Universidade do Minho
Escola de Ciências

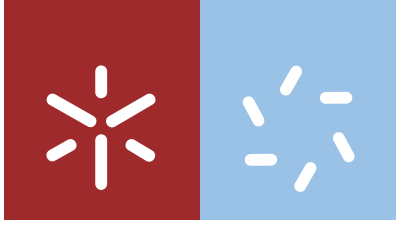
Maria Manuel Alves Soares

Clustering hierárquico numa plataforma de *smart cities*

outubro de 2017

Maria Manuel Alves Soares *Clustering* hierárquico numa plataforma de *smart cities*

UMinho | 2017



Universidade do Minho
Escola de Ciências

Maria Manuel Alves Soares

Clustering hierárquico numa plataforma de *smart cities*

Relatório de Estágio
Mestrado em Matemática e Computação

Trabalho realizado sob orientação do
Doutor Stéphane Clain
e do
Doutor Luís Pinto

outubro de 2017

Nome: Maria Manuel Alves Soares

Endereço eletrónico: soares.mariamanuel@gmail.com

Número do bilhete de identidade: 12424787

Título de Relatório de Estágio:
"*Clustering* hierárquico numa plataforma de *smart cities*"

Orientadores:
Doutor Stéphane Clain
Doutor Luís Pinto

Ano de conclusão: 2017

Designação do Mestrado:
Matemática e Computação

Universidade do Minho, 31 de outubro de 2017

Maria Manuel Alves Soares

Clustering hierárquico numa plataforma de *smart cities*

Resumo

Através do conceito de *smart cities* os cidadãos têm, nos dias de hoje, a oportunidade de serem cidadãos mais ativos e tirarem partido das tecnologias de informação e comunicação para melhorar a qualidade de vida das suas comunidades. Neste sentido, a plataforma móvel *JuntarAJunta* tem como objetivo aproximar os cidadãos das suas Juntas de Freguesia, permitindo-os reportar problemas de uma localidade ou fazer uma sugestão.

A utilização deste tipo de aplicações informáticas remete-nos para a necessidade da gestão dos seus dados de forma eficiente. Ao longo dos anos, cada vez mais se percebe que a análise adequada dos dados permite identificar padrões e encontrar novas correlações entre os dados e estas podem ser uma mais valia ao nível dos processos de tomadas de decisão e de desenvolvimento de estratégias.

Este relatório resulta de um estágio desenvolvido numa empresa e tem como tema principal a aplicação e o desenvolvimento de técnicas de *machine learning* na plataforma *JuntarAJunta*. O objetivo é identificar participações semelhantes, reportadas pelos cidadãos, com o intuito de reduzir informação redundante. Numa primeira abordagem recorreu-se ao método de *clustering* hierárquico, fazendo uso dos dados geográficos fornecidos pela aplicação. Através da análise dos resultados dos índices de validação dos *clusters* foi possível perceber até que ponto a localização geográfica é suficiente para agrupar participações semelhantes de forma correta. Os resultados obtidos remetem para uma segunda abordagem na qual são utilizados novos atributos da participação.

Palavras-chave: *Machine Learning*, *Clustering* Hierárquico, Validação de *Clustering*, Plataforma Móvel

Hierarchical clustering on a smart cities platform

Abstract

Through the smart cities concept, citizens have today the opportunity to be more active and take advantage of information and communication technologies to improve the quality of life of their communities. In this sense the mobile application JuntarAJunta aims to bring citizens closer to their Town Councils, allowing them to report problems or make suggestions.

The increase in the use of this type of applications reminds us of the need to manage the data efficiently. Over the years it has become increasingly clear that an adequate analysis of the data allows the identification of patterns and finding new correlations between data, which can be an added value in the decision-making process and strategies development.

This report is based on a curricular internship and its main theme is the application and development of machine learning techniques in the JuntarAJunta application. The goal is to identify similar participations, reported by citizens, in order to reduce redundant information. In the first approach was used the Hierarchical clustering method, making use of the geographic data provided by the application. By analyzing the results of the clusters validation indexes, it was possible to see at what extent the geographic location is sufficient to group similar participations correctly. The results obtained refer to a second approach in which new information are attributed to participations.

Keywords: Machine Learning, Hierarchical Clustering, Cluster Validation, Mobile Application

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este relatório é relativo ao desenvolvimento de um estágio curricular na empresa de informática Codeangel, decorrido entre 5 de Dezembro de 2016 e 31 de Outubro de 2017.

Este estágio enquadra-se na unidade curricular "Estágio/Dissertação" do Curso de Mestrado em Matemática e Computação, área de especialização em Matemática e Ciências da Computação.

O tema inicial de estágio passava por usar técnicas de *machine learning* no contexto de uma plataforma de comércio eletrónico. Mais concretamente, pretendia-se identificar entidades de maior impacto nas vendas de produtos já existentes e, com base nesta informação, predizer o conjunto de entidades de maior impacto para novos produtos.

Em Março de 2017, face a um menor interesse da empresa em usar o comércio eletrónico como área de aplicação, os objetivos do estágio sofreram uma alteração relativamente ao que estava inicialmente definido. Estes passaram a estar relacionados com o desenvolvimento de técnicas de *machine learning* numa plataforma móvel designada por *JuntarAJunta* [1].

A plataforma móvel *JuntarAJunta* tem como finalidade dar a oportunidade à população de ter uma posição ativa na comunidade de uma Junta de Freguesia e assim poder reportar problemas, propondo intervenções e sugestões.

Com a redefinição dos objetivos, o objetivo principal do estágio passou a ser o de identificar participações semelhantes na aplicação *JuntarAJunta* através do uso de técnicas de *machine learning*. Esta tarefa foi estudada através de duas abordagens analisadas por mim e pela aluna Carolina Sousa [13]. Numa primeira fase fez-se uso dos dados fornecidos pela aplicação, designadamente a informação geográfica que se refere ao local de onde o evento a reportar foi fotografado. Através dos resultados obtidos, achou-se oportuno reformular o

[1] http://juntarajunta.pt/

problema e introduzir novos dados que não eram fornecidos inicialmente pela aplicação baseados noutros atributos relativos ao evento a reportar. Assim, este relatório apresenta de forma mais detalhada a primeira fase desta tarefa. A segunda fase é apresentada em detalhe no relatório [13].

Este relatório está organizado em 6 capítulos como se descreve a seguir. Neste primeiro capítulo é apresentada a empresa, é definido o tema de estágio e o seu enquadramento tendo em conta o objetivo final.

No segundo capítulo estão descritas as tarefas iniciais que foram realizadas durante os primeiros meses do estágio. Estas tarefas relacionavam-se com a aprendizagem de ferramentas de desenvolvimento *web* e tinham como principal objetivo a apresentação dos resultados numa *dashboard*.

No terceiro capítulo é apresentada a aplicação *JuntarAJunta* e os dados obtidos através desta. Também é feita uma breve análise à base de dados e são destacados alguns dos problemas encontrados nesta.

No quarto capítulo são apresentados os conceitos e fundamentos teóricos relativos à metodologia de *clustering* hierárquico adotada para a elaboração da tarefa proposta.

No quinto capítulo são inicialmente identificadas as variáveis a serem utilizadas no método de *clustering* hierárquico e é explicada a forma de criação dos cenários sintéticos que serão utilizados. Além disso, são também apresentados os resultados obtidos da aplicação do método para a identificação de participações semelhantes, baseados no estudo dos indicadores de validação de *clusters*.

No sexto capítulo, com base nos resultados obtidos, são apresentadas conclusões que justificam a necessidade de serem introduzidas novas variáveis e que nos remetem para a abordagem apresentada no relatório [13]. Neste capítulo são também sugeridas algumas melhorias na aplicação *JuntarAJunta*.

Capítulo 2

Tarefas iniciais

As tarefas iniciais realizadas neste estágio prenderam-se com os objetivos inicialmente traçados na área do comércio eletrónico (*eCommerce*): a identificação das entidades de maior impacto nas vendas de produtos já existentes e, com base nesta informação, a predição do conjunto de entidades de maior impacto para a venda de novos produtos.

Os objetivos iniciais relacionam-se com o facto de que, hoje em dia, o volume de dados armazenados aumenta exponencialmente e num curto espaço de tempo. Um dos grandes desafios é como gerir toda a informação de forma eficiente e de forma a tirar maior partido dela.

Ao longo dos anos, cada vez mais empresas percebem que a análise desses dados pode ser uma mais valia ao nível dos processos de tomadas decisão. Assim, as empresas investem neste tipo de análise, que permite a elaboração de novas estratégias de negócio. Uma análise adequada dos dados permite identificar padrões e encontrar novas correlações entre os dados.

Num contexto de *eCommerce*, o objetivo da análise pode passar por melhorar os processos de trabalho e inferir acerca das tendências de mercado e do comportamento e expectativas dos consumidores. É assim possível tomar decisões mais precisas e, sobretudo, antecipadas em relação à concorrência. Decisões que, se podem revelar muito importantes num cenário de competitividade. [1] [2] [3] [4] [6] [5]

De forma a gerir os dados de negócio e acompanhar o desempenho do *eCommerce*, podem ser utilizadas técnicas estatísticas e de *machine learning*. Os resultados podem ser apresentados através de *dashboards*, que permitem uma leitura mais fácil e uma gestão em tempo real.

Como tarefa inicial no âmbito do estágio foi feita uma aprendizagem ao nível das linguagens **HTML**, **CSS** e **JavaScript** [1]:

[1] https://www.w3schools.com/css/

- **HTML** *HyperText Markup Language*, linguagem utilizada na construção de páginas na *web*;
- **CSS** *Cascading Style Sheets* que permite melhorar a apresentação visual de informação;
- **JavaScript** que, em particular, permite dotar as páginas *web* de comportamentos interativos.

Ao longo desta aprendizagem foi utilizada a biblioteca **jQuery** [2] de funções **JavaScript** que interage com o **HTML** e a **framework Bootstrap** [3], estrutura de *web front-end* livre e de código aberto, para projetar sites e aplicações da *web*. De forma a aplicar e praticar a aprendizagem destes conceitos, foram construídos alguns sites e foram apresentados dados numa *dashboard*. Foi também feita a apresentação de informação contida numa base de dados **MySQL** [4] através de uma *dashboard*, em que para aceder e utilizar os dados foi utilizada a linguagem **PHP** *Hypertext Preprocessor* [5]. O objetivo era criar uma *dashboard* com informações importantes para funcionários de uma empresa fictícia. O resultado final desta tarefa está representado nas Figuras 2.1, 2.2 e 2.3.

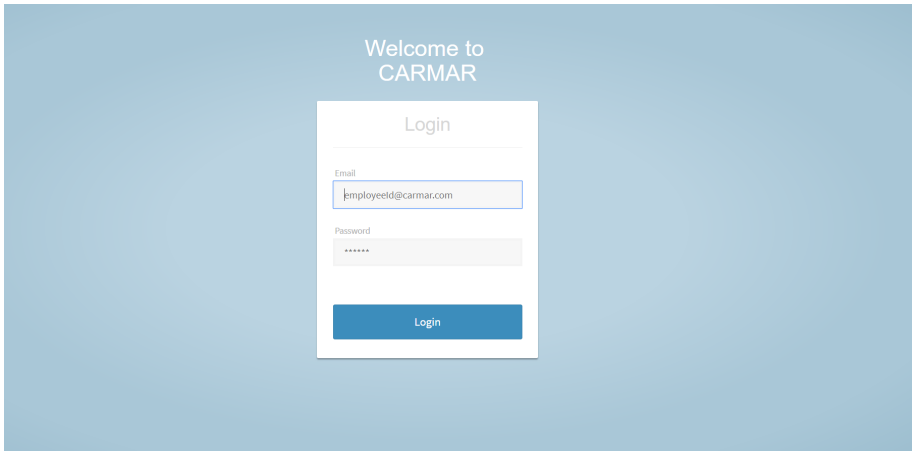


Figura 2.1: Tarefa de construção de uma página de *login* para aceder à *dashboard*

[2] https://jquery.com/
[3] http://getbootstrap.com/
[4] https://www.mysql.com/
[5] https://www.w3schools.com/php/

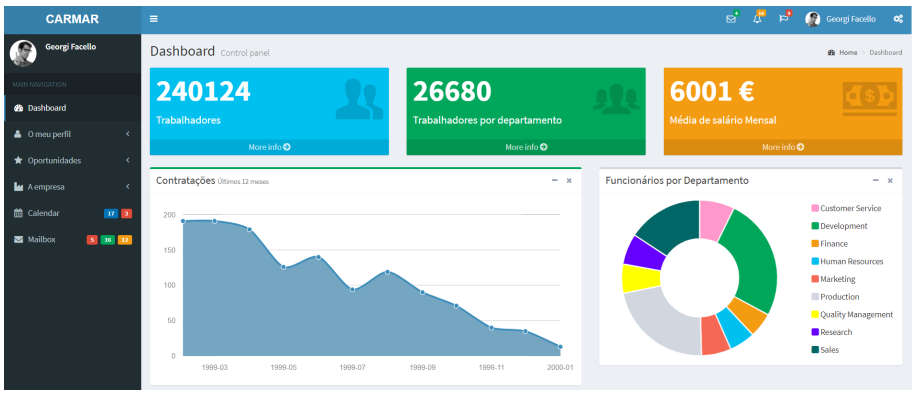


Figura 2.2: Tarefa de desenvolvimento de uma *Dashboard*

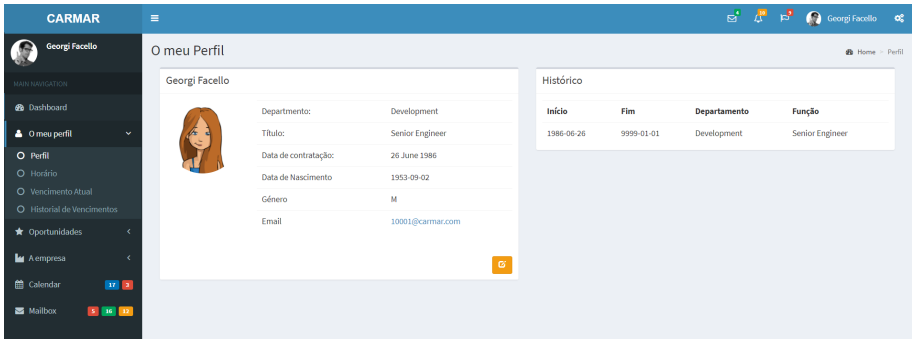


Figura 2.3: Tarefa de construção de um perfil de utilizador

A tarefa que se seguiu tratava da criação de formulários utilizando a *framework* **Django** [6]. Esta *framework* é escrita na linguagem **Python** [7] e permite criar formulários e manipular bases de dados, isto é, ajuda a desenvolver sites de forma mais rápida e mais fácil.

Concluídas estas tarefas iniciais de aprendizagem de diversas tecnologias relacionadas com a construção de páginas *web*, *dashboards* e formulários, que duraram cerca de 3 meses, o estágio prosseguiu com um estudo introdutório sobre *machine learning*. Para tal foi seguido o curso *Intro to machine learning* na plataforma *Udacity* [8]. Este curso permitiu adquirir algumas bases sobre técnicas de *machine learning* e, em particular, permitiu um primeiro contacto com técnicas de *clustering*, que posteriormente foram estudadas mais aprofundadamente e aplicadas no contexto da plataforma *JuntarAJunta*, conforme detalhado nos capítulos 4 e 5. Na realização deste curso foi utilizada a biblioteca **Scikit Learn** [9].

[6] https://www.djangoproject.com/
[7] https://www.python.org/
[8] https://www.udacity.com/
[9] http://scikit-learn.org/stable/

Capítulo 3

Plataforma *JuntarAJunta*

A plataforma móvel *JuntarAJunta* tem por finalidade dar a oportunidade à população de ter uma posição ativa na comunidade de uma Junta de Freguesia e assim poder reportar problemas propondo intervenções e apresentar sugestões.

Esta aplicação informática apresenta duas funcionalidades principais: intervenções e sugestões. A funcionalidade das intervenções permite reportar algum problema numa determinada zona, sendo necessário o utilizador fotografar o problema, descrevê-lo e escolher uma categoria adequada. A funcionalidade das sugestões possibilita ao utilizador sugerir alterações e apresentar sugestões para melhorar a sua freguesia, sendo a localização de cada participação obtida a partir da localização do utilizador no momento da submissão.

Dado que um grande número de participações poderia trazer problemas ao nível da sua gestão pelas Juntas de Freguesia, pensou-se que seria interessante identificar participações que exibissem algumas características semelhantes e pudessem estar a reportar o mesmo problema.

A identificação de participações semelhantes poderá ser uma capacidade útil à plataforma *JuntarAJunta* em duas vertentes. Por um lado, o agrupamento de participações permite que as decisões nas Juntas de Freguesia tenham acesso a uma lista de problemas reportados que evita repetições e, ao mesmo tempo, à frequência com que um problema for reportado. Por outro lado, a capacidade de agrupar participações poderá permitir que, ao submeter uma participação, a aplicação sugira ao utilizador a ou as participações existentes semelhantes à que está a ser submetida. Assim, o utilizador apoiaria a participação já existente dando-lhe mais visibilidade e credibilidade, em vez de submeter um novo pedido e, evitando assim participações repetidas. Assim, o objetivo principal do estágio passou por implementar técnicas e mecanismos de *machine learning*, de forma a identificar participações no contexto da plataforma *JuntarAJunta*.

A utilização da aplicação *JuntarAJunta* passa pelo preenchimento de um formulário, de forma a submeter uma intervenção ou uma sugestão. A aplicação

faz uso das coordenadas GPS e da freguesia relativa ao local onde é realizada a submissão da participação.

Para cada participação submetida na aplicação, (intervenção ou sugestão) existe um conjunto de atributos associados. Estes atributos e a sua descrição apresentam-se de seguida:

- **ID**. Identificação do autor da participação;
- **Latitude**. Obtida ao submeter a participação;
- **Longitude**. Obtida ao submeter a participação;
- **Freguesia**. Obtida pelas coordenadas geográficas;
- **Foto**. Tirada através da aplicação;
- **Descrição**. Descrição do problema;
- **Categoria**. Categoria na qual se insere a intervenção (*Apenas para as intervenções*);
- **Estado**. Resolvido, não resolvido ou reportado;
- **Data**. Data de submissão da participação;
- **Data de alteração**. Data em que o estado da participação sofreu alteração;

As categorias sugeridas ao utilizador quando é reportada uma intervenção são as seguintes:

1. Acessibilidade;
2. Árvores (Perigo);
3. Estacionamento;
4. Estrutura Perigosa;
5. Grafitti;
6. Iluminação Pública;
7. Jardins Públicos;
8. Limpeza de ruas;

9. Mobiliário urbano;
10. Obstrução da estrada ou passeio;
11. Plantação de árvores;
12. Recolha Carros abandonados;
13. Recolha de lixo;
14. Recolha Reciclagem;
15. Reparação estrada ou pavimento;
16. Saneamento;
17. Sinalética;
18. W.C. público;
19. Outros;
20. Transportes Públicos;
21. Animais abandonados;
22. Maus tratos de animais.

De seguida será feita uma breve descrição da base de dados associada à plataforma *JuntarAJunta*. A informação obtida a partir da plataforma encontra-se numa base de dados da empresa Codeangel que, a 17 de março de 2017 continha 142 entradas e a 3 de agosto de 2017 continha 421 entradas. Depois de uma análise da base de dados foram detetadas algumas inconsistências nos dados. Detetaram-se como inconsistências as seguintes situações:

- Entradas repetidas, isto é, entradas pertencentes ao mesmo utilizador e que continham a mesma informação;
- Entradas associadas a freguesias erradas, isto é, entradas em que as coordenadas da freguesia não correspondiam à freguesia associada à participação;
- Entradas de teste e que por isso não têm informação relevante;
- Entradas em que a latitude e a longitude são zero.

Capítulo 4

Metodologia de análise de *clustering*

O *clustering* de dados é um conjunto de técnicas para agrupamento de dados em *clusters*, baseado no grau de semelhança dos dados. Por isso, a escolha deste tipo de técnicas para abordar o problema da identificação de participações semelhantes, no contexto da plataforma *JuntarAJunta*, tornou-se uma escolha natural. Tendo como base os passos para análise de *clusters* sugeridos em [9] e [12], o processo seguido neste trabalho segue os seguintes passos:

1. *Escolha das variáveis a serem utilizadas*;
2. *Escolha do método de clustering*;
3. *Representação gráfica do processo de clustering*;
4. *Comparação de dendrogramas*;
5. *Escolha da medida de dissemelhança*;
6. *Escolha da partição*;
7. *Teste e interpretação dos resultados*;

Pelas inconsistências referidas anteriormente e também pelo facto do volume de dados na plataforma *JuntarAJunta* ser muito reduzido, quando se iniciou esta fase dos trabalhos, os dados utilizados neste estudo foram criados de forma sintética.

Ao longo da realização deste trabalho, todas as implementações foram executadas na linguagem **Python** fazendo uso da biblioteca **Scikit-Learn** que é específica de *machine learning*, da biblioteca **Scipy** que é uma biblioteca científica de onde se utilizaram os pacotes **Numpy** para a gestão aleatória de

dados e **MatplotLib** para a construção de gráficos. Foram também utilizadas funções para o cálculo do índice *Dunn* obtidas em [11].

Nas secções seguintes deste capítulo será definida a notação utilizada ao longo deste relatório de acordo com [14] [8].

4.1 Conceitos principais

Nesta secção são apresentados alguns conceitos importantes para a análise de *clusters* e parte da notação utilizada ao longo da elaboração deste trabalho.

Seja X um conjunto de N objetos, aos quais se quer aplicar um método de *clustering*, e K o número de *clusters*. Define-se um *clustering* de X em K *clusters* como sendo uma partição de X em K conjuntos, $C_1, \ldots, C_K$, tal que as seguintes propriedades são satisfeitas:

1. $C_i \neq \emptyset, \quad i = 1, \ldots, K;$
2. $X = \bigcup_{i=1}^{K} C_i;$
3. $C_i \cap C_j = \emptyset, \quad i \neq j, \quad i, j = 1, \ldots, K.$

Pretende-se que os objetos de um *cluster* sejam mais "similares"entre si do que com objetos de qualquer outro *cluster*.

Quando X está contido num espaço vetorial, como por exemplo $\mathbb{R}^n$, o *centróide* $\overline{C}$ é o ponto médio do conjunto C e é definido por

$$\overline{C} = \frac{1}{|C|} \sum_{x \in C} x$$

onde $|C|$ representa o número de objetos do conjunto C.

4.2 Medidas de dissimilaridade

Para identificar a noção de proximidade entre dois elementos, neste trabalho foi utilizada a distância euclidiana. Sejam x e y dois pontos definidos por coordenadas geográficas em $\mathbb{R}$, tal que $x = (latitude_x, longitude_x)$ e $y = (latitude_y, longitude_y)$.

A distância euclidiana é calculada de acordo com:

$$d(x, y) = \sqrt{(latitude_x - latitude_y)^2 + (longitude_x - longitude_y)^2}$$

Seja U o conjunto constituído por K subconjuntos de X, $U = \{C_1, \ldots, C_K\}$. A *medida de dissimilaridade*, d, em U é uma função

$$d : U \times U \to \mathbb{R}$$

tal que:

1. $\exists d_0 \in \mathbb{R} : \quad -\infty < d_0 \leq d(C_i, C_j) < +\infty, \quad \forall C_i, C_j \in U$

   sendo d_0 o grau mínimo de dissimilaridade entre elementos.

2. $\forall C_i \in U, \quad d(C_i, C_i) = d_0$

3. $\forall C_i, C_j \in U, \quad d(C_i, C_j) = d(C_j, C_i)$

d é uma métrica se também verificar as seguintes propriedades:

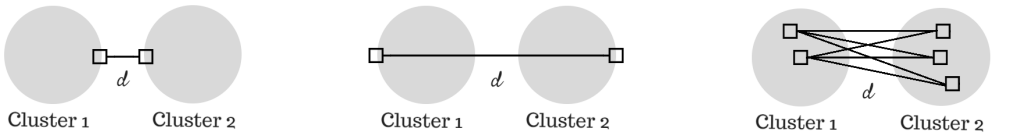
4. $\forall C_i, C_j \in U, \quad d(C_i, C_j) = d_0 \quad$ se e só se $\quad C_i = C_j$

5. $\forall C_i, C_z \in U, \quad d(C_i, C_z) \leq d(C_i, C_j) + d(C_j, C_z)$

Seja $X = \{x_1, \ldots, x_N\}$ e $1 \leq i, j \leq N$.

A matriz de dissimilaridade de X, $D(X)$, é uma matriz $N \times N$, simétrica, tal que as linhas e colunas representam os objetos de X e D_{ij} é o valor da dissemelhança entre os objetos x_i e x_j. A dissemelhança é o fator de decisão para a união ou não de *clusters* e é calculada a partir da distância entre pares de observações (métrica) e no caso da distância entre conjuntos de observações, isto é, distância entre *clusters*, é aplicado um critério de ligação.

Na Figura 4.1 apresentam-se as imagens que representam os critérios de ligação abordados neste estudo:



(a) *Single linkage* (b) *Complete linkage* (c) *Group average linkage*

Figura 4.1: Critérios de ligação

Seja $d : X \times X \to \mathbb{R}$ uma métrica. Os critérios de ligação referidos anteriormente estão definidos de seguida:

(a) **Single linkage (nearest neighbor)**: a distância entre dois *clusters* é determinada pela distância dos dois objetos mais próximos nos dois clusters;

$$d_{min}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

(b) **Complete linkage (furthest neighbor)**: a distância entre dois *clusters* é determinada pela distância dos dois objetos mais distantes nos dois clusters;

$$d_{max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

(c) **Group average linkage**: a distância entre dois *clusters* é determinada pela média das distâncias entre todos os pares dos dois clusters;

$$d_{avg}(C_i, C_j) = \sum_{x \in C_i, y \in C_j} \frac{d(x, y)}{|C_i||C_j|}$$

onde $|C_i|$ representa o número de objetos de C_i e $|C_j|$ representa o número de objetos de C_j.

Uma outra possibilidade seria a utilização das distâncias entre os centróides é dada por,

$$d_c(C_i, C_j) = d(\overline{C_i}, \overline{C_j})$$

4.3 *Clustering* hierárquico

O *clustering* hierárquico é uma técnica de *clustering* em que não se fixa um número prévio de *clusters* e é baseada na construção de uma hierarquia de fusão entre *clusters*. A estratégia pode ser aglomerativa ou divisiva.

Numa abordagem de aglomeração, designada por *bottom up*, inicialmente cada observação representa um *cluster* e vai-se fazendo a fusão de pares de *clusters*, onde em cada passo de fusão são agrupados num mesmo *cluster* os dois *clusters* mais próximos entre si. O processo termina quando é obtido um único *cluster*, do qual fazem parte todas as observações.

Numa abordagem de divisão, designada por *top down*, inicialmente todas as observações representa um único *cluster* e, à medida que se desce na hierarquia, há uma separação de *clusters* terminando cada *cluster* como uma única observação.

O método requer uma análise posterior para encontrar o número "ideal"de *clusters* conforme a secção 4.5.

Assim, este trabalho trata da identificação de participações semelhantes através do método de *clustering* hierárquico, dado que este método não necessita do número de *clusters*, *a priori*, para ser utilizado.

4.4 Representação gráfica

Supondo que se pretende aplicar o método de *clustering* hierárquico a um conjunto de N participações, utilizando uma estratégia aglomerativa, o algoritmo começa por identificar cada participação como sendo um *cluster*. De seguida, são calculadas todas as distâncias entre participações e são unidas num único *cluster* as duas participações que se encontram a uma menor distância. O processo continua verificando quais as participações ou *clusters* mais próximos, conforme a métrica e o critério de ligação escolhido, unindo-os num novo *cluster*. No final, existirá apenas um *cluster* que contém todas as participações.

Este processo pode ser apresentado na forma de um dendrograma que utiliza as medidas de dissimilaridade. A dissimilaridade entre duas observações é a altura do dendrograma onde dois *clusters* se fundem num único.

Através da aplicação deste processo é possível construir um dendrograma que permite reconstruir o histórico de fusões. No dendrograma 4.2 estão apresentadas linhas ligadas segundo os níveis de dissemelhança que agruparam pares de variáveis.

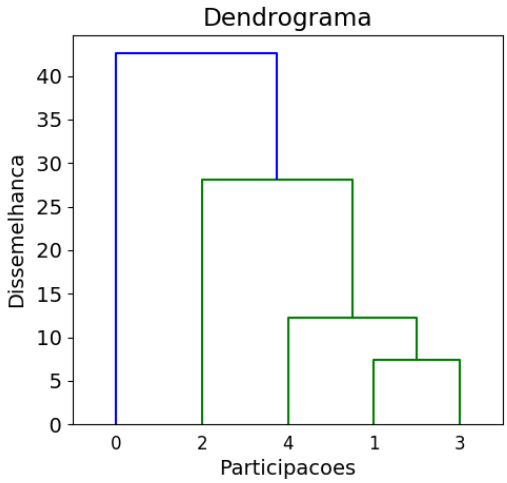


Figura 4.2: Representação de um dendrograma

As visualizações gráficas de dados multivariados são muito relevantes no que se refere à sua análise. Estas podem fornecer informações sobre a estrutura dos dados e podem ser úteis para sugerir que os dados podem conter *clusters*. A utilidade da visualização gráfica neste contexto decorre do poder do sistema visual humano na deteção de padrões, tal como referido em [9].

4.5 Índice de qualidade de um *cluster*

As técnicas de *clustering* hierárquico impõem uma estrutura hierárquica dos dados, sendo necessário verificar se o tipo de estrutura é aceitável, ou se introduz uma distorção inaceitável das relações originais entre os dados.

Dado que o dendrograma é uma simplificação em duas dimensões de uma relação n-dimensional é inevitável que haja algumas distorções quanto à dissemelhança. Essa distorção pode ser medida a partir do coeficiente de correlação cofenético, que mede o quanto o dendrograma preserva as distâncias da matriz inicial de dissemelhança das do dendrograma.

Quanto mais próximo de 1 for o valor deste coeficiente, melhor preserva as distâncias originais, o que quer dizer que não existem distorções significativas no dendrograma obtido. Relativamente à distância cofenética entre duas observações, esta é a altura do dendrograma onde dois *clusters* se unem. A partir do coeficiente de correlação cofenético é possível encontrar o melhor método de ligação.

Os índices de qualidade de *clusters* são usados no procedimento de avaliação dos resultados de um algoritmo de agrupamento de forma a avaliar se o *cluster* representa o agrupamento, que se pretende. Segundo [7] as estatísticas de validação de agrupamento podem ser divididas em 3 grupos:

- ***Validação interna do cluster***, que consiste em utilizar a informação interna do processo de *clustering* para avaliar a estrutura de *cluster* sem utilizar informações externas.

  Tendo em conta que o objetivo dos algoritmos de agrupamento é dividir o conjunto de dados em *clusters* de objetos, de modo que os objetos no mesmo *cluster* sejam tão semelhantes quanto possível e os objetos em diferentes *clusters* sejam tão distintos quanto possível.

  Isto é, interessa que a distância média dentro do *cluster* seja tão pequena quanto possível e que a distância média entre *clusters* seja o maior possível. Assim, as medidas de validação interna refletem a *compacidade*, a *conexão* e a *separação* do *cluster*. Em que:

  - *Compacidade ou coesão do cluster*: mede o quão próximos estão os objetos dentro do mesmo *cluster*. Uma pequena variação dos objetos dentro do *cluster* significa uma boa compacidade, isto é, um bom agrupamento. Os diferentes índices para avaliar a compacidade dos *clusters* são baseados em medidas de distância.
  - *Separação*: mede o quão bem separado um *cluster* está de outros *clusters*. Os índices utilizados como medidas de separação incluem,

por exemplo, distâncias entre centros de *cluster* e as distâncias mínimas entre pares entre objetos em diferentes *clusters*.

- *Conexão*: mede até que ponto um objeto e os seus objetos vizinhos são colocados no mesmo *cluster*.

- ***Validação externa do cluster***, que consiste em comparar os resultados da análise de um *cluster* com um resultado conhecido externamente.

- ***Validação relativa do cluster***, que consiste em avaliar a estrutura de *cluster* variando valores de parâmetros diferentes para o mesmo algoritmo (por exemplo: variando o número de *clusters* k). Geralmente, é usado para determinar o número correto de clusters.

Neste relatório serão utilizados os critérios de validação relativa do *cluster* para validar o critério de ligação, os critérios de validação interna do *cluster* para encontrar o número de *clusters* correto e o critério de validação externa para medir a precisão dos resultados. No caso de serem analisados dados reais não se possui informação acerca do verdadeiro agrupamento dos dados utilizados e por isso não se utilizam os critérios de validação externa.

Descrevem-se de seguida os índices de validação utilizados neste trabalho.

Índice de Silhouette

O *índice de Silhouette* é um índice de validação interno que mede o quão próximo os objetos num dado *cluster* estão dos objetos desse *cluster*, em comparação com os objetos doutros *clusters*. O valor de *Silhouette* S_k para k *clusters* é dado por:

$$S_k = \frac{1}{N} \sum_{i=1}^{k} \sum_{x \in C_i} \frac{\mathrm{b}(x, C_i) - \mathrm{a}(x, C_i)}{\max(\mathrm{b}(x, C_i) - \mathrm{a}(x, C_i))}$$

onde,

$$\begin{aligned} \mathrm{a}(x, C_i) &= \frac{1}{|C_i|} \sum_{y \in C_i} d(x, y) \\ \mathrm{b}(x, C_i) &= \min_{1 \leq j \leq k, i \neq j} \left(\frac{1}{|C_j|} \sum_{y \in C_j} d(x, y) \right) = \min_{1 \leq j \leq k, i \neq j} (\mathrm{a}(x, C_j)) \end{aligned}$$

- O valor de *Silhouette* varia entre -1 e 1;
- Um valor próximo de 1 significa que os objetos foram bem agrupados;
- Um valor próximo de 0 significa que os objetos estão entre dois *clusters*;

- Um valor negativo significa os objetos foram associados ao *cluster* errado.

Índice Dunn

O índice de validação interno *Dunn* D_k para k *clusters* é definido por :

$$D_k = \frac{\min\limits_{1\leq i<j\leq k} \delta(C_i, C_j)}{\max\limits_{1\leq z\leq k} \Delta(C_z)}$$

onde, $\delta(C_i, C_j)$ é dado a partir do cálculo do valor mínimo da dissemelhança entre os objetos de diferentes *clusters* e $\Delta(C_i)$ é dado pelo valor máximo de dissemelhança entre os objetos de um mesmo *cluster*, isto é,

$$\delta(C_i, C_j) = \min_{x\in C_i, y\in C_j} d(x, y)$$

$$\Delta(C_i) = \max_{x,y\in C_i} d(x, y)$$

- Se o conjunto de dados apresentar *clusters* compactos e bem separados então o diâmetro dos *clusters* deverá ser um valor pequeno e a distância entre os *clusters* deverá ser um valor grande. Assim, quanto maior é o valor do índice melhor.

Calinski-Harabaz

Para os k *clusters*, o valor de *Calinski-Harabaz* CH_k é um índice de validação interno dado pelo quociente entre a média das dispersões entre *clusters* e a dispersão no *cluster*. O índice Calinski-Harabasz para k *clusters* é definido por:

$$CH_k = \frac{N-k}{k-1} \times \frac{\sum\limits_{i=1}^{k} |C_i| d(\overline{C_i}, \overline{X})}{\sum\limits_{i=1}^{k} \sum\limits_{x\in C_i} d(x, \overline{C_i})}$$

onde N é o número total de objetos.

- Um maior resultado no *Índice Calinski-Harabaz* relaciona-se com um modelo com *clusters* melhor definidos.

Fowlkes-Mallows

O índice de validação externa *Fowlkes-Mallows* FM permite calcular a precisão de um resultado de *clustering* quando se sabe o verdadeiro agrupamento dos dados.

Este índice de *Fowlkes-Mallows* é definido por:

$$FM = \frac{VP}{\sqrt{(VP + FP)(VP + FN)}}$$

onde VP é o número de *verdadeiros positivos*, FP é o número de *falsos positivos* e FN é o número de *falsos negativos* de U.

Define-se um *Verdadeiro Positivo* quando dois objetos pertencem ao mesmo *cluster* no agrupamento verdadeiro e no resultado do método de *clustering*; um *Falso Positivo* quando dois objetos pertencem ao mesmo *cluster* no agrupamento verdadeiro e a *clusters* diferentes no resultado do método de *clustering*; um *Falso Negativo* quando dois objetos pertencem ao mesmo *cluster* no resultado do método de *clustering* e a *clusters* diferentes no agrupamento verdadeiro.

- O valor deste índice varia entre 0 e 1;
- Um valor próximo de 1 significa que existe uma grande similaridade entre os *clusters* obtidos e os *clusters* corretos.

Capítulo 5

Aplicação do método para a identificação de participações semelhantes

O objetivo desta fase é a identificação de participações semelhantes. Para tal, será utilizado o algoritmo de *clustering* hierárquico cujo objetivo é agrupar as participações mais próximas geograficamente.

Devido ao pouco volume de dados, todas as participações que serão utilizadas na secção seguinte foram criadas sinteticamente em **Python**.

Para o uso do método também é necessário a escolha da métrica, do critério de ligação e do valor máximo de dissemelhança. Dada a natureza dos dados, será utilizada a métrica euclidiana. O critério de ligação e o valor máximo de dissemelhança são valores mais difíceis de definir. Por esse motivo, serão criados diferentes cenários com o intuito de determinar, em geral, qual o melhor critério de ligação. De forma a ilustrar os possíveis cenários que se poderiam encontrar nos dados da aplicação, foram criados cenários isotrópicos, que se caracterizam pela sua uniformidade. Tendo em conta que numa situação real as fotografias são à volta do objeto, pensa-se que os cenários isotrópicos retratam bem esta situação.

5.1 As variáveis

A escolha da informação a ser utilizada para o processo de agregação é um passo importante no uso de mecanismos de *clustering*, pois é a partir dessa informação que serão inferidos os resultados. Tendo em conta o método utilizado e o objetivo desta fase, as informações mais importantes são a latitude e a longitude. Como foi visto anteriormente, cada participação tem associadas outras informações importantes, mas algumas delas são difíceis de analisar, tendo em

conta a informação contida na base de dados. Por exemplo, a descrição e a fotografia podem conter informações importantes que diferenciam participações que estejam muito próximas geograficamente, mas que não representam o mesmo problema. Neste caso, tais características não poderão ser utilizadas porque este tipo de informação requer métodos de aprendizagem supervisionada e um grande volume de dados. Por outro lado, a categoria e a freguesia são características que poderão ser utilizadas e que ajudarão a diferenciar participações.

Neste estudo será suposto que cada participação ficará definida pelas suas coordenadas geográficas e cada grupo de participações analisado pertence à mesma categoria e à mesma freguesia.

Seja n o número de eventos e $1 \leq i \leq n$. O *evento* i é denotado por $E_i \in \mathbb{R}^2$, e é definido por:

$$E_i = (latitude_i, longitude_i)$$

Seja m_i o número de participações associadas a um evento i e $1 \leq j \leq m_i$.

A *participação* j relativa ao evento i é denotada por P_{ij} tal que:

$$P_{ij} = (latitude_{ij}, longitude_{ij})$$

5.2 Criação de cenários isotrópicos

Os cenários isotrópicos (Figura 5.1) caracterizam-se pela uniformidade da distribuição das participações em todas as direções.

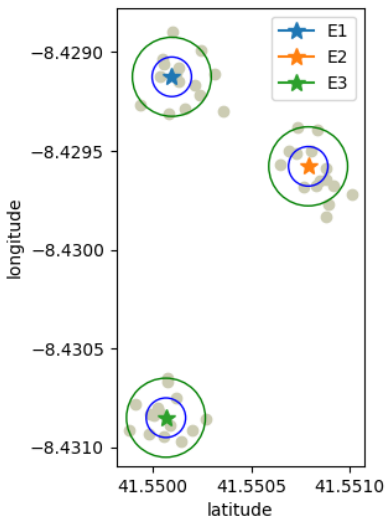


Figura 5.1: Cenário isotrópico

De um forma geral, teremos o seguinte:

> Seja E_i um evento tal que $E_i = (latitude_i, longitude_i)$.
> Seja $\mathbf{P}_i$ o conjunto das m_i participações associadas ao evento E_i, em que $1 \leq j \leq m_i$ e $P_{ij} \in \mathbf{P}_i$. As participações P_{ij} são provenientes de uma distribuição normal bivariada, tal que cada $P_{ij} \sim N_2(E_i, \sigma_i I_2)$ e σ_i está relacionado com a distância máxima entre as participações e o evento E_i. Assim, $\mathbf{P}_i$ forma o *cluster* C_i.

Na Figura 5.1, cada evento E_1, E_2 e E_3 é representado por uma estrela e tem associadas participações, representadas a cinzento, sendo que o círculo azul de raio σ_1, σ_2 e σ_3, respetivamente, representa a área onde estão, aproximadamente, 68% das participações e o círculo verde de raio $2\sigma_1$, $2\sigma_2$ e $2\sigma_3$, respetivamente, representa a área onde estão, aproximadamente, 95% das participações.

As participações podem ser vistas como participações semelhantes referentes ao evento que as originou. Como foi dito anteriormente, o conjunto de participações segue uma distribuição normal com valor médio dado pelo evento e a variância um valor σ escolhido.

Dado que cada participação é definida por uma latitude e uma longitude e estas não são dependentes e não existe correlação entre elas, a criação das participações P_{ij} associadas ao evento E_i consiste em criar amostras aleatórias *Latitude* e *Longitude* tal que $Latitude \sim N(lat_i, \sigma_i)$ e $Longitude \sim N(long_i, \sigma_i)$.

Assim, o objetivo é fazer variar a posição de E_i e alterar σ_i para determinar qual dos métodos consegue identificar melhor os *clusters*.

Cada cenário será composto por cinco eventos, E_1, E_2, E_3, E_4 e E_5, e cinco conjuntos de participações com 10 pontos cada, formando cinco *clusters* distintos. Cada evento representa a posição real do problema reportado e as participações representam participações da comunidade relativas ao evento. As participações de cada *cluster* serão identificadas da seguinte forma,

- As participações do *cluster* com centro E_1 serão representada por $P_{1,1}$, $P_{1,2}$, $P_{1,3}$, $P_{1,4}$, $P_{1,5}$, $P_{1,6}$, $P_{1,7}$, $P_{1,8}$, $P_{1,9}$, $P_{1,10}$.
- As participações do *cluster* com centro E_2 serão representada por $P_{2,1}$, $P_{2,2}$, $P_{2,3}$, $P_{2,4}$, $P_{2,5}$, $P_{2,6}$, $P_{2,7}$, $P_{2,8}$, $P_{2,9}$, $P_{2,10}$.
- As participações do *cluster* com centro E_3 serão representada por $P_{3,1}$, $P_{3,2}$, $P_{3,3}$, $P_{3,4}$, $P_{3,5}$, $P_{3,6}$, $P_{3,7}$, $P_{3,8}$, $P_{3,9}$, $P_{3,10}$.
- As participações do *cluster* com centro E_4 serão representada por $P_{4,1}$, $P_{4,2}$, $P_{4,3}$, $P_{4,4}$, $P_{4,5}$, $P_{4,6}$, $P_{4,7}$, $P_{4,8}$, $P_{4,9}$, $P_{4,10}$.

- As participações do *cluster* com centro E_5 serão representada por $P_{5,1}$, $P_{5,2}$, $P_{5,3}$, $P_{5,4}$, $P_{5,5}$, $P_{5,6}$, $P_{5,7}$, $P_{5,8}$, $P_{5,9}$, $P_{5,10}$.

Para todos os cenários serão apresentados os dendrogramas obtidos pelo algoritmo usando os diferentes critérios, assim como, a análise da variação da dissemelhança conforme o número de *clusters* formados em cada passo do método. Não estando ainda definido o valor de limite para a dissemelhança, será utilizado o método *linkage* para executar o método de *clustering* e o método *dendrogram* para apresentar toda a hierarquia produzida pelo método de *clustering*. Ambos os métodos são da biblioteca **Scipy**.

Cenário 1

O primeiro cenário que se apresenta na Figura 5.2 é composto por cinco eventos associados a conjuntos de participações onde se conseguem identificar perfeitamente cinco *clusters*. Para a construção deste cenário foi considerado $\sigma = 0.001$ para todos os conjuntos.

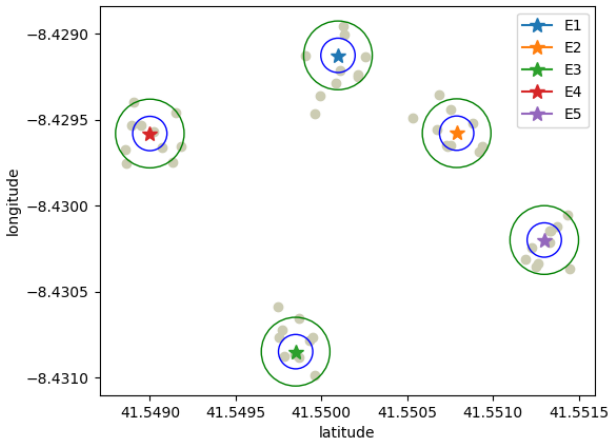
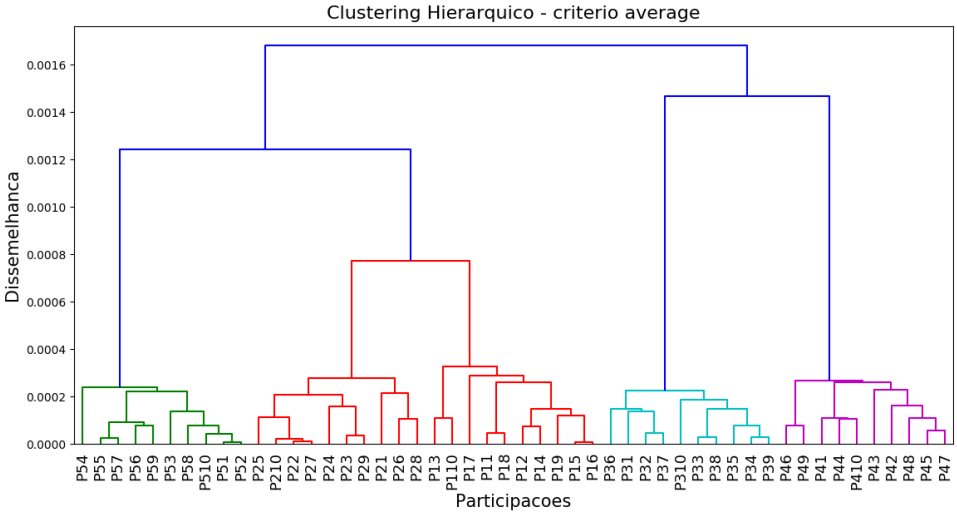


Figura 5.2: Representação gráfica do cenário 1

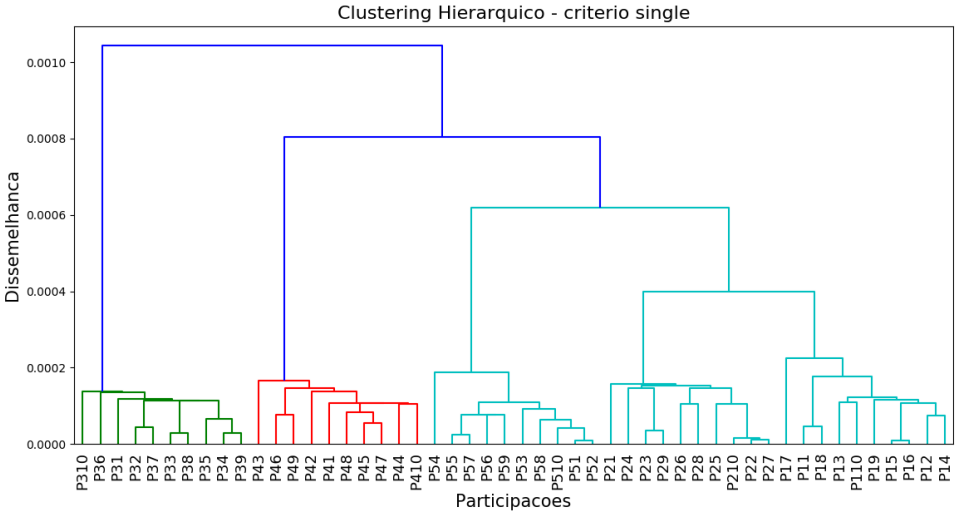
A figura seguinte (Figura 5.3) contém os resultados obtidos aplicando os diferentes critérios de ligação.

No que diz respeito aos valores de coeficiente cofenético, o critério que obteve um valor do coeficiente cofenético mais alto foi o critério *group average*, com um valor aproximado de 0.862, o seguinte foi o *critério complete*, com um valor aproximado de 0.829 e por fim, o critério *single* com um valor de coeficiente cofenético de aproximadamente 0.785.

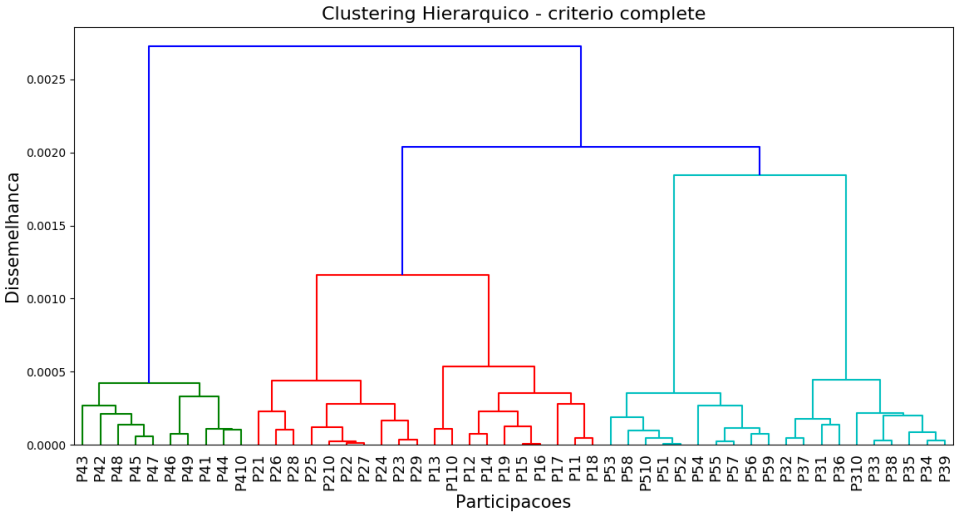
Ao analisar os resultados obtidos em 5.3 é possível concluir que os três critérios apresentam bons resultados, uma vez que foi possível identificar visualmente os *clusters*, que correspondem exatamente aos cinco *clusters* corretos.



(a) Dendrograma obtido utilizando o critério de ligação *group average*



(b) Dendrograma obtido utilizando o critério de ligação *single*



(c) Dendrograma obtido utilizando o critério de ligação *complete*

Figura 5.3: Resultados obtidos para o cenário 1

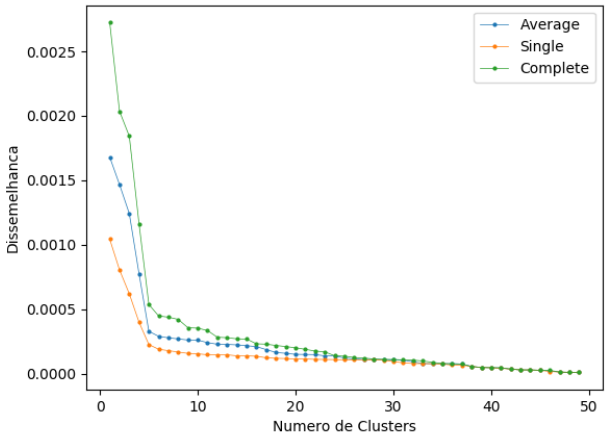


Figura 5.4: Variação do valor de dissemelhança no Cenário 1

No gráfico 5.4 está representada a variação da dissemelhança em função do número de *clusters*, para os três critérios utilizados. Tal como se pode verificar, as representações correspondentes aos três critérios têm o mesmo comportamento, isto é, ocorre uma diminuição acentuada do valor da dissemelhança até ao valor cinco e a partir desse valor a diminuição torna-se mais gradual. Graficamente é possível perceber que, para o número de *clusters* igual a cinco, a curvatura é máxima. Esta análise sugere o valor da dissemelhança máximo correspondente a cinco *clusters*, para os três critérios.

Cenário 2

O cenário da Figura 5.5 é composto por cinco eventos e apresenta conjuntos de participações mais próximos do que o cenário anterior. O valor de σ utilizado foi de 0.00019. O objetivo é avaliar se o processo de *clustering* consegue identificar os cinco conjuntos iniciais quando temos uma dispersão maior das participações.

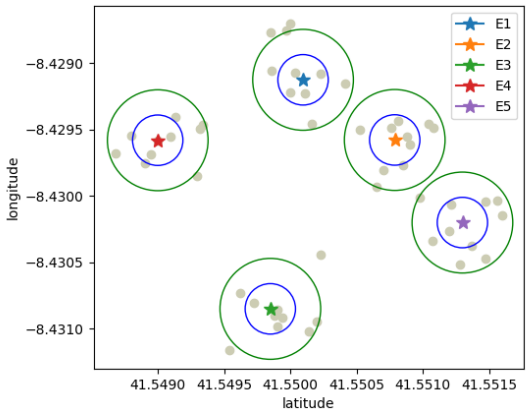
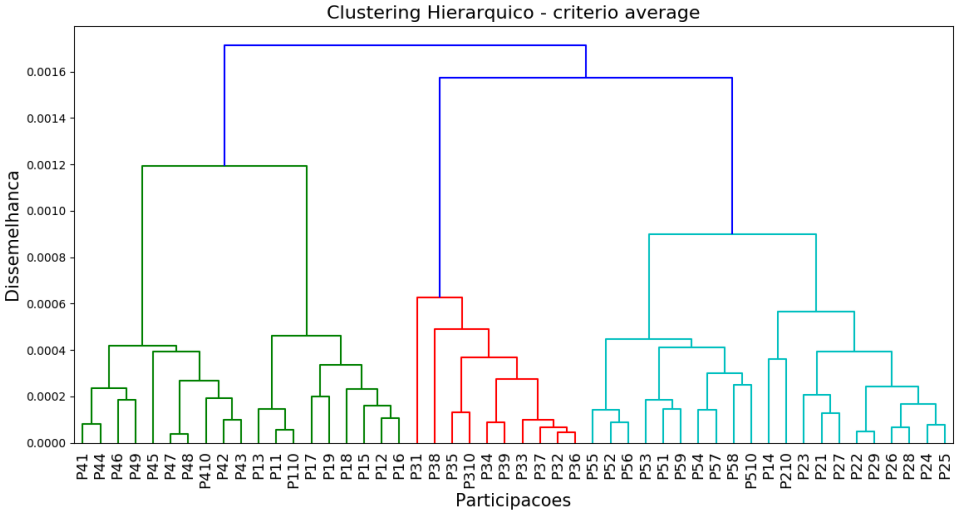
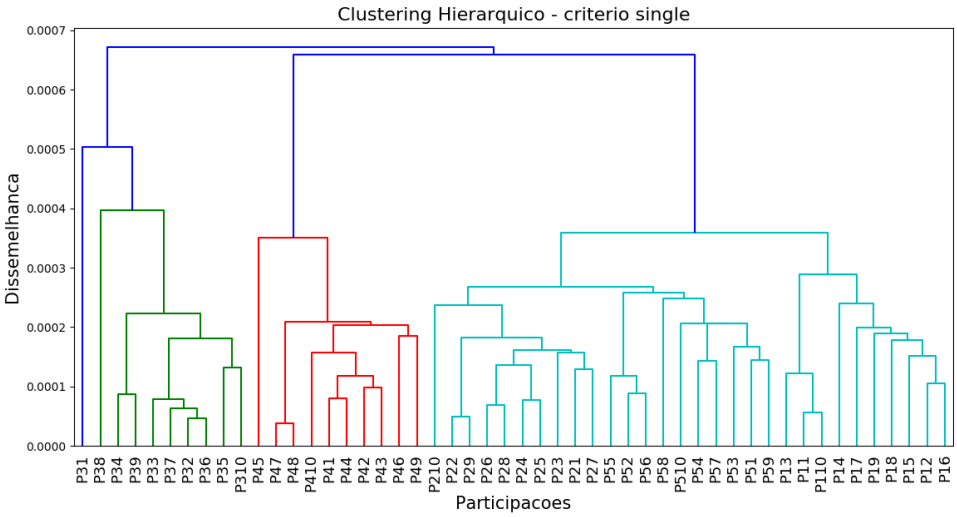


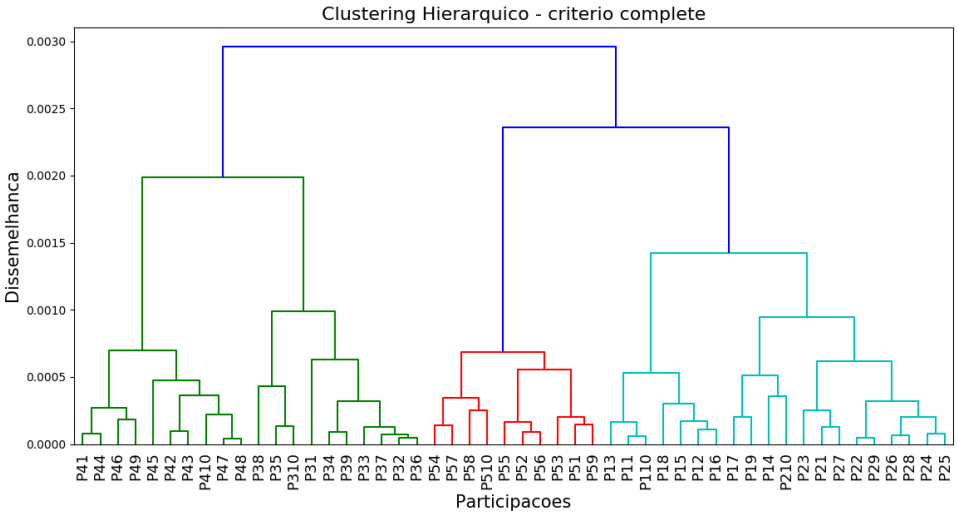
Figura 5.5: Representação gráfica do cenário 2



(a) Dendrograma obtido utilizando o critério de ligação *group average*



(b) Dendrograma obtido utilizando o critério de ligação *single*



(c) Dendrograma obtido utilizando o critério de ligação *complete*

Figura 5.6: Resultados obtidos para o cenário 2

A Figura 5.6 contém os resultados obtidos aplicando os critérios de ligação.

Ao analisar os resultados obtidos é possível concluir que os grupos identificados utilizando o critério *single*, não estão tão bem definidos como nos outros dois critérios. No que diz respeito ao critério *group average*, é de notar que a participação $P_{1,4}$ foi associada, de forma incorreta, ao *cluster* com centro E_2.

Quanto ao critério *complete* verifica-se que as participações $P_{1,4}$, $P_{1,7}$ e $P_{1,9}$ foram mal associadas ao *cluster* com centro E_2. Conclui-se assim, que o critério que apresentou melhores resultados foi o *group average*.

No que diz respeito aos valores de coeficiente cofenético, o critério que obteve um valor do coeficiente cofenético mais alto foi o critério *group average*, com um valor aproximado de 0.802, o seguinte foi o critério *complete*, com um valor aproximado de 0.769 e por fim o critério *single* com um valor de coeficiente cofenético de aproximadamente 0.739. Verifica-se uma diminuição dos valores dos coeficientes em relação ao cenário anterior, principalmente o critério *complete*.

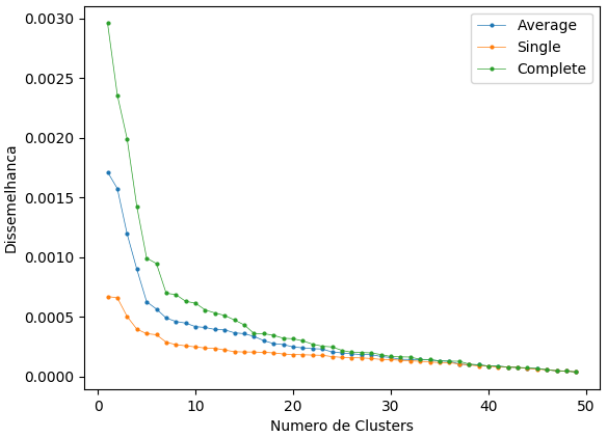


Figura 5.7: Variação do valor de dissemelhança no Cenário 2

Ao analisar o gráfico 5.7 pode concluir-se que o valor da dissemelhança apresenta uma diminuição mais gradual para valores superiores a 5 *clusters* para o critério *group average*. Quanto ao critério *complete* essa diminuição ocorre mais tarde, para valores superiores 7. O critério *single* apresenta um comportamento mais uniforme durante todo o processo de *clustering*, não havendo uma diminuição tão acentuada nos valores de dissemelhança. Consequentemente, com este critério não é possível concluir um valor ideal para o número de *clusters*. Em relação ao cenário 1 a mudança de comportamento das curvas é menos acentuado mas continua a ser possível identificar o ponto de curvatura máximo que corresponde ao número ideal de *clusters* para os critérios *group average* e *complete*.

Cenário 3

O cenário seguinte que se apresenta na Figura 5.8 é composto por cinco eventos e apresenta conjuntos de quatro *clusters* bem definidos e um *cluster* com participações mais espalhadas. O valor de σ utilizado para os *clusters* de centro E_1, E_3, E_4 e E_5 foi de 0.00019 e para o *cluster* com centro E_2 foi de 0.0003. Neste caso, temos uma interferência forte entre as participações associadas aos diferentes eventos, e por isso, com um risco elevado de um agrupamento errado.

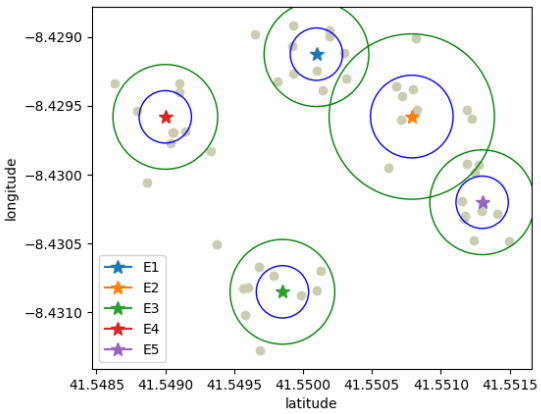
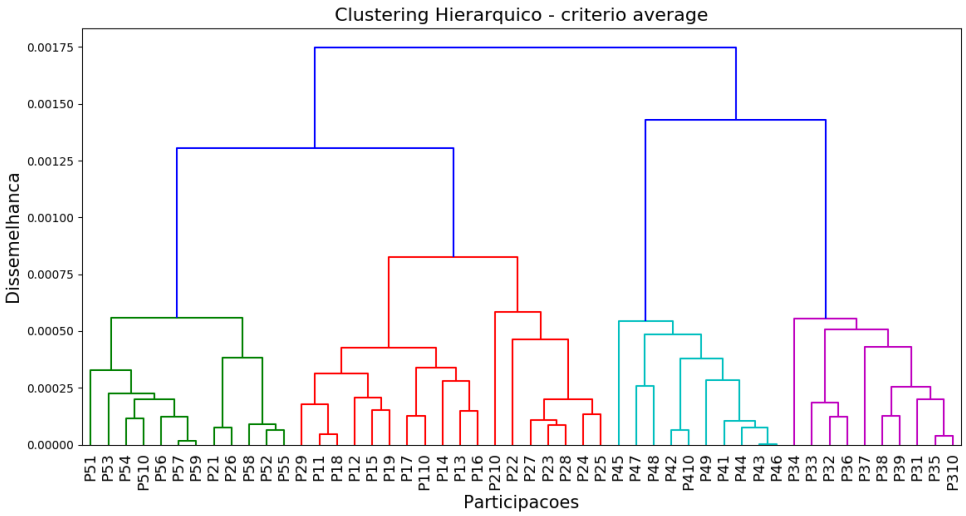


Figura 5.8: Representação gráfica do cenário 3

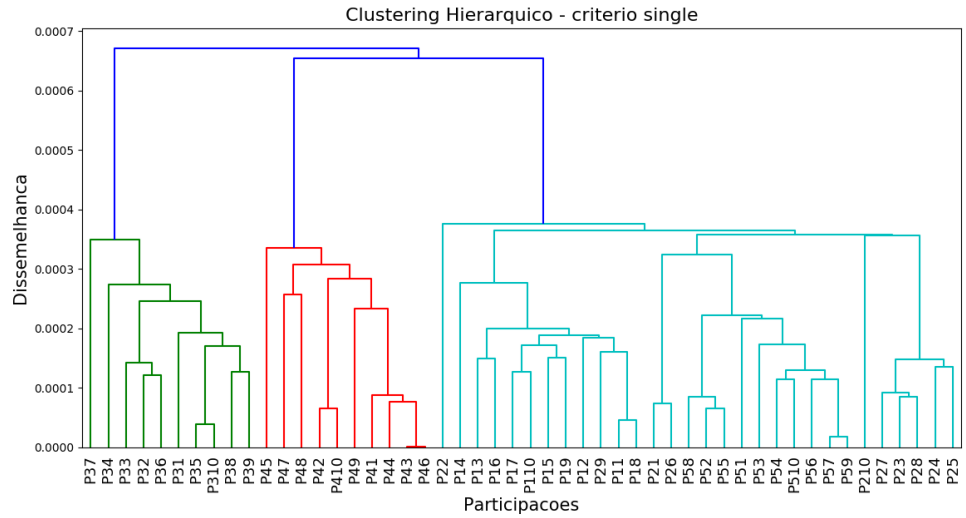
A figura seguinte (Figura 5.9) contém os resultados obtidos aplicando os diferentes critérios de ligação.

Como seria de esperar, algumas das participações correspondentes ao *cluster* com centro E_2 foram associadas aos *clusters* com centro E_1 e E_5. Estas situações foram observadas nos critérios *complete* e *group average*, sendo que o critério *complete* apenas associou uma participação ao *cluster* errado. Quanto ao critério *single* verificou-se que não é possível identificar os grupos de *clusters*.

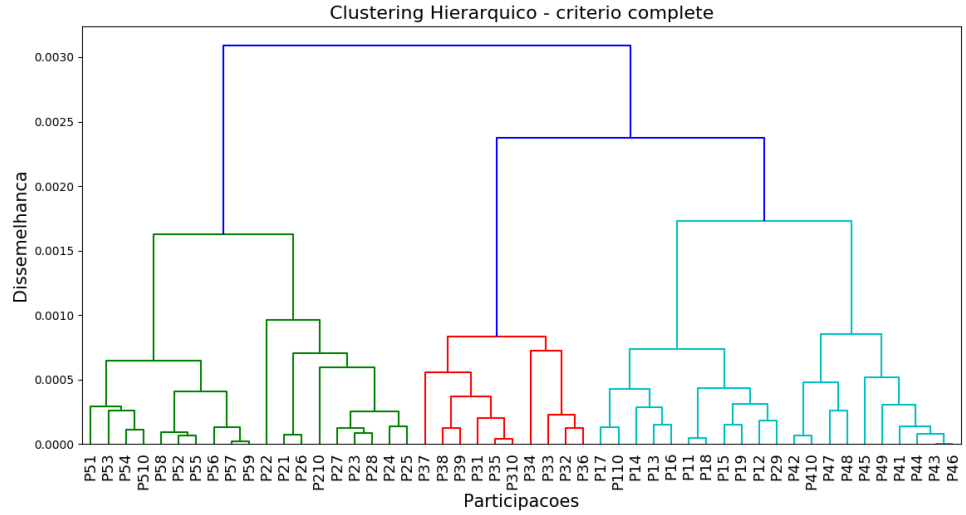
Apesar do critério *complete* ter produzido melhores resultados, o coeficiente cofenético mais alto foi o obtido pelo critério *group average* (aproximadamente, 0.812), seguido do critério *complete* (aproximadamente, 0.777) e, por último, o critério *single* (aproximadamente, 0.747).



(a) Dendrograma obtido utilizando o critério de ligação *group average*



(b) Dendrograma obtido utilizando o critério de ligação *single*



(c) Dendrograma obtido utilizando o critério de ligação *complete*

Figura 5.9: Resultados obtidos para o cenário 3

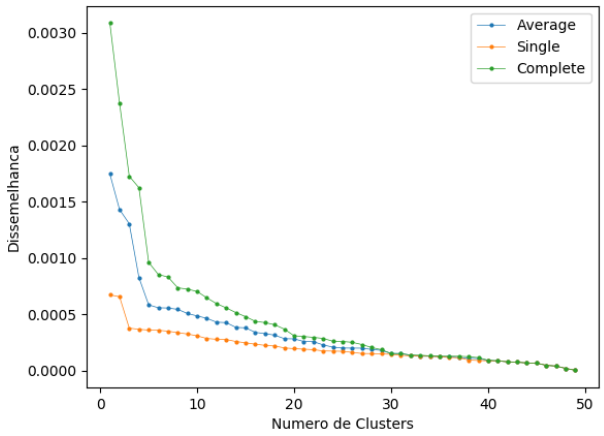


Figura 5.10: Variação do valor de dissemelhança no cenário 3

Ao analisar o gráfico 5.10 pode verificar-se que os valores da dissemelhança diminuem de forma mais rápida até ao valor correspondente a 5 *clusters* para os critérios *group average* e *complete*, apresentando em seguida uma diminuição mais gradual. No caso do critério *single* a mudança de comportamento é notada para valores correspondes a 3 *clusters*, o que é errado.

Cenário 4

O cenário da Figura 5.11 é composto por cinco eventos e apresenta três *clusters* bem definidos e dois *clusters* com participações mais espalhadas. O σ utilizado para os *clusters* de centro E_2, E_3 e E_4 foi de 0.00019 e para os *cluster* com centro E_1 e E_3 foi de 0.0003.

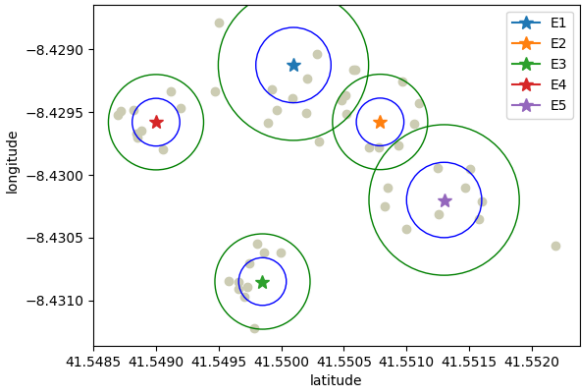
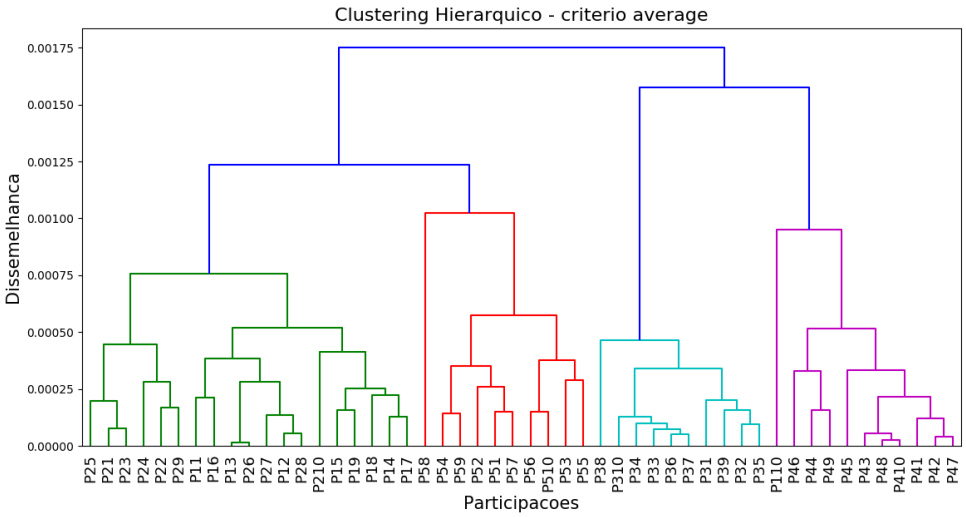
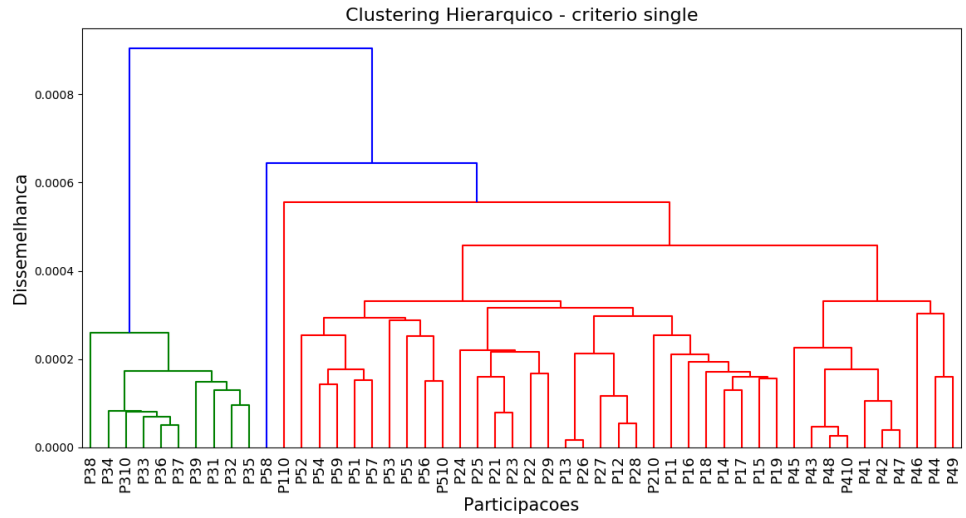


Figura 5.11: Representação gráfica do cenário 4

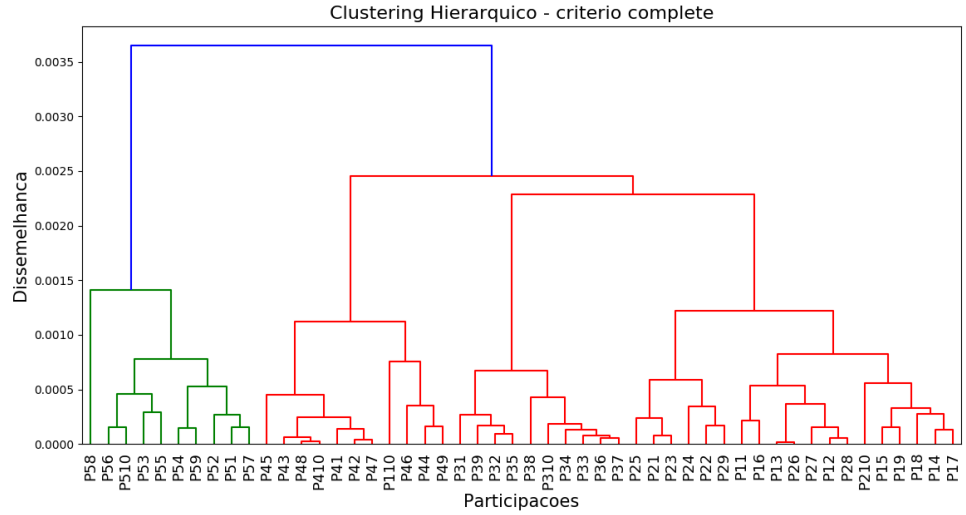
A Figura 5.12 contém os resultados obtidos com os diferentes critérios de ligação. Como era de esperar, algumas participações foram associadas ao grupo errado. Os critérios *group average* e *complete* apresentam melhores resultados, contudo nos dois casos ocorre uma associação incorreta entre as participações de centro E_1 e de centro E_2 e uma associação incorreta da participação $P_{1,10}$ às participações de centro E_4.



(a) Dendrograma obtido utilizando o critério de ligação *group average*



(b) Dendrograma obtido utilizando o critério de ligação *single*



(c) Dendrograma obtido utilizando o critério de ligação *complete*

Figura 5.12: Resultados obtidos para o cenário 4

No que diz respeito o coeficiente cofenético o valor mais alto foi o obtido pelo critério *group average* (aproximadamente, 0.780), seguido do critério *complete* (aproximadamente, 0.683) e, por último, o critério *single* (aproximadamente, 0.528). Usando o critério *group average* ainda é possível identificar que o número de *clusters* é cinco.

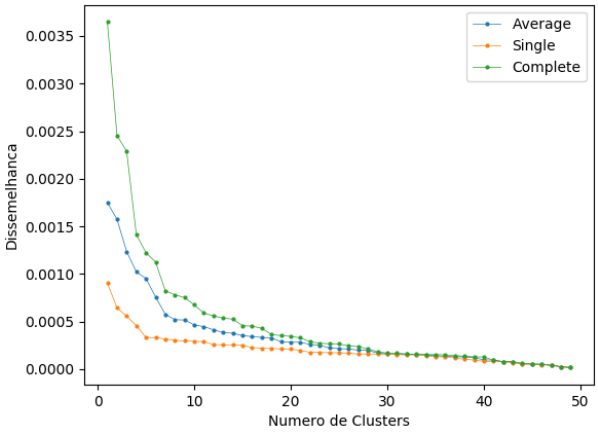


Figura 5.13: Variação do valor de dissemelhança no Cenário 4

O gráfico 5.13 permite concluir que os valores da dissemelhança diminuem de forma mais rápida até ao valor correspondente a 7 *clusters* para os critérios *average* e *complete*. No caso do critério *single* a mudança de comportamento é notada para valores correspondes a 5 *clusters*.

Cenário 5

O cenário seguinte (Figura 5.14) é composto por cinco eventos e apresenta participações espalhadas. O σ utilizado para os *clusters* de centro E_1, E_2, E_4 e E_5 foi de 0.0003 e para o *cluster* com centro E_3 foi de 0.0005. O objetivo passa por experimentar um conjunto de *clusters* muito misturado, com uma zona de influência muito diferente em que $\sigma_{E_3} >> \sigma_{E_1} = \sigma_{E_2} = \sigma_{E_4} = \sigma_{E_5}$.

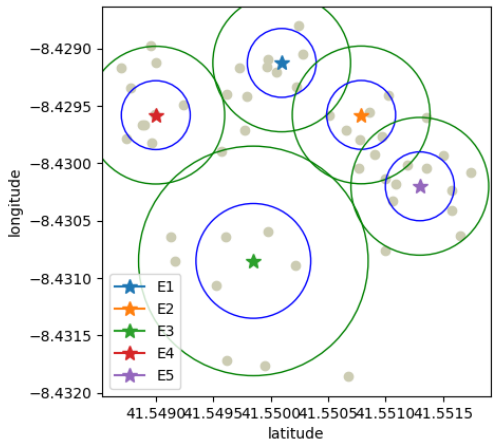
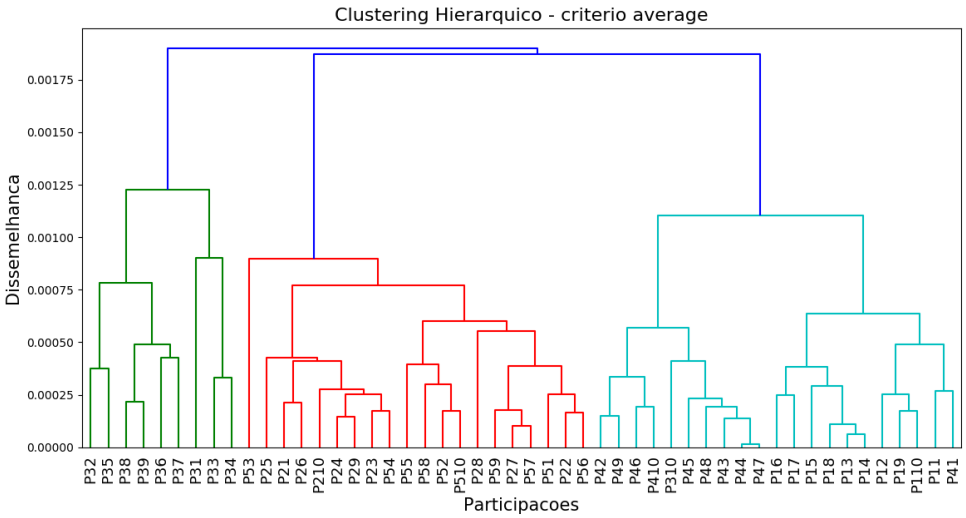
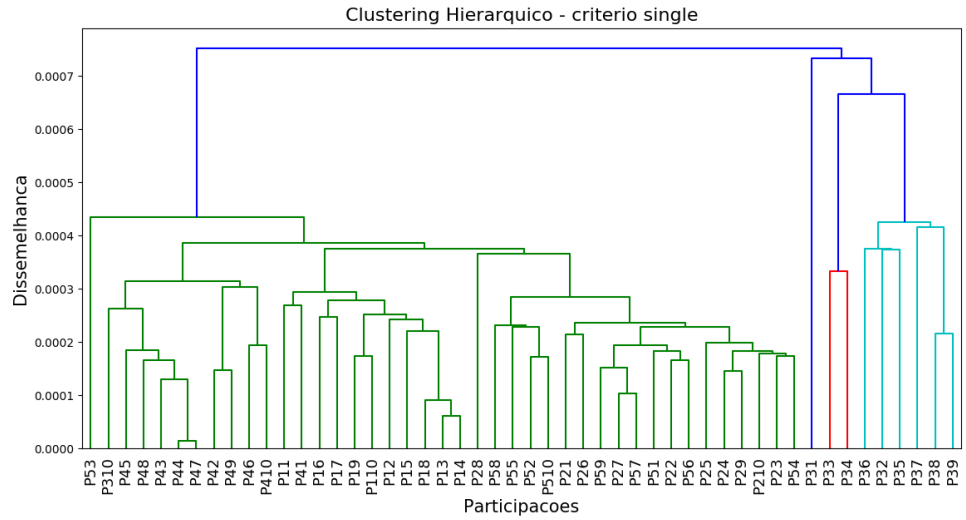


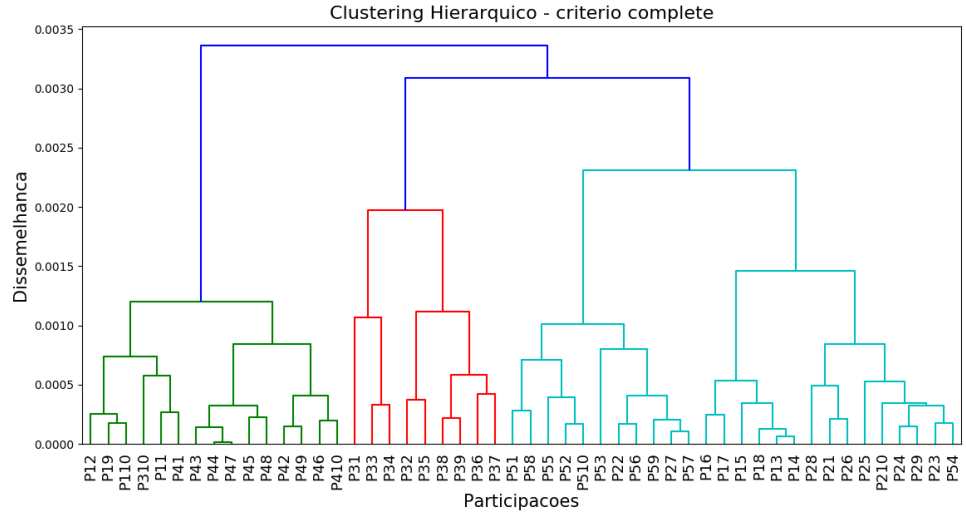
Figura 5.14: Representação gráfica do cenário 5



(a) Dendrograma obtido utilizando o critério de ligação *group average*



(b) Dendrograma obtido utilizando o critério de ligação *single*



(c) Dendrograma obtido utilizando o critério de ligação *complete*

Figura 5.15: Resultados obtidos para o cenário 5

Neste cenário pode concluir-se que foram produzidos melhores resultados para os critérios *group average* e *complete*. Contudo, de uma forma geral, os grupos identificados visualmente apresentam participações mal associadas, o que seria de esperar dado que as participações de centros diferentes estão sobrepostas.

O coeficiente cofenético mais alto foi obtido pelo critério *group average* (aproximadamente, 0.773), seguido do critério *complete* (aproximadamente, 0.708) e, por último, o critério *single* (aproximadamente, 0.500).

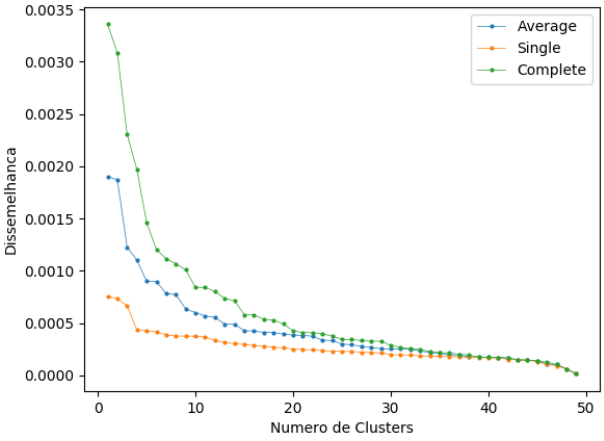


Figura 5.16: Variação do valor de dissemelhança no Cenário 5

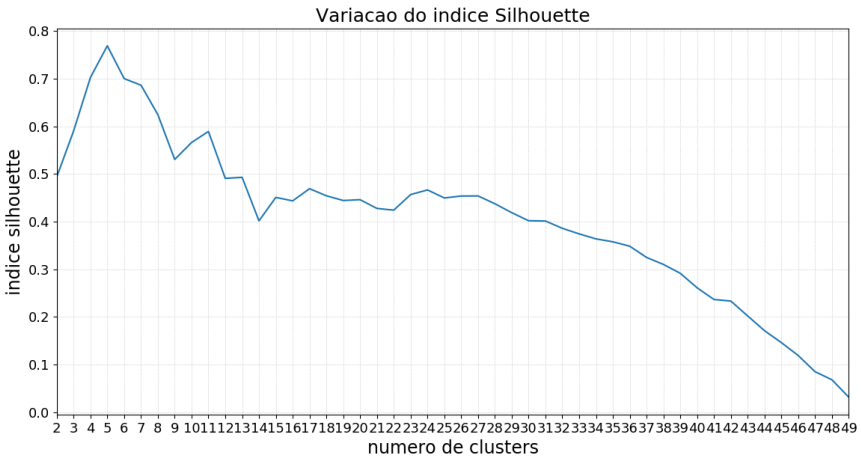
Ao analisar o gráfico 5.16 pode-se verificar que os valores da dissemelhança diminuem de forma mais rápida até ao valor correspondente a 4 *clusters* para o critério *single*, 9 para o critério *group average* e 6 no caso do critério *complete*.

5.3 Análise dos indicadores

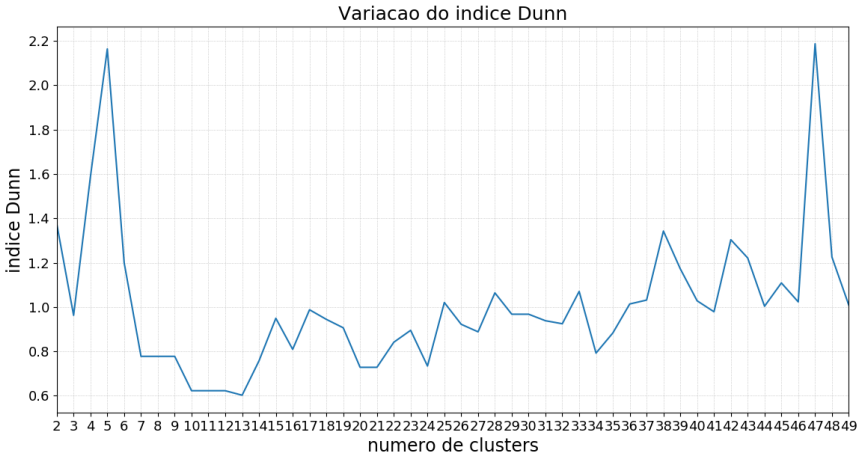
Do estudo apresentado na secção anterior é possível verificar que o critério *group average* apresenta, de uma forma geral, melhores resultados na identificação do número de *clusters*. Assim, para a análise dos *clusters* serão utilizados os resultados obtidos por este critério.

Para este estudo vão ser comparados os resultados para diferentes números de *clusters* nos diferentes cenários. Tendo em conta que o número de *clusters* que se pode concluir de forma intuitiva é 5, serão apresentados os resultados dos índices de validação num intervalo de valores próximos de 5.

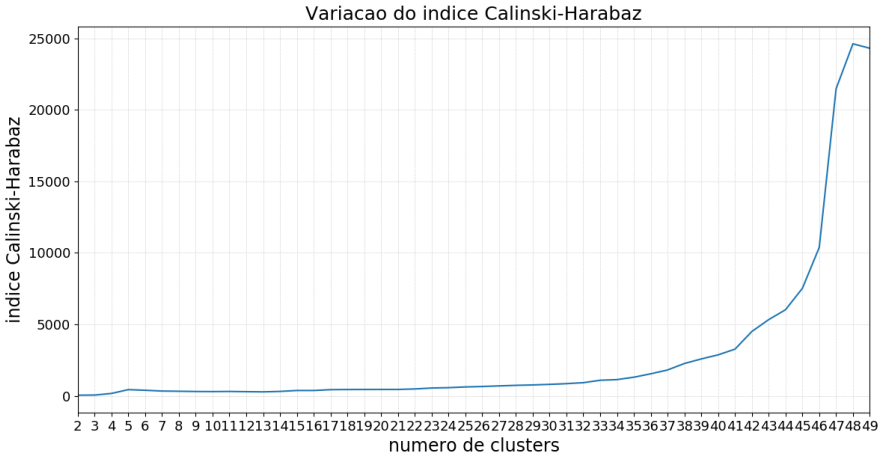
Cenário 1



(a) Variação do índice de Silhouette



(b) Variação do índice de Dunn



(c) Variação do índice de Calinski-Harabasz

Figura 5.17: Resultados dos índices obtidos para o cenário 1

Por análise do gráfico 5.17a pode verificar-se que é obtido um valor para o índice de *Silhouette* mais próximo de 1 para o valor correspondente a 5 *clusters*.

Quanto à análise do gráfico 5.17b verifica-se que os valores do índice *Dunn* mais elevados correspondem a 5 e a 47 *clusters*.

Por análise do gráfico 5.17c verifica-se que, de uma forma geral, o valor do índice *Calinski-Harabasz* aumenta à medida que o número de *clusters* aumenta. Contudo, numa inspeção mais próxima, é possível verifica-se um valor de índice máximo local para 5 *clusters*.

Comparação dos índices no cenário 1

nº clusters	**Silhoutte**	**Dunn**	**Calinski-Harabasz**
k=4	0.703	1.588	167.963
k=5	**0.769**	**2.164**	**436.569**
k=6	0.700	1.200	390.617
k=7	0.686	0.777	337.234
k=8	0.625	0.777	323.196

Tabela 5.1: Apresentação dos índices dos três métodos para o cenário 1

Através dos resultados apresentados na Tabela 5.1 pode verificar-se que os três métodos apresentam melhores resultados no caso em que o número de *clusters* é 5, o que corresponde à constituição inicial.

Índice *Fowlkes-Mallows*

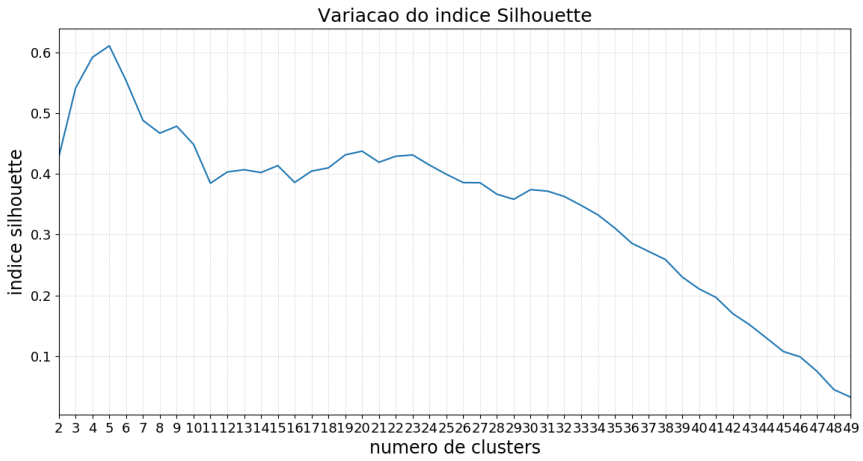
$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
0.832	**1.000**	0.964	0.948	0.897

Tabela 5.2: Apresentação do índice Fowlkes-Mallows para o cenário 1

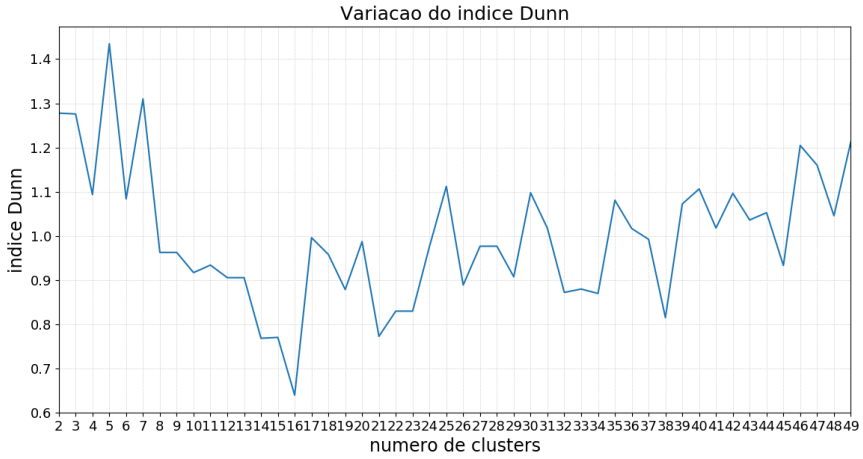
Através dos resultados apresentados na Tabela 5.2 pode verificar-se que o método apresenta o resultado de 1 quando o k é 5.

Portanto, os 5 *clusters* concluídos a partir dos índices de validação interna representam os *clusters* originais.

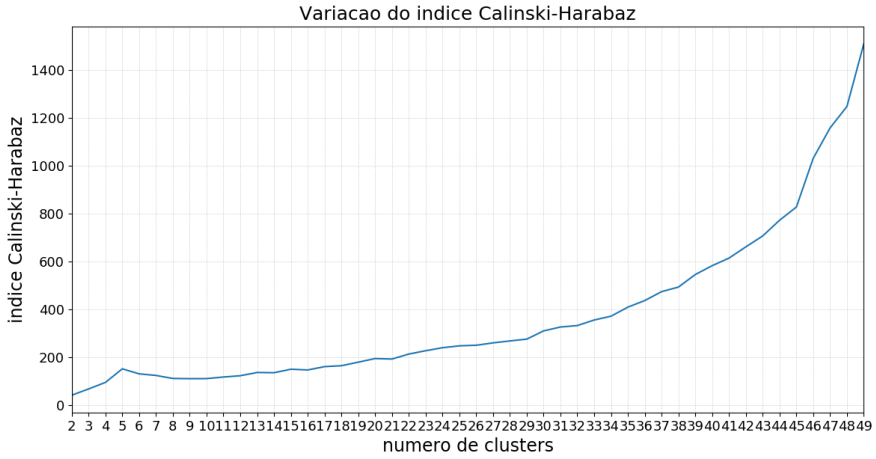
Cenário 2



(a) Variação do índice de Silhouette



(b) Variação do índice de Dunn



(c) Variação do índice de Calinski-Harabasz

Figura 5.18: Resultados dos índices obtidos para o cenário 2

Analisando o gráfico 5.18a, pode verificar-se um valor mais próximo de 1 para o índice *Silhouette* correspondente a 5 *clusters*. Quanto ao gráfico 5.18b verifica-se um valor do índice *Dunn* mais elevado para 5 *clusters*. Por análise do gráfico 5.18c verifica-se que o valor do índice *Calinski-Harabasz* aumenta à medida que o número de *clusters* aumenta, verificando-se também um máximo local para 5 *clusters*.

Comparação dos índices no cenário 2

nº clusters	**Silhoutte**	**Dunn**	**Calinski-Harabasz**
n=4	0.592	1.094	94.433
n=5	**0.610**	**1.434**	**151.025**
n=6	0.553	1.084	130.072
n=7	0.488	1.310	123.398
n=8	0.467	0.963	110.747

Tabela 5.3: Apresentação dos índices dos três métodos para o cenário 2

Através dos resultados apresentados na Tabela 5.3 pode verificar-se, tal como no cenário anterior, que os três métodos apresentam melhores resultados no caso em que o número de *clusters* é 5.

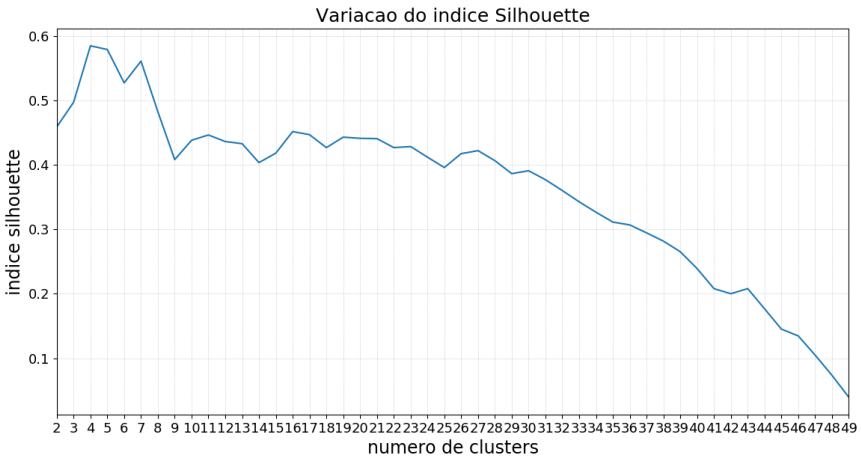
Índice *Fowlkes-Mallows*

$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
0.786	**0.958**	0.937	0.936	0.917

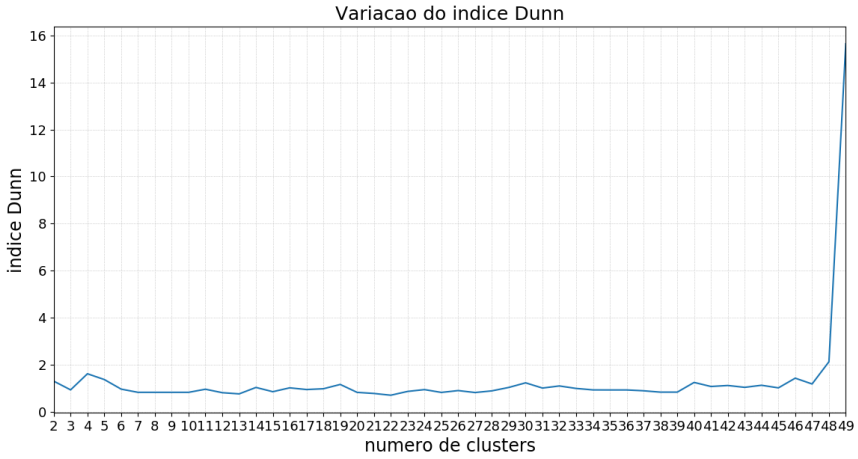
Tabela 5.4: Apresentação do índice Fowlkes-Mallows para o cenário 2

Através dos resultados apresentados na Tabela 5.4 pode verificar-se que o método apresenta o resultado mais próximo de 1 quando o k é 5 de onde se conclui que os *clusters* obtidos apresentam uma grande similaridade com os *clusters* corretos, com cerca de 5% de erro.
Aplicando o método *fcluster* da biblioteca **Scipy** com o valor de dissemelhança máximo correspondente a 5 *clusters*, obtiveram-se os seguintes resultados:
O *cluster* com todas as participações associadas ao evento E_4;
O *cluster* constituído pela maioria das participações associadas ao evento E_1;
O *cluster* com todas as participações associadas ao evento E_3;
O *cluster* com todas as participações associadas ao evento E_5;
O *cluster* com todas as participações do evento E_2 e a participação $P_{1,3}$.
Assim, todas as participações foram associadas corretamente à exceção de uma.

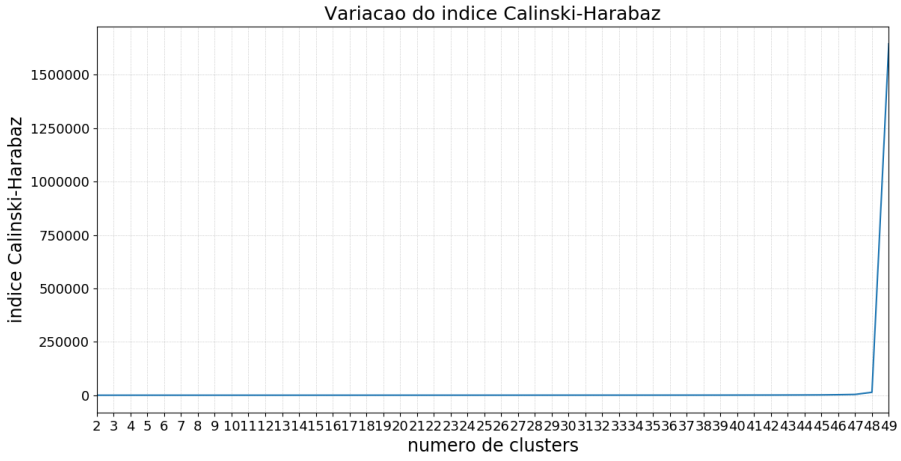
Cenário 3



(a) Variação do índice de Silhouette



(b) Variação do índice de Dunn



(c) Variação do índice de Calinski-Harabasz

Figura 5.19: Resultadosdos índices obtidos para o cenário 3

Por análise do gráfico 5.19a, verifica-se um valor do índice *Silhouette* mais próximo de 1 para 4 *clusters*. Quanto à análise do gráfico 5.19b, verifica-se que o valor de índice de *Dunn* é um máximo local para 4 *clusters*. Por análise do gráfico 5.19c, verifica-se que o valor do índice *Calinski-Harabasz* não sofre grande variação em quase todo o estudo, enquanto as participações não estão todas separadas em *clusters* singulares, embora, no intervalo entre 4 e 8 *clusters* seja possível observar um máximo local para 5 *clusters*.

Comparação dos índices no cenário 3

nº clusters	**Silhoutte**	**Dunn**	**Calinski-Harabasz**
n=4	**0.585**	**1.613**	105.333
n=5	0.579	1.367	**129.689**
n=6	0.527	0.960	109.435
n=7	0.561	0.823	115.623
n=8	0.483	0.823	104.783

Tabela 5.5: Apresentação dos índices dos três métodos para o cenário 3

Através da Tabela 5.5 verifica-se que os métodos *Silhouette* e *Dunn* apresentam melhores resultados para 4 *clusters*. O método *Calinski-Harabasz* apresenta-se melhor para 5 *clusters*.

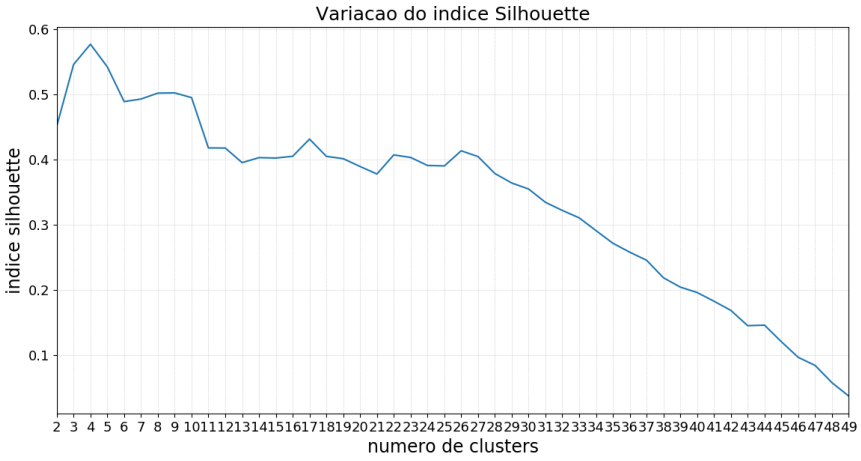
Índice *Fowlkes-Mallows*

$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
0.793	**0.884**	0.869	0.844	0.820

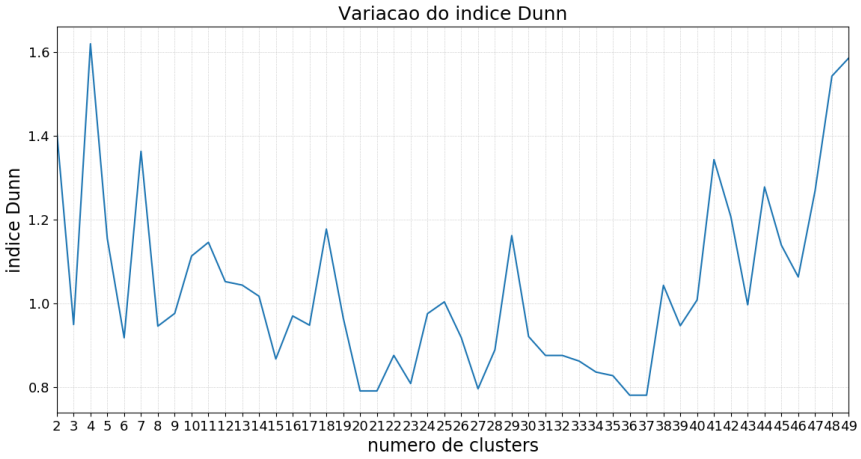
Tabela 5.6: Apresentação do índice Fowlkes-Mallows para o cenário 3

Pode verificar-se pelos resultados da Tabela 5.6 que o método apresenta o resultado melhor quando o k é 5. Conclui-se que quando k é 5 os *clusters* obtidos apresentam uma maior similaridade com os *clusters* corretos.
Aplicando o método *fcluster* com o valor de dissemelhança máximo de 5 *clusters*, obtiveram-se os seguintes resultados:
O *cluster* com as participações $P_{2,10}$, $P_{2,15}$ e as participações do evento E_5;
O *cluster* com todas as participações do evento E_1 e a participação $P_{2,18}$;
O *cluster* constituído pela maioria das participações do evento E_2;
Um *cluster* com as participações associadas ao evento E_4 e outro com as do evento E_3.
Assim, todas as participações foram bem associadas à exceção das participações do evento E_2 que se associaram a 3 *clusters* diferentes.

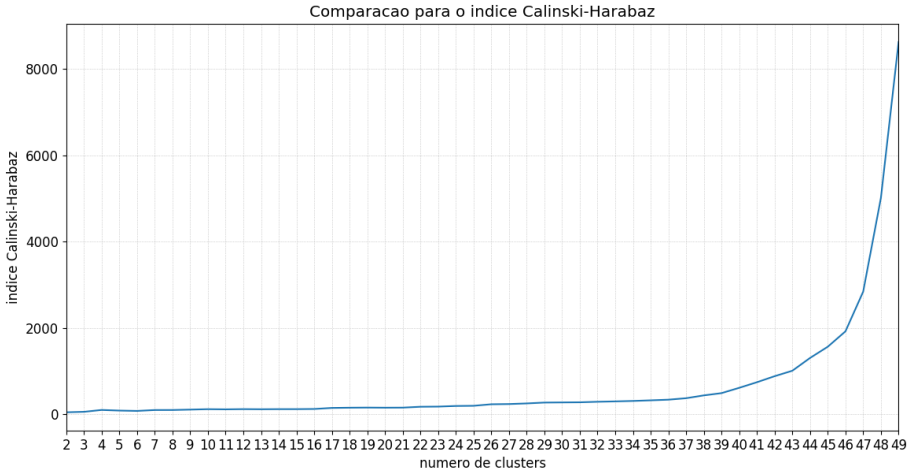
Cenário 4



(a) Variação do índice de Silhouette



(b) Variação do índice de Dunn



(c) Variação do índice de Calinski-Harabasz

Figura 5.20: Resultados dos índices obtidos para o cenário 4

Por análise do gráfico 5.20a pode verificar-se um valor mais próximo de 1 para o índice correspondente a 4 *clusters*. Quanto à análise do gráfico 5.20b verifica-se que o valor de índice é um máximo local para 4 *clusters*. Por análise do gráfico 5.20c verifica-se que o valor do índice aumenta à medida que o número de *clusters* aumenta, verificando-se um máximo local para 4 *clusters*.

Comparação dos índices no cenário 4

nº clusters	**Silhoutte**	**Dunn**	**Calinski-Harabasz**
n=4	**0.577**	**1.620**	**98.139**
n=5	0.543	1.155	83.584
n=6	0.489	0.919	75.610
n=7	0.493	1.364	96.466
n=8	0.502	0.947	97.190

Tabela 5.7: Apresentação dos índices dos três métodos para o cenário 4

Através dos resultados da Tabela 5.7, pode verificar-se que os três métodos apresentam melhores resultados no caso em que o número de *clusters* é 4.

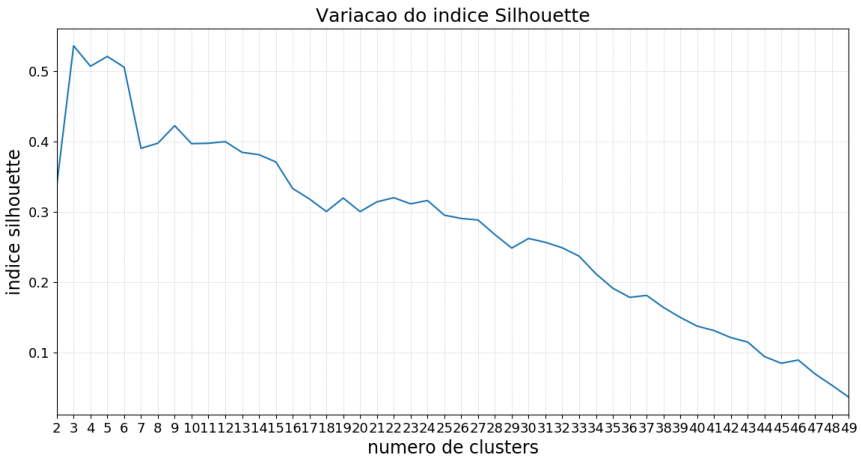
Índice *Fowlkes-Mallows*

$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
0.810	0.788	0.801	**0.824**	0.770

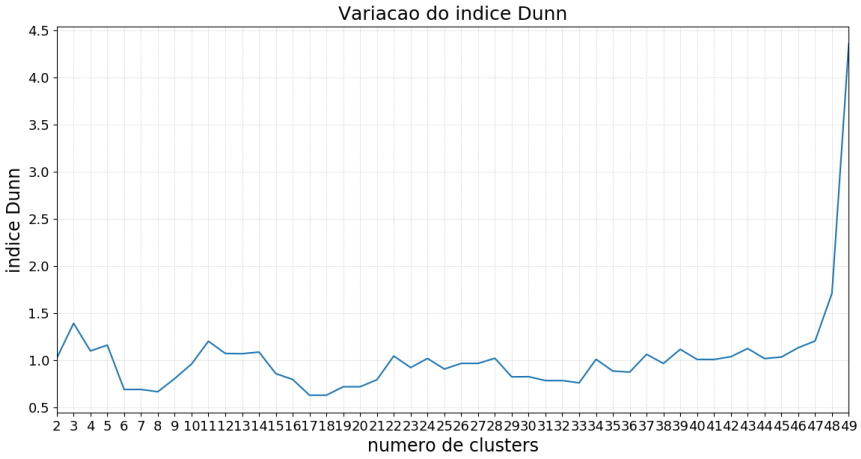
Tabela 5.8: Apresentação do índice Fowlkes-Mallows para o cenário 4

Pode verificar-se pela Tabela 5.8 que o método apresenta melhor resultado quando o k é 7, isto é, os *clusters* obtidos apresentam uma maior similaridade com os *clusters* corretos.
Aplicando o método *fcluster* com o valor de dissemelhança máximo de 5 *clusters*, obtiveram-se os seguintes resultados:
O *cluster* com todas as participações do evento E_2 e a maioria das participações associadas ao evento E_1;
O *cluster* constituído pela maioria das participações associadas ao evento E_5;
O *cluster* constituído pela participação $P_{1,9}$;
O *cluster* com todas as participações do evento E_3;
O *cluster* com todas as participações do evento E_4 e pela participação $P_{5,8}$.
Embora o número de *clusters* sugerido pelos critérios seja 4 ou 7, observa-se que no caso de 5 *clusters* quase todas as participações são associadas ao *cluster* correto à exceção das participações dos eventos E_1 e E_5 que se associaram a *clusters* diferentes.

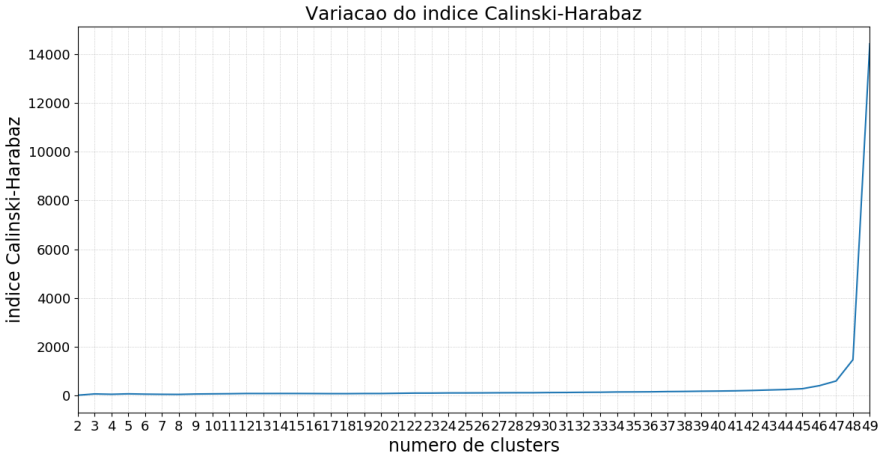
Cenário 5



(a) Variação do índice de Silhouette



(b) Variação do índice de Dunn



(c) Variação do índice de Calinski-Harabasz

Figura 5.21: Resultados dos índices obtidos para o cenário 5

Por análise do gráfico 5.21a, pode verificar-se que é obtido um valor mais próximo de 1 para o índice correspondente a 3 *clusters*. Quanto à análise do gráfico 5.21b, verifica-se que o valor de índice é um máximo local para 3 *clusters*. Por análise do gráfico 5.21c verifica-se que, em geral, o valor do índice aumenta à medida que o número de *clusters* aumenta.

Comparação dos índices no cenário 5

nº clusters	**Silhoutte**	**Dunn**	**Calinski-Harabasz**
n=2	0.338	1.020	16.253
n=3	**0.537**	**1.391**	68.994
n=4	0.507	1.097	55.176
n=5	0.521	1.159	**72.033**
n=6	0.506	0.688	60.471
n=7	0.391	0.688	53.608
n=8	0.398	0.644	50.082

Tabela 5.9: Apresentação dos índices dos três métodos para o cenário 5

Dados os resultados gráficos, nesta tabela houve a necessidade de analisar um maior intervalo para o número de *clusters*. Através dos resultados apresentados na Tabela 5.9 pode verificar-se que os métodos *Silhouette* e *Dunn* apresentam melhores resultados para 3 *clusters*. O método *Calinski-Harabasz* apresenta um melhor resultado no caso em que o número de *clusters* é 5.

Índice *Fowlkes-Mallows*

$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
0.492	0.690	0.646	**0.718**	0.713	0.700	0.679

Tabela 5.10: Apresentação do índice Fowlkes-Mallows para o cenário 5

Pode verificar-se pela Tabela 5.10 que o método apresenta melhor resultado quando o k é 5. Conclui-se assim que para 5 *clusters* os resultados obtidos apresentam uma maior similaridade com os *clusters* corretos.

Aplicando o método *fcluster* da biblioteca **Scipy** e utilizando o valor de dissemelhança máximo correspondente a 5 *clusters*, obtiveram-se os seguintes resultados:
O *cluster* constituído por grande parte das participações associadas ao evento E_3;
O *cluster* constituído por algumas das participações associadas ao evento E_3

sendo estas as participações $P_{3,1}$, $P_{3,3}$, $P_{3,4}$;
O *cluster* constituído por todas as participações associadas aos eventos E_2 e E_5;
O *cluster* constituído pela participação $P_{3,10}$ e grande parte das participações associadas aos eventos E_2;
O *cluster* constituído pela participação $P_{4,1}$ e por todas as participações associadas aos eventos E_1.
Assim, são bastantes as participações que foram associadas ao *cluster* errado.

Através dos resultados obtidos pelo cálculos dos índices de validação interna de um *cluster* pode-se concluir que o número "ideal"de *clusters* de participações se encontra entre 3 e 5. Em todo este estudo o resultado mais frequente é o de k igual a 5 *clusters*.

No que diz respeito aos resultados obtidos pelo cálculo do índice de validação externa de um *cluster* conclui-se que é obtida uma maior precisão nos resultados para 5 *clusters* de participações, havendo apenas um resultado que se apresenta melhor num caso de 7 *clusters* (cenário 4).

Tendo em conta os resultados dos índices de validação internos e externos pode verificar-se que com exceção de um caso, os resultados do índice de validação interno *Calinski-Harabasz* e os resultados do índice de validação externo *Fowlkes-Mallows* indicam as mesmas conclusões, isto é, um melhor resultado quando k é 5. Logo, pode dizer-se que o índice de validação interno *Calinski-Harabasz* apresenta uma maior fiabilidade nos resultados.

Capítulo 6

Conclusão

Para a elaboração do trabalho apresentado neste relatório, que tinha como objetivo identificar participações semelhantes na plataforma *JuntarAJunta*, foi aplicado o método de *clustering* hierárquico. Na aplicação deste método foram utilizados os dados latitude e longitude, que estão associados ao problema reportado.
Através da aplicação deste método e utilizando as coordenadas geográficas, pode concluir-se que o mesmo apresenta excelentes resultados nos casos em que os *clusters* estão bem definidos. Contudo, o método revela situações mal identificadas nos casos em que os *clusters* estão mais próximos, isto é, nos casos em que as participações estão geograficamente mais espalhadas e os *clusters* se sobrepõem.

Relativamente aos resultados apresentados pelos índices de validação interna e externa de *clusters* pode concluir-se que mesmo quando se identificam 5 *clusters*, que corresponde ao número correto, a constituição dos *clusters* nem sempre é a correta. À medida que os dados dos *clusters* ficam mais dispersos são encontrados resultados menos precisos, como se verificou a partir dos resultados do índice *Fowlkes-Mallows*.

O método de *clustering* hierárquico mostrou ser adequado ao estudo realizado, permitindo tirar conclusões úteis acerca da sua capacidade em identificar participações semelhantes. Além disso, permitiu concluir que, nos cenários onde não há grupos bem coesos e separados, a localização por si só pode favorecer a associação errada de participações. Logo, a posição geográfica, nestes casos, pode ser insuficiente para distinguir as participações no espaço e por isso, insuficiente para determinar um valor de dissemelhança máximo para a identificação de *clusters*. Assim, dado que as participações podem estar distantes do objeto fotografado e uma vez que as fotografias estão geralmente direcionadas para o centro do objeto fotografado, este trabalho tem uma segunda abordagem que tem por fundamento o método para identificação de participações semelhantes

descrito em [10]. O objetivo é obter informação relevante de forma a distinguir as participações no espaço, determinar um limite máximo para valor da dissemelhança e, consequentemente, obter resultados mais precisos. Para tal, foram consideradas duas novas variáveis, a orientação geográfica da fotografia e a distância entre o local onde o utilizador se encontra e o objeto a ser fotografado. Esta abordagem é apresentada em [13].

Ao nível das sugestões de melhoria da plataforma *JuntarAJunta* seria interessante e importante identificar participações falsas, ou seja participações que não correspondem ao relato de um problema ou apresentação de uma sugestão. Seria uma mais-valia conseguir identificá-las, uma vez que estas interferem na gestão e análise da informação contida na base de dados da aplicação. Uma outra sugestão de melhoria relaciona-se com a informação da categoria associada às intervenções reportadas na plataforma *JuntarAJunta*. A informação da categoria pode ajudar a identificar semelhanças nas participações mas podem existir casos em que duas participações são semelhantes e foram registadas com categorias diferentes. De forma a colmatar esta situação, e no caso de se trabalhar com dados reais, seria necessário construir grupos de categorias e integrar essa semelhança no algoritmo, de forma a melhorar o processo de identificação de participações semelhantes. Depois de uma análise detalhada das categorias atualmente consideradas na plataforma *JuntarAJunta*, concluiu-se que em alguns casos estas se sobrepõem. Propunha-se assim que as 22 categorias fossem agrupadas em 9 categorias, que passariam a ser: Espaços verdes; Estradas, passeios e caminhos; Higiene urbana; Iluminação; Sinalética; Estrutura perigosa; Outros; Animais e Serviços públicos.

No decorrer deste estudo houve alguma reflexão no sentido de automatizar a análise da variação do valor da dissemelhança, que exprime a relação entre o número de *clusters* e a dissemelhança conforme as Figuras 5.4, 5.7, 5.10, 5.13 e 5.16. Uma ideia que surgiu desta reflexão foi a de procurar aproximar esta variação por uma função potência de expoente negativo. O aprofundamento desta ideia requer contudo um estudo mais detalhado de métodos para encontrar o ponto de curvatura máximo, que representaria o número "ideal"de *clusters*.

A elaboração deste estágio permitiu trabalhar tecnologias informáticas transversais relacionadas com as construção de páginas *web*, *dashboards*, *formulários* e aplicar técnicas de *machine learning*, mais precisamente técnicas de *clustering*. Foi possível perceber a utilidade que as técnicas de *clustering* podem trazer como recurso para ajudar a gerir dados numa aplicação. Além disso, a experiência de estágio numa empresa da área de informática permitiu perceber como este tipo de problemas é abordado pelas pequenas empresas e que tipo de lacunas existe nas áreas de *data mining* e *machine learning*.

Bibliografia

[1] Big data. https://www.linknacional.com.br/12/big-data-volume-de-dados/. Acessado a 27/10/2017.

[2] Big data analytics: você sabe o que é? http://www.bigdatabusiness.com.br/voce-sabe-o-que-e-big-data-analytics/. Acessado a 27/10/2017.

[3] Conceitos fundamentais de machine learning. http://www.cienciaedados.com/conceitos-fundamentais-de-machine-learning/. Acessado a 27/10/2017.

[4] Machine learning. https://www.sas.com/pt_br/insights/analytics/machine-learning.html. Acessado a 27/10/2017.

[5] Tipos de data mining. https://paginas.fe.up.pt/~mgi99021/it/tipos.html. Acessado a 27/10/2017.

[6] Uma visão geral declusterização de dados. ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia368_02/topico5_02.pdf. Acessado a 27/10/2017.

[7] Cluster validation statistics: Must know methods. http://www.sthda.com/english/articles/29-cluster-validation-essentials/97-cluster-validation-statistics-must-know-methods/, 2017. Acessado a 23/10/2017.

[8] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1):243–256, 2013.

[9] B. S. Everitt, S. Landau, M. Leese, and D. Stahl. Cluster analysis. Wiley Online Library, 5th edition, 2011.

[10] Y. A. Lacerda, R. G. F. Feitosa, G. A. R. M. Esmeraldo, C. d. S. Baptista, and L. B. Marinho. Compass clustering: A new clustering method for

detection of points of interest using personal collections of georeferenced and oriented photographs. In Proceedings of the 18th Brazilian Symposium on Multimedia and the Web, WebMedia '12, pages 281–288, New York, NY, USA, 2012. ACM.

[11] D. D. R. Meneghetti. dunn-sklearn. `https://gist.github.com/douglasrizzo/cd7e792ff3a2dcaf27f6`, 2016.

[12] G. W. Milligan. Clustering validation: results and implications for applied analyses. In Clustering and classification, pages 341–375. World Scientific, 1996.

[13] C. Sousa. Modelação com técnicas de *clustering* de participações geoespaciais de cidadãos. Relatório de Estágio (Submetido), Universidade do Minho, 2017.

[14] S. Theodoridis and K. Koutroumbas. Pattern Recognition. Elsevier, 2nd edition, 2003.

Anexo A

Código em Python

A.1 Criação de cenários

criacaodasPerturbacoes.py

```
# -*- coding: utf-8 -*-
import numpy as np


# Criação das Perturbações
def criarPerturbacoes(eventos, lista_sigma,n):
    # Lista de Perturbações
    Pontos=[]

    for i in range(len(eventos)):
        sigma = lista_sigma[i]
        e = eventos[i]

        p_lat = np.random.normal(e[0], sigma,n)
        p_long = np.random.normal(e[1],sigma,n)
        lat_long = [list(a) for a in zip(p_lat, p_long)]
        Pontos = Pontos + lat_long

    return Pontos

# Participações
eventos=[[41.550097,-8.429125],[41.550789,-8.429578],[41.54985,-8.43085],
[41.5490,-8.429580],[41.5513,-8.4302]]

# Número de perturbações
n=10
```

A.2 Análise de cenários

AnalisedosCenarios.py

```
# -*- coding: utf-8 -*-
from matplotlib import pyplot as plt
```

```
from matplotlib.ticker import MaxNLocator
import numpy as np
from matplotlib import style,axis
from scipy.cluster.hierarchy import dendrogram, linkage,
fcluster,fclusterdata
from scipy.cluster.hierarchy import cophenet
from scipy.spatial.distance import pdist
from scipy.spatial.distance import squareform
from scipy import interpolate, optimize
from math import sqrt
from criacaoPerturbacoes import criarPerturbacoes_anisotropicas
from indicadores import indicadores
from sklearn.metrics import silhouette_samples, silhouette_score,
calinski_harabaz_score, fowlkes_mallows_score
from dynamicTreeCut import cutreeHybrid
from dunn_sklearn import dunn, min_cluster_distances, diameter

eventos=[[41.550097,-8.429125],[41.550789,-8.429578],[41.54985,-8.43085],
[41.5490,-8.429580],[41.5513,-8.4302]]

# Número de perturbações
n=10

#Maria
cenario1 = [[41.550134337954724, -8.4290016222474264], [41.550085334602308,
-8.4292839272343336],
[41.54996436736478, -8.4294667719548375],[41.550110903395193,
-8.4292147556555168],
[41.550215804642441, -8.4292473939805834], [41.550214918827749,
-8.4292382498400915],
[41.549904895921649, -8.4291245507714425],[41.550125767878079,
-8.4289555636302769],
[41.55025708659187, -8.4291300087354966], [41.549990760900755,
-8.4293610011788509],
[41.55053254608638, -8.4294882006979854], [41.550727523584406,
-8.4296478345972279],
[41.550937252821932, -8.4296537087672778],[41.550883953294104,
-8.4295184114353479],
[41.550672643808774, -8.4295575167303074],[41.550688268336451,
-8.4293566881328239],
[41.550737259601256, -8.4296534819736824],[41.550754509496123,
-8.4294370804737806],
[41.550918982325214, -8.4296826362490851],[41.550752132715971,
-8.4296472214978202],
[41.549869557003689, -8.4306575785496403],[41.54975565288121,
-8.4307647358983822],
[41.549930747676591, -8.4307860148353662], [41.549870025060081,
-8.4308809625883629],
[41.549782981941654, -8.4308742893903901], [41.549749383915163,
-8.430590355509894],
[41.549771086082899, -8.4307228922511896], [41.549951419015521,
-8.4307671072160737],
[41.549847353073993, -8.4308646666682723], [41.549959416666717,
-8.4309847210974826],
[41.549181424365585, -8.4296524170540845], [41.548903042552965,
-8.429397518601716],
[41.549152504288038, -8.4294597374301237], [41.549133283621316,
-8.4297491035139238],
```

```
42 [41.548948990338751, -8.4295336136603449], [41.548853653751372,
   -8.4296747670868868],
43 [41.548894076042593, -8.4295346185472191], [41.54902503992971,
   -8.4295672384729716],
44 [41.548862579666711, -8.4297502920239893], [41.549072432059084,
   -8.429662676283872],
45 [41.551339827142485, -8.4301479770805816], [41.551331755300694,
   -8.4301437518967859],
46 [41.55143259235183, -8.4300536312095602], [41.551449216407192,
   -8.4303649955213515],
47 [41.551250786020759, -8.4303552520864908], [41.551187477529666,
   -8.430311106734024],
48 [41.551263653237797, -8.4303336024427153], [41.551331881956578,
   -8.4302119543239566],
49 [41.551227244843879, -8.4302450652937111], [41.551372614887043,
   -8.4301235630301097]]
50 cenario2 = [[41.550002783589512, -8.4287047123249383], [41.550108632704273,
   -8.4292267824117033],
51 [41.549850871833094, -8.4287697215203181], [41.550166175844353,
   -8.4294590171809727],
52 [41.550039297487885, -8.4290736682501439], [41.550003090374211,
   -8.4292205907325712],
53 [41.550414878418046, -8.4291543217133409], [41.549861759577034,
   -8.4290578416435444],
54 [41.550229955207683, -8.4290819530779366], [41.549971978117171,
   -8.4287522888781528],
55 [41.550654879741629, -8.4299279812768173], [41.551040787376301,
   -8.4294566503084134],
56 [41.550853990228831, -8.429769771007166], [41.55075968704999,
   -8.4294837929540698],
57 [41.550815966771665, -8.429431217943824], [41.550883535592838,
   -8.4295497814069993],
58 [41.550701837827319, -8.4298081540435437], [41.550901187016002,
   -8.4296157446278226],
59 [41.551080863953068, -8.4294845941959995], [41.550523146092559,
   -8.429501308533327],
60 [41.550227529455704, -8.430445215573684], [41.549883056797938,
   -8.4308986352557724],
61 [41.549903795761303, -8.4309860716280163], [41.550142492442703,
   -8.4310163380141141],
62 [41.549619206211695, -8.4307322300931435], [41.549903053166886,
   -8.4308564136038786],
63 [41.549943290404109, -8.430917506016236], [41.549537109928075,
   -8.4311572305006628],
64 [41.55019643244939, -8.4309474293905176], [41.549726920813853,
   -8.4308082854953383],
65 [41.548950383176141, -8.4296889986656574], [41.54900239259554,
   -8.4295836501341679],
66 [41.549097040699763, -8.4295561029847157], [41.548899994666826,
   -8.4297508863657526],
67 [41.549299124231908, -8.4298458140818209], [41.548679056009057,
   -8.4296817686662155],
68 [41.549343203265735, -8.4294652558896814], [41.549319288808825,
   -8.4294954486199298],
69 [41.548803131215749, -8.4295452106156308], [41.549132759801651,
   -8.429402334241491],
70 [41.551477132596396, -8.4304750950015759], [41.551473797353665,
   -8.4300433082028796],
```

```
71 [41.551280198616702, -8.4305182059183501], [41.55107586219534,
   -8.4303392338169942],
72 [41.55159960089491, -8.4301482138377555], [41.551562663769722,
   -8.4300369393364178],
73 [41.551198194768425, -8.4302660798576259], [41.550974138250069,
   -8.4300087887697881],
74 [41.551371095755457, -8.4303774086704042], [41.551216973033462,
   -8.4300609114463416]]
75 cenario3 = [[41.550193026896636, -8.4289964505001063], [41.550314484058546,
   -8.4293012855085507],
76 [41.549917651037653, -8.4290642938507698], [41.549653833576841,
   -8.4289804681598497],
77 [41.550098279162796, -8.4292414971872649], [41.549930534961781,
   -8.4289152360111324],
78 [41.549817176790015, -8.4293249526200587], [41.5501920737291,
   -8.428950935527487],
79 [41.550144995060734, -8.4293843748811526], [41.549928180564841,
   -8.4292638782223648],
80 [41.551223892687915, -8.4295960232010767], [41.550819071431441,
   -8.4290078641986845],
81 [41.550675745502858, -8.4293569212475585], [41.550714589814781,
   -8.4296036301433759],
82 [41.55082816095296, -8.4295302116114943], [41.551186677719983,
   -8.4295321674919848],
83 [41.550799085980117, -8.4293830670538039], [41.550720099337703,
   -8.4294297809592518],
84 [41.550297693497455, -8.4291175694019245], [41.550621578656262,
   -8.4299481664277192],
85 [41.549579816380451, -8.4310210633625751], [41.550101150431452,
   -8.4308403609995253],
86 [41.550131072925097, -8.4307006321609919], [41.549689540294189,
   -8.4312724385285378],
87 [41.549601274184212, -8.4308187228242133], [41.549985686033985,
   -8.430878437179123],
88 [41.549370734957819, -8.4305059221252545], [41.549679540273068,
   -8.430668040239933],
89 [41.549786858690794, -8.4307348444142001], [41.549563883845238,
   -8.4308284480078832],
90 [41.549143992162591, -8.4296828726638804], [41.549104201017862,
   -8.4293363145064575],
91 [41.549055336627653, -8.4296960405034635], [41.549038950249951,
   -8.429770403387554],
92 [41.54886861653651, -8.4300594325698803], [41.54905649167705,
   -8.4296959073867388],
93 [41.548794557039706, -8.4295335627935675], [41.54863016521746,
   -8.4293359113759045],
94 [41.549326813029857, -8.4298271350092815], [41.549103419333015,
   -8.4294020563069072],
95 [41.551495993889532, -8.4304842473145 58], [41.551272074075541,
   -8.4299257167319599],
96 [41.551236118374263, -8.4304754419780163], [41.551411531168718,
   -8.4302843156238652],
97 [41.551248133182945, -8.4299860477587547], [41.551155145344175,
   -8.4301885046565186],
98 [41.551165598860571, -8.4303173058768675], [41.55118736594919,
   -8.4299178094964162],
99 [41.55117497046583, -8.4303017174809458], [41.551299285625284,
   -8.4302661983187352]]
```

```
100 cenario4 = [[41.550286893766717, -8.4290349571677989], [41.550479002130913,
    -8.4294081121234683],
101 [41.550578213508729, -8.429162681886293], [41.54996717852157,
    -8.4294846382251656],
102 [41.550092823598092, -8.4293870041454539], [41.550209713568002,
    -8.4292314868846869],
103 [41.549889527092354, -8.4295875126500608], [41.549926924875855,
    -8.4293175667877982],
104 [41.550198200769977, -8.4295030852861075], [41.549500257734891,
    -8.4287823837388505],
105 [41.550696222282241, -8.4297804514212125], [41.551102548021404,
    -8.4294278441180523],
106 [41.550774710616629, -8.4297820127611018], [41.550964519698248,
    -8.4292565111471589],
107 [41.550933007671603, -8.4297657505206018], [41.550594230658554,
    -8.4291596742974111],
108 [41.550521503978302, -8.4295161492926525], [41.550510129054047,
    -8.4293641299335924],
109 [41.55105652242549, -8.42958880804877], [41.55029665267832,
    -8.4297370528730049],
110 [41.549994588782752, -8.4306195451554728], [41.549865334087784,
    -8.4306174418823954],
111 [41.549729356204267, -8.430893360408124], [41.54970836953683,
    -8.4309733573571997],
112 [41.549803286726636, -8.4305454413322956], [41.549660719961537,
    -8.4308548494541427],
113 [41.549661662909287, -8.4309054868260649], [41.549786248807976,
    -8.4312200802129507],
114 [41.549743364644094, -8.430703518620474], [41.549581148148469,
    -8.430843716769175],
115 [41.548821122769553, -8.429479703292488], [41.548717714328426,
    -8.4294937856392664],
116 [41.548879583961977, -8.4296454792119597], [41.549117864540975,
    -8.4293308705735654],
117 [41.549052830408861, -8.4297967408659975], [41.549469309601093,
    -8.4293372577847503],
118 [41.548692850816998, -8.4295249816437376], [41.548841963495391,
    -8.4296740133284622],
119 [41.549195949386061, -8.4294696443431398], [41.548849195709487,
    -8.429698832743604],
120 [41.551503108242969, -8.4299541560616902], [41.551249954177919,
    -8.4299451227347415],
121 [41.551259737845598, -8.4303098355607808], [41.551579144650525,
    -8.4303552214090995],
122 [41.550998488763156, -8.4304313765702439], [41.550821742880565,
    -8.4302517661297891],
123 [41.551465297277204, -8.4301013961125513], [41.552188058560368,
    -8.4305659791199261],
124 [41.551601253259356, -8.4302149539105073], [41.550851994421564,
    -8.4301043733263228]]
125 cenario5 = [[41.549774293288522, -8.4297126461034502], [41.549725190844491,
    -8.429172005181222],
126 [41.549978235754729, -8.4290983849352621], [41.549967286506529,
    -8.4291581122160508],
127 [41.550222782827852, -8.4293374064349731], [41.550279566700212,
    -8.4290460815920145],
128 [41.550245167401791, -8.428801131117627], [41.550047684960532,
    -8.4292019605774655],
```

```
129 [41.54961672883384, -8.4293990752195924], [41.549790178026363,
    -8.4294187204162299],
130 [41.550862298523299, -8.4295517146409331], [41.55119188963203,
    -8.4300106184180503],
131 [41.550970192164968, -8.4297619151106957], [41.550659670597035,
    -8.4297118446362891],
132 [41.550511201273906, -8.4295782458065833], [41.55102216251678,
    -8.4294094038644491],
133 [41.551090961564796, -8.4301765266160942], [41.551352641570745,
    -8.4295966771918227],
134 [41.550781012278748, -8.4297924775059876], [41.550771265307269,
    -8.4300387745995575],
135 [41.550678183973687, -8.4318584922849684], [41.549978233833116,
    -8.4305985228421765],
136 [41.549951262035428, -8.4317596000206496], [41.549620176128165,
    -8.431721117133355],
137 [41.550218801223068, -8.4308858416163037], [41.549605662077397,
    -8.4306435931980346],
138 [41.549527301941502, -8.4310611980688872], [41.549131404765525,
    -8.430636943902666],
139 [41.549169154328595, -8.4308503666186549], [41.549240904168094,
    -8.4294901947428595],
140 [41.549573849581954, -8.4298915471710991], [41.54900120334316,
    -8.4291172065638076],
141 [41.548991708297393, -8.4295747569439889], [41.548883983183508,
    -8.4296659230617568],
142 [41.548739992285874, -8.429782053681814], [41.548776358691406,
    -8.429345808534034],
143 [41.548897914855125,-8.4296657412670797], [41.548964400140754,
    -8.4298184793613107],
144 [41.548950887026884, -8.4289779854822697], [41.548702084050767,
    -8.4291660427361617],
145 [41.551503431429488, -8.4299306528484639], [41.551576124180556,
    -8.4302380367687029],
146 [41.550995670507689, -8.4307555337686537], [41.550908454026953,
    -8.4299239116880837],
147 [41.551652164601428, -8.4306280383099814], [41.551355600533732,
    -8.430039555019798],
148 [41.550995711926276, -8.4301355164113954], [41.551745646660315,
    -8.4300804403941498],
149 [41.551061971823373, -8.4303260072375004], [41.551581583179583,
    -8.4304101527200359]]
150
151
152 cenario1_sigma = [0.0001, 0.0001, 0.0001, 0.0001, 0.0001]
153 cenario2_sigma = [0.00019, 0.00019, 0.00019, 0.00019, 0.00019]
154 cenario3_sigma = [0.00019, 0.0003, 0.00019, 0.00019, 0.00019]
155 cenario4_sigma = [0.0003, 0.00019, 0.00019, 0.00019, 0.0003]
156 cenario5_sigma = [0.0003, 0.0003, 0.0005, 0.0003, 0.0003]
157 cenario_nome = ['P1'+str(i+1) for i in range(0,10)] + ['P2'+str(i+1) for i
    in range(0,10)] +
158 ['P3'+str(i+1) for i in range(0,10)] + ['P4'+str(i+1) for i in range(0,10)]
    +
159 ['P5'+str(i+1) for i in range(0,10)]
160
161 # Representação gráfica do cenário
162 def representacaoCenario(eventos, perturbacoes, sigma):
163     for i in range(len(eventos)):
164         e = eventos[i]
```

```
165         plt.plot(e[0],e[1],marker='*', markersize=10, label= 'E'+str(i+1) )
166         if sigma!=[]:
167       circle = plt.Circle((e[0], e[1]), sigma[i], color='b', fill=False)
168       circle1 = plt.Circle((e[0], e[1]), 2*sigma[i], color='g', fill=False)
169       ax = plt.gca()
170       ax.add_artist(circle)
171       ax.add_artist(circle1)
172 
173     for p in perturbacoes:
174        plt.scatter(p[0],p[1], c='#ccccb3', marker='o')
175 
176     plt.xlabel('latitude')
177     plt.ylabel('longitude')
178     plt.gca().set_aspect('equal', adjustable='box')
179     plt.legend()
180     plt.show()
181 
182     return True
183 
184 #Escolha do cenário
185 Pontos = cenario1
186 nomecenario = 'cenario1'
187 sigma = cenario1_sigma
188 print 'representacao do cenario '+str(nomecenario)+' com o sigma =
    '+str(sigma)
189 matrizDistancias=pdist(Pontos)
190 representacaoCenario(eventos, Pontos, sigma)
191 
192 
193 # Dendrogramas
194 
195 # Definição da métrica
196 metrica ='euclidean'
197 
198 # Criação da matriz das distâncias tendo em conta o criterio de ligação
    escolhido
199 Z_single = linkage(Pontos, 'single',  metric=metrica)
200 Z_complete = linkage(Pontos, 'complete', metric=metrica)
201 Z_average = linkage(Pontos, 'average', metric=metrica)
202 
203 
204 plt.title('Clustering Hierarquico - criterio single',fontsize=16)
205 plt.xlabel('Participacoes',fontsize=15)
206 plt.ylabel('Dissemelhanca',fontsize=15)
207 
208 d_single = dendrogram(
209     Z_single,
210     labels=cenario_nome,
211     leaf_font_size=13,
212 )
213 
214 plt.show()
215 
216 c_single, coph_dists_single = cophenet(Z_single, matrizDistancias)
217 print 'Coeficiente cofenetico (single): '+str(c_single)
218 
219 
220 plt.title('Clustering Hierarquico - criterio complete',fontsize=16)
221 plt.xlabel('Participacoes',fontsize=15)
```

```
222 plt.ylabel('Dissemelhanca',fontsize=15)
223
224 d_complete = dendrogram(
225     Z_complete,
226     labels=cenario_nome,
227     leaf_font_size=13.,
228 )
229
230 plt.show()
231
232 c_complete, coph_dists_complete = cophenet(Z_complete, matrizDistancias)
233 print 'Coeficiente cofenetico (complete): '+str(c_complete)
234 #plt.subplot(1,3,3)
235
236
237 plt.title('Clustering Hierarquico - criterio average',fontsize=16)
238 plt.xlabel('Participacoes',fontsize=15)
239 plt.ylabel('Dissemelhanca',fontsize=15)
240
241 d_average = dendrogram(
242     Z_average,
243     labels=cenario_nome,
244     leaf_font_size=13.,
245 )
246
247 plt.show()
248
249 c_average, coph_dists_average = cophenet(Z_average, matrizDistancias)
250 print 'Coeficiente cofenetico (average): '+str(c_average)
251
252
253 # Função representativa do comportamento do clustering hierárquico
254 alturas_average = [x[1] for x in d_average['dcoord']]
255 alturas_complete = [x[1] for x in d_complete['dcoord']]
256 alturas_single = [x[1] for x in d_single['dcoord']]
257
258 y_average = sorted(alturas_average, reverse=True)
259 y_complete = sorted(alturas_complete, reverse=True)
260 y_single = sorted(alturas_single, reverse=True)
261
262 x =range(1, len(y_average)+1)
263
264
265 # Representação gráfica
266 plt.plot(x,y_average, marker='o', linewidth=0.5, markersize=2,
    label='Average')
267 plt.plot(x,y_single,marker='o', linewidth=0.5, markersize=2, label='Single')
268 plt.plot(x,y_complete, marker='o', linewidth=0.5, markersize=2,
    label='Complete')
269 plt.xlabel('Numero de Clusters')
270 plt.ylabel('Dissemelhanca')
271 plt.legend()
272 plt.show()
273
274
275 #Analise dos clusters
276 nclusters = 5
277
278 resultados = []
```

```
279 lista_silhouette =[]
280 lista_dunn =[]
281 lista_CH =[]
282 x = []
283
284 for index in range(1,len(y_average)):
285   d = y_average[index]
286   r = fcluster(Z_average, d, criterion='distance')
287   resultados.append(r)
288   numero_clusters = len(set(r))
289   x.append(numero_clusters)
290   #print 'Resultado para ' + str(numero_clusters) + ' com um valor maximo de
291   #dissemelhanca igual a '+str(d)
292   #print r
293   s = silhouette_score(squareform(matrizDistancias), r,
      metric="precomputed")
294   lista_silhouette.append(s)
295   #print 'Indice Silhouette: ' + str(s)
296   d = dunn(r, squareform(matrizDistancias))
297   lista_dunn.append(d)
298   #print 'Indice Dunn: ' + str(d)
299   ch = calinski_harabaz_score(Pontos, r)
300   lista_CH.append(ch)
301   #print 'Indice Calinski-Harabaz: ' + str(ch)
302
303
304 plt.rcParams.update({'font.size':13})
305 plt.plot( x, lista_silhouette)
306 plt.title("Variacao do indice Silhouette",fontsize=18)
307 plt.grid(linestyle=':', lw=0.5)
308 plt.xlabel("numero de clusters",fontsize=17)
309 plt.ylabel("indice silhouette",fontsize=17)
310 plt.gca().set_xlim([2,len(x)+1])
311 plt.xticks(range(2,len(x)+2))
312 plt.show()
313
314 plt.plot(x, lista_dunn)
315 plt.title("Variacao do indice Dunn",fontsize=18)
316 plt.grid(linestyle=':', lw=0.5)
317 plt.xlabel("numero de clusters",fontsize=17)
318 plt.ylabel("indice Dunn",fontsize=17)
319 plt.gca().set_xlim([2,len(x)+1])
320 plt.xticks(range(2,len(x)+2))
321 plt.show()
322
323 plt.plot(x, lista_CH)
324 plt.title("Variacao do indice Calinski-Harabaz",fontsize=18)
325 plt.grid(linestyle=':', lw=0.5)
326 plt.xlabel("numero de clusters",fontsize=17)
327 plt.ylabel("indice Calinski-Harabaz",fontsize=17)
328 plt.gca().set_xlim([2,len(x)+1])
329 plt.xticks(range(2,len(x)+2))
330 plt.show()
331
332 print 'silhouette'
333 print lista_silhouette[0]
334 print lista_silhouette[1]
335 print lista_silhouette[2]
```

```
print lista_silhouette[3]
print lista_silhouette[4]
print lista_silhouette[5]
print lista_silhouette[6]


print 'dunn'
print lista_dunn[0]
print lista_dunn[1]
print lista_dunn[2]
print lista_dunn[3]
print lista_dunn[4]
print lista_dunn[5]
print lista_dunn[6]


print 'CH'
print lista_CH[0]
print lista_CH[1]
print lista_CH[2]
print lista_CH[3]
print lista_CH[4]
print lista_CH[5]
print lista_CH[6]


print 'participacoes'
print resultados[0]
print resultados[1]
print resultados[2]
print resultados[3]
print resultados[4]
print resultados[5]
print resultados[6]


#label true
labeltrue=[1 for i in range(0,10)]+[2 for i in range(0,10)]+[3 for i in
range(0,10)]+
[4 for i in range(0,10)]+[5 for i in range(0,10)]
print fowlkes_mallows_score(labeltrue,resultados[0])
print fowlkes_mallows_score(labeltrue,resultados[1])
print fowlkes_mallows_score(labeltrue,resultados[2])
print fowlkes_mallows_score(labeltrue,resultados[3])
print fowlkes_mallows_score(labeltrue,resultados[4])
print fowlkes_mallows_score(labeltrue,resultados[5])
print fowlkes_mallows_score(labeltrue,resultados[6])
```