RESOLUÇÃO DO PROBLEMA DE AGRUPAMENTO SEGUNDO O CRITÉRIO DE MINIMIZAÇÃO DA SOMA DE DISTÂNCIAS

Vinicius Layter Xavier

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Compuação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Felipe Maia Galvão França

Rio de Janeiro
Março de 2012

RESOLUÇÃO DO PROBLEMA DE AGRUPAMENTO SEGUNDO O CRITÉRIO DE MINIMIZAÇÃO DA SOMA DE DISTÂNCIAS

Vinicius Layter Xavier

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Felipe Maia Galvão França, Ph.D.

_____

Dr. Mario Veiga Ferraz Pereira, D.Sc.

_____

Prof. Carlos Eduardo Pedreira, Ph.D.

_____

Prof. Antonio Alberto Fernandes de Oliveira, D. Sc.

RIO DE JANEIRO, RJ - BRASIL
MARÇO DE 2012

Xavier, Vinicius Layter

Resolução do Problema de Agrupamento Segundo o Critério de Minimização da Soma de Distâncias / Vinicius Layter Xavier. – Rio de Janeiro: UFRJ/COPPE, 2012.

IX, 78 p.: il.; 29,7 cm.

Orientador: Felipe Maia Galvão França

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2012.

Referências Bibliográficas: p. 71-75.

1. *Clustering*, Análise de Agrupamento. 2. Suavização Hiperbólica. 3. Programação Não-Diferenciável. 4 Problema de Fermat-Weber I. França, Felipe Maia Galvão II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Aos meus pais, Adilson e Solange.

AGRADECIMENTOS

Agradeço a Deus, a Jesus e a Francisco de Paula por todas as coisas boas que vivi e aprendi nessa fase da minha vida.

Agradeço aos meus pais, Adilson e Solange, pois foram fundamentais na minha formação, com ensinamentos diários ao longo da minha vida. Sem o apoio deles eu não conseguiria chegar até aqui.

Agradeço o incentivo, carinho e companheirismo dos meus amigos, em especial o meu irmão Leonardo que sempre foi muito amigo e companheiro.

Agradeço ao meu orientador, professor Felipe França, pela orientação, apoio e pela relação de confiança e amizade.

Agradeço a empresa PSR por ter me concedido uma bolsa de mestrado durante o período de um ano.

Agradeço a todos os integrantes da banca pelo aceite de imediato da participação na Banca de Defesa de Tese.

E por fim, a todos aqueles que de alguma forma colaboraram com este trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.sc.)

RESOLUÇÃO DO PROBLEMA DE AGRUPAMENTO SEGUNDO O CRITÉRIO DE MINIMIZAÇÃO DA SOMA DE DISTÂNCIAS

Vinicius Layter Xavier

Março/2012

Orientador: Felipe Maia Galvão França

Programa: Engenharia de Sistemas e Ciência da Computação

Este trabalho considera a resolução do problema de agrupamento correspondente à minimização da soma das distâncias euclideanas das observações aos seus centróides, através do uso da técnica de suavização hiperbólica. A modelagem matemática deste problema leva a uma formulação *min-sum-min* que, além de sua intrínseca natureza bi-nível, tem a importante característica de ser um problema fortemente não-diferenciável e não-convexo, com um grande número de mínimos locais. A estratégia de suavização hiperbólica resolve uma seqüência de subproblemas diferenciáveis de otimização irrestrita de baixa dimensão, que gradualmente se aproxima do problema original. A confiabilidade e a eficiência do método são ilustradas através de um conjunto de experimentos computacionais. Deve-se enfatizar que a metodologia proposta pode ser aplicada para a resolução do problema de localização de Fermat-Weber, que é análogo ao problema tratado.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

RESOLUTION OF THE CLUSTERING PROBLEM ACCORDING TO THE SUM OF DISTANCE CRITERIA

Vinicius Layter Xavier

March/2012

Advisor: Felipe Maia Galvão França

Department: Systems Engineering and Computer Science

This work considers the clustering problem corresponds to the minimization of the sum of the euclidean distances of observations to their cluster centroids. The mathematical modeling of this problem leads to a min-sum-min formulations which in addition to its intrinsic bi-level nature, has the significant characteristic of being strongly non-differentiable and non-convex problem with a large number of local minima. The hyperbolic smoothing strategy solves a sequence of low dimension differentiable unconstrained optimization sub-problems, which gradually approaches the original problem. The reliability and efficiency of the method are illustrated via a set of computational experiments. It must be emphasized that the proposed methodology can be analogously applied to the solving of the Fermat-Weber location problem.

1 – Introdução

1.1 - Análise de agrupamentos

Existe uma crescente necessidade por processos automáticos que executam o particionamento de dados em conjunto de grupos. Por exemplo, as bibliotecas digitais e a internet vêm crescendo exponencialmente e a habilidade de encontrar informações úteis depende dos algoritmos de classificação. Técnicas de Clustering podem ser usadas para descobrir grupos naturais em conjuntos de dados e identificar estruturas abstratas que possam existir, sem ter qualquer conhecimento prévio das características dos dados. (KOGAN, 2007)

Análise de agrupamento ou clustering consiste em agrupar um conjunto de observações de modo que as observações que pertençam a um mesmo grupo sejam parecidas entre si e diferentes das dos demais grupos. Desta forma temos dois princípios básicos da análise de agrupamento que são homogeneidade e separação. Sendo assim, quanto mais homogêneos são elementos dentro de um grupo mais separados ou diferentes são os grupos.

O procedimento de classificação constitui umas das mais básicas atividades culturais da humanidade e é fundamental para a ciência (ANDERBERG, 1973). De acordo com EVERITT e LEESE (2001), nomear e classificar são essencialmente sinônimos. Desta forma, quando damos nome a algo muitas vezes realizamos até mesmo sem perceber o ato de classificar, pois o nome representa uma espécie de rótulo para uma classe, ou grupo, e por outro lado, quando temos a informação da classe ou o rótulo, podemos inferir as propriedades de um específico objeto baseado na categoria ao qual pertence.

Na história da humanidade, a astronomia iniciou-se nomeando as estrelas e planetas, a biologia moderna se iniciou com o sistema de nomenclatura de Lineu, o qual basicamente ainda é usado nos dias atuais. (HAMPEL, 2002). Em qualquer campo do conhecimento humano, um dos mais importantes procedimentos de análise de dados é

classificar os dados em um conjunto de categorias, formando classes ou grupos, conforme registra Xu e Wunsch (2009).

Sistemas de classificação podem ser divididos em supervisionados ou não supervisionados. Os sistemas supervisionados partem da hipótese de que no espaço das observações existem grupos pré-definidos, ou seja, existe um conhecimento prévio da estrutura dos dados em que a especificação da pertinência a um dos grupos é conhecida a priori para cada uma das observações. Assim, na classificação supervisionada, o desafio é dada uma nova observação, poder classificá-la dentro de um dos grupos usando a estrutura intrínseca aos dados de entrada. (XU; WUNSCH, 2009).

A Figura 1.1 se refere a um exemplo de classificação supervisionada apresentada por DUDA, HART e STORK (2001), em que se tem dois grupos definidos a priori. Podemos ver a simplicidade do classificador dada pela reta desenhada: um novo peixe será classificado como *salmon*, se estiver à esquerda dessa linha e como *sea bass* em caso contrário.

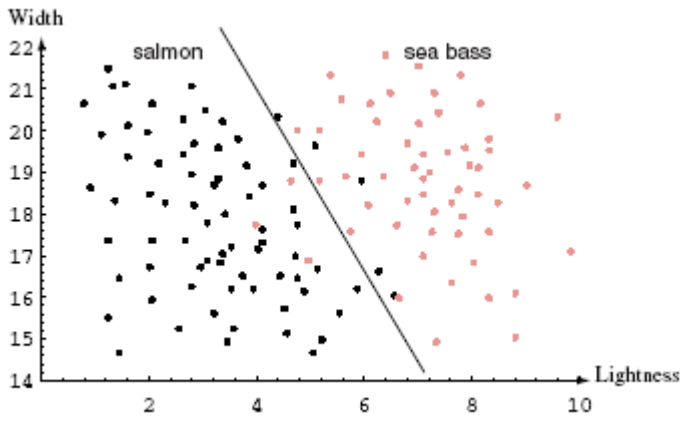


Figura1.1: Classificador de peixes: *salmon* e *sea bass*

Os sistemas não supervisionados, também conhecidos como agrupamento ou clustering, têm o objetivo de separar um conjunto de observações não classificadas em um número discreto de grupos que são definidos pela estrutura natural dos dados, sem uso de qualquer informação prévia sobre os grupos (XU; WUNSCH, 2009). Deste modo, quando se tem necessidade de explorar a desconhecida natureza dos dados independente

de se ter uma pré-informação de pertinência, a análise de clustering é a ferramenta mais adequada (XU; WUNSCH, 2009).

Em análise de clustering, quando a representação de dados é feita com poucos clusters, necessariamente se perde certos detalhes mais finos, semelhantemente ao processo de perda de dados por compressão. De outro lado, consegue-se simplificação, pois é possível representar muitos dados em um pequeno número de grupos. (BERKHIN, 2002).

Agrupamento também é um processo subjetivo, deste modo, é necessário atenção extra ao se realizar uma análise de cluster nos dados. A subjetividade está presente em diversos aspectos, entre eles nas hipóteses estabelecidas sobre os dados, a definição da medida de proximidade, a determinação do número de grupos, a seleção do algoritmo de agrupamento e a determinação dos índices de validação (XU; WUNSCH, 2009).

Além disso, para mesmo conjunto de dados, objetivos diferentes geralmente levam a diferentes partições. Um exemplo simples e direto é na partição de animais: uma águia, um canário, um leão, uma pantera e um carneiro. Se os animais são divididos com base no critério de poder ou não voar, temos dois clusters: com a águia e o canário em um grupo e o restante em outro grupo. No entanto, se mudarmos o critério e avaliarmos se eles são ou não carnívoros, temos uma partição completamente diferente com o canário e o carneiro em um cluster e os outros três no segundo grupo (XU; WUNSCH, 2009).

Sendo assim, de modo geral, toda classificação depende da capacidade de avaliação e do discernimento do usuário. A figura 1.2, apresentada originalmente em (XU; WUNSCH, 2009), mostra uma ilustração do papel da subjetividade nos resultados de uma análise de agrupamento. Nesse exemplo, é possível agrupar os dados em quatro grandes grupos, entretanto uma análise mais detalhada poderia nos levar a nove grupos.

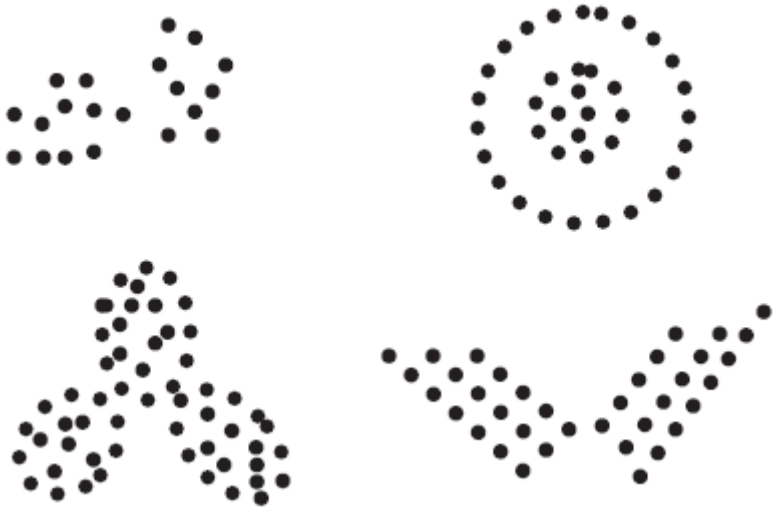


Figura 1.2. Papel da Subjetividade em análise de agrupamentos

Os diferentes critérios de agrupamento podem atribuir um indivíduo específico a diferentes grupos e, portanto, produzir diferentes partições. No entanto, em geral não há absolutamente qualquer maneira de se determinar qual é o melhor critério. De fato, cada critério tem seu uso adequado a situações particulares, embora alguns deles possam ser aplicados às mais vastas situações do que outros (XU; WUNSCH, 2009).

Uma questão importante refere-se ao critério a ser utilizado para se decidir até que ponto dois elementos do conjunto de dados podem ser considerados como semelhantes ou não. Para responder a esta questão é necessário considerar medidas que descrevem a similaridade entre os elementos de acordo com as características que foram medidas. Existe uma grande variedade de medidas, determinadas pelas especificidades das diferentes aplicações, cada uma delas produzindo um determinado resultado de agrupamento, caso aplicadas a um mesmo conjunto de observações. HANSEN e JAUMARD (1997) dissertam em detalhes sobre essa questão.

1.2 - Agrupamento segundo Critério de Minimização Soma de Distâncias

O objetivo central do presente trabalho considera a resolução do problema de agrupamento segundo o critério de minimização da soma das distâncias. O problema considerado é especificado a seguir.

Seja $S = \{s_1, s_2, \ldots, s_m\}$ um conjunto de m observações pertencentes a um espaço euclideano com n dimensões. Essas observações devem ser agrupadas em um determinado número q de grupos. Para cada grupo temos um centróide associado, $x_i \in \Re^n, i = 1, \cdots, q$. O conjunto das coordenadas dos centróides, $x_i, i = 1, \ldots, q$, é representado por um vetor $X \in \Re^{nq}$.

O procedimento de separação do conjunto das m observações no numero q de grupos deve ser efetuado de modo a minimizar a soma das distâncias das observações aos centróides do grupo em que estão associadas. Dessa forma, temos o problema de agrupamento ou clustering segundo o critério de minimização da soma de distâncias como um problema do tipo *min-sum-min:*

$$X^* = \arg\min_{X \in \Re^{nq}} \sum_{j=1}^{m} \min_i \left\| s_j - x_i \right\|_2 \qquad (1)$$

1.3 - Problema de localização de Fermat-Weber

Decisões sobre onde se deve localizar as facilidades, que podem ser fábricas, depósitos, escolas e antenas, considerando a população que deve ser servida, são de estrema importância para a sociedade e fundamentais para os prestadores de serviços. "Em geral, várias facilidades serão localizadas, que por sua vez, serão alocadas aos seus clientes. Desta forma tais problemas são também conhecidos como problemas de localização-alocação", conforme registram ARAKAKI e LORENA (2006).

Esforços significativos de pesquisa têm sido dedicados a problemas de localização, porque estes problemas são de grande importância prática, ao lado dos interesses de natureza teórica. Esses modelos não são usados apenas para otimizar a localização de instalações, mas também aparecem como subproblemas de um espectro mais amplo de problemas de logística. Em outras áreas científicas, sob aparência diferente, problemas equivalentes também são estudados. Por exemplo, em localização discreta, duas de suas variações, o problema p-mediana e o problema de localização de uma única simples fábrica, podem ser considerados os dois problemas mais estudados, segundo GARCIA, LABBÉ e MARÍN (2010).

Para problemas reais de localização, uma freqüente questão é localizar um objeto ou mais objetos, comumente chamados de facilidades, de modo a minimizar a soma de distâncias aos pontos de atendimento ponderadas pelos seus valores de demanda. Esse critério, natural e intuitivo, é o mais freqüentemente adotado nos estudos de localização.

Adotando a mesma notação da secção anterior e denominando por w_j a demanda associada ao ponto de atendimento j , esse problema de localização pode ser colocado sob a forma:

$$X^* = \arg\min_{X \in \Re^{nq}} \sum_{j=1}^{m} w_j \min_i \left\| s_j - x_i \right\|_2 . \qquad (2)$$

Essa formulação constitui-se no famoso problema de Fermat-Weber ou problema contínuo da p-mediana ou *Multisource Weber Problem*, que busca minimizar o custo total de transporte das facilidades às cidades, ponderando o peso específico de cada uma delas.

Dessa forma, pode ser visto facilmente que o problema de agrupamento segundo o critério de minimização da soma de distâncias (1) equivale ao problema de Fermat-Weber (2), unicamente considerando as demandas com valores unitários, $w_j = 1,\ j = 1, \cdots, m$.

Finalmente, deve-se ressaltar que todos os procedimentos metodológicos apresentados no desenvolvimento deste trabalho para resolução do problema (1) podem ser utilizados para resolução do problema de Fermat-Weber, bastando única e exclusivamente incluir os valores das demandas $w_j = 1,\ j = 1, \cdots, m$ nos lugares pertinentes.

2 – Métodos de Agrupamento

Nesse capítulo, faremos uma descrição sucinta de algumas das diversas alternativas metodológicas utilizadas para se resolver problemas de agrupamento. Sendo assim, o objetivo deste capítulo é fornecer uma revisão bibliográfica abrangendo os diferentes métodos de agrupamento.

Há atualmente uma vasta quantidade de algoritmos de agrupamento. Uma excelente revisão sobre métodos de agrupamento pode ser encontrada no artigo *Data Clustering: A Review* (JAIN; MURTY; FLYNN, 1999). Mais recentemente, os artigos *Recent Advances in Clustering: A Brief Survey* (KOTSIANTIS; PINTELAS, 2004) e *Survey of clustering data mining techniques* (BERKHIN, 2002) abordam as principais técnicas de agrupamento, incluindo alguns métodos mais novos.

Como ressaltado em BERKHIN, 2002, a especificação de classificações para os algoritmos de agrupamento não é uma tarefa simples, nem canônica, de modo que pode ocorrer uma certa sobreposição entre as diferentes classes definidas. Os três trabalhos citados acima divergem nesse tópico.

Na classificação clássica, os algoritmos de agrupamento são comumente divididos em Hierárquicos e em Particionais. A categorização, a seguir apresentada, é resultado de uma compilação daquelas estabelecidas pelas três referências acima destacadas.

Métodos Hierárquicos
Métodos de Particionamento e realocação
    k-means
    k-medoides
    CLARA
Métodos de Particionamento fundamentados em modelos probabilísticos
    EM
Métodos de Particionamento com base em concentração de pontos
    DBSCAN

Métodos de Particionamento baseados na estrutura de grade Grid based method

STING

CLIQUE

A seguir, cada uma das categorias será descrita de uma forma sucinta.

2.1 - Métodos Hierárquicos

Os agrupamentos do tipo hierárquico produzem uma hierarquia entre os grupos. Essa hierarquia pode ser representada por uma árvore de grupos, que é conhecida como dendograma. Nessa representação as observações de dados individuais são as folhas da árvore e os nós do interior são aglomerados de grupos. Os métodos hierárquicos permitem, assim, a exploração dos dados em diferentes níveis de granularidade.

Os métodos de agrupamento hierárquico são classificados em métodos aglomerativos (*bottom-up*) e métodos de divisão ou divisivos *(top-down*). Na abordagem dos métodos aglomerativos, inicia-se com cada observação do conjunto de dados em um distinto grupo e sucessivas incorporações de observações são realizadas até que todas as observações pertençam a um único grupo. Na abordagem dos métodos divisivos inicia-se com um único grupo contendo todas as observações do conjunto de dados e recursivas divisões são executadas até que o número de grupos seja igual ao número de observações.

Os métodos hierárquicos são muito naturais e intuitivos, tendo sido utilizados por Aristóteles, já na Antiga Grécia. Algumas vantagens do agrupamento hierárquico são:

- Flexibilidade em relação ao nível de granularidade dos grupos
- Não é necessário definir o número de grupos a priori
- Facilidade de lidar com diversas funções de similaridade ou distância
- Aplicabilidade a qualquer tipo de atributo
- Bons resultados gráficos estão integrados nos métodos

Os métodos hierárquicos apresentam, entretanto, importantes desvantagens:

• A maioria dos algoritmos hierárquicos não possui habilidade de executar ajustes uma vez que os grupos são criados, os elementos agrupados não podem ser alocados a novos grupos.

• Não é aplicável a grandes massas de dados.

Nos agrupamentos hierárquicos, a representação tradicional ponto por atributo de dados, onde são especificadas as coordenadas de cada observação, é muitas vezes de importância secundária. Em vez disso, é utilizado uma matriz NxN , onde N é o número de observações, correspondente a matriz de distâncias (dissimilaridades) ou de semelhanças, por vezes chamada de matriz de conectividade.

Com a matriz de conectividade podemos associar um grafo $G(V, A)$, cujos vértices são as observações de dados e cujas arestas são representadas pela matriz de conectividade. Com isso, se estabelece uma conexão entre os problemas de agrupamento hierárquico e o problema de particionamento em grafos.

A metodologia de agrupamentos hierárquicos apresenta dificuldades quando aplicada para um grande conjunto de dados, pois é necessário manter a matriz de conectividade na memória.

A adequação dos grupos para a fusão, no caso aglomerativo, ou divisão dos grupos, no caso divisivo, é influenciada pela escolha da função de similaridade dos elementos dos grupos, pois existe a presunção geral de que os grupos consistem em pontos similares. Para mesclar ou dividir subconjuntos de pontos ao invés de pontos individuais, a distância entre os indivíduos tem que ser generalizada para uma distância entre subconjuntos. Deste modo, uma medida de proximidade derivada, chamada de ligação métrica (*linkage metrics*) é utilizada. Assim, o tipo de ligação métrica utilizada tem um impacto significativo em algoritmos hierárquicos, pois reflete um conceito de proximidade e conectividade.

Umas das mais simples métricas de ligação são o Encadeamento Simples (Single linkage), conhecido como vizinho mais próximo, e o Encadeamento completo (complete linkage), conhecido como vizinho mais longe.

A métrica de Encadeamento Simples (*Single linkage*) é frequentemente adotada e está relacionada com o problema de encontrar a árvore geradora de peso mínimo, cuja resolução tem complexidade $O(N^2)$ onde N corresponde ao número de observações. O método usa a menor distância inter-clusters definida em termos de pares de nós, um em cada grupo respectivo, e, dessa forma, está naturalmente relacionado com o grafo de conectividade $G(V, A)$, apresentado anteriormente. O algoritmo MST (*Minimum Spanning Tree*), é um exemplo de algoritmo hierárquico divisivo que lida directamente como grafo de conectividade (JAIN; DUBES, 1988)

Outros métodos utilizam informação geométrica, onde a distância inter-cluster é definida em termos de todos os elementos do grupo, ou, alternativamente, onde o grupo é representado pelo seu ponto central. Alguns métodos desse tipo de ligação são abaixo relacionados:

- método de ligação por centróide (*Centroid method*)
- método de ligação por mediana
- método *Ward* de ligação
- método de ligação por média (A*verage linkage*)

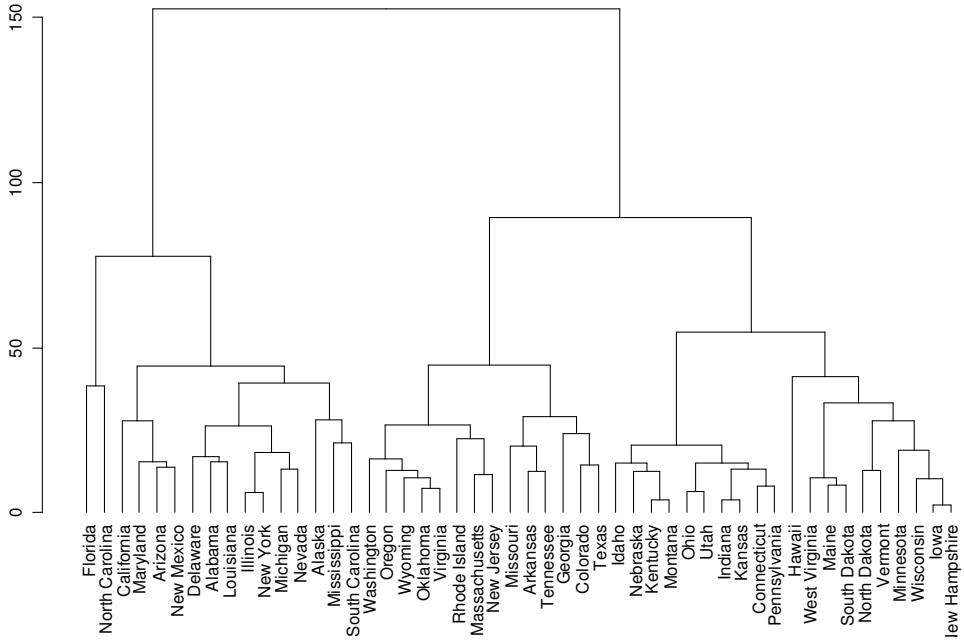


Figura 2.2 - Dendograma dos estados Americanos

A figura 2.2 ilustra uma classificação obtida pela utilização de um método hierárquico. Esse gráfico tem como base um método aglomerativo utilizando a distância euclidiana conjugada à métrica de encadeamento médio. A sua geração foi feita através do software estatístico **R**, pela função **hclust** da biblioteca **stats,** conforme código mostrado no Anexo 1.1.

2.2 1 - Métodos de Partição e realocação

Como relatado, uma das principais desvantagens dos métodos hierárquicos é a falta de habilidade de executar ajustes. Ao contrário dos tradicionais métodos hierárquicos, em que os grupos não são revisitados depois de terem sido construídos, os algoritmos de realocação melhoram gradualmente os grupos, resultando em clusters de melhor qualidade.

Basicamente, os algoritmos de particionamento e realocação, dividem os dados em vários subgrupos. Entretanto, a verificação de todos os possíveis subconjuntos é computacionalmente inviável. Deste modo, foram desenvolvidas diferentes Heurísticas que executam a realocação de forma iterativa, re-atribuindo as observações entre os grupos.

2.2.1 - Algoritmo k-means

Entre os algoritmos de agrupamento do tipo partição, o algoritmo *k-means* é destacadamente o mais importante, originalmente proposto por MacQueen (1967). O algoritmo *k-means* é amplamente utilizado nas mais diversas aplicações. Na área de mineração de dados, por exemplo, entre inúmeros algoritmos, foi classificado como o segundo mais importante em *data mining*, conforme WU *et al.* (2007).

O algoritmo *k-means* trata do problema de agrupamento segundo o critério de mínima soma de quadrados:

$$X^* = \arg\min_{X \in \Re^{nq}} \sum_{j=1}^{m} \min_i \left\| s_j - x_i \right\|_2^2, \quad (3)$$

onde x_i é o centro de gravidade do grupo i , sendo s_j uma observação genérica j, que pertence ao grupo *i*.

Desde o surgimento do algoritmo *k-means*, inúmeros aprimoramentos e variações têm sido propostos, como aqueles mostrados em HANSEN e MLADENOVIC (2001), LIKAS *et al.* (2003), HANSEN *et al.* (2005) e BAGIROV (2008)..

O algoritmo *k-means* é um método iterativo simples para particionar um conjunto de dados em um número de grupos especificado pelo usuário. Basicamente o algoritmo, em suas diversas formas, possui a seguinte estrutura de passos.

Algoritmo k-means

Passo 1: seleção de *k* pontos para sementes iniciais dos centróides dos *k grupos.*
Passo 2: cada observação é associado a um *cluster,* para o qual a dissimilaridade entre o objeto e o centro do cluster é menor que as demais.
Passo 3: os centros dos clusters são recalculados, redefinindo cada um em função da média de todas as observações atribuídas ao grupo.
Passo 4: retorna ao passo 2 até que os centros dos clusters se estabilizem.

O passo 2 tem como resultado um particionamento das observações. A cada iteração, os objetos são agrupados em função do centro do grupo mais próximo, ocorrendo realocação dos objetos entre os grupos. Por conseqüência, os centros dos clusters, também chamados de centróides, são modificados e recalculados no passo 3, prosseguindo iterativamente até que os centróides não sejam mais deslocados, ou seja, até que haja uma convergência da sequência de centróides.

A distância euclidiana é utilizada como medida de proximidade padrão. Nesse caso pode-se facilmente mostrar que a função de custo (3), não-negativa, irá sempre decrescer quando houver uma mudança nos passos 2 e 3. Sendo assim, a convergência é

garantida em um número finito de iterações, desde que o número de particionamentos seja finito.

O algoritmo é normalmente bastante sensível em relação às sementes iniciais dos centróides, sendo a convergência garantida apenas para um ótimo local. Existem várias técnicas para a seleção destas sementes iniciais, sendo a mais simples a amostragem de forma aleatória do conjunto de dados, definindo as observações da amostra como os centróides iniciais

Esse algoritmo é muito sensível a ruídos, uma vez que o método considera a soma dos desvios ao quadrado em relação aos centros de gravidade. Além disto, é inadequado para descobrir grupos com formas não convexas ou de tamanhos muito diferentes. Este método é muito utilizado, quando se tem a hipótese dos grupos serem esféricos.

2.2.2 - Algoritmo k-Medoids

A diferença básica em relação ao *k-means* está na utilização de uma das observações do conjunto original como elemento representativo, chamado *medoid*, localizado mais no meio do *cluster,* ao invés da tradicional escolha do centro de gravidade do grupo.

Nos algoritmos *k-medoids* os grupos são definidos como subconjuntos de pontos que estão mais próximos dos seus elementos representativos, que são chamados de medoides. O *medoid* pode ser definido como o objeto do grupo, cuja soma das dissimilaridades a todos os objetos do mesmo grupo seja mínima, o que é equivalente a média dessa soma ser mínima.

Assim como o k-means, esse método é bem adequado na hipótese dos grupos serem esféricos, ocupando cada medoide uma observação mais central do grupo. Entretanto, esse método é menos sensível a ruídos de que o *k-means,* pois não avalia os

desvios das observações aos centróides ao quadrado, como é feito no *k-means*. Além disso, possui a característica de ser capaz de lidar com qualquer tipo de atributo.

O algoritmo *Partitioning Around Medoids* (PAM) é um algoritmo clássico da família dos métodos *k-medoids*.

Algoritmo *Partitioning Around Medoids* (PAM)

1. Initialização: seleciona aleatoriamente q das m observação do conjunto de dados como os medoids;
2. Associa cada observação do conjunto de dados ao medoid mais próximo usando qualquer distância métrica válida;
3. Para cada medoid:
   Para cada não-medoid:
   calcule o custo total da configuração do grupo como se esta observação
   fosse o medoid
4. Selecione a configuração com o menor custo.
5. Repita os passos 2-5 até que não haja mudança nos medoid.

Em relação ao tempo de processamento, o k-medoid método é menos eficiente do que o k-means, pois o cálculo do medoide é mais custoso computacionalmente do que o cálculo do centro de gravidade, resultando num maior tempo de processamento.

2.2.3 - Algoritmo CLARA (*Clustering Large Applications*)

O algoritmo PAM, descrito na secção anterior, efetua as diversas combinações das m observações formando os k grupos. Assim, no pior caso, pode envolver $C_m^q = m!/((m-q)!\,q!)$ operações. Dessa forma, o algoritmo PAM só pode ser

efetivamente utilizado para pequenas bases de dados, em função de sua alta complexidade de natureza combinatória.

Dadas essas dificuldades, para grandes bases de dados KAUFMAN e ROUSSEEUW (1986) introduziram o método CLARA, cuja idéia básica é trabalhar com amostras menores do conjunto original em vez de se usar todas as observações. Essas amostras ou sub-conjuntos de dados têm um tamanho fixo, quanto menor o tamanho, menor a complexidade.

Cada amostra de dados é particionada em q grupos, usando o mesmo princípio do algoritmo PAM. Após o algoritmo convergir, calcula-se a soma das distâncias das observações aos seus respectivos medoides, sendo esta soma uma medida de qualidade do agrupamento associado a essa particular amostra.

Seleciona-se os medoides cuja amostra possui a menor soma, ou seja, o melhor agrupamento. Uma vez que q medoides foram selecionados a partir da melhor amostra de dados, cada observação do conjunto de dados inteiro é atribuída para o medoide mais próximo.

Este método tem certa fragilidade, pois nem sempre um bom agrupamento com base em amostras representa um bom agrupamento para os dados como um todo. Dessa forma, a eficiência do CLARA também depende do número de amostras utilizadas. Essa dificuldade pode ser minorada tomando-se um grande número de amostras. De toda maneira, há sempre o compromisso entre complexidade versus qualidade.

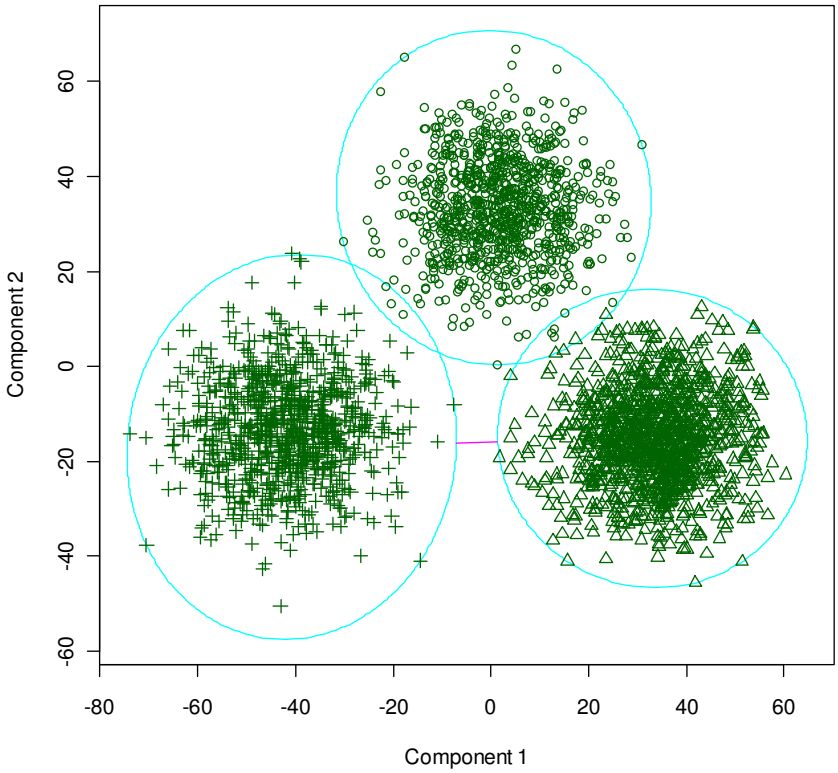


Figura 2.3 - Algoritmo CLARA,: resultados para 3 grupos sintéticos

A Figura 2.3 mostra os resultados obtidos pela utilização do algoritmo CLARA disponibilizado no software estatístico R, através da função CLARA, aplicado a um conjunto de dados sintéticos. O conjunto completo de observações consiste de 3000 pontos em um espaço planar, gerados por três distribuição normais bi-variadas, sendo 1000 pontos para cada uma. Foram utilizadas 100 amostras de 42 observações cada uma. O código que gerou a figura está disponível no anexo 1.2.

2.3 - Métodos Probabilísticos

Na abordagem de agrupamento probabilístico, os dados são considerados uma amostra independente extraída de um modelo de misturas de várias distribuições de probabilidade, conforme descrito no livro de McLachlan e Basford (1988).

Existem diversos algoritmos que tratam de misturas de distribuições. A maioria deles consiste de modificações e aprimoramentos do algoritmo EM (*Expectation Maximization*) apresentado no artigo de DEMPSTER, LAIRD e RUBIN (1977), proclamado como seminal pela literatura. Esses algoritmos assumem um modelo de probabilidades subjacente aos dados. Dessa forma, é possível se calcular a probabilidade de uma observação pertencer a um certo grupo desde que as distribuições sejam conhecidas.

2.3.1 - Algoritmo EM - Expectation Maximinization

Dada uma família de distribuições densidades probabilidade:

$$f\left(x; p; \theta\right) = \sum_{j=1}^{c} p_j g_j\left(x; \theta_j\right)$$

Onde x é uma variável aleatória p-dimensional e p_j é a proporção da função densidade probabilidade g_j na mistura de distribuições, cada uma com os seus parâmetros $\theta^t = \left(\theta_1^t, \theta_2^t, ..., \theta_c^t\right)$.

O algoritmo EM começa com estimativas iniciais para os parâmetros das distribuições. Esses valores são então usados para calcular a estimativa de probabilidade a posteriori, ou valor esperado da probabilidade a posteriori.

$$P(cluster\ j \mid x_i) = \frac{\hat{p}_j\ g_j(x_i; \overset{\wedge}{\theta}_j)}{f\left(x_i; \hat{p}; \overset{\wedge}{\theta}\right)}, \quad j = 1, ..., c.$$

As observações são alocadas ao grupos no qual a estimativa de probabilidade a posteriori é máxima. Tendo como base as observações alocadas aos grupos, estima-se novamente os parâmetros das distribuições. Uma forma ortodoxa para efetuar essa estimação é através da maximização da função de verossimilhança.

O algoritmo prossegue processando iterativamente as suas duas fases: cálculo dos valores esperados de pertinência a cada grupo (Expectation) e estimação dos parâmetros das distribuições de probabilidade (Maximization), até que não ocorra

mudança nos parâmetros das distribuições. Considerando essa natureza iterativa entre duas fases, o algoritmo EM tem uma estratégia similar ao algoritmo *k-means*.

Os algoritmos de agrupamento probabilístico têm algumas características importantes:

• Podem ser interrompidos e retomados com lotes consecutivos de dados, pois os grupos possuem representação totalmente distinta dos conjuntos de dados

• Em qualquer fase do processo iterativo do modelo, distribuições intermediárias podem ser usadas para se fazer alocações provisórias das observações aos diferentes grupos.

• Produzem um sistema de agrupamento de facil interpretação

• O modelo conceitual de mistura tem fundamento probabilístico claro, portanto a determinação do número de grupos mais adequado é uma tarefa simples e natural.

Uma desvantagem de tais algoritmos do tipo EM é que eles tendem a ser computacionalmente caros, devido à complexidade intrínseca à fase *Maximization*, que normalmente corresponde à resolução de um problema de programação não-linear. Outra dificuldade encontrada é o *overfitting*, que corresponde a uma dependência excessivamente alta do conjunto de valores dos parâmetros estimados para à específica base de dados utilizada. Esta dificuldade pode estar presente, em particular, quando um grande número de grupos é especificado ou quando as distribuições de probabilidades possuem muitos parâmetros. Em resumo, as dificuldades crescem com o número total de parâmetros a serem estimados.

A figura 2.4 mostra um conjunto de observações geradas por duas distribuições normais multivariadas para constituir os dados de entrada do algoritmo EM.

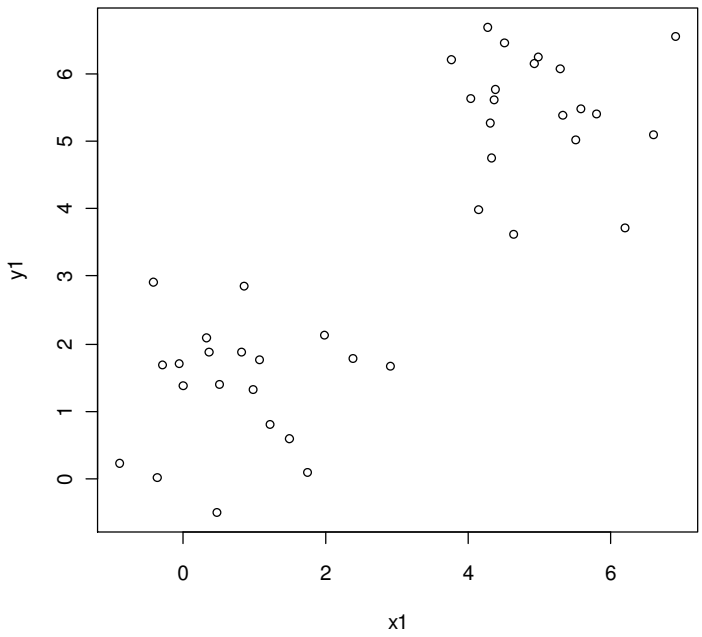


Figura 2.4 – Dados de entrada para o algoritmo EM

A figura 2.5 mostra os resultados da classificação obtida pelo algoritmo EM. O experimento foi efetuado usando o software estatístico R, através da função Mclust, da biblioteca mclust (FRALEY; RAFTERY,2006) Na figura, as observações são classificadas em dois grupos, representados por símbolos diferentes, quadrados e triângulos, e cores diferentes, respectivamente em cores vermelhas e azuis. O código que gerou as figuras desta seção encontra-se no anexo 1.3.

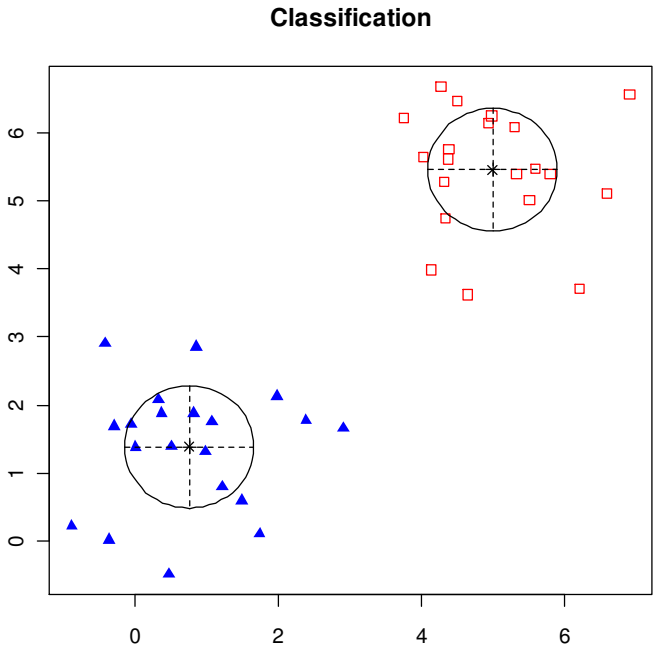


Figura 2.5 – Agrupamentos produzidos pelo algoritmo EM

A figura 2.6 exibe, as observações representadas conforme a qualidade da classificação. Quanto menor e mais claro for o ponto do gráfico melhor foi a classificação.

Em destaque estão as observações fora dos dois círculos, que é um subconjunto de observações com classificação de natureza incerta. Essas observações são comumente denominadas por *outliers*.

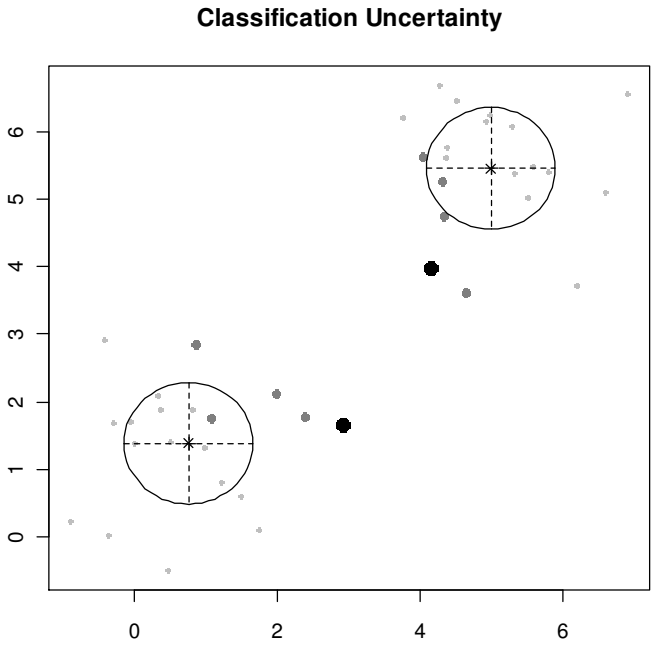


Figura 2.6 – Classificações incertas produzidas pelo algoritmo EM

A figura 2.7 apresenta as curvas de nível associadas às distribuições densidade de probabilidade estimadas pelo algoritmo EM.

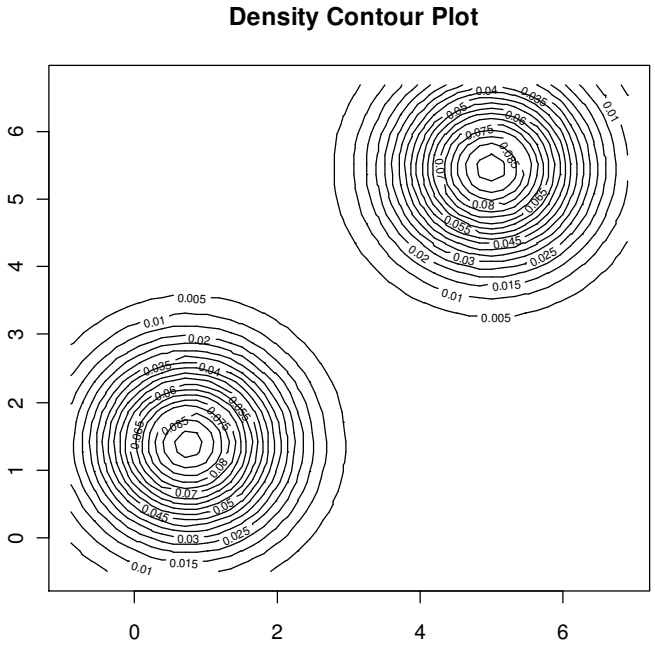


Figura 2.7 - Curvas de nível das distribuições produzidas pelo algoritmo EM

2.4 - Métodos com base na concentração de pontos

A Figura 2.9, apresentada originalmente por BERKHIN (2002), ilustra algumas formas de grupos que apresentam dificuldades para algoritmos de particionamento e realocação, por exemplo, para algoritmos tipo k-means. Em contraposição, esses grupos são tratados adequadamente por algoritmos de agrupamento com base na concentração de pontos.

Figura 2.9 -Agrupamentos com forma irregular

Algoritmos de agrupamento com base na concentração de pontos são capazes de descobrir agrupamentos de formas arbitrárias, não possuindo nenhuma restrição intrínseca quanto à forma, pois um grupo pode crescer em qualquer direção desde que seja densa. Além disso, uma estratégia fundamentada em concentração fornece uma natural proteção contra *outliers*, pois normalmente são pontos isolados.

2.4.1 - Algoritmo DBSCAN

Um dos mais conhecidos algoritmos baseados em concentração de pontos é o DBSCAN (*Density Based Spatial Clustering of Applications with Noise*). O algoritmo DBSCAN se fundamenta em dois critérios avaliados em conjunto: critério do raio e critério de concentração. No primeiro critério, busca-se encontrar grupos de observações onde as observações são classificadas com base no critério de vizinhança, considerando um pré-definido raio (Eps). No segundo critério, um grupo tem que possuir um atributo de concentração, ou seja, tem que conter um número mínimo de observações (MinPts).

A partir dos parâmetros definidos Eps e MinPts são definidos os conceitos de vizinhança, densidade, conectividade e fronteira. Esse algoritmo basicamente realiza a separação do conjunto de observações em três classes:

• Pontos Núcleo. Estes são pontos que estão no interior de um grupo. Um ponto é interior, se há um número de pontos suficientes (MinPts) em sua vizinhança com raio (Eps).

• Pontos Fronteira. Um ponto de fronteira é um ponto que não é um ponto núcleo, ou seja, não há um número suficiente de pontos (MinPts) em sua vizinhança, mas está dentro da vizinhança de um ponto núcleo, considerando um dado raio (Eps),

• Pontos Ruído: Um ponto de ruído é qualquer ponto que não é um ponto núcleol e nem um ponto da fronteira, ou seja , corresponde a um *outlier*.

Para encontrar os agrupamentos, o algoritmo DBSCAN faz uma varredura nas observações determinando todos os pontos núcleo. Faz-se a seguir uma varredura dos pontos núcleo fazendo as conexões a todos os pontos que estejam a uma distância menor do que (Eps). Cada subconjunto de pontos conectados entre si, conceito de conectividade, forma um cluster.

A figura 2.8 mostra um exemplo de resultados obtidos através da função dbscan da biblioteca fcp do software estatístico R. Este gráfico foi elaborado com base no exemplo da função dbscan, o conjunto de dados contém 600 observações e o código R encontra-se no anexo 1.4. As observações foram classificadas em 11 grupos, de modo que existe uma cor associada a cada grupo. Os pontos núcleo são representados por triângulos e os pontos fronteira são representados por círculos, possuindo eles a respectiva cor do grupo. Os pontos ruído, *outliers,* também são representados por círculos, na cor preta.

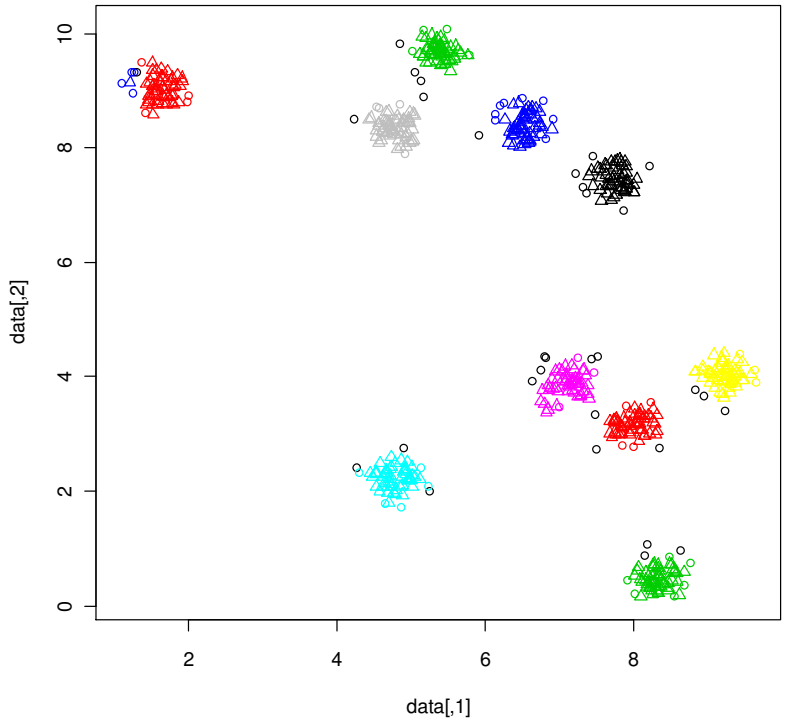


Figura 2.8 – Resultados da classificação através do algoritmo DBSCAN

2.5 - Métodos baseados na estrutura de grade (*Grid based method*)

Métodos baseados em estrutura de grade primeiramente executam a divisão do espaço das observações em um número finito de células (Hiper-retângulos). Em seguida, executam as operações básicas pertinentes no nível das células: Células que contêm mais do que um certo número de pontos são tratadas como densas, bem como, conjunto de células densas são conectadas para formar os clusters.

A Figura 2.10 ilustra o processo iterativo de subdivisão de um espaço com duas dimensões. No primeiro nível (1 st layer), tem-se o retângulo original contendo todas as observações. Nos níveis inferiores é efetuado o gradual refinamento do particionamento em retângulos, cada vez menores, como mostrado na figura pelos níveis ((i-1) st layer) e (i th layer).

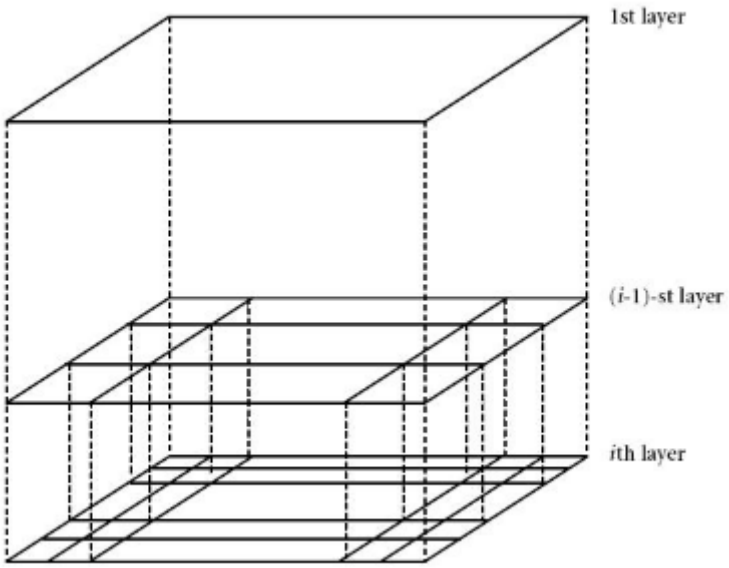


Figura 2.10 – Subdivisão do espaço pelos algoritmos de estrutura de grade

A seguir, são apresentados dois algoritmos baseados nessa idéia de estrutura de grade: algoritmo STING e algoritmo CLIQUE.

2.5.1 - Algoritmo STING

Um dos principais algoritmos de agrupamento baseados em grade é o algoritmo STING (*STatistical INformation Grid-based*). A tradução literal "informação estatística baseada em grade" já dá uma idéia parcial.

Como em outros algoritmos desta classe, a área espacial é dividida em células retangulares. O algoritmo STING inicia dividindo a área espacial em vários níveis de células retangulares, correspondentes a níveis de resolução, formando uma estrutura hierárquica. As células de nível mais alto são compostas a partir de células de nível mais baixo. Isso gera uma estrutura hierárquica das células produzidas pela subdivisão consecutiva do espaço.

A Figura 2.11, originalmente apresentada por BERKHIN (2002), ilustra o processo da subdivisão de espaços e da conseqüente formação da hierarquia das células

efetuado pelo algoritmo STING. A estrutura hierárquica de células representa a informação em diferentes níveis de agrupamento. Em cada nível, são registradas informações estatísticas básicas sobre os atributos de cada célula da grade como: freqüência, média, desvio padrão. valor máximo e valor mínimo.

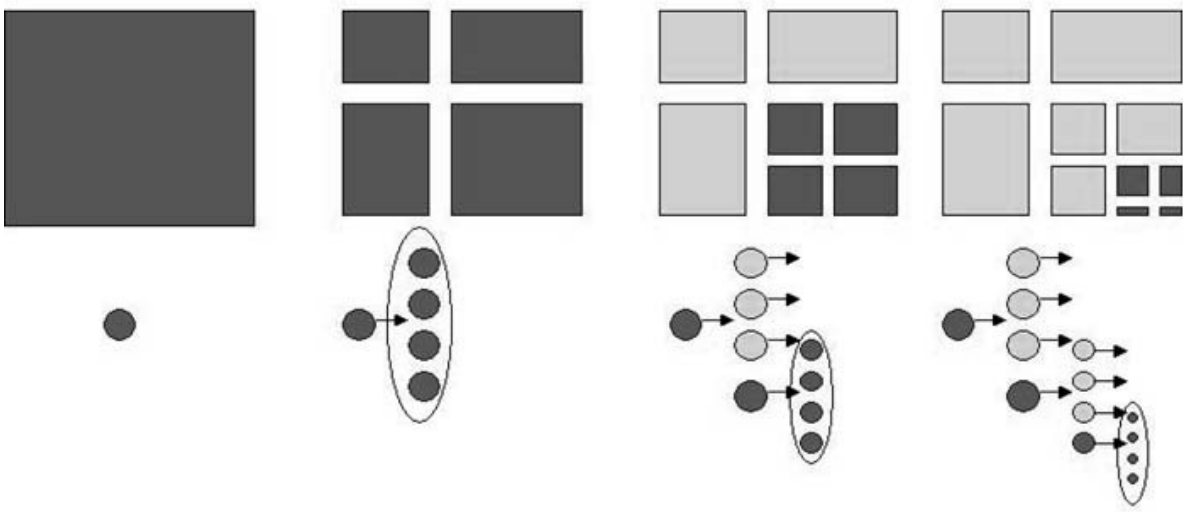


Figura 2.11 – Algoritmo STING: subdivisão do espaço e hierarquia das células

O algoritmo STING tem a propriedade de gerar bons resultados de agrupamento em um curto espaço de tempo de execução, entretanto existem duas grandes dificuldades com este algoritmo. Em primeiro lugar, o desempenho de STING depende da granularidade do nível mais baixo da estrutura de grade. Em segundo lugar, os grupos resultantes são somente limitados horizontalmente ou verticalmente, não existindo qualquer limitação na diagonal. Esta lacuna pode afetar significativamente a qualidade dos grupos.

2.5.2 - Algoritmo CLIQUE

Outro método de agrupamento baseado em grade é o algoritmo CLIQUE (AGRAWAL *et al*., 1998). Esse algoritmo começa por encontrar todas as células nos espaços unidimensionais correspondentes a cada atributo produzidas pelas projeções em um dado número (NumPartes) de partes iguais. São mantidas somente células que contenham número de elementos acima de um patamar (MinPts), denominadas células densas.

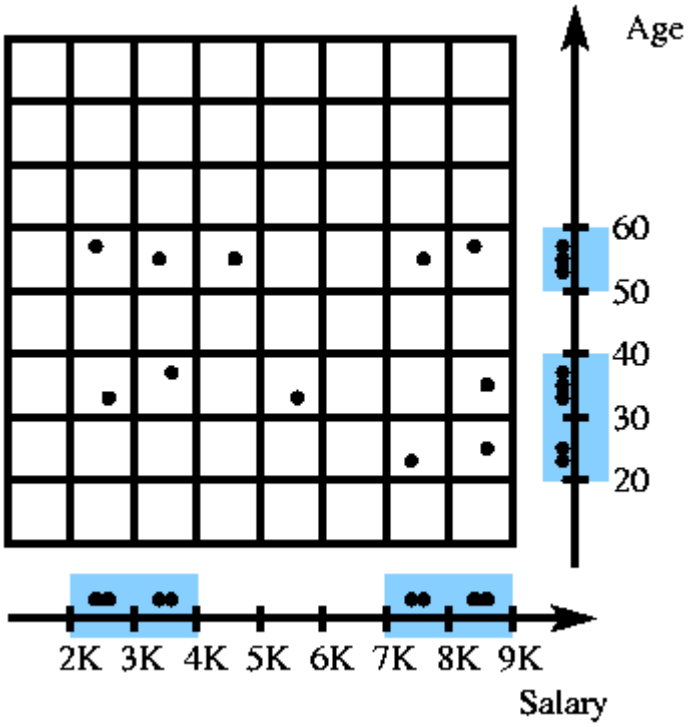


Figura 2.12 – Algoritmo CLIQUE: projeção nos espaços univariados

A Figura 2.12, disponibilizada pela Free University of Bozen, ilustra o procedimento de projeções efetuado pelo algoritmo CLIQUE em um exemplo com dois atributos: *Age* (eixo vertical) e *Salary* (eixo horizontal). A figura também mostra a grade produzida pelo parâmetro (NumPartes). Cada observação é representada por um pequeno círculo. Neste exemplo, foi considerado MinPts = 2. Assim, na projeção dos dados na dimensão *Salary,* as células [2K,3K], [3K,4K], [7K,8K] e [8K,9K] são células cluster. As células cluster para a dimensão *Age* são [20,30], [30,40] e [50,60]. Uma vez que as células cluster são identificadas, elas são juntadas pelo critério de adjacência. No exemplo, são formados os clusters : [2K,4K] , [7K,9K] (para o atributo *Salary*) e [20,40] , [50,60]] (para o atributo *Age*).

No caso geral, o algoritmo faz a identificação dos clusters em cada um dos atributos, como explicado no exemplo. Os atributos que não tem clusters, ou seja, que não atendem ao critério (MinPts), são eliminados. Os atributos restantes são combinados entre si, verificando-se a existência de clusters nesses espaços de mais alta dimensão. O processo continua até que não haja mais formação de clusters. Como

resultado, os grupos são formados pela união de hiper-retângulos entre subespaços projetados atendendo fundamentalmente ao critério de densidade (MinPts).

Voltando ao exemplo da figura 2.12, os atributos *Salary* e *Age* não são eliminados, desde que ambos contenham clusters. No passo seguinte, o algoritmo junta os dois atributos e efetua a seleção de células no espaço bidimensional. Neste exemplo, não existe qualquer célula bidimensional com dois ou mais pontos, portanto o critério de mínimo de pontos não é satisfeito. Assim, a dimensionalidade dos clusters é igual a um, e o algoritmo dá como resultado quatro clusters: [2K,4K], [7K,9K] para *Age* e [20,40], [50,60] para *Salary*.

3 - Descrição da Metodologia

Dado um conjunto de observações, $S = \{s_1, s_2, \ldots, s_m\}$, pertencentes ao $\Re^n$, deseja-se separar as observações em um número de grupos, q , definido a priori. Para cada grupo $i = 1, \ldots, q$ temos um centróide associado, $x_i \in \Re^n$. O conjunto das coordenadas dos centróides, $x_i, i = 1, \ldots, q$, é representado por $x \in \Re^{nq}$. Dada uma observação $s_j \in S$, inicialmente é calculada a distância deste ponto a cada centróide componente de $x = (x_1, \cdots, x_q)$. Vamos definir z_j como a menor distância entre a observação e os centróides, ou seja:

$$z_j = \min_{i=1,\cdots,q} \left\| s_j - x_i \right\|_2 \qquad (1)$$

A Figura 3.1, mostra uma observação s_j conectada aos centróides. A distância z_j está destacada em vermelho.

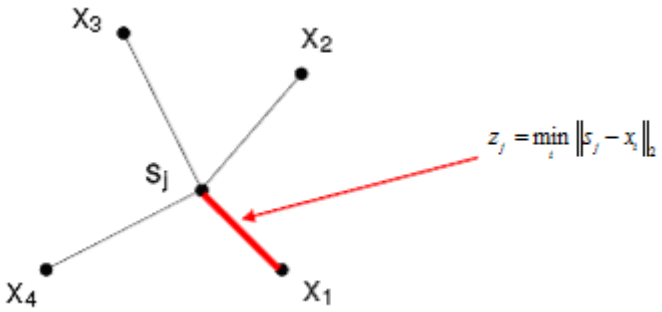


Figura 3.1 – Cálculo da distância ao centróide mais próximo

No presente trabalho, como medida da qualidade do agrupamento associado a um dado conjunto $x = (x_1, \cdots, x_q)$ de posições dos centróides, vamos definir o valor:

$$D(x) = \sum_{j=1}^{m} z_j . \qquad (2)$$

Dessa forma, o valor $D(x)$ corresponde à soma das distâncias das observações a seus centróides mais próximos. Como referido na Introdução, esta métrica é natural e muito utilizada. Ademais, remete-se também ao problema de Fermat-Weber, talvez o mais importante no tópico conhecido como *Facility Location.* O problema de Weber-Fermat pretende minimizar o custo total de entrega de produtos. Tal custo está

associado diretamente à distância Euclideana dos centros de demanda até as instalações centrais produtoras ponderadas pelas demandas $w_j,\ j = 1, \ldots, m$.

A localização ótima dos centróides deve fornecer o melhor valor da medida de qualidade $D(x)$. Como $D(x)$ depende unicamente de x , o problema é encontrar x^* , localização ótima dos centróides, pela seguinte especificação:

$$x^* = \arg\min_{x \in \Re^{nq}} D(x) \tag{3}$$

Aplicando 1,2 e 3, temos o problema de clustering como um problema do tipo *min-sum-min*

$$x^* = \arg\min_{x \in \Re^{nq}} \sum_{j=1}^{m} \min_i \left\| s_j - x_i \right\|_2 . \tag{4}$$

3.1 - Transformação do problema

O problema (4), acima, pode ser formulado de uma maneira equivalente como:

$$\begin{aligned} &minimizar \ \sum_{j=1}^{m} z_j \\ &sujeito\ a: \ z_j = \min_{i=1,\ldots,q} \left\| s_j - x_i \right\|_2 , \quad j = 1, \ldots, m \end{aligned} \tag{5}$$

O problema (5), além de ser definido em dois níveis, apresenta a dificuldade adicional de ser não-diferenciável. Com o objetivo de se obter um problema diferenciável, procederemos uma série de transformações análogas às apresentadas em SOUSA(2005) e XAVIER(2010). Considerando a definição de z_j , cada valor z_j tem que necessariamente satisfazer ao seguinte conjunto de desigualdades:

$$z_j - \left\| s_j - x_i \right\|_2 \leq 0, \quad i = 1, \ldots, q. \tag{6}$$

Substituindo o conjunto de restrições de igualdade do problema (5) pelas restrições de desigualdades (6), obtemos o problema relaxado:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j \\ &\textit{sujeito a}: \ z_j - \left\| s_j - x_i \right\|_2 \leq 0, \quad j = 1,\ldots,m, \ i = 1,\ldots,q \,. \end{aligned} \tag{7}$$

Este problema não é equivalente ao (5), pois as variáveis z_j não são limitadas inferiormente e, desse modo, não há qualquer empecilho para que $z_j \to -\infty$. Pela definição, z_j é uma distância, assim, o correto seria termos z_j limitado inferiormente em zero. Para obtermos tal equivalência ao problema original, o problema (7) tem que ser modificado. Para isso, preliminarmente faremos a definição da função $\psi(y) = max\{0, y\}$. Usando a função ψ para substituir o conjunto de restrições de desigualdades de (7), obtemos o problema:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j \\ &\textit{sujeito a}: \ \sum_{i=1}^{q} \psi(z_j - \left\| s_j - x_i \right\|_2) = 0, \quad j = 1,\ldots,m \,. \end{aligned} \tag{8}$$

O problema (8) continua sem equivalência ao original, pois mantém a indesejada propriedade de z_j, $j = 1,\ldots,m$ não serem limitados inferiormente. Como a função objetivo do problema (8) minimiza cada valor z_j , continua não havendo qualquer empecilho para que $z_j \to -\infty$.

Vamos, então, introduzir uma segunda modificação para limitar z_j : a substituição do sinal = por > nas restrições em (8) . Desta forma, obtemos o problema não canônico:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j \\ &\textit{sujeito a}: \ \sum_{i=1}^{q} \psi(z_j - \left\| s_j - x_i \right\|_2) > 0, \quad j = 1,\ldots,m \,. \end{aligned} \tag{9}$$

Para uma observação s_j fixa, definindo $d_i = \left\| s_j - x_i \right\|_2$ e assumindo $d_1 < \cdots < d_q$, a Figura 3.2 mostra a representação de três parcelas do somatório da equação (8) em função de z_j .

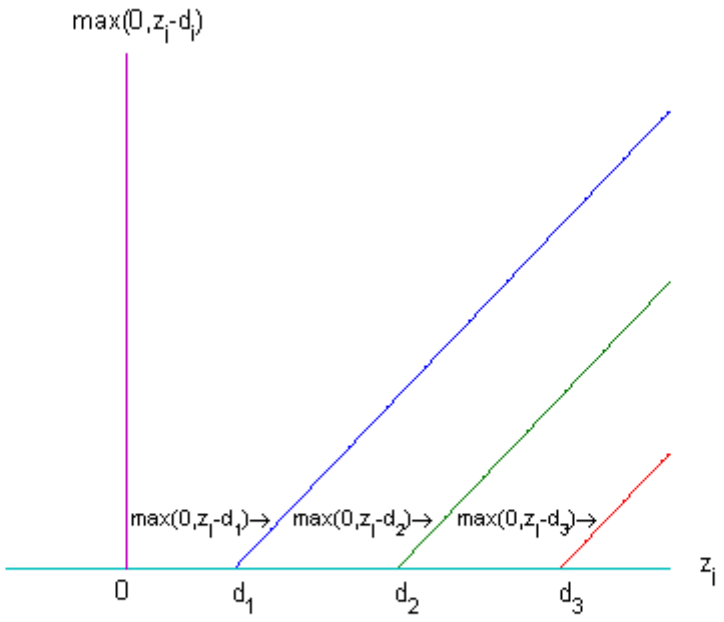


Figura 3.2 Três parcelas do somatório de restrição do problema (8) em função de z_j

A formulação canônica pode ser recuperada através da inclusão de uma perturbação ε nas desigualdades em (9), obtendo-se o problema modificado:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j \\ &\textit{sujeito a}: \ \sum_{i=1}^{q} \psi(z_j - \left\| s_j - x_i \right\|_2) \geq \varepsilon, \quad j = 1,\ldots,m \quad \varepsilon > 0 \end{aligned} \tag{10}$$

Para qualquer posição dos centróides x , é evidente que o conjunto viável do problema (10) está contido no conjunto viável do problema (9). Quando $\varepsilon \to 0_+$, o conjunto viável do problema (10) tende àquele do problema (9). Pode-se, então, pensar em resolver (9) através da resolução de uma sequência de problemas iguais a (10) para uma sequência de valores positivos decrescentes de ε que gradualmente se aproximam de 0_+.

3.2 - Suavização do problema

Analisando o problema (10), temos a presença da função ψ, que impõe ao mesmo uma estrutura não-diferenciável muito rígida, fazendo com que tenha uma resolução computacional muito difícil. Para contornar essa dificuldade é necessário se fazer uma adaptação do problema (10). Em vista disso, o método numérico adotado neste trabalho se fundamenta na suavização do problema. Para suavizar o problema, para $y \in \Re$ e $\tau > 0$, definimos a função $\phi(y,\tau)$, que além de ser uma boa aproximação da função $\varphi(y)$, também é diferenciável.

$$\phi(y,\tau) = \left(y + \sqrt{y^2 + \tau^2} \right) / 2 \qquad (11)$$

A função ϕ possui as seguintes propriedades:

(a) $\phi(y,\tau) > \psi(y), \quad \forall \tau > 0;$

(b) $\lim_{\tau \to 0} \phi(y,\tau) = \psi(y);$

(c) $\phi(y,\tau)$ é uma função convexa crescente que pertence à classe de funções C^{∞}.

(d) $\phi'(y,\tau) = \dfrac{d\phi(y,\tau)}{dy} = \dfrac{1}{2}\left[1 + \dfrac{y}{\sqrt{y^2 + \tau^2}} \right]$

(e) $\phi''(y,\tau) = \dfrac{d^2\phi(y,\tau)}{dy^2} = \dfrac{\tau^2}{\left(y^2 + \tau^2 \right)^{\frac{3}{2}}}$

(f) $\phi''(0,\tau) = \dfrac{d^2\phi(0,\tau)}{dy^2} = \dfrac{1}{\tau}$

(g) $\lim_{\tau \to 0} \dfrac{d^2\phi(0,\tau)}{dy^2} \to \infty$

A propriedade (b) da função $\phi(y,\tau)$ mostra a sua aproximação assintótica à função $\psi(y)$. A propriedade (c) permite o uso de métodos de otimização mais poderosos. Devido à propriedade (f), a curvatura da função $\phi(y,\tau)$ no ponto $y = 0$ é crescentemente atenuada por valores crescentes do parâmetro τ. De modo análogo, pelas propriedades (b) e (g), podemos ver que no problema original, essa curvatura tende a infinito, fato que provoca o surgimento de pontos de mínimos locais.

A Figura 3.3 mostra a aproximação assintótica da função da função $\phi(y,\tau)$ à função $\psi(y)$ quando $\tau \to 0$.

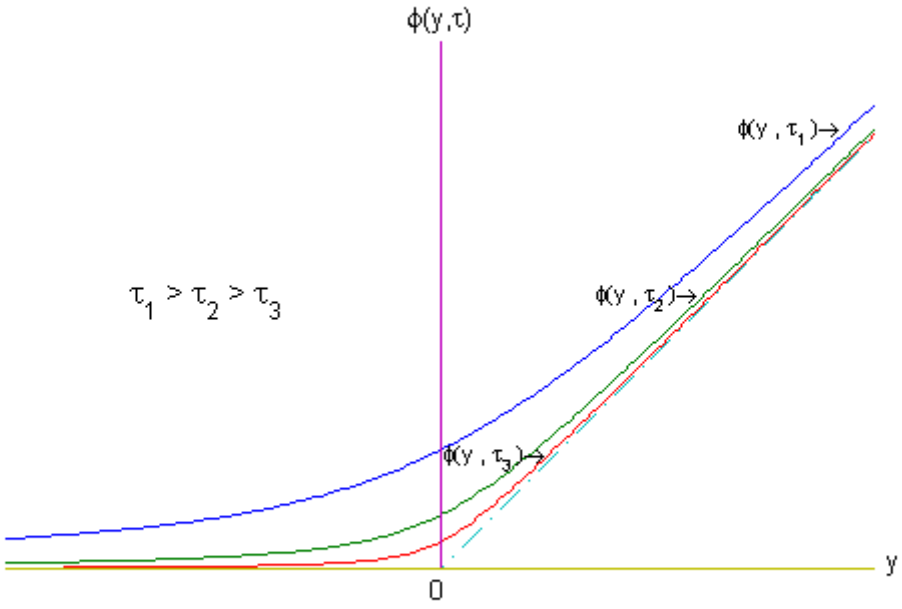


Figura 3.3 – Comportamento da $\phi(y,\tau)$ quando $\tau \to 0$.

Dadas as propriedades acima descritas, a função ψ, presente no problema (10), pode perfeitamente ser substituída pela ϕ, originando o seguinte problema modificado:

$$\begin{aligned} &minimizar \sum_{j=1}^{m} z_j \\ &sujeito\ a: \sum_{i=1}^{q} \phi(z_j - \left\| s_j - x_i \right\|_2, \tau) \geq \varepsilon, \quad j = 1,\ldots,m. \end{aligned} \tag{12}$$

Assumindo as mesmas hipóteses descritas para a apresentação da Figura 3.1, a Figura 3.4 mostra a representação de três parcelas dos somatórios das restrições do problema (8) e do problema (12) em função de z_j, ou seja, são exibidas as situações antes e após a suavização de cada uma parcela.

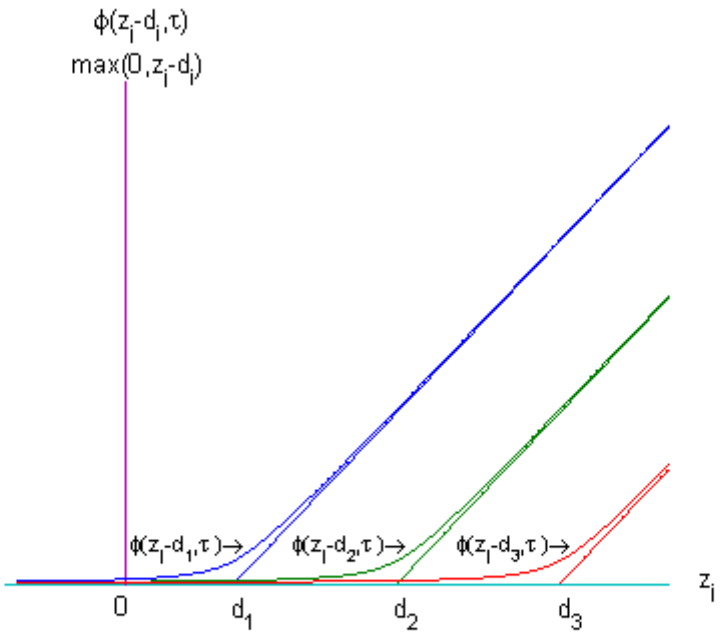
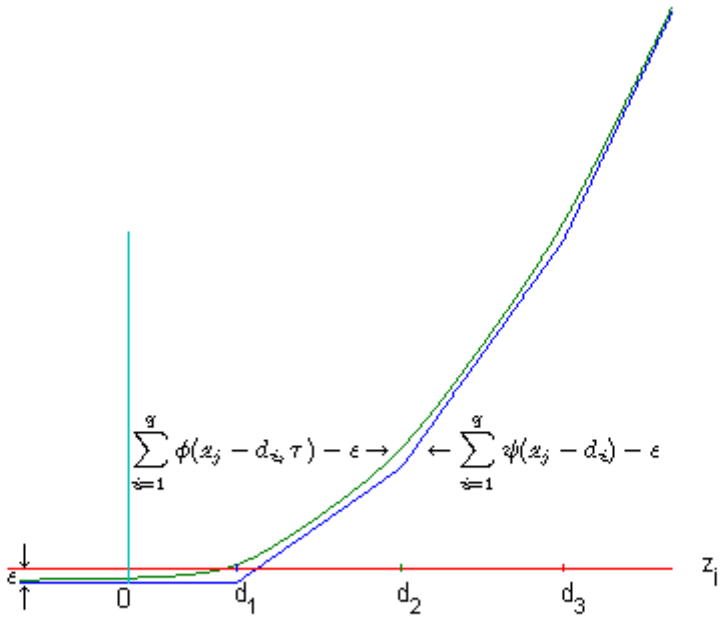


Figura 3.4 Representação de parcelas das restrições de (8) e de (12) em função de z_j.

Considerando o efeito agregado, temos que quando $\tau \to 0$, uma restrição $\sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau)$ tende a $\sum_{i=1}^{q} \psi(z_j - \|s_j - x_i\|_2)$. A Figura 3.5, abaixo, mostra os gráficos de uma restrição do problema perturbado (10) e de sua aproximação dada no problema (12). Nessa figura, podemos ver que ambas são funções crescentes de z_j. Adicionalmente, devido à propriedade (c) da função $\phi(y, \tau)$, para $\tau > 0$, as restrições do problema (12) são monotônicamente crescentes em função de z_j.



'Figura 3.5 – Restrições do problema perturbado (10) e do problema aproximado (12)

Para obtermos um problema completamente diferenciável, ainda é necessário suavizar a distância Euclideana das observações até os centróides $\|s_j - x_i\|_2$. Para tal propósito, para valores do parâmetro $\gamma \neq 0$, introduzimos a função $\theta(s_j, x_i, \gamma)$:

$$\theta(s_j, x_i, \gamma) = \sqrt{\sum_{l=1}^{n} (s_{jl} - x_{il})^2 + \gamma^2} \ . \qquad (13)$$

A função θ possui as seguintes propriedades:

(a) $\lim_{\gamma \to 0} \theta(s_j, x_i, \gamma) = \|s_j - x_i\|_2$;

(b) θ é uma função que pertence à classe de funções C^{∞}.

O gráfico da Figura 3.6 ilustra o comportamento de $\theta(s_j, x_i, \gamma)$ quando $\gamma \to 0$. Nesta figura para facilidade de ilustração, é adotada a convenção $u = \|s_j - x_i\|_2$.

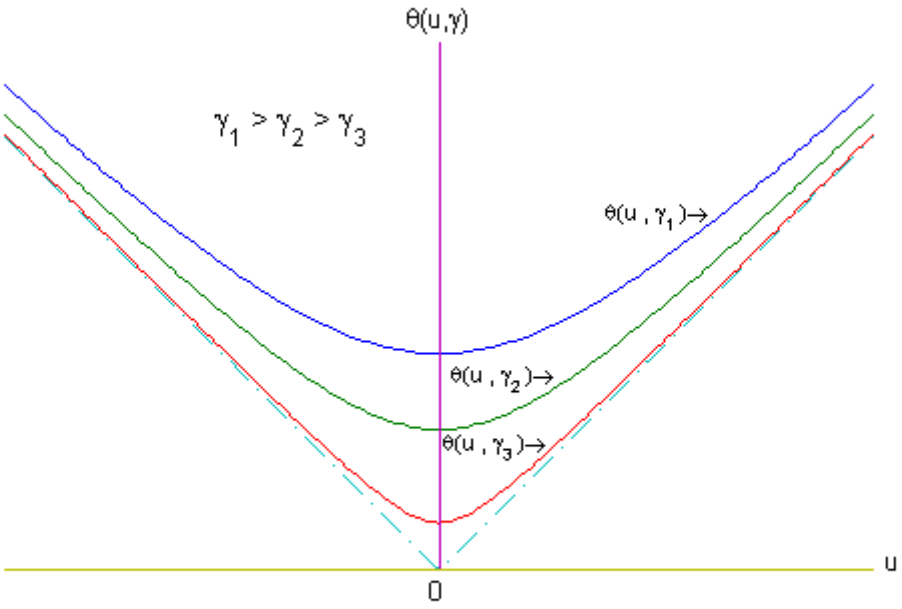


Figura 3.6 Comportamento de $\theta(s_j, x_i, \gamma)$ quando $\gamma \to 0$.

Finalmente, ao se substituir as distâncias Euclideanas $\|s_j - x_i\|_2$ do problema (12) pelas suas aproximações $\theta(s_j, x_i, \gamma)$, obtemos o seguinte problema completamente diferenciável:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j \\ &\textit{sujeito a}: \ \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) \geq \varepsilon, \quad j = 1, \ldots, m. \end{aligned} \tag{14}$$

Nesse problema, a função objetivo trabalha no sentido de minimizar z_j ao extremo, como as restrições do problema são monotônicamente crescentes em função de z_j, todas as restrições serão ativas. Devido a isso, o problema (14) pode ser escrito segundo a seguinte forma equivalente:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j \\ &\textit{sujeito a}: \ h(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \ldots, m \,. \end{aligned} \tag{15}$$

A dimensão do espaço no domínio das variáveis do problema (15) é igual a $(nq + m)$. Assim, o problema possui um número muito grande de variáveis, pois o valor do parâmetro m, que representa a cardinalidade do conjunto S das observações, é geralmente muito grande.

Analisando o problema (15), podemos observar que possui uma estrutura separável , tendo em vista que cada variável z_j aparece somente em uma única restrição de igualdade.

Além da estrutura separável, a derivada parcial de $h_j = (x, z_j)$ em relação a z_j é diferente de zero, logo todas as condições necessárias para o uso do Teorema da Função Implícita são estabelecidas, possibilitando calcular cada componente $z_j, j = 1, \ldots, m$ como função dos centróides $x_i, i = 1, \ldots, q$. Desse modo, é possível formular o problema irrestrito:

$$\textit{minimizar} \ \ f(x) = \sum_{j=1}^{m} z_j(x) \ , \tag{16}$$

onde cada $z_j(x)$ é obtido através do cálculo da raiz de cada equação:

$$h_j(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \ldots, m \tag{17}$$

Devido a propriedade (c) da função $\phi(y, \tau)$, cada termo de (17) é estritamente crescente com z_j, por isso $h_j(x, z_j)$ possui uma única raiz. Novamente, devido ao teorema da função implícita, z_j possui todas as derivadas em relação $x_i, i = 1, \ldots, q$ possibilitando o cálculo do gradiente da função objetivo do problema (16):

$$\nabla f(x) = \sum_{j=1}^{m} \nabla z_j(x).\,, \tag{18}$$

Derivando a equação (17), temos:

$$\frac{\partial h(x, z_j)}{\partial z_j} \nabla z_j(x) + \nabla h(x, z_j) = 0,$$

de onde obtemos:

$$\nabla z_j(x) = -\nabla h(x, z_j) / \frac{\partial h(x, z_j)}{\partial z_j}. \tag{19}$$

O termo $\frac{\partial h(x, z_j)}{\partial z_j}$ é obtido usando as equações (11) e (17)., como segue abaixo:

$$\frac{\partial h(x, z_j)}{\partial z_j} = \sum_{i=1}^{q} \phi'(z_j - \theta(s_j, x_i, \gamma), \tau), \quad j = 1, \ldots, m$$

$$\phi'(z_j - \theta(s_j, x_i, \gamma), \tau) = 1 + \frac{z_j - \theta(s_j, x_i, \gamma)}{\sqrt{z_j - \theta(s_j, x_i, \gamma) + \tau^2}}$$

O termo $\nabla h(x, z_j)$ é obtidos usando as equações (11), (13) e (17)., como segue abaixo

$$\nabla h(x, z_j) = \phi'(z_j - \theta(s_j, x_i, \gamma), \tau) \left[-\nabla \theta(s_j, x_i, \gamma) \right],$$

$$\nabla \theta(s_j, x_i, \gamma) = \frac{1}{(\sum_{l=1}^{n} (s_{jl} - x_{il})^2 + \gamma^2)^{1/2}} \; v,$$

onde $v \in R^{nq}$ é do tipo $v = \left(0, \ldots, 0, x_{i1} - s_{j1}, \ldots, x_{in} - s_{jn}, 0, \ldots, 0\right)^T$ em que as $(i-1)n$ primeiras posições são nulas, seguindo de n posições iguais às componentes dos deslocamentos $(x_i - s_j)$ e seguido de $(i+1)n$ posições iguais a zero.

Desta maneira, é possível resolver o problema (16) fazendo o uso de qualquer método baseado na informação de derivadas de primeira ordem ou de segunda ordem. É importante enfatizar que o problema (16) está definido em um espaço de dimensão nq,

de modo que o problema torna-se pequeno, desde que o número de grupos seja pequeno, como é o casa da maioria das aplicações reais.

3.3 - Algoritmo HSCM

A solução do problema original de clustering pode ser obtida usando o algoritmo HSCM (Hyperbolic Smoothing Clustering Method), descrito de modo simplifico a seguir:

Algoritmo HSCM (Hyperbolic Smoothing Clustering Method)

Passo de Inicialização:

Escolha valores iniciais para: x^0, γ^1, τ^1, ε^1.

Escolha $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, $0 < \rho_3 < 1$, e defina $k = 1$.

Passo Principal: Repita indefinidamente até que uma regra de parada seja atendida.

1) Resolva o problema (16) com $\gamma = \gamma^k$, $\tau = \tau^k$, $\varepsilon = \varepsilon^k$, iniciando em um ponto inicial x^{k-1}, e seja x^k a solução obtida.
2) Calcule os novos parâmetros $\gamma^{k+1} = \rho_1 \gamma^k$, $\tau^{k+1} = \rho_2 \tau^k$, $\varepsilon^{k+1} = \rho_3 \varepsilon^k$.
   Faça $k = k + 1$.

De modo análogo a outros métodos de suavização, a solução para o problema de agrupamento é obtida através da resolução de uma sequência infinita de subproblemas de minimização irrestritos.

Note que o algoritmo faz τ e γ se aproximarem de zero, de modo que as restrições (14) dos subproblemas resolvidos tendem às do problema (10), simultaneamente o o algoritmo faz ε se aproximar de zero, de modo que o problema (10) se aproxime gradualmente do problema (9).

A seguir, apresentaremos o algoritmo de Weiszfeld destinado a localização de um único centróide para o problema de Fermat-Weber. Esse algoritmo será articulado

ao algoritmo HSCM para oferecer um aprimoramento local nas soluções intermediárias produzidas pelo HSCM.

3.4 - Algoritmo de Weiszfeld

Por volta de 1600 Pierre de Fermat propôs o seguinte problema: "Dados três pontos num plano, encontre um quarto ponto de modo que a soma das distâncias deste ponto aos demais seja mínima". Este problema é conhecido como o problema de Fermat (STRECK, 2009)

A mediana geométrica de um conjunto de pontos em um espaço Euclideano é o ponto que minimiza a soma das distâncias aos demais pontos do conjunto. Essa é uma generalização da mediana tradicional, que minimiza a soma em uma única dimensão.

O problema de encontrar a mediana geométrica ficou conhecido como problema de Fermat-Weber (WESOLOWSKY, 1993). O nome Fermat está associado ao problema de Fermat com um único centro, que é um caso particular do problema de Fermat-Weber. Já o nome de Weber é devido a Alfred Weber que discutiu o problema com diversos centros em 1909, no seu livro sobre localização de facilidades.

Em 1937, WEISZFELD propôs o seguinte método iterativo para a resolução do problema de Fermat-Weber com um único grupo:

$$w^{k+1} = \frac{\left(\sum_{j=1}^{m} \frac{s_j}{s_j - w^k} \right)}{\left(\sum_{j=1}^{m} \frac{1}{s_j - w^k} \right)}$$

A partir de um centro w^k, calcula-se um novo valor, w^{k+1}, que se aproxima gradativamente do valor da mediana.

3.5 - Articulação do Algoritmo HSCM com o Algoritmo de Weiszfeld

No passo 2 do algoritmo Suavização hiperbólica, na iteração k uma solução intermediária x^k é obtida. Evidentemente, esses q centróides constituintes de x^k definem uma partição do conjunto inicial das observações, em q partes, grupos ou clusters.

Nesse contexto, a utilização do algoritmo de Weiszfeld isoladamente para cada um dos q grupos produzirá uma solução de melhor qualidade do que esse ponto x^k intermediário. O valor da função objetivo original será reduzido, pois será obtida a localização ótima da mediana geométrica separadamente para cada grupo.

A solução do problema original de clustering pode ser obtida usando o Algoritmo Hyperbolic Smoothing Clustering articulado com o algoritmo de Weiszfeld. Essa articulação é descrita de modo simplificado a seguir:

Algoritmo HSCM articulado ao Algoritmo de Weiszfeld

Passo de Inicialização:

Escolha valores iniciais para: x^0, γ^1, τ^1, ε^1.

Escolha $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, $0 < \rho_3 < 1$, e defina k=1

Passo Principal: Repita indefinidamente até que uma regra de parada seja atendida.

1) Resolva o problema (16) com $\gamma = \gamma^k$, $\tau = \tau^k$, $\varepsilon = \varepsilon^k$, iniciando no ponto inicial x^{k-1}, e seja x^k a solução intermediária obtida.
2) Faça partição do conjunto de observações em q grupos usando a solução intermediária x^k obtida acima.
3) Use o algoritmo de Weiszfeld separadamente em cada um dos q grupos, obtendo ponto central formado pelas medianas $w^k = (w_1^k, \ldots, w_q^k)$.
4) Calcule a função objetivo original, definida pelas equações (1) e (2), usando o novo ponto central w^k construído através do algoritmo Weiszfeld.

5) Calcule os novos parâmetros $\gamma^{k+1} = \rho_1 \gamma^k$, $\tau^{k+1} = \rho_2 \tau^k$, $\varepsilon^{k+1} = \rho_3 \varepsilon^k$.

Faça $k = k + 1$.

4 - Resultados Computacionais

Neste capítulo serão apresentados os resultados computacionais obtidos pela proposta metodológica apresentada no capítulo anterior. O capítulo será dividido em duas partes: descrição geral dos experimentos e apresentação de cada um dos resultados.

4.1 - Descrição geral dos experimentos

Nesta primeira parte, alguns detalhes dos experimentos serão abordados, como a configuração do computador, o compilador utilizado, a rotina de otimização, a escolha do ponto inicial, a especificação dos parâmetros de suavização e a regra de parada.

4.1.1 - Características do Computador, Compilador e Rotina de otimização

Alguns experimentos computacionais foram realizados e comparados com os melhores resultados que se encontram na literatura. Para os experimentos computacionais foi utilizado um LAPTOP POSITIVO PREMIUM, com processador da Intel, Pentium Dual-Core T4300, 2.1 GHz e 4 GBytes de memória RAM.

O algoritmo foi implementado na linguagem Fortran 77, utilizando-se o compilador Compaq Visual Fortran versão 6.1.0. Para a realização dos procedimentos de minimização irrestrita multidimensional, foi utilizado o método Quase-Newton, com atualização da matriz hessiana dada pela forma BFGS, na implementação VA13C da Harwell Library, biblioteca disponibilizada gratuitamente em: http://www.cse.scitech.ac.uk/nag/hsl/

4.1.2 - Escolha dos pontos iniciais

A geração dos centróides iniciais é uma questão importante para qualquer algoritmo, pois enfrentamos uma grande dificuldade, que é a presença de inúmeros pontos de mínimo local. Assim, dado o problema contemplado ser de natureza de otimização global, foi adotada a estratégia de se fazer várias tentativas de resolução através de diferentes pontos iniciais. Basicamente neste trabalho, foram adotadas duas opções: 10 tentativas para alguns problemas e 100 tentativas para os demais.

Se pudermos iniciar o processo iterativo com um bom conjunto de centróides iniciais, certamente facilitaria na obtenção de uma boa solução final, entretanto não se tem essa informação a priori. Diante disso, na primeira tentativa, utilizamos o critério de gerar o primeiro conjunto de centróides iniciais através de perturbações aleatórias em torno da média das observações, segundo as seguintes expressões:

$$\bar{s} = \frac{\sum_{j=1}^{m} s_j}{m}$$

$$\sigma_s^2 = \frac{\sum_{i=1}^{n} (s_j - \bar{s})^2}{m}$$

$$x_i^0 = \bar{s} + a\,\sigma_s \;,\; i = 1,\ldots,q$$

As duas primeiras expressões calculam a média e a variância das observações. A última expressão calcula cada um dos q centróides iniciais, $x_i^0 \;,\; i = 1,\ldots,q$, onde a é um vetor cujas componentes são variáveis aleatórias com distribuição uniforme no intervalo $[-0.5, +0.5]$. Logo temos um variação em torno da média $\bar{s}$, que é proporcional ao desvio padrão σ_s, multiplicado pelo número aleatório a.

Nas demais tentativas, a partir da solução obtida anterior, podemos usar uma perturbação nesta solução para gerar um novo conjunto de pontos iniciais. Temos, todavia, um dilema entre uma pequena perturbação ou uma grande perturbação. Uma pequena perturbação é excelente quando a solução que foi perturbada é boa, pois o novo

ponto inicial herda boas propriedades estruturais intrínsecas ao anterior. Em contrapartida, se a solução não é boa, ou seja, um mínimo local não profundo, uma grande perturbação seria mais adequada.

Durante as atividades preliminares de calibração do algoritmo, mostrou-se mais adequada uma perturbação que provocasse mudanças substantivas em parte dos centróides, enquanto, outra parte mantivesse nas vizinhanças bem próximas aos anteriores. Esses dois objetivos simultâneos foram alcançados pela seguinte expressão:

$$x_i^0 = x_i^* + 4\,a\,(0.75\,R_i + 0.25\,\overline{R}\,), \quad i = 1,\ldots,q,$$

onde, x_i^0 é o novo centróide genérico inicial i na próxima tentativa, x_i^* é o centróide ótimo obtido na tentativa anterior, a é um vetor cujas componentes são variáveis aleatórias com distribuição uniforme no intervalo $[-0.5, +0.5]$, R_i é a maior distância das observações do grupo i ao centróide do grupo x_i^* e $\overline{R}$ é a média dos valores R_i.

4.1.3 - Especificação dos parâmetros de suavização

No desenvolvimento da presente implementação do algoritmo, um esforço foi dirigido na escolha adequada dos parâmetros iniciais, bem como nos valores dos fatores de decrescimento dos parâmetros.

Os experimentos empíricos indicaram a seguinte escolha. O parâmetro da Suavização Hiperbólica τ é por excelência um parâmetro vital no processo de suavização, pois dá diretamente o nível da mesma. A escolha de um valor pequeno implica obviamente num problema muito nervoso ou duro, próximo à natureza não diferenciável dada pela função patamar ψ presente na equação (8). A escolha de um valor grande implica suavização excessiva, distanciamento grande do problema original e falta de sensibilidade (PAULA Jr., 2010). O parâmetro da perturbação foi fixado como $\varepsilon = \tau / 4$. O parâmetro da suavização da distância euclidiana foi fixado como

$\gamma = \tau / 100$. Em relação aos fatores de redução dos parâmetros, adotamos um mesmo valor para todos: $\rho_1 = \rho_2 = \rho_3 = \sqrt{10}/10$.

4.1.4 - Regra de Parada

Os testes foram realizados adotando-se um único critério de parada, o número máximo de iterações do algoritmo. Para as instâncias testadas, a convergência ocorre entre a primeira iteração e a sétima iteração, sendo que, para a maioria dos casos ocorre a convergência em até 4 iterações. Por segurança, o número máximo de iterações escolhido foi superestimado e fixado em oito iterações.

A configuração dos dados influencia diretamente a convergência do algoritmo. Por exemplo, quando temos claramente três grupos bem separados, é natural que o método convirja para a solução em um número menor de iterações quando se especifica 3 grupos do que, quando se especifica 2 ou 4 grupos.

4.2 - Apresentação dos resultados

Para mostrar a eficiência e confiabilidade do algoritmo, serão apresentados nesta seção resultados computacionais obtidos na resolução de algumas instâncias da biblioteca TSPLIB, (REINELT, 1991). Não obstante ser destinada originalmente a problemas TSP, caixeiro-viajante, essa biblioteca é amplamente utilizada para diferentes problemas de agrupamento, como o das p-medianas. As instâncias P654 e U1060 são as mais utilizadas na literatura, por isso, são consideradas neste trabalho. Além dessas, foram consideradas as instâncias PCB3038, PLA7397, USA13509, D15112, D18512, PLA33810 e PLA85900.

4.2.1 - Problema teste P654

Algumas propriedades do método serão apresentadas tendo com exemplo o conjunto de dados P654. Esse problema teste foi escolhido, como primeiro exemplo,

pois é um dos mais referenciados na literatura. A figura 4.1 mostra as observações deste conjunto.

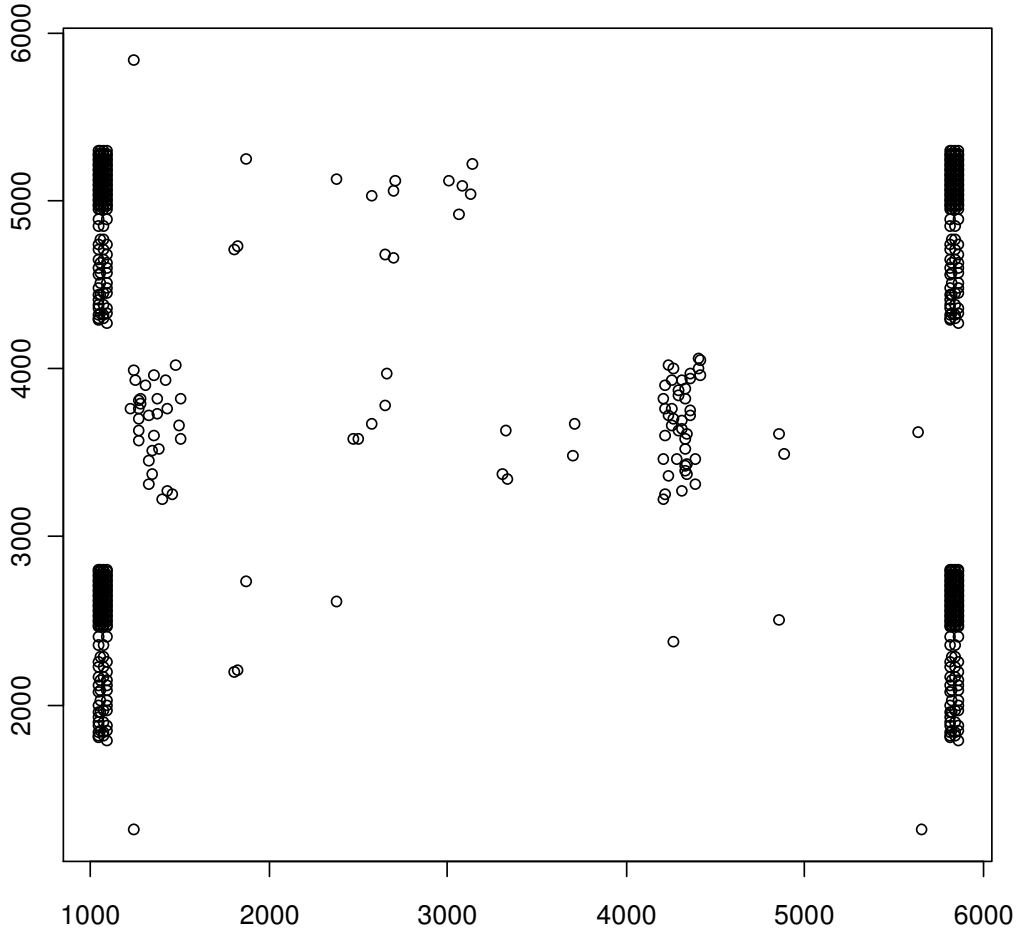


Figura 4.1- Conjunto de dados P654

A Figura 4.2, apresenta este conjunto de dados agrupado em dois grupos. As observações estão ligadas aos seus respectivos centróides por segmentos de reta, com uma cor associada a cada grupo.

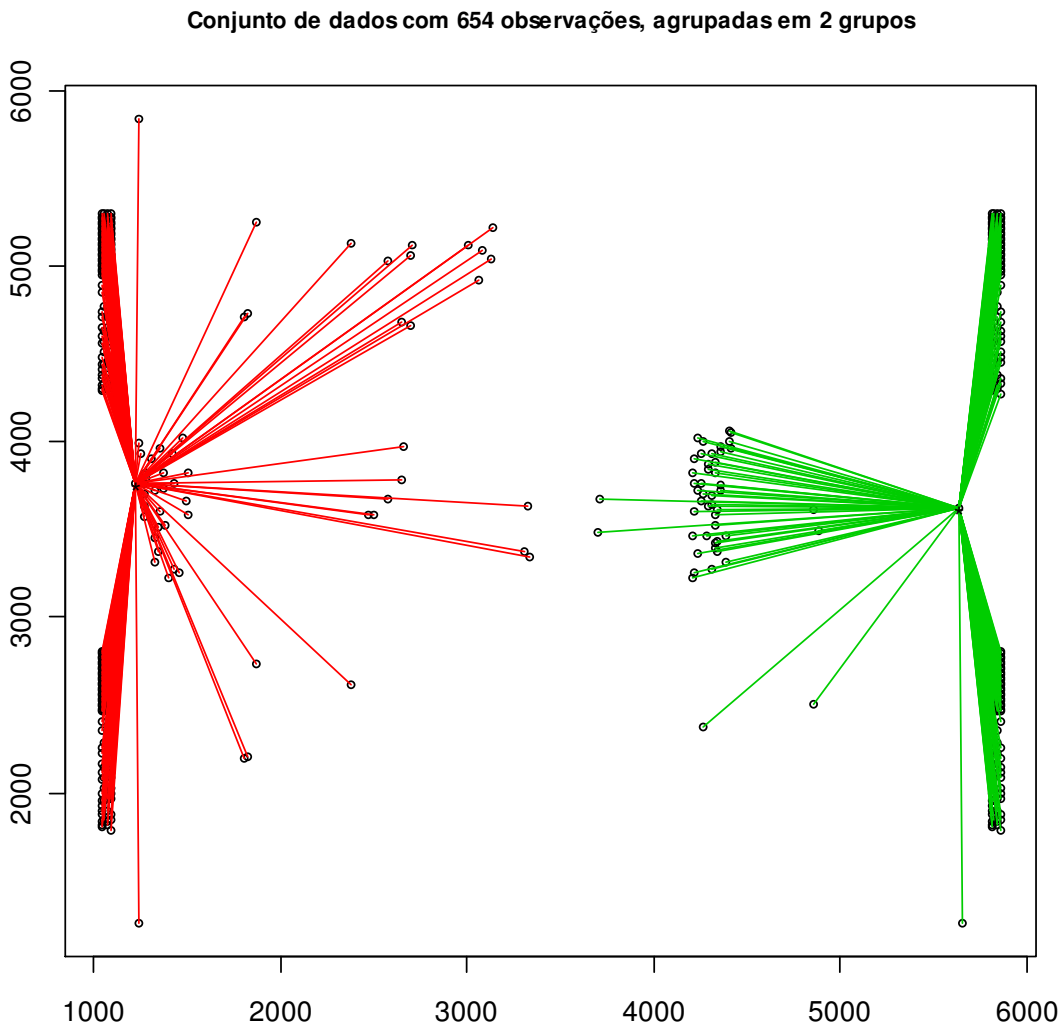


Figura 4.2 – Resultado para P654, caso 2 grupos

A Tabela 4.1 mostra uma típica sequência de centróides intermediários obtidos no processo iterativo no passo 1 do algoritmo. As cinco colunas da tabela representam, na ordem, a iteração e as duas componentes dos centróides.

A sequência de centróides, por sua vez, define uma sequência de partições e suas respectivas sequências de medianas, obtidas no passo 3 pelo algoritmo de Weiszfeld. A Tabela 4.2 mostra para cada iteração o valor da função objetivo suavizada calculada pela equação (16), usando os centróides, e o valor exato da função objetivo calculado pelas medianas, usando as equações (1) e (2). A linha $k = 0$ mostra os centróides do ponto inicial.

.

Tabela 4.1 - Sequência dos centróides no processo iterativo.

Iteração k	Coordenadas do Centróide 1		Coordenadas do Centróide 2	
	x_{11}	x_{12}	x_{21}	x_{22}
0	2256,21	3212,92	3610,97	4185,99
1	1219,18	3754,70	5620,76	3643,17
2	1221,13	3767,12	5628,49	3618,00
3	1224,29	3766,00	5633,07	3618,93
4	1224,32	3766,00	5633,11	3618,94
5	1224,32	3766,00	5633,11	3618,94
6	1224,32	3766,00	5633,11	3618,94
7	1224,32	3766,00	5633,11	3618,94
8	1224,32	3766,00	5633,11	3618,94

Podemos ver na Tabela 4.2 que o valor da função objetivo suavizada diminui monotonamente no decorrer das iterações, na medida em que os parâmetros da suavização vão diminuindo. Além disso, podemos ver que a sequência de valores exatos da função objetivo calculados pelas medianas converge na primeira iteração. Esse fenômeno é decorrente de os centróides desta primeira iteração já produzirem o particionamento ótimo.

Tabela 4.2- Valores da função objetivo para cada iteração

Iteração	*Fsuav,*	$F_{medianas}$
1	1389530,5	815313,3
2	960083,2	815313,3
3	851561,7	815313,3
4	824377,6	815313,3
5	817579,5	815313,3
6	815879,9	815313,3
7	815454,9	815313,3
8	815348,7	815313,3

Os gráficos apresentados nas Figuras 4.3 e 4.4, que seguem abaixo, mostram as soluções de particionamento obtidas pelo algoritmo proposto para o conjunto de dados P654, classificados de 3 a 10 grupos. Nesses gráficos, cada grupo possui uma cor

diferente. Novamente os segmentos de reta ligam cada observação ao seu respectivo centróide.

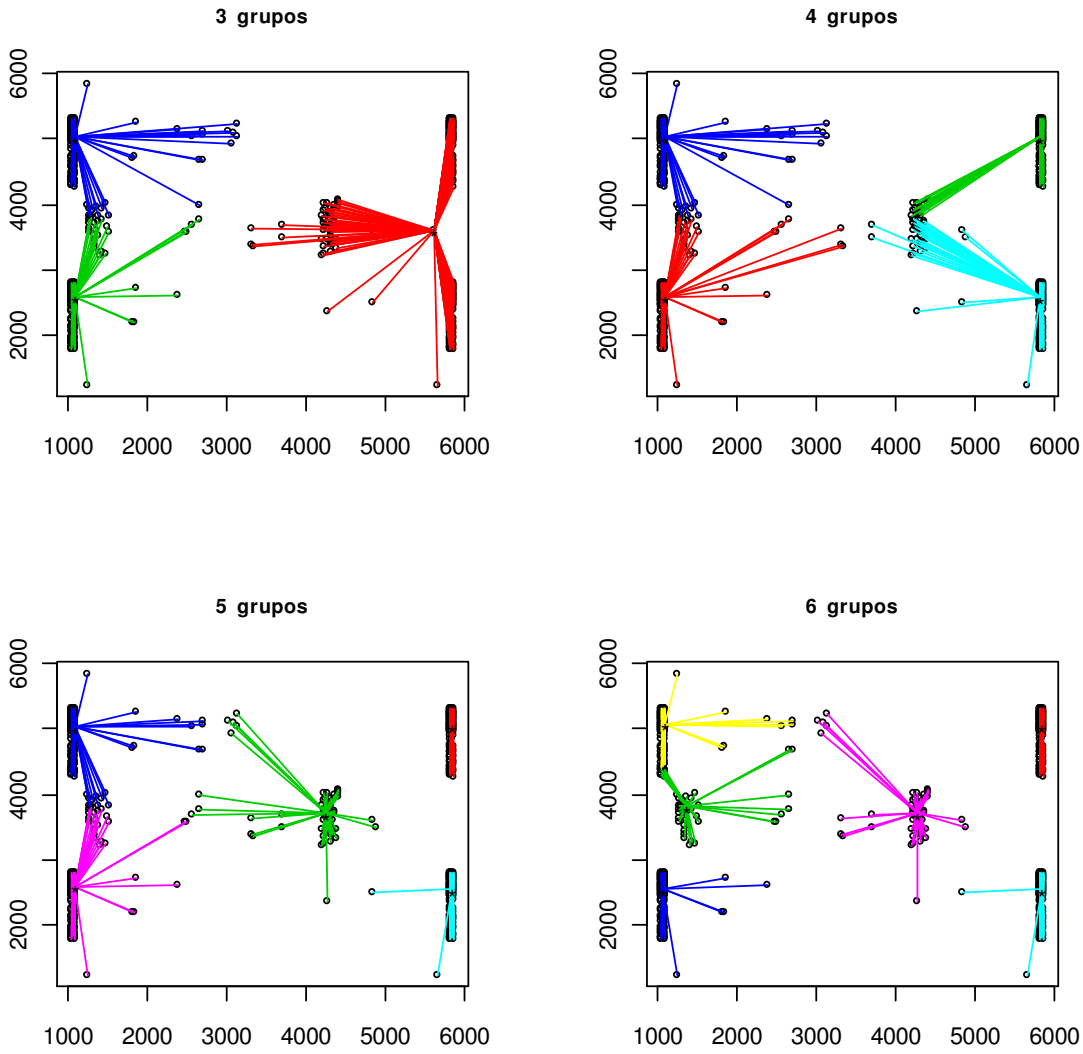


Figura 4.3- Observações agrupadas 3 a 6 grupos.

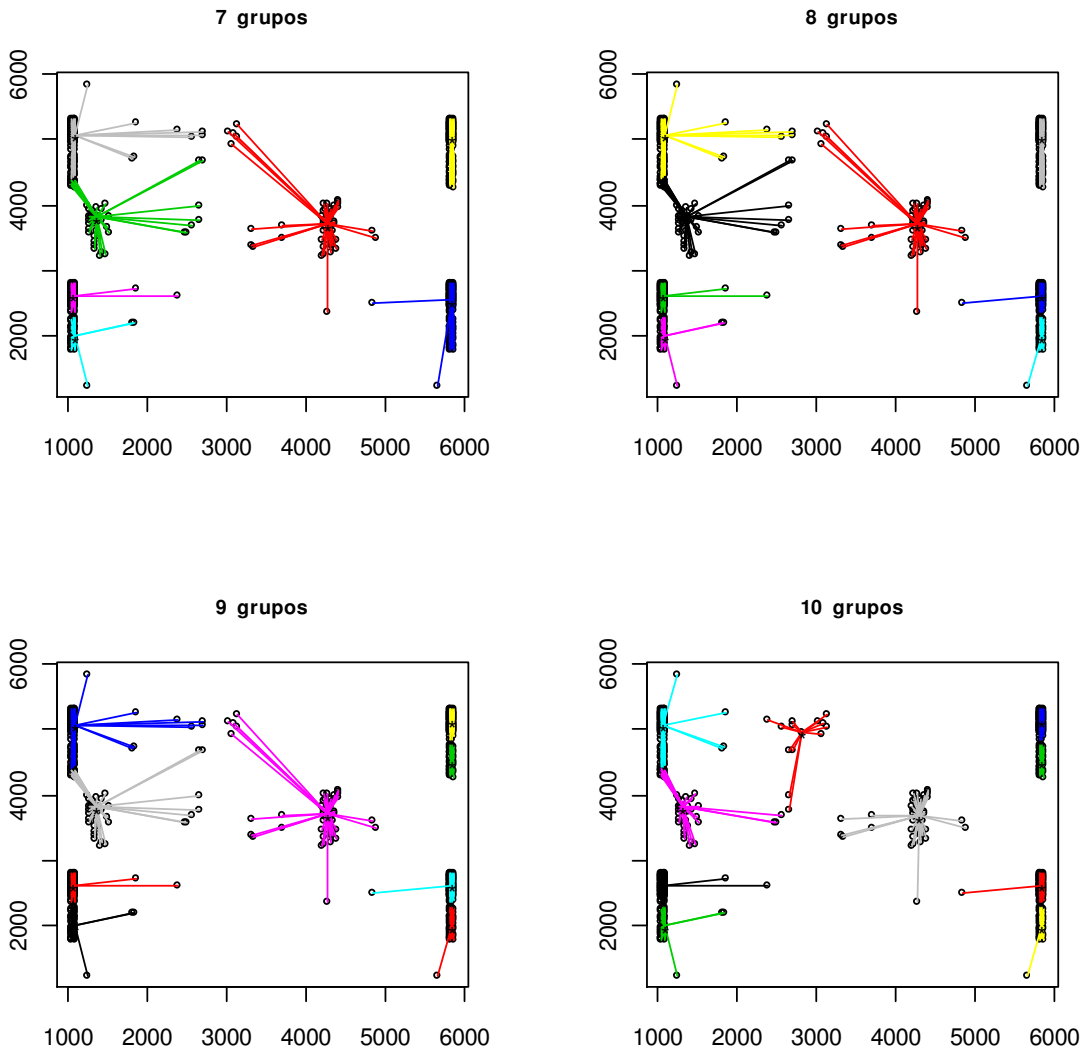


Figura 4.4- Observações agrupadas de 7 a 10 grupos.

O artigo "Improvements and Comparison of Heuristic for Solving the Multisource Weber Problem" (BRIMBERG *et al.*, 2000) apresenta uma coletânea dos melhores resultados, para os problemas P654 e U1060, obtidos pelos diferentes métodos disponíveis na literatura.

A Tabela 4.3 mostra os resultados obtidos pelo método proposto para o problema P654. Nas cinco primeiras colunas temos: q = número de grupos; f_{min} = o melhor valor da função objetivo encontrado em 100 tentativas; Ocorrência = número de ocorrências do menor valor da função objetivo; Erro Médio = erro médio percentual das

tentativas e Tempo Médio = tempo médio de CPU em segundos. Nas duas últimas colunas, temos os valores dos melhores resultados obtidos em BRIMBERG *et al*. (2000) e o erro comparativo percentual do método proposto em relação aos melhores resultados.

O erro médio percentual das tentativas é calculado em relação a melhor solução pela seguinte fórmula:

$$Erro\ Médio\ = \sum_{t=1}^{T} \frac{\frac{(f_t - f_{\min})}{f_{\min}}.100}{T}$$

onde f_t é o valor da função objetivo na tentativa t, $f_{\min}$ é o menor valor encontrado em todas as tentativas e T é o número de tentativas. Nesse particular exemplo, $T = 100$.

O erro comparativo percentual é calculado pela seguinte fórmula:

$$Erro\ Comparativo = \frac{(f_{\min} - f_{ot})}{f_{ot}}.100,$$

onde f_{ot} foi obtido em BRIMBERG *et al*. (2000).

Para os casos de 2 até 14 grupos, o método proposto obteve os melhores resultados conhecidos na literatura, com apenas 100 pontos iniciais. Nos demais casos, a solução obtida foi bem próxima da melhor solução conhecida, com erro relativo máximo de 5,08, apresentado no caso de 70 grupos. Para 2 grupos, todas as 100 tentativas levaram a mesma solução.

Tabela 4.3 - Resultados para o problema P654

Hyperbolic Smoothing Clustering					Melhor solução conhecida F_{ot}	
q	Fmin	Ocorrência	Erro Médio	Tempo Médio	Fot	Erro Comparativo
2	815313,0	100	0.00	0.07	815313,0	0,00
3	551063,0	64	0.07	0.13	551063,0	0,00
4	288191,0	81	14.12	0.20	288191,0	0,00
5	209069,0	71	12.40	0.27	209069,0	0,00
6	180488,0	29	5.73	0.39	180488,0	0,00
7	163704,0	13	5.64	0.55	163704,0	0,00
8	147051,0	11	4.83	0.73	147051,0	0,00
9	130936,0	11	5.91	0.89	130936,0	0,00
10	115339,0	20	7.61	1.13	115339,0	0,00
11	100133,0	14	13.87	1.37	100133,0	0,00
12	94152,1	12	11.03	1.58	94152,05	0,00
13	89454,8	6	8.82	1.96	89454,76	0,00
14	84807,7	3	7.56	2.26	84807,69	0,00
15	80198,0	9	7.41	2.65	80177,04	0,03
20	63640,9	3	9.04	5.97	63389,02	0,40
25	54619,5	1	11.11	9.26	52209,51	4,62
30	46409,8	1	7.78	15.04	44705,19	3,81
35	40300,9	1	10.67	24.27	39257,26	2,66
40	36364,1	1	3.11	29.20	35704,40	1,85
45	33807,0	1	6.87	43.50	32306,97	4,64
50	30372,9	1	6.31	50.77	29338,01	3,53
55	27060,4	5	1.57	60.86	26699,16	1,35
60	25488,2	3	2.91	62.20	24504,39	4,01
65	23377,4	3	2.59	88.08	22747,09	2,77
70	22559,4	1	2.25	100.30	21468,15	5,08
75	20900,3	1	2.71	117.13	20312,96	2,89
80	19936,2	1	1.97	153.02	19193,88	3,87
85	18814,6	1	4.26	192.87	18316,53	2,72
90	18395,6	1	2.94	222.41	17544,35	4,85
95	17526,7	1	1.98	229.22	16786,38	4,41
100	16505,3	1	1.99	280.44	16087,68	2,60

4.2.2 - Problema teste U1060

No problema teste U1060, de modo análogo ao problema anterior, também foram realizadas comparações com os melhores resultados conhecidos na literatura.

Tabela 4.4 – Resultados para o problema U1060

Hyperbolic Smoothing Clustering					Melhor solução conhecida$_t$	
q	Fmin	Ocorrência	Erro médio	Tempo médio	f_{ot}	Erro Comparativo
2	3011160	100	0.00	0.30	X	X
3	2426090	100	0.00	0.61	X	X
4	2064240	100	0.00	0.90	X	X
5	1852610	98	0.01	1.35	1851879,9	0,04
6	1692620	5	0.08	1.06	X	X
7	1558660	42	0.28	1.45	X	X
8	1425090	94	0.21	1.79	X	X
9	1326990	78	0.41	2.17	X	X
10	1250450	44	0.21	3.01	1249564,8	0,07
15	980617	9	1.11	6.23	980132,1	0,05
20	829249	1	1.55	11.39	828802,0	0,05
25	721850	4	2.09	18.12	722061,2	-0,03
30	638686	1	2.60	26.89	638263,0	0,07
35	579280	1	2.89	37.58	577526,6	0,30
40	531706	1	2.91	52.40	529866,2	0,35
45	492658	1	2.57	78.72	489650,0	0,61
50	458641	1	2.47	83.41	453164,0	1,21
55	428909	1	2.68	104.34	422770,0	1,45
60	405151	1	2.57	124.67	397784,4	1,85
65	381491	1	2.60	150.06	376759,5	1,26
70	361719	1	2.97	176.02	357385,0	1,21
75	346522	1	2.74	208.66	340242,0	1,85
80	333114	1	2.61	250.60	326053,2	2,17
85	318698	1	3.07	279.82	313738,2	1,58
90	310766	1	2.43	324.98	302837,0	2,62
95	297720	1	3.16	386.20	292875,1	1,65
100	287294	1	2.70	423.36	283113,0	1,48

Na Tabela 4.4, por convenção, os campos onde não se tem a melhor solução conhecida, f_{ot} , foram preenchidos com a letra X. O resultado obtido para 25 grupos ($q = 25$) , é menor do que o respectivo valor da melhor solução conhecida. Nos demais casos, o algoritmo proposto, usando somente 100 pontos iniciais, produziu resultados muito próximos aos melhores resultados conhecidos com o erro comparativo máximo de 2,62.

4.2.3 - Problema teste PCB3038

Este problema teste é frequentemente utilizado em problemas de agrupamento onde se minimiza a soma das distâncias ao quadrado. Alguns resultados são encontrados na literatura para o problema discreto da p-mediana. Para o caso contínuo, apenas quando o número de centróides é muito grande, maior ou igual a 100, somente poucos resultados são registrados através de métodos heurísticos (BRIMBERG *et al*., 2000).

Tabela 4.5 - Resultados para o problema PCB3038

q	F_{min}	Ocorrência	Erro médio	Tempo médio
2	2898540	10	0.00	0.88
3	2374450	8	0.12	1.74
4	1977870	10	0.00	2.54
5	1778510	6	0.01	2.09
6	1605930	10	0.00	4.43
7	1485820	1	0.50	6.12
8	1385190	3	0.27	6.91
9	1296830	8	0.25	5.45
10	1212050	6	0.31	6.70
15	969868	3	0.50	16.19
20	839442	4	0.17	26.61
25	748343	1	0.19	40.07
30	676691	1	0.28	61.74
35	616658	1	0.46	75.58
45	536030	1	0.29	164.03
50	506081	1	0.63	165.90

A Tabela 4.5 apresenta os resultados para o problema PCB3038. Podemos ver que em dez tentativas, o método convergiu em todas as tentativas para as mesmas soluções para os casos de 2, 4 e 6 grupos. Pode-se também observar que, à medida que o número de grupos aumenta, há uma tendência de decréscimo desses valores. De outro lado, os baixos valores exibidos pela coluna Erro médio, mostram, de forma clara, a consistência computacional do método proposto.

4.2.4 - Problema teste PLA7397

A Tabela 4.6 apresenta os resultados para o problema PLA7397. Podemos ver que nas dez tentativas realizadas, o método convergiu para as mesmas soluções para 2 e 3 grupos. À medida que o número de grupos aumenta, de uma forma análoga ao teste anterior, o número de ocorrências da melhor solução encontrada diminui. Esse fenômeno observado no conjunto de experimentos é muito natural, pois a complexidade do problema e o número de mínimos locais crescem com o número de grupos.

Tabela 4.6 - Resultados para o teste PLA7397

q	F_{min}	Ocorrência	Erro médio	Tempo médio
2	1009631094,88	10	0.00	1.44
3	775739054,15	10	0.00	3.01
4	604453117,00	6	8.95	5.40
5	531898862,70	5	10.91	8.62
6	470587831,11	7	2.47	11.42
7	436259767,42	4	4.27	14.77
8	402525675,30	5	2.01	19.97
9	384292455,95	9	0.65	27.22
10	367205709,70	3	1.97	32.48
15	304017244,50	1	1.74	81.59
20	247183124,65	1	1.73	137.12
25	220151679,01	1	1.14	226.17
30	201378489,09	1	1.40	365.42

4.2.5 - Problema teste USA13509

Este problema teste possui 13509 observações e foi criado através da localização das cidades dos Estados Unidos com mais de 500 habitantes.

A Tabela 4.7 mostra resultados para o problema USA13509. De 2 a 5 grupos o erro médio é zero, pois o método convergiu para a mesma solução. Para os demais casos com maior número de grupos o erro médio permaneceu muito pequeno. Esse comportamento indica, uma vez mais, a consistência computacional da metodologia proposta.

Tabela 4.7 - Resultados para o teste USA13509

q	F_{min}	Ocorrência	Erro médio	Tempo médio
2	1,07194E+09	10	0.00	6.64
3	7,99993E+08	10	0.00	9.52
4	6,80198E+08	10	0.00	11.27
5	5,85900E+08	10	0.00	18.29
6	5,25966E+08	9	0.39	22.12
7	4,84741E+08	9	0.21	28.32
8	4,50157E+08	3	0.78	35.47
9	4,21673E+08	5	1.30	41.96
10	4,08167E+08	3	0.71	55.71
15	3,17216E+08	5	0.45	125.02
20	2,66171E+08	3	0.96	251.79
30	2,10216E+08	1	2.27	593.38
40	1,79911E+08	1	1.27	1195.68

A Figura 4.5 (http://www.tsp.gatech.edu/gallery/idata/usa13509.html) mostra este conjunto de dados. É possível ver claramente um esboço do mapa dos Estados Unidos.

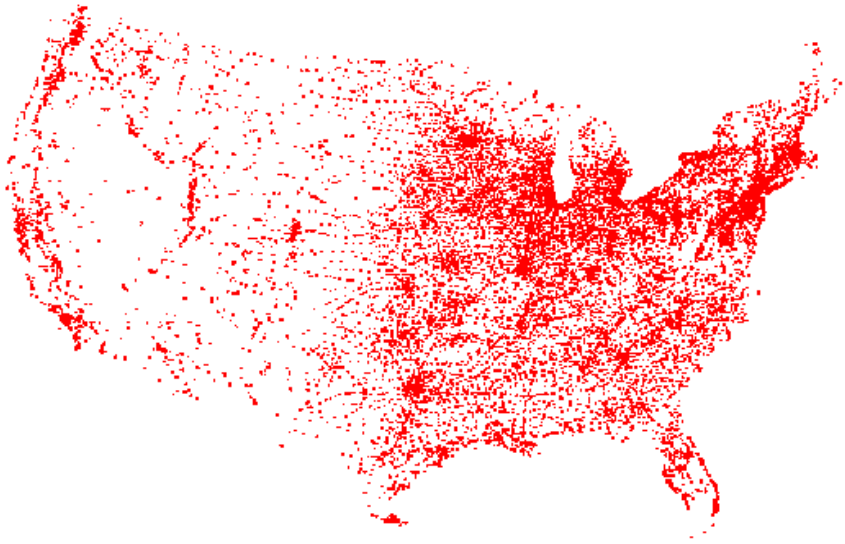


Figura 4.5- Conjunto de dados USA13509

4.2.6 - Problema teste D15112

A Tabela 4.8 mostra resultados para o problema teste D15112, que corresponde a um conjunto de 15112 cidades da Alemanha. As colunas têm os mesmos significados previamente definidos. De modo análogo aos resultados anteriores, o erro médio é muito pequeno, mostrando consistência da proposta metodológica.

Tabela 4.8 - Resultados para o problema D15112

q	F_{min}	Ocorrência	Erro médio	Tempo médio
2	6,88513E+07	10	0.00	2.77
3	5,49316E+07	10	0.00	5.37
4	4,59830E+07	10	0.00	7.62
5	4,01359E+07	9	0.79	12.04
6	3,66573E+07	6	0.89	16.79
7	3,39779E+07	10	0.00	21.48
8	3,17781E+07	7	0.91	28.70
9	3,01750E+07	7	0.44	36.77
10	2,85082E+07	5	0.61	44.59
15	2,32074E+07	3	0.17	95.26
20	1,98848E+07	1	0.19	189.48
25	1,77769E+07	1	0.48	288.57

4.2.7 - Problema teste D18512

A figura 4.6 mostra o conjunto de dados D18512, que corresponde a um conjunto de 18512 cidades da Alemanha.

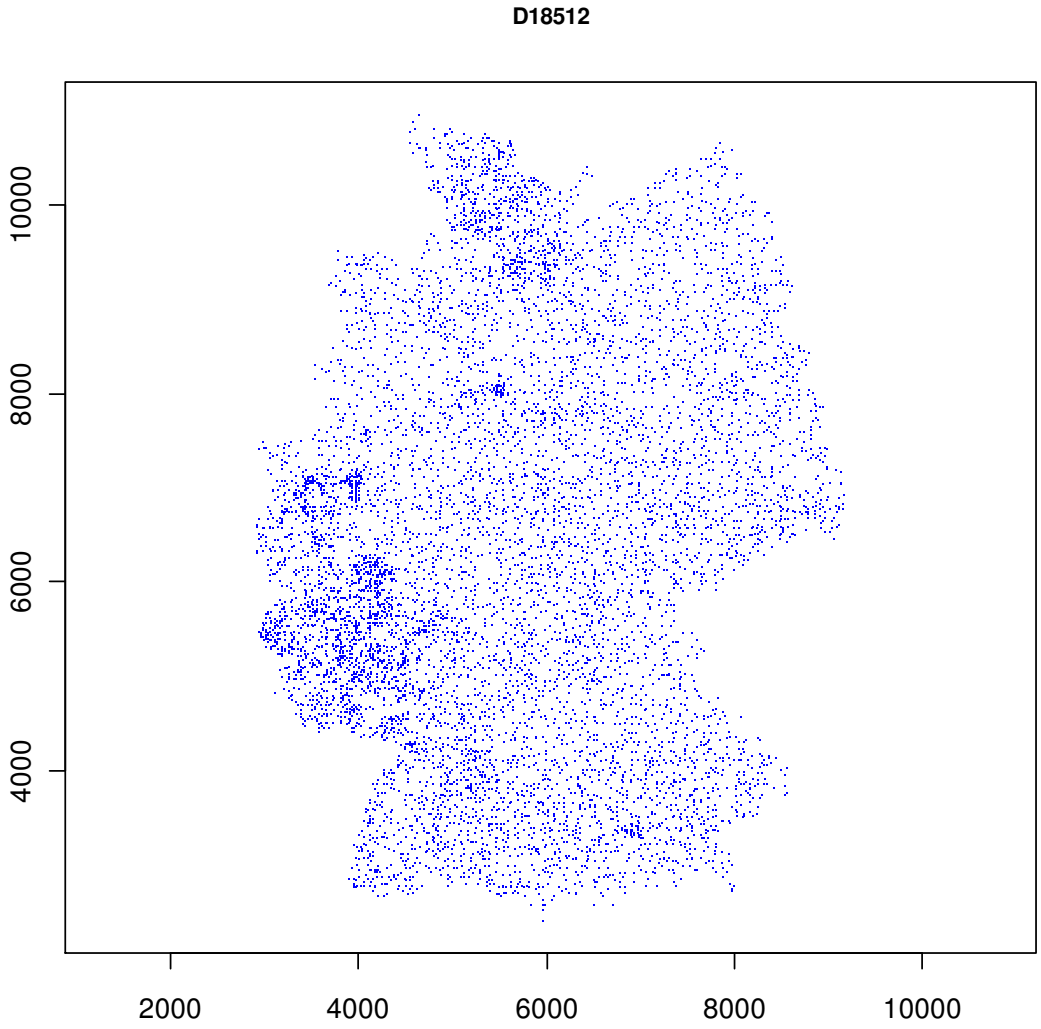


Figura 4.5- Conjunto de dados D18512

A Tabela 4.9 mostra resultados para o problema D18512. Para os casos de 2 a 5 grupos o erro médio é zero, pois o método convergiu para a mesma solução. Para os demais casos o erro permaneceu muito pequeno. Para esse problema teste, o número de tentativas foi fixado em 100 para cada caso.

Tabela 4.9 - Resultados para o problema D18512

q	F_{min}	Ocorrência	Erro médio	Tempo médio
2	30769651,65	100	0.00	3.26
3	24712136,20	100	0.00	5.68
4	20836111,94	100	0.00	9.31
5	19004859,63	13	0.33	14.41
6	17231136,70	50	0.36	18.91
7	15782176,55	96	0.02	23.59
8	14689649,03	23	0.04	30.13
9	13583885,41	99	0.05	34.81
10	12966043,08	22	0.55	44.72
15	10640348,71	49	0.26	99.39
20	9080013,54	50	0.19	174.23
25	8108354,79	7	0.13	289.54

A figura 4.7 ilustra uma solução para o caso com 2 grupos. Nessa figura, cada grupo possui uma cor diferente e os centróides são os pontos na cor preta.

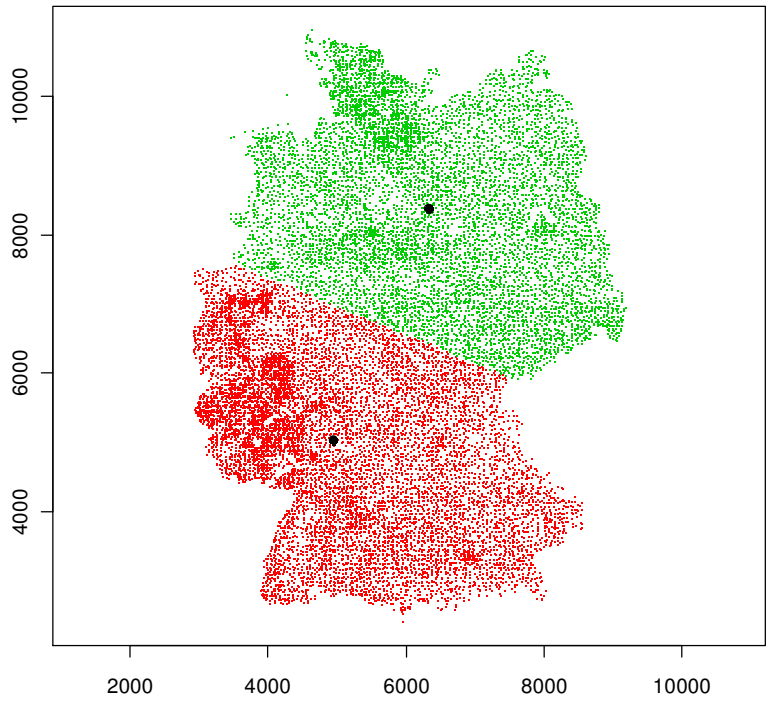


Figura 4.7- Conjunto de dados D18512 agrupados em 2 grupos

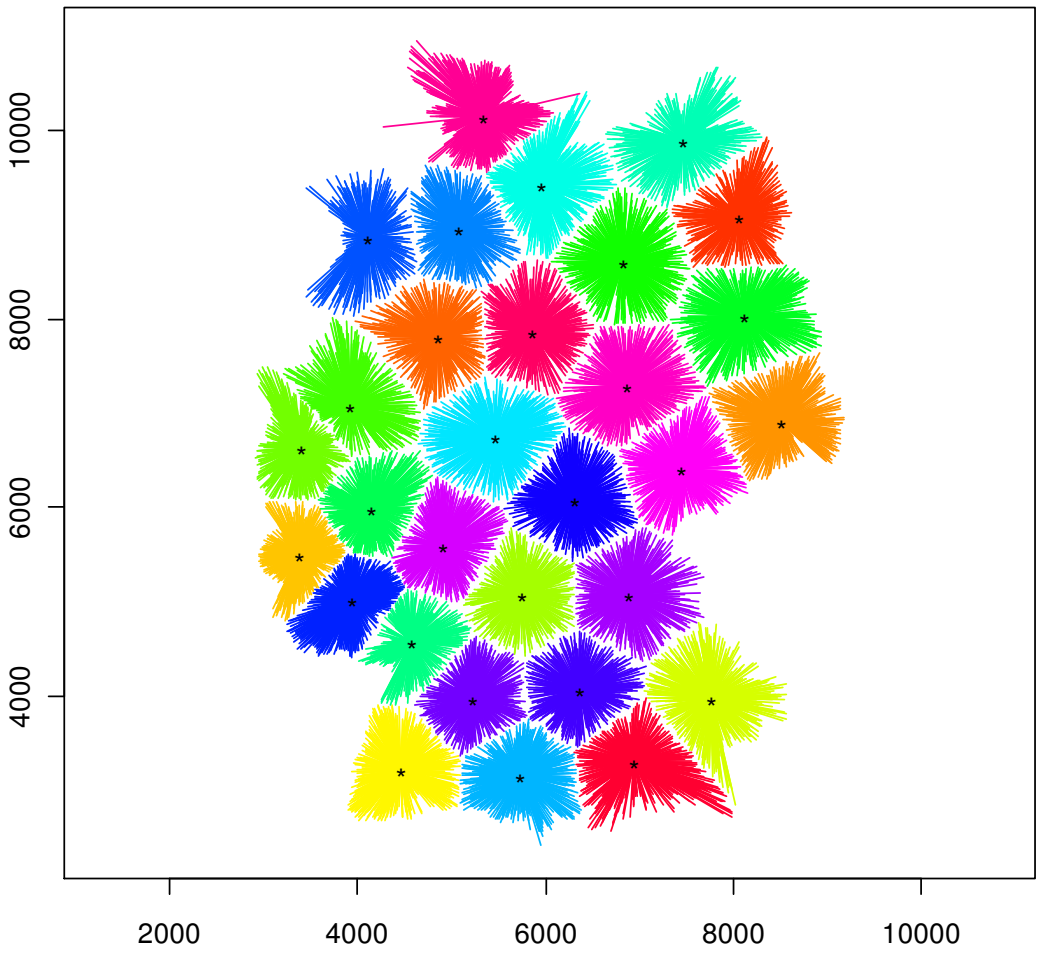


Figura 4.8- Conjunto de dados D18512 agrupados em 30 grupos

A figura 4.8 ilustra uma solução para o caso de 30 grupos, onde cada observação está conectada por um segmento de reta ao seu centróide.

4.2.8 - Problema teste PLA33810

Usando as mesmas convenções anteriores, a Tabela 4.10 mostra os resultados obtidos para esse conjunto de dados, exibindo um comportamento do método análogo àqueles anteriores. A consistência do método é confirmada novamente.

Tabela 4.10 - Resultados para o problema PLA33810

q	F_{min}	Ocorrência	Erro médio	Tempo médio
2	5,13203E+09	10	0,00	10,28
3	4,10346E+09	10	0,00	18,25
4	3,42138E+09	10	0,00	24,95
5	3,09212E+09	1	0,18	42,35
6	2,81971E+09	10	0,00	53,44
7	2,60162E+09	5	0,23	70,66
8	2,42045E+09	9	0,13	88,33
9	2,27991E+09	8	0,28	113,38
10	2,17520E+09	8	0,06	137,16
15	1,79317E+09	6	0,08	358,53
20	1,54650E+09	3	0,20	637,96
25	1,37892E+09	3	0,12	937,35
30	1,25795E+09	1	0,41	1506,22
35	1,16645E+09	1	0,10	3576,52

4.2.9 - Problema teste PLA85900

O problema teste PLA85900 é uma das maiores instâncias da biblioteca TSPLIB. Esse conjunto de dados, mostrado pela Figura 4.9, define o maior problema do caixeiro viajante atualmente resolvido de forma exata.

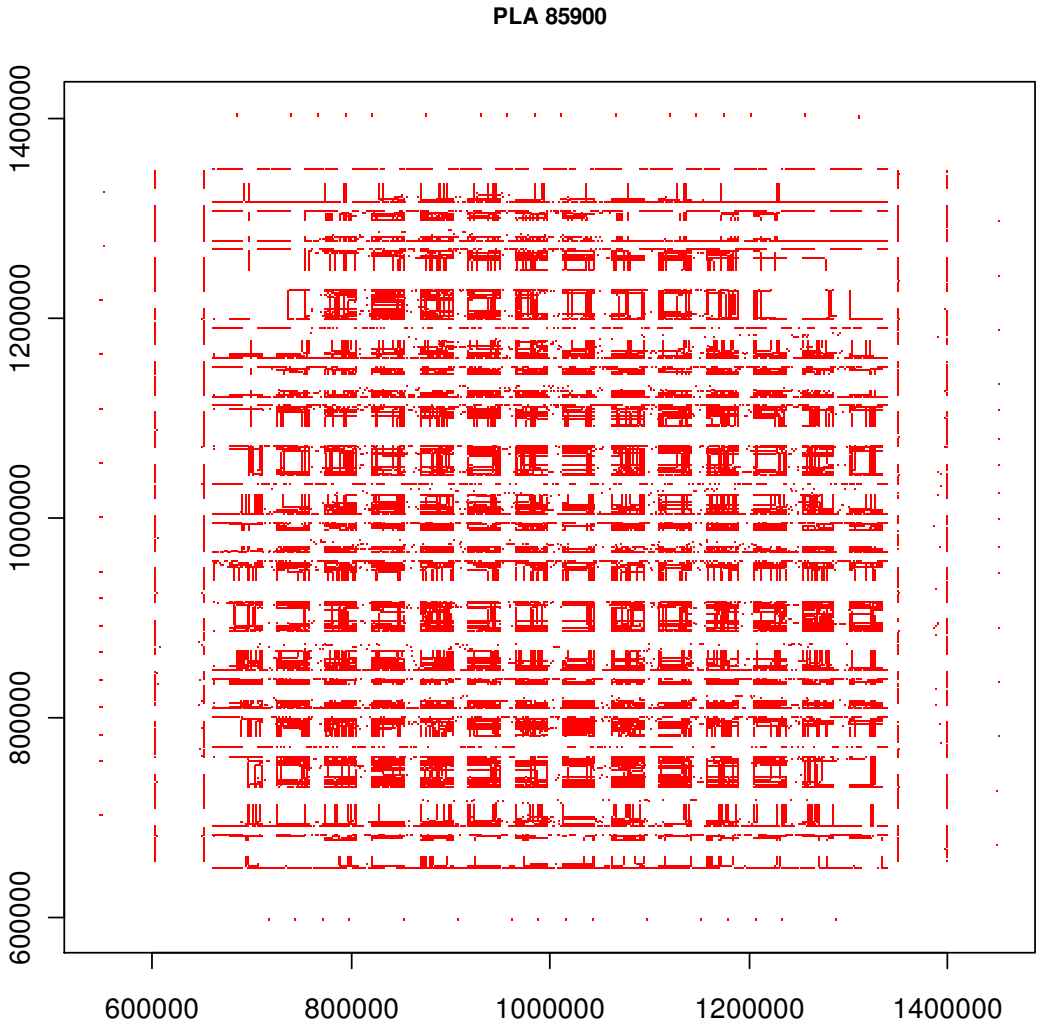


Figura 4.9- Conjunto de dados PLA85900

Usando as mesmas convenções anteriores, a Tabela 4.11 mostra os resultados obtidos para esse conjunto de dados, exibindo um comportamento do método análogo àqueles anteriores. A consistência do método é confirmada novamente. Não foi possível encontrar soluções publicadas para essa instância. Esse mesmo fato igualmente

aconteceu para os problemas teste PCB3038, PLA7397, USA13509, D15112, D18512 e PLA33810..

Tabela 4.11- Resultados para o problema PLA85900

q	$F_{min.}$	Ocorrência	Erro médio	Tempo médio
2	1,63625E+10	6	0,27	25,33
3	1,27835E+10	10	0,00	50,91
4	1,08063E+10	10	0,00	74,62
5	9,84539E+09	7	0,11	121,02
6	9,02515E+09	10	0,00	156,63
7	8,36416E+09	3	0,18	206,71
8	7,78239E+09	10	0,00	260,89
9	7,37264E+09	9	0,09	317,09
10	7,04126E+09	1	0,19	381,33
15	5,76935E+09	10	0,00	937,84
20	5,02191E+09	1	0,13	1690,06
30	4,11982E+09	2	0,08	4062,92
40	3,58238E+09	1	0,11	8169,64

A Figuras 4.10 ilustra as soluções produzidas para os casos de 2 a 5 grupos e a Figura 4.11 ilustra as soluções para os casos de 6 a 10 grupos.

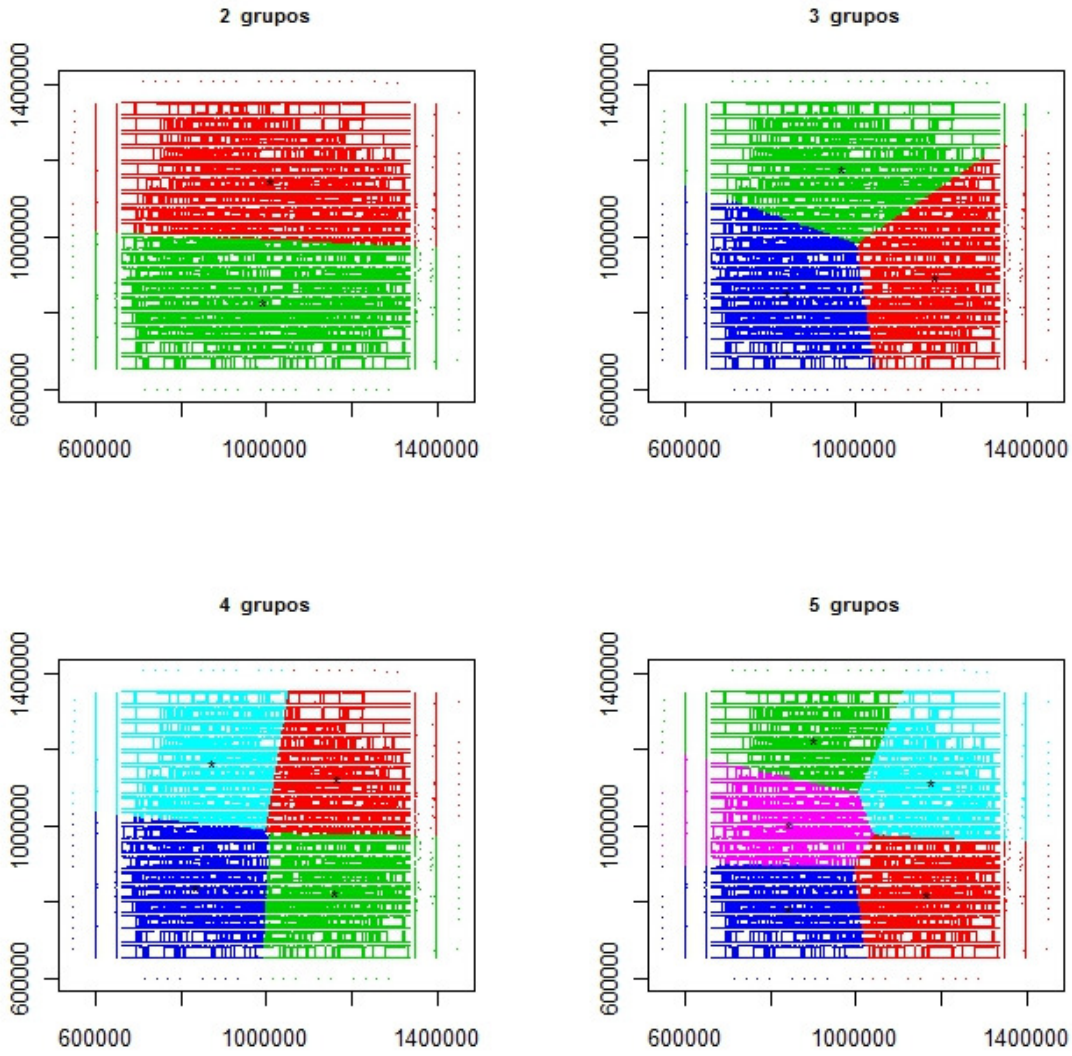


Figura 4.10 – Conjunto de dados PLA85900 agrupado de 2 a 5 grupos.

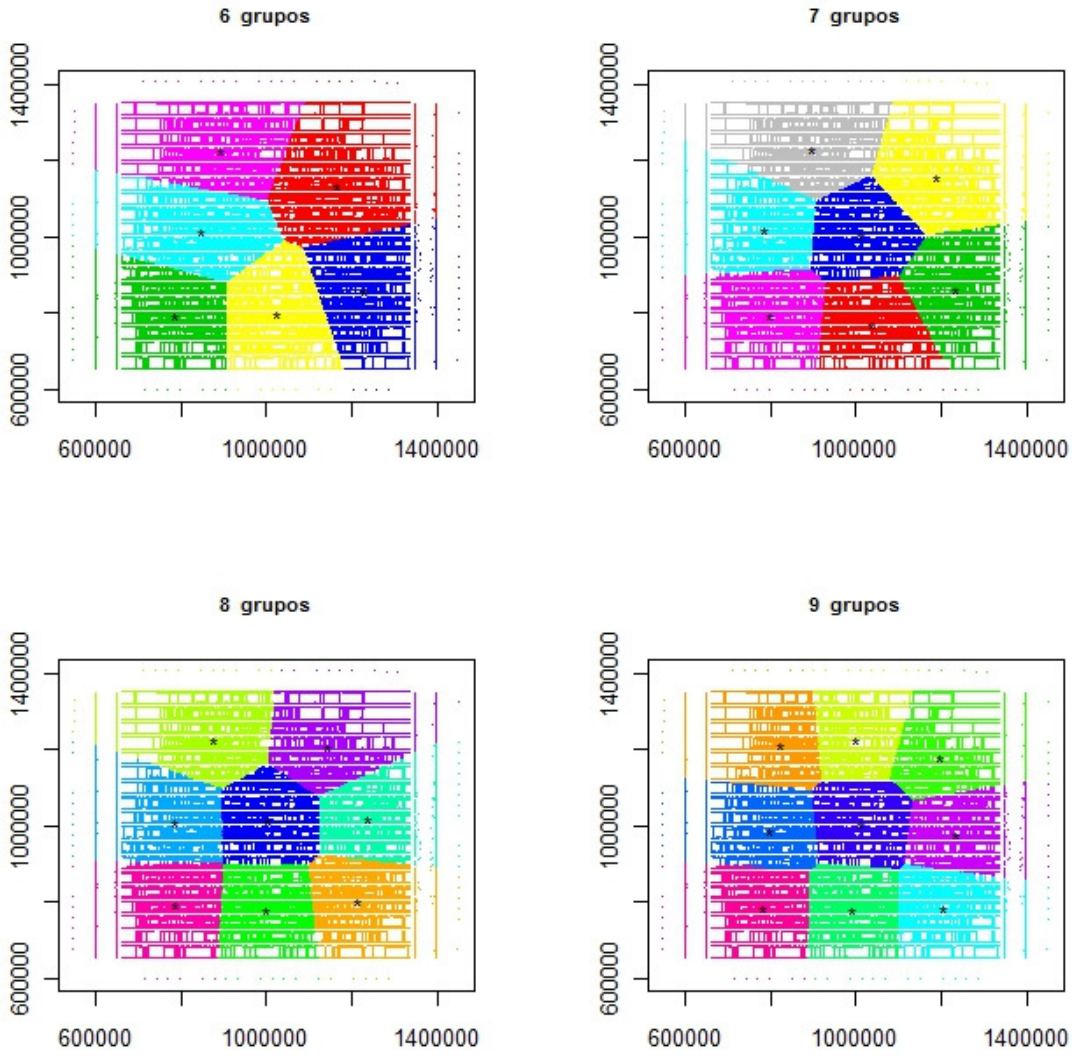


Figura 4.11 – Conjunto de dados PLA85900 agrupado de 6 a 10 grupos.

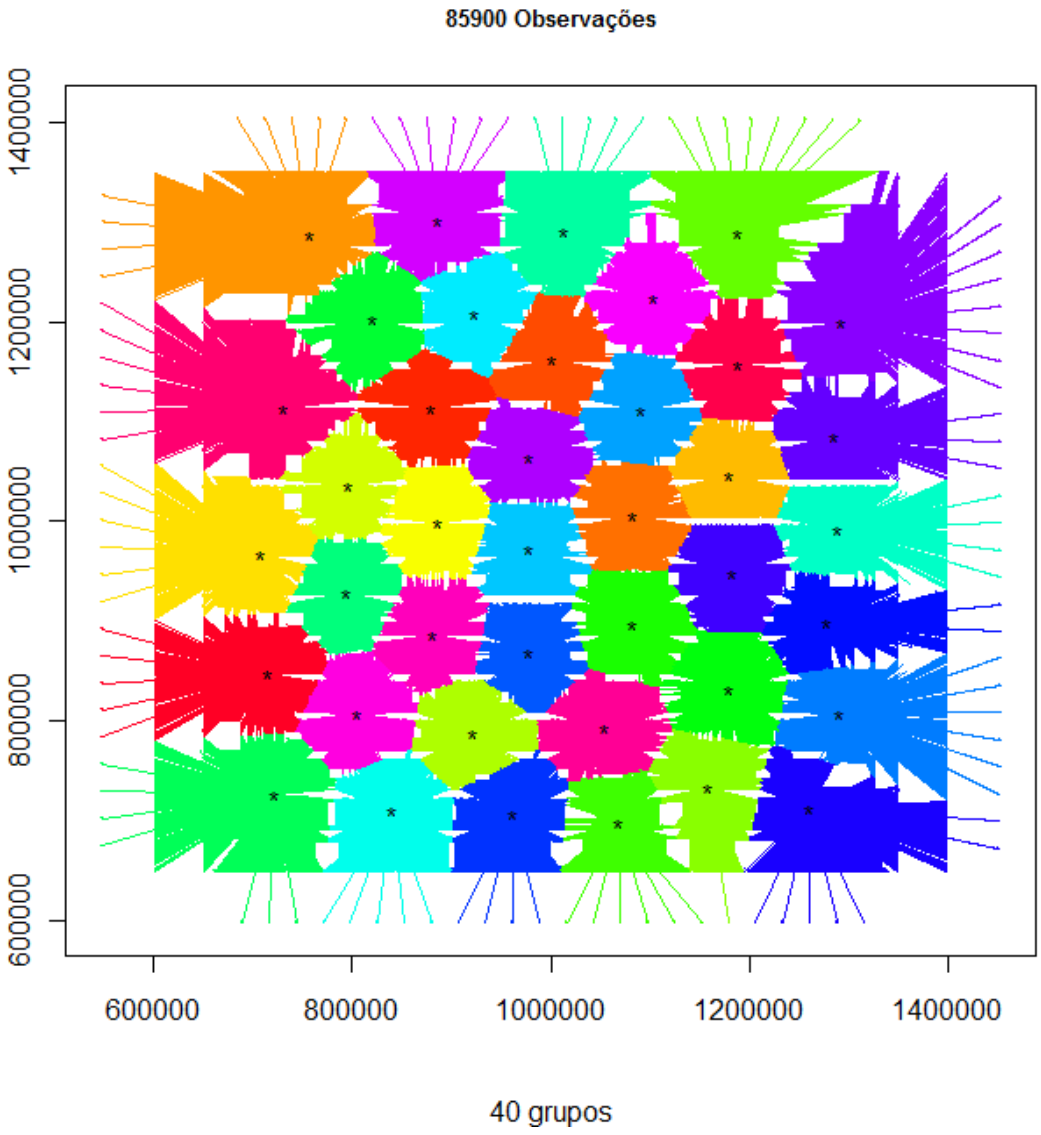


Figura 4.12 – Conjunto de dados PLA85900 agrupado em 40 grupos.

A Figura 4.12 ilustra a solução produzida para o caso com 40 grupos. Novamente as observações estão conectadas por segmentos aos centróides em diferentes cores.

5 - Conclusões

Neste trabalho foi apresentada uma nova proposta metodológica para resolução do problema de agrupamento segundo o critério de minimização da soma de distâncias.

Primeiramente, deve ser registrado que o algoritmo fundamentado nessa proposta teve um desempenho satisfatório. Sendo aplicado a pequenas instâncias, que são amplamente testadas, obteve resultados iguais e até melhores do que os conhecidos na literatura. Além disso, o algoritmo resolveu com consistência problemas maiores, problemas esses, para os quais não se têm resultados conhecidos na literatura.

Em todos os casos, o erro relativo apresentado foi sempre pequeno, mostrando que o algoritmo é consistente, pois para diferentes pontos iniciais, gera soluções muito próximas. Ademais, o algoritmo se mostrou robusto e eficiente, pois as soluções foram de boa qualidade, considerando os valores da solução e os tempos de execução.

Em suma: os resultados computacionais obtidos mostram que a abordagem metodológica foi exitosa. Mesmo assim, algumas melhorias podem ser feitas em trabalhos futuros. Abaixo, relacionamos quatro mais importantes:

- Uso do esquema de partição das observações em dois conjuntos; observações em faixas de fronteira entre dois ou mais centróides e observações situadas em regiões gravitacionais. Esse esquema foi usado com amplo sucesso na resolução do problema de agrupamento, segundo o critério de soma de mínimos quadrados em XAVIER e XAVIER (2011). Pela grande analogia entre os dois problemas, espera-se que a inclusão desse procedimento de partição oferecerá uma melhora substancial nos tempos de execução.

- Controle da ocorrência de agrupamentos vazios, ou seja, centróides aos quais não possuam observações associadas: A ocorrência de grupos vazios viola as hipóteses mais elementares da especificação do problema. Sob o ponto de vista computacional, destrói as propriedades naturais de convergência do algoritmo. Observamos empiricamente que as maiores dificuldades oferecidas ao desempenho do algoritmo

foram derivadas dessas ocorrências. Dessa maneira, a inclusão de salvaguardas internas no algoritmo para evitá-las é essencial.

- Escolha de pontos inicias: Como registrado amplamente na literatura, por exemplo por BRIMBERG *et al.* (2000), a escolha de um bom ponto inicial é de fundamental importância para a obtenção de uma solução final. Na presente implementação do método, esse procedimento foi feito da forma mais simples. Certamente, a definição de um ponto inicial, levando em consideração a distribuição espacial das observações e as particularidades do problema, produzirá melhoras de desempenho do algoritmo. Para esse fim, pode-se, por exemplo, usar o esquema incremental usado com sucesso por Bagirov (2008).

- Especificação dos valores dos parâmetros e manipulação desses valores: Como toda a proposta metodológica é fundamentada nos parâmetros τ , ε , γ e nas taxas de redução desse valores, é natural que as escolha inicial e a atualização do conjunto desses valores definem o desempenho global do algoritmo. No desenvolvimento da implementação do método, foi feito um conjunto de testes visando uma harmonização entre os parâmetros. Há muito a ser pesquisado e melhorado nesse assunto.

- Regra de parada: O critério de parada adotado, número fixo de iterações, é muito singelo. Há inúmeras formas de aprimorá-lo

Referências Bibliográficas

AGRAWAL, R.; GEHRKE, J.; GUNOPULOS, D.; RAGHAVAN, P. Automatic Subspace Clustering on High Dimensional Data for Data Mining Applications In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, p. 94-105, Seattle, Washington, 1998

ARAKAKI, R. G. I.;LORENA, L. A. N. Uma Heurística de Localização-Alocação (HLA) para Problemas de Localização de Facilidades, **Revista Produção**, São Paulo, v. 16, n. 2, p. 319-328, 2006.

ANDERBERG, M. R. **Cluster Analysis for Applications**. New York: Academic Press Inc, 1973.

AVELLA, P.; SASSANO, A.; VASSIL'EV, I. Computational Study of Large Scale p-Median Problems. **Mathematical Programming**, n.109 , p.89–114, 2007.

BAGIROV, A. M. Modified Global k-Means Algorithm for Minimum Sum-of-Squares Clustering Problems, **Pattern Recognition**, v. 41 Issue 10 , p. 3192-3199, 2008

BERKHIN, P. A Survey of Clustering Data Mining Techniques. Technical Report, Accrue Software, 2002.

BRIMBERG, J.; HANSEN, P.; MLADENOVIC, N.; TAILLARD, E. D. Improvements and Comparison of Heuristics for Solving the Multisource Weber Problem, **Operations Research**, v. 48, p. 129-135, 2000

DEMPSTER, A.; LAIRD, N.; RUBIN, D. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**, Series B, v. 39, n.1, p. 1-38, 1977.

DUDA, R.O.; HART, P.E.; STORK, D. G. **Pattern Classification**, 2 ed. New York: John Wiley & Sons Ltd., 2001

EVERITT, B.; LANDAU, S.; LEESE, M. **Cluster Analysis**. 4.ed. Londres: Edward Arnold Ltd.,2001.

FRALEY, F.; RAFTERY A. E. MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering.Technical Report No. 504, Department of Statistics, University of Washington, 2006

FREE UNIVERSITY OF BOZEN, Clique clustering, acessado em 08/08/2011, http://www.inf.unibz.it/dis/teaching/msc/project-clique.html,.

GARCIA, S.; LABBÉ, M.; MARÍN, A. Solving large p-median problems with a radius formulation, INFORMS Journal on Computing, v. 23, No. 4, p. 546-556, 2011.

HAMPEL, F. Some thoughts about classification. Invited keynote lecture, 8th Conference of the International Federation of Classification Societies. In: JAJUGA, K.; SOKOLOWSKI, A.; H. H. BOCK. **Classification, Clustering, and Data Analysis. Recent Advances and Applications**. 2002.

HANSEN, P.; JAUMARD B. Cluster Analysis and Mathematical Programming, **Mathematical Programming**, v. 79, p.191-215, 1997

HANSEN, P.; MLADENOVIC, N. J-Means: A New Heuristic for Minimum Sum-of-Squares Clustering, **Pattern Recognition**, v. 34, p. 405-413, 2001

HANSEN, P.; NGAI, E.; CHEUNG, B.K.; MLADENOVIC, N. Analysis of Global k-Means: an Incremental Heuristic for Minimum Sum-of-Squares Clustering. **Journal of Classification,** v. 22, p. 287-310, 2005

JAIN, A. K.;DUBES, R. C. Algorithms for Clustering Data, Prentice-Hall Inc., Upper Saddle River, 1988

JAIN, A.K.; MURTY, M.N.; FLYNN, P.J. Data Clustering: A Review, ACM Computing Surveys, v.31, set. , p. 264-323, 1999

KAUFMAN, L.; ROUSSEEUW, P. J. **Clustering Large Data Sets.** Pattern Recognition in Practice II, 1986"

KOGAN, J. **Introduction to Clustering Large and High-Dimensional Data**. New York: Cambridge, 2007.

KOTSIANTIS, S.B.; PINTELAS, P.E. Recent Advances in Clustering: A Brief Survey. **WSEAS Transactions on Information Science and Applications**. v. 1, p. 73-81, 2004

LIKAS, A.;VLASSIS, M.; VERBEEK, J. The Global k-means Clustering Algorithm, Pattern Recognition, v. 36, p. 451-461, 2003

MacQueen, J. B. Some Methods for Classification and Analysis of Multivariate Observations, In: Proceedings of the Fifth Berkeley Symposiums on Mathematical Statistics and Probability, 1967, University of California Press, v. 1, p. 281-297.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A; HUBERT, M. Cluster Analysis Basics and Extensions, 2005

MCLACHLAN, G.; BASFORD, K. **Mixture Models: Inference and Applications to Clustering**. New York: Marcel Dekker, 1988.

METZ, J.; MONARD, M. C. Clustering Hierárquico: uma Metodologia para Auxiliar na Interpretação dos Clusters. In: Congresso da SBC XVIII, 2005, São Leopoldo.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Pearson Addison Wesley, 2006.

PAULA Jr., J. R. Resolução do Problema de Empacotamento de Circunferências na Superfície de uma Esfera Utilizando Suavização Hiperbólica, M.Sc. thesis, Federal University of Rio de Janeiro, 2010, Rio de Janeiro.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2012, http://www.R-project.org

REINELT, G. TSPLIB A Traveling Salesman Library, **INFORMS Journal on Computing**, n.4, v.3, 1991, p. 376-384

STRECK S. The Fermat Problem. In: 15th Annual Natural Sciences Academic Festival, 2009

SOUSA, L.C.F. **Desempenho Computacional do Método de Agrupamento Via Suavização Hiperbólica**, M.Sc. thesis, Federal University of Rio de Janeiro, Rio de Janeiro., 2005.

WEBER, A. Ueber den Standort der Industreien. Erster Teil. Reine Theorie der Standorte. Mit einen Mathematischen Anhang von G. Pick, Tübingen, Germany, Verlag J. C. B. Mohr, 1909

WEISZFELD, E. Sur le Point Pour Lequel la Somme des Distances de n Points Donnes Est Minimum, **Tohoku Mathematical Journal**, 43, p. 355-386 , 1937

WESOLOWSKY, G. O. The Weber Problem: History and Perspectives, **Location Sciences**, n.1 ,p. 5-23, 1993

XAVIER, A.E. The Hyperbolic Smoothing Clustering Method. **Pattern Recognition**, v. 43, p. 731-737, 2010

XAVIER, A.E.; XAVIER V. L. Solving the Minimum Sum-of-Squares Clustering Problem by Hyperbolic Smoothing and Partition into Boundary and Gravitational Regions. **Pattern Recognition**, v. 44, p. 70-77, 2011

XU, R.; WUNSCH, D. C. **Clustering.**. New Jersey: John Wiley & Sons, 2009

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, Springer, v.14, p. 1-37, 2007

Anexo 1- Códigos R

1.1 Código do método Hierárquico

```
hc <- hclust(dist(USArrests), "ave")
(dend1 <- as.dendrogram(hc))
plot(dend1)
```

1.2 Código do algoritmo CLARA

```
library(cluster)
x <- rbind(cbind(rnorm(200,0,8), rnorm(200,0,8)),
           cbind(rnorm(300,50,8), rnorm(300,50,8)))
clarax <- clara(x, 2, samples=50)
clarax
clarax$clusinfo
plot(clarax)

data(xclara)
(clx3 <- clara(xclara, 3))
plot(clx3)
```

1.3 Código do algoritmo EM

```
library(mclust)
x1 = rnorm(n=20, mean=1, sd=1)
y1 = rnorm(n=20, mean=1, sd=1)
 x2 = rnorm(n=20, mean=5, sd=1)
y2 = rnorm(n=20, mean=5, sd=1)
 rx = range(x1,x2)
```

```
 ry = range(y1,y2)
 plot(x1, y1, xlim=rx, ylim=ry)
 points(x2, y2)
mix = matrix(nrow=40, ncol=2)
mix[,1] = c(x1, x2)
mix[,2] = c(y1, y2)
mixclust = Mclust(mix)
plot(mixclust, data = mix)
```

1.4 Código do algoritmo DBSCAN

```
n <- 600
library(fpc)
x <- cbind(runif(10, 0, 10)+rnorm(n, sd=0.2), runif(10, 0, 10)+rnorm(n, sd=0.2))
ds <- dbscan(x, 0.2)
plot(ds, x,xlab="",ylab="",main="")
```

Publicações

Durante o período do mestrado, foram publicados os seguintes trabalhos:

XAVIER, A. E.; XAVIER V. L. Solving the Minimum Sum-of-Squares Clustering Problem by Hyperbolic Smoothing and Partition into Boundary and Gravitational Regions. **Pattern Recognition**, v. 44, p. 70-77, 2011

XAVIER V. L. A system developed for solving the matching problem in the Brazilian Census Post Enumeration Survey. In: **58th World Statistics Congress ISI2011**, Contributed Papers, Dublin, Irlanda 2011

XAVIER V. L. Solving the Sum of Euclidean Distances Clustering Problem by the Hyperbolic Smoothing Method. In: **Young Statisticians Meeting (YSI 2011) - Satellite meeting to the 2011 ISI World Statistics Congress**, Dublin, Irlanda 2011

SILVA, A. D.; ROMEO, O.S M; SOARES, T. S.; XAVIER, V. L. Study of Record Linkage Software for the 2010 BrazilianCensus Post Enumeration Survey In: **58th World Statistics Congress ISI2011**, Invited Papers, Dublin, Irlanda 2011