e pelas nossas idas à missa, pois Deus foi fundamental na realização desse projeto.
- Ao amigo **Gutenberg**, pelas boas conversas, pelas caronas e pela amizade sincera que construímos, pois são amizades como a sua que nos fazem crescer.
- Ao grande amigo **Carlos Renato,** pela sua amizade e pelo apoio que tem me dado, tanto no meu projeto de vida profissional, quanto pessoal.
- Ao Secretário **Marco Antônio dos Santos**, pela assistência.
- Aos amigos e professores da Biometria, pela hospitalidade e pelo conhecimento que me proporcionaram.
- Ao amigo e professor **Yuri,** pela sua amizade leal e pelos convites aos cursos por ele organizados em Ciência do Solo.
- Aos amigos **André Luiz, Ewerton Pereira, Fabio Sandro, Jonas, Albaro Paiva, Edyniesky e Rodrigo,** pelas boas risadas e cooperação no nosso ambiente de estudos.
- Às amigas **Leda Valéria, Evelyn Chagas, Hérica Silva, Kerolly, Nathiele** e **Edneide,** pela amizade, pelo companheirismo e cooperação no nosso ambiente de estudo.
- Ao **Josué**, pelo cafezinho.
- Ao grande amigo **Rodrigo Lins**, por todo o companheirismo, pela ajuda e força que me deu ao longo desses dois anos de Luta.
- Ao amigo **Fabrizio Barbosa,** pelo apoio e ajuda.
- Às minhas amigas "irmãs" **Armanda Maria**, **Lidia Melo**, **Daphne Gilly** e **Roberta Benuccy**, pela fraternidade, pelo companheirismo e pelos vários momentos felizes e divertidos que passamos juntos.
- Aos meus *brothers* **Rômulo Tenório** e **Nikolas Cardoso,** por todo o apoio e companheirismo em todos os momentos.

- Aos amigos e companheiros de luta, **Fabio Azevedo** e **Elisangela Rodrigues**, por todo apoio e força que sempre me deram.
- À UFRPE e ao programa de Pós-Graduação em Biometria e Estatística Aplicada, pela oportunidade que me proporcionou de adquirir e aprimorar meus conhecimentos e pela estrutura oferecida para que eu pudesse concluir meu trabalho e obter o titulo de Doutor com êxito.
- À**CAPES**, pelo apoio financeiro.

Resumo

O monitoramento da qualidade da água é de extrema importância para a humanidade, seja em regiões rurais, seja em regiões urbanas, visto que todos os seres vivos dependem desse líquido para sobrevivência, embora ele esteja ficando cada vez mais escasso. Os estudos realizados em variáveis de qualidade da água ainda se beneficiam muito pouco dos métodos estatísticos, para detectar problemas relacionados a mesma. Neste trabalho, objetivou-se desenvolver uma nova abordagem para a análise de agrupamento, a partir da combinação dos métodos hierárquicos e não-hierárquicos de agrupamento, mostrando que essas técnicas estatísticas podem contribuir de forma eficaz em estudos relacionados a variáveis de qualidade da água.Nesse estudo,utilizamos uma base dedados relacionada a variáveis de qualidade da água, coletada em cisternas de placas, em quatro assentamentos localizados no Município de Serra Talhada, na Região do Pajeú, no sertão do Estado de Pernambuco. A metodologia aqui utilizada foram os métodos estatísticos multivariados, mais especificamente a análise de agrupamento hierárquica e não hierárquica, e os índices da estatística da silhueta como critério de validação dos agrupamentos obtidos. A ideia consistiu basicamente em comparar os vários métodos de agrupamento quanto à "qualidade" dos grupos obtidos, relacionados as variáveis de qualidade da água das cisternas de placas, pelos vários métodos utilizados. Os agrupamentos obtidos para as cisternas ora estudadas mostraram que, no geral, as cisternas C14, C49, C90e C98 são similares entre si e dissimilares em relação aos demais agrupamentos, de acordo com os métodos hierárquicos de agrupamento. Com relação ao tamanho dos grupos, observou-se que geralmente o grupo 1 classificou o maior número de cisternas tanto nos métodos hierárquicos com a distância euclidiana, quanto no método não hierárquico do k – medoid, enquanto que, nos métodos hierárquicos de agrupamento com a métrica de camberra e nos agrupamentos não hierárquicos de k – média e agrupamento fuzzy, esse comportamento não se verificou, e, consequentemente, as cisternas ficaram melhor distribuídas entre todos os grupos.A Correção Cofenética indicou como melhor agrupamento o método da ligação média. No agrupamento fuzzy, apenas dois grupos ficaram com sete e oito cisternas respectivamente, já no agrupamento não hierárquico de k – média, houve uma formação de três grupos que classificaram aproximadamente o mesmo número de cisternas. Ainda se verificou que este método de agrupamento classificou um grupo de cisternas dissimilares das demais cisternas classificadas em outros grupos, pois estas cisternas são facilmente identificadas, ou seja, as cisternas C64, C65, C66, C71, C79 e C100 apresentam taxas "elevadas" de coliformes fecais,além de anormalidade em outras variáveis de qualidade da água. O método de k – medoid também formou três grupos: o grupo 1 alocou muitas cisternas; já os grupos 2 e 3 ficaram com poucas observações,pois as cisternas C75, C64 e C53 são os centróides ou as cisternas que representam cada grupo respectivamente. Portanto, conclui-se que os métodos hierárquicos e não hierárquicos de agrupamento auxiliados pelo índice da estatística da silhueta podem,de forma bastante satisfatória, monitorar a qualidade da água de cisternas de placas em uma região semiárida do Nordeste Brasileiro.

Palavras-chave: Métodos multivariados, análise de agrupamento, agrupamento fuzzy, estatística da silhueta, cisterna de placa, qualidade de água.

Abstract

The monitoring of water quality is of utmost importance for humanity, whether in rural or urban areas, since all living things depend on this liquid for survival, and it is getting increasingly scarce. The studies on water quality variables still benefits very little statistical methods to detect problems the same. This work aimed to develop a new approach to cluster analysis, from the combination of hierarchical methods and non-hierarchical clustering, showing that these statistical techniques can contribute effectively in studies related to quality variables gives water. This study will use a database related to quality variables gives water collected in cisterns plaques in four settlements located in the city of Serra Talhada Pajeu in the region in the interior of Pernambuco. The methodology used here were the multivariate statistical methods specifically the hierarchical cluster analysis and non-hierarchical and the silhouette of statistical indices as validation criteria of the obtained clusters. The idea is basically to compare the various clustering methods as the "quality" of the groups obtained, relating the quality of water variables of the plaques of cisterns, the various methods used. The groups obtained for cisterns now studied showed that in general the cisterns C14, C49 and C90, C98 are similar among themselves, and dissimilar to other groups according to hierarchical methods of grouping. Regarding the size of the groups, which is generally observed Group 1 scored the highest number of cisterns both in the hierarchical method with the Euclidean distance as the non-hierarchical method k - medoid, while the grouping of hierarchical methods with the metric of canberra and non hierarchical clustering k - mean and fuzzy grouping this behavior was not observed, and consequently the cisterns were better distributed among all groups. The Correlation cophenetic has indicated as better grouping of the average bond method. In fuzzy clustering only two groups were left with seven end eight cisterns respectively, as in the non-hierarchical clustering k - mean hear the formation of three groups which roughly classified the same number of cisterns, and yet it has been found that this grouping method rated a group of dissimilar cisterns of other cisterns classified in other groups where these tanks are easily identified, the is the cisterns C64, C65, C66, C71, C79 and C100 have rates fecal coliforms beyond abnormality in other quality variables gives water. The k method - medoid also formed three groups, wherein the group 1 allocated many cisterns while groups 2 and 3 were with few observations where cisterns C75, C64 and C53 are the centroids or cisterns which represent each group respectively. Therefore, it is concluded that the hierarchical methods and non-hierarchical clustering aided by the silhouette of statistical index can, quite satisfactorily monitor the water quality boards tanks in a semiarid region of Nordeste Brasileiro.

Keywords: Multivariate methods, cluster analysis, fuzzy clustering, statistical silhouette, plaques cistern, water quality.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 – INTRODUÇÃO

A evolução da humanidade sempre teve uma relação direta,uma dependência inquestionável dos recursos hídricos. A dependência é tão significativa que a qualidade da água usada por determinada sociedade pode ser determinante para sua qualidade de vida. Assim, na sociedade de um modo geral, a qualidade da água tem papel extremamente relevante nas regiões áridas e semiáridas, onde os recursos hídricos são escassos, e dessa forma, os reservatórios de água devem ter um monitoramento contínuo e observar atentamente algumas variáveis de qualidade da água.

O monitoramento contínuo das variáveis de qualidade da água pode colaborar de maneira significativa na gestão dos recursos hídricos, tanto na economia de água, quanto no seu uso eficiente sem desperdícios. Além disso, o acompanhamento dos dados relacionados à qualidade da água permite que a água de tais reservatórios seja utilizável sem danos à saúde humana, como preconiza o Conselho Nacional do Meio Ambiente (CONAMA, 1986).

Observa-se que os estudos relacionados ao monitoramento da qualidade da água, muitas vezes, subutilizam a estatística,relegando-a a um papel secundário ou sequer a utilizam. Assim, nesta tese, um dos objetivos é verificar este aspecto, mostrando que a estatística pode ser melhor usada e contribuir de maneira significativa no processo de análise e estudo de variáveis de qualidade da água.

A partir do levantamento bibliográfico desenvolvido neste trabalho, situamos que, em 2014, surgiu uma das primeiras pesquisas utilizando métodos estatísticos para variáveis de qualidade da água, nos assentamentos localizados na região do Pajeú em Serra Talhada–PE. Trata-se de uma dissertação feita no Programa de Pós-Graduação em Biometria e Estatística Aplicada. O referido trabalho é de Lima (2014), que ressalta a importância de se estudar e analisar estatisticamente variáveis de qualidade de água em cisternas de placas, visando a observá-las quantitativa e qualitativamente, assim como uma localização das mesmas do ponto de vista geográfico na região do Pajeú–PE. Portanto, a proposta desta tese é utilizar uma combinação de métodos estatísticos multivariados existentes na literatura que não foram anteriormente explorados em variáveis de qualidade da água.

O termo análise multivariada, corresponde a várias técnicas e métodos que utilizam simultaneamente as informações existentes de todas as variáveis respostas

na interpretação de uma base de dados, considerando as correlações existentes entre elas (FERREIRA, 2008). Nesse contexto, a intensificação do uso dos métodos multivariados pode melhorar a qualidade das pesquisas, facilitando a interpretação da estrutura dos dados, resultando em menos perda de informação. Assim, destacamos que os métodos de análise multivariada têm sido regularmente aplicados em vários estudos científicos nas diversas áreas da pesquisa.

A região semiárida, como já é de conhecimento de todos, incluindo órgãos de pesquisas e poder governamental, é escassa de recursos hídricos (OLIVEIRA *et al.*, 2015). Nos últimos anos, com a redução dos índices pluviométricos, a falta de água tem se tornado um problema de utilidade pública, em virtude dos grandes reservatórios estarem com indisponibilidade hídrica para o consumo humano. Assim, diante dessa realidade, torna-se ainda mais pertinente a realização de estudos relacionados a variáveis de qualidade da água na região do Pajeú–PE.

A estiagem é um fenômeno natural conhecido no mundo inteiro pela ausência de precipitação. Modelos climáticos globais prevê em 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 por causa direta da falta de água em regiões assoladas pela seca ou estiagem (OLIVEIRA *et al.*, 2015).

As regiões mais afetadas pelo fenômeno da estiagem são aquelas com clima semiárido. Nas zonas áridas, a razão entre a pluviosidade e a evapotranspiração fica entre 0,20 e 0,50. A precipitação média anual nas regiões semiáridas varia de 300 a 800 mm. Assim, a evapotranspiração é um 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 estabelece regimes intermitentes aos rios e vegetação sem folhas em boa parte de ano, isto porque o semiárido do nordeste Brasileiro é propenso a secas, experimentando uma curta estação chuvosa em torno de janeiro a abril (HASTENRATH, 2012).

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 consu-

mo doméstico. Apesar de mais de dois milhões de pessoas que vivem em regiões semiáridas do Brasil consumirem água da chuva armazenada em cisternas, pouca informação está disponível sobre a qualidade da água (AVES *et al.*, 2014).

Essa 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 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 para a saúde humana, especialmente em países economicamente menos desenvolvidos (NNANE *et al.*, 2011). Os parâmetros de qualidade da água abordados para estudo neste trabalho foram os Cátions, os Anions, o Potencial hidrogeniônico (pH) e os 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 (CONAMA, 1986).

A Análise de Agrupamento Hierárquico e Agrupamento não Hierárquico, como também seus respectivos índices de validação, foi utilizada nesse trabalho, para uma avaliação sobre a qualidade da água nos assentamentos localizados na região do Pajeú, no município de Serra Talhada–PE. Em decorrência desse processo, pudemos empreender uma análise sobre problemas relacionados às variáveis de qualidade da água.

O estudo começou com a obtenção dos agrupamentos das cisternas de placas localizadas nos assentamentos Catolé, Poldrinho, Poço do Serrote e Três Irmãos, de acordo com as suas similaridades obtidas pela Análise de Agrupamento. Após a construção dos dendogramas onde se visualizam os grupos, fez-se necessário encontrar um ponto de corte neste gráfico, ou seja, obter o número de grupos ideal que gerariam o melhor agrupamento.

Na literatura, existem vários critérios para formar agrupamentos. Porém, nenhum deles gera uma formação ideal de grupos, de acordo com Mingoti (2005), ou seja, os referidos métodos somente auxiliam na tomada de decisão final quanto ao número de grupos. Logo, neste trabalho, a proposta de escolha do número de grupos é através do índice da estatística da silhueta. Este índice é calculado em cada um dos métodos hierárquicos de agrupamento, com base na matriz de distância euclidiana ou métrica de camberra.

A utilização do índice da estatística da silhueta na análise de agrupamento teve como objetivo encontrar o número ideal ou “ótimo” de grupos, isto é, a ideia aqui ora exposta é obter além do número de agrupamento ideal, outros agrupamentos pertinentes ou candidatos a grupos que classifiquem bem os seus objetos ou cisternas, considerando o índice da estatística da silhueta. Essa iniciativa possibilita ao pesquisador a escolha de acordo com a natureza de cada pesquisa.

Dando continuidade à pesquisa, analisaram-se os métodos não hierárquicos de agrupamentos aplicados aos dados de cisternas de placas dos assentamentos já mencionados, localizados na região do Pajeú no município de Serra Talhada–PE. Um dos métodos não hierárquicos de agrupamento aqui proposto é o Agrupamento Fuzzy, cuja classificação consiste em atribuir cada cisterna ao grupo cujo grau de pertinência é maior que para os demais grupos, ou seja, utilizou-se uma matriz de grau de pertinência para classificar todas as cisternas em um determinado grupo.

Os agrupamentos visam a obter grupos de cisternas que sejam homogêneos dentro de cada grupo e heterogêneos entre os grupos, de acordo com cada uma das características relacionadas às variáveis de qualidade de água. Nesta análise, assim como nos métodos de agrupamento hierárquicos, utilizaram-se os índices da estatística da silhueta para encontrar o número de grupos “ótimo”,além da utilização do coeficiente de partição (V_{PC}), da partição entropia (V_{PE}), do índice de Xie e Beni (V_{XB}) e do parâmetro de fuzificação (m) para validar o número de agrupamentos obtidos pelo método não hierárquico de Agrupamento Fuzzy.

Levando-se em consideração o contexto do estudo apresentado neste trabalho, observa-se que as várias cisternas aqui consideradas podem ser usadas para avaliações das variáveis de qualidade de água das cisternas de placas localizadas nos referidos assentamentos. No entanto, através dos agrupamentos obtidos, torna-se pertinente usar alguns deles para estudos quanto às características relacionadas à qualidade da água. Em outras palavras, poderiam ser tomadas medidas cabíveis com relação à qualidade da água de cisternas alocadas em determinado grupo, e, assim, ações mais efetivas poderiam ser tomadas com mais segurança.

Considerando a classificação das cisternas nos seus respectivos grupos, tanto pelos métodos hierárquicos, quanto pelos métodos não hierárquicos, medidas podem ser tomadas pelas autoridades ou órgãos responsáveis no sentido de fiscalizar e melhorar a qualidade da água das cisternas, que estão fora dos padrões para o consumo doméstico. Isso pode ser possível, uma vez que, observando-se os grupos

obtidos, pode-se ver junto à secretaria de saúde do estado ou município ou mesmo junto a outro órgão competente um tratamento de água com diferentes intensidades, de acordo com cada grupo.

Com relação às características inerentes a cada grupo, obtidas pelos métodos hierárquicos e não hierárquicos de agrupamento, espera-se obter informações mais precisas ou refinadas não obtidas ou exploradas por outros métodos, que sejam capazes de fornecer informações suficientes para lidar com questões relacionadas à qualidade da água, reforçando mais ainda a eficiência desses métodos estatísticos aqui abordados, na avaliação de variáveis de qualidade da água.

Portanto, neste trabalho, objetivou-se mostrar que a estatística multivariada pode colaborar de forma eficiente na "condução" ou solução de problemas relacionados à qualidade da água. Através dele, podemos perceber como os métodos estatísticos colaboraram com a análise da qualidade da água,seja em cisternas de uma comunidade rural em uma região semiárida do Nordeste Brasileiro, como a área de estudos desta tese,seja em cisternas de localidades em outras regiões do país,seja em outros reservatórios em região rural ou urbana.

No que se refere ao estudo ou monitoramento de variáveis relacionadas à qualidade de água, a metodologia estatística aqui proposta nesta tese tem os seguintes objetivos:

- Formar grupos, considerando as variáveis de qualidade de água das cisternas de placas localizadas nos assentamentos da região do Pajeú–PE.
- Considerar os agrupamentos obtidos, identificando quais são os significativos, e qual o mais relevante para cada análise, segundo o índice da estatística da silhueta.
- Identificar as cisternas consideradas "diferentes", isto é, com qualidade da água comprometida para o consumo doméstico, considerando as variáveis de qualidade da água.
- Realizar uma interpretação prática das informações que podem ser retiradas de cada grupo.
- Identificar os grupos que podem ser formados, a partir do Agrupamento Fuzzy, e quais os grupos mais relevantes,de acordo com os índices de validação utilizados.
- Obter as informações relevantes nos grupos formados pelos métodos não hierárquicos de k–médias e k–medoid.

De maneira geral, espera-se que o conjunto de métodos estatísticos multivariados propostos nesta tese proporcione um melhor entendimento com relação à qualidade da água das cisternas de placas localizadas nos assentamentos da região do Pajeú–PE. Dessa forma, vamos poder empreender um monitoramento, de forma mais eficiente, colaborando, assim, com uma fiscalização mais efetiva por parte de órgãos competentes. Acreditamos que iniciativas de pesquisas como esta apresentam não só relevância acadêmica, mas social, uma vez que contribuem, através de suas análises, para evitar doenças que possam ser contraídas por consumo de água imprópria para o consumo humano.

2 - REVISÃO DE LITERATURA

2.1 Estatística Multivariada

O estudo investigativo surge da necessidade de novas descobertas, que são obtidas de pesquisas em várias áreas, de variáveis que são mensuradas em geral de maneira conjunta.Uma das ferramentas utilizadas em análise estatística de variáveis conjuntas é a estatística multivariada (ALBUQUERQUE, 2013). Os métodos multivariados são um conjunto de técnicas que permitem ao investigador interpretar grandes conjuntos de dados, que podem ser referentes a indivíduos ou variáveis. Esses métodos buscam encontrar relações entre variáveis, entre indivíduos ou entre ambos (DOCAMPO *et al.*, 2013).

De acordo com Reis (2001), um dos objetivos da Estatística Multivariada é simplificar os dados, descrevendo a informação através de um reduzido número de dimensões de análise.Estes objetivos incluem métodos que analisam a relação de dependência e interdependência entre conjuntos de variáveis ou indivíduos, quer seja descritivo, quer seja inferencial. Um aspecto essencial, contudo, da análise estatística multivariada é a dependência entre as diferentes variáveis. Assim, a mensuração e análise de dependência entre variáveis, entre conjuntos de variáveis, e entre variáveis e conjunto de variáveis constitui-se em processo fundamental para a análise multivariada (ANDERSON, 2003).

As patologias de um modo geral podem ser observadas e classificadas de acordo com suas características e sintomas nas pessoas que as portam. As técnicas multivariadas têm se consagrado como uma ferramenta estatística sofisticada para classificação de patologias e seleção de pessoas com doenças em comum (FENG *et al.*, 2013). Contudo, a análise de *Clusters* tem sido uma das mais evidentes ferramentas, dentre as técnicas estatísticas multivariadas na seleção de variáveis patológicas,assim como em grupos de pacientes acometidos por alguma patologia (MCGUIRE *et al.*, 2013).

A análise de *Clusters* torna-se confiável quando realizada em sucessivas etapas. O método da ligação completa e a distância euclidiana quadrática foram utilizados no agrupamento hierárquico por Hair *et al.* (2010), para determinar o número de *clusters* em uma base de dados de perfis efetivos de atletas, antes e durante a competição, principalmente para examinar as diferenças entre esses perfis na competi-

ção e os objetivos dos atletas que foram realizados com relação aos esportes que praticam.

O método de K – médias foi aplicado por Armstrong et al. (2012) com o objetivo de examinar a heterogeneidade e de identificar padrões previamente desconhecidos de características clínicas de pacientes idosos, que utilizam serviços de reabilitação e são atendidos em domicílio, em Ontário, no Canadá. Martinent *et al.* (2013), utilizaram o método hierárquico de Ward e o método não hierárquico de k – médias para agrupar e analisar grupos de atletas antes e durante a competição. Siou *et al*. (2011) avaliaram padrões alimentares e realizaram estudos comparativos de variáveis relacionadas a estes padrões alimentares entre grupos, e dentro dos grupos formados por homens e mulheres adultos, com a utilização dos métodos hierárquicos da variância mínima de Ward e o método Flexível – Beta, e o método não hierárquico de k – médias, chegaram a resultados conclusivos.

Assim, vemos que a análise de *clusters* tem uma ampla aplicabilidade nas ciências médicas, no que se refere a detectar grupos ou subgrupos de pacientes com características homogêneas com respeito a determinadas variáveis de saúde(SHAW *et al*., 2008). Nas últimas décadas, tem crescido de forma significativa a prevalência de doenças cardíacas, principalmente em pacientes idosos, e a análise de *clusters* tem sido utilizada para auxiliar na qualidade de vida desses pacientes (FUKUOKA *et al*., 2007). A análise de clusters foi aplicada na variável pesopré-gestacional, no ganho de peso durante a gravidez e na redução de peso no período pós-parto, com o objetivo de desenvolver campanhas de saúde mais eficazes, assim como de auxiliar procedimentos de intervenção em mulheres gestantes (MACKERT e WALKER, 2011).

2.2 Análise de Agrupamento

A classificação é uma atividade conceitual básica dos seres humanos. As crianças aprendem desde muito cedo a classificar os objetos pertencentes ao seu ambiente, e a associar os resultados dessa classificação ao seu dia a dia (REIS, 2001).A análise de agrupamento foi utilizada em um estudo observacional para detectar grupos de risco em alemães com 50 anos ou mais, portadores de algumas doenças causadas pelo uso de tabaco, pelo consumo excessivo de álcool, pelo sedentarismo e pelos hábitos alimentares inadequados (SCHNEIDER *et al*,. 2009).

A análise de Agrupamento ou Análise de Clusters, que também é denominada classificação não supervisionada, é a classificação de objetos em diferentes grupos, sendo que cada um deles deve conter os objetos semelhantes, segundo alguma função de distância estatística.Ressaltamos que essa metodologia foi criada há mais de sete décadas(JOHNSON e WICHERN, 1998).

A Estatística Multivariada é uma das áreas mais importantes e utilizadas no universo das análises estatísticas pelo seu poder e resultados sofisticados no estudo de grandes massas de dados que contêm muitas variáveis. Esta rica área da estatística é contemplada por várias subáreas, dentre elas, a Análise de Agrupamento ou *Clusters*. A análise de agrupamento é uma técnica exploratória usada como um método classificador e que pode ter outras denominações ou definições de acordo com a literatura. Dentre as denominações existentes, as mais citadas são a Taxonomia Numérica, o reconhecimento de padrões, a Análise de grupos, o Agrupamento ou Aglomerados e a Análise de *Clusters*.A vasta nomenclatura da Análise de Agrupamento deve-se certamente à sua importância e à intensiva aplicação em diversas áreas de estudo, inclusive em estudos de saúde. Mesmo com a grande variedade da nomenclatura existente, a Análise de Agrupamento tem ganhado posição de destaque em trabalhos científicos que fazem uso dessa metodologia, e,por isso, nós a escolhemos como aporte metodológico deste trabalho de tese.

A análise estatística multivariada tem sido aplicada com êxito em inúmeros estudos. Uma das técnicas multivariadas, a Análise de Agrupamento ou *Clusters,* é um método de análise estatística multivariada não – supervisionada que tem como atributo a identificação de estruturas hierárquicas em um grande número de observações menores e semelhantes. Assim, os objetos dentro de um mesmo grupo são muito semelhantes, e objetos em grupos diferentes são significativamente diferentes em suas características (IRAWAN *et al*., 2009). Segundo Thyne *et al*. (2004), os métodos de estatística multivariada têm sido empregados para extrair informações críticas de conjuntos de dados em sistemas complexos.

Muitas técnicas de Agrupamento são baseadas em encontrar a partição que otimiza uma função objetiva chamada de particionamento. Uma partição é um conjunto de subconjuntos mutuamente exclusivos da população. Uma vez que é impraticável pesquisar todas as partições possíveis, métodos de agrupamento, são usadas várias estratégias para obter uma solução ideal, ou um ótimo local (MICHAUD, 1997).

De acordo com Michaud (1997), uma população de n elementos descrita por m atributos,para ter uma descrição intuitiva desejável, deve apresentar pequenas distâncias no espaço m – dimensional de atributos entre elementos do mesmo grupo e grandes distâncias entre elementos de grupos distintos. Algoritmos de agrupamento são usados extensivamente em compreensão de dados e descoberta de conhecimento em áreas como astronomia, biologia, psicologia e ciências da saúde em geral, ou seja, o agrupamento pode ser um problema tão universal, que podem existir inúmeras variantes do mesmo (RAJU *et al*., 2011).

Segundo Passarino *et al*. (2007), o termo Análise de Agrupamento abrange diferentes algoritmos e métodos para agrupar objetos de forma que o grau de associação entre dois objetos é máximo se eles pertencem ao mesmo grupo e mínimo se pertencem a grupos diferentes. A Análise de Agrupamento refere-se a um conjunto diversificado de técnicas estatísticas que agregam casos em grupos onde os casos de um mesmo grupo são semelhantes e dissimila para casos em outros grupos (COID *et al*., 2012).

Para Picard *et al*. (2010), os métodos clássicos para agrupamentos de um conjunto de observações que podem ser pessoas, espécies ou objetos, como análise de agrupamento hierárquica,não leva em consideração a variabilidade das características das “observações” utilizados para formar o *Cluster*. Assim, é necessário que uma medida de similaridade ou dissimilaridade seja estabelecida para associar duas “observações”devido às suas semelhanças, de forma que os *Clusters* serão o mais homogêneo possível.

Existem diversos algoritmos para formar agrupamento. De fato, algoritmos diferentes podem produzir distintos agrupamentos. Além disso, a análise de agrupamento pode detectar grupos que não existem na realidade. Uma forma de avaliar os métodos de agrupamento é mediante os métodos gráficos.

O tipo de gráfico a ser utilizado está relacionado ao número de variáveis de interesse, ou seja, quando se tem apenas duas variáveis de interesse ($p = 2$), um diagrama de dispersão entre elas permitirá uma visualização de possíveis agrupamentos entre os indivíduos. Quando o número de variáveis de interesse é maior do que dois ($p > 2$), pode-se implementar uma análise de componentes principais. Assim, se a proporção de variáveis explicadas pelas primeiras componentes for significativa, isto é, aproximadamente 80% ou mais, procede-se a um diagrama de disper-

são das primeiras componentes principais para visualizar a existência de possíveis grupos.

De forma genérica, a análise de *clusters* compreende cinco etapas:

1. a seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. a definição de conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. a definição de uma medida de semelhança ou distância entre cada dois indivíduos;
4. a escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição de um algoritmo de classificação;
5. e, por fim, a validação dos resultados obtidos.

A seleção das variáveis deve proceder levando-se em consideração o conhecimento prévio sobre o assunto a ser estudado, ou seja, o objetivo do estudo permitirá escolher entre os dados disponíveis, as variáveis mais significativas na abordagem do problema, e outra consideração não menos importante é de ordem estatística, que está relacionada aos tipos de variáveis utilizadas no estudo. De acordo com Vialle *et al*. (2011), a análise de agrupamento é utilizada para procurar agrupamentos naturais entre os objetos e descobrir as estruturas latentes presentes nos dados.

2.3 Distribuição Normal Multivariada

A estatística multivariada é uma generalização da univariada e tem uma importância particular, a generalização da conhecida Distribuição Normal. A sua devida importância deve-se ao fato de muitos métodos estatísticos multivariados se basearem no pressuposto de que os dados são extraídos de uma população com distribuição normal multivariada. Sabe-se que, mesmo que os dados não sigam uma distribuição exatamente normal, em geral, é possível aproximar a distribuição real à normal.

Para Reis (2001), a importância do estudo da distribuição normal justifica-se por três razões: muitos fenômenos reais, tais como físicos, biológicos e econômicos, seguem uma distribuição normal; muitas distribuições amostrais são normais ou, devido ao teorema central do limite, podem ser consideradas aproximadamente normais; e por fim, a relativa facilidade matemática dessa distribuição.

A distribuição normal multivariada é uma generalização da univariada para *p* – variáveis ($p \geq 2$) dimensões. De acordo com Ohlson *et al*. (2011), uma matriz Xé normalmente distribuída com notação dada por $X \sim N_{p,n}(\mu, \Sigma, I_n)$, em que a matriz de parâmetro Σ representa a covariância entre as colunas de X, e I_n, a matriz identidade de dimensão *n*, indicando que as colunas são independentes identicamente distribuídas. Sua função densidade de probabilidade é dada por

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \qquad (1)$$

em que$\boldsymbol{\mu}' = (\mu_1, \mu_2, \ldots, \mu_p)$é o vetor de médias,e a matriz de covariâncias é dada por:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{pmatrix} \qquad (2)$$

A matriz X é particionada da seguinte maneira

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix} = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \end{pmatrix} \qquad (3)$$

e ainda o produto $(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ é conhecido como a Distância Generalizada de Mahalanobis.

A suposição de multinormalidade pode ser verificada com o auxílio das análises das distribuições marginais univariadas e bivariadas. Vale ressaltar que a normalidade de todas as distribuições univariadas e bivariadas não representa nenhuma garantia de normalidade multivariada da matriz X (SCHLAMM e MESSINGER, 2011). Note que, se a matriz X segue uma distribuição normal, então todas as distribuições univariadas e bivariadas certamente seguem distribuição normal. No entanto, quando as distribuições univariadas e bivariadas seguem distribuição normal, a probabilidade da matriz X ser normal é alta.

A partir do exposto, percebemos que os métodos estatísticos são auxiliares importantes para detectar tendências, explorar relações e tirar conclusões a partir de dados experimentais. No entanto, não é difícil verificar que muitos pesquisadores aplicam testes estatísticos sem antes verificarem se eles são adequados para a aplicação pretendida. Assim, existem técnicas estatísticas paramétricas e não paramétricas,univariadas e bivariadas, para verificar a normalidade e homogeneidade de variâncias antes da comparação de dois ou mais conjuntos de amostras em testes

inferenciais, em correlação e análise de regressão (GRANATO *et al*., 2014). As suposições de normalidade podem ser verificadas por:

i) **Normalidade Univariada:** Esse tipo de normalidade pode ser verificada através de gráficos de probabilidade normal, tais como o *normal plot* ou *Q-Qplot*, por meio de histogramas ou *boxplot's* e ainda por meio de testes estatísticos baseados na assimetria empírica ou curtose como a estatística de Jarque – Bera, Cramér – von Mises, Anderson–Darling e Kolmogorov – Smirnov (QUESSY e MAILHOT, 2011). Uma maneira alternativa é a transformação dos dados originais quando eles não são retirados de uma população com distribuição normal univariada.

ii) **Normalidade Bivariada:** A normalidade bivariada é verificada por gráficos de dispersão entre as variáveis X_i e X_j, para $i \neq j$ com i, j = 1, 2, ..., p,isto é, uma conhecida verificação gráfica de normalidade bivariada baseia-se na concentração elíptica de pontos no gráfico de dispersão. No entanto, em muitas situações, esse tipo de gráfico não é suficiente e um teste formal de significância vai dar uma visão mais objetiva quanto à adequação da distribuição normal bivariada (BEST e RAYNER, 1988).

iii) **Normalidade Multivariada:**Assim como no caso univariado, no caso multivariado, a normalidade é verificada pelo gráfico de probabilidade *Q-Q plot*. Assim, quando os dados amostrais seguem uma distribuição normal multivariada, é possível visualizar uma nuvem de pontos próxima à reta ajustada. No entanto, se os pontos ficarem dispersos, isto é, sem nenhuma tendência visível, existem fortes indícios de ausência de normalidade.

Segundo Atkinson e Riani (2004), uma alternativa usada para detecção de normalidade multivariada é a transformação dos dados para auxiliar na suposição das hipóteses de normalidade. E ainda se verificou que a extensão de Box e Cox (1964), à família de respostas multivariada em geral, leva a um aumento significativo da normalidade, e, consequentemente, a uma análise simplificada dos dados amostrais.

2.4 Outliers Multivariado

Ao se fazer um estudo observatório dos dados, em muitos casos, é possível visualizar características que não necessitam de análises mais sofisticadas para que sejam visualizadas. Tais características, como pontos discrepantes ou *outliers* ou mesmos informações inexistentes, podem trazer informações preciosas em investigações mais detalhadas de um conjunto de dados.

Em contraste com os *outliers* univariados, cuja identificação de observações extremas é simples, para *outliers* multivariados, a estrutura de covariância do conjunto de dados deve ser considerada.Então, é importante que antes do uso de qualquer método multivariado se faça uma investigação da existência desses valores extremos ou *outliers* que podem afetar ou mesmo distorcer os resultados finais obtidos por esses métodos.A visualização da nuvem de pontos elípticos resultantes de uma distribuição normal multivariada na geometria euclidiana pode ser importante para comparar com métodos de detecção de outliers multivariados (FILZMOSER *et al*., 2012).

Os valores extremos ou *outliers* multivariados não são necessariamente os extremos de uma única variável, mas podem ser em toda gama, isto é, no espaço p-dimensional definido por todas as variáveis em estudo (PREARO et al., 2012). Assim, em muitas situações, requer uma atenção especial com esses extremos multivariados, uma vez que podem ser considerados como extremos apenas do ponto de vista multivariado e não univariado.

As observações discrepantes podem ser caracterizadas em termos univariados, bivariados ou multivariados. Estas observações são vistas como univariadas, quando a análise considera uma única variável e podem ser identificadas por um simples *boxplot* (ABUZAID *et al*., 2012). As observações discrepantes bivariadas podem ser detectadas com o auxílio de gráficos de dispersão bidimensionais e das elipses de confiança (JOHNSON e WICHERN, 2007). No âmbito multivariado, as observações discrepantes são identificadas pela estrutura de covariância e por alguma medida de distância de uma observação especial ao ponto central de todas as observações, em que serão consideradas discrepantes as observações que estão muito distantes do ponto central (Todorov *et al*., 2011), além da utilização dos gráficos de dispersão tridimensionais e dos gráficos Q-Q *plots* na identificação de *outliers*.

De acordo com Southworth (2008), um método robusto para identificar observações discrepantes em dados multivariados é a distância de Mahalanobis$D_{i,}$ aliada a outros métodos estatísticos também considerados robustos. Assim, para detectar valores discrepantes, é sugerida uma estatística de teste com distribuição F $[p(n-1)/(n-p)]F_{(p,n-p,\alpha)}$, com p, $n-p$ graus de liberdade a um nível de significância α, considerando que valores de D_i maiores do que o valor crítico obtido por esta estatística de teste são valores discrepantes.Na prática a distribuição χ^2 é uma aproximação pobre para a distribuição de distâncias robustas, sendo superada pela distribuição F para distâncias estimadas, usando a estimativa determinada pela mínima covariância da matriz de covariância (HARDIN e ROCKE, 2005).

Os pontos discrepantes ou *outliers* devem ser identificados cuidadosamente, uma vez que pontos peculiares ou similares ao restante da população devem ser selecionados para que sejam analisados estatisticamente ou inferencialmente, e pontos para dissimilar a população podem ser descartados, se for de interesse do pesquisador.

Existem diversas técnicas estatísticas multivariadas utilizadas para detectar observações discrepantes, ou *outliers* em massas de dados. Segundo Baxter (1999), métodos de análises de agrupamento são as técnicas mais utilizadas para identificar observações discrepantes. Ainda de acordo com o mesmo autor, a Análise de Componentes Principais também é útil para essa finalidade e, principalmente, para mostrar como os *outliers* se relacionam com outros casos em uma amostra.

1. Análise de Agrupamento: A análise de agrupamento é um método multivariado mais comumente utilizado em análises de dados e é quase sempre utilizado como um método de identificação de grupos, mais também parece ser um dos principais métodos utilizados para identificar *outliers* multivariados, visto que nenhum grupo foi considerado similar para ser colocado esses *outliers*.
2. Análise de Componentes Principais: Os casos que são considerados discrepantes em relação a uma ou mais variáveis podem ter uma forte influência sobre os primeiros componentes principais e serem evidentes em parcelas com base nessas influências. Outra maneira de colocar esse ponto de vista é que a Análise de Componentes Principais é muito útil para i-

dentificar valores atípicos ou *outliers* que são identificados por métodos mais simples e evidentes (BAXTER, 1999).

3. Uma observação importante é que agrupamentos oriundos de grandes amostras podem apresentar grupos atípicos, mas que não necessariamente podem ser considerados observações discrepantes ou *outliers.* Vale ressaltar que existem vários algoritmos para formar grupos, e algoritmos diferentes podem produzir distintos grupos. Além disso, podem detectar grupos que não existam na realidade, sugerindo, dessa forma, várias restrições para a inclusão ou remoção de algum grupo na análise.

2.5 Técnicas de Agrupamento

Muitas técnicas de agrupamento são baseadas em encontrar uma solução ótima (MICHAUD, 1997).As técnicas de agrupamento podem avaliar de forma eficiente grandes massas de dados, para que os objetos que tenham características semelhantes sejam agrupados e diferenciados de grupos que apresentem características diferentes (BERGE *et al*., 2003). As diferentes técnicas de agrupamentos podem levar a diferentes soluções e estas técnicas de agrupamentos podem ser hierárquicas ou não hierárquicas (FERREIRA, 2008).

A técnica de agrupamento mais adequada é escolhida mediante o problema estudado e o tipo de dado do referido problema. Muitos algoritmos existem para formar agrupamentos e podem ser todos utilizados em uma mesma massa de dados se o objetivo da pesquisa for meramente exploratória e avaliativa no sentido de comparar os resultados das diferentes técnicas. Neste estudo, o método de Ward mostrou-se mais eficiente por distribuir as cisternas entre os grupos de maneira mais igualitária assim como a obtenção de ótimos valores da estatística da silhueta.

De acordo com Barroso e Artes (2003), na medida do possível, pode-se utilizar mais de um método em uma mesma massa de dados e, através da comparação dos grupos formados, pode-se adotar a solução que melhor representa a situação em estudo. Serão divididas e apresentadas, neste trabalho, as técnicas de agrupamentos, denominadas hierárquicas e não hierárquicas.

2.5.1 Técnicas hierárquicas

Os agrupamentos hierárquicos são realizados por sucessivas fusões ou por sucessivas divisões. Os métodos hierárquicos aglomerativos iniciam com tantos grupos quanto os objetos, ou seja, um agrupamento é formado por um único objeto ou indivíduo. Inicialmente, os objetos mais similares são agrupados e fundidos formando um único grupo. Eventualmente o processo é repetido, e, com o decréscimo da similaridade, todos os subgrupos são fundidos, formando um único grupo com todos os objetos.

Os métodos hierárquicos divisivos trabalham na direção oposta. Um único subgrupo inicial existe com todos os objetos ou indivíduos e estes são subdivididos em dois subgrupos de tal forma que exista o máximo de semelhança entre os objetos dos mesmos subgrupos e a máxima dissimilaridade entre objetos ou indivíduos de subgrupos distintos (ZHONG *et al.,* 2008). Estes subgrupos são posteriormente subdivididos em outros subgrupos dissimilares. O processo é repetido até que haja tantos subgrupos quanto objetos ou indivíduos.

Os resultados finais desses agrupamentos podem ser apresentados por gráficos denominados dendogramas. Os dendogramas apresentam os elementos e os respectivos pontos de fusão e divisão dos grupos formados em cada estágio.

Os esforços desta seção serão concentrados na discussão dos métodos hierárquicos aglomerativos ("*Linkage Methods*"). Serão discutidos os métodos de ligação simples (mínima distância ou vizinho mais próximo), ligação completa (máxima distância ou vizinho mais distante), ligação média (distância média), Ward.

Um algoritmo geral para os agrupamentos hierárquicos aglomerativos com *n* objetos (itens, indivíduos ou variáveis) é o seguinte:

1. começa com *n* grupos, cada um contendo um só indivíduo;
2. calcula a distância entre cada um dos grupos e determina os grupos com distância mínima, digamos U e V, cuja distância se denota como d_{UV};
3. une os grupos U e V e nomeia um novo grupo (UV), calculando novas distâncias entre este novo grupo recém formado e os demais grupos;
4. repete os passos 2 e 3 um total de *n-1* vezes, isto é, até que todos os indivíduos pertençam a um mesmo grupo,registrando os grupos que vão se unindo e as distâncias em que ocorreu a união.

Existe uma grande variedade de métodos aglomerativos, que têm como características o critério utilizado para definir a distância entre os grupos. Contudo, a maioria dos métodos parecem ser extensões de três grande conceitos de agrupamento(ANDERBERG, 1973):

1) métodos de ligação (single linkage, complete linkage, average linkage, median linkage);
2) método do Centroide;
3) métodos de soma de erros quadráticos ou variância (método de Ward).

De uma maneira geral, os métodos aglomerativos seguem os passos de um algoritmo padrão como descrito anteriormente. A diferença entre os métodos ocorre no passo 3, pois, neste momento, a função distância é definida de acordo com cada método (JOHNSON, 1992).As distâncias estão definidas para cada método como segue.

a) Método *Single Linkage* ou ligação por vizinho mais próximo:

Baseia-se na menor distância entre um objeto (UV) e um objeto W, ou seja:

$$d_{(UV)W} = \min\left(d_{UW}, d_{VW}\right)$$

b) Método *Complete Linkage* ou ligação por vizinho mais distante:

Baseia-se na maior distância entre um objeto (UV) e um objeto W, ou seja:

$$d_{(UV)W} = \max\left(d_{UW}, d_{VW}\right)$$

c) Método *Average Linkage* ou ligação por média:

Baseia-se na média entre todos os objetos ou indivíduos de (UV) e os de W, isto é:

$$d_{(UV)W} = \frac{\left(N_U \times d_{UW} + N_V \times N_{VW}\right)}{N_U + N_V}$$

em que N_U e N_V são os números de elementos dos grupos U e V, respectivamente, e d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

d) Método *Centroid Linkage* ou ligação por centroide:

Baseia-se na distância entre os centroide dos grupos (UV) e W, que são definidos como a média das coordenadas de todos os objetos de um grupo, ou seja,

$$d_{(UV)W} = \frac{(N_U \times d_{UW} + N_V \times N_{VW})}{N_U + N_V} - \frac{N_N \times N_V \times d_{UV}}{(N_U + N_V)^2}$$

em que N_U e N_V são os números de elementos dos grupos U e V, respectivamente, e d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

e) Método *Median Linkage* ou ligação por mediana:

Baseia-se na mediana das distâncias entre todos os objetos ou indivíduos do grupo (UV) e do grupo W, ou seja,

$$d_{(UV)W} = \frac{d_{UW} + d_{VW}}{2} - \frac{d_{UV}}{4}$$

em que d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

f) Método da ligação de Ward

Baseia-se na noção de que os grupos das observações multivariadas devem ser agrupados seguindo a forma de uma elipse. Nesse caso, um elemento é alocado a um determinado grupo de forma que minimize a homogeneidade dentro dos grupos, isto é, minimizando a soma de quadrado dos erros dentro dos grupos. Nesse método, a função distância é dada por:

$$d_{(UV)W} = \frac{((N_W + N_U) \times d_{UW} + (N_W + N_V) \times d_{VW} - N_W \times d_{UV})}{N_W + N_U + N_V}$$

em que N_U e N_V são os números de elementos dos grupos U e V, respectivamente, e d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

Existem algumas observações a serem feitas acerca dos métodos hierárquicos aglomerativos, como, por exemplo, os métodos da ligação simples, completa e média puderem ser aplicados tanto para variáveis qualitativas, quanto para quantitativas, enquanto os métodos de centroide e de Ward serem apropriados apenas para

variáveis quantitativas, pelo fato de compararem vetores de médias (BARROSO e ARTE, 2003).

Os métodos aglomerativos hierárquicos são utilizados para fundir grupos intermediários e daí formar uma hierarquia de grupos com base no critério de agregação de cada um desses métodos, gerando um *dendograma* e estatísticas como critério de agrupamentos (SUN *et al.*,2012). A escolha de um número final de grupos em uma massa de dados é subjetiva. O objetivo é que o número de grupos encontrados esteja de acordo com a "partição" natural dos objetos ou indivíduos agrupados ou comparados (MINGOTI, 2005).

Os métodos divisivos trabalham em sentido oposto aos métodos aglomerativos. Inicia-se tomando todos os indivíduos em um só grupo com *n* indivíduos ou objetos. Este único grupo divide-se em dois subgrupos de tal maneira que os indivíduos em um dos subgrupos se encontram distantes dos indivíduos que estão no outro subgrupo. O processo continua até que o número de indivíduos ou objetos seja o mesmo número de grupos.

De acordo com Chavent *et al*.(2007), o agrupamento hierárquico divisivo reverte o processo de agrupamento hierárquico aglomerativo, começado com todos os objetos em um grupo, e sucessivamente dividindo cada grupo em grupos menores. Uma abordagem natural para dividir grupos em dois subgrupos não vazios seria considerar todas as possíveis bipartições.

2.5.2 Técnicas não hierárquicas

Este tipo de método consiste em produzir um número fixo de grupos, digamos K. O número K pode estar preestabelecido ou pode ser obtido como parte do processo. Este tipo de método pode iniciar com uma partição inicial de indivíduos em grupos, ou uma seleção inicial de pontos similar que vai formar um centróide nos grupos.

Os agrupamentos não hierárquicos procuram a partição de *n* objetos em k grupos. Os métodos exigem a pré-fixação de centróides que produzam medidas sobre a qualidade da partição produzida. Um dos mais populares métodos é o das K-médias. Segundo Kassomenos *et al.*(2010), K-médias é um algoritmo não hierárquico amplamente utilizado em muitas aplicações que requerem agrupamentos divisivos. K-médias é um algoritmo particional que determina interativamente todos os grupos em uma única etapa (DORLING e DAVIS, 1995).

O algoritmo das K-médias, de uma forma bastante simplificada, é dividido em três passos:

a) particionar os itens em K grupos iniciais arbitrariamente;
b) percorrer a lista de itens e calcular as distâncias de cada um deles para o centroide (médias) dos grupos, fazer a realocação do item para o grupo em que ele apresentar mínima distância, obviamente se não for o grupo ao qual este pertença, e recalcular os centróides dos grupos que ganharam ou perderam itens ou objetos;
c) repetir o passo (b) até que nenhuma alteração seja feita.

Em suma, a abordagem de agrupamento hierárquico aglomerativo constrói conjuntos juntando várias vezes e fundindo os objetos separados pela distância mais curta. Já,nos métodos não hierárquicos, comumente com o uso do método de K-médias, os grupos são definidos a priori e *n* observações são divididas em K grupos (CLIFFORD *et al.,* 2011).

2.5.3 Agrupamento *Fuzzy*

Além das técnicas estatísticas de análise de agrupamento hierárquica e não hierárquica, outras técnicas como algoritmos evolutivos (Jain, 1999) e agrupamentos fuzzy rede neural podem ser empregadas para formação de agrupamentos.O agrupamento *fuzzy* é uma generalização dos métodos por particionamento, então como nos métodos que usam partição, também é necessário indicar o número inicial de grupos.

Nos métodos por particionamento, os elementos são alocados em cada grupo de forma clara, enquanto os agrupamentos *fuzzy* permitem visualizar o grau de associação de cada elemento a cada grupo, que geralmente se verifica em domínios de dados reais, nos quais cada elemento pertence a distintos grupos com diferentes graus de associação.

O agrupamento fuzzy leva vantagem em relação a outros métodos por particionamento, por fornecer informações mais detalhadas sobre a estrutura dos dados, pois apresenta os graus de associação de cada elemento a cada grupo e, consequentemente, não tem uma alocação clara de elementos para formar grupos. A principal desvantagem desse método consiste no crescimento da quantidade de coeficientes de associação com o aumento do número de elementos e de grupos. Contu-

do, é uma técnica válida, pois ela associa graus de incerteza aos elementos nos grupos, tratando-se portanto de uma situação estocástica, que se aproxima mais das características reais dos dados (KAUFMAN, 1990).

No agrupamento *fuzzy*, cada objeto pertence a mais de um grupo com diferentes graus de pertinência, como em lógica *fuzzy,*em vez de pertencer apenas um grupo. Um determinado objeto perto do centro de um grupo pertence a esse grupo com um grau mais elevado do que um objeto que está situado na extremidade desse grupo. Para cada objeto *Y*, o grau de pertinência descreve o quão forte esse objeto pertence a um determinado grupo (ZALIK, 2010).

2.6 Modelo Linear Multivariado

O modelo linear multivariado é considerado uma generalização do caso univariado. Este modelo trata do estudo com mais de uma variável resposta (p ≥ 2), isto é, do relacionamento entre *p* variáveis respostas Y_1, Y_2, ..., Y_p e um único conjunto de variáveis explicativas X_1, X_2, ... , X_r, em que:

$$Y_1 = \beta_{01} + \beta_{11}x_{11} + \cdots + \beta_{r1}x_{1r} + \varepsilon_1$$
$$\vdots$$
$$Y_p = \beta_{0p} + \beta_{1p}x_{n1} + \cdots + \beta_{rp}x_{nr} + \varepsilon_p$$

em que

β_{rp} são os parâmetros do modelo;

eε_p é erro associado ao modelo.

Assim, de acordo com Johnson e Wichern (2002) e Cuadras (2006), o modelo linear multivariado tem a seguinte forma matricial

$$\mathbf{Y = X\beta + E}$$

em que

Y é a matriz de variáveis (*n*×*p*), que contém *n* observações multivariadas sobre *p* variáveis dependentes;

X é a matriz de delineamentos (*n*× (*r* + 1));

E é a matriz de erros aleatórios (*n*×*p*), com vetor de médias $\boldsymbol{\theta}$ e matriz de variância covariância $\boldsymbol{\Sigma}$, com $E(\varepsilon_i)$ = 0 e $Cov(\varepsilon_j, \varepsilon_{j'}) = s_{jj'}$, para j, j' = 1, 2, ... , *p*.

sendo:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{21} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = [\boldsymbol{y}_1 \quad \vdots \quad \ldots \quad \vdots \quad \boldsymbol{y}_p]$$

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1r} \\ x_{20} & x_{21} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nr} \end{bmatrix} = [\boldsymbol{X}_0 \quad \vdots \quad \boldsymbol{X}_1 \quad \vdots \quad \ldots \quad \vdots \quad \boldsymbol{X}_r]$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rp} \end{bmatrix} = [\beta_1 \quad \vdots \quad \ldots \quad \vdots \quad \beta_p]$$

$$\mathbf{E} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \ldots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix} = [\boldsymbol{\varepsilon}_1 \quad \vdots \quad \ldots \quad \vdots \quad \boldsymbol{\varepsilon}_p]$$

A matriz $\boldsymbol{\beta}$ de parâmetros pode ser estimada através do método de mínimos quadrados, de forma similar ao caso univariado. Uma outra maneira de estimar a matriz $\boldsymbol{\beta}$,assim como a matriz de variância covariância $\boldsymbol{\Sigma}$, é pelos estimadores de máxima verossimilhança (ZHANG e RAO, 2006).

2.7 Definição do Número de Grupos

A determinação do número de grupos envolve a escolha do algoritmo de agrupamento e a decisão quanto ao número de grupos. Essa determinação da quantidade de grupos em uma massa de dados é um dos maiores problemas encontrados no processo de agrupamento.

Segundo Fernandez *et al.* (2010), pode-se modelar preferências de algum agente de decisão para agrupar objetos que são suficientemente semelhantes em sentido de preferência e suficientemente diferentes de outros objetos, estabelecendo uma ordem de preferência sobre um conjunto de agrupamentos.Uma primeira etapa na formação de grupos envolve o uso de agrupamento hierárquico para desenvolver uma partição inicial, e, em seguida, diferentes métodos estatísticos são utilizados em

combinação para determinar o número de grupos que resultam da análise de agrupamento hierárquico (XU e FUREY, 2007).

De acordo com Peng *et al*. (2012), determinar o número de grupos em um conjunto de dados é essencial, mais difícil na análise de agrupamento. Uma vez que essa tarefa envolve mais de um critério, pode ser modelada como um problema de múltiplos critérios de decisão. Tibshirani *et al*. (2001), propôs o método da diferença estatística para estimativa do número de agrupamentos em conjunto de dados. Já Dudoit e Fridlyand (2002) estimou o número de grupos usando o método de reamostragem com base em predição. Para Lattin *et al*.(2003), a determinação apropriada do número de grupos pode considerar diversas abordagens possíveis, isto é, o pesquisador pode especificar antecipadamente o número de grupos, e, talvez por motivos teóricos e lógicos, esse número seja conhecido. O pesquisador também pode ter razões práticas para estabelecer o número de agrupamentos, com base no uso que pretende fazer deles.

De acordo com Pal e Biswas (1997), um algoritmo de agrupamento ótimo não existe, isto é, diferentes algoritmos ou mesmo diferentes configurações do mesmo algoritmo produzem partições diferentes, e nenhum deles provou ser melhor em todas as situações. Assim, em um processo de agrupamento, muitos algoritmos de agrupamento não são capazes de calcular o número de grupos natural existente nos dados, fazendo-se necessário que essa informação seja fornecida inicialmente,a qual é frequentemente conhecida como parâmetro K (ALBALATE *et al*., 2011).

2.7.1 Validação dos grupos

As técnicas de validação de grupos ou *clusters* são frequentemente utilizadas em combinação com algoritmos de agrupamento, a fim de identificar o número ideal de grupos. A abordagem para validação dos grupos é executada várias vezes utilizando um algoritmo de agrupamento, com um grande número de candidatos a grupos, e, em seguida, é selecionado um valor tal que otimiza a qualidade da solução do agrupamento de acordo com um determinado critério de qualidade (ARBELAITZ *et al*., 2013).

De acordo com Jain e Dubes (1988), evidenciado por Falasconi *et al.* (2007), o procedimento de validação pode ser expresso em termos de três tipos de critérios: externo, interno e relativo.

O critério *externo* de validação principal avalia o grau de associação em duas partições do conjunto de dados, isto é, aplica-se à comparação de partições que têm um número arbitrário de grupos cada. Uma partição vem a partir de uma solução do agrupamento. A segunda partição é designada *a priori* e independentemente da primeira partição, que pode ser selecionada pela verdadeira característica da amostra, ou essa partição pode ser obtida por outro método de agrupamento diferente.

O critério *interno* determina se a estrutura de agrupamento é essencialmente apropriada para os dados, usando apenas informações contidas nos dados propriamente ditos. O teste de hipótese fornece um quadro para decidir o quanto apropriada é essa estrutura de agrupamento. Uma vez apropriada, o resultado da análise de *cluster* é comparado com a distribuição dos resultados obtidos na hipótese nula apropriada, ou hipótese alternativa, que é a distribuição de referência ou básica. Quando aplicada a uma partição de dados, o critério de validação interna dá o número ideal de grupos no conjunto de dados, o que se constitui em um grande desafio na Análise de *Cluster*.

O critério da validação relativa serve para comparar duas estruturas de agrupamento e para decidir qual delas se adequada melhor aos dados. A diferença do critério interno é apenas pelo modo como ele é aplicado, portanto,cada índice interno é usado como índice relativo. Normalmente, obtém-se uma sequência de partições de dados, e, em seguida, investigam-se os valores do índice sobre a sequência que observa algum aspecto em comum, tal como o máximo, o mínimo ou uma observação discrepante, que indique um melhor ajuste aos dados propriamente ditos.

2.7.2 Passos para validação dos Grupos

De acordo com Lattin *et al.* (2003) os passos para validação dos grupos são:

1. Dividir os dados em duas amostras aleatórias: Calibração e Validação.
2. Usar método de agrupamento para dados de calibração, determinar o número apropriado de grupos e calcular os centróides.
3. Uma vez encontrados os centróides dos grupos a partir dos dados de calibração, atribuir cada observação da amostra de validação para o centroide mais próximo. Este grupo será chamado de solução S_1.

4. Utilizar o mesmo método de agrupamento do passo 2 para agrupar os dados de validação, escolher a solução com o mesmo número de grupos determinado no passo 2 e denotar este grupo de solução S_2;
5. Uma vez que as soluções dos grupos S_1 e S_2representam diferentes atribuições do mesmo conjunto de dados para os grupos, cruzar uma tabela de S_1 versus S_2, para avaliar a concordância entre as duas soluções.

2.7.3 Índice de Rand

O Índice de Rand faz distinções entre duas partições com base na mediana ou centroide de cada partição. Várias funções de similaridade são utilizadas para essa finalidade. A escolha mais comum é a utilização de funções baseadas na mediana, entre as decisões tomadas pelas duas partições em pares individuais de objetos. Essa vertente de pesquisa foi desenvolvida por Rand (1971) e evidenciada por Carpineto e Romano (2012). Seja o conjunto A de n objetos e duas partições A_1 e A_2 do conjunto A, então o Índice de Rand (IR) é definido como

$$IR = \frac{n_{11} + n_{00}}{n_{00+}n_{01}+n_{10} + n_{11}} = \frac{n_{11} + n_{00}}{\binom{n}{2}}$$

em que:

n_{11}: Número de pares de objetos que se encontram no mesmo grupo em ambas as partições A_1 e A_2;

n_{10}: Número de pares de objetos que se encontram em diferentes grupos em ambas as partições A_1 e A_2;

n_{01}: Número de pares de objetos que se encontram no mesmo grupo em A_1, mas em diferentes grupos em A_2;

n_{10}: Número de pares de objetos que se encontram em diferentes grupos em A_1, mas no mesmo grupo em A_2.

Intuitivamente, pode-se pensar em n_{11}+ n_{00} como sendo o número de acordos entre as partições A_1 e A_2 e n_{10} + n_{01} como sendo o número de desacordos entre as partições A_1 e A_2. O Índice de Rand está entre 0 e 1, isto é $0 \leq IR \leq 1$, sendo 0 (zero) quando não ocorre acordo em qualquer par de objetos, pois só ocorre quando uma partição consiste de um único grupo, enquanto a outra partição é composta apenas

de grupos contendo objetos individuais, sendo 1 (um) quando as partições são coincidentes.

2.7.4 Índice de Rand Ajustado

O Índice de Rand Ajustado proposto por Hubert e Arabie (1985) assume a distribuição hipergeométrica generalizada como modelo probabilístico de aleatoriedade, isto é, as partições A_1 e A_2 são escolhidas aleatoriamente de forma que o número de objetos nos grupos e o número de grupos são fixos. O Índice de Rand Ajustado assume valor 0 (zero) quando o índice é igual ao valor esperado e é limitado por 1 (um). Tanto no índice de Rand quanto no índice de Rand Ajustado, os dois tipos de acordos, isto é, n_{11} e n_{00} têm a mesma importância. No entanto, em muitos agrupamentos, colocar um par de objetos em diferentes grupos geralmente não é tão informativo quanto colocar os dois objetos em um mesmo grupo. De acordo com Campello (2007), os pares do tipo n_{00} não são claramente indicativos de similaridade ou dissimilaridade, opondo-se aos pares do tipo n_{11} que são mais informativos. Nesse sentido, os pares n_{10} e n_{01} também não são explícitos com relação aos pares de objetos nomeados a cada grupo.Este Índice é obtido de maneira similar ao Índice Jaccard (HUR et al., 2002).

$$JC = \frac{n_{11}}{n_{11+}n_{01}+n_{10}}$$

Uma outra variante do Índice de Rand é a diferença da distância simétrica (SDD), fazendo o uso apenas de divergências, e é definida como

$$SDD = n_{01}+n_{10} = \binom{n}{2} - (n_{11} + n_{00})$$

A medida SDD tem sido utilizada como função objetivo a ser otimizada em vários trabalhos de consenso em agrupamento, incluindo uma versão com agrupamentos selecionados de forma ponderada.

2.7.5 Estatística Silhouette

A Estatística da Silhueta foi proposta originalmente por Rousseeuw(1987), para qualidade da análise de agrupamento, e evidenciada por Lin *et al*. (2013), para classificação de subtipos de leucemia linfoblástica aguda. A Estatística da Silhueta surgiu da necessidade de classificar um número de classes “K” plausível no intuito de otimizar a qualidade dos agrupamentos.Assim, para uma observação i pertencente a uma dada classe “A”, calcula-se a distância média entre i e os demais indivíduos de “A”, a(i). Logo, para cada uma das classes “C” diferentes de “A”, calcula-se a distância média entre i e os indivíduos de “C”, e seja b(i) o mínimo dessas distâncias. Portanto, para o indivíduo i, o valor da Estatística da Silhueta é dada por

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i),\ b(i)\right)}$$

A Estatística da Silhueta está no intervalo de -1 a 1, isto é, -1 ≤ s(i) ≤ 1, então valores negativos de s(i) dão fortes indícios de que o indivíduo i seja semelhante a indivíduos de outras classes.Valores de s(i) perto de 1 sugerem que i esteja bem classificado.

2.7.6 Gráfico da Silhueta

O gráfico de Silhueta é um procedimento descritivo para verificar a qualidade dos agrupamentos formados. A ideia do método é verificar se um ponto está mais próximo dos elementos do seu próprio grupo ou de elementos de grupos vizinhos. Esse gráfico é baseado no cálculo de duas medidas vistas anteriormente na Estatística da Silhueta, que é a distância média entre o objeto i os elementos do seu próprio grupo a(i) e a distância média entre o objeto i e os elementos do grupo mais próximo de i que é denotado por b(i), que não seja o seu próprio grupo.

Seja G(i) o grupo ao qual pertence o objeto i, supunha que existe $n_{G(i)}$ observações nesse grupo. Assim, temos que

$$a(i) = \frac{\sum_{j \epsilon G(i),\ j \neq i} d_{ij}}{n_{G(i)} - 1}$$

em que d_{ij} é a distância euclidiana entre os objetos i e j.

Para cada diferente grupo de G(i), determina-se a distância média entre seus elementos e i. Define-se o grupo W(i) como o que apresenta menor distância média entre seus elementos e o ponto i, admite-se ainda que o número de objetos de W(i) seja $n_{W(i)}$. O grupo W(i) é denominado de vizinho de i. Logo, temos

$$b(i) = \frac{\sum_{j \epsilon W(i),\ j \neq i} d_{ij}}{n_{W(i)}}$$

Nos gráficos de Silhuetas, representa-se o número de grupos no eixo das ordenadas por barras, cujo comprimento é igual aos respectivos valores de s(i) que podem ser lidos no eixo das abscissas. As barras no gráfico de Silhueta são classificadas de acordo com o número de grupos formados, e, em cada um deles, essas barras estão em ordem decrescentes de acordo com o valor de s(i). À sua direita, está a identificação dos grupos com as suas respectivas quantidades de elementos e seus valores de Silhuetas.

A silhueta foi proposta por Rousseeuw (1987), para aplicação a métodos supervisionados, contudo Kaufman e Rousseeuw (1990) afirmam que a mesma pode ser estendida para qualquer método de obtenção de agrupamento.

2.7.7 Índice de Kappa

Segundo Rosenfield e Fitzpatrick-Lins (1986), afirmado uma década depois por Brites *et al.* (1996), o índice de Kappa é um coeficiente de concordância para escalas nominais que mede o relacionamento entre a concordância, além da casualidade e concordância esperada. O índice de Kappa é a proporção de concordância depois que a concordância devido à casualidade é desconsiderada, ou seja,

$$K = \frac{p_0 - p_c}{1 - p_c}$$

em que:

p_0: proporção de unidades que concordam;

p_c: proporção de unidades que concordam aleatoriamente.

Essas proporções são dadas por:

$$p_0 = \sum_{i=1}^{R} \frac{n_i^2}{n}$$

e

$$p_c = \sum_{i=1}^{R} \frac{n_i^2}{n^2}$$

As proporções p_0 e p_c podem ser consideradas como proporção observada de pares concordantes e como proporção esperada de pares concordantes devido ao acaso(CAMPBELL, 1987).

2.7.8 Índice de Jaccard

De acordo com Reis (2001), o índice de Jaccard é definido como

$$S_{ij} = \frac{a}{a+b+c}$$

A distância de Jaccard pode ser utilizada também para dados binários, de acordo com Ferreira (2008):

$$d_{ij} = \frac{b+c}{a+b+c}$$

Essas distâncias evitam a contribuição da ausência conjunta de uma determinada característica para o cálculo da similaridade ou dissimilaridade entre dois indivíduos. Existem muitos outros índices de similaridade que foram propostos, todos com justificativas plausíveis para suas utilizações (JOHNSON e WICHERN, 1998).

2. 8 Tipos de Distância

Em estudos de análise de agrupamento, o ponto de partida é uma matriz de dissimilaridade entre as n observações ou cisternas. A natureza da medida de dissimilaridade é o ponto da análise dos dados que irá induzir ou condicionar ao agru-

pamento que se sucederá, sendo portanto de extrema importância que se faça uma reflexão a respeito da temática em estudo.

De maneira geral, a dissimilaridade d_{ij} entre as cisternas i e j são medidas que refletem as maiores ou menores diferenças entre os valores que essas cisternas podem assumir em um conjunto de p variáveis. Existem várias medidas que podem ser utilizadas como medidas de distâncias ou dissimilaridade entre observações (cisternas) de uma matriz de dados.

De acordo com Cormack (1971), existe uma série de medidas de dissimilaridade possível, a saber: distância euclidiana, distância euclidiana generalizada, distância de Mahalanobis, distância de Minkowski, distância ou métrica de Camberra.

2.8.1 Distância Euclidiana

Segundo Hair *et al.* (2010), este é o coeficiente de dissimilaridade mais conhecido e utilizado para indicar a proximidade entre objetos. É simplesmente a distância geométrica entre duas cisternas em um espaço multidimensional.A distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações ou cisterna de i e j para todas as p variáveis (MCROBERTS *et al.*, 2007).

$$\begin{aligned} d_{ij} &= \left\| x_i - x_j \right\| \\ &= \sqrt{(x_i - x_j)^t (x_i - x_j)} \\ &= \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \end{aligned}$$

onde: x_i é o vetor da i-ésima observação, e x_j é o vetor da j-ésima observação.

2.8.2 Distância Euclidiana Generalizada

Esta distância deriva da distância euclidiana que está bastante ligada à análise de agrupamento, ponderando os objetos ou cisternas nos agrupamentos, ou seja,são atribuídos pesos aos objetos considerados mais importantes no estudo realizado. A referida distância é dada pela seguinte expressão

$$d_{ij} = \|x_i - x_j\|_W$$
$$= \sqrt{(x_i - x_j)^t W (x_i - x_j)}$$
$$= \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} w_{kl}(x_{ik} - x_{jk})(x_{il} - x_{jl})}$$

onde W é uma matriz definida positiva, x_i é o vetor da i-ésima observação, e x_j é o vetor da j-ésima observação. Ainda há $W = D^{-2}$, onde D^{-2} é a matriz diagonal das variâncias, o que corresponde a tomar a distância euclidiana usual entre os dados normalizados, e $W = \Sigma^{-1}$, onde Σ é a matriz de variâncias-covariâncias das variáveis. Essa escolha gera a chamada distância de Mahalanobis, que é invariante a mudanças de escala nas variáveis.

2.8.3 Distância de Mahalanobis

Esta medida, ao contrário da distância euclidiana, considera a matriz de covariância Σ para o cálculo das distâncias, com o objetivo de corrigir o problema de escala entre as variáveis. A distância de Mahalanobis entre os grupos i e j é usualmente estimada, de acordo com Rao (1952),

$$d_{ij} = (\mathbf{X}_i - \mathbf{X}_j)' \Sigma^{-1} (\mathbf{X}_i - \mathbf{X}_j)$$

em que: $\boldsymbol{X}_i$ e $\boldsymbol{X}_j$ são respectivamente os vetores de valores das variáveis para os indivíduos; ei e j e Σ^{-1} é a matriz de variância/covariância. A distância de mahalanobis é usada em análises de *clusters*, como técnica de classificação, para escolha de um vetor médio mais próximo à vizinhança de um ponto de teste (RABAL *et al.,* 2012).

2.8.4 Distância de Minkowski

A distância de Minkowski é uma generalização das demais distâncias, pois as distâncias são basicamente normas de vetores. Esta distância é dada pela seguinte expressão

$$d_{ij} = \left(\sum_{k=1}^{p} | x_{ik} - x_{jk} |^{\lambda} \right)^{1/\lambda}$$

em que:

λ = 1 designa a distância de Manhattan ou distância ℓ1;

λ = 2 designa a métrica euclidiana usual ou distância ℓ2;

$\lambda \to \infty$ obtém $d_{ij} = \max_k |x_{ik} - x_{jk}|$ que se designa a distância máxima ou distância.

2.8.5 Métrica de Camberra

Para variáveis que apenas possam tomar valores não-negativos, pode definir-se a métrica

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

em que o objetivo é obter uma medida de dessemelhança invariante a transformações de escala diferenciadas em cada variável. Esse objetivo resulta da relativização de cada diferença através do denominador cujas unidades de medida são iguais às do denominador.

2. 9 Parâmetros de Qualidade da Água

Na gestão dos recursos hídricos, os aspectos de quantidade e qualidade não podem ser desintegrados, o que reforça a importância da avaliação da disponibilidade hídrica, em termos qualitativos, de águas superficiais e subterrâneas. Essa avaliação é tão importante que indicadores sociais e de qualidade de vida podem ser mostrados através de dados de qualidade de água. Os principais parâmetros que indicam poluição nos recursos hídricos são: temperatura da água, potencial hidrogeniônico, condutividade elétrica, sólidos dissolvidos totais e coliformes totais e fecais.

A temperatura pode ser considerada a característica mais importante do meio aquático. A temperatura é responsável por alterações em grande parte dos outros parâmetros físicos da água, tais como a densidade, a viscosidade, a pressão de vapor e a solubilidade dos gases dissolvidos (TUCCI, 2004). A temperatura é um im-

portante fator modificador da qualidade da água, pela influência direta sobre o metabolismo dos organismos aquáticos e pela relação com os gases dissolvidos. Assim, os aumentos de temperatura diminuem as concentrações de oxigênio dissolvido, de gás carbônico, de pH e a viscosidade, entre outras propriedades (HAMMER,1979 e SAWYER et al., 1994).

O pH pode ser alterado de acordo com o tipo de solo com que a água tem contato, ou pode subir muito em lagoas com grande população de algas, nos dias ensolarados, chegando a 9 ou mais. Isso porque as algas, ao realizarem fotossíntese, retiram muito gás carbônico, que é a principal fonte natural de acidez da água. Geralmente um pH muito ácido ou muito alcalino está associado à presença de despejos industriais.

Segundo Esteves (1998), é comum encontrar altos valores de pH em regiões de balanço hídrico negativo, como ocorre com os açudes do semiárido no Nordeste brasileiro. Na época de estiagem, esse fato é acentuado pelos altos valores de carbonatos e bicarbonatos encontrados nas águas, os quais se tornam mais concentrados pela evaporação.

O parâmetro condutividade é de fundamental importância no estudo de qualidade da água, no entanto, não determina, especificamente, quais os íons que estão presentes em determinada amostra de água, mas pode contribuir para possíveis reconhecimentos de impactos ambientais que ocorram no reservatório de água ocasionados por lançamentos de resíduos industriais, de detritos de mineração, de esgotos etc. A condutividade elétrica da água pode variar de acordo com a temperatura e com a concentração total de substâncias ionizadas dissolvidas. Em águas cujos valores de pH se localizam fora do intervalo ($5 < pH < 9$), os valores de condutividade são devidos apenas às altas concentrações de poucos íons em solução, dentre os quais os mais frequentes são o H^+ e o OH^-.

A concentração de sólidos dissolvidos totais (S.D.T.) é um grande problema, pois, em excesso de S.D.T., a água é imprópria para o consumo humano,pois apresenta paladar desagradável e problemas de corrosão de tubulações, além de seu consumo poder causar o acúmulo de sais na corrente sanguínea e,consequentemente, cálculos renais.

De acordo com Lima (2014), os coliformes totais incluem todas as bactérias na forma de bastonetes gram-negativos, não esporogênicos, aeróbicos ou anaeróbicos facultativos, capazes de fermentar a lactose com produção de gás, em 24 a 48 horas

a 35°C. Segundo Hitchins *et al.*(1996), esta definição é a mesma para o grupo dos coliformes fecais, mas restringindo-se aos membros capazes de fermentar a lactose com produção de gás, em 24 horas a 45,5°C. O índice de coliformes totais avalia as condições higiênicas, já as taxas de coliformes fecais são utilizadas como indicadores apenas de contaminação fecal, avaliando assim as condições higiênico-sanitárias deficientes, tendo em vista que a população deste grupo é constituída de uma alta proporção de *Escherichia coli* (SIQUEIRA, 1995).

Portanto, para os coliformes totais e coliformes fecais, a interpretação é baseada na ausência ou presença de micro-organismos, sendo necessário o estudo de cada amostra de forma individual.

3 - MATERIAIS E MÉTODOS

3.1 Área de Estudo

Este estudo foi realizado junto a comunidades do Sertão do Pajeú – PE, localizado no município de Serra Talhada, situado na microrregião do Pajeú, no Sertão do Estado de Pernambuco.

A área de estudo está localizada nas comunidades de Serra Grande – Poço do Serrote, Poldrinho, Catolé e Três Irmãos, sendo este próximo à Serra da Lagartixa, pertencente ao limite municipal entre as cidades de Serra Talhada (Figura 1) e Floresta, região do alto Sertão do Pajeú, ambientes semiáridos do Estado de Pernambuco, de coordenadas geográficas de 38°23'55.51" longitude Oeste e 8°07'06.72" latitude Sul.

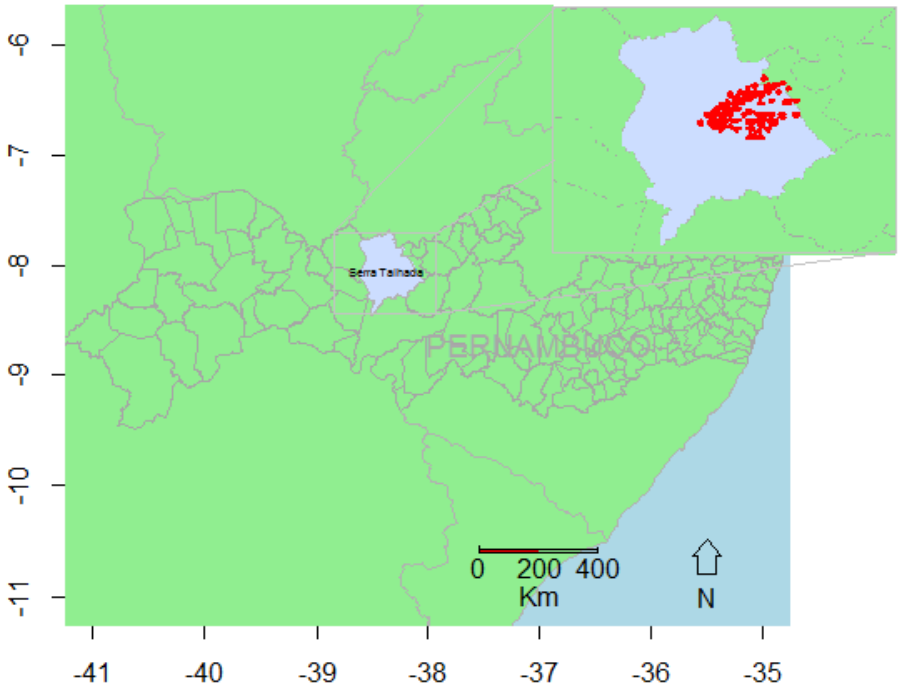

Figura 1. Localização Geográfica do município de Serra Talhada – PE

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

Os dados foram obtidos junto ao projeto de extensão aprovado pelo edital Nº 11/2014 CNPq/MDA/SPM, intitulado “Potencialidade do uso da água, manejo florestal e suas implicações na qualidade de vida das mulheres, adultas e idosas de comunidades do Sertão do Pajeú – PE, localizado no município de Serra Talhada, situado na microrregião do Pajeú no Sertão do Estado de Pernambuco”. A coleta dos dados foi obtida a partir de 100 cisternas de placas alocadas nas referidas comunidades. Os dados foram analisados no Software Estatístico R, em sua versão 3.2.2, com a utilização dos pacotes *cluspam*, *fclust, hclust* e *lattice.*

As cisternas de placas, comparadas a outros sistemas de captação de água, apresenta um baixo custo e um curto período de construção devido à simplicidade e à 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.

Segundo Lima (2014), o nome cisterna de placa surgiu por causa do material utilizado em sua construção, isto é, placas de cimentos pré-moldados, confeccionadas à parte, antes da montagem da estrutura da cisterna. A cisterna apresenta forma cilíndrica conforme 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 de novembro de 2014 a março de 2016, período esse considerado de maior incidência de eventos pluviométricos nas regiões semiáridas, em especial no sertão pernambucano. Essas 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 para colocação da água da cisterna, sendo 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 no Assentamento Poço do Serrote, no Assentamento Poldrinho, no Assentamento Catolé e no 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, foram levadas para a análise da qualidade da água no laboratório de análise química do Instituto Federal de Educação, Ciência e Tecnologia (IFPE), *Campus* de Vitória de Santo Antão – PE, na zona da mata pernambucana.

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 detectados coliformes fecais e totais, utilizando-se a técnica de Tubos Múltiplos, de acordo com o Standard Methods (APHA, 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 para consumo humano.

De acordo com o Conama (1986), os padrões de referências das variáveis de qualidade da água para o consumo humano são estabelecidos da seguinte maneira: parâmetros de condutividade elétrica cátions e ânions $<\mp 0,5$; o potencial hidrogeniônico (pH) está no intervalo de $6 < pH < 9,5$; os Sólidos Dissolvidos Totais (S.D.T) devem apresentar taxas inferior a 500 mg L^{-1}. Já, nos coliformes fecais e totais, é necessário que, em 40 amostras ou mais coletadas a cada mês,haja ausência de 100 ml de microorganismos em 95% das amostras examinadas por mês. Na Tabela 1, encontram-se os valores das estatísticas descritivas das variáveis de qualidade da água estudadas e analisadas nesta tese.

Tabela 1. Estatísticas descritivas das variáveis em estudo

Variáveis de Qualidade da Água	Mínimo	Média	Maximo
Cátions	0,3952	0,8603	2,1738
Ânions	0,3519	0,8527	2,2087
Potencial hidrogeniônico pH	0,8986	0,9930	1,1264
Sólidos Dissolvidos Totais (S.D.T.)	0,6244	0,9839	1,2345
Coliformes totais	0,0827	0,6963	2,6264
Coliformes fecais	0,3811	0,7347	2,9214

Como se observa na Tabela 1, foram medidos os valores mínimos, médio e máximo das seis variáveis estudadas neste trabalho.

3.2 Métodos

A partir dos dados da água contida em cisternas na região do Pajeú – PE, com variáveis de qualidade da água, e da aplicação da metodologia para análise de agrupamento, consideraram-se as seguintes análises: obtenção das matrizes de distância, realização de uma análise exploratória dos dados através de boxplots para verificar a existência de grupos naturais ou *outliers*, realização do teste de normalidade, utilização dos métodos de agrupamento hierárquico para construção dos grupos com os métodos da média das distâncias, da ligação simples, da ligação completa, do centroide, de Ward, da mediana e de mcquitty através da construção dos dendogramas com a distância euclidiana e a distância ou métrica de camberra, sendo as medidas de proximidade utilizadas na correlação cofenética, e na estatística da silhueta. Ainda se utilizaram os métodos não hierárquicos de agrupamento com agrupamento Fuzzy, *k* - médias ou *k* – *means* e o *k* – *medoid*, todos utilizando a distância euclidiana como a medida de proximidade.

3.2.1 Medidas de Distância

3.2.1.1 Distância Euclidiana

A distância entre dois objetos ou duas cisternas (i e j) foi utilizada para medir a similaridade ou dissimilaridade entre duas cisternas de acordo com as características das variáveis de qualidade da água. Essa distância é dada pela expressão

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

em que:

d_{ij} = distância euclidiana entre a i-ésima e j-ésima cisterna;

X_{ik}: característica observada na cisterna i;

X_{jk}: característica observada na cisterna j;

p: número de variáveis em estudo.

A distância euclidiana foi uma das medidas utilizada nos métodos hierárquicos e foi também utilizada nos métodos não hierárquicos de agrupamento.

3.2.1.2 Métrica de Camberra

Para variáveis que apenas possam tomar valores não-negativos, pode definir-sea métrica

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

em que:

d_{ij} = distância euclidiana entre a i-ésima e j-ésima cisterna;

X_{ik}: característica observada na cisterna i;

X_{jk}: característica observada na cisterna j;

p: número de variáveis em estudo.

O objetivo é obter uma medida de dessemelhança invariante a transformações de escala diferenciadas em cada variável.Esse objetivo resulta da relativização de cada diferença através do denominador cujas unidades de medida são iguais às do denominador.

3.2.2 Agrupamento Hierárquico

A classificação em grupos procede por etapas, em geral se determinando, a partir de *n* subgrupos de um único indivíduo cada, sucessivas fusões de subgrupos considerados mais “semelhantes”. Cada fusão reduz, em uma unidade, o número de subgrupos. Dados dois grupos U e V, para medir a semelhança/dessemelhança entre eles, os seguintes algoritmos de agrupamento são utilizados:

1) Método da Médias das Distâncias entre Grupos(*average linkage*)

Consiste em considerar que a distância entre dois grupos é a média de todas as distâncias entre pares de observações ou cisternas (um em cada grupo):

$$D_{UV} = \frac{1}{n_U . n_V} \sum_{i=1}^{n_U} \sum_{j=1}^{n_V} d_{ij}$$

2) Método do Vizinho Mais Próximo(***single linkage***)

Consiste em considerar que a distância entre dois grupos é a menor distância entre um objeto de um grupo e um objeto ou cisterna do outro grupo:

$$D_{UV} = \min_{\substack{i \in U \\ j \in V}} d_{ij}$$

3) Método do Vizinho Mais Distante(*complete linkage*)

Consiste em considerar que a distância entre dois grupos é a maior distância entre um objeto ou cisterna de um grupo e um objeto ou cisterna do outro grupo:

$$D_{UV} = \max_{\substack{i \in U \\ j \in V}} d_{ij}$$

4) Método dos Centróides (*Centroid method*)

Neste caso, toma-se a distância entre dois grupos como sendo a distância entre as médias, ou outros pontos considerados “representativos” (centróides) dos grupos:

$$D_{UV} = \|\overline{x_U} - \overline{x_V}\|$$

5) Método de Ward

Considera-se a imobilidade de um grupo U, isto é, a soma de quadrados das diferenças entre cada objeto ou cisterna e o “objeto médio” desse grupo, dado por:

$$I_U = \sum_{j=1}^{p}\left[\sum_{i\in U}(x_{ij} - \overline{x^{U_j}})^2\right]$$

em que $\overline{x^{U_j}}$ é a média dos valores da variável jpara os objetos ou cisternas do grupo U. Toma-se agora a distância entre os grupos U e V como sendo o aumento na soma total das imobilidades provocado pela divisão dos grupos U e V,isto é, seja I_U a imobilidade do grupo U, I_V a imobilidade do grupo V e $I_{U\cup V}$ a imobilidade do grupo resultante da divisão dos grupos U e V. Então,uma vez que a divisão de U e V não afeta a imobilidade dos grupos remanescentes, temos:

$$D_{UV} = I_{U\cup V} - (I_U + I_V)$$

Note que, se $\overline{\mathbf{x_U}}$ e $\overline{\mathbf{x_V}}$são os vetores de média dos grupos U e V respectivamente, então

$$D_{UV} = \frac{n_U n_V}{n_U + n_V}\|\overline{\mathbf{x_U}} - \overline{\mathbf{x_V}}\|^2$$

O algoritmo de Ward, também conhecido como método da variância mínima, proposto por Orlóci (1978), tem tendência a produzir grupos com um número aproximadamente igual de objetos.

6) Método da Mediana

O método da mediana é semelhante ao método do Centroide. Porém, a atualização dos centróides é feita pela média aritmética, sem a ponderação pelo tamanho dos grupos, ou seja, um novo cluster (uv)w será atualizado por

$$\overline{D}_{(UV)W} = \frac{\overline{D}_{UV} + \overline{D}_{W}}{2}$$

onde $\overline{D}_{UV}$ e $\overline{D}_{W}$ são as distâncias média entre os elementos do grupo UW e do grupo W, respectivamente, e, como mencionado, com base na metodologia proposta por Orlóci (1978) e Gama (1980), esse algoritmo é um caso particular do método do centroide.

7) Método de Mcquitty

Um método de agrupamento também utilizado nas análises em estudo, e bastante difundido, é o método hierárquico de McQuitty, definido como

$$D_{(UV)W} = \frac{D_{UW} + D_{VW}}{2}$$

onde $D_{(UV)W}$ é a distância entre o agrupamento (UV) e o agrupamento W, D_{UW} e D_{VW} são as distâncias entre a maior distância dos elementos dos agrupamentos U e W e os elementos dos agrupamentos V e W.

3.2.3 Agrupamento não Hierárquico

3.2.3.1 Agrupamento Fuzzy

O estudo de agrupamento é de grande interesse quando se deseja classificar um conjunto de dados de acordo com suas características ou variáveis mensuradas.Assim, o termo classe é pertinente, dada a informação de quantas partições e quais são essas partições em um conjunto de dados, bem como cada observação ou cisterna que pertence à partição. Nesse contexto, o trabalho de análise de dados é denominado *classificação* e objetiva determinar uma função que busque de maneira eficaz o mapeamento entre os dados e suas classes,

bem como mapear corretamente novos dados que venham a compor o conjunto de dados, trazendo ou não a informação sobre sua classificação no agrupamento.

O termo grupo deve ser usado quando não existe qualquer informação sobre como é a organização dos dados. Nesse caso, o trabalho de análise de dados é denominado agrupamento e tem por objetivo estudar as relações de similaridade entre os dados ou cisternas, determinando quais dados formam quais grupos.Os grupos são formados de maneira que se maximize a similaridade entre as cisternas de um grupo (similaridade dentro do grupo) e se minimize a similaridade entre cisternas de grupos diferentes (similaridade externa aos grupos). Então, formalmente, dado um conjunto de dados de entrada ($\vec{x} \in \mathbb{R}^{p}$), é encontrada uma função

$$f: \mathbb{R}^{p} \times W \longrightarrow G$$

onde W é um vetor de parâmetros ajustáveis, por meio de um algoritmo de aprendizado supervisionado ou não supervisionado, que determina c-grupos a partir da matriz de dados originais X, e, segundo Xu e Wunsch (2005), tem-se $G = G_1, G_2, \dots, G_c$ $(c \leq n)$, tal que:

i) $G_i \neq \emptyset, i = 1, \dots, c$;

ii) $\cup_{i=1}^{c} G_i = X$;

iii) $G_i \cap G_j = \emptyset, i, j = 1, \dots, c$ e $i \neq j$, supondo a abordagem clássica de classificação ou agrupamentos.

3.2.3.2 Método de K – média (*K – means*)

O método das k-médias consiste em agrupar os objetos ou cisternas em k grupos mutuamente excludentes ou distintos, sendo que, para encontrar esses grupos, o algoritmo utiliza um processo interativo com o objetivo de minimizar a soma das distâncias de cada cisterna em relação ao centroide de cada grupo que será a cisterna mais representativa do respectivo grupo. Assim como em outros métodos por particionamento, a principal diferença entre este método e outros por partição é que o centroide de cada grupo é dado pela média dos mesmos.

Para agrupar as cisternas em grupos, utilizando o método das k-médias, os seguintes passos devem ser seguidos:

1. Particionar as cisternas em k grupos iniciais, aleatoriamente.
2. Encontrar as k médias de cada grupo.
3. Determinar,para cada cisterna,o grupo mais próximo usando uma medida de dissimilaridade(distância), que, neste trabalho, foi a distância euclidiana.
4. Calcular a média de cada grupo. Se houver mudança na média dos grupos, voltar ao passo 2.
5. Definir as médias dos k grupos, isto é, não realocar nenhuma cisterna a outro grupo.

Segundo Vale (2005), a solução obtida pelo método das k – médias, em geral, depende do ponto de partida, uma vez que o algoritmo encontra um mínimo local, como ocorre em vários problemas de minimização.Esse método é prático e computacionalmente poderoso, porém é sensível a *outliers e* também não é indicado a agrupamentos não convexos. Indica-se a sua utilização em dados contínuos.

3.2.3.3 K – medoid (PAM)

O algoritmo PAM (*Partitioning Around Medoids*), como os demais métodos de agrupamento por particionamento, minimiza uma função custo em relação a um determinado vetor contendo k centróides. No entanto, nesse algoritmo, os referidos centróides são objetos ou cisternas denominados *medoids*. Os *medoids* são objetos representativos de cada agrupamento e contêm as características nas quais a dissimilaridade média dos objetos ou cisternas pertencentes a um dado grupo é mínima (VALE, 2005).

Esse algoritmo é dividido em duas etapas:

1ª Etapa: Construção do algoritmo

Essa etapa é inerente à construção dos objetos candidatos a *medoids*, os quais são construídos através de k seleções de objetos representativos. O primeiro *medoid* é o objeto no qual a soma das dissimilaridades entre todos os objetos é mínima. Os *medoids* seguintes são selecionados, de forma a minimizar a função objetivo o máximo possível. Essa função é dada pela seguinte expressão**:**

$$F = \sum_{i=1}^{n} d(x_i, m(x_i))$$

em que n é o total de cisternas no conjunto de dados, x_i é a i-ésima cisterna do conjunto de dados, $m(x_i)$ é considerado o medoide mais próximo à cisterna x_i, e $d(x_i, m(x_i))$ é a dissimilaridade ou distância entre x_i e $m(x_i)$.

Os seguintes passos devem ser seguidos para encontrar os medoids:

1. Considera-se um objeto x_i que não tenha sido selecionado anteriormente.

2. Considera-se um objeto x_j não selecionado anteriormente e se calcula a diferença entre a sua dissimilaridade em relação ao último objeto selecionado (D_j) com a dissimilaridade do objeto x_i selecionado no passo anterior ($d(x_j, x_i)$).
3. Caso a diferença seja positiva, calcula-se:

   $C_{ij} = \max(\left(D_j - d\left(x_i, x_j\right)\right), 0)$
4. Calcula-se o total por selecionar o objeto xi

   Total = $\sum_{j=1} C_{ij}$
5. Por fim, seleciona-se o objeto x_i que minimize a expressão anterior.

2ª Etapa: Troca de objetos entre medoids

Nesta etapa, busca-se melhorar o conjunto de medoids trocando os objetos ou cisternas entre eles. Assim, se houver uma minimização da função objetivo, mantém-se a troca, caso contrário, ela é desfeita.

O resultado que se busca é medido pela Distância Média Final (DMF) dado pela seguinte expressão:

$$DMF = \frac{1}{n}\sum_{i=1}^{n} d(x_i, m(x_i))$$

em que n é o total de objetos ou cisternas no conjunto de dados, x_i é a i-ésima cisterna do conjunto de dados, $m(x_i)$ é o medoide mais próximo à cisterna x_i e $d(x_i, m(x_i))$ é dissimilaridade entre x_i e $m(x_i)$. O k – medoide, proposto por Vinod (1969), foi implementado no algoritmo PAM.

3.3 Comparação dos Métodos Hierárquicos

3.3.1 Correlação Cofenética

A Correlação Cofenética é a correlação entre as distâncias previstas e as distâncias observadas. A qualidade do agrupamento será melhor quanto mais próximo for de um (1) o coeficiente de Correlação Cofenética. O coeficiente de Correlação Cofenética entre a matriz de distâncias originais e a matriz cofenética proposta por Bussab *et al* (1990) é dado pela expressão

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

em que

$$\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij} \text{ e } \bar{d} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}$$

em que:

c_{ij}: distância entre os objetos ou cisternas i e j, na matriz cofenética;

d_{ij}: distância entre os objetos ou cisternas i e j, na matriz original de distâncias ou matriz de dissimilaridade;

n: dimensão da matriz.

3.4 Validação dos Métodos

3.4.1 Índice da Silhueta

O índice ou estatística da silhueta foi proposto por Rousseeuw (1987), com o intuito de avaliar métodos de particionamento. Nesse caso, cada objeto (cisterna)é representado por um valor s(i) chamado de *silhueta*, que é baseado na comparação da homogeneidade e na “separação” de cada grupo.Assim,para um objeto i, o valor da silhueta é dado por

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad \text{onde } -1 \leq s(i) \leq 1$$

em que:
a(i) é a distância média do objeto i aos objetos do seu grupo;
b(i) é a distância média do objeto i aos objetos dos outros grupos.

Valores negativos de s(i) negativos sugerem que o indivíduo i seja semelhante a indivíduos de outras classes. Valores de s(i) na vizinhança de 1 dão indícios de que i esteja bem classificado.

3.4.2 Índice de Rand

Nesse contexto, considera-se o problema de, dadas duas diferentes propostas de agrupamento de n indivíduos em k grupos, quantificar o grau de similaridade entre esses agrupamentos. Um índice desse tipo será de grande utilidade quando se pretende validar uma classificação de um conjunto de objetos ou cisternas que pertencem a k diferentes grupos. Então, será útil poder quantificar em que medida, a classificação foi capaz de reproduzir essa estrutura ou realizar este agrupamento.

Um dos mais conhecidos índices para comparação de duas classificações ou partições é o Índice de Rand. Existem duas expressões alternativas porém equivalentes para o índice de Rand associado a duas diferentes classificações dos mesmos n objetos ou cisternas em k grupos.

A primeira expressão para o índice de Rand (IR) envolve a consideração de quantos, entre os $\binom{n}{2}$ paresde cisternas diferentes, são classificados de forma análoga nas duas classificações ou partições,ou por que pertencem a um mesmo grupo nas duas classificações, ou porque pertencem a grupos diferentes nas duas classificações. Assim, tem-se:

$$IR = \frac{A + B}{\binom{n}{2}}$$

em que A é o número de pares de observações ou cisternas que pertencem a um mesmo grupo nas duas classificações ou partições,e B é o número de pares de ob-

servações ou cisternas que pertencem, nas duas classificações ou partições, a grupos diferentes.

O índice de Rand toma valores no intervalo 0 ≤ IR ≤1. O valor máximo (IR = 1) corresponderá a uma situação na qual as duas classificações ou partições coincidam, não havendo pares de cisternas que estejam em um mesmo grupo em um caso, e em grupos diferentes no outro (RAND, 1971).

3.4.3 Índices de validação do agrupamento Fuzzy

Existem vários índices de validação que têm sido utilizados com muita intensidade para determinar o número ótimo de grupos (c) em um conjunto de dados(PAL e BEZDEK,1995; THEODORIDIS e KOUTROUBAS, 1999; HALKIDI *et al.*, 2001). Neste estudo são utilizadas três medidas de validação para o agrupamento fuzzy, que são o coeficiente de partição (V_{PC}), a partição entropia (V_{PE}) e o índice de Xie e Beni (V_{XB}), testados para diferentes números de agrupamento c e diferentes valores do parâmetro de fuzificação m, para verificar a sua adequação na obtenção de agrupamentos homogêneos. Esses três índices são discutidos em seguida.

Bezdek (1974), em duas publicações distintas, propôs os índices do coeficiente de partição (V_{PC}) e de partição entropia (V_{PE})para validação de agrupamento fuzzy, que são definidos como:

$$\mathrm{V_{PC}}(\mathbf{U}) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{c}}\sum_{\mathrm{k=1}}^{\mathrm{N}}(\mathrm{u_{ik}})^2$$

A gama de variação do índice V_{PC} é $\left[\frac{1}{c}, 1\right]$, e a partição ótima corresponde ao valor máximo de V_{PC}, o que implica a sobreposição mínima entre os agrupamentos. O valor máximo de V_{PC} é 1,enquanto o mínimo corresponde à quantidade 1/c.

$$\mathrm{V_{PE}}(\mathbf{U}) = -\frac{1}{\mathrm{N}}\left[\sum_{\mathrm{i=1}}^{\mathrm{c}}\sum_{\mathrm{k=1}}^{\mathrm{N}}\mathrm{u_{ik}}\log_{\mathrm{a}}\mathrm{u_{ik}}\right]$$

Na equação acima, o logaritmo tem base $a \in (1, \infty)$, em que o intervalo de variação de VPE é $[0, \log_a(\mathrm{c})]$. O valor mínimo de V_{PE} indica uma boa classificação ou agru-

pamento ótimo correspondendo a uma partição mais realista. Teoricamente, o valor mínimo de VPE é 0 (zero), em que:

U: matriz de partição fuzzy que contém a composição de cada vetor redimensionado em cada grupo;

$\mathbf{u_{ik}}$: denota o grau de adesão do vetor característico de variáveis, redimensionado no agrupamento fuzzy, representado pelo centroide dos respectivos grupos, com $\mathbf{u_{ik}} \in [0, 1]$.

A medida de validação de agrupamento fuzzy proposta por Xie e Beni (1991) é a proporção de compactação de uma c-partição fuzzy. Pode ser considerada como sendo uma função do conjunto de dados e o centroide dos grupos.

$$V_{XB}(\mathbf{U}, \mathbf{V}: \mathbf{X}) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^2 \|v_i - v_k\|^2}{N \min_{i \neq k} \|v_i - v_k\|^2}$$

em que o termo do numerador da equação acima é a soma dos quadrados dos desvios fuzzy de cada vetor característico $\mathbf{X_k}$ (k = 1,..., N) para o centróide fuzzy de cada grupo $\mathbf{v_i}$(i = 1,...,c). A magnitude do termo diminui à medida em que os agrupamentos ficam mais compactos. O denominador que, por sua vez, mede a distância mínima entre os centróides dos grupos de cisternas tem valor mais elevado para grupos que estão bem separados. O valor mínimo de V_{XB} sugere uma boa partição ou agrupamento ótimo, o que corresponde a grupos de cisternas compactos e bem separados. Temos ainda que **X** é a matriz de dados, N é o número de vetores de características ou variáveis de qualidade da água das cisternas e $\mathbf{V = (v_1,...,v_c)}$ representa um c-tripla de protótipos $\mathbf{v_i}$, cada qual caracterizando o baricentro de um dos grupos c.

3.5 Modelagem de Séries Temporais e formação de Agrupamentos

A metodologia de séries temporais proposta por Box-Jenkins é utilizada para estudar o comportamento das séries temporais das variáveis de qualidade da água analisadas neste estudo. Os modelos ARIMA são utilizados na modelagem de séries que apresentam comportamento seguindo um processo ruído branco, com média zero e variância constante. Dessa forma, a metodologia proposta por Box *et al.* (1994) pode ser empregada para avaliar e estudar o comportamento das séries tem-

porais das variáveis de qualidade da água das cisternas de placas, ou seja, se essas variáveis apresentam sazonalidade, tendência, ou se são estacionárias, ainda se apresentam estrutura de autocorrelação.

As variáveis coliformes fecais e totais bem como os sólidos dissolvidos totais (S.D.T.) também apresentam variações em cada período de coleta, conforme o índice pluviométrico e a temperatura. Já a variável pH, na maioria dos estudos relacionados à qualidade da água, são invariantes em relação à série temporal de índice pluviométrico.

Verificou-se que as séries temporais das variáveis de qualidade da água das cisternas de placas analisadas no período de novembro de 2014 a março de 2016 podem variar a cada período de coleta, isto é, os dados obtidos nas campanhas de monitoração podem apresentar grandes variações ou pequenas flutuações quando são comparadas em cada período de medição. Por exemplo, os cátions e os ânions podem ser mais elevados no período de estiagem, quando a diluição da água da chuva tem seu efeito menos acentuado.

Dessa forma, a depender do período que a coleta dos dados for feita na série temporal, podem-se formar agrupamentos bem distintos. Os agrupamentos gerados a partir de uma amostra realizada no mês de março, por exemplo, podem ser bem diferente dos agrupamentos gerados a partir de uma amostra realizada em outubro. Segundo Oliveira *et al.* (2015), as características da série temporal que causam essa diferença nos agrupamentos é a sazonalidade, a tendência, assim como a variabilidade ou estacionaridade que as séries temporais das variáveis de qualidade da água apresentam a cada período.

4 – RESULTADOS E DISCUSÕES

Com os dados originais referentes às variáveis estudadas, realizou-se uma padronização, visando a uma escala homogênea entre as variáveis, posteriormente obteve-se a matriz de distância (euclidiana). A seguir, têm-se os resultados obtidos a partir das amostras coletadas nas cisternas de placas nos assentamentos Catolé, Poço do Serrote, Poldrinho e Três Irmãos, localizados no município de Serra Talhada, região do Pajeú do Estado de Pernambuco.

4.1 Análise de Agrupamento e medidas de Similaridade

Foram apresentados os métodos da média das distâncias, da ligação simples, da ligação completa, do Centroide, de Ward, da mediana e de mcquitty e obtidos os respectivos dendogramas.

4.1.1 Variáveis analisadas

Na Figura 3,observam-se os *boxplots* dos dados originais (a) e dos dados padronizados (b)das variáveis de qualidade da água em estudo que foi coletada nas cisternas de placas localizadas nos assentamentos.Sendo necessário uma padronização das variáveis utilizando a distribuição normal padrão, objetivou-se corrigir o problema de escala existente nas variáveis estudadas a saber: cátions, ânions, pH, Sólidos dissolvidos totais (S.D.T.), coliformes fecais e coliformes totais.Já Docampo *et al.* (2013) binarizaram as variáveis de dados clínicos para correção de escala visando a formar grupos mais homogêneos.A normalidade dos dados não foi considerada, uma vez que o objetivo é meramente obter agrupamentos. Esta normalidade pode ser considerada caso seja necessário obter os estimadores dos parâmetros utilizados neste estudo.

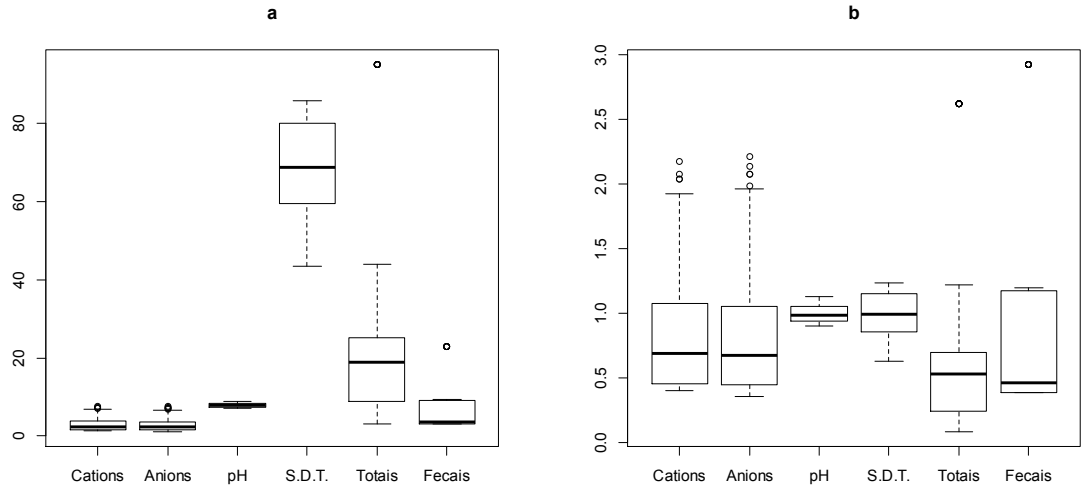

Figura 3. *Boxplots* das variáveis dos dados originais (a) e dos dados padronizados (b)

Observou-se, na Figura 3 (b), que as variáveis cátions e ânions apresentaram a maior variabilidade com relação à água das cisternas, além de se observar a existência de três *outliers* nos *boxplots* que representam os cátions e os ânions, mostrando o excesso desses dois minerais em algumas cisternas. Já o pH é menos disperso, ou seja, possui a menor variância, sugerindo que o pH é mais homogêneo nas cisternas observadas.

Os Sólidos dissolvidos totais (S.D.T.) e os coliformes totais apresentam comportamentos semelhantes (Figura 3), uma vez que suas medianas nos *boxplots,* que representam as duas variáveis, estão próximas do segundo quartil apresentando variabilidades similares. Isso indica que nas cisternas há quantidades próximas de sólidos dissolvidos totais e coliformes fecais. Observa-se ainda a existência de um *outlier* no *boxplot* que representa os coliformes totais.

Os coliformes fecais apresentam um comportamento diferenciado, conforme observado no seu *boxplot* Figura 3, assim como verificado por Dovoedo e Chakraborti(2015), em relação às outras cinco variáveis, uma vez que há mediana na vizinhança do primeiro quartil denotando assimetria nos dados relacionados a essa variável. Ainda se verificou um *outlier*, o que dá fortes indícios de existirem cisternas com medidas elevadas de coliformes fecais.

Observou-se, na Figura 4, o Q-Q plot, um *outlier* que se confirmou pelo *Boxplot* dos cátions, dos ânions, dos coliformes fecais e totais da Figura 3, na qual 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 4, o que não se verificou entre outras variáveis que tiveram relação dispersa.

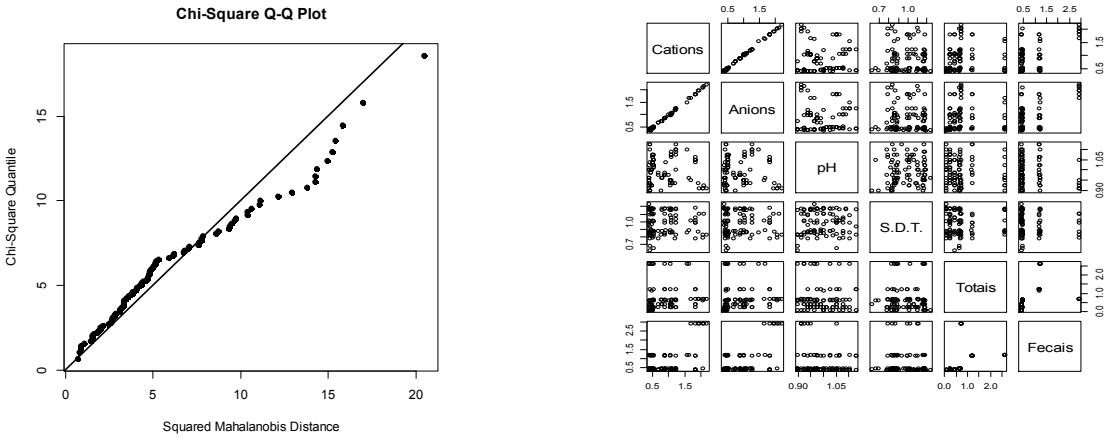

Figura 4. *Q-Q plot* e *Scatterplot* das variáveis cátions, ânions, pH, sólidos dissolvidos totais, coliformes fecais e totais.

A partir da matriz de dissimilaridade, foram utilizados os seguintes métodos hierárquicos aglomerativos, isto é,o método da média das distâncias, o método do vizinho do próximo, o método do vizinho mais distante, o método do centroide, o método de Ward, o método da mediana e o método de mcquitty.Com base nas correlações cofenéticas dos agrupamentos hierárquicos apresentados na Tabela 2, verificou-se que o método das médias das distâncias proporcionou o melhor agrupamento das cisternas, apresentando a maior correlação cofenética0,9562.

Tabela 2. Correlações cofenéticas dos agrupamentos hierárquicos

Método Hierárquico	Correlação Cofenética
Vizinho mais próximo	0,9381
Vizinho mais distante	0,9514
Média das distâncias	0,9562
Centroide	0,9506
Ward	0,8320
Mediana	0,9388
Mcquitty	0,9531

4.2 Métodos Hierárquicos

Os métodos hierárquicos de agrupamento foram utilizados na formação dos grupos das cisternas de placas, segundo as variáveis de qualidade da água contidas nas cisternas de placas da região do Pajeú – PE. Os métodos hierárquicos dividem-se em aglomerativos e divisivos.

4.2.1 Normalidade da distância euclidiana entre as cisternas

As observações extremas ou *outliers* podem distorcer a estrutura dos agrupamentos, logo, faz-se necessário uma inspeção em busca dessas observações, e, para essa finalidade, utilizou-se o *Q-Q plot* e o *boxplot* da matriz de distância euclidiana, assim como proposto por Baxter(1999).

Observa-se, na Figura 5,que o *boxplot* e o *Q-Q plot* da matriz de distância euclidiana das 100 cisternas apresentam falta de normalidade, isto é, a distância euclidiana entre as observações (cisternas) não tem distribuição normal.

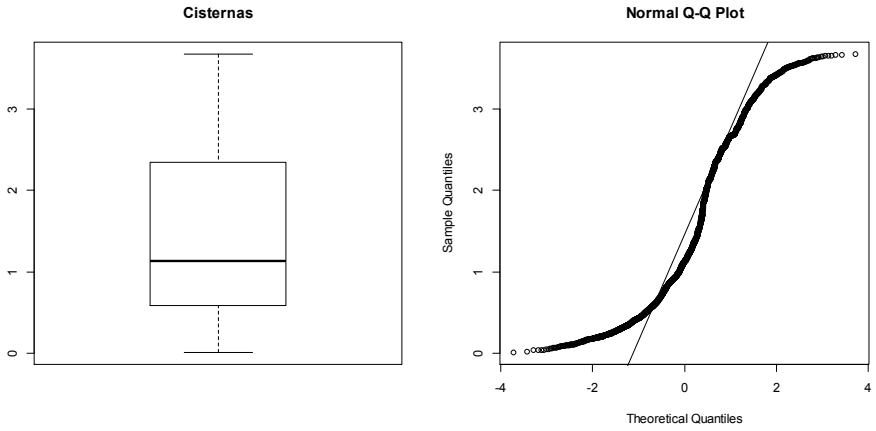

Figura 5. *Boxplot* e *Q-Q plot* da distância euclidiana das cisternas para detectar *outlier*.

Assim, uma padronização das cisternas relacionadas às variáveis em estudo é recomendada para a correção de escala, como observado em Cao *et al.* (1999). No entanto, aqui nesta seção da tese, não foi necessária a padronização, pois nosso intuito é meramente obter grupos.

4.2.2 Dendogramas obtidos para os Métodos Hierárquicos

A Figura 6 mostra o dendograma para as cisternas. Observa-se que as cisternas mais próximas ou similares entre si, de acordo com as variáveis estudadas, são as que estão no grupo 1, que, por sua vez, é o que possui o maior número de observação (cisternas), tendo como particularidade as características homogêneas com relação às seis variáveis mensuradas.

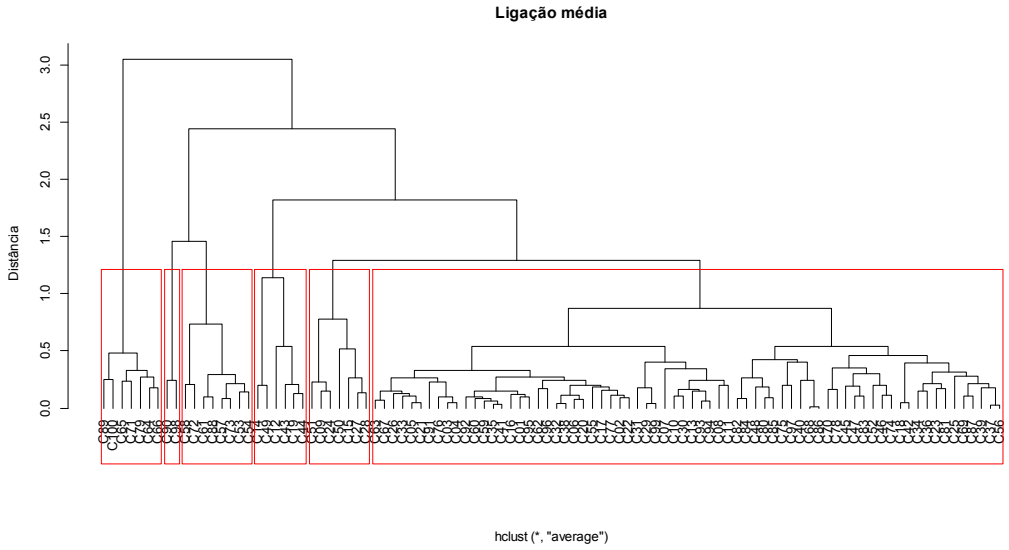

Figura 6. Dendograma resultante do método da média das distâncias

O dendograma apresentado aqui resultou em uma correlação cofenética de 0,9562, mostrando uma boa qualidade no agrupamento. A distância que foi utilizada na construção desse dendograma foi a distância euclidiana, pois é a distância mais real do ponto vista prático. O método utilizado na construção do dendograma foi o da ligação média, em conformidade com Borysov *et al.* (2014), e o que obteve dentre todos os métodos a maior correlação cofenética.

Pelo método da média das distâncias, foram formados seis grupos (Tabela 3). Observou-se que o grupo 1 é considerado com um número grande de observações em função do número de cisternas analisadas. Observa-se também que, entre esses grupos, dois deles têm o mesmo número de observações, que são os grupos 2 e 5 com sete cisternas cada um, enquanto o grupo 3 tem seis cisternas, o grupo 4 tem oito cisternas, e o grupo 6 é formado por apenas duas cisternas, podendo ser um grupo discrepante ou *outlier* merecendo uma atenção especial do pesquisador. As cisternas (C90 e C98) alocadas no grupo 6 podem estar com a qualidade da água

comprometida em relação a alguma, ou mesmo, a todas as variáveis analisadas, podendo causar danos à saúde das pessoas que as usam.

Espera-se que as cisternas do 1 tenham características similares em relação à qualidade da água, de acordo com as variáveis estudadas, e ainda que essa qualidade esteja de acordo com o que preconiza o Ministério da Saúde e o CONAMA para águas potáveis destinadas ao abastecimento humano.

Tabela 3. Grupos de cisternas obtidos por meio do método da média das distâncias

Grupos	n	Cisternas
Grupo 1	70	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C18 C20 C21 C22 C23 C25 C26 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C45 C46 C47 C48 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	7	C9 C15 C24 C27 C28 C50 C51
Grupo 3	6	C12 C14 C19 C43 C44 C49
Grupo 4	8	C53 C54 C57 C58 C61 C72 C73 C88
Grupo 5	7	C64 C65 C66 C71 C79 C89 C100
Grupo 6	2	C90 C98

Observa-se, na Figura 7, que a estatística da silhueta indica a consistência dos grupos, o que se evidencia nos grupos 4, 5 e 6 com uma boa homogeneidade, tendo estatísticas da silhueta 0,64, 0,84 e 0,83, respectivamente, indicando que as cisternas foram bem classificadas nesses três grupos. Os grupos 1, 2 e 3 apresentaram estatística da silhueta de 0,49, 0,54 e 0,52, respectivamente, que apesar de não ter valores elevados, como para os outros três grupos, ainda apresenta indícios de uma boa classificação das cisternas nos grupos para os quais elas foram alocadas. De maneira geral, de acordo com o gráfico da silhueta e a estatística da silhueta média (Figura 7), que é de 0,54, as cisternas estão bem classificadas em todos os grupos, não sendo necessário realocar observações,do mesmo modo que encontrado por Lin *et al.* (2013).

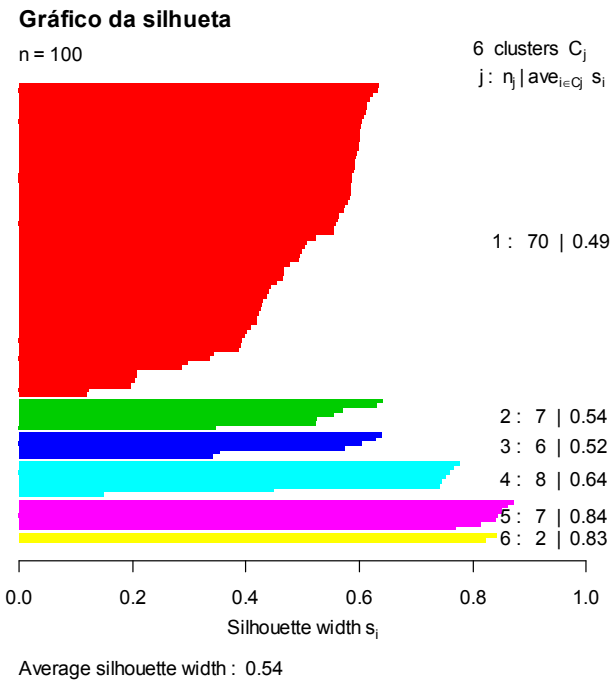

Figura 7. Gráfico e estatística da silhueta obtidos pelo método da média das distâncias

De acordo com o dendograma apresentado na Figura 8, obtido pelo método da ligação simples, como encontrado em Borysov *et at.* (2014), observa-se a presença de seis grupos. O método da ligação simples forma os grupos 1, 2, 3, 4, 5 e 6, Tabela 4, com as respectivas cisternas.

Tabela 4. Grupos de cisternas obtidos por meio do método da ligação simples

Grupos	n	Cisternas
Grupo 1	70	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C18 C20 C21 C22 C23 C25 C26 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C45 C46 C47 C48 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	7	C9 C15 C24 C27 C28 C50 C51
Grupo 3	4	C12 C19 C43 C44
Grupo 4	2	C14 C49
Grupo 5	10	C53 C54 C57 C58 C61 C72 C73 C88 C90 C98
Grupo 6	7	C64 C65 C66 C71 C79 C89 C100

O grupo 1, foi similar ao método da média das distâncias, ou seja, todas essas cisternas são similares segundo as variáveis em estudo, levando a indícios de que a qualidade da água de todos essas cisternas é homogênea. Nos grupos 2 e 6, foi a-

locada a mesma quantidade de cisternas com sete unidades cada um, os grupos 3 e 5 apresentaram quatro e dez cisternas, respectivamente, e o grupo 4 foi formado apenas por duas unidades, que foram as cisternas 14 e 49. O grupo 4, por ter classificado apenas duas cisternas, deve ser observado com uma atenção especial, uma vez que pode se tratar de um *outlier,*e as águas dessas duas cisternas podem estar com a qualidade imprópria para o uso doméstico.

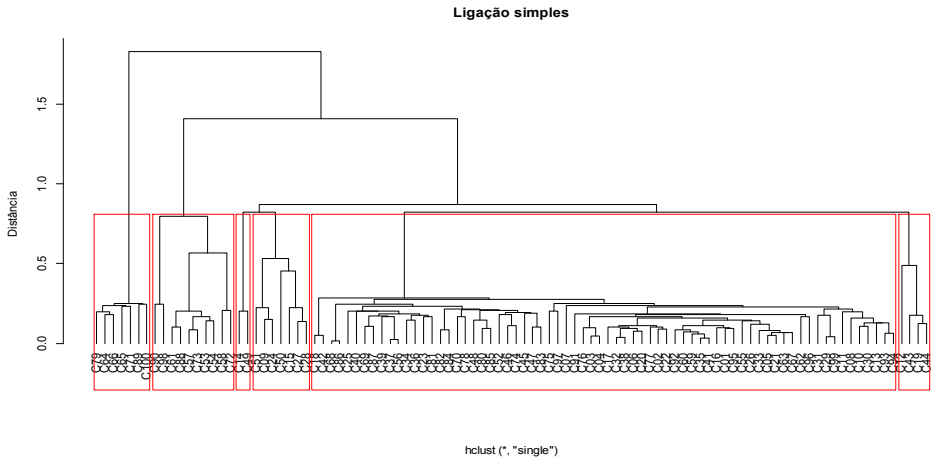

Figura 8. Dendograma resultante do método da ligação simples

O gráfico e as estatísticas da silhueta, como calculadas por Arbelaitz *et al.* (2013), para o método da ligação simples, são apresentados na Figura 9. Observa-se que os grupos 3, 4 e 6 classificaram bem as cisternas, pois apresentaram estatísticas da silhueta 0,68, 0,82 e 0,80, respectivamente, indicando homogeneidade dentro desses três grupos, enquanto os grupos 1, 2 e 5 obtiveram estatísticas da silhueta 0,47, 0,49 e 0,47, respectivamente, que, mesmo sendo inferiores aos outros três grupos, ainda garantem uma boa classificação das cisternas neles. Isto é, os grupos 1, 2 e 5, de acordo com a estatística da silhueta, devem ter cisternas com características similares, até mesmo ao grupo 1. Portanto, o método da ligação simples obteve uma boa classificação das unidades (cisternas) em cada um dos seis grupos por ele formado, por ter apresentado valores “elevados” para as estatísticas da silhueta de todos os grupos e estatística da silhueta média de 0,51 (Figura 9).

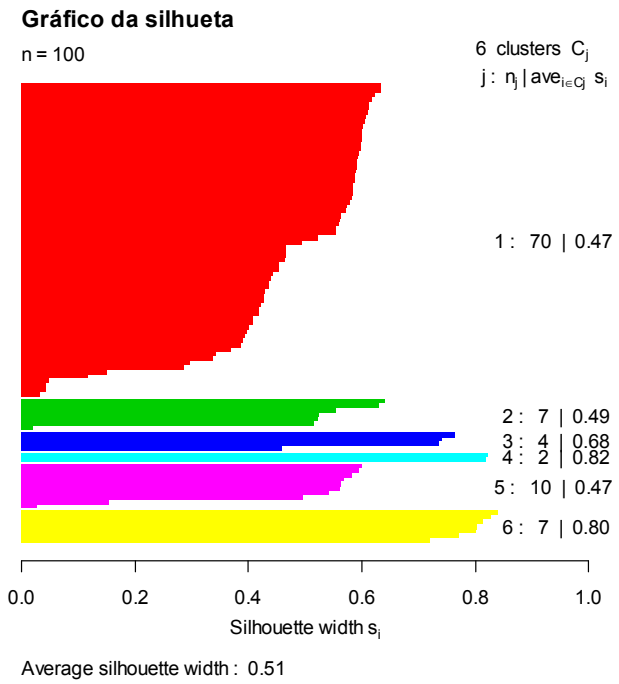

Figura 9. Gráfico e estatística da silhueta obtidos pelo método da ligação simples

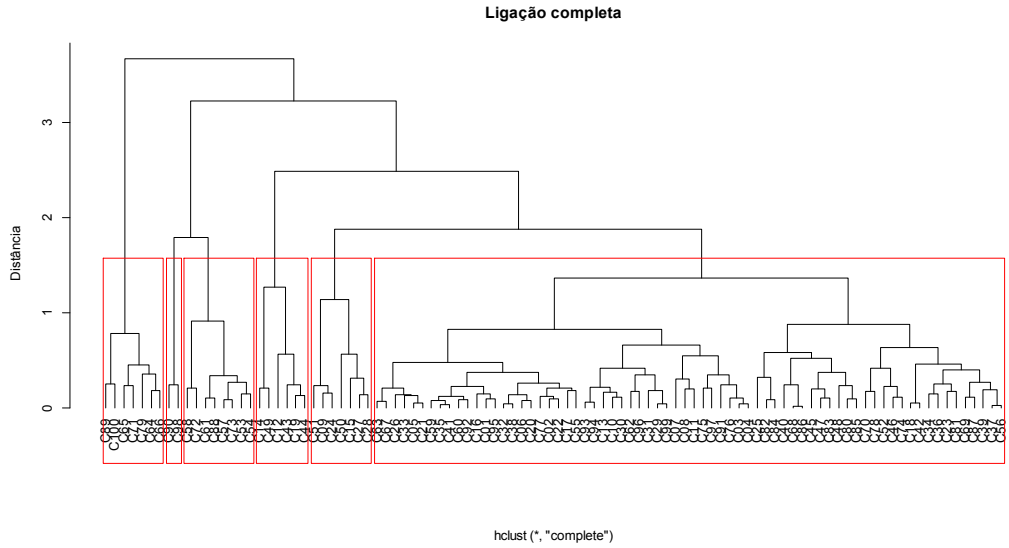

Figura 10. Dendograma resultante do método da ligação completa

Numa observação visual feita pelo Dendograma apresentado na Figura 10, percebe-se que o método da ligação completa ou do vizinho mais distante indica a presença de seis grupos com número distinto de cisternas (Tabela 5) em cada grupo.

Tabela 5. Grupos de cisternas obtidos por meio do método da ligação completa

Grupos	n	Cisternas
Grupo 1	70	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C18 C20 C21 C22 C23 C25 C26 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C45 C46 C47 C48 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	7	C9 C15 C24 C27 C28 C50 C51
Grupo 3	6	C12 C14 C19 C43 C44 C49
Grupo 4	8	C53 C54 C57 C58 C61 C72 C73 C88
Grupo 5	7	C64 C65 C66 C71 C79 C89 C100
Grupo 6	2	C90 C98

O grupo 1 é formado como nos métodos da ligação completa e da ligação média, ou seja, de acordo com o método da ligação completa, todas as cisternas alocadas nesse grupo têm características similares de acordo com as variáveis estudadas. Os grupos 2 e 5, por sua vez, classificaram sete cisternas cada um, o que corresponde a 10% do grupo 1; os grupos 3 e 4 alocaram seis e oito cisternas respectivamente; enquanto o grupo 6 obteve apenas duas cisternas (C90 e C98), como obtido pelo método da média das distâncias, podendo este grupo ser um *outlier*. Isso mais uma vez alerta que o pesquisador deve ter uma atenção especial com essas duas cisternas, pois elas poderiam estar com a qualidade da água imprópria para uso doméstico, assim como alertou Fan *et al.* (2013).

O gráfico com as respectivas estatísticas da silhueta para o método da ligação completa estão apresentados na Figura 11, em que as estatísticas da silhueta para cada um dos grupos são apresentadas individualmente, além da estatística média da silhueta. Observa-se que os grupos 4, 5 e 6 têm ótimas estatísticas da silhueta com valores de 0,64, 0,84 e 0,83, respectivamente, ou seja, esses grupos alocaram de maneira adequada as cisternas que neles estão, assim essas cisternas têm probabilidade baixa de serem realocadas em outros grupos. Os grupos 1, 2 e 3, por sua vez, obtiveram valores da estatística da silhueta inferiores aos outros três grupos, mas o suficiente para garantir uma boa classificação, de acordo com esse critério de validação, pois seus valores foram 0,49, 0,54 e 0,52, respectivamente, para os grupos 1, 2 e 3, observados na Figura 11, em um intervalo de -1 a 1 que esta estatística pode assumir. A estatística da silhueta média foi de 0,54, garantindo assim uma boa classificação para os grupos obtidos pelo método da ligação completa ou do vizinho mais distante.

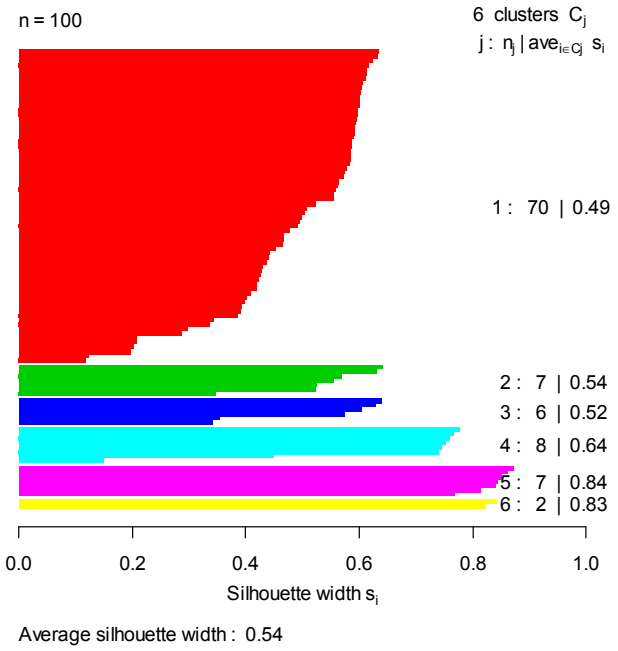

Figura 11. Gráfico e estatística da silhueta obtidos pelo método da ligação completa

O Dendograma apresentado na Figura 12, obtido pelo método do centroide,apresenta seis grupos que tem número distinto de cisternas,como observado na Tabela 6, com as respectivas cisternas alocadas em cada grupo.

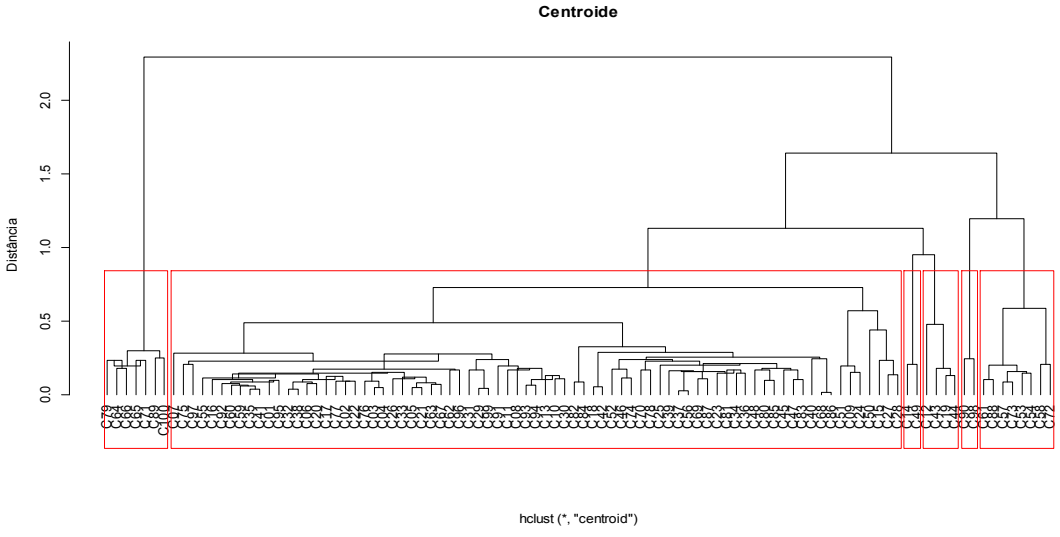

Figura 12. Dendograma resultante do método do centroide

Tabela 6. Grupos de cisternas obtidos por meio do método do centroide

Grupos	n	Cisternas
Grupo 1	77	C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C13 C15 C16 C17 C18 C20 C21 C22 C23 C24 C25 C26 C27 C28 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C45 C46 C47 C48 C50 C51 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	4	C12 C19 C43 C44
Grupo 3	2	C14 C49
Grupo 4	8	C53 C54 C57 C58 C61 C72 C73 C88
Grupo 5	7	C64 C65 C66 C71 C79 C89 C100
Grupo 6	2	C90 C98

Observa-se que o grupo 1 alocou setenta e sete cisternas de acordo com o método do centroide, diferindo das outras técnicas de agrupamento apresentadas anteriormente. Então, de acordo com o método do centroide, 77% das cisternas têm características similares,conforme as variáveis estudadas, corroborando com o que afirma You e Seo (2009), o que é um percentual bastante significativo perante a amostra analisada. Outro diferencial observado nesse método de agrupamento é que os grupos 3 e 6 foram formados apenas por duas observações cada um deles, com as cisternas (C14 e C49) e (C90 e C98), respectivamente, o que não havia ocorrido em outros métodos de agrupamento. Os grupos 2, 4 e 5 foram formados de maneira similar a outros métodos de agrupamento, com quatro, oito e sete cisternas, respectivamente. Nessas condições, o que merece maior atenção são os grupos 3 e 6, em especial, o grupo 6, que já foi formado em outros métodos de agrupamento pelas mesmas cisternas.

O gráfico da silhueta com as respectivas estatísticas da silhueta estão apresentados na Figura 13, para o método do centroide,pelo qual se observa que os grupos 3, 5 e 6 têm ótimas estatísticas da silhueta, com respectivos valores de 0,82, 0,80 e 0,83, sugerindo que as cisternas alocadas nesses três grupos possivelmente não serão realocadas para outros grupos. Os grupos 1, 2 e 4 tiveram estatísticas da silhueta, respectivamente, 0,51, 0,68 e 0,65, indicando boas classificações das cisternas nesses três grupos, pois é superior a 0,50, mesmo tendo estatísticas da silhueta inferiores aos outros três grupos. Ou seja, essas cisternas alocadas nestes grupos não serão realocadas a outros grupos. A estatística média da silhueta de 0,56 confirma que o método do centroide fez uma boa classificação das cisternas

nos seus respectivos grupos, sugerindo que as cisternas que estão em um determinado grupo têm baixa probabilidade de serem realocadas em outro grupo.

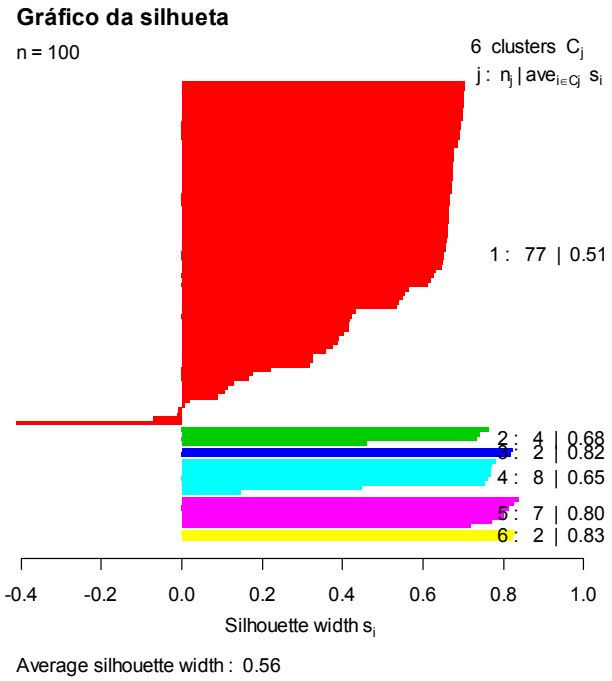

Figura 13. Gráfico e estatística da silhueta obtidos pelo método do centroide

Em uma observação visual do dendograma feita com base na Figura 14, observa-se a existência de seis grupos nessa figura que foi obtida pelo método de Ward. Na Tabela 7, estão apresentados os grupos 1, 2, 3, 4, 5 e 6 com as respectivas cisternas alocadas em cada um desses grupos.

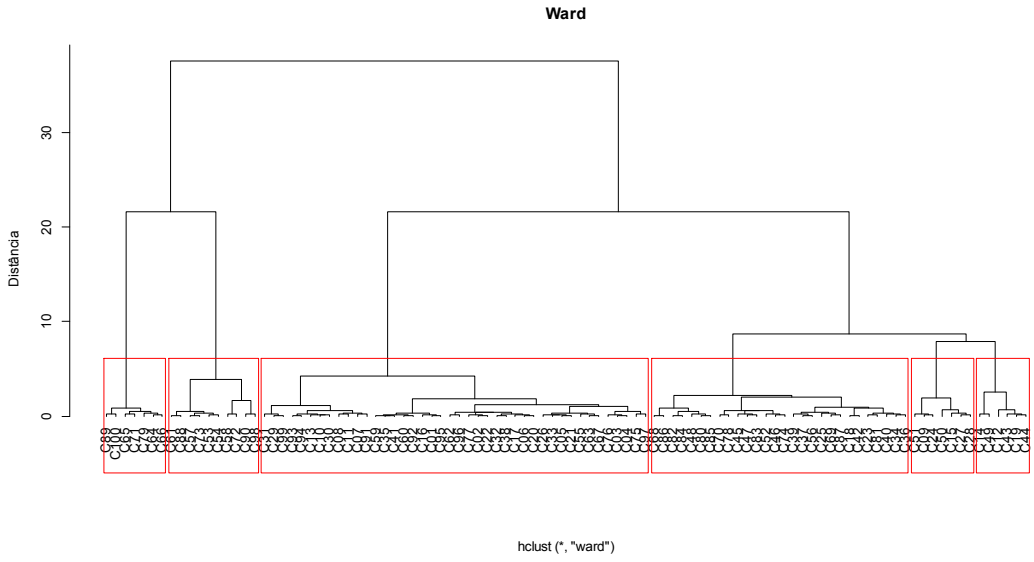

Figura 14. Dendograma resultante do método de Ward

Tabela 7. Grupos de cisternas obtidos por meio do método do Ward

Grupos	n	Cisternas
Grupo 1	42	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C20 C21 C22 C26 C29 C30 C31 C32 C33 C35 C38 C41 C55 C59 C60 C62 C63 C67 C75 C76 C77 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	7	C9 C15 C24 C27 C28 C50 C51
Grupo 3	6	C12 C14 C19 C43 C44 C49
Grupo 4	28	C18 C23 C25 C34 C36 C37 C39 C40 C42 C45 C46 C47 C48 C52 C56 C68 C69 C70 C74 C78 C80 C81 C82 C83 C84 C85 C86 C87
Grupo 5	10	C53 C54 C57 C58 C61 C72 C73 C88 C90 C98
Grupo 6	7	C53 C54 C57 C58 C61 C72 C73 C88 C90 C98

O método de Ward apresentou algumas diferenças em relação aos outros métodos estudados até então. Uma delas é que o grupo 1 agrupou quarenta e duas cisternas, ou seja, um número bem inferior em relação ao mesmo grupo obtido pelos outros métodos de agrupamento. Observou-se também que o grupo 4 ficou com vinte e oito cisternas, o que também diferiu dos outros métodos de agrupamento, uma vez que em nenhum deles um segundo grupo alocou um número maior de observações. Uma outra observação é que o método de Ward não formou grupos com duas cisternas, como nos outros métodos de agrupamento. Os grupos 2 e 6 foram formados por sete cisternas cada, enquanto os grupos 3 e 5 classificaram seis e dez cisternas respectivamente. De maneira geral, o método de Ward classificou as cisternas nos seis grupos de maneira mais parcimoniosa, corroborando com Borysov *et al.*(2014), isto é, não houve grupos com um número elevado de cisternas, tampouco grupos com poucas observações, ou seja, o método de Ward obteve um melhor resultado, comparado aos demais métodos de agrupamento.

O método de Ward apresentou características peculiares em relação aos outros métodos utilizados para esta análise, como bem destacadas anteriormente. Essas características também foram notáveis no Gráfico e nas estatísticas da silhueta observadas na Figura 15, pois o único grupo que obteve estatística da silhueta ótima com valor de 0,84 foi o grupo 6, enquanto os grupos 1, 2, 3, 4 e 5 obtiveram estatísticas da silhueta 0,55, 0,50, 0,43, 0,48 e 0,51, respectivamente, dando indícios de boas classificações das cisternas em seus respectivos grupos, mesmo com estatísticas da silhueta inferiores ao grupo 6. A estatística média da silhueta obtida pelo método de Ward foi de 0,54, corroborando com Clifford *et al.* (2011), valor que não cau-

sa suspeita de que este método não classificou adequadamente as cisternas nos respectivos grupos.

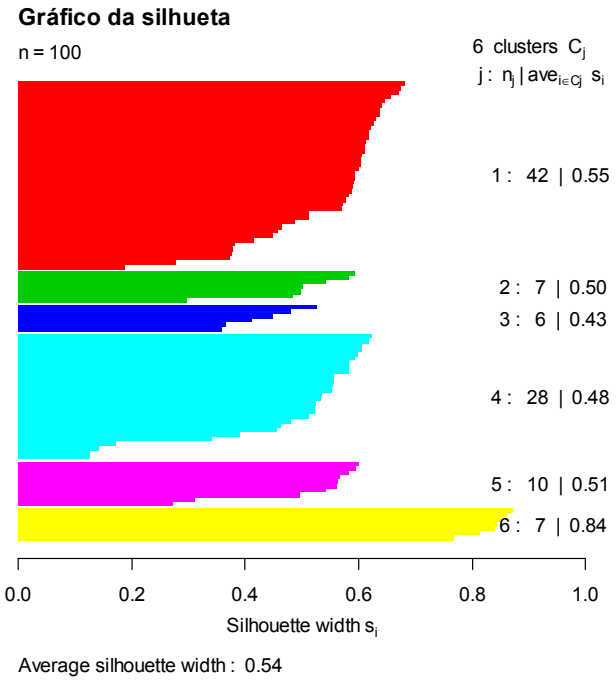

Figura 15. Gráfico e estatística da silhueta obtidos pelo método de Ward

No Dendograma da Figura 16, observa-se a existência de seis grupos que foram obtidos pelo método da mediana.

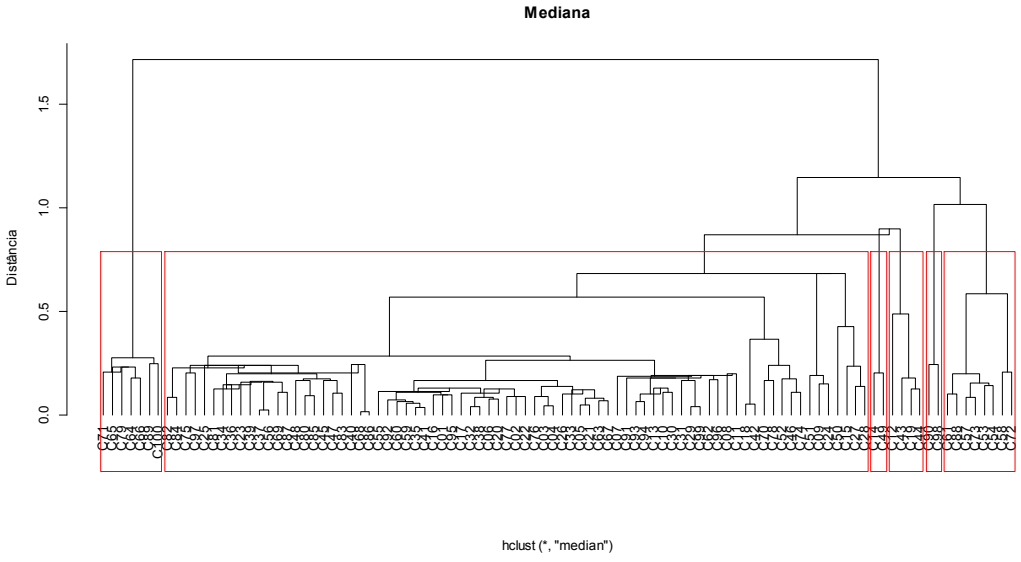

Figura 16. Dendograma resultante do método da mediana

A Tabela 8 apresenta os grupos 1, 2, 3, 4, 5 e 6 com suas respectivas observações ou cisternas alocadas em cada grupo pelo método da mediana.

Tabela 8. Grupos de cisternas obtidos por meio do método da mediana

Grupos	n	Cisternas
Grupo 1	77	C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C13 C15 C16 C17 C18 C20 C21 C22 C23 C24 C25 C26 C27 C28 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C45 C46 C47 C48 C50 C51 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	4	C12 C19 C43 C44
Grupo 3	2	C14 C49
Grupo 4	8	C53 C54 C57 C58 C61 C72 C73 C88
Grupo 5	7	C64 C65 C66 C71 C79 C89 C100
Grupo 6	2	C90 C98

Observa-se que o grupo 1 alocou setenta e sete cisternas, de acordo com o método da mediana. Em outras palavras, segundo o método da mediana, 77% das cisternas têm características similares de acordo com as variáveis estudadas, o que é um percentual muito considerável perante a amostra analisada. Outro diferencial observado nesse método de agrupamento é que os grupos 3 e 6 alocaram apenas duas cisternas em cada um deles, com as cisternas C14, C49 e C90, C98, respectivamente, o que não havia ocorrido em outros métodos de agrupamento. Os grupos 2, 4 e 5 foram formados de maneira similar a outros métodos de agrupamento, com quatro, oito e sete cisternas, respectivamente. Nessas condições, o que merece maior atenção são os grupos 3 e 6, em especial o grupo 6 que já foi formado em outros métodos de agrupamento pelas mesma cisternas.

O método da mediana apresentou características semelhantes ao método do centroide. Essas características também foram notáveis no gráfico e nas estatísticas da silhueta observadas na Figura 17, na qual os grupos 3, 5 e 6obtiveram estatísticas da silhueta ótima com valores de 0,82, 0,80 e 0,83, respectivamente, enquanto os grupos 1, 2 e 4 obtiveram estatísticas da silhueta 0,51, 0,68 e 0,65, respectivamente,dando indícios de boas classificações das cisternas em seus respectivos grupos, mesmo com estatísticas da silhueta inferiores aos grupos 3, 5 e 6. A estatística média da silhueta obtida pelo método da mediana foi de 0,56, valor que não dá indícios de que esse método classificou incorretamente as cisternas nos respectivos grupos.

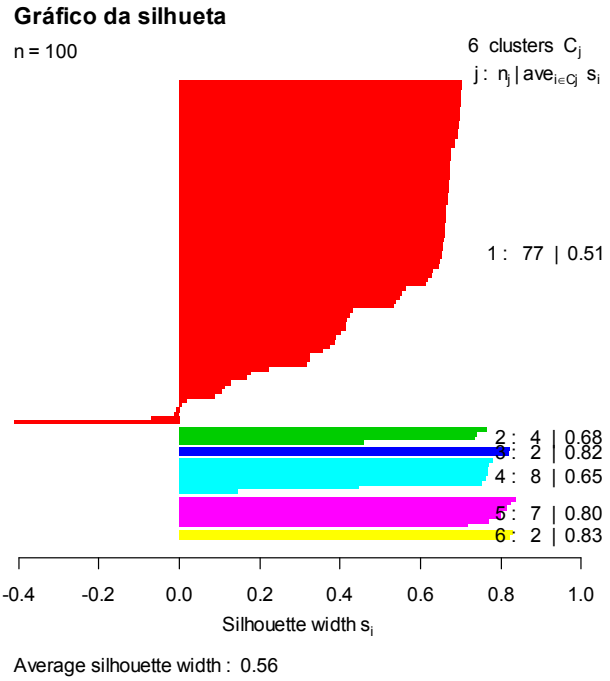

Figura 17. Gráfico e estatística da silhueta obtidos pelo método da mediana

O Dendograma obtido pelo método de mcquitty está apresentado na Figura 18.Observa-se a existência de seis grupos nessa figura que foram obtidos pelo método de mcquitty.

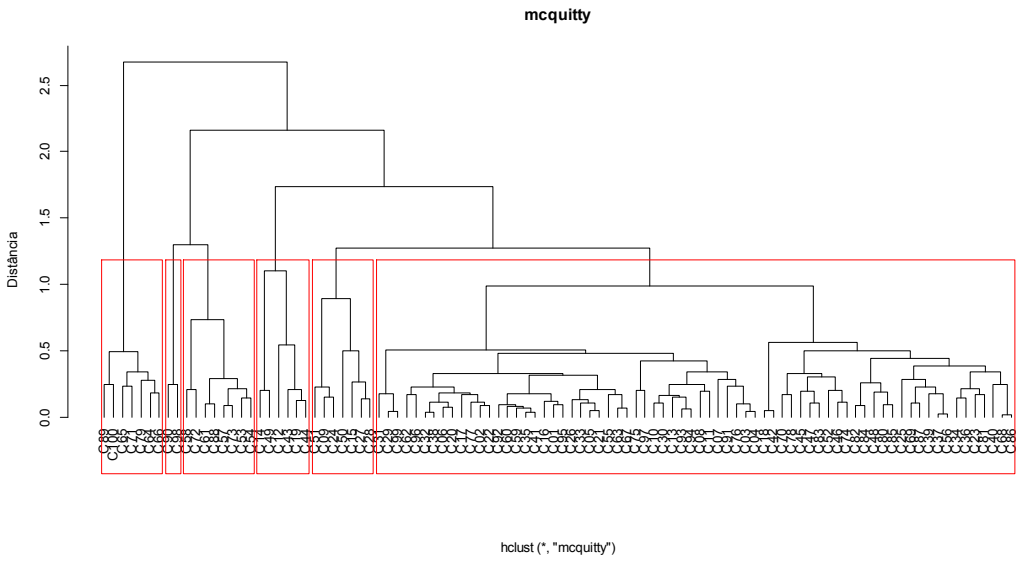

Figura 18. Dendograma resultante do método de mcquitty

Na Tabela 9, estão apresentados os grupos 1, 2, 3, 4, 5 e 6, e cada um desses grupos têm as respectivas cisternas alocadas pelo método.

Tabela 9. Grupos de cisternas obtidos por meio do método de mcquitty

Grupos	n	Cisternas
Grupo 1	70	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C18 C20 C21 C22 C23 C25 C26 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C45 C46 C47 C48 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	7	C9 C15 C24 C27 C28 C50 C51
Grupo 3	6	C12 C14 C19 C43 C44 C49
Grupo 4	8	C53 C54 C57 C58 C61 C72 C73 C88
Grupo 5	7	C64 C65 C66 C71 C79 C89 C100
Grupo 6	2	C90 C98

Observa-se que grupo 1 é constituído por setenta cisternas, como nos métodos da média das distâncias, da ligação completa e da ligação simples, ou seja, de acordo com o método de mcquitty, todas essas cisternas alocadas nesse grupo têm características similares,conforme as variáveis mensuradas. Os grupos 2 e 5, por sua vez, classificaram 10% das cisternas no grupo 1, ou seja, sete cisternas cada um; os grupos 3 e 4 alocaram seis e oito cisternas, respectivamente.O grupo 6 obteve apenas duas cisternas (C90 e C98), como obtido pelos métodos anteriores, podendo este grupo ser um *outlier*, pois há uma forte evidência de que essas duas cisternas têm características distintas das demais, e, desse modo, elas poderiam estar com a qualidade da água imprópria para uso doméstico.

O Gráfico com as respectivas estatísticas da silhueta para o método de mcquitty está apresentado na Figura 19, em que as estatísticas da silhueta para cada um dos grupos são apresentadas individualmente, além da estatística média da silhueta. Observa-se que os grupos 4, 5 e 6 têm ótimas estatísticas da silhueta com valores de 0,64, 0,84 e 0,83, respectivamente, isto é, esses grupos alocaram de maneira adequada as cisternas que neles estão, assim essas cisternas não devem ser realocadas em outros grupos pelo algoritmo de agrupamento. Os grupos 1, 2 e 3, por sua vez, obtiveram valores da estatística da silhueta menores que os outros três grupos, mas o suficiente para garantir uma boa classificação, de acordo com este critério de validação, pois seus valores foram 0,49, 0,54 e 0,52, respectivamente, para os grupos 1, 2 e 3 (Figura 19), em um intervalo de -1 a 1 que essa estatística pode assumir. A estatística da silhueta média foi de 0,54,número que é considerado significativo para uma boa classificação das cisternas nos grupos obtidos pelo método de mcquitty.

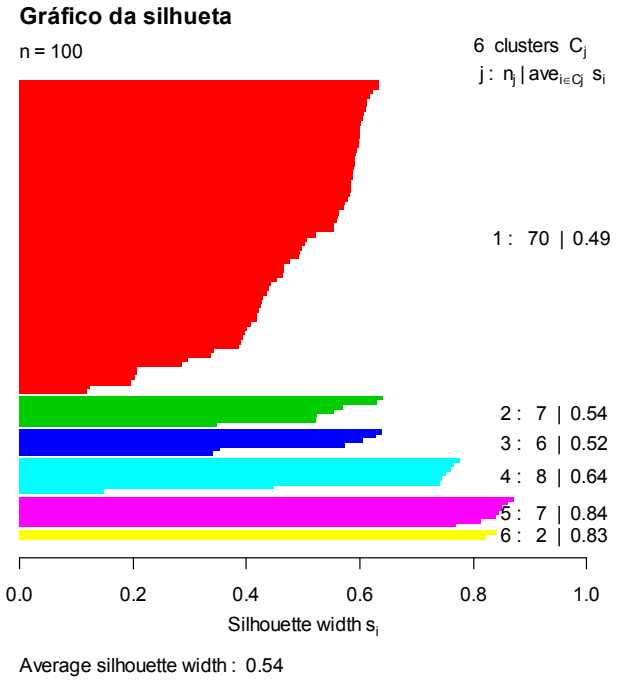

Figura 19. Gráfico e estatística da silhueta obtidos pelo método de mcquitty

Os Dendogramas obtidos pelos algoritmos de agrupamento ora apresentados têm aspectos diferentes, no entanto, poucas cisternas foram alocadas em grupos diferentes, apesar de o número de cisternas incluídas em cada um dos grupos ter sido bem distinto.O método de Ward, concordando com Borysov *et al.* (2014), apresentou a solução mais adequada ao problema, sugerindo assim a construção de seis grupos com números diferentes de cisternas.

Observa-se que, mesmo a estrutura de todos os agrupamentos sendo similar, existem algumas alterações na maneira como as cisternas são agrupadas, ou seja, as cisternas que estão dentro de um mesmo grupo podem ser agrupadas em outra ordem, quando se mudam os algoritmos. Entretanto, isso não causa maiores problemas práticos.

No geral, os grupos devem ter a menor dispersão (variância) interna possível e serem “solitários” ou isolados em relação aos demais grupos, ou seja, espera-se que os grupos sejam distintos em relação aos demais grupos. Um bom agrupamento produz grupos internamente homogêneos e mantém a heterogeneidade entre eles. É desejável ainda que a distribuição das observações (cisternas) entre os grupos ocorra de maneira mais uniforme possível, com todos os grupos alocando aproximadamente o mesmo número de cisternas.

Um breve resumo sobre a matriz de distâncias euclidiana foi realizado, obtendo-se as seguintes estimativas: a média foi 1,43; a mediana foi 1,13; o desvio padrão foi 0,98; as distâncias máxima e mínima foram 3,67 e 0,01, respectivamente; e o coeficiente de variação foi 68,53%.

Na Tabela 10, observam-se os dados da matriz de distância euclidiana com a existência de seis grupos obtidos pelo método da média das distâncias.

Tabela 10. Estatísticas descritivas da matriz de distância euclidiana obtidas pelo método da média das distâncias

Grupos	1	2	3	4	5	6
Número de cisternas	70	7	6	8	7	2
Média	0,64	0,58	0,72	0,45	0,38	0,24
Mediana	0,60	0,57	0,99	0,33	0,35	0,24
Desvio padrão	0,31	0,30	0,44	0,27	0,16	-
Mínimo	0,01	0,13	0,13	0,08	0,18	0,24
Máximo	1,36	1,14	1,27	0,91	0,78	0,24

O grupo 3 foi o que apresentou maior dispersão com um desvio padrão de 0,44, enquanto o grupo 6 alocou apenas duas observações ou cisternas, podendo este grupo ser um *outlier*, o que requer uma maior atenção do pesquisador.

A matriz de distância euclidiana e o método da ligação simples apresentaram seis grupos com seus respectivos valores que estão descritos na Tabela 11. Os grupos 2 e 6 alocaram sete cisternas; os grupos 3 e 5, por sua vez, agruparam quatro e dez cisternas, respectivamente; o grupo 1 alocou a maioria das cisternas com 70 observações; enquanto o grupo 4 ficou com apenas duas cisternas. Logo, de acordo com a Tabela 11 e com o observado no Dendograma da Figura 8, o grupo 4, por classificar poucas cisternas, pode ser considerado um *outlier*, ou seja, as cisternas (C14 e C49) que estão no grupo 4 podem ter qualidade da água imprópria para o consumo doméstico.

Tabela 11. Estatísticas descritivas da matriz de distância euclidiana obtidas pelo método da ligação simples

Grupos	1	2	3	4	5	6
Número de cisternas	70	7	4	2	10	7
Média	0,64	0,59	0,36	0,20	0,80	0,38
Mediana	0,60	0,57	0,36	0,20	0,73	0,35
Desvio padrão	0,31	0,30	0,20	-	0,57	0,16
Mínimo	0,01	0,13	0,13	0,20	0,08	0,18
Máximo	1,36	1,14	0,57	0,20	1,79	0,78

Observou-se, na Tabela 12, considerando-se a matriz de distância euclidiana dos dados das cisternas relacionados às variáveis estudadas e o método da ligação completa com seis grupos, que os grupos 2 e 5 têm sete cisternas, os grupos 3 e 4 ficaram com seis e oito cisternas, respectivamente, e o grupo 1 classificou setenta cisternas, isto é, 70% das observações, enquanto o grupo 6 alocou apenas duas cisternas (C90, C98), conforme observado no dendograma apresentado na Figura 10. Este grupo, portanto, pode-se tratar de um *outlier* e estar com a qualidade da água comprometida para uso doméstico. Ainda vale ressaltar que o grupo 3 é o mais provável de conter cisternas com características distintas, uma vez que obteve a maior variância interna (Tabela 12), comparada com as variâncias do outros cinco grupos.

Tabela 12. Estatísticas descritivas da matriz de distância euclidiana obtidos pelo método da ligação completa.

Grupos	1	2	3	4	5	6
Número de cisternas	70	7	6	8	7	2
Média	0,64	0,58	0,76	0,45	0,38	0,24
Mediana	0,60	0,57	0,99	0,33	0,35	0,24
Desvio padrão	0,31	0,30	0,44	0,27	0,16	-
Mínimo	0,01	0,13	0,13	0,08	0,18	0,24
Máximo	1,36	1,14	1,27	0,91	0,78	0,24

A estatística descritiva, apresentada na Tabela 13, é referente à matriz de distância euclidiana e do método do centroide com seis grupos, em que os grupos 2, 4 e 5 agruparam quatro, oito e sete cisternas, respectivamente, o grupo 1, por sua vez, ficou com setenta e sete cisternas, ou seja, 77% das observações, enquanto os grupos 3 e 6 alocaram duas cisternas cada um (C14, C49) e (C90, C98), respectivamente. Ou seja, estes dois grupos podem se tratar de *outliers,* como já observado em outros métodos hierárquicos de agrupamento e no dendograma apresentado na Figura 12.

Tabela 13. Estatísticas descritivas da matriz de distância euclidiana obtidas pelo método do centroide

Grupos	1	2	3	4	5	6
Número de cisternas	77	4	2	8	7	2
Média	0,75	0,36	0,20	0,45	0,38	0,24
Mediana	0,71	0,36	0,20	0,33	0,35	0,24
Desvio padrão	0,38	0,20	-	0,27	0,16	-
Mínimo	0,01	0,13	0,20	0,08	0,18	0,24
Máximo	1,88	0,57	0,20	0,91	0,78	0,24

O grupo 1 é o candidato mais evidente a apresentar cisternas com características dissimilares, uma vez que sua variância interna (Tabela 13)é superior à variância interna dos demais grupos classificados pelo método do centroide.

Observam-se os valores descritivos da matriz de distância euclidiana e do método de Ward na Tabela 14, com seis grupos, em que os grupos 2 e 6 ficaram com sete cisternas cada um, os grupos 3 e 5 agruparam seis e dez cisternas, respectivamente, enquanto os grupos 1 e 4 foram os que ficaram com maior número de observações, alocando quarenta e duas e vinte oito cisternas, respectivamente.

Tabela 14. Estatísticas descritivas da matriz de distância euclidiana obtidas pelo método de Ward

Grupos	1	2	3	4	5	6
Número de cisternas	42	7	6	28	10	7
Média	0,39	0,58	0,76	0,43	0,80	0,38
Mediana	0,38	0,57	0,99	0,43	0,73	0,35
Desvio padrão	0,17	0,30	0,44	0,15	0,57	0,16
Mínimo	0,03	0,13	0,13	0,01	0,08	0,18
Máximo	0,82	1,14	1,27	0,87	1,79	1,78

O método de Ward apresentou um bom resultado, se comparado aos demais métodos hierárquicos de agrupamento, do ponto de vista de ter melhor dividido as cisternas entre os grupos (Tabela 14), de acordo com o dendograma apresentado na Figura 14. Vale ainda ressaltar que os grupos 3 e 5 são os mais heterogêneos, isto é, classificaram cisternas com características dissimilares, por terem as maiores variâncias internas.

De acordo com as estatísticas descritivas observadas na Tabela 15, relacionadas à matriz de distância euclidiana e ao método da mediana com seis grupos, os grupos 2, 4 e 5 agruparam quatro, oito e sete cisternas, respectivamente, o grupo 1

obteve 77% das observações com 77 cisternas, enquanto os grupos 3 e 6 alocaram duas cisternas cada um deles, isto é, (C14, C49) e (C90, C98), respectivamente.

Tabela 15. Estatísticas descritivas da matriz de distância euclidiana obtidas pelo método da mediana

Grupos	1	2	3	4	5	6
Número de cisternas	77	4	2	8	7	2
Média	0,75	0,36	0,20	0,45	0,38	0,24
Mediana	0,71	0,36	0,20	0,33	0,35	0,24
Desvio padrão	0,38	0,20	-	0,27	0,16	-
Mínimo	0,01	0,13	0,20	0,08	0,18	0,24
Máximo	1,88	0,57	0,20	0,91	0,78	0,24

Assim, as cisternas alocadas nos grupos 3 e 6 estão com características distintas das classificadas nos demais grupos, além do mais, essas cisternas que formaram esses dois grupos já formaram os mesmos agrupamentos em outros métodos hierárquicos estudados anteriormente, podendo esses dois grupos serem *outliers*. O grupo 1, por ter alocado um grande número de cisternas, obteve a maior variabilidade interna, sendo possivelmente o grupo mais heterogêneo entre todos os grupos formados pelo método da mediana.

A Tabela 16 apresenta os valores descritivos da matriz de distância euclidiana e do método de mcquitty com seis grupos,com a seguinte descrição: os grupos 2 e 5 alocaram sete cisternas cada um, os grupos 3 e 4 ficaram com seis e oito cisternas respectivamente, o grupo 1, por sua vez, classificou 70% das observações com setenta cisternas, enquanto o grupo 6 classificou apenas duas cisternas (C90, C98), como observado no dendograma apresentado na Figura 18, levando a indícios de que este grupo é um *outlier,* como observado em todos os métodos de análise de agrupamento hierárquico estudados anteriormente. O grupo mais disperso (Tabela 16) é o grupo 3 por ter variância interna superior aos demais grupos ora formados por esse método, ou seja, as duas cisternas alocadas nesse grupo devem ter características distintas ou dissimilares às cisternas alocadas nos demais grupos, formados pelo método de mcquitty.

Tabela 16. Estatísticas descritivas da matriz de distância euclidiana obtidas pelo método de mcquitty

Grupos	1	2	3	4	5	6
Número de cisternas	70	7	6	8	7	2
Média	0,64	0,58	0,76	0,45	0,38	0,24
Mediana	0,60	0,57	0,99	0,33	0,35	0,24
Desvio padrão	0,31	0,30	0,44	0,27	0,16	-
Mínimo	0,01	0,13	0,13	0,08	0,18	0,24
Máximo	1,36	1,14	1,27	0,91	0,78	0,24

De um modo geral, observa-se que, para todos os métodos hierárquicos de análise de agrupamento, o desvio padrão mínimo foi observado (Tabelas 10, 11, 12, 13, 14, 15 e 16), nos grupos que alocaram seis ou sete cisternas, isto é, os grupos com esses números de observações parecem serem os mais homogêneos ou terem características similares em relação à qualidade da água para todas as cisternas que estão neles alocados. Os métodos do centroide e da mediana foram os únicos que formaram dois grupos com apenas duas cisternas, enquanto o método de Ward foi o único que não formou grupos com apenas duas cisternas, apresentando assim uma divisão mais igualitária entre os seis grupos formados. Observou-se, em todos os algoritmos de agrupamento, nos quais se formaram grupos com duas cisternas que essas cisternas sempre foram (C14, C49) e C(90, C98). Vale ressaltar que o maior desvio padrão foi observado no grupo 5 para o método da ligação simples e de Ward, e, em ambos os métodos, o grupo 5 alocou dez cisternas, ou seja, essas cisternas alocadas nesse grupo têm a maior probabilidade de terem características distintas com relação à qualidade da água.

4.3 Métrica de Camberra

4.3.1 Análise dos dados utilizando a métrica de Camberra nos métodos de agrupamento hierárquico

Os dados de qualidade da água das cisternas de placas serão novamente analisados pelos métodos de agrupamento hierárquicos estudados na seção anterior, diferindo apenas pela distância utilizada, que, nesta seção, será a distância ou métrica de camberra, em vez da distância euclidiana. O objetivo é verificar se, com a utilização da distância de Camberra, que é uma medida de dissimilaridade insensível

às transformações de escala nas variáveis, os métodos de agrupamento hierárquicos apresentam resultados muito distintos dos apresentados com a utilização da distância euclidiana, que é mais sensível às diferenças de escala nas variáveis. A ideia é que, mesmo com a utilização das duas distâncias, cada um dos métodos não apresente soluções tão diferentes.

4.3.2 Dendogramas obtidos para os Métodos Hierárquicos

Observou-se, pelo dendograma apresentado na Figura 20, obtido pelo método da média das distâncias com a formação de seis grupos, os grupos 1, 2, 3, 4, 5 e 6 (Tabela 17), com as respectivas cisternas alocadas em cada grupo.

Tabela 17. Grupos de cisternas obtidos por meio do método da média das distâncias com métrica de camberra

Grupos	n	Cisternas
Grupo 1	41	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C20 C21 C22 C26 C29 C30 C31 C32 C33 C35 C38 C41 C55 C59 C60 C62 C63 C67 C76 C77 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 3	4	C12 C19 C43 C44
Grupo 4	10	C14 C15 C27 C28 C49 C50 C58 C72 C90 C98
Grupo 5	29	C18 C23 C25 C34 C36 C37 C39 C40 C42 C45 C46 C47 C48 C52 C56 C68 C69 C70 C74 C75 C78 C80 C81 C82 C83 C84 C85 C86 C87
Grupo 6	7	C64 C65 C66 C71 C79 C89 C100

O grupo 1 foi formado por quarenta e uma cisternas, ou seja, todas essas cisternas são similares segundo as variáveis em estudo, levando a indícios de que a qualidade da água de todos essas cisternas é semelhante, e ainda foi o que alocou a maior quantidade de cisternas. O grupo 5, por sua vez, classificou vinte e nove cisternas, os grupos 2 e 4 apresentaram nove e dez cisternas respectivamente, o grupo 3 foi formado apenas por quatro unidades, que foram as cisternas (C12, C19, C43, C44). O grupo 3,por ter classificado apenas quatro cisternas, deve ser observado com uma atenção especial,uma vez que as águas dessas cisternas podem estar com a qualidade imprópria para o uso doméstico, o grupo 6 classificou sete cisternas, uma quantidade de observações que não é alta se comparada ao grupo 1.

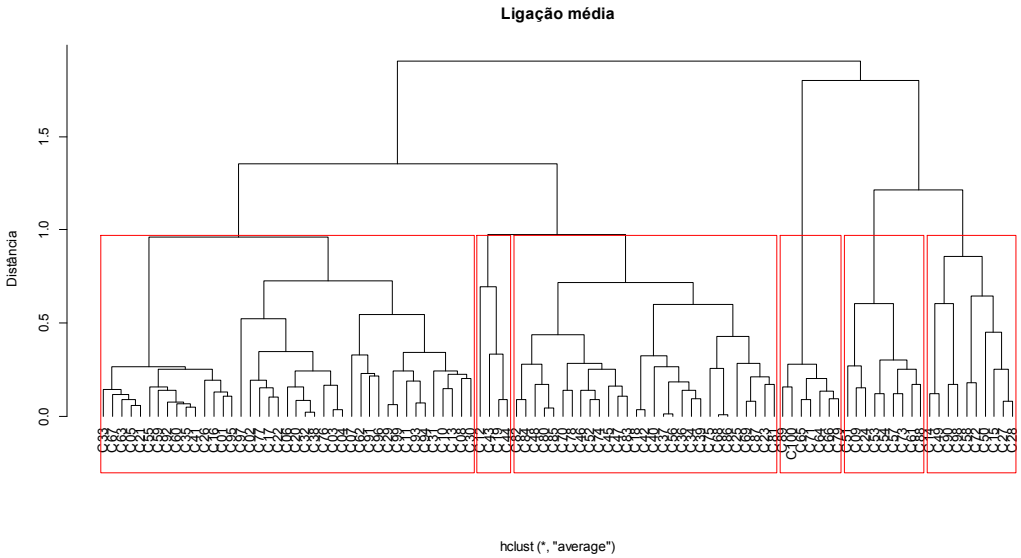

Figura 20. Dendograma resultante do método da média das distâncias com a métrica de camberra

Observa-se, na Figura 21, o gráfico da silhueta com as respectivas estatísticas da silhueta obtidas pelo método da média das distâncias, utilizando-se a métrica de camberra, isto é, nessa figura se encontra a estatística da silhueta para os seis grupos individualmente, assim como a estatística da silhueta média. Os grupos 2, 3 e 6 têm ótimas estatísticas da silhueta com valores de 0,64, 0,50 e 0,82, respectivamente, isto é, esses grupos alocaram de maneira adequada as cisternas que neles estão, assim essas cisternas não devem ser realocadas em outros grupos poreste algoritmo de agrupamento. Os grupos 1, 4 e 5, por sua vez, obtiveram valores da estatística da silhueta menores que os outros três grupos, em particular o grupo 4, que obteve a mais baixa estatística da silhueta, seguido pelo grupo 5, enquanto no grupo 1 ainda é possível uma classificação regular de acordo este critério de validação, pois seus valores foram 0,43, 0,35 e 0,40, respectivamente para os grupos 1, 4 e5 (Figura 21), em um intervalo de -1 a 1 que essa estatística pode assumir. A estatística da silhueta média foi de 0,46, que é considerado baixo para uma boa classificação das cisternas nos grupos obtidos pelo método da média da distância, utilizando a distância ou métrica de camberra.

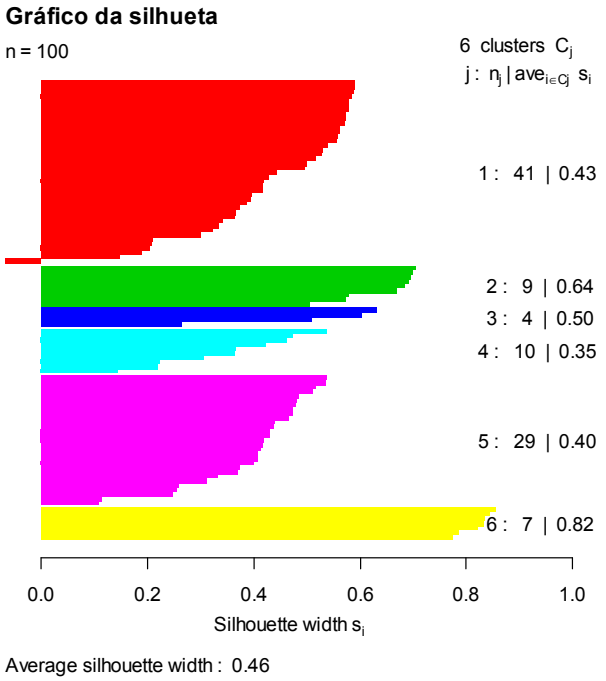

Figura 21. Gráfico e estatística da silhueta obtidos pelo método da média das distâncias com a métrica de camberra

O Dendograma obtido pelo método da ligação simples, utilizando a métrica de Camberra, está apresentado na Figura 22:

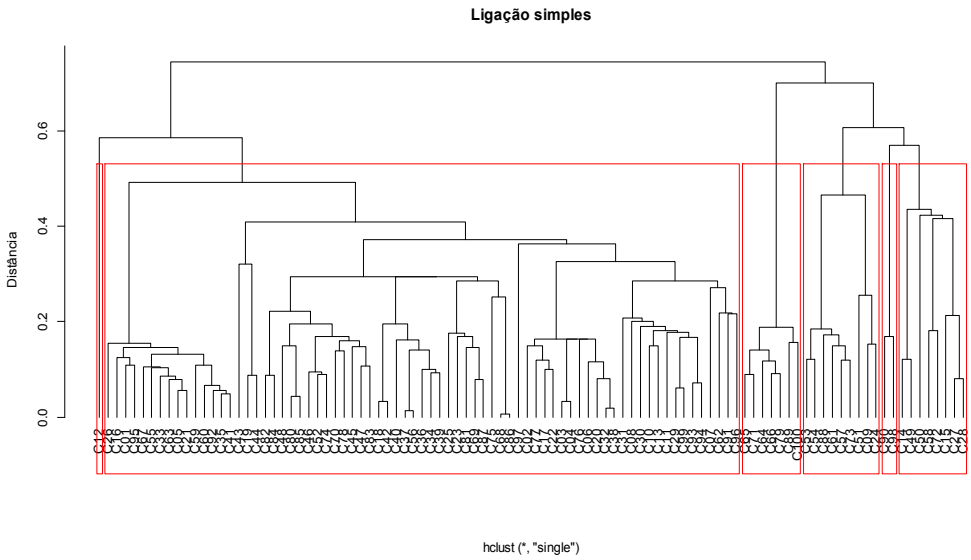

Figura 22. Dendograma resultante do método da ligação simples com a métrica de camberra

Observa-se a existência de seis grupos nessa figura que foi obtida pelo método da ligação simples, ressaltando-se um caso especial que deve ser observado, o grupo 3 com uma única observação que é a cisterna (C12). Na Tabela 18, estão a-

presentados os grupos 1, 2, 3, 4, 5 e 6, e cada um deles tem as respectivas cisternas alocadas pelo método.

Tabela 18. Grupos de cisternas obtidos por meio do método da ligação simples com métrica de camberra

Grupos	n	Cisternas
Grupo 1	73	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C18 C19 C20 C21 C22 C23 C25 C26 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C43 C44 C45 C46 C47 C48 C52 C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 3	1	C12
Grupo 4	8	C14 C15 C27 C28 C49 C50 C58 C72
Grupo 5	7	C64 C65 C66 C71 C79 C89 C100
Grupo 6	2	C90 C98

O grupo 1 é constituído por setenta e três cisternas, ou seja, segundo o método da ligação simples utilizando a métrica de camberra, todas as cisternas alocadas nesse grupo têm características similares de acordo com as variáveis estudadas. Os grupos 2, 4 e 5, por sua vez, classificaram nove, oito e sete cisternas, respectivamente, os grupos 3 e 6 alocaram uma e duas cisternas, respectivamente, isto é, as cisternas (C12) e(C90, C98), podendo esses dois grupos serem *outliers*, em especial o grupo 3,com cuja cisterna o pesquisador deve ter uma atenção especial, pois ela poderia estar com a qualidade da água imprópria para uso doméstico, conforme Fan *et al.* (2013). Por outro lado, a cisterna (C12) poderia estar alocada no grupo errado uma vez que sua estatística da silhueta (Figura 23) foi zero, ou seja, este grupo poderia ser inserido em outro grupo com características similares.

O Gráfico da silhueta com as respectivas estatísticas da silhueta estão apresentados na Figura 23, para o método da ligação simples com a métrica de Camberra. Observa-se que os grupos 2, 5 e 6 obtiveram boas estatísticas da silhueta com respectivos valores de 0,62, 0,77 e 0,79, sugerindo que as cisternas alocadas nesses três grupos possivelmente não serão realocadas para outros grupos, como observado por Peng *et al*. (2012).

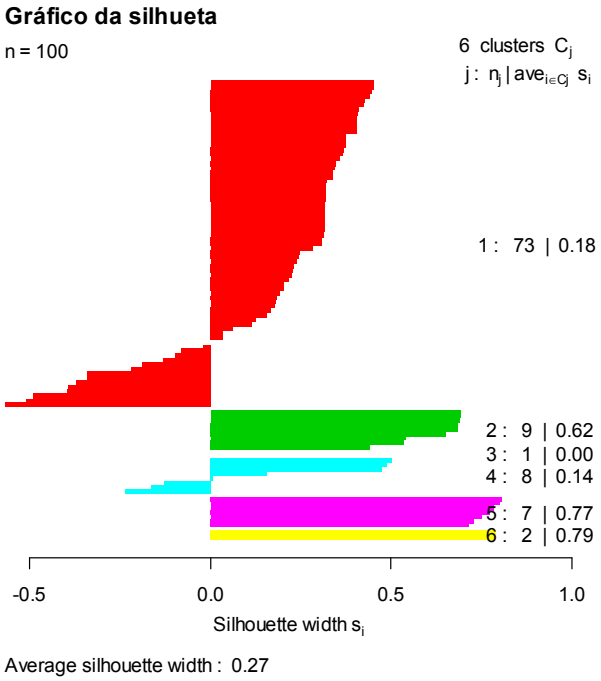

Figura 23. Gráfico e estatística da silhueta obtidos pelo método da ligação simples com a métrica de camberra

Os grupos 1, 3 e 4 tiveram estatísticas da silhueta respectivamente 0,18, 0,00 e 0,14, indicando classificações duvidosas das cisternas nesses três grupos, uma vez que vez as estatísticas da silhueta foram baixas. O grupo 3, em especial, obteve estatística da silhueta zero, sugerindo que esse grupo não exista, logo a cisterna (C12) que forma o grupo 3 deve ser alocada em outro grupo que seja mais similar.A estatística média da silhueta de 0,27 confirma que o método da ligação simples fez uma classificação moderada das cisternas nos seus respectivos grupos, corroborando com Lin *et al.* (2013) e sugerindo que cisternas que estão em um determinado grupo podem migrar para outro grupo.

De acordo com o Dendograma apresentado na Figura 24, obtido pelo método da ligação completa, utilizando a métrica de camberra, observa-se a presença de seis grupos, isto é, o método da ligação completa forma os grupos 1, 2, 3, 4, 5 e 6 (Tabela 19), com as respectivas cisternas.

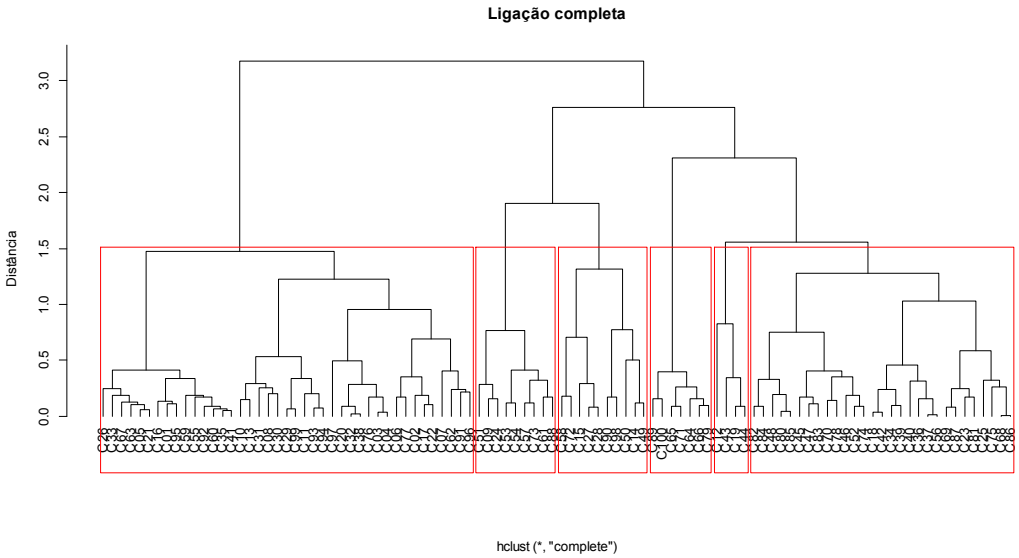

Figura 24. Dendograma resultante do método da ligação completa com a métrica de camberra

Tabela 19. Grupos de cisternas obtidos por meio do método da ligação completa com métrica de camberra

Grupos	n	Cisternas
Grupo 1	41	C1 C2 C3 C4 C5 C6 C7 C8 C10 C11 C13 C16 C17 C20 C21 C22 C26 C29 C30 C31 C32 C33 C35 C38 C41 C55 C59 C60 C62 C63 C67 C76 C77 C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 3	4	C12 C19 C43 C44
Grupo 4	10	C14 C15 C27 C28 C49 C50 C58 C72 C90 C98
Grupo 5	29	C18 C23 C25 C34 C36 C37 C39 C40 C42 C45 C46 C47 C48 C52 C56 C68 C69 C70 C74 C75 C78 C80 C81 C82 C83 C84 C85 C86 C87
Grupo 6	7	C64 C65 C66 C71 C79 C89 C100

O método da ligação completa utilizando a métrica de camberra apresentou visíveis diferenças dos outros métodos estudados até então. Uma delas é que o grupo 1 agrupou quarenta e uma cisternas, ou seja, um número bem diferente em relação ao mesmo grupo obtido pelos outros métodos de agrupamento. Observou-se também que o grupo 5 ficou com vinte e nove cisternas, o que também diferiu dos outros métodos de agrupamento, utilizando a métrica de camberra. Uma outra observação é que o método da ligação completa não formou grupos com duas cisternas como nos outros métodos de agrupamento. Os grupos 2, 3, 4 e 6 classificaram nove, quatro, dez e sete cisternas, respectivamente. De maneira geral, o método da ligação completa classificou as cisternas nos seis grupos de maneira mais igualitária,

isto é, não houve grupos com um número elevado de cisternas, nem grupos com poucas observações.

O método da ligação completa utilizando a métrica de camberra apresentou características peculiares em relação aos outros métodos utilizados para essas análises, como bem destacamos anteriormente. Essas características também foram notáveis no gráfico e nas estatísticas da silhueta observadas na Figura 25, pois o único grupo que obteve estatística da silhueta ótima com valor de 0,82 foi o grupo 6, como observado em Lin *et al.* (2013). Já os grupos 1, 2, 3 e 5 obtiveram estatísticas da silhueta 0,43, 0,64, 0,50 e 0,40,respectivamente, dando indícios de boas classificações das cisternas em seus respectivos grupos, mesmo com estatísticas da silhueta inferiores ao grupo 6, enquanto o grupo 4 obteve o menor valor da estatística da silhueta que foi 0,35. A estatística média da silhueta, obtida pelo método da ligação completa utilizando a métrica de Camberra, foi de 0,46, valor que dá uma boa credibilidade ao método,por ter classificado adequadamente as cisternas nos seus respectivos grupos.

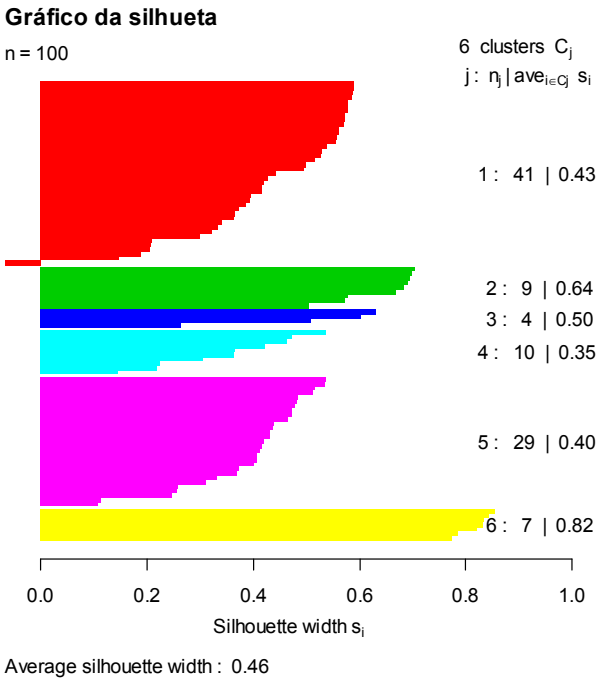

Figura 25. Gráfico e estatística da silhueta obtidos pelo método da ligação completa com a métrica de camberra

Em uma observação visual feita pelo Dendograma apresentado na Figura 26, percebe-se que o método do centroide utilizando a métrica de camberra indica a

presença de seis grupos com número distinto de cisternas (Tabela 20), em cada grupo, isto é, os grupos 1, 2, 3, 4, 5 e 6, com as respectivas cisternas.

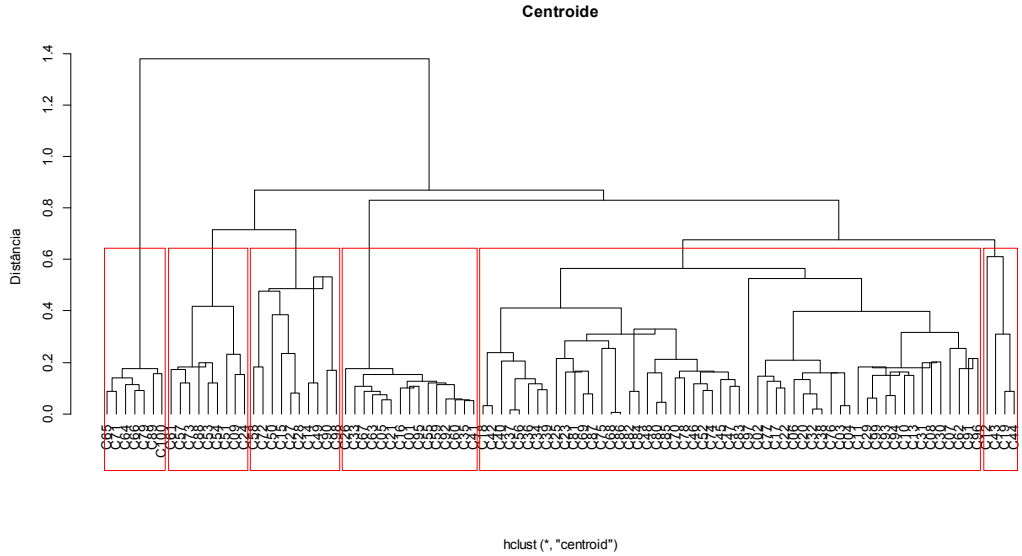

Figura 26. Dendograma resultante do método do centroide com a métrica de camberra

Tabela 20. Grupos de cisternas obtidos por meio do método do centroide com a métrica de camberra

Grupos	n	Cisternas
Grupo 1	15	C1 C5 C16 C21 C26 C33 C35 C41 C55 C59 C60 C63 C67 C92 C95
Grupo 2	55	C2 C3 C4 C6 C7 C8 C10 C11 C13 C17 C18 C20 C22 C23 C25 C29 C30 C31 C32 C34 C36 C37 C38 C39 C40 C42 C45 C46 C47 C48 C52 C56 C62 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C93 C94 C96 C97 C99
Grupo 3	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 4	4	C12 C19 C43 C44
Grupo 5	10	C14 C15 C27 C28 C49 C50 C58 C72 C90 C98
Grupo 6	7	C64 C65 C66 C71 C79 C89 C100

O grupo 1 é constituído por quinze cisternas como no método da ligação média, ou seja, conforme o método do centroide utilizando a métrica de camberra, todas as cisternas alocadas nesse grupo têm características similares,de acordo com as variáveis estudadas e com o que observaram You e Seo (2009). O grupo 2, por sua vez, classificou cinquenta e cinco cisternas, correspondendo a 55% das observações,enquanto os grupos 3, 4, 5 e 6 alocaram nove, quatro, dez e sete cisternas, respectivamente. Vê-se que o método do centroide apresentou característi-

cas idênticas às apresentadas pelo método da ligação média ou média das distâncias, em que as mesmas cisternas foram alocadas nos mesmo grupos, como observado nos dendogramas apresentados nas Figuras 20 e 26.

O Gráfico com as respectivas estatísticas da silhueta para o método do centroide estão apresentados na Figura 27, na qual as estatísticas da silhueta para cada um dos grupos são apresentadas individualmente, além da estatística média da silhueta.

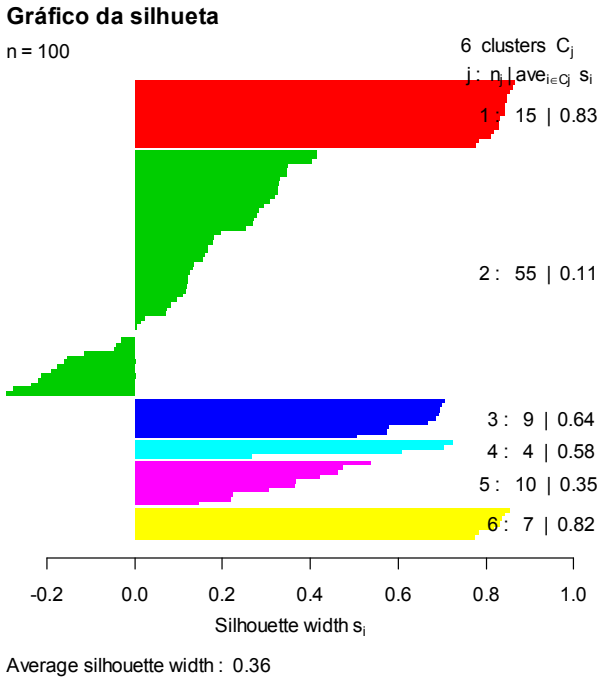

Figura 27. Gráfico e estatística da silhueta obtidos pelo método do centroide com a métrica de camberra

Observa-se que os grupos 1, 3 e 6 têm ótimas estatísticas da silhueta com valores de 0,83, 0,64 e 0,82, respectivamente, ou seja, esses grupos alocaram de maneira adequada as cisternas que neles estão, assim essas cisternas têm probabilidade baixa de serem realocadas em outros grupos. O grupo 4, por sua vez, obtive valor da estatística da silhueta inferior aos outros três grupos, mas o suficiente para garantir uma boa classificação, de acordo com esse critério de validação, pois seu valor foi de 0,58. Já os grupos 2 e 5 obtiveram estatísticas da silhueta mais baixas com valores de 0,11 e 0,35, respectivamente (Figura 27), podendo ser questionável a classificação desses dois grupos. A estatística da silhueta média foi de 0,36,o que pode resultar em uma classificação regular para os grupos obtidos pelo método do centroide.

O Dendograma obtido pelo método de Ward com a métrica de camberra está apresentado na Figura 28. Observa-se a existência de seis grupos nessa figura que foi obtida pelo método de Ward.

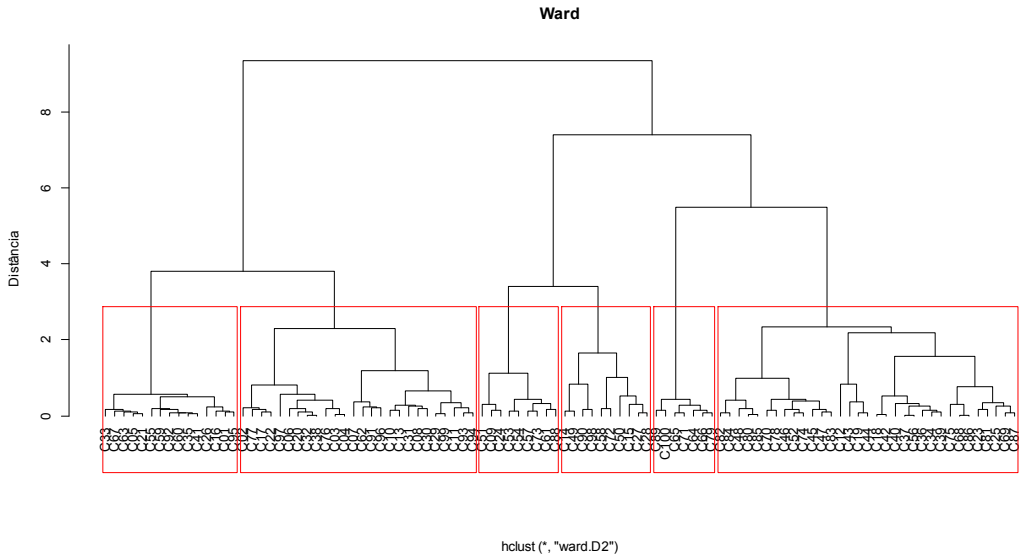

Figura 28. Dendograma resultante do método de Ward com a métrica de camberra

Na Tabela 21, estão apresentados os grupos 1, 2, 3, 4, 5 e 6, e, em cada um desses grupos, foram alocadas as respectivas cisternas pelo método de Ward com a métrica de camberra.

Tabela 21. Grupos de cisternas obtidos por meio do método de Ward com a métrica de camberra

Grupos	n	Cisternas
Grupo 1	15	C1 C5 C16 C21 C26 C33 C35 C41 C55 C59 C60 C63 C67 C92 C95
Grupo 2	26	C2 C3 C4 C6 C7 C8 C10 C11 C13 C17 C20 C22 C29 C30 C31 C32 C38 C62 C76 C77 C91 C93 C94 C96 C97 C99
Grupo 3	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 4	33	C12 C18 C19 C23 C25 C34 C36 C37 C39 C40 C42 C43 C44 C45 C46 C47 C48 C52 C56 C68 C69 C70 C74 C75 C78 C80 C81 C82 C83 C84 C85 C86 C87
Grupo 5	10	C14 C15 C27 C28 C49 C50 C58 C72 C90 C98
Grupo 6	7	C64 C65 C66 C71 C79 C89 C100

Observa-se que o grupo 1 é constituído por quinze cisternas, como nos métodos da média das distâncias e do centroide utilizando a métrica de camberra, ou seja, de acordo com o método de Ward, todas as cisternas alocadas nesse grupo têm características similares,conforme as variáveis mensuradas. Os grupos 2 e 4, por

sua vez, classificaram o maior número de cisternas, ou seja, vinte e seis e trinta e três cisternas, respectivamente, os grupos 3 e 5 alocaram nove e dez cisternas respectivamente, enquanto o grupo 6 obteve o menor número de observações com sete cisternas.Portanto, o método de Ward com a métrica de camberra (Figura 28) e com a distância euclidiana (Figura 14) divide as cisternas entre os seis grupos obtidos pelo método de forma mais "justa", se comparado aos outros métodos, isto é, não existem grupos com muitas cisternas ou grupos com uma ou duas cisternas apenas.

O método de Ward com a métrica de camberra apresentou características peculiares em relação aos outros métodos. Essas características também foram notáveis no Gráfico e nas estatísticas da silhueta observadas na Figura 29, na qual os grupos 1, 3 e 6 obtiveram boas estatísticas da silhueta com valores de 0,77, 0,64 e 0,84, respectivamente, enquanto os grupos 2, 4 e 5 obtiveram estatísticas da silhueta 0,39, 0,40 e 0,35, respectivamente, dando indícios de boas classificações das cisternas em seus respectivos grupos, como encontrado por Clifford *et al.* (2011), mesmo com estatísticas da silhueta inferiores aos grupos 1, 3 e 6. A estatística média da silhueta obtida pelo método de Ward com a métrica de camberra foi de 0,50, dando indícios de que esse método fez uma boa classificação das cisternas nos seus respectivos grupos.

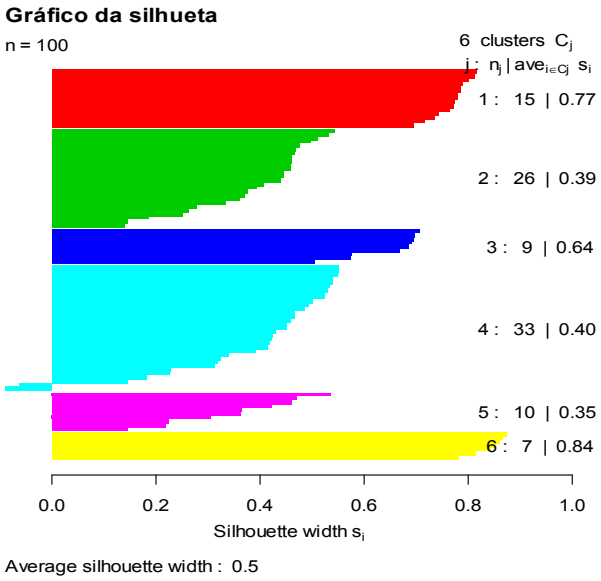

Figura 29: Gráfico e estatística da silhueta obtidos pelo método de Ward com a métrica de camberra

De acordo com o Dendograma apresentado na Figura 30, obtido pelo método da mediana com a métrica de camberra, observa-se a presença de seis grupos, isto

é, o método da mediana forma os grupos 1, 2, 3, 4, 5 e 6 (Tabela 22), com as respectivas cisternas alocadas nos grupos por esse método.

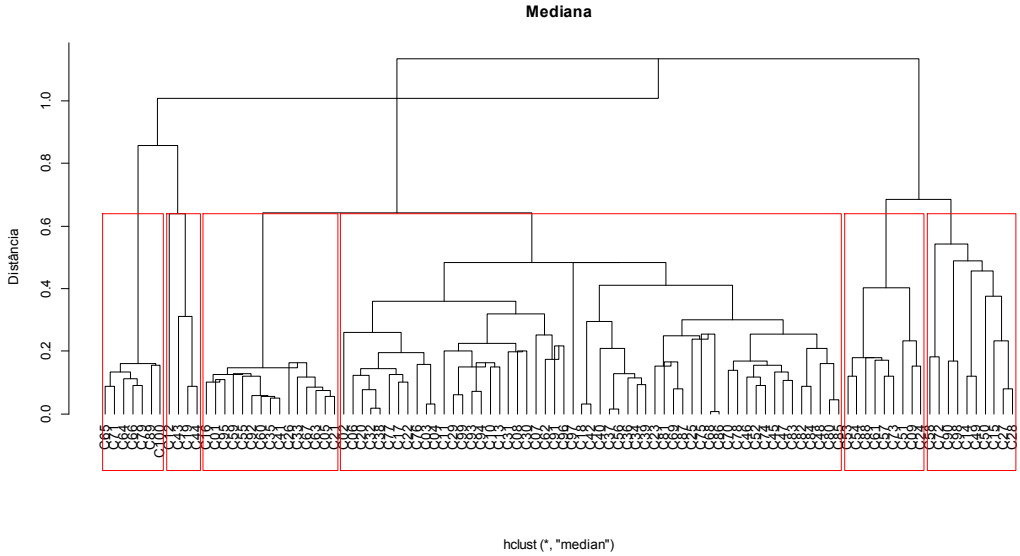

Figura 30. Dendograma resultante do método da mediana com a métrica de camberra

Tabela 22. Grupos de cisternas obtidos por meio do método da mediana com a métrica de camberra

Grupos	n	Cisternas
Grupo 1	15	C1 C5 C16 C21 C26 C33 C35 C41 C55 C59 C60 C63 C67 C92 C95
Grupo 2	55	C2 C3 C4 C6 C7 C8 C10 C11 C13 C17 C18 C20 C22 C23 C25 C29 C30 C31 C32 C34 C36 C37 C38 C39 C40 C42 C45 C46 C47 C48 C52 C56 C62 C68 C69 C70 C74 C75 C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87 C91 C93 C94 C96 C97 C99
Grupo 3	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 4	4	C12 C19 C43 C44
Grupo 5	10	C14 C15 C27 C28 C49 C50 C58 C72 C90 C98
Grupo 6	7	C64 C65 C66 C71 C79 C89 C100

Pelo método da mediana com a métrica de camberra foram formados seis grupos (Figura 30). Observou-se que o grupo 1 é formado por quinze cisternas, o grupo 2 alocou cinquenta e cinco cisternas, o que é considerado um número grande de observações. Os grupos3 e 5 agruparam nove e dez cisternas, respectivamente, enquanto o grupo 6 ficou com sete cisternas, e o grupo 4 formado por apenas quatro cisternas, sendo este o grupo que alocou o menor número de cisternas, de acordo com método da mediana.

Observou-se, na Figura 31, que as estatísticas da silhueta indicam a consistência dos grupos 1, 3 e 6 com uma boa homogeneidade, tendo estatísticas da silhueta 0,83, 0,64 e 0,82, respectivamente, apontado que as cisternas foram bem classificadas nesses três grupos.

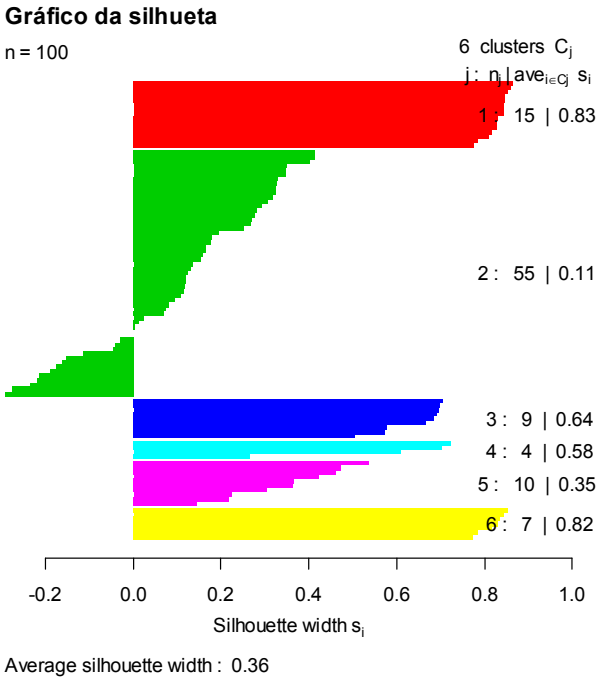

Figura 31. Gráfico e estatística da silhueta obtidos pelo método da mediana com a métrica de camberra

Os grupos 2, 4 e 5 apresentaram estatísticas da silhueta de 0,11, 0,58 e 0,35, respectivamente, que, apesar de não terem valores elevados, como para os outros três grupos, ainda apresentam indícios de uma boa classificação das cisternas nos grupos para os quais foram alocadas, em especial o grupo 4. Já o grupo 2 tem uma classificação questionável, uma vez que sua estatística da silhueta foi a mais baixa de todos. De maneira geral, de acordo com o gráfico da silhueta e a estatística da silhueta média (Figura 31), que é de 0,36, as cisternas estão classificadas de forma regular nos seus respectivos grupos, sendo possível realocação de observações para outros grupos.

O Dendograma apresentado na Figura 32, obtido pelo método de mcquitty com a métrica de Camberra, apresenta seis grupos que têm número distinto de cisternas,como observado na Tabela 23, isto é, os grupos 1, 2, 3, 4, 5 e 6, com as respectivas cisternas alocadas em cada grupo.

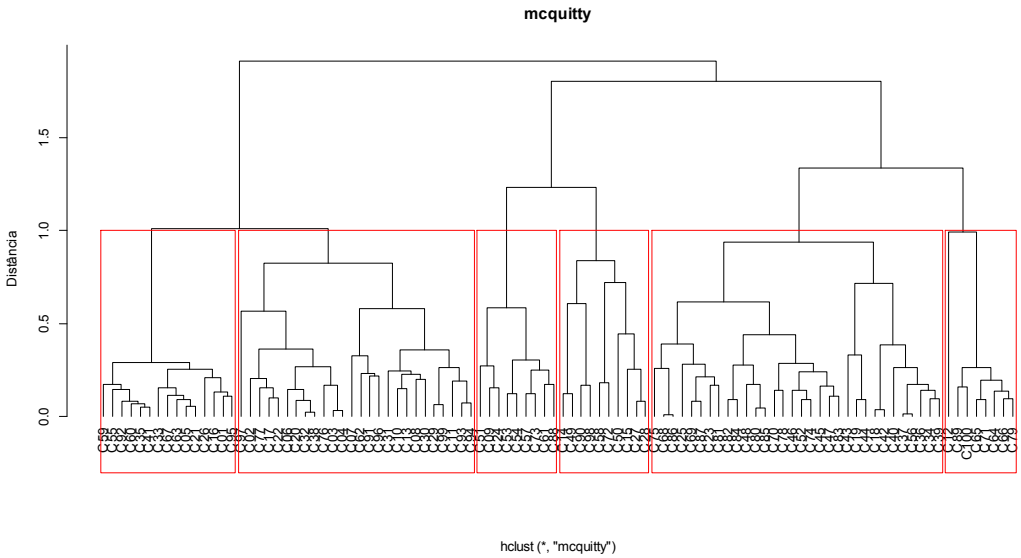

Figura 32. Dendograma resultante do método de mcquitty com a métrica de camberra

Tabela 23. Grupos de cisternas obtidos por meio do método de mcquitty com a métrica de camberra

Grupos	n	Cisternas
Grupo 1	15	C1 C5 C16 C21 C26 C33 C35 C41 C55 C59 C60 C63 C67 C92 C95
Grupo 2	26	C2 C3 C4 C6 C7 C8 C10 C11 C13 C17 C20 C22 C29 C30 C31 C32 C38 C62 C76 C77 C91 C93 C94 C96 C97 C99
Grupo 3	9	C9 C24 C51 C53 C54 C57 C61 C73 C88
Grupo 4	8	C12 C64 C65 C66 C71 C79 C89 C100
Grupo 5	10	C14 C15 C27 C28 C49 C50 C58 C72 C90 C98
Grupo 6	32	C18 C19 C23 C25 C34 C36 C37 C39 C40 C42 C43 C44 C45 C46 C47 C48 C52 C56 C68 C69 C70 C74 C75 C78 C80 C81 C82 C83 C84 C85 C86 C87

Observa-se que o grupo 1 alocou quinze cisternas de acordo com o método de mcquitty utilizando a métrica de camberra, diferindo das outras técnicas de agrupamento apresentadas anteriormente com essa distância. Outro diferencial observado nesse método de agrupamento é que os grupos 2 e 6 foram formados por um número elevado de observações, com vinte e seis e trinta e duas cisternas,respectivamente, o que não havia ocorrido em outros métodos de agrupamento, em especial com relação ao grupo 6. Os grupos 3, 4 e 5 foram formados de maneira similar a outros métodos de agrupamento, com nove, oito e dez cisternas, respectivamente. Nessas condições o que merece atenção especial é o grupo 6, já que foi

formado em outros métodos de agrupamento por poucas cisternas, o que não ocorreu nesse método com a métrica de camberra.

O gráfico da silhueta com as respectivas estatísticas da silhueta estão apresentados na Figura 33, para o método de mcquitty, utilizando-se a métrica de Camberra, pois observa-se que os grupos 1, 3 e 4 têm boas estatísticas da silhueta, com respectivos valores de 0,77, 0,64 e 0,68, sugerindo que as cisternas alocadas nesses três grupos têm probabilidades baixas de serem realocadas para outros grupos, o que corrobora com Arbelaitz *et al.*(2013). Os grupos 2, 5 e 6 tiveram estatísticas da silhueta respectivamente de 0,39, 0,35 e 0,42, indicando classificações regulares das cisternas nesses três grupos, mesmo tendo estatísticas da silhueta inferiores aos outros três grupos. A estatística média da silhueta de 0,50 confirma que o método de mcquitty com a métrica de camberra teve uma boa classificação das cisternas nos seus respectivos grupos.

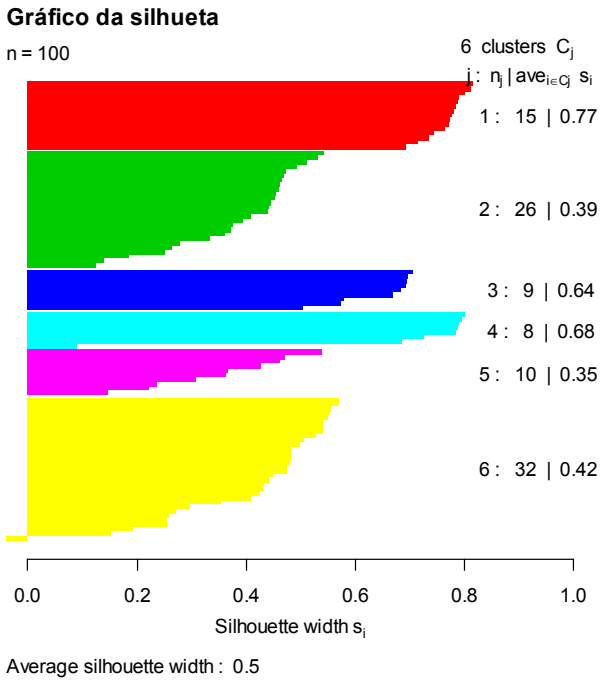

Figura 33. Gráfico e estatística da silhueta obtidos pelo método de mcquitty com a métrica de camberra

Após uma análise individual dos resultados dos métodos de ligação média ou média das distâncias, de ligação simples, de ligação completa, de centroide, de Ward, de mediana e de mcquitty, realizou-se uma comparação dos agrupamentos obtidos por cada método, focando as características desses grupos, assim como o número de grupos, tal qual fez Clifford *et al.* (2011).

Os agrupamentos obtidos pelo método de Ward apresentaram um melhor desempenho de acordo com a estatística média da silhueta e individual de cada grupo em relação aos métodos de ligação média, de ligação simples, de centroide, de mediana e de mcquitty. O método de mcquitty também obteve resultado similar ao método de Ward, com estatística média da silhueta igual a este método, diferindo apenas pelo fato do grupo 6 do método de mcquitty apresentar estatística da silhueta com apenas 50% do valor da estatística da silhueta do mesmo grupo para o método de Ward.Já os agrupamentos obtidos pelo método da ligação simples foram os que obtiveram desempenho mais fraco, segundo a estatística média da silhueta e as estatísticas da silhueta individual para cada grupo, pois o grupo 4 obteve estatística da silhueta (Figura 23) igual a zero, sugerindo que esse grupo classificou incorretamente ou mesmo que esse grupo não existe.

Um breve resumo sobre a matriz da distância de camberra ou métrica de camberra foi realizado, obtendo-se as seguintes estimativas: a média foi de 1,38; a mediana, 1,35; o desvio padrão, 0,67; o mínimo e o máximo da métrica de camberra, 0,01 e 3,18, respectivamente; e o coeficiente de variação, 48,55%, indicando uma homogeneidade regular.

De acordo com as estatísticas descritivas observadas na Tabela 24, relacionadas à matriz da distância ou métrica de camberra e ao método da média das distâncias com seis grupos, os grupos 2, 4 e 6 agruparam nove, dez e sete cisternas, respectivamente, enquanto os grupos 1 e 5 alocaram quarenta e uma e vinte nove cisternas, respectivamente, já o grupo 3 obteve 4% das observações com quatro cisternas,isto é (C12, C19, C43, C44).

Tabela 24. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método da média das distâncias

Grupos	1	2	3	4	5	6
Número de cisternas	41	9	4	10	29	7
Média	0,39	0,83	0,36	1,21	0,44	0,38
Mediana	0,37	0,87	0,36	1,42	0,44	0,35
Desvio padrão	0,17	0,62	0,20	0,53	0,16	0,16
Mínimo	0,03	0,08	0,13	0,14	0,01	0,18
Máximo	0,82	1,49	0,57	1,98	0,88	0,78

Assim, as cisternas alocadas no grupo 3 estão com características distintas das classificadas nos demais grupos, podendo esse grupo ser um *outlier*. O grupo 2 ob-

teve a maior variabilidade interna, sendo possivelmente o grupo mais heterogêneo entre todos os grupos, seguido pelo grupo 4, formado pelo método da média das distâncias com a métrica de camberra.

A Tabela 25 apresenta os valores descritivos da matriz da distância de camberra ou métrica de Camberra, utilizando o método da ligação simples com seis grupos e apresentando a seguinte descrição: os grupos 2, 4 e 5 alocaram nove, oito e sete cisternas, respectivamente; o grupo 1, por sua vez, classificou 73% das observações com setenta e três cisternas, enquanto o grupo 6 classificou duas cisternas (C90, C98), e o grupo 3 alocou apenas uma cisterna (C12), como observado no dendograma apresentado na Figura 22, levando a indícios de que esses dois grupos são *outliers*.

Tabela 25. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método da ligação simples

Grupos	1	2	3	4	5	6
Número de cisternas	73	9	1	8	7	2
Média	0,72	0,83	-	1,11	0,38	0,24
Mediana	0,64	0,87	-	1,34	0,35	0,24
Desvio padrão	0,42	0,62	-	0,57	0,16	-
Mínimo	0,01	0,08	-	0,13	0,18	0,24
Máximo	2,06	1,49	-	1,98	1,78	0,24

O grupo mais disperso (Tabela 25) é o grupo 2 por ter variância interna superior à dos demais grupos formados por esse método, ou seja, as nove cisternas alocadas nesse grupo devem ter características distintas ou dissimilares às cisternas alocadas nos demais grupos formados pelo método da ligação simples. Já o grupo 3, que alocou uma cisterna, deve ser verificado com atenção, pois este grupo pode não existir, e a cisterna (C12) alocada no grupo 3 unir-se a outro grupo que seja mais similar a ele.

A descrição apresentada na Tabela 26 é referente à matriz da distância de camberra ou métrica de camberra e do método da ligação completa com seis grupos. Nesse caso, os grupos 2, 4 e 6 agruparam nove, dez e sete cisternas, respectivamente, o grupo 3, por sua vez, ficou apenas com quatro cisternas, ou seja, 4% das observações (C12, C19, C43, C44), enquanto os grupos 1 e 5 alocaram quarenta e uma e vinte nove cisternas, respectivamente, ou seja, estes dois grupos ficaram com um maior número de observações, como já observado no método da

ligação média pelo dendograma da Figura 20. O grupo 2 é o que apresenta cisternas com características mais dissimilares, uma vez que seu desvio padrão interno (Tabela 26) é superior ao desvio padrão interno dos demais grupos classificados pelo método da ligação completa.

Tabela 26. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método da ligação completa

Grupos	1	2	3	4	5	6
Número de cisternas	41	9	4	10	29	7
Média	0,39	0,83	0,36	1,21	0,44	0,38
Mediana	0,37	0,87	0,36	1,41	0,44	0,35
Desvio padrão	0,17	0,62	0,20	0,53	0,16	0,16
Mínimo	0,03	0,08	0,13	0,13	0,01	0,18
Máximo	0,82	1,49	0,57	1,98	0,88	0,78

A Tabela 27 apresenta os valores descritivos da matriz da distância ou métrica de camberra com método do centroide, com seis grupos apresentando a seguinte descrição: os grupos 1 e 2 alocaram quinze e cinquenta e cinco cisternas,respectivamente; os grupos 3, 5 e 6 ficaram com nove, dez e sete cisternas, respectivamente;e o grupo 4, por sua vez, classificou o menor número de observações do método do centroide, utilizando a métrica de camberra com quatro cisternas (C12, C19, C43, C44), como observado no Dendograma apresentado na Figura 26. O grupo mais disperso (Tabela 27) é o grupo 3 por ter desvio padrão interno superior ao dos demais grupos ora formados por esse método, isto é, as nove cisternas alocadas nesse grupo devem ter características distintas ou dissimilares, mesmo estando alocadas no mesmo grupo.

Tabela 27. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método do centroide

Grupos	1	2	3	4	5	6
Número de cisternas	15	55	9	4	10	7
Média	0,24	0,62	0,83	0,36	1,21	0,38
Mediana	0,23	0,56	0,87	0,36	1,42	0,35
Desvio padrão	0,12	0,29	0,62	0,20	0,53	0,16
Mínimo	0,03	0,01	0,08	0,13	0,13	0,18
Máximo	0,48	1,31	1,49	0,57	1,98	0,78

De acordo com as estatísticas descritivas observadas na Tabela 28, relacionadas à matriz da distância ou métrica de camberra e do método de Ward com seis

grupos, os grupos 3, 5 e 6 agruparam nove, dez e sete cisternas, respectivamente, já os grupos 1, 2 e 4 alocaram quinze, vinte e seis e trinta e três cisternas, respectivamente. Assim, as cisternas alocadas nos grupos 1, 2 e 6 estão com características bastante similares por apresentarem os menores desvios padrões interno sem relação aos dos demais grupos, como encontrado por Berge *et al.* (2003). O grupo 3 obteve a maior variabilidade interna, sendo possivelmente o grupo mais heterogêneo entre todos os grupos formados pelo método de Ward com a métrica de camberra.

Tabela 28. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método de Ward

Grupos	1	2	3	4	5	6
Número de cisternas	15	26	9	33	10	7
Média	0,24	0,36	0,83	0,62	1,21	0,38
Mediana	0,23	0,38	0,87	0,50	1,42	0,35
Desvio padrão	0,12	0,14	0,62	0,40	0,53	0,16
Mínimo	0,03	0,04	0,08	0,01	0,13	0,18
Máximo	0,48	0,65	1,49	1,96	1,98	0,78

Na Tabela 29, observam-se os dados da matriz da distância ou métrica de camberra com existência de seis grupos obtidos pelo método da mediana. O grupo 1 é formado por quinze cisternas, o grupo 2 é formado por cinquenta e cinco cisternas, o grupo 3 é formado por nove cisternas, o grupo 4 é formado por quatro cisternas, o grupo 5 é formado por dez cisternas, e o grupo 6 é formado por sete cisternas. O grupo 3 foi o que apresentou maior dispersão com um desvio padrão de 0,62, enquanto o grupo 4 alocou apenas quatro observações ou cisternas, podendo este grupo ser um *outlier*, o que requer uma maior atenção do pesquisador.

Tabela 29. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método da mediana

Grupos	1	2	3	4	5	6
Número de cisternas	15	55	9	4	10	7
Média	0,24	0,62	0,83	0,36	1,21	0,38
Mediana	0,23	0,56	0,87	0,36	1,42	0,35
Desvio padrão	0,12	0,29	0,62	0,20	0,53	0,16
Mínimo	0,03	0,01	0,08	0,13	0,13	0,18
Máximo	0,48	1,31	1,49	0,57	1,98	0,78

As estatísticas descritivas apresentadas na Tabela 30 são referentes à matriz da distância ou métrica de camberra e do método de mcquitty com seis grupos. Dentre eles, os grupos 3, 4 e 5 agruparam nove, oito e dez cisternas, respectivamente. O grupo 1, por sua vez, ficou com quinze cisternas, ou seja, 15% das observações, enquanto os grupos 2 e 6 alocaram vinte seis e trinta e duas cisternas, respectivamente, ou seja, esses dois grupos foram os que alocaram os maiores números de cisternas através desse método de agrupamento, como observado no dendograma apresentado na Figura 32. O grupo 4 é o candidato mais evidente a apresentar cisternas com características dissimilares, uma vez que seu desvio padrão interno (Tabela 30) de 0,94 é superior ao desvio padrão interno dos demais grupos classificados pelo método de mcquitty, assim como nos valores encontrados por Berge *et al.* (2003), na técnica hierárquica de Ward.

Tabela 30. Estatísticas descritivas da matriz de distância ou métrica de camberra obtidas pelo método da mcquitty

Grupos	1	2	3	4	5	6
Número de cisternas	15	26	9	8	10	32
Média	0,24	0,36	0,83	0,91	1,21	0,57
Mediana	0,23	0,38	0,87	0,42	1,42	0,48
Desvio padrão	0,12	0,14	0,62	0,94	0,53	0,34
Mínimo	0,03	0,04	0,08	0,18	0,13	0,01
Máximo	0,48	0,65	1,49	2,53	1,98	1,63

Uma vez que este trabalho tem como principal objetivo a formação de grupos de cisternas que tenham a menor variabilidade interna possível, o método de Ward apresentou a solução mais apropriada para o nosso problema, pois formou grupos com cisternas mais similares ou homogêneas, segundo as variáveis de qualidade da água mensuradas.

De acordo com Mingote (2007), um grupo pode representar uma observação ou cisterna, sendo possível conhecer as observações incluídas nos grupos além da relação entre elas, assim como a estimação da hierarquia entre as observações. Então, a análise de agrupamento hierárquico de dados organiza os grupos numa árvore hierárquica, facilitando a navegação entre os dados.

As técnicas de agrupamentos hierárquicos possuem outras características e vantagens em conformidade com Silva (2005), isto é, com o conhecimento da relação entre os dados e a descoberta dos grupos que alocaram as observações

existentes e suas ligações, ainda é possível verificar o número “ideal” de grupos para técnicas não hierárquicas abordadas mais adiante, entre outras vantagens.

Comparando-se os resultados obtidos através dos métodos hierárquicos de agrupamento utilizados no agrupamento de cisternas de placas quanto à qualidade da água nelas armazenadas, o método de Ward e o método de mcquitty obtiveram as melhores soluções para o problema, quando levada em consideração a divisão das cisternas entre os grupos. O método de Ward obteve bons resultados tanto com a distância euclidiana quanto com a métrica de Camberra, como encontrados em Berge *et al.* (2003) e Clifford *et al.* (2011). O método de mcquitty, por um lado, obteve um bom resultado apenas com a métrica de camberra. Por outro lado, o método de mcquitty com a métrica de camberra foi o que apresentou a maior variância interna para um grupo, quando comparado a todos os outros métodos apresentados, utilizando a distância euclidiana e a métrica de camberra.

4.4 Métodos Não Hierárquicos

Existem uma gama de métodos não hierárquicos fundamentados na combinação de distribuições, na teoria da estimação de densidades e partição. A principal diferença entre os métodos hierárquicos discutidos em seções anteriores e os métodos não hierárquicos é que naqueles, em geral, não se conhece o número de grupos *k*, pois sendo um objeto ou cisterna alocado em um determinado grupo, ele não deve mais ser realocado em outro grupo. Já, nos métodos não hierárquicos, o processo de agrupamento é aplicado diretamente na matriz de dados, e não na matriz de dissimilaridade como nos métodos hierárquicos.

Nos métodos não hierárquicos, fixou-se *a* priori o número *k* de grupos, e assim as cisternas foram alocadas nos k grupos, utilizando uma função objetivo como critério. Dessa forma, a realocação das cisternas encerrou-se quando uma regra de parada especificada a priori for satisfeita como observado por Ferreira (2008). Esse método em geral é definido como método de otimização em vez de partição, em virtude de suas características. A seguir, serão analisados e discutidos os resultados obtidos dos dados das cisternas de placas da região do Pajeú – PE, localizadas no sertão do estado de Pernambuco, utilizando-se três métodos não hierárquicos de agrupamento, a saber: os agrupamentos *fuzzy*, o *k* – média e o *k* – medoide.

4.4.1 Agrupamento Fuzzy

Os sistemas Híbridos caracterizam-se pela combinação de duas ou mais técnicas de Inteligência Computacional em um só modelo,objetivando-se utilizar o que existe de mais robusto para obter a melhor solução para um dado problema. Nesse contexto, existem os modelos híbridos construídos com Redes Neurais Artificiais (RNA).Portanto, os modelos híbridos têm como base a adição da Teoria de Conjuntos *Fuzzy* utilizando o algoritmo *c – Means*.

Nessa linha, Bezdek *et al.* (1992) propôs o algoritmo *Fuzzy c – Means,*o qual funciona como base para outros modelos híbridos utilizados na tarefa de agrupamento ou classificação *fuzzy.* Segundo esse modelo, uma variação não supervisionada da RNA é adicionada de característica *fuzzy*.

O agrupamento apresentado a seguir é obtido pelo método não hierárquico de agrupamento *fuzzy*,no qual se observa a presença de seis grupos. O método de agrupamento *fuzzy* forma os grupos 1, 2, 3, 4, 5 e 6 (Tabela 31), com as respectivas cisternas.

Tabela 31. Grupos de cisternas obtidos por meio do método não hierárquico de agrupamento *fuzzy*

Grupos	n	Cisternas
Grupo 1	7	C73 C45 C46 C56 C69 C78 C87
Grupo 2	12	C12 C14 C18 C19 C42 C43 C44 C49 C50 C52 C70 C74
Grupo 3	29	C03 C04 C05 C07 C08 C10 C11 C13 C21 C23 C25 C26 C30 C32 C33 C38 C39 C48 C63 C67 C75 C76 C77 C82 C84 C91 C93 C94 C97
Grupo 4	29	C01 C02 C06 C16 C17 C20 C22 C29 C31 C34 C35 C36 C40 C41 C47 C55 C59 C60 C62 C68 C80 C81 C83 C85 C86 C92 C95 C96 C99
Grupo 5	15	C09 C15 C27 C28 C51 C53 C54 C57 C58 C61 C72 C73 C88 C90 C98
Grupo 6	8	C24 C64 C65 C66 C71 C79 C89 C100

O agrupamento apresentado na Tabela 31 foi obtido pela matriz do grau de pertinência de cada cisterna ao seu respectivo grupo. Assim, um objeto ou cisterna é atribuído a cada grupo de acordo com o grau de adesão máxima aos grupos, e essa matriz produz a partição de agrupamentos o mais próximo possível da realidade. Em outras palavras, uma cisterna é atribuída a cada grupo de acordo com o grau de pertinência máxima, desde que tal grau de pertinência seja igual ou maior a 0,5 (≥ 0,5), o que foi observado em todas as cisternas. Caso contrário, presume-se que a cister-

na que obteve grau de pertinência inferior a 0,5 não pertence a nenhum grupo, o que não se verificou neste agrupamento *fuzzy*.

As cem cisternas consideradas no estudo (Tabela 31) foram analisadas para extrair características para a análise de agrupamento. As correlações entre as características e as variáveis relacionadas com a qualidade da água foram analisadas, ou seja, as seis variáveis de qualidade da água ora analisadas foram relacionadas aos sais minerais e às condições higiênicas e sanitárias.

As variáveis de localização, latitude e longitude então inclusas no estudo para identificar as regiões onde estão localizadas as cisternas que são geograficamente próximas. As seis variáveis foram padronizadas, como já mencionado neste trabalho, e foi atribuída a mesma escala de importância para todas as variáveis, o que significa que nenhuma variável em particular influenciou no resultado dos agrupamentos, exceto pela característica natural da variável.

Para analisar a sensibilidade do resultado do algoritmo *Fuzzy c – Means* (FCM), a variação no parâmetro de fuzificação m é de 1,4 a 2,5, no entanto,Pal e Bezdek (1995) mencionam que o FCM oferece melhor desempenho para m no intervalo de 1,5 – 2,5.

A homogeneidade dos grupos é obtida a partir do algoritmo *Fuzzy c – Means* (FCM) e testada através de medidas de heterogeneidades de Hosking e Wallis (1993). Por um lado, quando o conjunto inteiro de cem cisternas foi considerado como um único grupo, este grupo tornou-se bastante heterogêneo. Por outro lado, quando o número de grupos *c* aumenta além de um, o algoritmo apresenta grupos que são relativamente homogêneos. Além disso, o tamanho dos grupos diminui com o aumento do número de grupos. Assim, neste estudo, o número máximo de grupos foi fixado pela matriz do grau de pertinência das cisternas em cada grupo, e os resultados obtidos para c não superior a 6 são apresentados e discutidos em seguida.

O número ideal de grupos no conjunto de dados de cisterna é identificado através das medidas de validação do agrupamento *fuzzy*. O índice de Xie e Beni (V_{XB}) indica como número ótimo de grupos (Tabela 32) c = 2 para m = 1,5; 1,6; 1,7; 2,2; 2,3 e 2,4, indica c = 3 para m = 1,4, indica c = 4 para m = 2,5 e c = 5 para m = 1,8; 1,9; 2,0 e 2,1. As medidas de heterogeneidade de Hosking e Wallis mostram que para c = 5, isto é, quando o número de grupos é maior, o algoritmo de agrupamento (FCM) fornece grupos mais homogêneos. A escolha de um número maior de grupos, no caso c = 5, permite uma melhor comparação dos resultados, como definido ante-

riormente. Os valores ótimos do índice de Xie e Beni sugerem que a melhor escolha de m está no intervalo 1,8 – 2,1, ou seja, o número de grupo indicado pelo algoritmo (FCM) é c = 5 para m na vizinhança de 2,0.

Tabela 32. Comparação das medidas de validação dos grupos para o conjunto de dados relativos às cisternas de placas na região do Pajeú em Serra Talhada – PE

c	m	V_{PC}	V_{PE}	V_{XB}	c	m	V_{PC}	V_{PE}	V_{XB}
2	1,4	0,932	**0,123**	0,315	2	2,0	**0,753**	**0,400**	0,265
3		**0,939**	**0,123**	**0,209**	3		0,628	0,662	0,389
4		0,834	0,312	0,575	4		0,565	0,831	0,345
5		0,863	0,256	0,367	5		0,582	0,847	**0,232**
6		0,854	0,290	0,570	6		0,545	0,962	0,318
2	1,5	**0,903**	**0,174**	**0,311**	2	2,1	**0,729**	**0,432**	0,255
3		0,829	0,311	0,438	3		0,595	0,714	0,378
4		0,783	0,409	0,540	4		0,529	0,895	0,312
5		0,818	0,352	0,350	5		0,542	0,929	**0,209**
6		0,803	0,403	0,527	6		0,504	1,050	0,282
2	1,6	**0,872**	**0,226**	**0,304**	2	2,2	**0,707**	**0,460**	**0,244**
3		0,789	0,388	0,435	3		0,566	0,760	0,369
4		0,736	0,503	0,499	4		0,497	0,951	0,281
5		0,771	0,454	0,329	5		0,505	1,001	0,187
6		0,748	0,523	0,483	6		0,467	1,128	0,248
2	1,7	**0,840**	**0,276**	**0,295**	2	2,3	**0,688**	**0,484**	**0,234**
3		0,746	0,465	0,424	3		0,539	0,800	0,361
4		0,690	0,594	0,458	4		0,469	1,000	0,253
5		0,722	0,559	0,306	5		0,424	1,161	0,314
6		0,693	0,643	0,437	6		0,396	1,279	0,262
2	1,8	**0,809**	**0,322**	0,285	2	2,4	**0,670**	**0,505**	**0,224**
3		0,704	0,537	0,412	3		0,516	0,835	0,353
4		0,646	0,680	0,419	4		0,444	1,044	0,227
5		0,674	0,661	**0,281**	5		0,400	1,209	0,287
6		0,640	0,758	0,394	6		0,371	1,333	0,234
2	1,9	**0,780**	**0,363**	0,275	2	2,5	**0,655**	**0,524**	0,214
3		0,665	0,603	0,400	3		0,495	0,865	0,347
4		0,603	0,759	0,380	4		0,422	1,082	**0,204**
5		0,627	0,758	**0,256**	5		0,378	1,251	0,259
6		0,591	0,865	0,355	6		0,349	1,380	0,209

Os valores em negrito denotam valores ótimos das medidas de validação.
$V_{PC:}$ coeficiente de partição; V_{PE}: partição entropia; V_{XB}: índice de Xie e Beni; m: parâmetro de fuzificação; c: número de grupos.

O coeficiente de partição (V_{PC}) e partição entropia (V_{PE}) indicam c = 2 como melhor partição, isto é, sugerem dois grupos como resultado ótimo e são ineficazes. Em geral, V_{PC} é maximizado, e V_{PE} é minimizado para c = 2, independentemente do

valor que o parâmetro de fuzificação m assume (Tabela 32). Isso ocorre porque essas duas medidas de validação necessitam de uma relação direta com alguma propriedade dos dados. Em conformidade com Xie e Beni (1991) e Halkidi *et al.*(2001), nota-se que, à medida que o coeficiente de partição decresce monotonamente, com o aumento do número de agrupamentos, a partição entropia apresenta um crescimento monótono à medida que o número de grupos aumenta.

Segundo Bargaoui *et al.* (1998) e Hall e Minns (1999), ambas as medidas frequentemente indicam como resultado c = 2, sendo uma partição ótima. No caso dos dados das cisternas de placas na região do Pajeú – PE, a tendência monótona é claramente observada na partição entropia (V_{PE}) para valores de m superior a 1,9 (Tabela 32). O índice de Xie e Beni (V_{XB}) não exibe nenhuma tendência monofônica, e, portanto, é eficaz na identificação da partição ótima ou dos grupos formados pelas cisternas de placas utilizadas no estudo da qualidade da água na região do Pajeú – PE.

De acordo com Xu e Wunsch (2005), se o parâmetro de fuzificação m é definido em m = 2, o grau de pertinência das observações ou cisternas de um determinado grupo é obtido unicamente em função das razões entre as distâncias entre o objeto e os centros de grupos.Porém, se o valor do parâmetro de fuzificação m é diferente de 2 ($m \neq 2$), observa-se que existe uma alteração na influência das relações entre as distâncias dos dados aos centros dos grupos.

Quando o parâmetro de fuzificação m cresce, isto é, para ($m \rightarrow \infty$), o grau de pertinência não é calculado mais em função das distâncias entre os dados os centros dos grupos, e sim em função do número de grupos c. Observa-se que, para altos valores de m, os vetores protótipos que compõem os elementos da matriz de atualização dos dados tendem a se aproximar do centro do conjunto de dados, ou seja, para valores altos do parâmetro de fuzificação m, o algoritmo (FCM) apresenta resultados com grupos menos definidos. Portanto, o “melhor” valor que m pode assumir no algoritmo (FCM) é aquele que minimiza a movimentação da matriz de atualização.

4.4.2 K – *Means*

Os métodos não hierárquicos baseados em partição são os mais utilizados, sendo o método de *k* – médias ou (*k* – *means*) um dos métodos obtidos por partição que se tornou consagrado nas últimas décadas. A técnica de agrupamento não hie-

rárquico k – média busca particionar os objetos ou cisternas com suas respectivas variáveis de qualidade da água em *k* grupos (G_1, G_2, ... , G_k), em que G_i denota os grupos de cisternas em *k* grupos, minimizando algum critério numérico que, resultando em valores baixos, dá indícios de bons resultados.

A implementação mais comum do método de agrupamento não hierárquico de *k* – médias é aquela que busca uma partição das cem cisternas em *k* grupos como observado em Joshua *et al.* (2012), minimizando a soma dos quadrados dos desvios dentro dos grupos. Uma maneira simples e clara para a utilização do método de *k* – média é observado na Figura 34, nos dados das cisternas como um todo, isto é, quando se plota cada uma das variáveis contra as outras.

Inicialmente vamos observar a dispersão dos dados no Scatterplot apresentado na Figura 34. O Scatterplot sugere que pelo menos uma das cisternas é considerada diferente das outras em relação à sua qualidade de água, em especial a taxa de coliformes fecais. As cisternas são facilmente identificadas, ou seja, as cisternas (C64, C65, C66, C71, C79, C100) apresentam taxas “elevadas” de coliformes fecais. Verificou-se ainda que as outras variáveis de qualidade da água apresentaram taxas elevadas nessas seis cisternas. E observou-se claramente que essas cisternas apresentam taxas elevadas em relação à maioria das variáveis de qualidade da água.

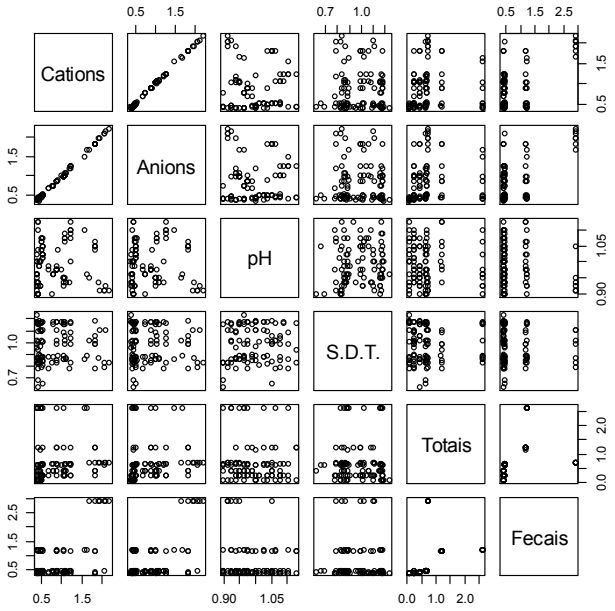

Figura 34. *Scatterplot* da matriz de dados de cisternas de placas da região do Pajeú – PE

Uma vez que os dados foram padronizados para a correção de escala de medida das variáveis, as suas variâncias são muito semelhantes, e assim prosseguimos no processo de agrupamentos. Primeiro, será realizada a soma de quadrado dentro dos grupos, que varia de um a seis grupos, como sugerido pelos métodos de agrupamento anteriores, cujo objetivo é obter a indicação do número mais plausível de grupos. Esta solução é mostrada na Figura 35, e os alto mais significativo na curva da Figura 35 ocorre quando se passa de 2 para 3 grupos, a partir do que se pode concluir que o agrupamento ótimo se verifica na formação de três grupos, em conformidade com o estudo de Kassomenos *et al.* (2010).

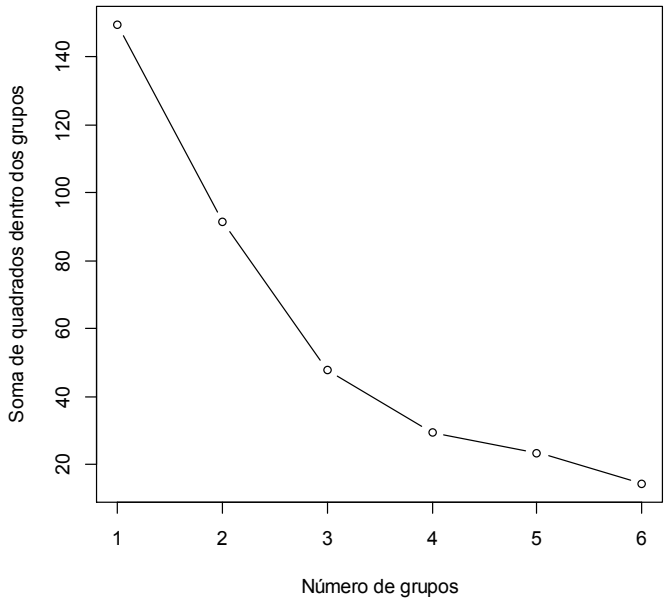

Figura 35. Soma de quadrados dentro dos grupos para diferentes grupos, usando o método de *k* – médias para os dados de cisternas de placas na região do Pajeú – PE.

Observa-se a existência de três grupos, que foram obtidos pelo método de agrupamento não hierárquico de *k* – média. Formaram-se os grupos 1, 2 e 3, conforme observado na (Tabela 33), com as respectivas cisternas.

Tabela 33. Grupos de cisternas obtidos por meio do método não hierárquico de *k* – média

Grupos	n	Cisternas
Grupo 1	28	C12 C18 C19 C23 C25 C34 C36 C37 C39 C40 C42 C43 C44 C45 C46 C47 C48 C52 C56 C69 C70 C74 C78 C80 C81 C83 C85 C87
Grupo 2	23	C14 C15 C27 C28 C49 C50 C53 C54 C57 C58 C61 C64 C65 C66 C71 C72 C73 C79 C88C89 C90 C98 C100
Grupo 3	49	C01 C02 C03 C04 C05 C06 C07 C08 C09 C10 C11 C13 C16 C17 C20 C21 C22 C24 C26 C29 C30 C31 C32 C33 C35 C38 C41 C51 C55 C59 C60 C62 C63 C67 C68 C75 C76 C77 C82 C84 C86 C91 C92 C93 C94 C95 C96 C97 C99

Nesse método, não se sabe inicialmente o número de grupos, assim a indicação do número ótimo de grupos foi sugerida pela soma de quadrados dentro dos grupos ou pelo coeficiente de fusão,em conformidade com os estudos de Kassomenos *et al.*(2010). Observa-se ainda uma boa alocação das cisternas nos três grupos feita pelo método de *k* – média.

Esse método por ser interativo nos leva a crer que esses três grupos deveriam ter aproximadamente o mesmo número de cisternas, ou seja, 33 ou 34 cisternas em cada grupo. Como isso não aconteceu, evidentemente o método de *k* – média agrupou da maneira mais pertinente a que as características de cada cisterna poderiam se assemelhar dentro de um mesmo grupo e diferente dos outros dois grupos. Isso pode ser observado no Scatterplot apresentado na Figura 36, na qual os grupos aparecem em diferentes cores para cada variável de qualidade da água mensurada nas cisternas em estudo.

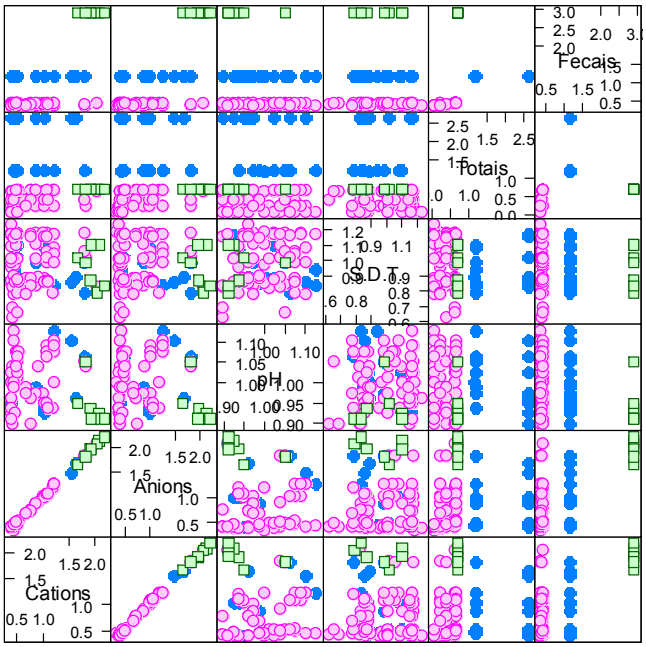

Figura 36. Scatterplot da matriz de dados de cisternas de placas da região do Pajeú – PE, com os agrupamentos obtidos pelo método de *k* – média com 10 simulações iniciais

O resultado não é totalmente satisfatório em decorrência do número de cisternas nos grupos. Essa classificação foi realizada pelo método de *k*– média, respeitando-se a similaridade de cada cisterna aos grupos, e não se sabe de fato uma maneira que melhor atribua cada cisterna a cada grupo. Portanto, o método não hierárquico de agrupamento de *k* – média realizou um agrupamento das cisternas que não gerou nenhum grupo com poucas cisternas ou *outliers*, dando indícios de uma classificação parcimoniosa, do ponto de vista de similaridade ou homogeneidade das cisternas que estão no mesmo grupo e de dissimilaridade das cisternas que estão nos demais grupos. Isto é, a qualidade da água dessas cisternas atendem a três escalas de medidas diferentes, segundo as variáveis que estão sendo consideradas para este estudo.

4.4.3 K – *Medoid*

O método de agrupamento não hierárquico de *k* – médias, para realizar a tarefa para a qual é destinado,isto é, agrupar, precisa de acesso à matriz de dados e uso da distância euclidiana. Porém, em se tratando de métodos não hierárquicos de a-

grupamento, existe um outro método semelhante ao *k* – médias, diferenciando-se apenas por utilizar um elemento do grupo que representa os demais, sendo um ponto de referência. Este método é conhecido como critério de k – *medoid* proposto por Vinod (1969).

O número de grupos sugerido pela soma de quadrados dentro dos grupos para diferentes grupos (Figura 35) são três grupos, e, para cada um desses grupos, obteve-se um elemento representativo ou cisterna que foi o ponto de referência de cada um dos grupos, que é a média ou medoid dos grupos.Assim considerando k = 3 grupos, como indicado na Figura 35, o algoritmo *pam* sugere como medoids iniciais ou centro dos grupos as cisternas (C75, C64, C53), para os grupos 1, 2 e 3, respectivamente, que estão apresentados na Tabela 34.

Tabela 34. Grupos de cisternas obtidos por meio do método não hierárquico de *k – medoid*

Grupos	n	Cisternas
Grupo 1	81	C01 C02 C03 C04 C05 C06 C07 C08 C09 C10 C11 C12 C13 C15 C16 C17 C18 C19 C20 C21 C22 C23 C24 C25 C26 C27 C28 C29 C30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40 C41 C42 C43 C44 C45 C46 C47 C48C50 C51 C52C55 C56 C59 C60 C62 C63 C67 C68 C69 C70 C74 **C75** C76 C77 C78 C80 C81 C82 C83 C84 C85 C86 C87C91 C92 C93 C94 C95 C96 C97 C99
Grupo 2	9	C14 C49 **C64**C65 C66 C71 C79 C89 C100
Grupo 3	10	**C53** C54C57 C58 C61 C72 C73 C88 C90 C98

Com base na Tabela 34, observa-se que as cisternas C75, C64 e C53, que estão em negrito, representam os medoids de cada grupo. Verificou-se que oitenta e uma cisternas são alocadas no grupo 1, pois estão mais próximas do medoid C75, o grupo 2, por sua vez, alocou nove cisternas que estão mais próximas do medoidC64, já o grupo 3 classificou dez cisternas que estão próximas ao medoid C53 e, portanto, não aos outros dois grupos por estarem mas distantes dos seus medoids.

A medida das estatísticas da silhueta calculada para cada grupo representa a qualidade dos grupos encontrados. Quanto maior esse valor, melhor a qualidade dos agrupamentos, assim como proposto por Albalate *et al.* (2011). Logo as estatísticas da silhueta dos grupos 1, 2 e 3 foram 0,63, 0,62 e 0,67, respectivamente, enquanto a estatística média da silhueta foi de 0,64, como observado na Figura 37. Esses valo-

res dão fortes indícios de que foram obtidos bons agrupamentos das cisternas, segundo a similaridade ou semelhança em suas qualidades de água.

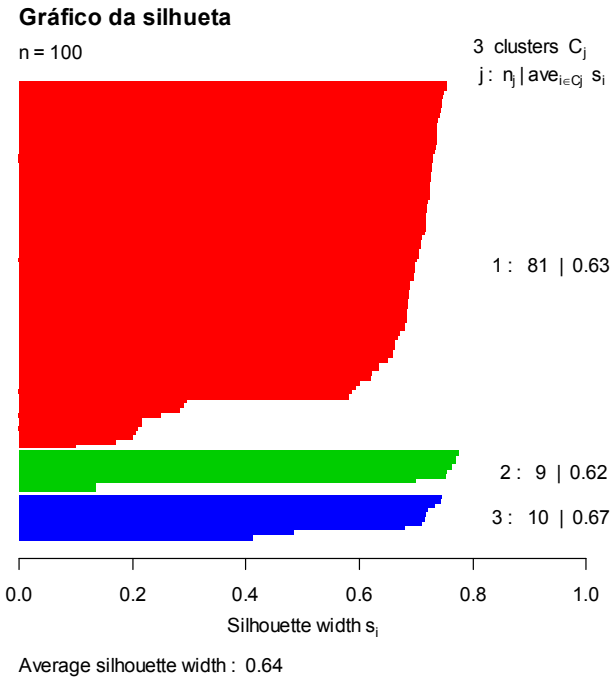

Figura 37. Gráfico e estatística da silhueta obtidos pelo método de K – medoid

De acordo com Kaufman e Rousseau (1990), *PAM – Partioning Around Medoids* difere do método de *k* – média ou *k* – *means* no que diz respeito à escolha dos k representantes de cada grupo e na função a ser minimizada. Os k representantes dos grupos são escolhidos entre os indivíduos observados. Na Figura 38, observa-se a representação gráfica do algoritmo *pam* utilizado pelo método de *k* – medoid na obtenção de três grupos formados pelas cisternas de placas da região do Pajeú – PE.

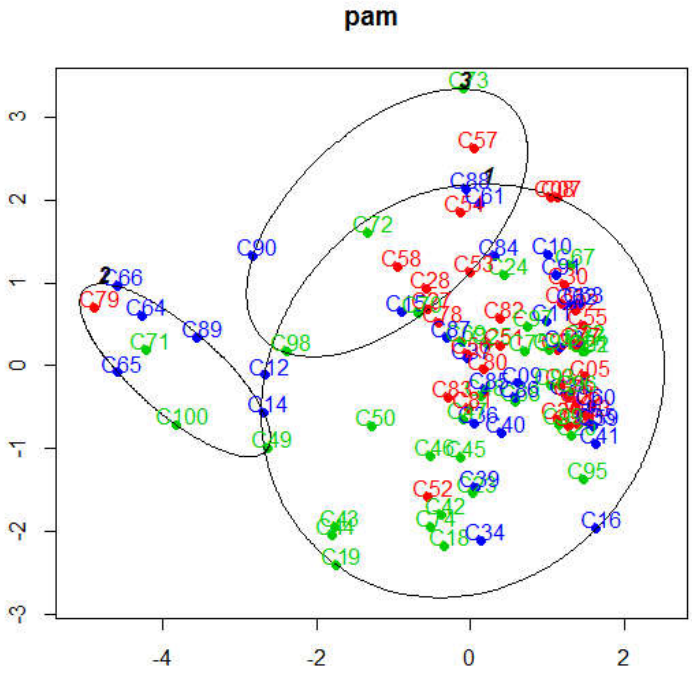

Figura 38: Agrupamentos obtidos pelo método do *k – medoid* utilizando *pam.*

Observa-se a formação de três grupos – 1, 2 e 3 – representados pelas cores vermelha, azul e verde, respectivamente. Porém, como visualizado na Figura 38, a cisterna C90, por exemplo, está no grupo 3, sendo vizinha próximo do grupo 2, por isso, ela aparece na cor azul e na borda do grupo 3. Também a cisterna C100 está no grupo 2, mais com uma vizinhança considerável de pertencimento ao grupo 3, e assim por diante, como observado por Docampo *et al.* (2013), em uma Análise de Componentes Principais. Sendo assim as cisternas que estão nos seus devidos grupos e com as cores originais do grupo seriam candidatas a medoid deste grupo, mas não fortes o suficiente para representar o grupo como as cisternas que os representaram, como citado anteriormente.

Portanto, a qualidade da água das cisternas que estão no grupo 1 deve representar um padrão de qualidade de água das cisternas de placas da região do Pajeú – PE, uma vez que oitenta e uma cisternas ou 81% dessa amostra estão com qualidade da água similar por estarem em um mesmo grupo. Com relação aos grupos 2 e 3,não é descartável a possibilidade dos mesmos representarem a qualidade da água dessa região, haja vista que a qualidade do agrupamento desses grupos está bastante satisfatória, quando considerado como critério de suas validações as estatísticas da silhueta.

5 CONCLUSÕES

A utilização dos métodos multivariados de análise de agrupamento podem contribuir de maneira muito significativa com o pesquisador na formação dos grupos de cisternas de placas, com base em uma ou mais variáveis de qualidade da água ora analisada. A obtenção da “melhor” solução foi realizada observando-se a qualidade dos agrupamentos através da comparação da variância interna de cada grupo e do índice da estatística da silhueta. A utilização de diferentes métodos de agrupamento hierárquico acarretaram uma maior confiabilidade nos resultados obtidos, no entanto, esses resultados sugeriram grupos de tamanho distintos.

As técnicas multivariadas de análise de agrupamento proposto responderam questionamentos a respeito de variáveis de qualidade da água em regiões semiáridas do Nordeste Brasileiro, como a região do Pajeú de Pernambuco.

Além dos métodos hierárquicos de análise de agrupamento, neste trabalho também foi proposta a utilização de métodos não hierárquicos, como é o caso do agrupamento Fuzzy usado para a análise de agrupamento. Este método utilizou a matriz de grau de pertinência que alocou cada cisterna para o grupo no qual ela obteve a mais elevada pertinência. Assim, possibilitou a utilização do método de agrupamento não hierárquico Fuzzy, cujos resultados se mostraram satisfatórios.

A proposta de utilizar e comparar os métodos hierárquicos e os métodos não hierárquicos em análise de agrupamento mostrou-se eficiente, possibilitando, dessa forma, a obtenção do “melhor” número de grupos como sugerido pela literatura, além da obtenção de outros agrupamentos que poderiam ser utilizados no estudo. Logo, o pesquisador tem uma maior autonomia na escolha do número de grupos, podendo considerar como número “ótimo” de grupos em determinada situação aquele sugerido pelos métodos hierárquicos, pelos métodos não hierárquicos ou por uma combinação de ambos os métodos.

Dentre todos os métodos de agrupamentos hierárquicos utilizados com a distância euclidiana, os que indicaram melhor qualidade no agrupamento, segundo os índices da estatística da silhueta, foram os métodos do centroide e da mediana, pois estes métodos indicam que a qualidade da água das cisternas alocadas em cada grupo é similar acerca das variáveis analisadas, no entanto, estes dois métodos de agrupamento são penalizados por produzirem grupos muito distintos, ou seja, um

grupo tem um grande número de cisternas, e outros grupos têm apenas duas cisternas.

O método da ligação simples utilizando a distância euclidiana apresentou o resultado mais insatisfatório segundo a estatística da silhueta, indicando que os grupos formados por esse método podem não ter classificado cisternas com qualidade da água similar o quanto os grupos produzidos por outros métodos de agrupamento.

O método de Ward utilizando a distância euclidiana produziu o melhor resultado, quando considerado que nenhum dos grupos ficou apenas com duas cisternas, nem algum dos grupos ficou com uma “grande” quantidade de cisternas. Os métodos de agrupamento hierárquicos utilizando a distância euclidiana produziram, em sua maioria, grupos com duas cisternas, que, em geral, foram as cisternas C14 e C49, podendo esses grupos ser *outliers,* e as cisternas que foram neles classificadas possuírem água imprópria para o consumo humano.

Os métodos hierárquicos de agrupamento utilizando a distância euclidiana, o método do centroide e o método da mediana foram os que produziram as melhores soluções, quando considerado o desvio padrão interno de cada grupo, ou seja, esses métodos obtiveram grupos mais homogêneos, sugerindo que as cisternas que foram alocadas em cada um desses grupos apresentam qualidade da água similar em relação às variáveis mensuradas.

Quanto aos métodos de agrupamentos hierárquicos utilizando a distância ou métrica de camberra, os que obtiveram melhor qualidade nos agrupamentos, segundo os índices da estatística da silhueta, foram os métodos de Ward e de mcquitty, isto é, esses métodos dão indícios de que a qualidade da água das cisternas alocadas em cada grupo é similar em relação às variáveis analisadas. Além disso, esses dois métodos produziram bons resultados no que tange ao tamanho dos grupos. Evidentemente pode não ter sido o “melhor” resultado, no entanto, formaram grupos mais parcimoniosos no sentido de não haver grupos com muitas ou poucas cisternas, como em outros métodos de agrupamento.

No que concerne à heterogeneidade dos grupos, os métodos hierárquicos de agrupamento utilizando a métrica de camberra apresentaram basicamente as mesmas características, isto é, não se observou em nenhum dos métodos a formação de grupos mais homogêneos, assim como não ocorreram grupos heterogêneos em nenhum dos métodos. Verificou-se pelo menos um dos grupos heterogêneos ou

homogêneos em todos os métodos de agrupamento hierárquicos utilizando a métrica de camberra.

O agrupamento Fuzzy sugeriu a formação de seis grupos de acordo com a matriz do grau de pertinência das cisternas de cada grupo. No entanto, levando em consideração o índice de Xie e Beni, a indicação foi de cinco grupos, o que se tornaria mais plausível por ser um índice de validação.

O método de *k* – média ou *k* – *means* sugeriu a formação de três grupos, de acordo com o Scatterplot das variáveis de qualidade da água ora estudadas, o que foi confirmado pela soma de quadrados dentro dos grupos, ou gráfico de fusão ou cotovelo referenciado na literatura, no qual se observa que esses agrupamentos são parcimoniosos, segundo o número de cisternas alocadas em cada grupo.

Dentre todas as cisternas analisadas, as cisternas C64, C65, C66, C71, C79, C100, agrupadas pelo método não hierárquico de *k* – médias, apresentam taxas "elevadas" de coliformes fecais, ou seja, essas cisternas devem ser acompanhadas pelo pesquisador com mais atenção, pois as águas nelas armazenadas podem causar danos à saúde das pessoas que a consomem. Ainda se verificou que as outras variáveis de qualidade da água estudadas apresentaram taxas elevadas nessas seis cisternas.

O método não hierárquico de *k* – *medoid* formou três grupos, que, apesar de apresentarem uma discrepância em relação ao número de cisternas alocadas em cada grupo, em especial o grupo 1, que classificou muitas cisternas em relação aos grupos 2 e 3, a qualidade desses agrupamentos foi satisfatória segundo o índice da estatísticas da silhueta. Nesses grupos, os medoids foram as cisternas C75, C64 e C53, respectivamente, isto é, essas cisternas têm qualidade da água similar às cisternas dos seus grupos, uma vez que representam as cisternas de cada grupo e são dissimilares de outros grupos.

Os métodos de agrupamento hierárquicos utilizados nesta tese formaram seis grupos, assim como o método não hierárquico de agrupamento Fuzzy. Já os métodos não hierárquicos de K – means e K – *medoid* formaram apenas três grupos.

A metodologia proposta nesta tese pode contribuir com o comportamento da qualidade da água na região do Pajeú pernambucano, reduzindo a incidência de cisternas com água poluída, uma vez que medidas cabíveis podem ser tomadas nas cisternas que forem detectadas e consideradas com qualidade da água imprópria para o consumo humano, evitando ainda a reincidência do problema.

Portanto, os métodos de análise de agrupamento hierárquicos e não hierárquicos podem de forma complementar ser uma maneira eficiente de monitorar a qualidade da água de cisternas de placas na região do Pajeú – PE ou em qualquer outra região.

Como prosseguimento deste trabalho, sugerimos utilizar outras metodologias multivariadas no problema de agrupamento de cisternas, de placas e outros reservatórios, considerando variáveis de qualidade da água e outras variáveis de interesse, assim despertando novos sensos críticos nos pesquisadores em relação a tais temáticas e contribuindo para a sofisticação e o enriquecimento dos resultados de natureza multivariada relacionados a esse fenômeno.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ABUZAID, A. H.; MOHAMED, I. B.; HUSSIN, A. G. Boxplot for circular variables. **Comput Stat,** v. 27, p. 381 – 392, 2012.

ALBALATE, A.; SUENDERMANN, D.; MINKER, W. On cluster validation for detecting the number of clusters in a data set. **International Journal on Artificial Intelligence Tools**, v. 20, n. 5, p. 941 – 953, 2011.

ALBUQUERQUE, M. A. **análise de agrupamento hierárquica e incremental-estudo de caso em ciências florestais.** 1v. 160f. Tese (Doutorado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Recife, 2013.

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.

ANDERSON, T. W. **An introduction to multivariate statistical analysis.** California; John Wiley & Sons, 2003. 721 p.

ARBELAITZ, O.; GURRUTXAGA, I.; MUGUERZA, J.; PÉREZ, J, M.; PERONA, I. An extensive comparative study of cluster validity indices. **Pattern Recognition,** v. 46, p. 243 – 256, 2013.

APHA Standard Methods for the examination of water wastewater, 19ª Ed. Washington, American Public Health Association/AWWA/WEF. 1995.

ARMSTRONG, J. J.; ZHU, M.;HIRDES, J. P.; STOLEE, P. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. **Archives of Physical Medicine and Rehabilitation,** p. 2198 – 2205, 2012.

ATKINSON, A. C.; RIANI, M. The Forward Search and Data Visualisation. **Computational Statistics,** v. 19,p. 29 – 54, 2004.

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

BARGAOUI, ZK.; FORTIN, V.; BOBÉE, B.; DUCKSTEIN, L. A fuzzy approach to the delineation of region of influence for hydrometric stations. Revue des sciences de l'eau v. 11 n. 2, p. 255 – 282, 1998.

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.

BAXTER, M. J. Detecting Multivariate Outliers in Artefact Compositional Data. **Archaeomerry,** The Nottingham Trent University, Clifon Campus, v. 41, n. 2, p. 321 – 338, 1999.

BERGE, A. C. B.; ATWILL, E. R.; SISCHO, W. M. Assessing antibiotic resistance in fecal Escherichia coli in young calves using cluster analysis techniques. **Preventive Veterinary Medicine,** v. 61, p. 91 – 102, 2003.

BEST, D. J.; RAYNER, J. C. W. a test for bivariate normality. **Statistics and Probability Letters**, North-Holland, v. 6, p. 407 – 412, 1988.

BEZDEK, J. C., TSAO, E. C. E PAL, N. R. Fuzzy kohonen clustering networks. In *IEEE International Conference on Fuzzy Systems*, p. 1035 – 1043, 1992.

BEZDEK, J. C. Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology,v.1, p. 57–71, 1974.

BEZDEK, J. C. Cluster validity with fuzzy sets. Journal of Cybernetics, v.3 n. 3, p. 58 – 72, 1974.

BORYSOV, P.; HANNIG, J.; MARRON, J. S. Asymptotics of hierarchical clustering for growing dimension. Journal of Multivariate Analysis: v. 142, pp. 465 – 469, 2014.

BOX, G, E. P.; COX, D. R. An Analysis of Transformations. Journal of the Royal Statistical Society, v. 26, n. 2, pp 211 – 252, 1964.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis – Forecasting and Control.**3. ed. New Jersey: Prentice Hall, 1994.

BRITES, R. S.; SOARES, V. P.; RIBEIRO, C. A. A. S. Comparação de Desempenho entre Três Índices de Exatidão Aplicados a Classificações de Imagens Orbitais. INPE, p. 813 – 821, 1996.

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.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D.**Introdução a análise de agrupamento**. São Paulo: Associação Brasileira de Estatística, 1990. 105p.

CAMPBELL, J. B. Introduction to remote sensing. New York, The Guilford Press, 1987. 551 p.

CAMPELLO, R. J. G. B. A Fuzzy Extension of the Rand Index and Other Related Indexes for Clustering and Classification Assessment. **Pattern Recognition Letters,** v. 28, n. 7, p. 833 – 841, 2007.

CAO, YONG.; WILLIAMS, D. D.; WILLIAMS, N, E. Data Transformation and Standardization in the Multivariate Analysis of River Water Quality. Ecological Applications: v. 9, n. 2, pp. 669 – 677, 1999.

CARPINETO, C.; ROMANO, G. Consensus Clustering Based on a New Probabilistic Rand Index with Application to Subtopic Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 34, n. 12, p. 2315 – 2326, 2012.

CHAVENT, M.; LECHEVALLIER, Y.; BRIANT, O. DIVCLUS-T: A monothetic divisive hierarchical clustering method. **Computational Statistics & Data Analysis,** v. 52, p. 687 – 701, 2007.

CHERWIN, K.; KNAPP, A. Unexpected patterns of sensitivity to drought in three semi-arid grasslands. Global Change Ecology - Original Research, 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 of genetics,** v. 2, n. 88, p. 1 – 11, 2011.

COID, J.; FREESTONE, M.; ULLRICH, S. Subtypes of psychopathy in the British household population: findings from the national household survey of psychiatric morbidity. **Soc Psychiatry Psychiatr Epidemiol,** p. 879 – 891, 2012.

CONAMA – Conselho Nacional do Meio Ambiente.Resolução CONAMA nº 357/05. Disponível em http://www.mma.gov.br/port/conama/res/res05/res35705.pdf. Acesso em 06/05/2014. n. 53, pp. 58 – 63, 1986.

CORMACK, R. A review of classification. Journal of the Royal **Statistical Society** (Series A), v. 134, p. 321 - 367, 1971.

CUADRAS, C. M. Models Estadístics Multivariants. Apostila. Barcelona, 2006. 249p.

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.

DOCAMPO, E.; COLLADO, A.; ESCARAMIÍS, G.; CARBONELL, J.; RIVERA, J. et al. Cluster Analysis of Clinical Data Identifies Fibromyalgia Subgroups. **PLoS ONE,** Iran University of Medical Sciences, Iran (Republic of Islamic), p. 1 – 8, 2013.

DORLING, S. R.; DAVIS, T. D. Extending cluster analysis-synoptic meteorology links to characterize chemical climates at six northwest European monitoring stations. **Atmos Environ**, v. 29, p. 145–167, 1995.

DOVOEDO, Y. H.;CHAKRABORTI, S. Boxplot-Based Outlier Detection for the Location-Scale Family. Communications in Statistics: Simulation and Computation: v. 44, n. 6, pp. 1492 – 1513, 2015.

DUDOIT, S.; FRIDLYAND, J. A prediction-based resampling method for estimating the number of clusters in a dataset. **Genome Biology,**v. 3, n. 7, p. 1 – 21, 2002.

ESTEVES, F. A. **Fundamentos de Limnologia** - 2a Ed. Rio de Janeiro, Interciência/INEP, 1998, 575p.

FALASCONI, M.; PARDO, M.; VEZZOLI, M.; SBERVEGLIERI, G. Cluster validation for electronic nose data. ScinceDirect, v. 125, n. 2, p. 596 – 606, 2007.

FAN L.; LIU G.; WANG F.; GEISSEN V.; RITSEMA, C. J. Factors Affecting Domestic Water Consumption in Rural Households upon Access to Improved Water Supply: Insights from the Wei River Basin, China: v. 8, n. 8, pp. 1 – 10, 2013.

FENG, S. et al. Esophageal cancer detection based on tissue surface-enhanced Raman spectroscopy and multivariate analysis. **Applied Physics Letters 102,** p. 1 – 5, 2013.

FERNANDEZ, E.; NAVARRO, J.; BERNAL, S. Handling multicriteria preferences in cluster analysis. **European Journal of Operational Research**, n. 202, p. 819 – 827, 2010.

FERREIRA, D. F. Estatística Multivariada. 1ª. ed, Lavras: Ed. UFLA, 2008, 662 p.

FILZMOSER, P.; HRON, K.; REIMANN, C. Interpretation of multivariate outliers for compositional data. **Computers and Geosciences,** v. 39, p. 77 – 85, 2012.

FUKUOKA, Y.; LINDGREN, T. G.; RANKIN, S. H.; COOPER, B. A.; CARROLL, D. L. Cluster analysis: A useful technique to identify elderly cardiac patients at risk for poor quality of life. Quality of Life Research, n. 16, pp. 1655 – 1663, 2007.

GAMA M. de P. Bases da análise de agrupamentos ("Cluster Analysis"). Brasília: UnB, 1980. 229f. Dissertação (Mestrado em Estatística e Métodos Quantitativos) - Universidade de Brasília, 1980.

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

GRANATO, D.; CALADO, V. M. A.; JARVIS, V. Observations on the use of statistical methods in Food Science and Technology. **Food Research International,** v. 55, p. 137 – 149, 2014.

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

HAIR, J. F. et al. **Multivariate Data Analysis**. 7. ed. Pearson Prentice Hall, 2010. 593 p.

HAIR, J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R. E. Multivariate data analysis. Prentice-Hall, Englewood Cliffs-NJ, v. 7, 2010.

HALL, M. J.; MINNS, A. W. The classification of hydrologically homogeneous regions. Hydrological Sciences Journal, v. 44, n. 5, p. 693–704, 1999.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. Journal of Intelligent Information Systems, v. 17, p. 107–145, 2001.

HAMMER, M. J. **Sistemas de Abastecimento de Água e Esgotos**. Livros Técnicos e Científicos, Editora S.,A.., SP, 1979, 561p.

HARDIN, J.; ROCKE, D. M. The distribution of robust distances. **Journal of Computational and Graphical Statistics,** v. 14, p. 928–946, 2005.

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

HITCHINS, A.D.; HARTMAN, P.A.; TODD, E.C.D. Compendium of methods for the microbiological examination of foods: Coliforms-Escherichia coli and its toxins. 3.ed. Washington: American Public Health Association, 1996.

HOSKING, J.R. M.; WALLIS, J.R. Some statistics useful in regional frequency analysis. Water Resources Research (Correction: Water Resources Research v. 31 n. 1,p.251,1995),v. 29 n. 2, p. 271–281, 1993.

HUBERT, L.; ARABIE, P. Comparing Partitions. Classification. v. 2, n. 1, p. 193 – 218, 1985.

HUR, A.; ELISSEEFF, A.; GUYON, I. A Stability Based Method for Discovering Structure in Clustered Data. *Proc. Pacific Symp. Biocomputing,* p. 6 – 17, 2002.

IRAWAN, D. E.; PURADIMAJA, D. J.; NOTOSISWOYO, S.; SOEMINTADIREDJA, P. Hydrogeochemistry of volcanic hydrogeology based on cluster analysis of Mount Ciremai, West Java, Indonesia. **Journal of Hydrology**, p. 221–234, 2009.

JAIN, A.; DUBES, R. Algorithms for Clustering Data, Prentice Hall, Chapter 4, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys,** New York, v. 31, n. 3, p. 265 – 323, 1999.

JALALI, M. Groundwater geochemistry in the Alisadr, Hamadan. Environ Monit Assess, Western Iran. 2009.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 4 ed. New Jersey: Prentice Hall, 1992.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis.4ed. New Jersey: Prentice Hall, 1998, 816 p.

JOHNSON, R. A.; WICHERN, D. W. Applied Multivariate Statistical Analysis. 6 ed. New Jersey: Upper Saddle River, 2002. 767 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6 ed. New Jersey: Upper Saddle River, 2007. 767 p.

JOSHUA, J. A.; MU, Z.;JOHN, P. H.; PAUL, S. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population, **Arch Phys Med Rehabil,** v. 93, n. 2, pp. 198 – 205, 2012.

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

KAUFMANN, L.; ROUSSEEUW, P. J. Finding groups in data: an introduction to cluster analysis. New York: Jonh Wiley, 1990. 342 p.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. Analyzing Multivariate Data. Canada: 2003, 545 p.

LIMA, ISYS P. de A. Aplicação do controle estatístico de qualidade de água em cisternas instaladas em comunidades na região do Sertão do Pajeú – Semiárido Pernambucano. Recife: Universidade Federal Rural de Pernambuco, 2014. 68p. Dissertação de Mestrado.

LIN, T. C.; LIU, R. S.; CHAO, Y. T.; CHEN, S. Y. Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms. Gene, v. 518, n. 1, p. 159 – 163, 2013.

MACKERT, M.; WALKER, L. Cluster Analysis Identifies Subpopulations for Health Promotion Campaign Design. **Public Health Nursing**, v. 28 n 5, pp. 451–457, 2011.

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.

MCGUIRE, J. F. et al. A cluster analysis of tic symptoms in children and adults with Tourette syndrome: Clinical correlates and treatment outcome. **Psychiatry Research,** p. 1198 – 1204, 2013.

MCROBERTS, R. E. et al. Estimating areal means and variances of forest attributes using the k-nearest neighbours technique and satellite imagery. **Remote Sens. Environ**. v. 111, p. 466 – 480, 2007.

MICHAUD, P. Clustering techniques. **European Centre for Applied Mathematics,** Paris, France, p. 135 – 147, 1997.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 1ª reimpressão. 2007. 297p.

MINGOTI, S. A. Análise de dados atravéz de métodos de estatística multivariada. uma abordagem aplicada. Belo Horizonte: UFMG, 2005 295p.

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.

OHLSON, M.; ANDRUSHCHENKO, Z.; ROSEN D, V. Explicit estimators under *m*-dependence for a multivariate normal distribution. Ann Inst Stat Math, v. 63, pp. 29–42, 2011.

OLIVEIRA, M. R. G.; CANTALICE, J. R. B.; FERREIRA, T. E. A.; CUNHA FILHO, M.; CRUZ, D. V.; FALCÃO, A. P. S. T. Estudo Estatístico do Coeficiente de Escoamento Superficial da Bacia Hidrográfica do Riacho Jacu no sertão do Pajeú – PE: v. 33, n. 3, pp. 277 – 290, 2015.

ORLÓCI, L. **Multivariate analysis in vegetational research**. 2. ed. The Hague: Dr. W. Junk B. V. Publishers, 1978. 451 p.

PAL, N. R.; BEZDEK, J. C. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems v. 3 n. 3, p. 370–379, 1995.

PAL, N, R.; Biswas, J. Cluster validation using graph theoretic concepts. **Pattern Recognition**, v. 30, p. 847 – 857, 1997.

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

PENG, Y.; ZHANG Y.; KOU, G.; SHI, Y. A Multicriteria Decision Making Approach for Estimating the Number of Clusters in a Data Set. **journal. pone,** n. v. 7, n. 7, p. 1 – 9, 2012.

PICARD, N.; MORTIER, F.; ROSSI, V.; FLEURY, S. G. Clustering species using a model of population dynamics and aggregation theory. **Ecological Modelling,** p. 152 – 160, 2010.

PREARO, L. C.; GOUVÊA, M. A.; ROMEIRO, M. C. Avaliação da adequação da aplicação de técnicas multivariadas de dependência em teses e dissertações de algumas instituições de ensino superior. Ensaios FEE, Porto Alegre, v. 33, n. 1, p. 261 – 290, 2012.

QUESSY, J. F.; MAILHOT, M. Asymptotic power of tests of normality under local alternatives. **Journal of Statistical Planning and Inference,** v. 141, p. 2787 – 2802, 2011.

RABAL, H.; CAP, N.; CRIADO, C.; ALAMO, N. Holodiagrams using Mahalanobis distance. Optik. p. 1725 – 1731, 2012.

RAJU, R.; APPRAO, A.; VALLIKUMAI, A. Comparative evaluation of HCM and FCM clustering techniques. **International Journal of Computational Intelligence Research,** v. 7, n. 1, p. 41 – 49, 2011.

RAND, W. M. Objective Criteria for the Evaluation of Clustering Methods. Journal **of the American Statistical Association**, v. 66, p. 846 – 850, 1971.

RAO, C. R. **Advanced statistical methods in biometric research**. New York: John Wiley & Sons, 1952. 390 p.

REIS, E. **Estatística Multivariada Aplicada.** Lisboa: Edições Silabo, 2001. 342 p.

ROSENFIELD, G.H.; FITZPATRICK-LINS, K.A. A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 52, n. 2, p. 223 – 227, 1986.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal Computational Applied Mathematics. V. 20, p. 53–65, 1987.

SAWYER, C. N.; MACCARTY, P. L.; PARKIN, G. F. **Chemistry of Enviromental Engeneering**. 4° Ed., International Student Edition, MacGraw-Hill Book Company, 1994, 858p.

SCHLAMM, A.; MESSINGER, D. An empirical estimate of the multivariate normality of hyperspectral image data. **Rochester Institute of Technology,** p. 1 – 11, 2011.

SCHNEIDER, S.; HUY, C.; SCHUESSLER, M.; DIEHL, K.; SCHWARZ, S. Optimising lifestyle interventions: identification of health behaviour patterns by cluster analysis. **European Journal of Public Health,** v. 19, n. 3, p. 271 – 277, 2009.

SHAW, S. Y.; SHAH, L.; JOLLY, A. M.; WYLIE, J. L. Identifying heterogeneity among injection drug users: A cluster analysis approach. American Journal of Public Health, n. 98, pp. 1430 – 1437, 2008.

SILVA, H. C. M. B. Métodos de Partição e Validação em Análise Classificatória Baseados em Teoria de Grafos. Porto. 2005. 275p. Tese (Doutorado em Matemática Aplicada). Faculdade de Ciências da Universidade do Porto, março de 2005.

SIOU, G. L.; YASUI, Y.; CSIZMADI, L.; MCGREGOR, S. E.; ROBSON, P. J. Exploring Statistical Approaches to Diminish Subjectivity of Cluster Analysis to Derive Dietary Patterns. **American Journal of Epidemiology,** Oxford University, v. 173, n. 8, p. 956 – 967, 2011.

SIQUEIRA, R. S. Manual de microbiologia de alimentos. Brasília: EMBRAPA, 1995. 159p.

SOUTHWORTH, H. Detecting Outliers in Multivariate Laboratory Data. **Journal of Biopharmaceutical Statistics,** v. 18, p. 1178 – 1183, 2008.

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

SUN, J.; BI, J.; CHAN, G.; OSLIN, D.; FARRER, L.; GELERNTER, J.; KRANZLER, H. R. Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. **Addictive Behaviors**, v. 37, p. 1138 – 1144, 2012.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society**, n. 63, p. 411 – 423, 2001.

THEODORIDIS, S.; KOUTROUBAS, K. Pattern Recognition. Academic Press, New York, 1999.

THYNE, G.; GULER, C.; POETER, E. Sequential analysis of hydrochemical data for watershed characterization, ground water. **Dublin**, v. 711, n. 5, p. 1 – 13, 2004.

TODOROV, V.; TEMPL, M.; FILZMOSER, P. Detection of multivariate outliers in business survey data with incomplete information. **Adv Data Anal Classif,** v. 5, p. 37 – 56, 2011.

TUCCI, C. E. M. Hidrologia: Ciência e Aplicação, 3ª edição, Porto Alegre, **Editora da UFRGS/ABRH**, 2004.

VALE, M. N. **Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos.** 120 f. Dissertação (Mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, Rio de Janeiro, 2005.

VIALLE, C.; SABLAYROLLES, C.; LOVERA, M.; JACOB, S.; HUAU, M. C.; VIGNOLES, M. M. Monitoring of water quality from roof runoff: Interpretation using multivariate analysis. **water research,** p. 3765 – 3775, 2011.

VINOD, H. Integer programming and the theory of grouping. *Journal of the American Statistical Association,* n.64, p. 506 – 517, 1969.

XIE, X. L.; BENI, G. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence v. 13,n. 8, p. 841–847, 1991.

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

XU, R.; WUNSCH, D. II survey of clustering algorithms. *IEEE Transactions on Neural Networks,* v. 16, n. 3, p. 645 – 678, 2005.

YOU, S. H.; SEO, J, W. Storm surge prediction using an artificial neural network model and cluster analysis. Natural Hazards: v. 15, n. 1, pp. 97 – 114, 2009.

ZALIK, K. R. Cluster validity index for estimation of fuzzy clusters of different sizes and densities. **Pattern Recognition,** v. 43, p. 3374 – 3390, 2010.

ZHANG, H.; RAO, M. B. Maximum Likelihood Estimation in Linear Models with equi-Correlated random errors. Aust. N. Z. J. Stat, p. 48, v. 1, p. 79 – 93, 2006.

ZHONG, C.; MIAO, D.; WANG, R.; ZHOU, X. DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points. **Pattern Recognition Letters,** v. 29, p. 2067 – 2077, 2008.