Universidade do Minho
Escola de Ciências

Adriana Loureiro Vilas Boas Lopes

Classificação de clientes com aprendizagem automática

março de 2021

Classificação de clientes com aprendizagem automática

Adriana Lopes

UMinho | 2021

Universidade do Minho
Escola de Ciências

Adriana Loureiro Vilas Boas Lopes

Classificação de clientes com aprendizagem automática

Dissertação de Mestrado
em Matemática e Computação

Trabalho efetuado sob a orientação do
Professor Doutor Stéphane Louis Clain
e do
Professor Doutor Gaspar José Machado

março de 2021

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Agradecimentos

Embora este seja um trabalho individual feito por mim, por trás houve várias pessoas envolvidas que tornaram possível este feito. Esta dissertação acaba por ser fruto de um grande trabalho de equipa.

Em primeiro lugar queria agradecer aos meus dois orientadores, Doutor Stéphane Louis Clain e Doutor Gaspar José Machado, por todo o apoio e ajuda que me deram, pela grande disponibilidade que sempre tiveram e por todo o incentivo e partilha de conhecimento.

Quero também agradecer à empresa *Yarilabs* por tornar possível este projeto, ao seu representante Emanuel Mota e ao meu tutor da empresa Rui Fonseca que também estiveram sempre disponíveis para me ajudar e me proporcionaram as condições para a realização deste trabalho.

Por fim, mas não menos importante, quero agradecer à minha família, ao meu namorado e aos meus amigos que me foram acompanhando durante o meu percurso académico.

Declaração de Integridade

Declaro ter actuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

As empresas querem obter informações sobre os seus clientes através das suas bases de dados que sejam vantajosas para o negócio. Um processo de o fazer é recorrer a técnicas de inteligência artificial.

Neste trabalho o foco é a utilização de técnica de *clustering*, um método de aprendizagem automática não supervisionado, e a sua aplicação a uma base de dados específica, passando primeiro por bases de dados criadas artificialmente.

Foram propostas diversas métricas adequadas ao contexto da empresa de café *Pact Coffee* inicialmente avaliadas em bases de dados sintéticas simples, aplicando-se o *clustering* hierárquico.

Inseriu-se a questão temporal, ou seja, em vez de se aplicar o *clustering* para a base de dados inteira, primeiro dividiu-se esta por faixas de tempo e só depois se aplicou o *clustering* para cada uma. Depois criou-se uma continuidade entre as faixas, e identificaram-se trajetórias e clientes ao longo do tempo.

Os dados sintéticos começaram então a obter resultados favoráveis e por isso aplicou-se o mesmo método aos dados reais da empresa *Pact Coffee*.

Palavras chave: Aprendizagem automática, Métricas , Clustering Hierárquico, Cluster dinâmico, Streaming Data.

Abstract

Companies want to obtain information about their customers through their databases that are beneficial to the business. One solution is to use artificial intelligence, which is what I intend to do.

In this work the focus is on the use of clustering techniques, an unsupervised machine learning method, and its application to a specific database, passing first through artificially created databases.

Several metrics were proposed, appropriate to the context of the coffee company *Pact Coffee*, which were evaluated using simple synthetic databases, applying hierarchical clustering.

The time issue was inserted, that is, instead of applying the clustering to the entire database, it was first divided into time bands and only then was clustering applied to each one. Then there was a continuity between the bands, and trajectories and customers were identified over time.

The synthetic data then began to obtain favorable results and therefore the same method was applied to the real data of the company *Pact Coffee*.

Keywords: Machine learning, Metrics, Hierarchical clustering, Dinamic cluster, Streaming Data.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação do trabalho desenvolvido

A empresa Yarilabs desenvolve actividade de consultoria em informática e gestão de base de dados. Um dos desafios que a empresa encontra é elaborar ferramentas de análise de dados, nomeadamente a identificação do perfil dos consumidores e assim melhorar a competitividade das empresas que usam os seus serviços. Por exemplo, um dos seus clientes, que é uma empresa de venda *online* de café, apresentou o problema da classificação do comportamento dos clientes na compra de cápsulas de café. A gestão adequada do portfólio de clientes requer uma atitude pró-ativa da empresa que vende as cápsulas por forma a reter os seus clientes e aumentar as vendas. Para tal, fazer propostas ou ações de sensibilização muito dirigidas ao perfil de cada cliente para optimizar os resultados é um aspeto muito importante. A classificação dos comportamentos associados aos interesses do consumidor permite estabelecer um conjunto de ações com alvos perfeitamente identificados. A construção de ferramentas capazes de identificar diferentes perfis de clientes com base na informação disponível torna-se, assim, fundamental para aumentar a competitividade das empresas.

Numa base de dados encontram-se conjuntos de arquivos que estão relacionados entre

si e contêm registos tanto de pessoas, lugares ou outras coisas. Estas bases de dados estão organizadas de forma a fazerem algum sentido, ou seja, de modo a serem úteis para algum estudo ou investigação. Uma das aplicações que se pode fazer numa base de dados é o *clustering*, que consiste em fazer agrupamentos automáticos dos dados pertencentes a uma base seguindo a ideia da maior semelhança entre os dados. Por exemplo, uma empresa de venda *online* de café pode utilizar o *clustering* com o objetivo de conhecer melhor os grupos dos seus clientes, para posteriormente fazer promoções, publicidade a certos produtos, etc.

Pretende-se assim neste trabalho elaborar ferramentas de identificação de perfis de consumidor usando a tecnologia de *Machine Learning* e desenvolver aplicações para os casos apresentados pela empresa Yarilabs.

1.2 A empresa Yarilabs

Este trabalho foi desenvolvido no contexto de uma dissertação em empresa, neste caso a empresa Yarilabs. Criada a 19 de abril de 2017, é uma *startup* em tecnologia, que atualmente aceita projetos em todo o mundo a partir do seu escritório português que se situa em Braga, onde está localizada a equipa operacional, com cerca de 14 funcionários.

A Yarilabs ajuda os seus clientes a desenvolver novos produtos e também empresas estabelecidas que precisam de ajuda de consultoria para lançar novos projetos ou melhorar produtos existentes.

Neste momento, a empresa foca-se essencialmente em linguagens de programação funcionais, sendo os seus principais serviços o *design* de novos produtos, desenvolvimento *Web* e *Mobile*, projeto e desenvolvimento rápido de protótipos, desenvolvimento de *back-end*, implementações de nuvem em escala e *DevOps*, e também treinamento.

1.3 Estrutura da dissertação

Esta dissertação divide-se em duas partes, uma mais teórica e outra dedicada à aplicação para os dados reais.

Em relação à parte teórica, introduz-se no Capítulo 2 a noção de métrica e no Capítulo 3 apresenta-se as noções fundamentais de *clustering*. O Capítulo 4 é dedicado à criação de uma primeira métrica, com destaque para a separação do atributo do tempo dos restantes. Apresenta-se no Capítulo 5 um conjunto de bases de dados que tem como objetivo avaliar a capacidade desta métrica proposta em particionar as bases de dados de uma maneira coerente em relação aos aspetos comerciais. Foram identificadas muitas fragilidades na métrica do Capítulo 4 para realizar uma boa partição e por consequência apresenta-se no Capítulo 6 uma nova tecnologia de *clustering*, onde a integração do atributo do tempo no procedimento de *clustering* passa por uma estrutura em faixas temporais. Neste mesmo capítulo apresenta-se uma série de bases de dados para validar esta nova proposta.

A tecnologia desenvolvida no Capítulo 6 é aplicada no Capítulo 7 às bases de dados provenientes da empresa *Pact Coffee*. Neste capítulo houve necessidade de introduzir uma métrica mais sofisticada por haver dados nominativos e aplicou-se a técnica de janela temporal.

Finalmente no Capítulo 8 apresentam-se as conclusões do trabalho desenvolvido.

Capítulo 2

Similaridade e Dissimilaridade

Os algoritmos de optimização que surgem na aprendizagem automática necessitam uma quantificação da noção de erro por forma a definir a função custo. Há necessidade, assim, de se introduzir métricas para quantificar a diferença entre dois eventos de uma base de dados, quaisquer que sejam os tipos dos seus atributos. A noção de "dissimilaridade" estabelece as condições mínimas para esta quantificação, que têm por contraponto a noção de "similaridade". A noção de "distância", que acrescenta condições à definição de dissimilaridade, é a que na prática se usa. [3]

2.1 Definições

Seja $\mathcal{A}$ um conjunto. Diz-se que a função $d : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ é uma **dissimilaridade** se

- $d(x, x') \geq 0, \forall x, x' \in \mathcal{A}$.
- $d(x, x) = 0, \forall x \in \mathcal{A}$.

Estas duas condições impõem que não há eventos mais próximos a um qualquer evento do que o próprio evento.

O contraponto faz-se com a noção de similaridade. Seja $\mathcal{A}$ um conjunto. Diz-se que a função $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ é uma **similaridade** se

- $s(x, x) \geq s(x, x'), \forall x, x' \in \mathcal{A}$.
- $s(x, x) \geq s(x', x), \forall x, x' \in \mathcal{A}$.

A dissimilaridade ao contrário da similaridade, calcula a diferença entre dois objectos da base de dados, que deve ser baixa quando dois eventos estão próximos e alta quando estão distantes.

Uma métrica mais restritiva do que a de dissimilaridade é dada pela noção de distância. Seja $\mathcal{A}$ um conjunto. Diz-se que a função $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ é uma **distância** se

- $d(x, x') \geq 0, \forall x, x' \in \mathcal{A}$.
- $d(x, x') = d(x', x), \forall x, x' \in \mathcal{A}$.
- $d(x, x'') \leq d(x, x') + d(x', x''), \forall x, x', x'' \in \mathcal{A}$.
- $d(x, x') = 0 \Leftrightarrow x = x', \forall x, x' \in \mathcal{A}$.

Nas duas secções seguintes apresentam-se exemplos simples e clássicos de distâncias que podem ser usadas em *Machine Learning* para a criação de *clusters*.

2.2 Métricas para dados numéricos

Nesta secção considera-se que os eventos são sequências ordenadas de $I(\in \mathbb{N})$ números reais, ou seja, $\mathcal{A} = \mathbb{R}^I$. Sejam, então, $x = (x_1, \ldots, x_I)^{\mathrm{T}}$ e $x' = (x'_1, \ldots, x'_I)^{\mathrm{T}}$ dois elementos de $\mathbb{R}^I$. Seja, ainda, $r > 0$. Pode-se provar que a função

$$d_r(x, x') = \left(\sum_{i=1}^{I} |x_i - x'_i|^r \right)^{\frac{1}{r}},$$

a que se chama **Distância de Minkowsky**, é uma distância. Dois casos particulares são a **Distância de Manhattan**, onde se considera $r = 1$,

$$d_1(x, x') = \sum_{i=1}^{I} |x_i - x'_i|,$$

e a **Distância Euclidiana**, onde se considera $r = 2$,

$$d_2(x, x') = \sqrt{\sum_{i=1}^{I} |x_i - x'_i|^2}.$$

Pode-se também mostrar que

$$\lim_{r \to +\infty} d_r(x, x') \equiv d_\infty(x, x') = \max_{i=1}^{I} |x_i - x'_i|$$

e que d_∞ também é uma distância, a que se chama **Distância de Chebyshev**.

No caso de os atributos de um evento não terem todos a mesma importância e/ou grandeza, pode ser útil reforçar o impacto de algum deles. Assim, podemos usar pesos $\alpha_i > 0$ e definir uma métrica ponderada dada por

$$d_\alpha(x, x') = \left(\sum_{i=1}^{I} \alpha_i |x_i - x'_i|^r \right)^{\frac{1}{r}}.$$

Um caso particular é a normalização/dimensionamento dos dados com um valor de referência para cada atributo. Por exemplo em contexto de atributos mistos (tempo, distância, volume, ...) a normalização permite eliminar o viés associado às unidades escolhidas para fazer as medidas.

2.3 Métricas para dados nominativos

No caso de os dados serem nominativos, ou seja, dados que não podem ser manipulados algebricamente, há necessidade de definir novas métricas. Nesta secção considera-se, então, que os eventos são sequências ordenadas de $I(\in \mathbb{N})$ elementos, ou seja,

$\mathcal{A} = A_1 \times A_2 \times \cdots \times A_I$, em que A_i, $i = 1, \ldots, I$, são conjuntos de dados nominativos (por exemplo, $A_1 = \{\text{vermelho}, \text{azul}, \text{branco}\}$, $A_2 = \{\text{vaca}, \text{gato}\}$, $A_3 = \{\text{grande}, \text{médio}, \text{pequeno}\}$). Sejam, então, $x = (x_1, \ldots, x_I)^{\mathrm{T}}$ e $x' = (x'_1, \ldots, x'_I)^{\mathrm{T}}$ dois elementos de $\mathcal{A}$. A **Distância de Hamming** (ou *Matching Distance*) é dada por

$$d(x, x') = \frac{1}{I} \sum_{i=1}^{I} d_i(x, x'),$$

em que

$$d_i(x, x') = \begin{cases} 1 & \text{se } x_i \neq x'_i, \\ 0 & \text{caso contrário.} \end{cases}$$

Uma variante é considerar novamente métricas ponderadas usando o cardinal de cada atributo (*i.e.* o número de valores possíveis para cada atributo), como por exemplo

$$d(x, x') = \frac{1}{I} \sum_{i=1}^{I} \omega_i d_i(x, x'),$$

onde $\omega_i = |A_i|$ ou $\omega_i = 1/|A_i|$.

Capítulo 3

Partição e Clustering

Seja dada uma base de dados constituída por eventos não rotulados em que se pretende determinar uma sua partição por forma a que os elementos dos elementos da partição sejam similares entre si mas dissimilares em relação aos elementos dos outros elementos da partição. Ao processo de criar a partição de uma base de dados chama-se *clustering* e aos elementos da partição (subconjuntos da base de dados) chama-se *clusters*.

Formalmente, dada uma base de dados D, pretende-se criar uma sua partição com $K(\in \mathbb{N})$ elementos, ou seja, $\mathcal{P} = \{C^1, \ldots, C^K\}$ tal que

$$D = \bigcup_{k=1}^{K} C^k$$

e

$$C^k \cap C^{k'} = \emptyset, \text{com } k, k' = 1, \ldots, K, k \neq k'.$$

Existem muitas técnicas para a construção dos *clusters*, seguindo cada metodologia o seu conjunto de regras para definir a semelhança entre os elementos da base de dados. Dois dos grandes grupos de algoritmos de *clustering* são os algoritmos particionais e os algoritmos hierárquicos, que se introduzem nas duas secções seguintes.

3.1 Algoritmos particionais

Os algoritmos particionais baseiam-se no que hoje em dia se chama "Algoritmo de Lloyd" e que foi introduzido em 1957 por Stuart P. Lloyd mas apenas publicado em 1982 [6]. O algoritmo de Lloyd, que surgiu no contexto da modulação por códigos de pulso, tinha por objetivo determinar um conjunto finito de valores que representam um domínio contínuo de valores. Como técnica de *clustering*, o algoritmo de Lloyd consiste em, dada uma base de dados D e um conjunto $M(0) = \{m^1, \ldots, m^K\}$ com $K(\in \mathbb{N})$ elementos do espaço dos atributos, aplicar iterativamente os seguintes dois passos até se atingir a convergência:

Passo 1 – *assignment* (atribuição): determinar os K *clusters* associados a M

$$M(t) = \{m^1, \ldots, m^K\} \rightarrow \mathcal{P}(t+1) = \{C^1, \ldots, C^K\}$$

Passo 2 – atualização: determinar os K novos representantes associados a $\mathcal{P}$

$$\mathcal{P}(t+1) = \{C^1, \ldots, C^K\} \rightarrow M(t+1) = \{m^1, \ldots, m^K\}$$

O algoritmo *K-means*, termo introduzido por James MacQueen em 1967 [7] é o mais conhecido algoritmo de *clustering* que se baseia no Algoritmo de Lloyd. O objetivo do *K-means* é minimizar a distância a que se encontram os pontos de um *cluster* do centróide desse *cluster*. O algoritmo é dado por:

Passo 0 – inicialização

$D = (x^n)_{n=1}^N, x^n \in \mathcal{A} = \mathbb{R}^I$: base de dados

$K \in \mathbb{N}$: número de *clusters*

$M(0) \leftarrow \{m^1, \ldots, m^k, \ldots, m^K\}$: K pontos distintos de $\mathcal{A}$

d : dissemelhança (usualmente a distância euclidiana ou de Manhattan)

$t \leftarrow 0$

Passo 1 – *assignment*

$$M(t) = \{m^1, \ldots, m^k, \ldots, m^K\} \rightarrow \mathcal{P}(t+1) = \{C^1, \ldots, C^k, \ldots, C^K\}$$
$$C^k \leftarrow \{x^n \in D : d(x^n, m^k) \leq d(x^n, m^j), j \in \{1, \ldots, k-1, k+1, \ldots, K\}\}$$

Passo 2 – atualização dos representantes

$$\mathcal{P}(t+1) = \{C^1, \ldots, C^k, \ldots, C^K\} \rightarrow M(t+1) = \{m^1, \ldots, m^k, \ldots, m^K\}$$
$$m^k \leftarrow \frac{1}{|C^k|} \sum_{x \in C^k} x$$

Passo 3 – fim?

terminar se $\mathcal{P}(t+1) = \mathcal{P}(t)$

caso contrário, $t \leftarrow t+1$ e regressar ao **Passo 1**

Num algoritmo de Loyd, o valor de K é um meta-parâmetro indicado pelo utilizador, colocando-se a questão de se saber qual o seu melhor valor. A estratégia é quantificar a qualidade do *clustering* para cada valor de K com uma função de custo $E(K)$. Obtém-se, assim, uma curva $K \rightarrow E(K)$ que permite determinar o valor ótimo de K.

3.2 Algoritmos hierárquicos

Os algoritmos hierárquicos, ao contrário dos particionais, não assumem um número pré-definido de *clusters* mas antes consideram uma estrutura hierárquica de partições tomando como ponto de partida um dos dois casos extremos: uma partição com tantos elementos quantos os elementos da base de dados nos chamados algoritmos hierárquicos aglomerativos, ou uma partição com um único elemento que é a própria base de dados nos chamados algoritmos hierárquicos divisivos. Além do mais, os algoritmos hierárquicos não precisam necessariamente de um representante dos *clusters*. Apenas algumas situações, tais como a utilização dos centróides para avaliar a distância entre *clusters*, pode

requerer um representante. Os algoritmos hierárquicos aglomerativos, mais eficientes que os algoritmos hierárquicos divisivos, vão ser os considerados neste trabalho.

Um algoritmo hierárquico aglomerativo, dada uma base de dados $D = (x^n)_{n=1}^N$ e uma dissemelhança d, considera $\mathcal{P} = \{C^1, \ldots, C^n, \ldots, C^N\}$, $C^n = \{x^n\}$, como a partição inicial (cada evento da base de dados constitui um *cluster*), e repete os seguintes dois passos até existir um único *cluster*:

Passo 1 – calcular todas as distâncias inter-*clusters* (matriz de distâncias)
Passo 2 – agrupar os dois *clusters* cuja distância inter-*clusters* é menor

Para calcular a distância entre dois *clusters* C e C', operação necessária para calcular as distâncias inter-*clusters* no Passo 1 do algoritmo, irão ser consideradas neste trabalho as seguintes estratégias:

- *single linkage* – a distância entre os *clusters* é dada pela distância dos dois elementos mais próximos dos dois *clusters*, ou seja,

$$d(C, C') = \min_{x \in C, x' \in C'} d(x, x').$$

- *complete linkage* – a distância entre os *clusters* é dada pela distância dos dois elementos mais afastados dos dois *clusters*, ou seja,

$$d(C, C') = \max_{x \in C, x' \in C'} d(x, x').$$

- *average linkage* – a distância entre os *clusters* é dada pela média das distâncias entre todos os pares dos dois *clusters*, ou seja,

$$d(C, C') = \frac{1}{|C||C'|} \sum_{x \in C} \sum_{x \in C'} d(x, x').$$

Para ilustrar o algoritmo, considere-se a base de dados com espaço de atributos $\mathcal{A} = \mathbb{R}^2$ dada por

$$D = \big\{(0,1), (2,0), (1,2), (2,0.5), (1,2.75)\big\}.$$

Consideremos d a distância de *Manhattan* sendo a entrada (ℓ, c) da matriz de dissimilaridade dada por $M[\ell, c] = d(x^\ell, x^c)$ e o método *Single linkage* para calcular a dissimilaridade entre os *clusters*, i.e. seleciona-se o menor valor, que vamos denotar por E e que representa a "energia de associação" para agrupar dois *clusters* na partição.

- A partição inicial é $\mathcal{P}(0) = \big\{\{x^1\}, \{x^2\}, \{x^3\}, \{x^4\}, \{x^5\}\big\}$ e a matriz de distâncias está representada na Tabela 3.1. Como a menor dissimilaridade é o valor da entrada $(4, 2)$, que é a distância entre o *cluster* $\{x^2\}$ e o *cluster* $\{x^4\}$, tem-se que $E = 0.5$ e o próximo passo é juntar estes dois *clusters* num único. Assim, a nova partição é dada por $\mathcal{P}(1) = \big\{\{x^1\}, \{x^2, x^4\}, \{x^3\}, \{x^5\}\big\}$, com 4 *clusters*.

Tabela 3.1: $M(\mathcal{P}(0))$.

	x^1	x^2	x^3	x^4	x^5
x^1	0	3	2	2.5	2.75
x^2	3	0	3	0.5	3.75
x^3	2	3	0	2.5	0.75
x^4	2.5	0.5	2.5	0	3.25
x^5	2.75	3.75	0.75	3.25	0

- Recalcula-se a matriz, desta vez, associada à partição $\mathcal{P}(1)$. Para tal, quando se calcula a distância do *cluster* $\{x^1\}$ com o novo *cluster* $\{x^2, x^4\}$, tem de se selecionar o $\min(d(\{x^1\}, \{x^2\}), d(\{x^1\}, \{x^4\}))$. O mesmo acontece, para calcular a distância entre os restantes *clusters* e o *cluster* $\{x^2, x^4\}$, obtendo-se a matriz de distâncias dada na Tabela 3.2. Atendendo a que $E = 0.75$, que é a distância entre os *clusters* $\{x^3\}$ e $\{x^5\}$, estes dois *clusters* vão-se fundir num só *cluster*. Obtém-se assim, uma nova partição $\mathcal{P}(2) = \big\{\{x^1\}, \{x^2, x^4\}, \{x^3, x^5\}\big\}$.

Tabela 3.2: $M(\mathcal{P}(1))$.

	x^1	x^2, x^4	x^3	x^5
x^1	0	2.5	2	2.75
x^2, x^4	2.5	0	2.5	3.25
x^3	2	2.5	0	0.75
x^5	2.75	3.25	0.75	0

- A matriz das distâncias $M(\mathcal{P}(2))$ associada à partição $\mathcal{P}(2)$ está representada na Tabela 3.3. Agora, tem-se $E = 2$, que é a distância entre os *clusters* $\{x^1\}$ e $\{x^3, x^5\}$, obtendo-se a partição $\mathcal{P}(3) = \big\{\{x^2, x^4\}, \{x^1, x^3, x^5\}\big\}$.

Tabela 3.3: $M(\mathcal{P}(2))$.

	x^1	x^2, x^4	x^3, x^5
x^1	0	2.5	2
x^2, x^4	2.5	0	2.5
x^3, x^5	2	2.5	0

- Por fim, a matriz das distâncias $M(\mathcal{P}(3))$ associada à partição $\mathcal{P}(3)$ está representada na Tabela 3.4. Assim, $E = 2.5$ e já só resta um passo, que é juntar os dois *clusters* existentes, $\{x^2, x^4\}$ e $\{x^1, x^3, x^5\}$, obtendo-se assim a partição final que é constituída apenas por um *cluster* que inclui todos os pontos de D, $\mathcal{P}(4) = \big\{\{x^1, x^2, x^3, x^4, x^5\}\big\}$.

Tabela 3.4: $M(\mathcal{P}(3))$.

	x^2, x^4	x^1, x^3, x^5
x^2, x^4	0	2.5
x^1, x^3, x^5	2.5	0

Construídas todas as partições e as respectivas distâncias inter-*clusters* (energias de associação), através de uma curva que relaciona o número de *clusters* de cada partição e a sua energia de associação, é possível escolher o número de *clusters* ideal, visualizando nesta curva qual o ponto em que há uma subida mais abrupta tendo em consideração as variações da curva. Na Figura 3.1 apresentam-se as curvas de associação quando se considera o método *Single linkage*, *Complete linkage* e *Average linkage*, a partir das quais se poderá concluir que o número ideal de *clusters* será 3.

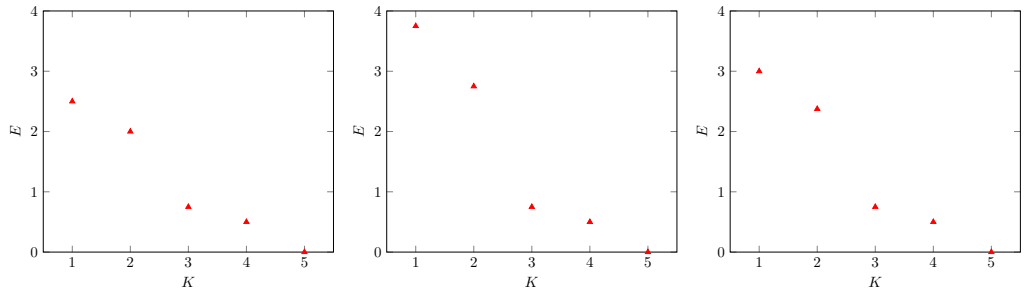


Figura 3.1: Curva de associação do *clustering* hierárquico para *Single linkage* (esquerda), *Complete linkage* (centro) e *Average linkage* (direita).

3.3 Avaliação da qualidade dos clusters

Perante a necessidade de avaliar a qualidade dos *clusters* que se podem construir para uma certa *base de dados*, recorre-se ao uso de métricas de avaliação. Uma delas é a *Inércia*, que calcula a soma das distâncias que estão os pontos de um *cluster* ao

centróide desse mesmo *cluster*, ou seja, avalia a "distância intra-*cluster*", que deve ser o mais baixa possível. Outra métrica é o *Índice de Dunn*, que para além de garantir o mesmo que a Inércia, assegura também que *clusters* distintos sejam tão diferentes um dos outros quanto possível. Assim, avalia-se a "distância inter-*cluster*" para além da "distância intra-*cluster*". Este índice deve ser o mais alto possível, sendo a sua expressão dada por

$$\text{Índice de Dunn} = \frac{\min(\text{distância inter-}cluster)}{\max(\text{distância intra-}cluster)}.$$

Na caracterização da qualidade de um processo de *clustering* pode também contar o número de *clusters* como elemento de penalização. Por exemplo, se $\mathcal{P} = \{C^1, \ldots, C^K\}$ é uma partição da base de dados, considera-se a função custo

$$E(\mathcal{P}) = aK^{\alpha} + \sum_{k=1}^{K} V(C^k),$$

onde a variância de um *cluster* é dada por

$$V(C^k) = \frac{1}{|C^k|} \sum_{x \in C^k} \|x - \bar{x}\|^2$$

com $\bar{x}$ a média dos elementos de C^k e a e α parâmetros do modelo.

Capítulo 4

Métrica de atributos temporais

A componente temporal dos dados tem um papel diferente dos outros atributos e deve ser quantificada de uma maneira especifica. Assim, seja $x = (id, t, y)$ um evento, onde id é a identificação do cliente que realizou esse evento, t é a data em que o realizou e y representa todos os outros atributos associados a este evento. Para a construção dos *clusters* irá ser preciso apenas considerar o par (t, y), ficando o atributo id de fora, ou seja, não contará como atributo. Uma métrica temporal separa o tempo t dos restantes atributos y introduzindo uma ponderação dependente de t.

Na literatura já existem algumas métricas que diferenciam o tempo dos outros atributos. Por exemplo, no modelo de janela amortecida, os atributos de um evento são afetados por um coeficiente que depende do tempo em que este ocorreu relativamente a um tempo referencial dado. Assim, quando um elemento é acrescentado à base dados é lhe atribuído o maior peso possível, sendo que esse peso diminui através de "funções de envelhecimento", como por exemplo a função de decaimento exponencial

$$f(t) = 2^{-\lambda|t-t_r|},$$

onde λ é o factor de envelhecimento e t_r é o tempo de referência. Este modelo de janela amortecida não descarta completamente os elementos, simplesmente considera que os

mais afastados do tempo de referência contribuem menos porque recebem pesos mais baixos.

4.1 Uma nova métrica temporal

Nesta secção vai-se apresentar uma nova proposta de métrica, criada de raiz, para tratar de eventos com uma componente temporal.

Consideremos uma métrica que toma em conta o tempo da observação $\bar{t}$ para avaliar a distância entre dois eventos. Assim, introduzimos a nova métrica $d((t,y),(t',y');\bar{t})$ para calcular a distância entre dois eventos $x = (id, t, y)$ e $x' = (id', t', y')$.

Pedimos que as seguintes condições sejam satisfeitas pela métrica.

1. *(H1) Condição de separação das variáveis.* Para qualquer par de eventos x e x' e tempos de referência $\bar{t}$, temos

   $$d((t,y),(t',y');\bar{t}) = \rho(t,t';\bar{t})d(y,y').$$

2. *(H2) Condição de invariância por translação temporal.* Para qualquer tempo $a \in \mathbb{R}$, temos

   $$\rho(t+a, t'+a; \bar{t}+a) = \rho(t,t';\bar{t})$$

   Em particular se $a = -\bar{t}$ então $\rho(t,t';\bar{t}) = \rho(t-\bar{t}, t'-\bar{t}; 0)$ o que implica

   $$\rho(t,t';\bar{t}) = \rho(t-\bar{t}, t'-\bar{t}).$$

3. *(H3) Invariância das distancias relativas ao tempo de referência.* O rácio das distâncias $d((t,y),(t',y');\bar{t})$ e $d((s,y),(s',y');\bar{t})$é independente de $\bar{t}$, ou seja

   $$\frac{\rho(t-\bar{t}, t'-\bar{t})}{\rho(s-\bar{t}, s'-\bar{t})} = C(t,t',s,s'). \tag{4.1}$$

Proposição

Supomos que a métrica satisfaz as três condição *(H1)*, *(H2)*, *(H3)*. Então temos

$$\rho(t-\bar{t},t'-\bar{t})d(y,y') = \exp\big(f(t,t')\big)\exp(A\bar{t})d(y,y')$$

com $A \in \mathbb{R}$.

Demonstração.

Derivamos a expressão 4.1 em ordem a $\bar{t}$ e obtemos

$$\begin{aligned}&\Big(\partial_1\rho(t-\bar{t},t'-\bar{t}) + \partial_2\rho(t-\bar{t},t'-\bar{t})\Big)\rho(s-\bar{t},s'-\bar{t})\\ &\quad = \Big(\partial_1\rho(s-\bar{t},s'-\bar{t}) + \partial_2\rho(s-\bar{t},s'-\bar{t})\Big)\rho(t-\bar{t},t'-\bar{t}),\end{aligned}$$

onde ∂_1 e ∂_2 representam as derivadas parciais em ordem ao primeiro e segundo argumento. Deduzimos que

$$\frac{\frac{d}{d\bar{t}}\rho(t-\bar{t},t'-\bar{t})}{\rho(t-\bar{t},t'-\bar{t})} = \frac{\partial_1\rho(t-\bar{t},t'-\bar{t}) + \partial_2\rho(t-\bar{t},t'-\bar{t})}{\rho(t-\bar{t},t'-\bar{t})} = A \in \mathbb{R}.$$

Primitivando em ordem a $\bar{t}$, obtém-se

$$P_{\bar{t}}\left(\frac{\frac{d}{d\bar{t}}\rho(t-\bar{t},t'-\bar{t})}{\rho(t-\bar{t},t'-\bar{t})}\right) = P_{\bar{t}}A \Leftrightarrow \ln\big(\rho(t-\bar{t},t'-\bar{t})\big) = f(t,t') + A\bar{t}.$$

Logo obtemos a fórmula. □

Este resultado indica que $A = -\frac{1}{\tau}$ corresponde a um *scaling* do tempo. A função $f(t,t')$ pode ser qualquer mas deve manter o *scaling*. Uma possibilidade consiste em utilizar o tempo mínimo entre t e t' e deduzimos que

$$f(t,t') = \frac{\min(t,t')}{\tau}.$$

Obtemos assim a métrica

$$\begin{aligned}d((t,y),(t',y');\bar{t}) &= \exp\left(\frac{\min(t,t')}{\tau}\right)\exp\left(-\frac{\bar{t}}{\tau}\right)d(y,y')\\ &= \exp\left(\frac{\min(t-\bar{t},t'-\bar{t})}{\tau}\right)d(y,y').\end{aligned}$$

A partir desta métrica podemos então construir os *clusters*. A ideia é observar as trajetórias dos eventos. Por exemplo, será que um evento de um cliente que ocorreu há 3 meses e está num determinado *cluster*, passado 1 mês, outro evento desse mesmo cliente ainda está no mesmo *cluster*? Ou a sua trajetória mudou para outro *cluster*?

4.2 Exemplo

Consideremos a base de dados dada na Tabela 4.1. Ora, para criar a matriz de dissimilaridade desta base de dados, é necessário calcular a distância de cada elemento a cada elemento, usando a nova métrica.

Tabela 4.1: Base de dados.

id	t	y_1	y_2
x^1	1	2	3
x^2	2	1	1
x^3	1	2	4
x^4	3	3	4

Sejam $\tau = 5$ e $\bar{t} = 5$. Para calcular, por exemplo, $d(x^1, x^2; \bar{t})$, temos:

$$\begin{aligned} d(x^1, x^2, \bar{t}) &= d((1,(2,3),(2,(1,1)),5) \\ &= \exp\left(\min\left(\frac{1-5}{5}, \frac{2-5}{5}\right)\right) d((2,3),(1,1)) \\ &\approx 2.01. \end{aligned}$$

O mesmo é feito para as restantes distâncias, de onde se obtém a Tabela 4.2 da matriz de dissimilaridade.

Tabela 4.2: Matriz de dissimilaridade usando a nova métrica.

	x^1	x^2	x^3	x^4
x^1	0	2.01	0.67	1.34
x^2	2.01	0	2.68	4.09
x^3	0.67	2.68	0	0.67
x^4	1.34	4.09	0.67	0

Tendo esta matriz de dissimilaridade calculada, pode ser utilizada como *input* no *clustering* hierárquico aglomerativo.

Capítulo 5

Aplicação a várias bases de dados sintéticos

Neste capítulo vai-se comparar a métrica introduzida no capítulo anterior com a distância de *Manhattan* quando aplicadas a bases de dados sintéticas. Irão ser criados 5 exemplos distintos, alguns mais simples, isto é, exemplos onde se consigam distinguir bem os *clusters* visualmente e outros mais complexos onde irão haver interseções entre os *clusters* em determinados momentos do tempo.

5.1 Análise do clustering

Inicialmente, para cada exemplo, foram criados *clusters* sintéticos com etiquetas (*labels*) associadas, onde cada elemento foi identificado como fazendo parte de tal *cluster*.

Após a criação dos dados, cada análise começará com uma observação de uma curva do cotovelo e um dendrograma, que derivam da aplicação do algoritmo de *Agglomerative Clustering* à base de dados por forma a encontrar-se o número ideal de *clusters* da base de dados.

O *clustering* foi feito de três maneiras diferentes:

- Primeiramente, usou-se o algoritmo *Agglomerative Clustering* com distância de *Manhattan*, onde se consideram todos os atributos.

- De seguida considera-se o *Agglomerative Clustering* com a distância de *Manhattan* mas apenas para os atributos de espaço.

- Por último outra vez o *Agglomerative Clustering* com a métrica criada no Capítulo 4.

Para cada um destes três casos, foram aplicados três métodos diferentes: *single linkage*, *complete linkage* e *average linkage*.

Para o número tido por ótimo de *clusters*, foi feita uma associação entre os *clusters* criados artificialmente e os preditos pelo algoritmo. Assim, cada elemento tem associado duas etiquetas, a real e a predita. Com estes dados foram calculadas métricas de avaliação dos *clusters* (*accuracy, precision* e *recall*), através da função *classification report*, que basicamente dizem o quão correta foi a predição dos *clusters* em relação aos reais.

5.2 Métricas de avaliação

Como já foi referido, irão ser calculadas 3 métricas de avaliação dos *clusters*. [4] Nesta secção será introduzida a definição de cada uma. Considere-se uma base de dados, que num todo corresponde a uma partição P_{ex}, onde a cada elemento corresponde uma etiqueta que se refere ao *cluster* em que este elemento está inserido. Quando se aplica um algoritmo de *clustering* hierárquico a esta base de dados, vai ser associada uma nova etiqueta a cada elemento que faz corresponder ao *cluster* em que este modelo inseriu o elemento, formando uma nova partição P com todos os elementos e a respetiva etiqueta. O objetivo depois, é comparar a etiqueta real da partição P_{ex} com a etiqueta prevista pelo modelo da partição P. Aqui surgem as métricas de avaliação dos *clusters*:

- *Precision*: Diz o quão preciso é o modelo em relação aos dados previstos comparando-os com os dados inseridos, ou seja, compara todos os elementos de P e vê se a sua etiqueta corresponde à etiqueta do mesmo elemento em P_{ex}. Esta métrica é calculada para todas as diferentes etiquetas que existam na partição.

- *Recall*: Diz quantas etiquetas reais o modelo encontrou, ou seja, compara todos os elementos de P_{ex} e vê se a sua etiqueta corresponde à etiqueta do mesmo elemento na partição P. Esta métrica é calculada para todas as diferentes etiquetas que existam na partição.

- *Accuracy*: Esta métrica ao contrário das outras duas, é feita para o resultado geral do modelo, e não para cada etiqueta. Basicamente, a *accuracy* calcula a percentagem de etiquetas que o modelo previu correctamente.

Considere-se de seguida um pequeno exemplo para ajudar a entender estas métricas. Seja $D = \{x_1, x_2, x_3, x_4, x_5\}$ uma base de dados e $E = [0, 1, 2, 2, 2]$ as etiquetas associadas a cada elemento de D. Aplicou-se o *clustering* hierárquico aglomerativo e obteve-se um novo conjunto de etiquetas $E' = [0, 0, 2, 2, 1]$. Na Tabela 5.1 apresentam-se os valores das três métricas de qualidade.

A *precision* da etiqueta 0 é de 50% pois em E apenas o elemento x_1 tem etiqueta 0, mas o modelo previu que tanto o elemento x_1 como o x_2 tem etiqueta 0, ou seja, dos elementos com etiqueta 0 que o modelo previu, apenas metade tem mesmo a etiqueta 0. Já em relação ao *recall*, todos os elementos que tinham realmente a etiqueta 0 (no caso apenas o x_1), o modelo previu correctamente a etiqueta. Enquanto por exemplo em relação à etiqueta 2, havia 3 elementos, e o modelo previu apenas 2 deles, daí o *recall* ser apenas de 67%. A *accuracy* é de 60% porque o modelo apenas acertou 3 de 5 elementos.

Tabela 5.1: Exemplo — métricas de qualidade com K *etiquetas*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.60	0.50	1.00	0.00	0.00	1.00	0.67

5.3 Exemplo 1

5.3.1 Construção da base de dados

A base de dados é constituída por 100 elementos e 3 *clusters* com 18, 40 e 42 elementos, respetivamente, cada um da forma (y_1, y_2, t). Os pontos de cada *cluster* pertencem a paralelepípedos que não se intersectam seguindo as seguintes leis probabilísticas uniformes:

- *cluster* 1: $y_1 \sim U([0.1, 0.3])$, $y_2 \sim U([0.6, 0.8])$, $t \sim U([0, 3])$.
- *cluster* 2: $y_1 \sim U([0.1, 0.4])$, $y_2 \sim U([0.05, 0.4])$, $t \sim U([0, 3])$.
- *cluster* 3: $y_1 \sim U([0.6, 0.9])$, $y_2 \sim U([0.1, 0.7])$, $t \sim U([0, 3])$.

Representa-se na Figura 5.1 a base de dados do Exemplo 1.

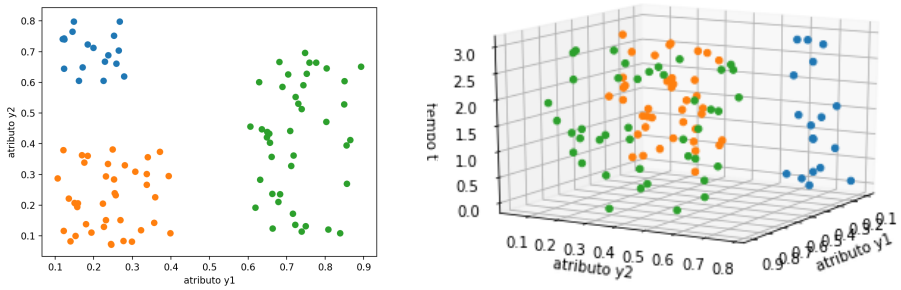
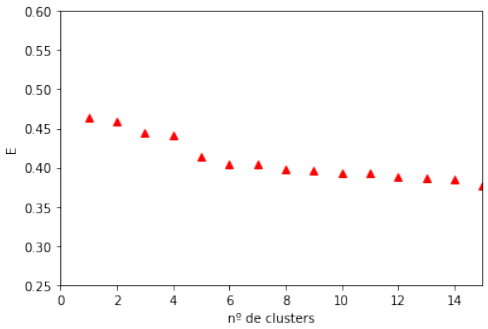


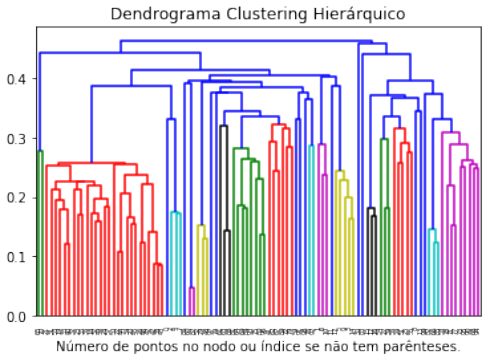
Figura 5.1: Exemplo 1 — visualização 2D e 3D dos dados.

5.3.2 Clustering hierárquico com distância de Manhattan

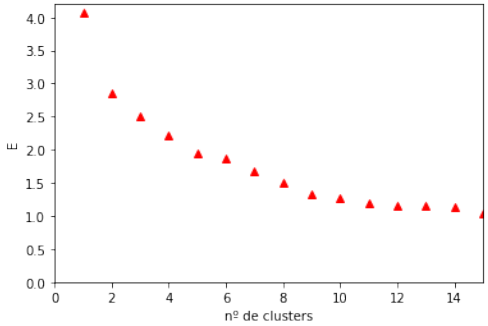
Apresentam-se os resultados para este exemplo nas Figuras 5.2, 5.3 e 5.4, e nas Tabelas 5.2 e 5.3.



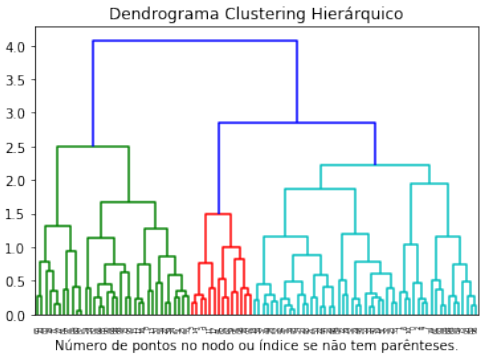
(a) *Single linkage* — curva do cotovelo



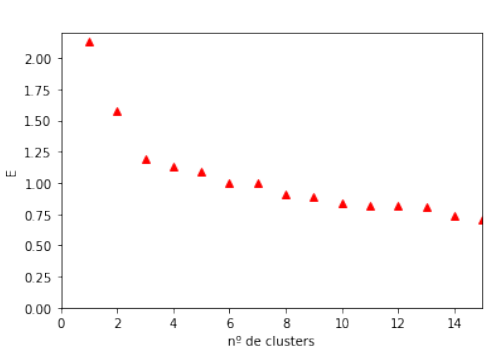
(b) *Single linkage* — dendrograma



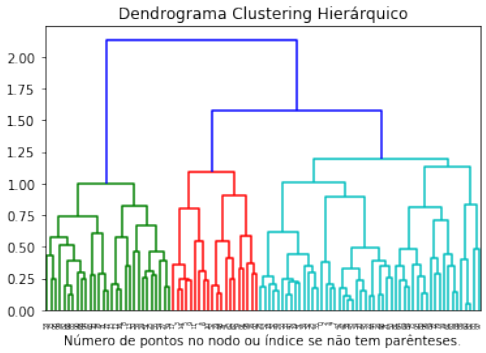
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

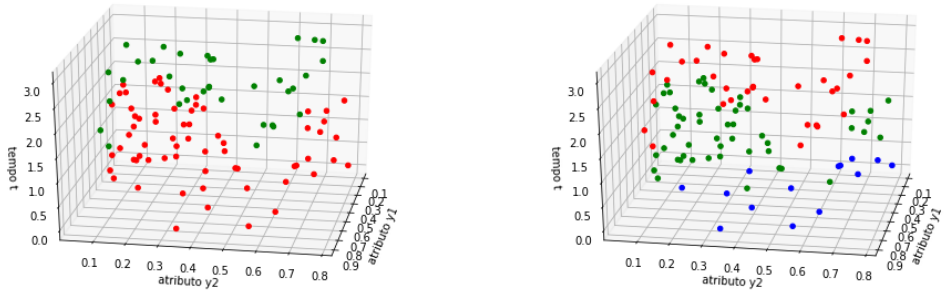
Figura 5.2: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan*.

Single linkage Quer considerando a curva do cotovelo, quer o dendrograma, não se consegue tirar grandes conclusões relativamente ao número ideal de *clusters* para o algoritmo.

Complete linkage Quer considerando a curva do cotovelo, quer o dendrograma, $K = 2$ e $K = 3$ parecem ser os valores ideais para o número de *clusters*.

Tabela 5.2: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.52	0.48	0.78	0.60	0.50	0.00	0.00
3	0.57	0.60	0.50	0.59	0.75	0.43	0.33



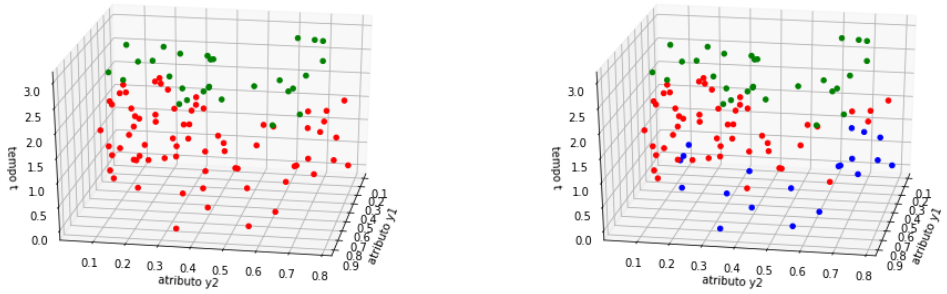
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.3: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*.

Average linkage O número ideal será 2, no entanto, 3 também parece uma boa opção, pelo que se observa da curva do cotovelo e do dendrograma.

Tabela 5.3: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.46	0.44	0.78	0.52	0.36	0.00	0.00
3	0.51	0.53	0.68	0.52	0.36	0.45	0.50



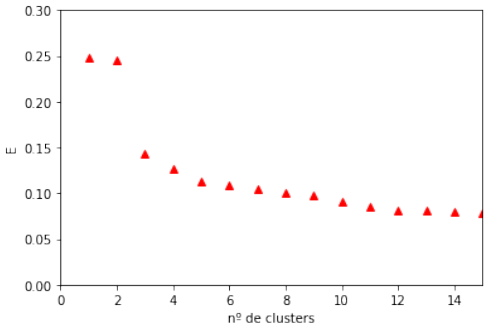
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.4: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*.

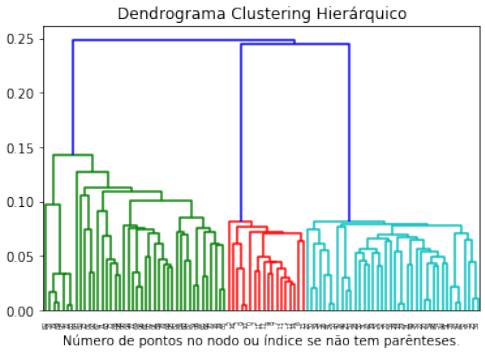
Conclusão Neste caso verificou-se que o *single linkage* não permitiu retirar grandes conclusões. Nos outros dois métodos embora com 3 *clusters* a *accuracy* seja ligeiramente melhor que usando 2, tanto um caso como o outro funciona bastante mal.

5.3.3 Clustering hierárquico com distância de Manhattan só para os atributos de espaço

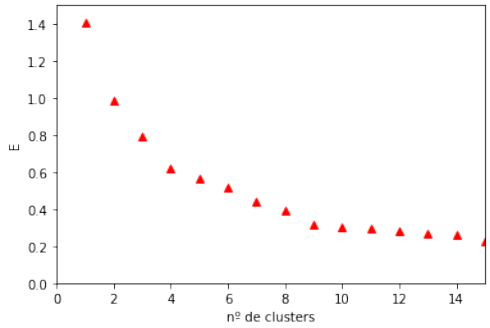
Apresentam-se os resultados para este exemplo nas Figuras 5.5, 5.6, 5.7 e 5.8, e nas Tabelas 5.4, 5.5 e 5.6.



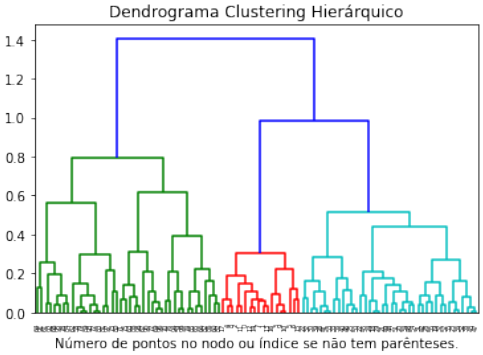
(a) *Single linkage* — curva do cotovelo



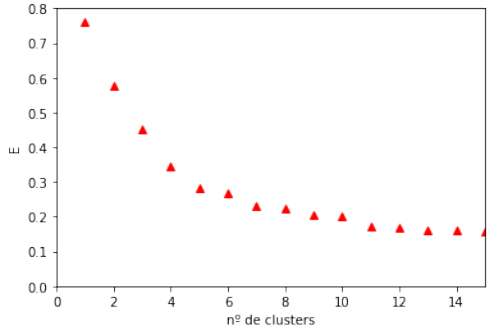
(b) *Single linkage* — dendrograma



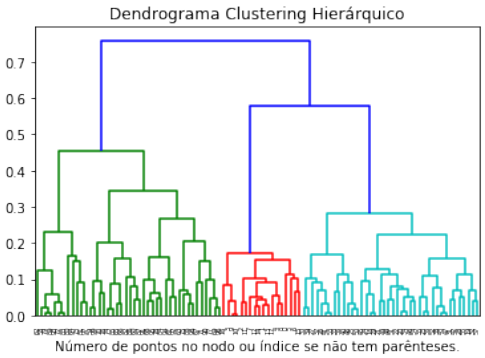
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.5: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço.

Single linkage. Pelo dendrograma e pela curva de cotovelo $K = 3$ é o número ideal de *clusters*.

Tabela 5.4: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00

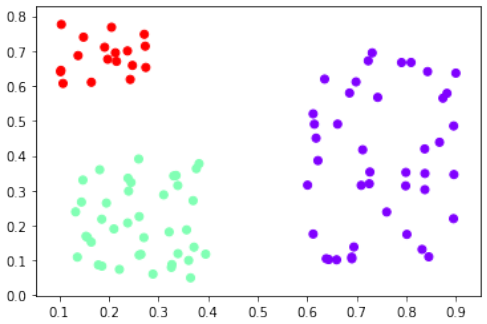


Figura 5.6: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: 3 *clusters*.

Complete linkage Mais uma vez, $K = 3$ é o número ideal de *clusters*.

Tabela 5.5: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00

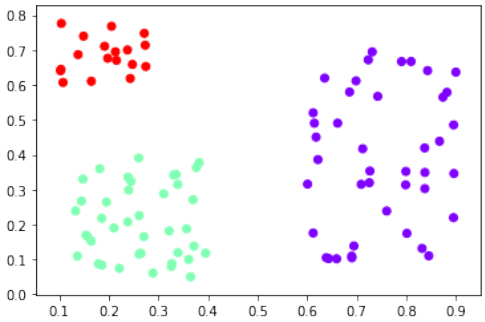


Figura 5.7: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*: 3 *clusters*.

Average linkage $K = 3$ é o número ideal de *clusters* a ser usado.

Tabela 5.6: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00

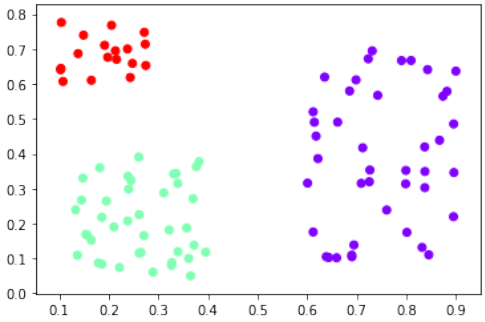
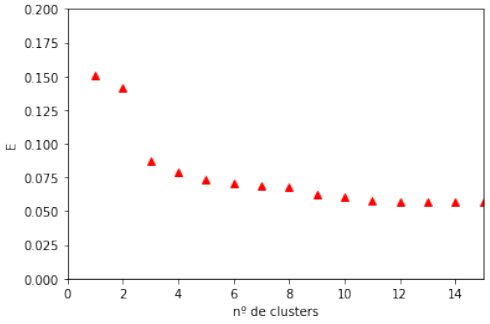


Figura 5.8: Exemplo 1 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*: 3 *clusters*.

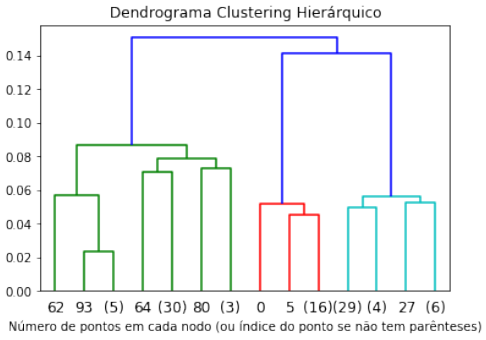
Conclusão Sem o atributo do tempo, o algoritmo consegue encontrar os *clusters* pretendidos.

5.3.4 Clustering hierárquico com a primeira tentativa de métrica

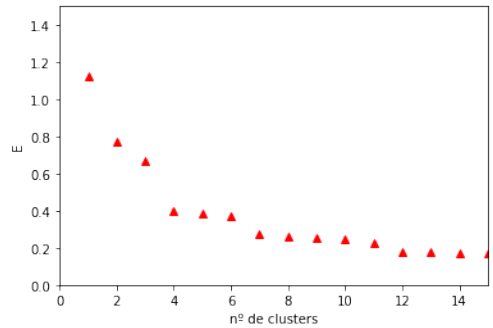
Vai-se considerar para tempo de observação $\bar{t} = 3$ e $\tau = 5$. Apresentam-se os resultados para este exemplo nas Figuras 5.9, 5.10, 5.11 e 5.12, e nas Tabelas 5.7, 5.8 e 5.9.



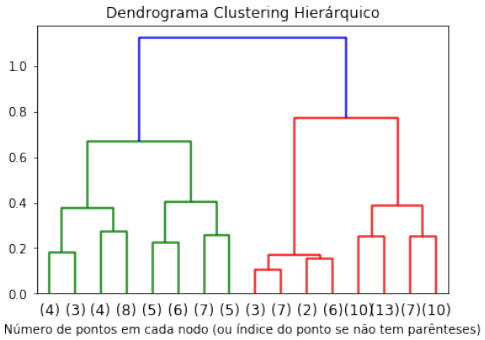
(a) *Single linkage* – curva do cotovelo



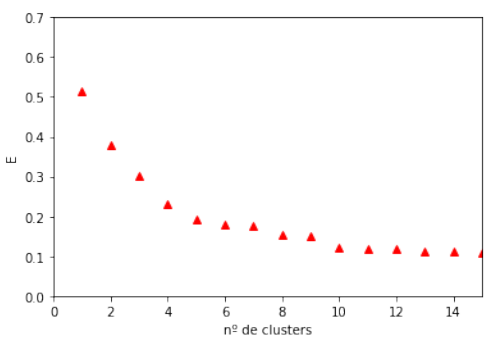
(b) *Single linkage* – dendrograma



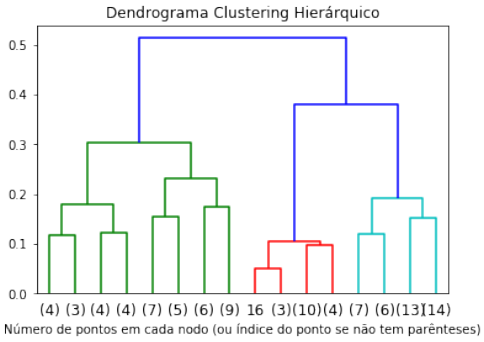
(c) *Complete linkage* – curva do cotovelo



(d) *Complete linkage* – dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.9: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica.

Single linkage O número ideal de *clusters* é $K = 3$, pelo que se observa da curva do cotovelo e do dendrograma.

Tabela 5.7: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00

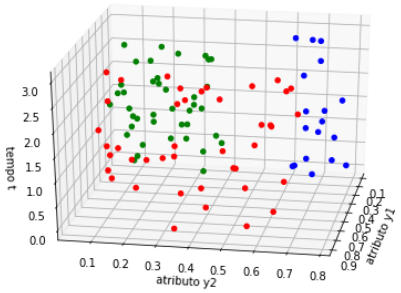
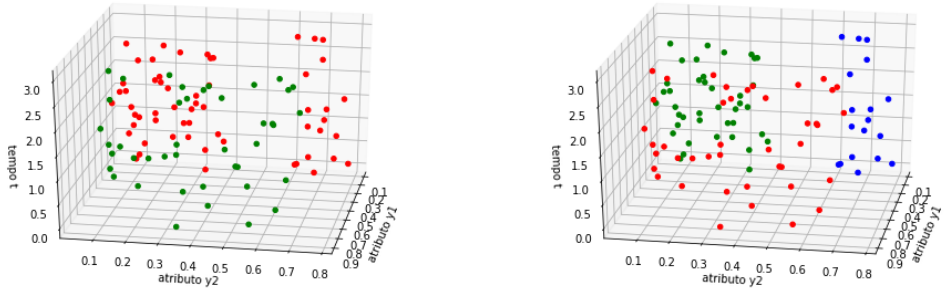


Figura 5.10: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: 3 *clusters*.

Complete linkage $K = 2$ parece ser o número ideal de *clusters*, no entanto aplicou-se também o algoritmo para $K = 3$.

Tabela 5.8: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.82	0.69	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) 2 *clusters* (b) 3 *clusters*

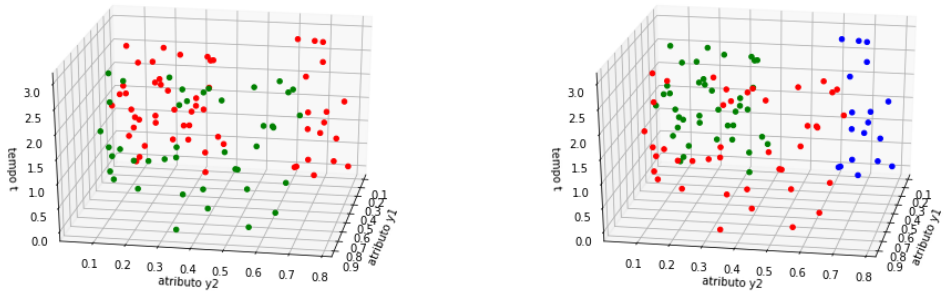
Figura 5.11: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*.

Average linkage

Tanto $K = 2$ como $K = 3$ parecem ser um número ideal de *clusters*.

Tabela 5.9: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.82	0.69	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.12: Exemplo 1 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*.

Conclusão Como se pode observar, a *accuracy* do modelo usando 3 *clusters* foi melhor, uma vez que acertou em tudo.

5.3.5 Conclusões

Nesta base de dados sintética, facilmente se conclui que quando se usa a métrica criada no Capítulo 4 existe um grande impacto no *Agglomerative Clustering*, uma vez que nos casos em que a métrica foi utilizada o modelo acertou totalmente. Isto deve-se ao facto da métrica dar uma importância diferente aos atributos de espaço em relação ao

atributo do tempo. Por isso, é que também quando se aplicou o algoritmo com distância de Manhattan só com os atributos de espaço a *accuracy* foi de 100%, uma vez que os *clusters* ficaram claramente distintos. Pelo contrário, quanto o tempo teve a mesma influência que os restantes atributos os resultados foram fracos, e com o nem se conseguiu chegar a uma conclusão, o que pode dever-se ao facto deste método ter em conta a distância mínima entre os elementos e neste exemplo eles estarem bastante dispersos.

5.4 Exemplo 2

5.4.1 Construção da base de dados

A base de dados é constituída por 35 elementos e 3 *clusters* com 15, 12 e 8 elementos, respetivamente, cada um da forma (y_1, y_2, t). Os pontos de cada *cluster* pertencem a cilindros que não se intersectam seguindo as seguintes regras:

- *cluster* 1: $m_1 = 0$, $m_2 = 0$ e $\sigma = 0.05$
- *cluster* 2: $m_1 = \frac{2}{3}$, $m_2 = \frac{1}{3}$ e $\sigma = 0.05$
- *cluster* 3: $m_1 = \frac{2}{3}$, $m_2 = \frac{1}{4}$ e $\sigma = 0.05$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $r = \sigma.\mathrm{rand}()$
- $\theta = 2\pi\mathrm{rand}()$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$
- $t = 3.\mathrm{rand}()$

Representa-se na Figura 5.13 a base de dados do Exemplo 2.

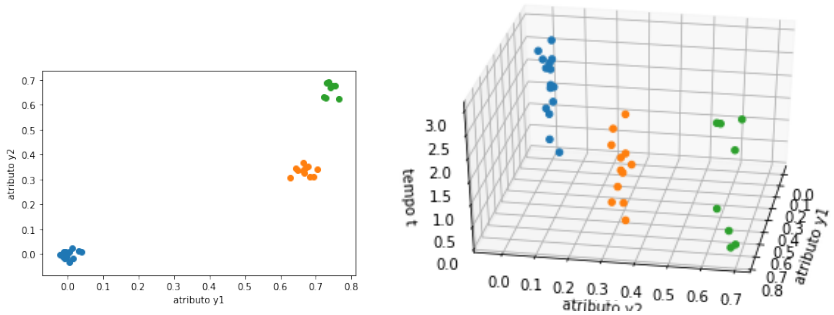
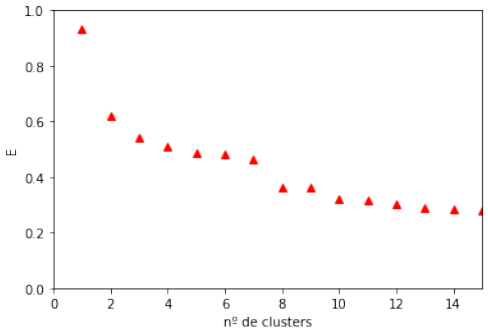


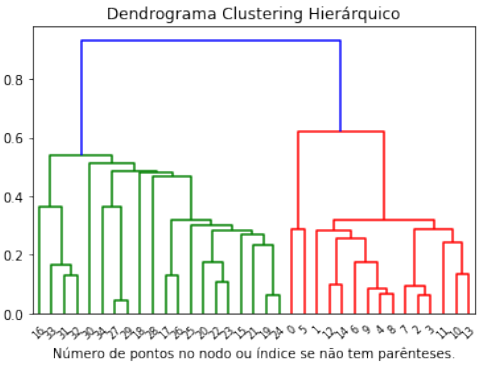
Figura 5.13: Exemplo 2 — visualização 2D e 3D dos dados.

5.4.2 Clustering hierárquico com distância de Manhattan

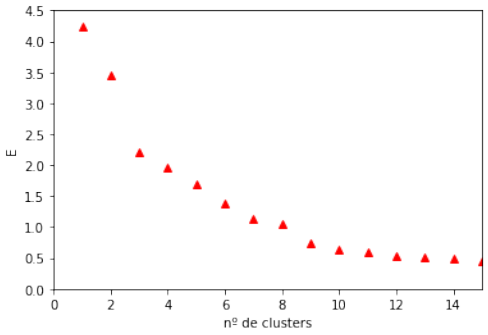
Apresentam-se os resultados para este exemplo nas Figuras 5.14, 5.15, 5.16 e 5.17, e nas Tabelas 5.10, 5.11 e 5.12.



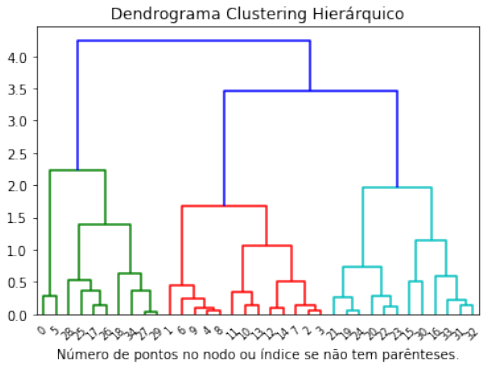
(a) *Single linkage* — curva do cotovelo



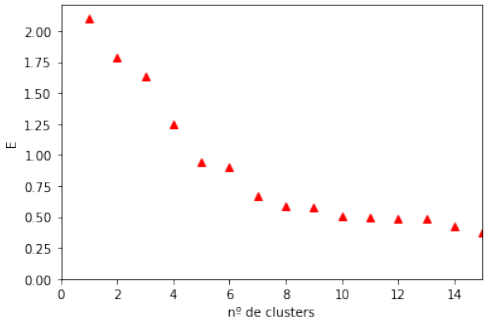
(b) *Single linkage* — dendrograma



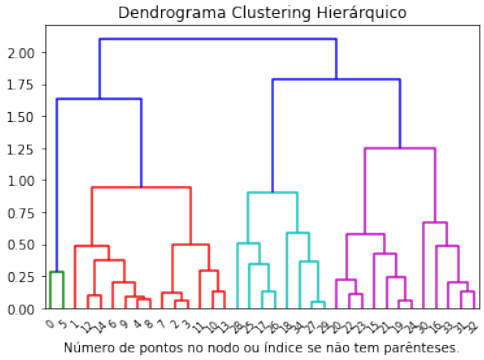
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.14: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan*.

Single linkage Quer considerando a curva do cotovelo, quer o dendrograma, conclui-se que $K = 2$ é o ideal.

Tabela 5.10: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	1.00	1.00	0.60	1.00	0.00	0.00

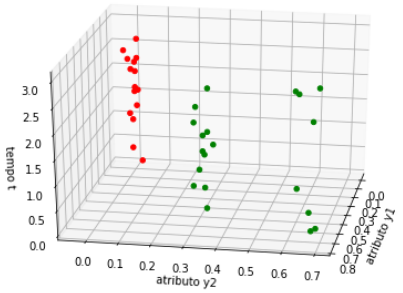


Figura 5.15: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: 2 *clusters*.

Complete linkage Considerando a curva do cotovelo e o dendrograma, $K = 3$ é o valor ideal para o número de *clusters*.

Tabela 5.11: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.71	0.40	0.50	0.67	0.67	1.00	0.87

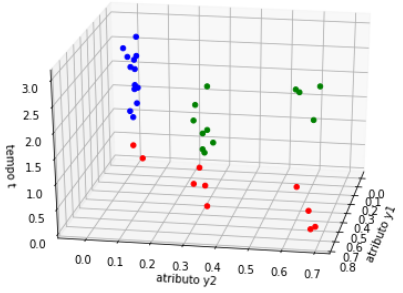


Figura 5.16: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: 3 *clusters*.

Average linkage $K = 2$ parece uma boa opção, pelo que se observa da curva do cotovelo e do dendrograma.

Tabela 5.12: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00

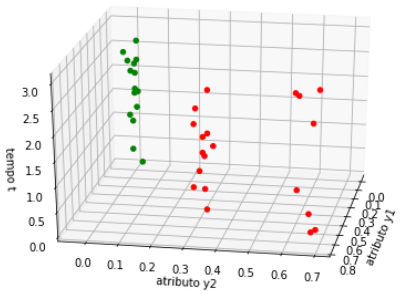
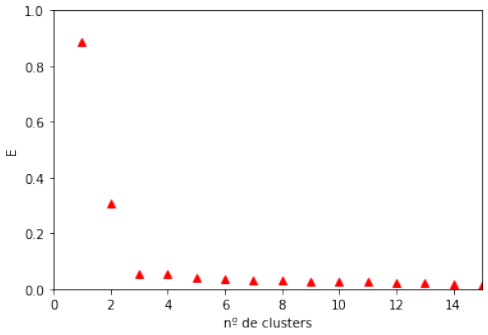


Figura 5.17: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: 2 *clusters*.

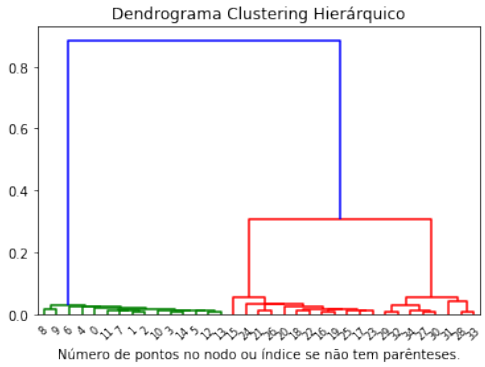
Conclusão O modelo funciona mal nos três casos.

5.4.3 Clustering hierárquico com distância de Manhattan só para os atributos de espaço

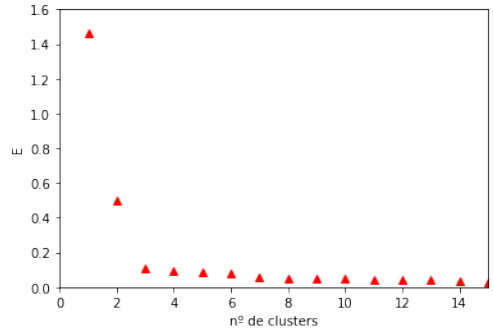
Apresentam-se os resultados para este exemplo nas Figuras 5.18, 5.19, 5.20 e 5.21, e nas Tabelas 5.13, 5.14 e 5.15.



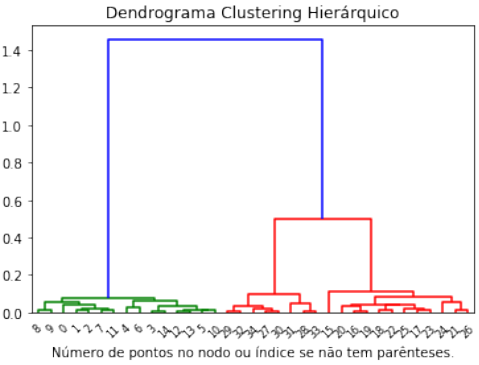
(a) *Single linkage* — curva do cotovelo



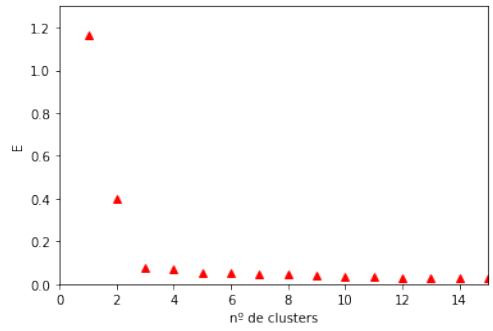
(b) *Single linkage* — dendrograma



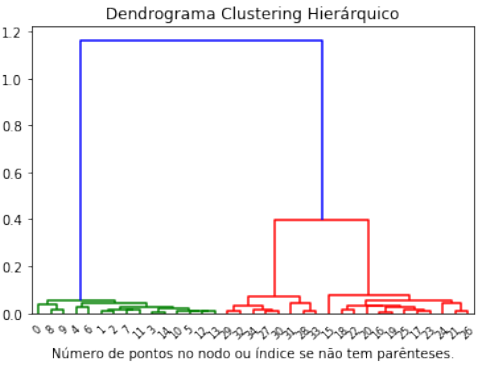
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



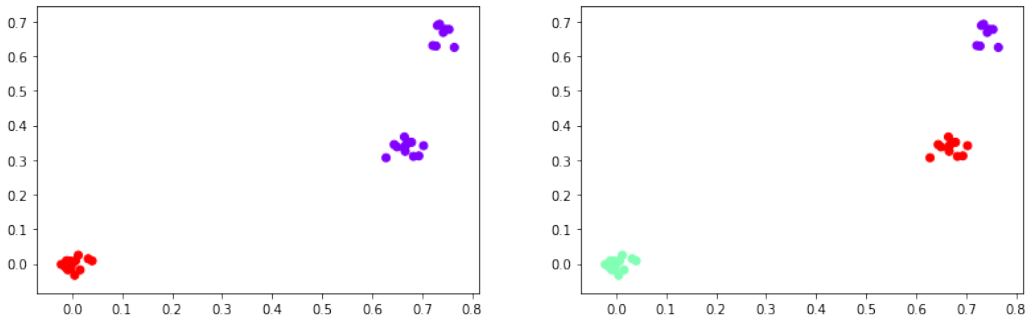
(f) *Average linkage* — dendrograma

Figura 5.18: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço.

Single linkage. Pelo dendrograma e pela curva de cotovelo tanto $K = 2$ como $K = 3$ podem ser o número ideal de *clusters*.

Tabela 5.13: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



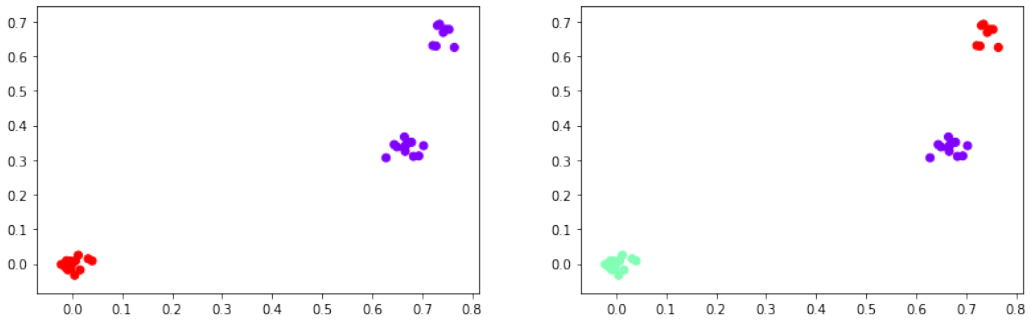
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.19: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*.

Complete linkage Mais uma vez, tanto $K = 2$ como $K = 3$ são números ideais de *clusters*.

Tabela 5.14: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



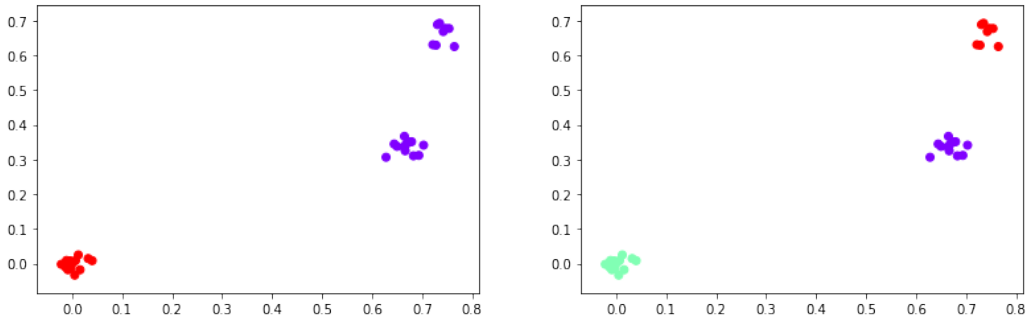
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.20: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*.

Average linkage $K = 2$ e $K = 3$ são um número ideal de *clusters* a ser usado.

Tabela 5.15: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



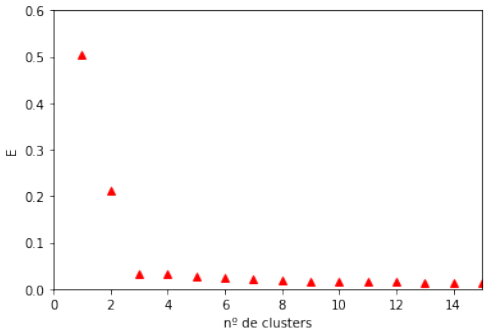
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.21: Exemplo 2 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*.

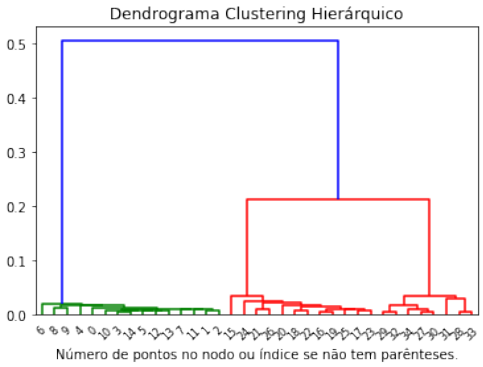
Conclusão Sem o atributo do tempo, o algoritmo consegue encontrar os *clusters* pretendidos, ao contrário de quando se utiliza o atributo t.

5.4.4 Clustering hierárquico com a primeira tentativa de métrica

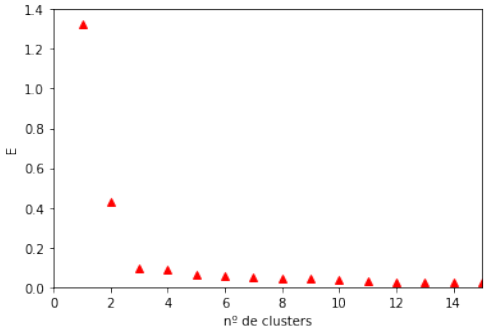
Vai-se considerar para tempo de observação $\bar{t} = 3$ e $\tau = 5$. Apresentam-se os resultados para este exemplo nas Figuras 5.22, 5.23, 5.24 e 5.25, e nas Tabelas 5.16, 5.17 e 5.18.



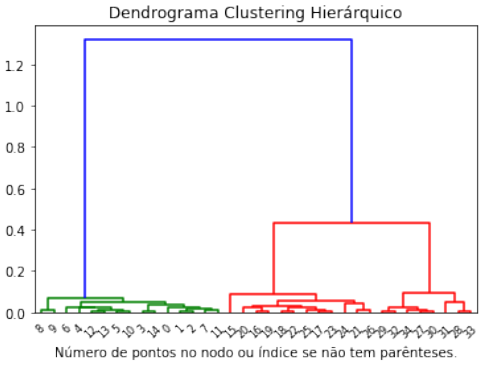
(a) *Single linkage* — curva do cotovelo



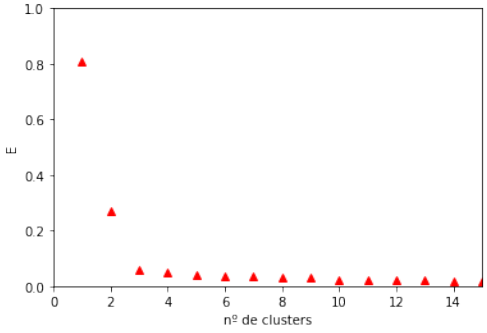
(b) *Single linkage* — dendrograma



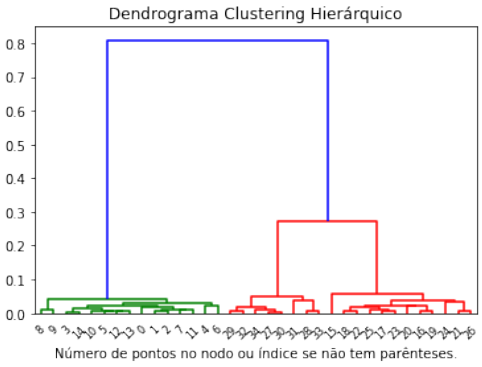
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



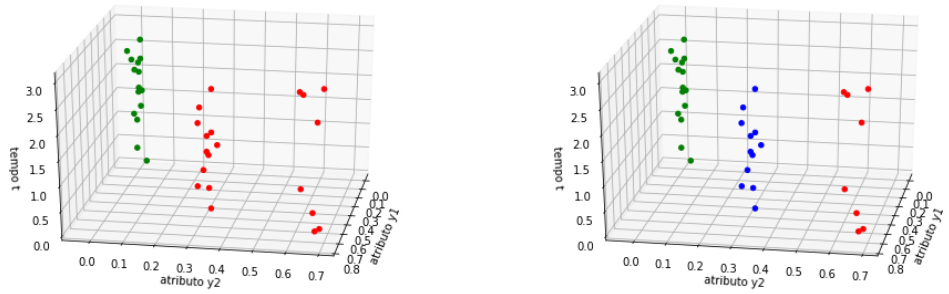
(f) *Average linkage* — dendrograma

Figura 5.22: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica.

Single linkage O número ideal de *clusters* é $K = 2$ ou $K = 3$, pelo que se observa da curva do cotovelo e do dendrograma.

Tabela 5.16: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



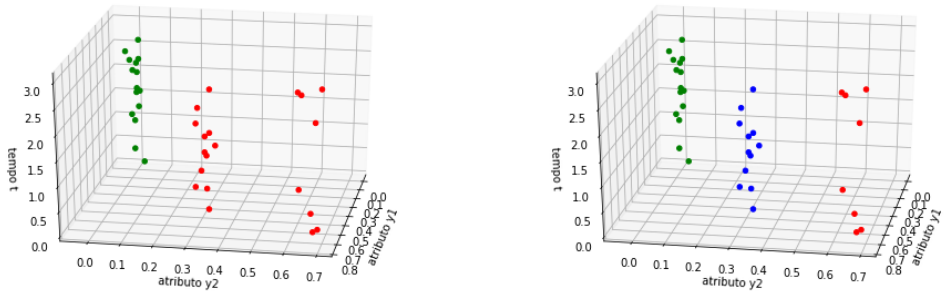
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.23: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*.

Complete linkage $K = 2$ e $K = 3$ parecem ser números ideais de *clusters*.

Tabela 5.17: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) 2 *clusters* (b) 3 *clusters*

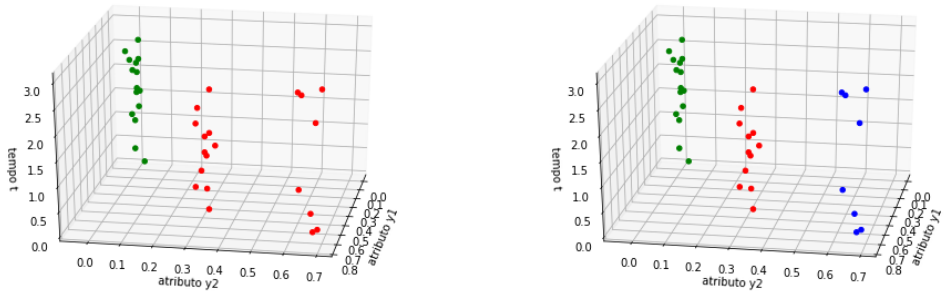
Figura 5.24: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*.

Average linkage

Tanto $K = 2$ como $K = 3$ parecem ser um número ideal de *clusters*.

Tabela 5.18: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.25: Exemplo 2 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*.

Conclusão Quando se usa $K = 3$ a *accuracy* do modelo é melhor, uma vez que acertou em tudo.

5.4.5 Conclusões

Também nesta base de dados se percebe a importância de separar os atributos de espaço do atributo do tempo, visto que usando a métrica criada no Capítulo 4 o modelo acertou totalmente, o que também aconteceu no caso do algoritmo ser usado apenas com os atributos de espaço.

5.5 Exemplo 3

5.5.1 Construção da base de dados

A base de dados é constituída por 35 elementos e 3 *clusters* com 15, 12 e 8 elementos, respetivamente, cada um da forma (y_1, y_2, t). Os pontos de cada *cluster* pertencem a cilindros, sendo que dois deles se intersectam, seguindo as seguintes regras:

- *cluster* 1: $m_1 = 0$, $m_2 = 0$ e $\sigma = \frac{1}{5}$
- *cluster* 2: $m_1 = \frac{2}{3}$, $m_2 = \frac{1}{3}$ e $\sigma = \frac{1}{3}$
- *cluster* 3: $m_1 = \frac{2}{3}$, $m_2 = \frac{1}{4}$ e $\sigma = \frac{1}{4}$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $r = \sigma \mathrm{rand}()$
- $\theta = 2\pi \mathrm{rand}()$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$
- $t = 3\mathrm{rand}()$

Representa-se na Figura 5.26 a base de dados do Exemplo 3.

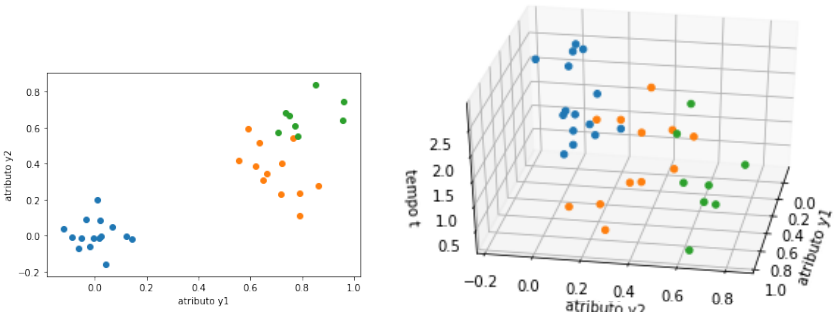
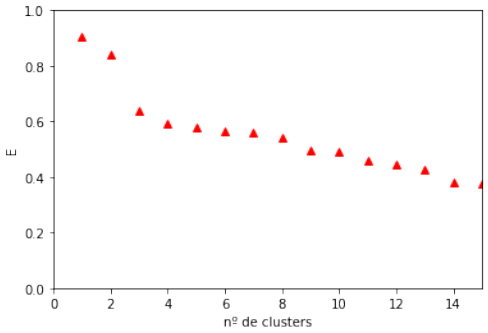


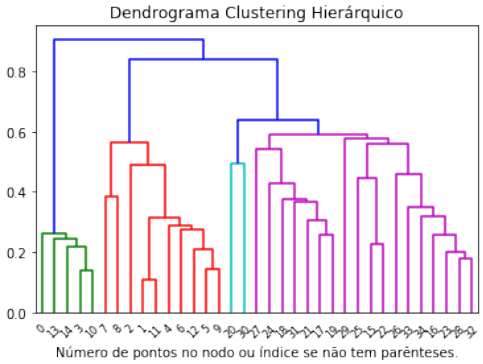
Figura 5.26: Exemplo 3 – visualização 2D e 3D dos dados.

5.5.2 Clustering hierárquico com distância de Manhattan

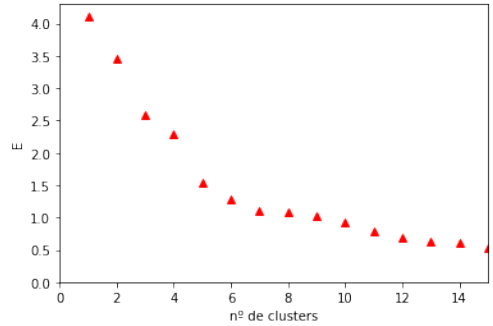
Apresentam-se os resultados para este exemplo nas Figuras 5.27, 5.28, 5.29 e 5.30, e nas Tabelas 5.19, 5.20 e 5.21.



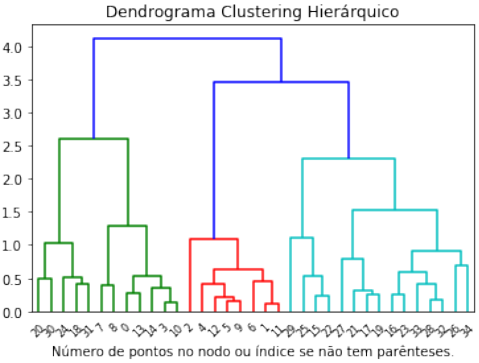
(a) *Single linkage* — curva do cotovelo



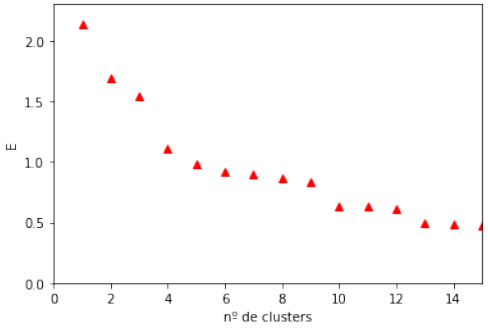
(b) *Single linkage* — dendrograma



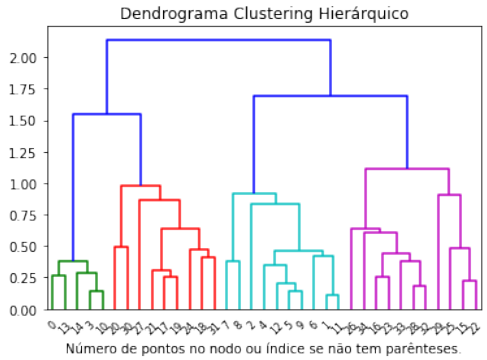
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.27: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan*.

Single linkage Quer considerando a curva do cotovelo, quer o dendrograma, conclui-se que $K = 2$ é o ideal.

Tabela 5.19: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.63	0.60	1.00	0.00	0.00	1.00	0.67

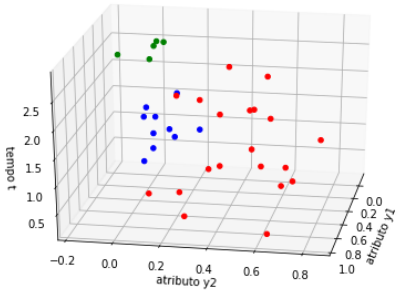


Figura 5.28: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: 3 *clusters*.

Complete linkage Considerando a curva do cotovelo e o dendrograma, $K = 3$ é o valor ideal para o número de *clusters*.

Tabela 5.20: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.54	0.17	0.25	0.60	0.75	1.00	0.53

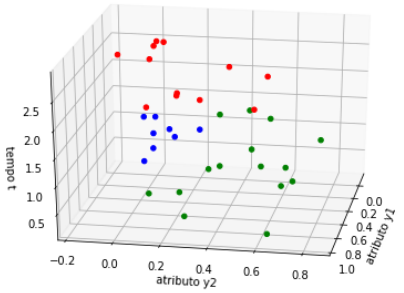


Figura 5.29: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: 3 *clusters*.

Average linkage $K = 2$ parece uma boa opção, pelo que se observa da curva do cotovelo e do dendrograma.

Tabela 5.21: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.46	0.48	0.67	0.43	0.50	0.00	0.00

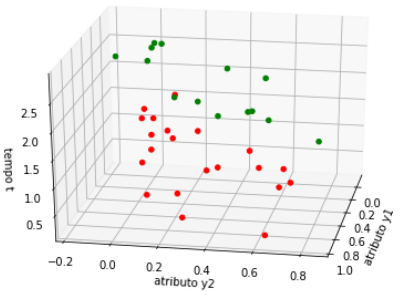
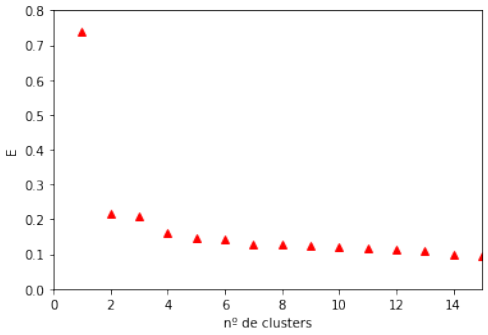


Figura 5.30: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: 2 *clusters*.

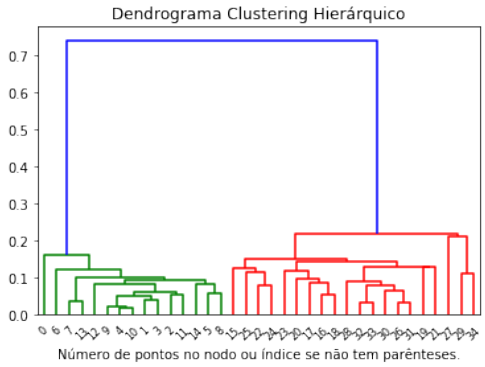
Conclusão O modelo funciona bastante mal nos três casos.

5.5.3 Clustering hierárquico com distância de Manhattan só para os atributos de espaço

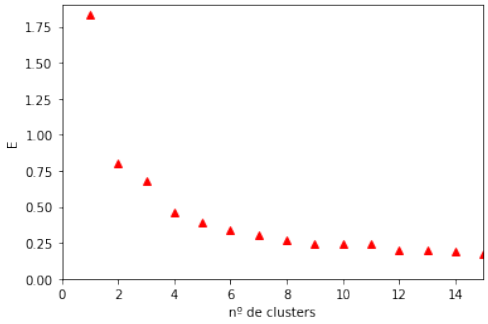
Apresentam-se os resultados para este exemplo nas Figuras 5.31, 5.32, 5.33 e 5.34, e nas Tabelas 5.22, 5.23 e 5.24.



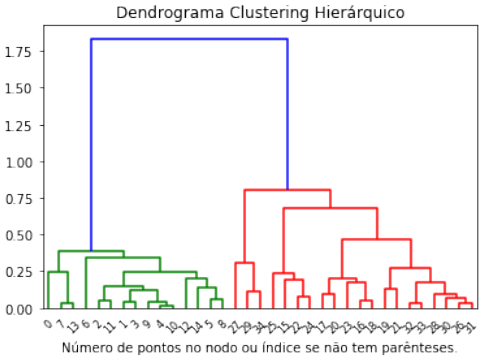
(a) *Single linkage* — curva do cotovelo



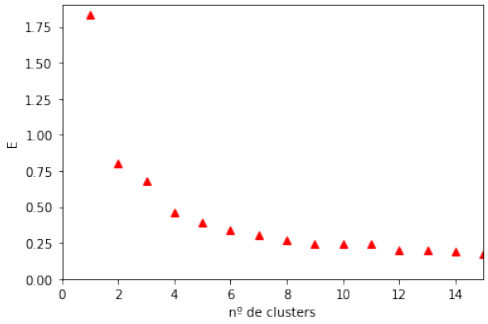
(b) *Single linkage* — dendrograma



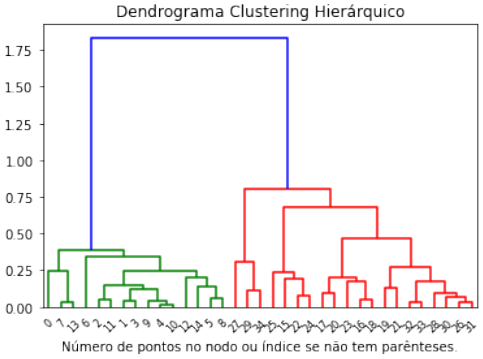
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



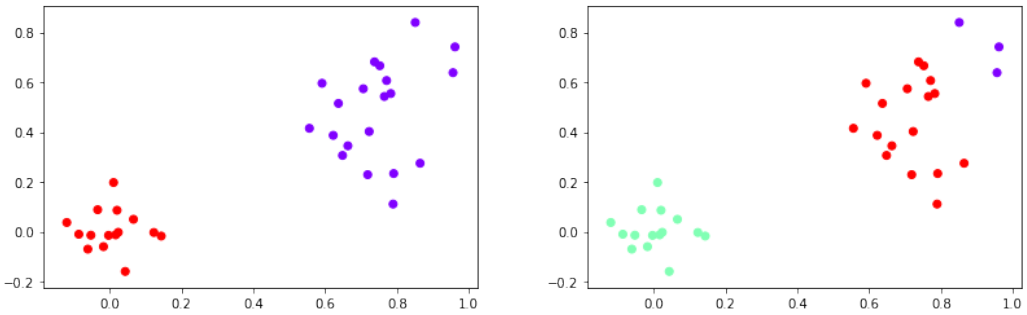
(f) *Average linkage* — dendrograma

Figura 5.31: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço.

Single linkage. Pelo dendrograma e pela curva de cotovelo tanto $K = 2$ como $K = 3$ podem ser o número ideal de *clusters*.

Tabela 5.22: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	0.86	1.00	0.38	1.00	1.00	0.71	1.00



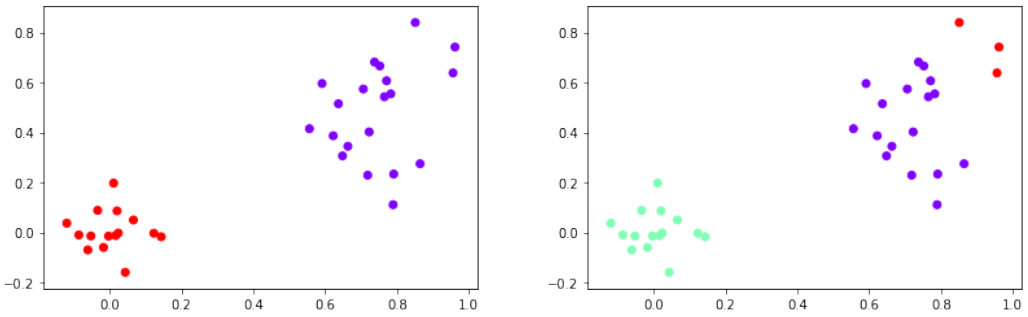
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.32: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*.

Complete linkage Mais uma vez, tanto $K = 2$ como $K = 3$ são números ideais de *clusters*.

Tabela 5.23: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	0.86	0.71	1.00	1.00	1.00	1.00	0.38



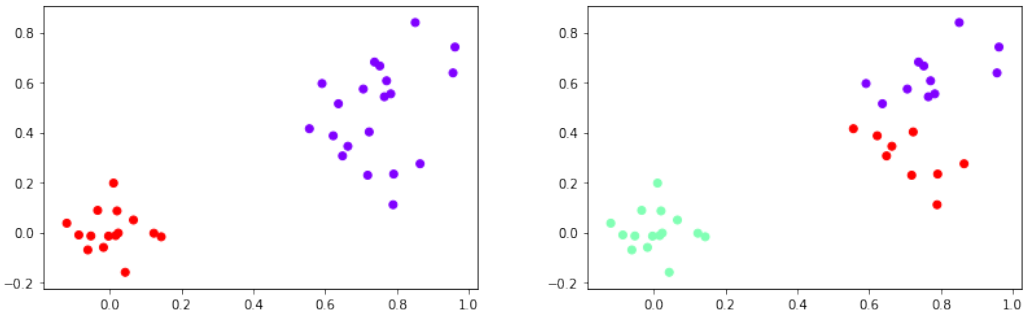
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.33: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*.

Average linkage Mais uma vez, tanto $K = 2$ como $K = 3$ são um número ideal de *clusters*.

Tabela 5.24: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	0.91	0.73	1.00	1.00	1.00	1.00	0.75



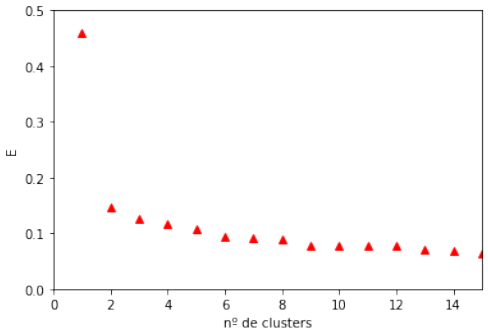
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.34: Exemplo 3 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*.

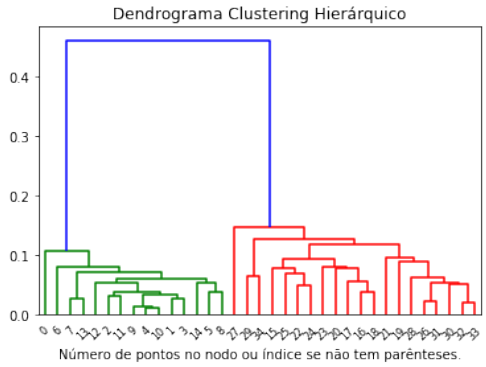
Conclusão Embora o modelo não acerte a 100%, quando se aplica apenas aos atributos de espaço obtém uma *accuracy* razoavelmente boa.

5.5.4 Clustering hierárquico com a primeira tentativa de métrica

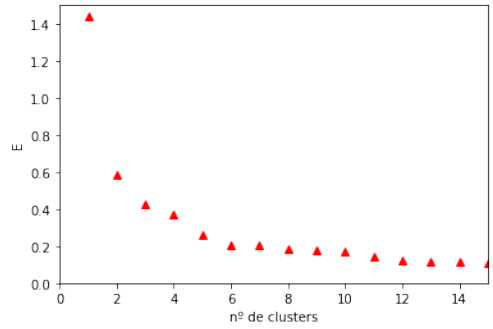
Vai-se considerar para tempo de observação $\bar{t} = 3$ e $\tau = 5$. Apresentam-se os resultados para este exemplo nas Figuras 5.35, 5.36, 5.37 e 5.38, e nas Tabelas 5.25, 5.26 e 5.27.



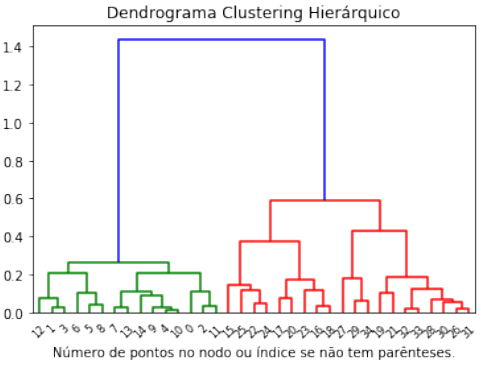
(a) *Single linkage* — curva do cotovelo



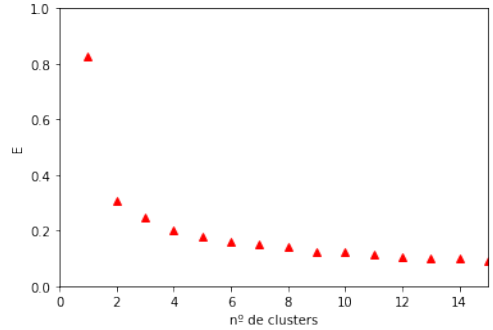
(b) *Single linkage* — dendrograma



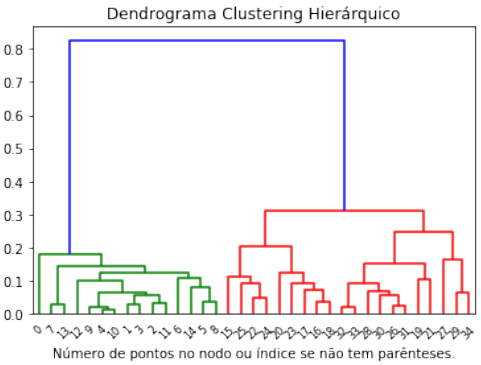
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



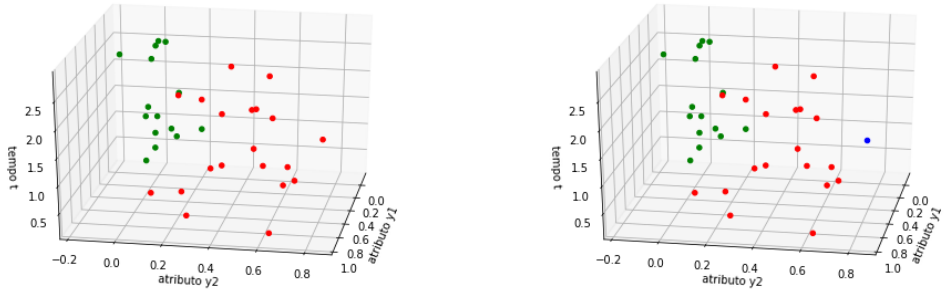
(f) *Average linkage* — dendrograma

Figura 5.35: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica.

Single linkage Embora apenas $K = 2$ pareça ser uma escolha ideal para o número de *clusters*, fez-se também o caso de $K = 3$.

Tabela 5.25: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	0.80	0.63	1.00	1.00	1.00	1.00	0.12



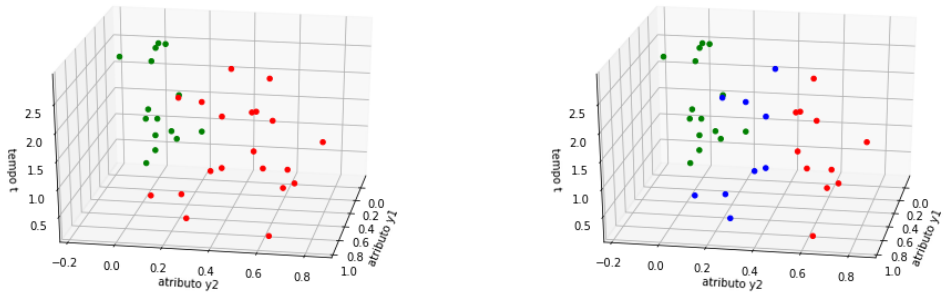
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.36: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*.

Complete linkage Neste caso, também se fez para $K = 2$ e $K = 3$.

Tabela 5.26: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	0.91	0.73	1.00	1.00	1.00	1.00	0.75



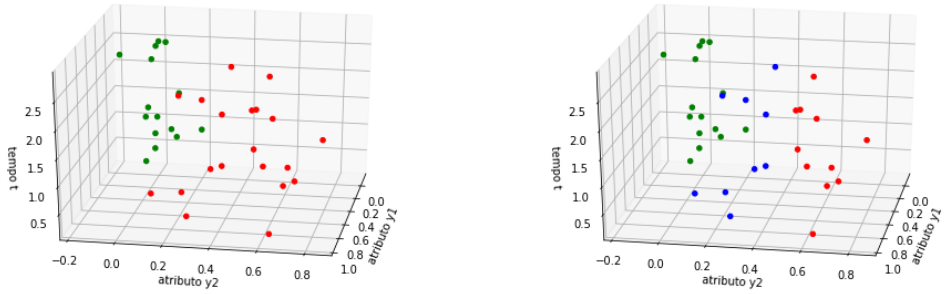
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.37: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*.

Average linkage $K = 2$ e $K = 3$ parecem ser um número ideal de *clusters*.

Tabela 5.27: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.77	0.60	1.00	1.00	1.00	0.00	0.00
3	0.91	0.73	1.00	1.00	1.00	1.00	0.75



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.38: Exemplo 3 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*.

Conclusão Quando se usa $K = 3$ a *accuracy* do modelo é melhor, no entanto não foi perfeita.

5.5.5 Conclusões

O que se pode concluir depois dos modelos aplicados a esta base de dados é que na verdade, o correto é existirem apenas dois *clusters*, uma vez que dois deles se intersectam e a distância ao *cluster* restante é bastante grande. E, mais uma vez, os modelos com melhores resultados foram os que usaram a métrica criada no Capítulo 4, seguidos dos que só usaram os atributos de espaço.

5.6 Exemplo 4

5.6.1 Construção da base de dados

A base de dados é constituída por 65 elementos e 3 *clusters* com 25, 22 e 18 elementos, respetivamente, cada um da forma (y_1, y_2, t). Os pontos de cada *cluster* pertencem a cilindros, sendo que dois deles se intersectam, seguindo as seguintes regras:

- *cluster* 1: $m = (m_1, m_2) = (0.5 + 0.4t, 0.1 + 0.2t)$
- *cluster* 2: $m = (m_1, m_2) = (5.2 - 2.7t, 0.7 + 1.55t)$
- *cluster* 3: $m = (m_1, m_2) = (1.8 + 0.5t, 7.3 + 1.2t)$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $t = 3\text{rand}()$
- $r = 0.5\text{rand}()$
- $\theta = 2\pi\text{rand}()$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$

Representa-se na Figura 5.39 a base de dados do Exemplo 4.

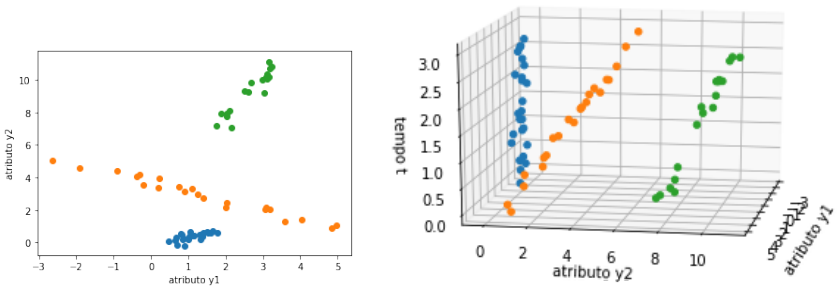
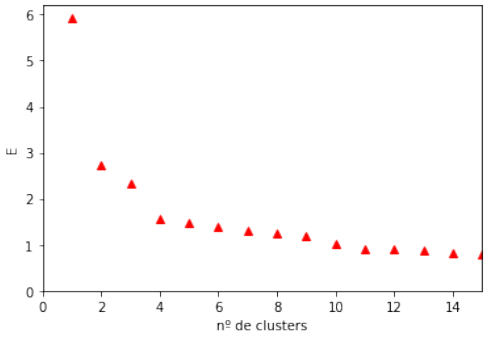


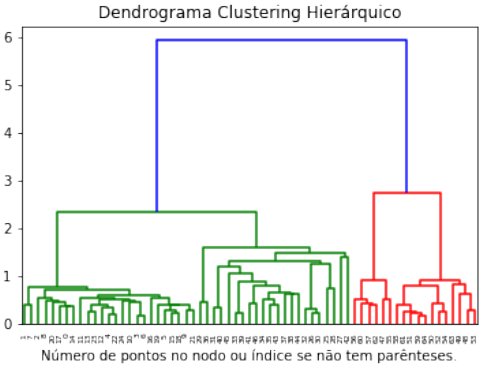
Figura 5.39: Exemplo 4 — visualização 2D e 3D dos dados.

5.6.2 Clustering hierárquico com distância de Manhattan

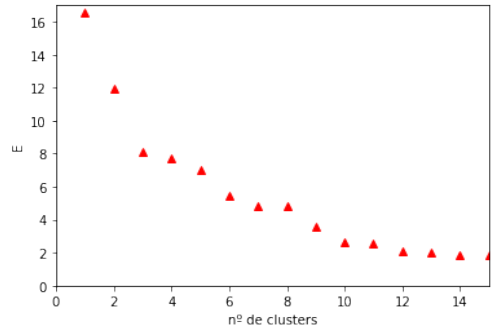
Apresentam-se os resultados para este exemplo nas Figuras 5.40, 5.41, 5.42 e 5.43, e nas Tabelas 5.28, 5.29 e 5.30.



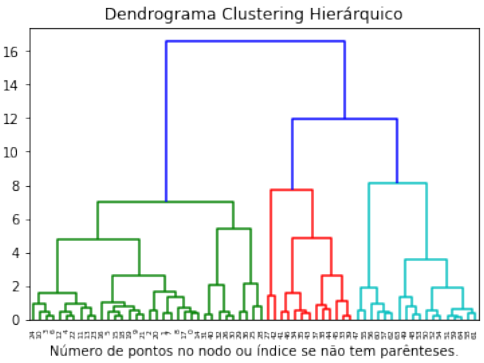
(a) *Single linkage* — curva do cotovelo



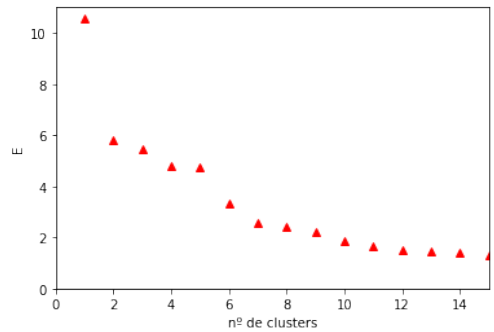
(b) *Single linkage* — dendrograma



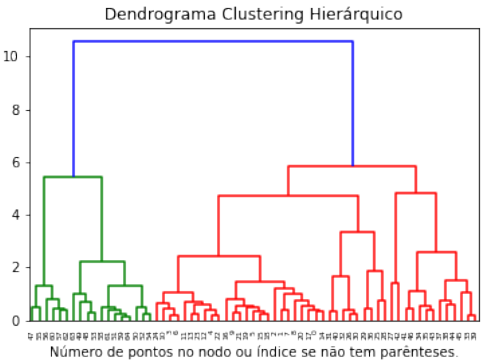
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.40: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan*.

Single linkage Considerando a curva do cotovelo e o dendrograma, conclui-se que $K = 2$ é o ideal.

Tabela 5.28: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.66	1.00	1.00	0.53	1.00	0.00	0.00

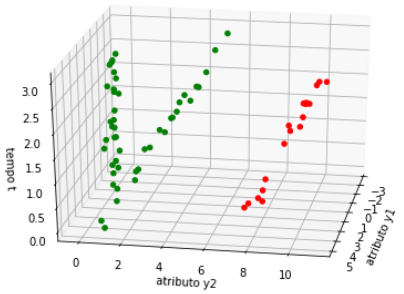
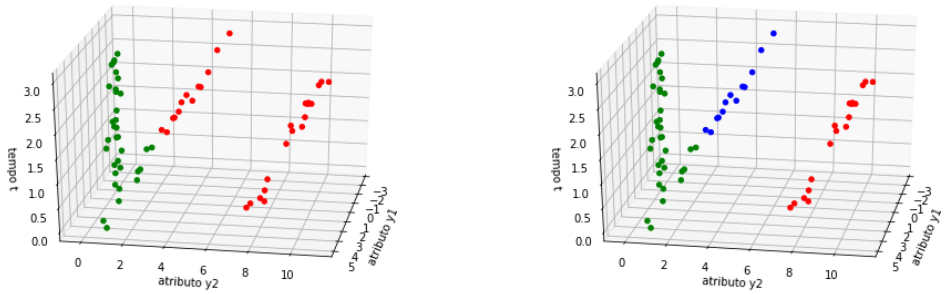


Figura 5.41: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: 2 *clusters*.

Complete linkage Considerando a curva do cotovelo e o dendrograma, $K = 3$ é o valor ideal para o número de *clusters*.

Tabela 5.29: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.66	0.58	1.00	0.74	1.00	0.00	0.00
3	0.86	1.00	1.00	0.74	1.00	1.00	0.59



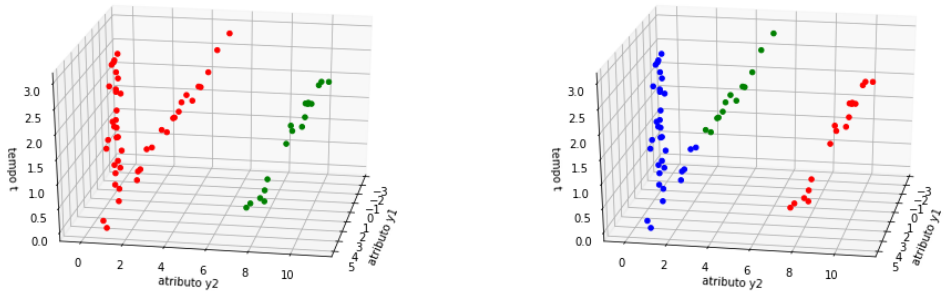
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.42: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*.

Average linkage $K = 2$ parece a melhor opção, no entanto aplicou-se o algoritmo também para $K = 3$..

Tabela 5.30: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.66	0.53	1.00	1.00	1.00	0.00	0.00
3	0.86	1.00	1.00	1.00	0.59	0.74	1.00



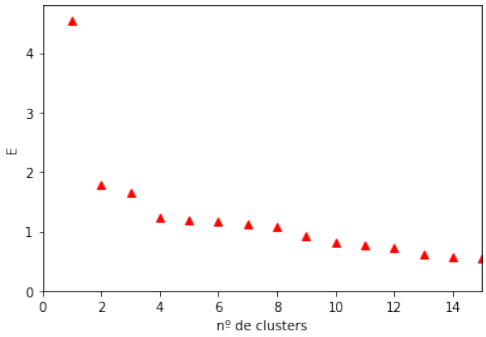
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.43: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*.

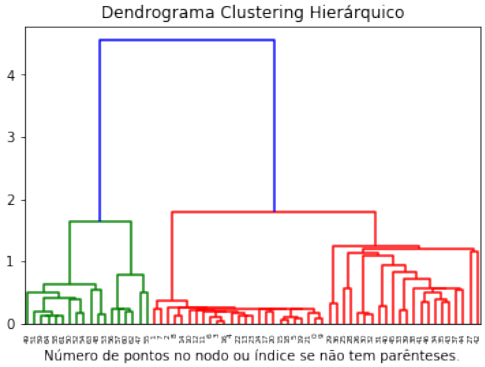
Conclusão O modelo funciona razoavelmente bem com $K = 3$.

5.6.3 Clustering hierárquico com distância de Manhattan só para os atributos de espaço

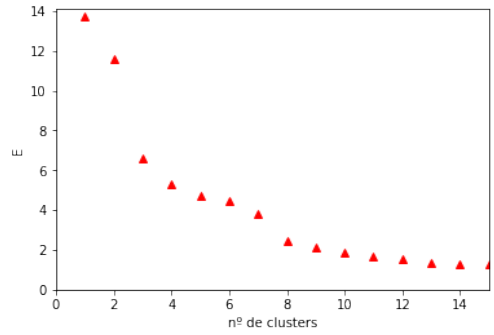
Apresentam-se os resultados para este exemplo nas Figuras 5.44, 5.45, 5.46 e 5.47, e nas Tabelas 5.31, 5.32 e 5.33.



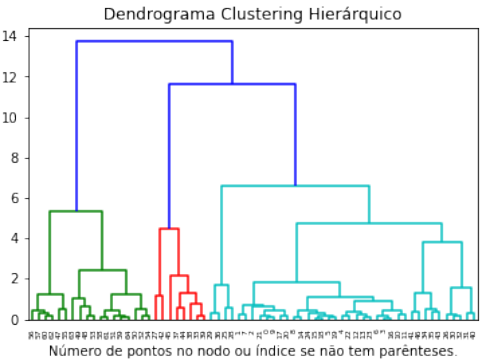
(a) *Single linkage* — curva do cotovelo



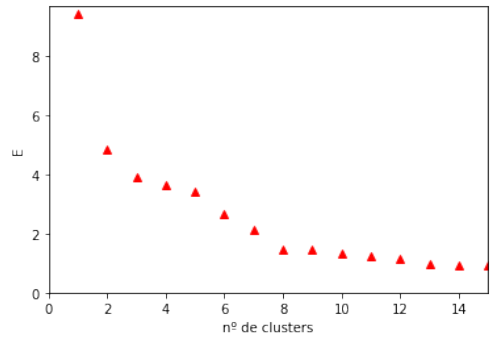
(b) *Single linkage* — dendrograma



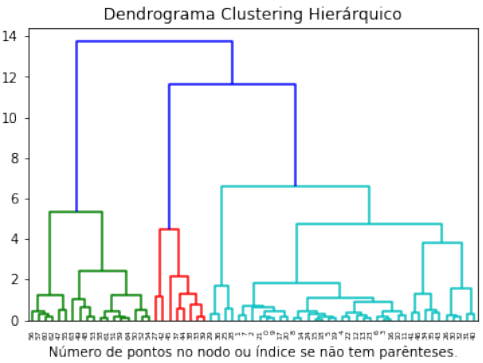
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.44: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço.

Single linkage. $K = 3$ parece ser uma boa opção para o número de *clusters*.

Tabela 5.31: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00

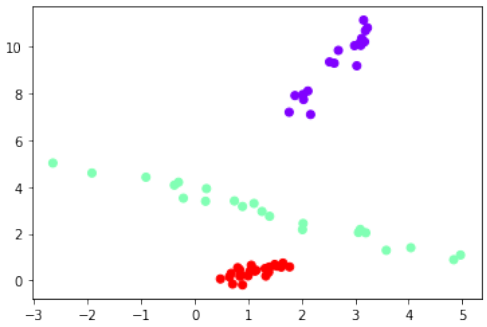


Figura 5.45: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: 3 *clusters*.

Complete linkage Mais uma vez, $K = 3$ é o número ideal de *clusters*.

Tabela 5.32: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.78	0.64	1.00	1.00	1.00	1.00	0.36

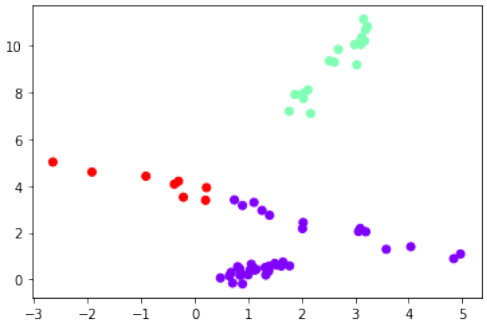


Figura 5.46: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: 3 *clusters*.

Average linkage $K = 3$ é o número ideal de *clusters*.

Tabela 5.33: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.86	1.00	0.59	1.00	1.00	0.74	1.00

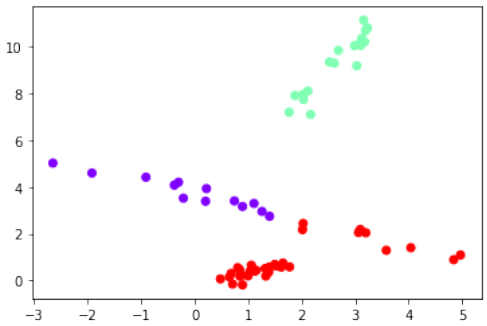
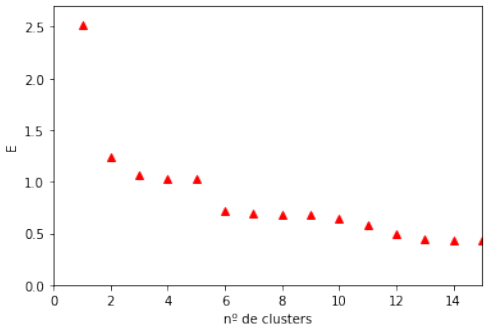


Figura 5.47: Exemplo 4 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: 3 *clusters*.

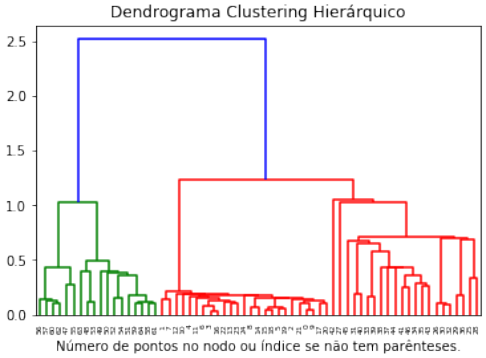
Conclusão Mesmo só aplicando o algoritmo nos atributos de espaço, este não foi capaz de acertar totalmente em todos os casos.

5.6.4 Clustering hierárquico com a primeira tentativa de métrica

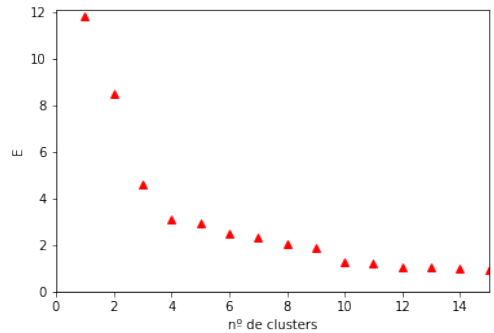
Vai-se considerar para tempo de observação $\bar{t} = 3$ e $\tau = 5$. Apresentam-se os resultados para este exemplo nas Figuras 5.48, 5.49, 5.50 e 5.51, e nas Tabelas 5.34, 5.35 e 5.36.



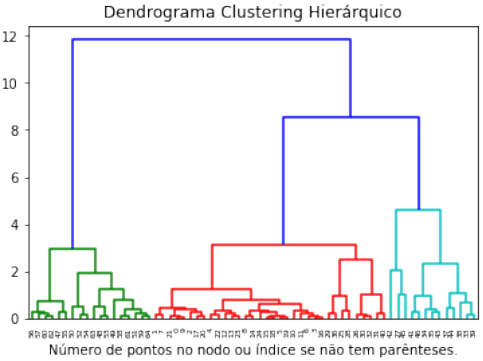
(a) *Single linkage* — curva do cotovelo



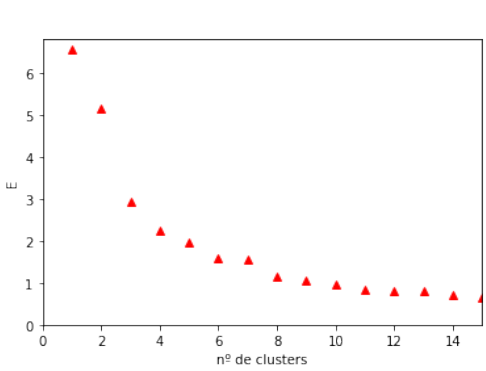
(b) *Single linkage* — dendrograma



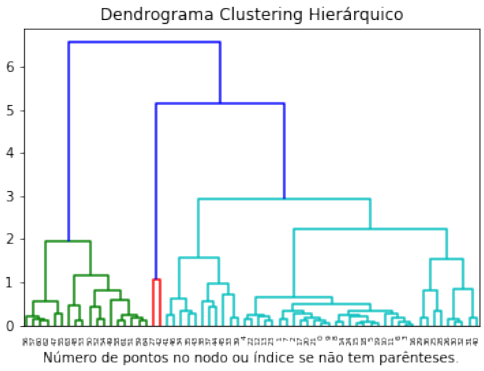
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



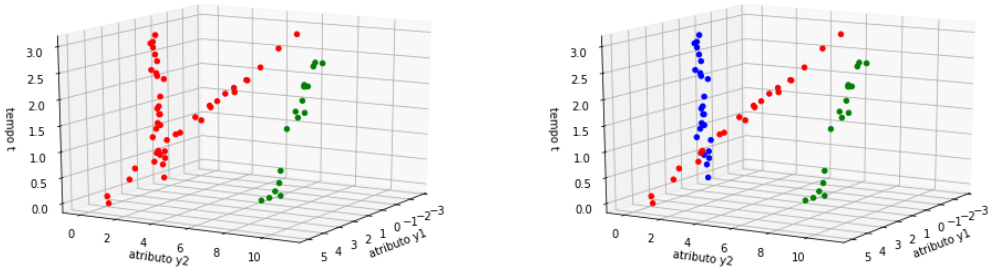
(f) *Average linkage* — dendrograma

Figura 5.48: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica.

Single linkage Embora apenas $K = 2$ pareça ser uma escolha ideal para o número de *clusters*, fez-se também o caso de $K = 3$.

Tabela 5.34: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.66	0.53	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.49: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*.

Complete linkage Neste caso, fez-se apenas para $K = 3$.

Tabela 5.35: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.86	1.00	0.59	1.00	1.00	0.74	1.00

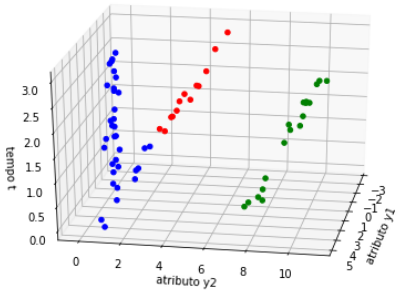


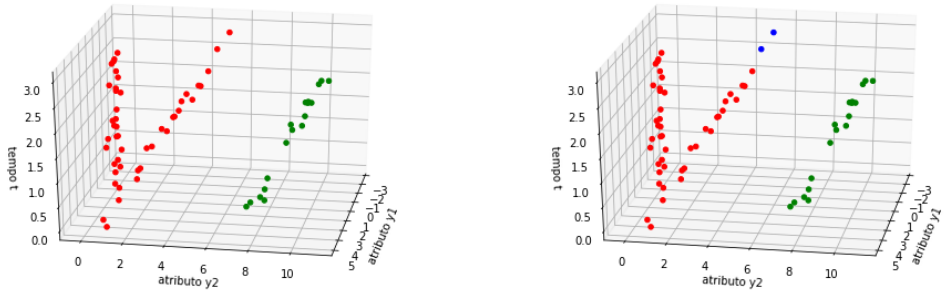
Figura 5.50: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*: 3 *clusters*.

Average linkage

$K = 2$ e $K = 3$ parecem ser um número ideal de *clusters*.

Tabela 5.36: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.66	0.53	1.00	1.00	1.00	0.00	0.00
3	0.69	0.56	1.00	1.00	1.00	1.00	0.09



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.51: Exemplo 4 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*.

Conclusão Apenas o *single linkage* acertou totalmente.

5.6.5 Conclusões

Para esta base de dados, observa-se que quando se usa o modelo com a distância de manhattan só para os atributos de espaço ou o modelo com a métrica criada no Capítulo 4, o método que trabalha melhor é o *single linkage*, sendo que no segundo caso acerta

totalmente. No entanto, quando se aplica o modelo com distância de manhattan para todos os atributos, o *single linkage* é o pior método, mas os outros também não funcionam a 100%. Posto isto, verifica-se que o melhor resultado foi usando a métrica criada no Capítulo 4, embora tenha funcionado bem apenas para um caso, o que provavelmente estará relacionado com o *linkage* usado.

5.7 Exemplo 5

5.7.1 Construção da base de dados

A base de dados é constituída por 3 *clusters* com 400 elementos cada um e da forma (y_1, y_2, t). Os *clusters* seguem as seguintes regras:

- *cluster* 1:
  - $t = \text{rand}()$
  - $\theta = \frac{\text{rand}()2}{\pi}$
  - $r = \frac{ran()}{6}$
  - $x = 4t(1-t)$
  - $y = x$
  - $y_1 = x + r\cos\theta$
  - $y_2 = y + r\sin\theta$
- *cluster* 2:
  - $t = \text{rand}()$
  - $\theta = 2\pi.\text{rand}()$
  - $r = 0.1\text{rand}()$

- $x = t$
- $y = \frac{1}{1+5t}$
- $y_1 = x + r\cos\theta$
- $y_2 = y + r\sin\theta$

- *cluster* 3:
  - $t = 0.4 + 0.1\exp(-5\text{rand}()^2)$
  - $y_1 = 0.5 + 0.1\exp(-5\text{rand}()^2)$
  - $y_2 = 0.5 + 0.1\exp(-5\text{rand}()^2)$

Representa-se na Figura 5.52 a base de dados do Exemplo 5.

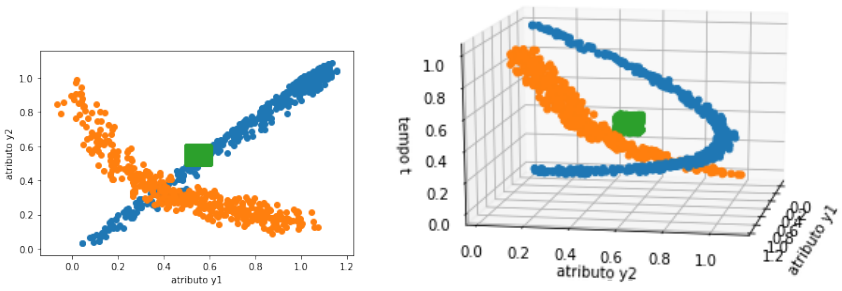
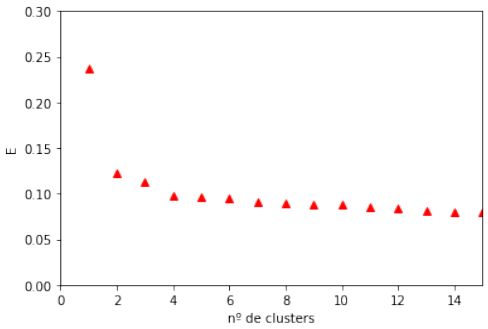


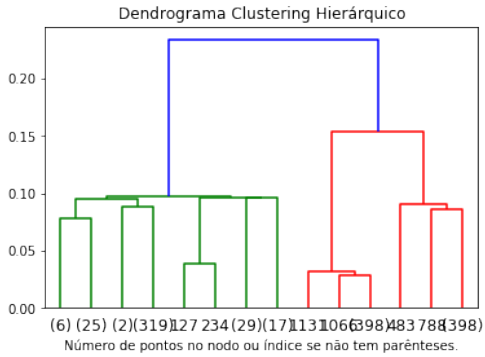
Figura 5.52: Exemplo 5 — visualização 2D e 3D dos dados.

5.7.2 Clustering hierárquico com distância de Manhattan

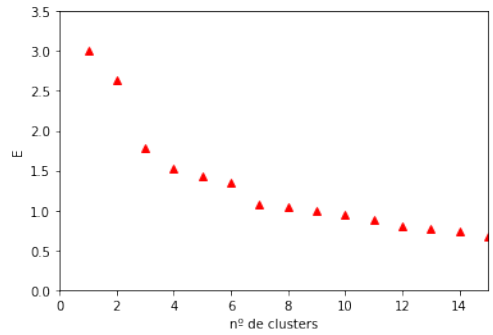
Apresentam-se os resultados para este exemplo nas Figuras 5.53, 5.54, 5.55 e 5.56, e nas Tabelas 5.37, 5.38 e 5.39.



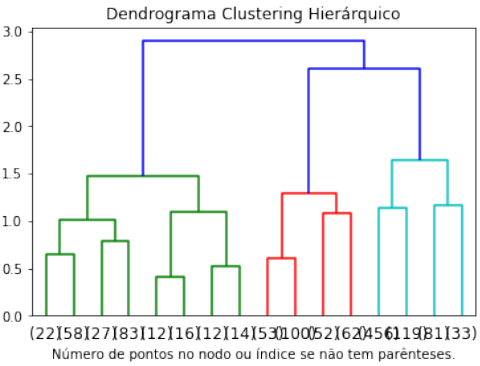
(a) *Single linkage* — curva do cotovelo



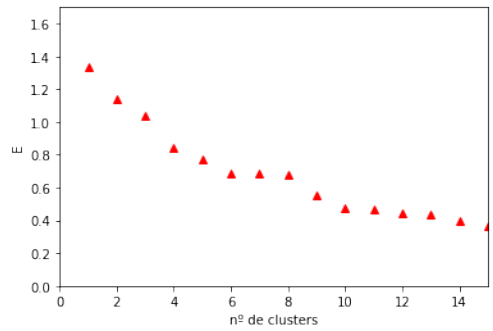
(b) *Single linkage* — dendrograma



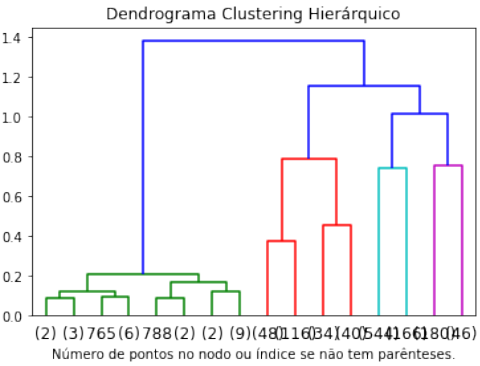
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



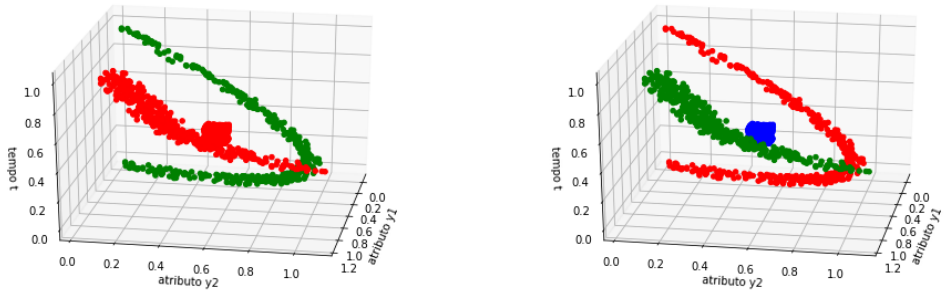
(f) *Average linkage* — dendrograma

Figura 5.53: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan*.

Single linkage Considerando a curva do cotovelo e o dendrograma, $K = 2$ e $K = 3$ parecem ser ambos uma boa opção.

Tabela 5.37: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.67	0.50	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.54: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* e *single linkage*.

Complete linkage Neste caso, apenas $k = 3$ parece uma boa opção.

Tabela 5.38: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
3	0.70	0.58	1.00	0.69	0.42	1.00	0.67

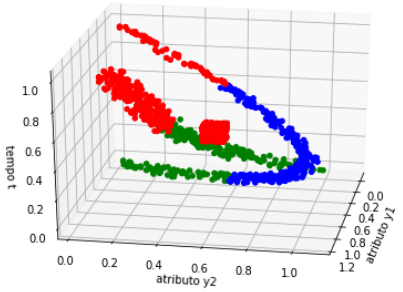
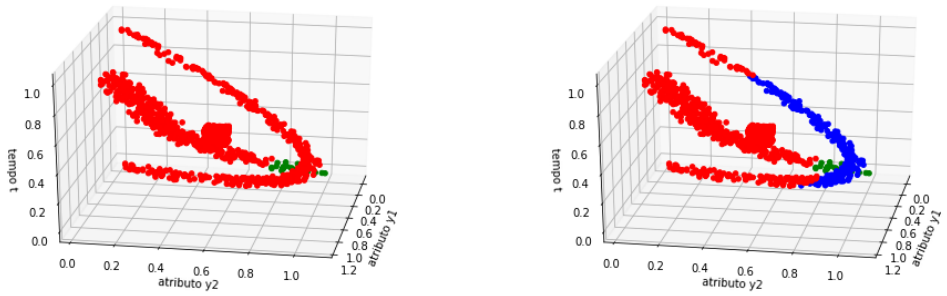


Figura 5.55: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* e *complete linkage*: 3 *clusters*.

Average linkage Neste caso, aplicou-se o algoritmo para $K = 2$ e $K = 3$.

Tabela 5.39: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.36	0.34	1.00	1.00	0.07	0.00	0.00
3	0.55	0.43	1.00	1.00	0.07	1.00	0.59



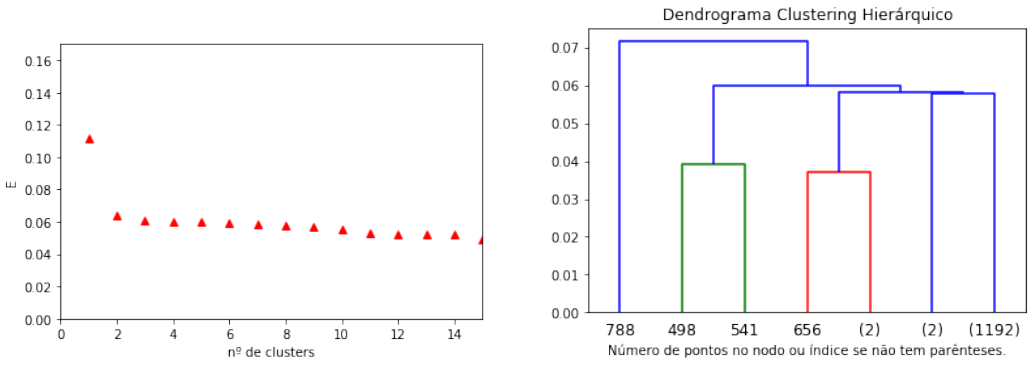
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.56: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* e *average linkage*.

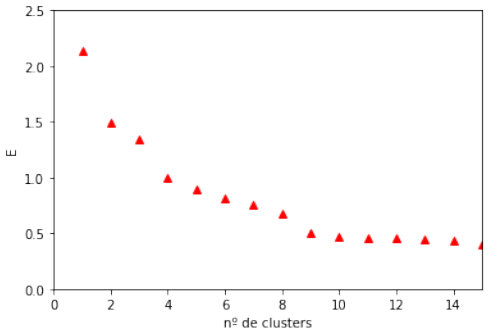
Conclusão O modelo funciona bastante mal.

5.7.3 Clustering hierárquico com distância de Manhattan só para os atributos de espaço

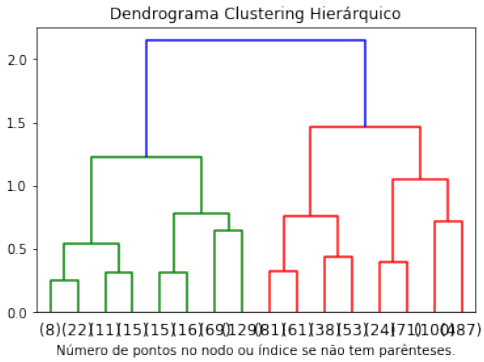
Apresentam-se os resultados para este exemplo nas Figuras 5.57, 5.58, 5.59 e 5.60, e nas Tabelas 5.40, 5.41 e 5.42.



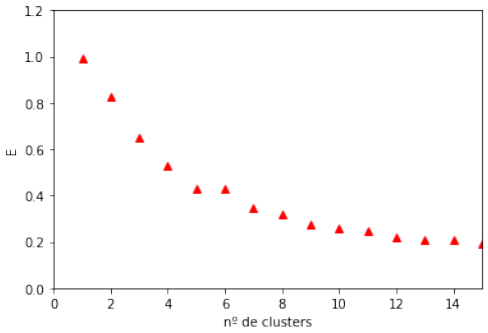
(a) *Single linkage* — curva do cotovelo (b) *Single linkage* — dendrograma



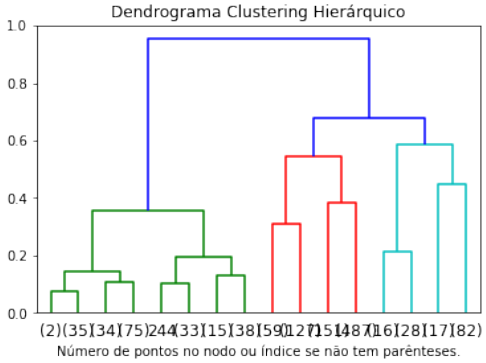
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.57: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço.

Single linkage Aplicou-se o algoritmo apenas para $K = 2$.

Tabela 5.40: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.33	0.33	1.00	1.00	0.00	0.00	0.00

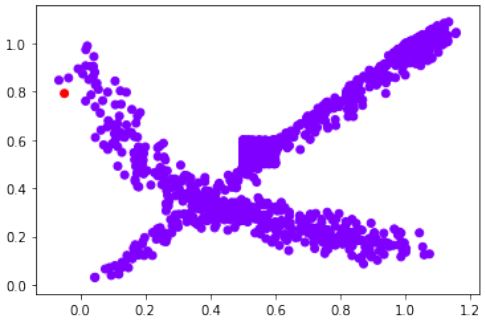
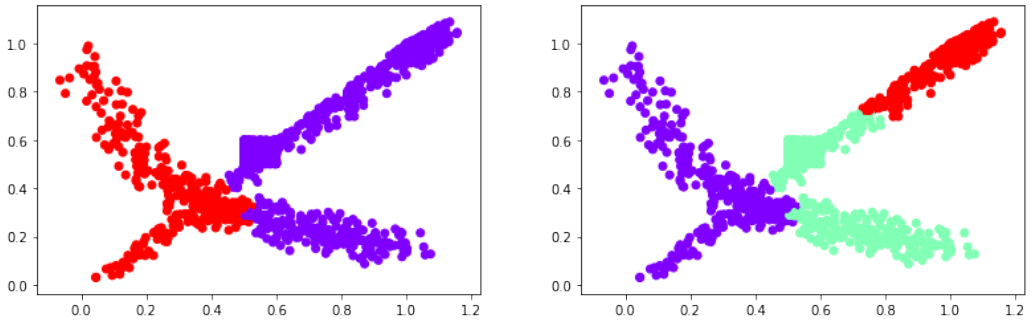


Figura 5.58: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *single linkage*: 2 *clusters*.

Complete linkage Desta vez, aplicou-se para $K = 2$ e $K = 3$.

Tabela 5.41: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.50	0.44	1.00	0.72	0.51	0.00	0.00
3	0.70	0.72	0.51	0.59	1.00	1.00	0.58



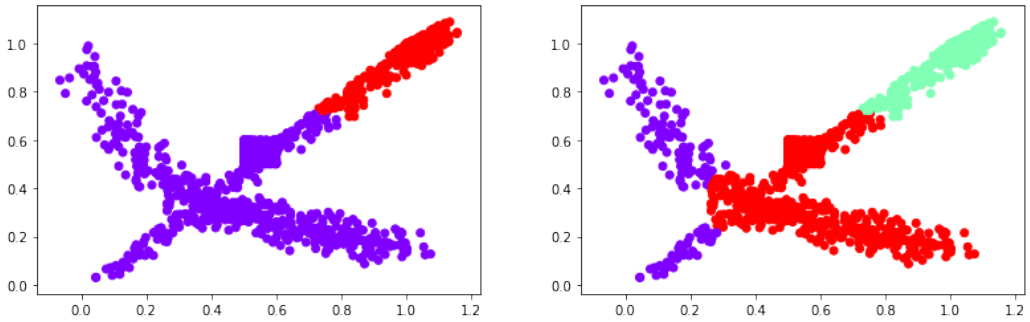
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.59: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *complete linkage*.

Average linkage Neste caso, aplicou-se para $K = 2$ e $K = 3$.

Tabela 5.42: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.53	0.41	1.00	1.00	0.58	0.00	0.00
3	0.61	0.69	0.25	1.00	0.58	0.49	1.00



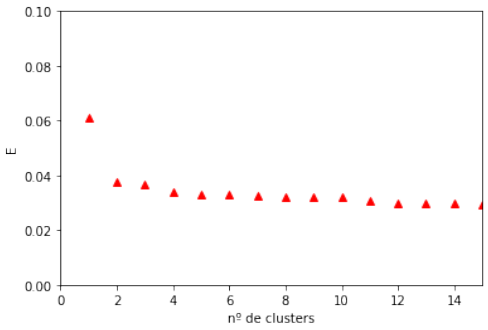
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.60: Exemplo 5 — *Agglomerative Clustering* com distância de *Manhattan* só para os atributos de espaço e *average linkage*.

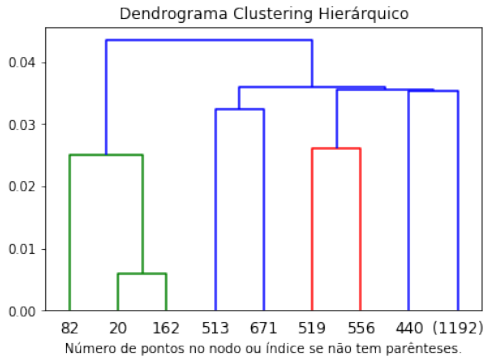
Conclusão Mesmo só aplicando o algoritmo nos atributos de espaço, o modelo não correu da melhor maneira.

5.7.4 Clustering hierárquico com a primeira tentativa de métrica

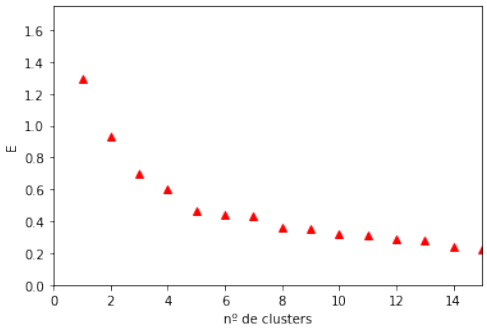
Vai-se considerar para tempo de observação $\bar{t} = 3$ e $\tau = 5$. Apresentam-se os resultados para este exemplo nas Figuras 5.61, 5.62, 5.63 e 5.64, e nas Tabelas 5.43, 5.44 e 5.45.



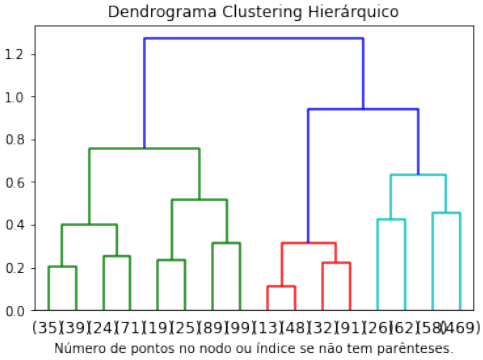
(a) *Single linkage* — curva do cotovelo



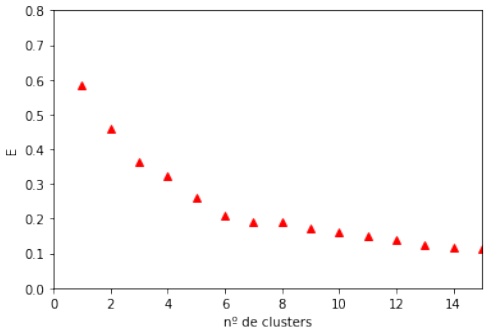
(b) *Single linkage* — dendrograma



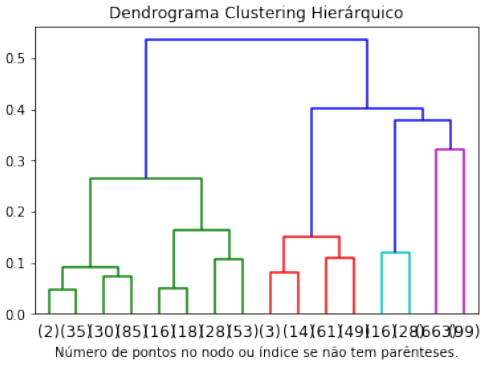
(c) *Complete linkage* — curva do cotovelo



(d) *Complete linkage* — dendrograma



(e) *Average linkage* — curva do cotovelo



(f) *Average linkage* — dendrograma

Figura 5.61: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica.

Single linkage Neste caso, considerou-se apenas o caso de $K = 2$ para o número de *clusters*.

Tabela 5.43: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.34	0.33	1.00	1.00	0.01	0.00	0.00

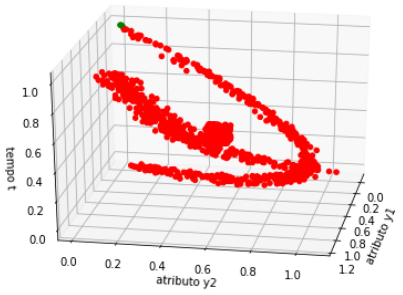
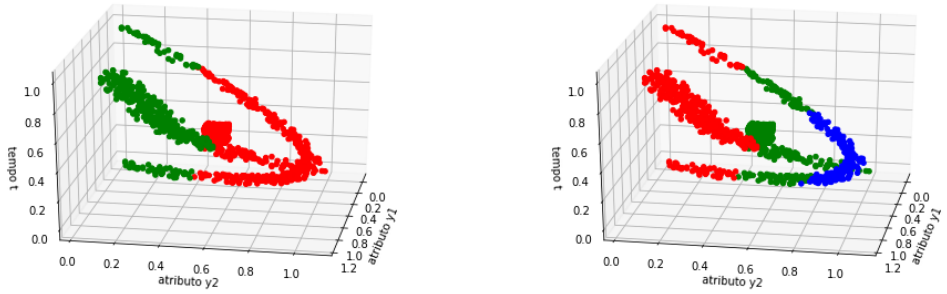


Figura 5.62: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica e *single linkage*: 2 *clusters*.

Complete linkage Neste caso, fez-se para $K = 2$ e $K = 3$.

Tabela 5.44: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.59	0.50	1.00	0.78	0.78	0.00	0.00
3	0.75	0.78	0.78	0.65	1.00	1.00	0.46



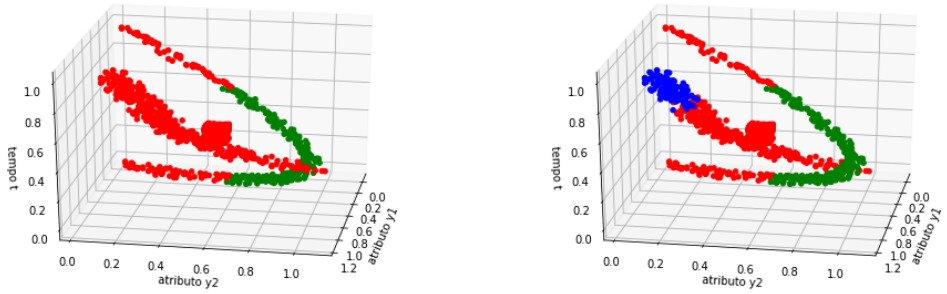
(a) 2 *clusters* (b) 3 *clusters*

Figura 5.63: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica e *complete linkage*.

Average linkage Neste caso, usou-se $K = 2$ e $K = 3$.

Tabela 5.45: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*: métricas de qualidade com K *clusters*.

K		etiqueta 0		etiqueta 1		etiqueta 2	
	accuracy	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
2	0.56	0.43	1.00	1.00	0.67	0.00	0.00
3	0.66	0.50	1.00	1.00	0.67	1.00	0.32



(a) 2 *clusters* (b) 3 *clusters*

Figura 5.64: Exemplo 5 — *Agglomerative Clustering* com a primeira tentativa de métrica e *average linkage*.

Conclusão Os resultados não foram bons.

5.7.5 Conclusões

Para este exemplo, apenas o algoritmo que tem em conta todos os atributos da mesma forma, com o *single linkage* obteve o resultado pretendido. De facto, esta base de dados era mais complexa e daí a probabilidade dos resultados serem bons ser menor. Aqui se percebeu que a métrica criada no Capítulo 4 não estava a funcionar como pretendido, uma vez que falhou em encontrar os *clusters* pretendidos.

5.8 Conclusões

Foi feito um estudo intensivo de várias métricas, incluindo a proposta de métrica definida no Capítulo 4. As primeiras bases de dados obtiveram bons resultados, o que fez pensar que esta métrica era uma boa aposta. Foi elaborada uma base de dados mais sofisticada que fez com que as conclusões preliminares sobre esta métrica mudassem. Conclui-se assim que a aplicação da nova métrica obteve resultados bons apenas para

os dados sintéticos mais simples, ou seja, os que visualmente tinham *clusters* distintos. Verificou-se também que a métrica criada no Capítulo 4 estaria apenas a fazer uma projeção dos elementos no plano, daí os seus resultados comparados com os do algoritmo apenas com os atributos de espaço serem bastante semelhantes. Assim, estes resultados motivaram a uma criação de uma nova alternativa, que é apresentada no capítulo seguinte.

Capítulo 6

O método das faixas temporais

6.1 Princípios

Este capítulo introduz um novo processo para a construção dos *clusters* de uma base de dados em que um dos atributos é o tempo. Seja, então, D uma base de dados constituída por elementos da forma $x = (y, t)$, onde $y = (y_1, y_2)$. Esta nova abordagem divide-se essencialmente em duas partes.

Parte 1

Primeiramente, é necessário dividir a base de dados em faixas ao longo do tempo t que se intersectam. Sejam, assim, $t^m = m\tau$, $m \in \mathbb{N}_0$, e τ a "altura" de cada faixa. Definem-se, então, as faixas primais por

$$F^m = \left\{ x = (y, t) \in D : m\tau - \frac{\tau}{2} \leq t \leq m\tau + \frac{\tau}{2} \right\}, \text{ com } m \in \mathbb{N},$$

e as faixas duais por

$$F^{m+\frac{1}{2}} = \{ x = (y, t) \in D : m\tau \leq t \leq m\tau + \tau \}, \text{ com } m \in \mathbb{N}_0.$$

Tem-se, então, que F^1 inclui a segunda metade entre $t^0 = 0$ e $t^1 = \tau$ e também a primeira metade entre $t^1 = \tau$ e $t^2 = 2\tau$. Tal observa-se na Figura 6.1.

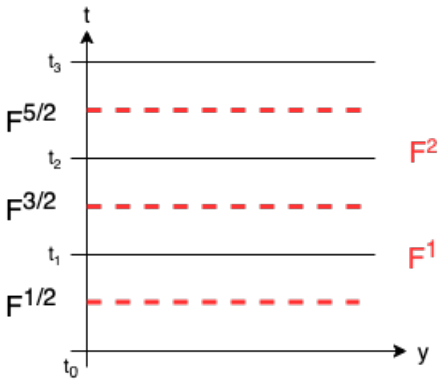


Figura 6.1: Faixas de divisão ao longo do atributo do tempo t.

Para cada faixa é identificado uma partição $P^n = \{C_1^n, C_2^n, \ldots, C_k^n\}$ que é um conjunto de k *clusters*, sendo que n identifica a faixa (primal ou dual). Aplica-se o *clustering* hierárquico aglomerativo com a distância de *Manhattan*, usando todos os atributos, para cada faixa. Um dos critérios definidos é que cada *cluster* numa faixa, deve ter pelo menos 5% dos elementos nessa faixa, de maneira a evitar a existência de *outliers*.

Parte 2

Sempre que novos elementos são adicionados à base de dados não é necessário fazer o *clustering* para todas as faixas, uma vez que já tinha sido feito anteriormente, apenas se faz para a última e penúltima faixa de modo a estabelecer-se uma correspondência entre os *clusters* de uma faixa e outra, sendo que a penúltima já tem correspondência com todas as anteriores.

A correspondência entre os *clusters* de duas faixas seguidas é feita tomando por base uma tabela de intersecção, ou seja, para cada *cluster* de uma faixa $F^{m-\frac{1}{2}}$ faz-se a intersecção com cada *cluster* da faixa seguinte F^m e deste modo consegue saber-se quais os *clusters* que tem mais elementos em comum, estabelecendo-se desta forma a correspondência entre uma faixa primal e uma faixa dual.

Considere-se por exemplo, a tabela de correspondência entre as faixas $F^{\frac{1}{2}}$ e F^1,

ambas com 3 *clusters*, dada na Tabela 6.1.

Tabela 6.1: Tabela de intersecção entre duas faixas.

		Faixa $F^{\frac{1}{2}}$		
		$C_1^{\frac{1}{2}}$	$C_2^{\frac{1}{2}}$	$C_3^{\frac{1}{2}}$
Faixa F^1	C_1^1	7	1	0
	C_2^1	0	2	9
	C_3^1	0	8	0

Como a faixa $F^{\frac{1}{2}}$ é mais antiga que a faixa F^1, então a intersecção tem de ser vista por esta ordem. Desta tabela de intersecção, observa-se que o *cluster* $C_1^{\frac{1}{2}}$ tem mais elementos em comum com o C_1^1, pois $C_1^{\frac{1}{2}} \cap C_1^1 = 7$, enquanto que $C_1^{\frac{1}{2}} \cap C_2^1 = 0$ e $C_1^{\frac{1}{2}} \cap C_3^1 = 0$. O que significa que o *cluster* $C_1^{\frac{1}{2}}$ corresponde ao C_1^1. Do mesmo modo, $C_2^{\frac{1}{2}}$ corresponde ao C_3^1 e $C_3^{\frac{1}{2}}$ corresponde ao C_2^1.

O mesmo é feito para as faixas seguintes. De seguida, para cada faixa calculam-se os centróides de cada *cluster*. Desta maneira, conseguimos saber a trajetória dos *clusters* fazendo a ligação dos seus centróides.

Apresentam-se nas secções seguintes vários exemplos com a abordagem agora apresentada, testando-se novamente os métodos *single linkage*, *complete linkage* e *average linkage*. Note-se que em alguns exemplos aparecem duas versões, isto acontece quando se verificou ser possível obter bons resultados usando números de *clusters* diferentes.

6.2 Exemplo 1

Para este exemplo, foram criados 3 *clusters* em forma de cilindros, que não se intersectam. Cada um deles tem 200 elementos e são da forma (y_1, y_2, t). Os pontos de cada *cluster* seguem as seguintes regras:

- *cluster* 1: $(m_1, m_2) = (0, 0)$
- *cluster* 2: $(m_1, m_2) = (\frac{2}{3}, \frac{1}{3})$
- *cluster* 3: $(m_1, m_2) = (\frac{3}{4}, \frac{2}{3})$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $r = 0.05\text{rand}(0, 1)$
- $\theta = 2\pi\text{rand}(0, 1)$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$
- $t = 3\text{rand}(0, 1)$

Representa-se na Figura 6.2 a base de dados do Exemplo 1.

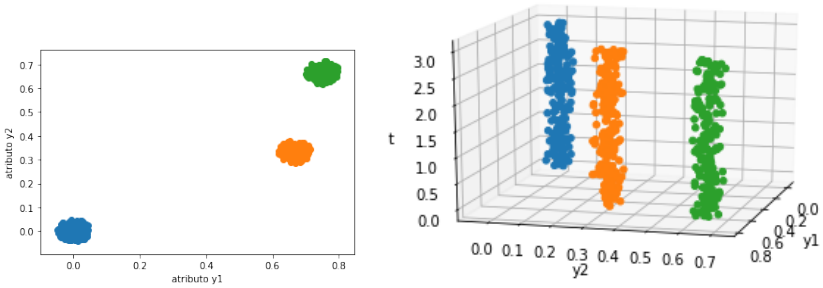


Figura 6.2: Elementos do exemplo 1.

Usou-se $\tau = 0.6$.

Single linkage

Apresentam-se os resultados na Tabela 6.2 e na Figura 6.3.

Tabela 6.2: Correspondência dos *clusters* (*single linkage*) — exemplo 1.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_2
do	C_2	C_2	C_3	C_2	C_3	C_1	C_1	C_3	C_1
Cluster	C_3	C_3	C_2	C_3	C_1	C_3	C_3	C_1	C_3

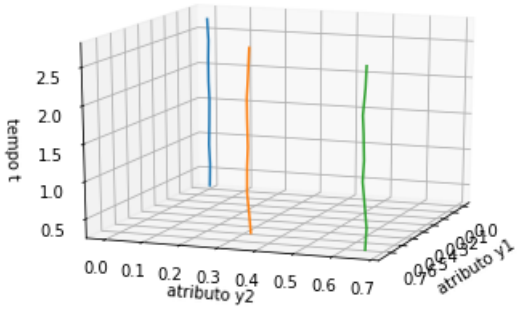


Figura 6.3: Trajetória dos *clusters* (*single linkage*) — exemplo 1.

Complete linkage

Apresentam-se os resultados na Tabela 6.3 e na Figura 6.4.

Tabela 6.3: Correspondência dos *clusters* (*complete linkage*) — exemplo 1.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_1
do	C_2	C_3	C_3	C_3	C_1	C_3	C_3	C_1	C_2
Cluster	C_3	C_1	C_1	C_1	C_3	C_1	C_1	C_3	C_3

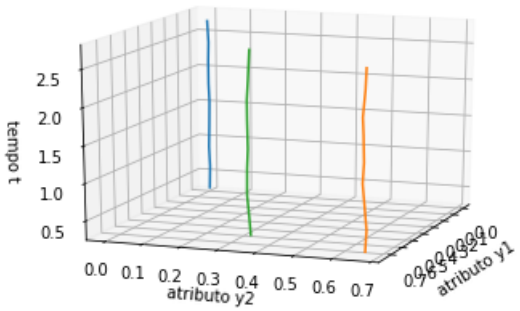


Figura 6.4: Trajetória dos *clusters* (*complete linkage*) — exemplo 1.

Average linkage

Apresentam-se os resultados na Tabela 6.4 e na Figura 6.5.

Tabela 6.4: Correspondência dos *clusters* (*average linkage*) — exemplo 1.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_1
do	C_2	C_3	C_1	C_3	C_1	C_3	C_3	C_1	C_2
Cluster	C_3	C_1	C_3	C_1	C_3	C_1	C_1	C_3	C_3

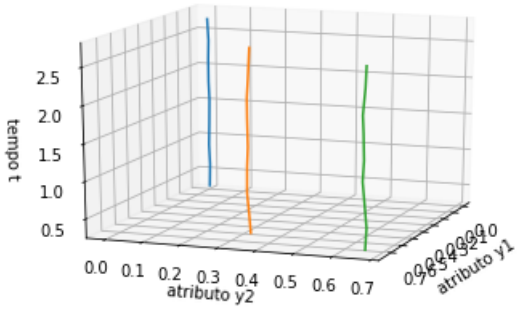


Figura 6.5: Trajetória dos *clusters* (*average linkage*) — exemplo 1.

Conclusão

Esta base de dados era simples — três *clusters* independentes que evoluem de forma constante ao longo do tempo. O objetivo com a estratégia das faixas é verificar isto mesmo, e de facto em cada faixa também foram encontrados três *clusters*.

6.3 Exemplo 2

Para este exemplo, foram criados 3 *clusters*, que não se intersetam, cada um limitado por um retângulo. O primeiro e o segundo *cluster* têm 250 elementos cada e o terceiro 350. Todos os elementos são da forma (y_1, y_2, t). Para cada *cluster* foram calculados y_1 e y_2 aleatórios que seguem as seguintes regras:

- *cluster* 1: $0.1 < y_1 < 0.3$ e $0.6 < y_2 < 0.8$
- *cluster* 2: $0.1 < y_1 < 0.4$ e $0.05 < y_2 < 0.4$
- *cluster* 3: $0.6 < y_1 < 0.9$ e $0.1 < y_2 < 0.7$

Para cada elemento foi calculado também um t aleatório entre 0 e 3.

Representa-se na Figura 6.6 a base de dados do Exemplo 2.

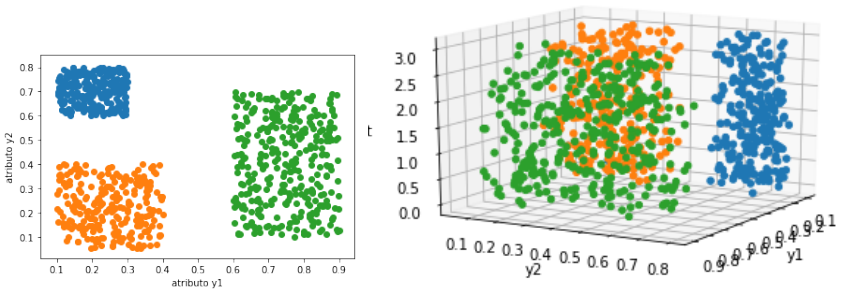


Figura 6.6: Elementos do exemplo 2.

Usou-se $\tau = 0.6$.

Single linkage

Apresentam-se os resultados na Tabela 6.5 e na Figura 6.7.

Tabela 6.5: Correspondência dos *clusters* (*single linkage*) — exemplo 2.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_2	C_2	C_2	C_1	C_1	C_2	C_2	C_2
do	C_2	C_1	C_1	C_3	C_3	C_2	C_1	C_1	C_3
Cluster	C_3	C_2	C_3	C_1	C_2	C_3	C_3	C_3	C_1

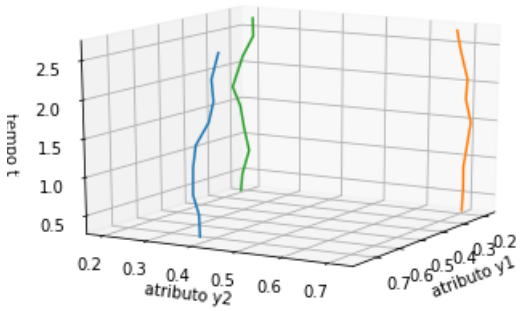


Figura 6.7: Trajetória dos *clusters* (*single linkage*) — exemplo 2.

Complete linkage

Apresentam-se os resultados na Tabela 6.6 e na Figura 6.8.

Tabela 6.6: Correspondência dos *clusters* (*complete linkage*) — exemplo 2.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_2	C_1	C_2	C_1	C_1	C_1	C_1	C_1
do	C_2	C_1	C_2	C_1	C_2	C_2	C_2	C_2	C_2
Cluster	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3

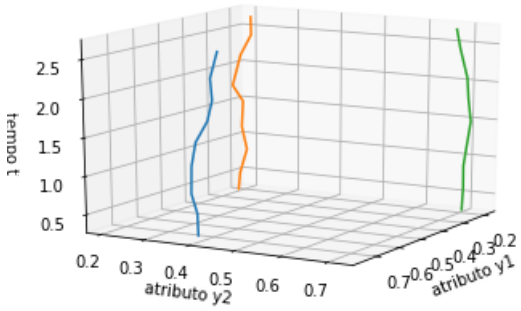


Figura 6.8: Trajetória dos *clusters* (*complete linkage*) — exemplo 2.

Average linkage

Apresentam-se os resultados na Tabela 6.7 e na Figura 6.9.

Tabela 6.7: Correspondência dos *clusters* (*average linkage*) — exemplo 2.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
do	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2
Cluster	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3

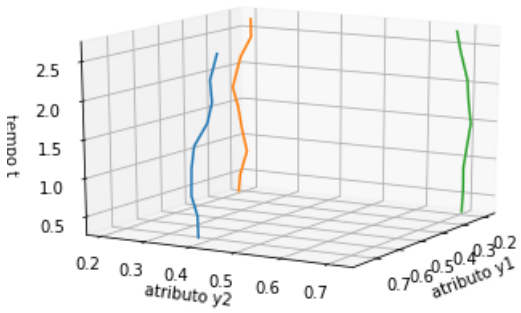


Figura 6.9: Trajetória dos *clusters* (*average linkage*) — exemplo 2.

Conclusão

Esta base de dados também era simples, uma vez que os três *clusters* estão bem separados, embora dispersos. Mais uma vez, o objetivo da estratégia das faixas era encontrar estes três *clusters*, o que se verificou bastante bem.

6.4 Exemplo 3

Para este exemplo, foram criados três *cluster*, cada um pertencente a um cilindro disperso, sendo que dois dos cilindros se intersetam. O primeiro *cluster* tem 155 elementos, o segundo 152 e o terceiro 158 e são da forma (y_1, y_2, t). Os pontos de cada *cluster* seguem as seguintes regras:

- *cluster* 1: $(m_1, m_2) = (0, 0)$ e $\sigma = \frac{1}{5}$
- *cluster* 2: $(m_1, m_2) = (\frac{2}{3}, \frac{1}{3})$ e $\sigma = \frac{1}{3}$
- *cluster* 3: $(m_1, m_2) = (\frac{3}{4}, \frac{2}{3})$ e $\sigma = \frac{1}{4}$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $r = \sigma \mathrm{rand}(0, 1)$
- $\theta = 2\pi \mathrm{rand}(0, 1)$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$
- $t = 3\mathrm{rand}(0, 1)$

Representa-se na Figura 6.10 a base de dados do Exemplo 3.

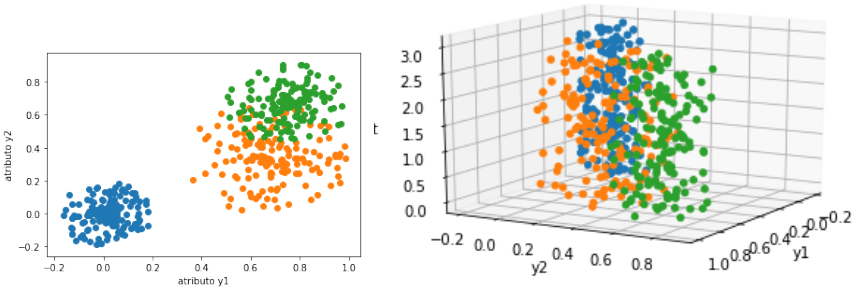


Figura 6.10: Elementos do exemplo 3.

Usou-se $\tau = 0.6$.

Single linkage

Apresentam-se os resultados na Tabela 6.8 e na Figura 6.11.

Tabela 6.8: Correspondência dos *clusters* (*single linkage*) — exemplo 3.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	$C1$
do Cluster	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2

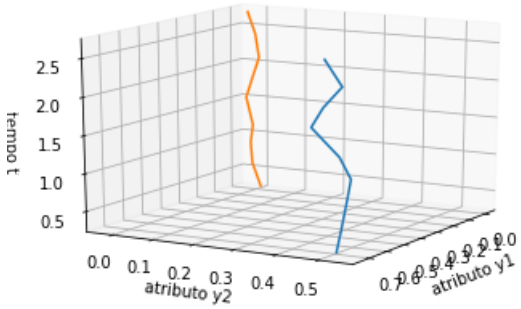


Figura 6.11: Trajetória dos *clusters* (*single linkage*) — exemplo 3.

Complete linkage

1ª versão

Apresentam-se os resultados na Tabela 6.9 e na Figura 6.12.

Tabela 6.9: Correspondência dos *clusters* (*complete linkage* 1ª versão) — exemplo 3.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_3	C_1
do	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2
Cluster	C_3	C_3	C_1	C_3	C_3	C_3	C_3	C_1	C_3

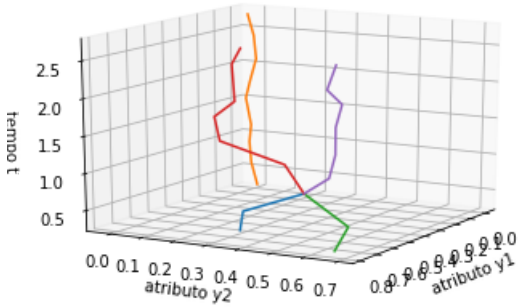


Figura 6.12: Trajetória dos *clusters* (*complete linkage* 1ª versão) — exemplo 3.

2ª versão

Apresentam-se os resultados na Tabela 6.10 e na Figura 6.13.

Tabela 6.10: Correspondência dos *clusters* (*complete linkage* 2ª versão) — exemplo 3.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	$C1$
do Cluster	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2

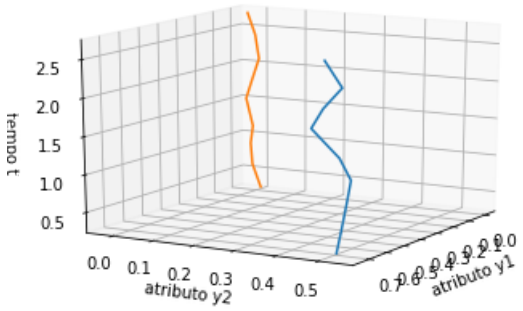


Figura 6.13: Trajetória dos *clusters* (*complete linkage* 2ª versão) — exemplo 3.

Average linkage

1ª versão

Apresentam-se os resultados na Tabela 6.11 e na Figura 6.14.

Tabela 6.11: Correspondência dos *clusters* (*average linkage* 1ª versão) — exemplo 3.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_3	C_1	C_1	C_3	C_1
do	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2
Cluster	C_3	C_1	C_3	C_3	C_1	C_3	C_3	C_1	C_1

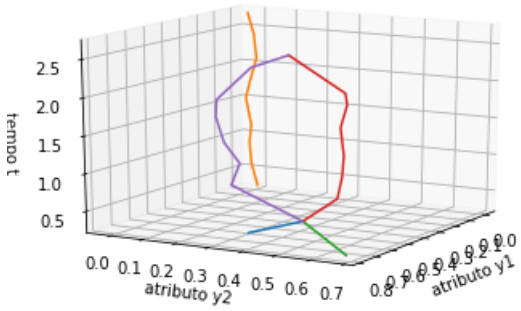


Figura 6.14: Trajetória dos *clusters* (*average linkage* 1ª versão) — exemplo 3.

2ª versão

Apresentam-se os resultados na Tabela 6.12 e na Figura 6.15.

Tabela 6.12: Correspondência dos *clusters* (*average linkage* 2ª versão) — exemplo 3.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	$C1$
do Cluster	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2

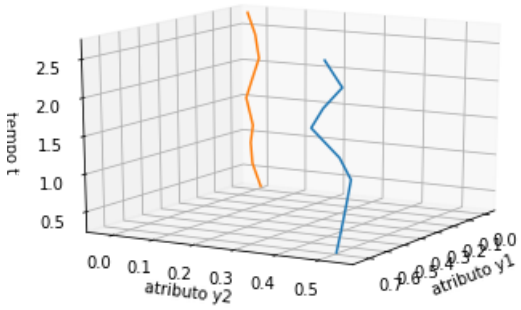


Figura 6.15: Trajetória dos *clusters* (*average linkage* 2ª versão) — exemplo 3.

Conclusão

Nesta base de dados, embora artificialmente se tenham criado três *clusters*, quando se observa a projeção parece haver apenas dois *clusters*. Neste caso, o objetivo da estratégia das faixas foi verificar realmente quantos *clusters* existem e de facto, verifica-se que na realidade o mais correto é haver apenas dois. No entanto algumas faixas também conseguiram encontrar três, só que não acontecia para todas as faixas.

6.5 Exemplo 4

Para este exemplo, foram criados três *cluster*, o primeiro com 85 elementos, o segundo com 82 e o terceiro com 88 e os elementos são da forma (y_1, y_2, t). Os pontos de cada *cluster* seguem as seguintes regras:

- *cluster* 1: $(m_1, m_2) = (0.5 + 0.4t, 0.1 + 0.2t)$
- *cluster* 2: $(m_1, m_2) = (5.2 - 2.7t, 0.7 + 1.55t)$
- *cluster* 3: $(m_1, m_2) = (1.8 + 0.5t, 7.3 + 1.2t)$

Para cada elemento de cada *cluster*, calculou-se:

- $t = 3\mathrm{rand}(0, 1)$
- $r = 0.5\mathrm{rand}(0, 1)$
- $\theta = 2\pi\mathrm{rand}(0, 1)$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$

Representa-se na Figura 6.16 a base de dados do Exemplo 4.

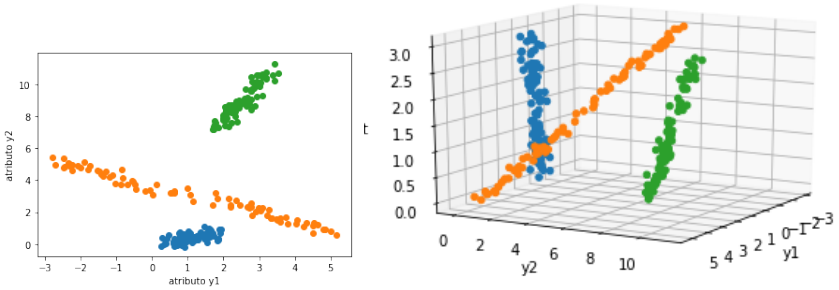


Figura 6.16: Elementos do exemplo 4.

Usou-se $\tau = 0.6$.

Single linkage

Apresentam-se os resultados na Tabela 6.13 e na Figura 6.17.

Tabela 6.13: Correspondência dos *clusters* (*single linkage*) — exemplo 4.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_3	C_1	C_1	C_1	C_1	C_1	C_2	C_2
do	C_2	C_1	C_2	C_2	C_2	C_2	C_2	C_1	C_1
Cluster	C_3	C_2	C_3	C_3	C_3	C_3	C_3	C_3	C_3

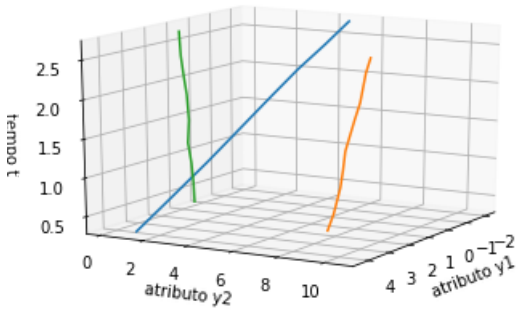


Figura 6.17: Trajetória dos *clusters* (*single linkage*) — exemplo 4.

Complete linkage

Apresentam-se os resultados na Tabela 6.14 e na Figura 6.18.

Tabela 6.14: Correspondência dos *clusters* (*complete linkage*) — exemplo 4.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
do	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2
Cluster	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3

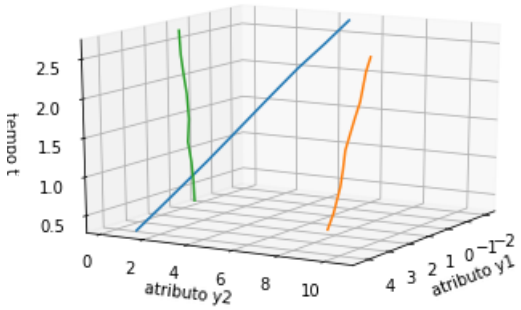


Figura 6.18: Trajetória dos *clusters* (*complete linkage*) — exemplo 4.

Average linkage

Apresentam-se os resultados na Tabela 6.15 e na Figura 6.19.

Tabela 6.15: Correspondência dos *clusters* (*average linkage*) — exemplo 4.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
do	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2
Cluster	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3

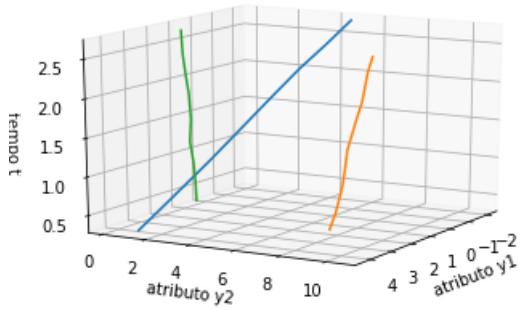


Figura 6.19: Trajetória dos *clusters* (*average linkage*) — exemplo 4.

Conclusão

Nesta base de dados também foram criados 3 *clusters* que não se intersectam. Com a abordagem das faixas também foram encontrados 3 *clusters*. No entanto, isto só foi possível devido ao valor de τ escolhido, se tivesse sido escolhido um valor de τ maior, talvez houvesse alguma ou algumas faixas onde existiriam apenas dois *clusters*, uma vez que pela projeção podemos ver que o *cluster* laranja está perto do azul.

6.6 Exemplo 5

Para este exemplo, foram criados três *cluster*, cada um com 400 elementos da forma (y_1, y_2, t). Os pontos de cada *cluster* seguem as seguintes regras:

- *cluster* 1:
  - $t = \text{rand}()$
  - $\theta = \frac{\text{rand}()2}{\pi}$
  - $r = \frac{\text{rand}}{6}$
  - $x = 4t(1-t)$
  - $y = x$

  - $y_1 = x + r\cos\theta$
  - $y_2 = y + r\sin\theta$

- *cluster* 2:
  - $t = \text{rand}()$
  - $\theta = 2\pi\text{rand}()$
  - $r = 0.1\text{rand}()$
  - $x = t$
  - $y = \frac{1}{1+5t}$
  - $y_1 = x + r\cos\theta$
  - $y_2 = y + r\sin\theta$

- *cluster* 3:
  - $t = 0.4 + 0.1\exp(-5\text{rand}()^2)$
  - $y_1 = 0.5 + 0.1\exp(-5\text{rand}()^2)$
  - $y_2 = 0.5 + 0.1\exp(-5\text{rand}()^2)$

Representa-se na Figura 6.20 a base de dados do Exemplo 5.

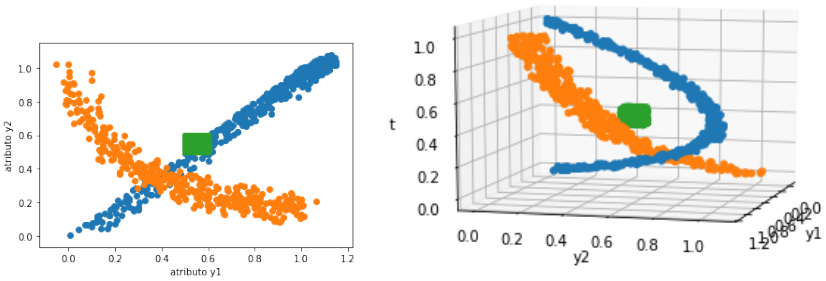


Figura 6.20: Elementos do exemplo 5.

Usou-se $\tau = 0.2$.

Single linkage

Apresentam-se os resultados na Tabela 6.16 e na Figura 6.21.

Tabela 6.16: Correspondência dos *clusters* (*single linkage*) — exemplo 5.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_2	C_2	C_1	C_1	C_1	C_1	C_1	C_2
do	C_2	C_1	C_1	C_2	C_2	C_2	C_2	C_2	C_1
Cluster				C_3	C_3				

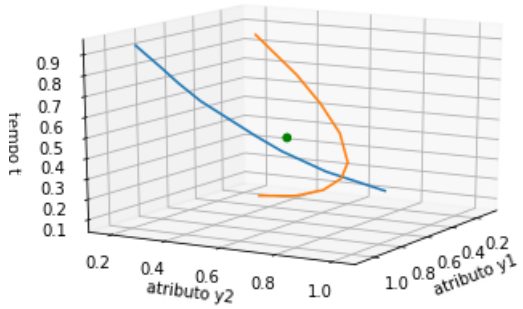


Figura 6.21: Trajetória dos *clusters* (*single linkage*) — exemplo 5.

Complete linkage

Apresentam-se os resultados na Tabela 6.17 e na Figura 6.22.

Tabela 6.17: Correspondência dos *clusters* (*complete linkage*) — exemplo 5.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice do Cluster	C_1	C_2	C_2	C_2					
	C_2	C_1	C_1	C_1	C_2	C_1	C_1	C_2	
				C_3	C_1				
					C_3	C_2	C_2	C_1	C_1
									C_2

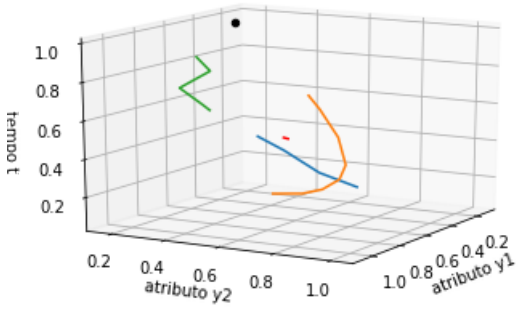


Figura 6.22: Trajetória dos *clusters* (*complete linkage*) — exemplo 5.

Average linkage

Apresentam-se os resultados na Tabela 6.18 e na Figura 6.23.

Tabela 6.18: Correspondência dos *clusters* (*average linkage*) — exemplo 5.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice do Cluster	C_1	C_2	C_2	C_2					
	C_2	C_1	C_1	C_1	C_2	C_1	C_1	C_2	
				C_3	C_1				
					C_3	C_2	C_2	C_1	C_1
									C_2

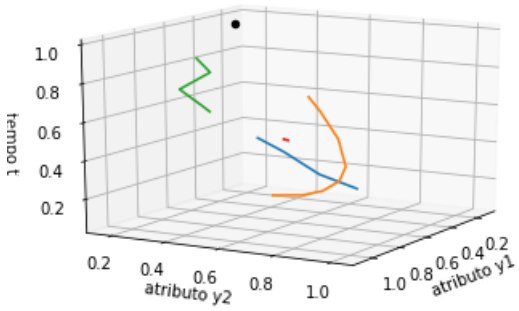


Figura 6.23: Trajetória dos *clusters* (*average linkage*) — exemplo 5.

Conclusão

Na projeção desta base de dados apenas se observa um único *cluster*, no entanto, quando se vê a evolução ao longo do tempo percebe-se que existem três *clusters*, dois do início do tempo até ao fim e outro que apenas aparece no meio. Com a abordagem das faixas, apenas o *single linkage* conseguiu encontrar estes três *clusters*. Os outros métodos encontraram bem no início, mas depois foram piorando. Isto pode estar relácionado com o *linkage*, mas também com o valor de τ escolhido, se este fosse menor poderia ter funcionado bem para os três *linkages*.

6.7 Exemplo 6

A base de dados deste exemplo é constituída por 3 *clusters* bem separados sendo que cada um deles tem 105, 102 e 108 elementos, respetivamente, e são da forma (y_1, y_2, t). Os pontos de cada *cluster* pertencem a cilindros com alturas diferentes e que não se intersectam seguindo as seguintes regras:

- *cluster* 1: $m_1 = 0$, $m_2 = 0$ e $\sigma = 0.05$
- *cluster* 2: $m_1 = \frac{2}{3}$, $m_2 = \frac{1}{3}$ e $\sigma = 0.05$

- *cluster* 3: $m_1 = \frac{3}{4}$, $m_2 = \frac{2}{3}$ e $\sigma = 0.05$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $r = \sigma \text{rand}()$
- $\theta = 2\pi \text{rand}()$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$
- *cluster* 1: $t = \text{rand}(0, 1)$.
- *cluster* 2: $t = \text{rand}(0, 2)$.
- *cluster* 3: $t = \text{rand}(1.5, 3)$.

Representa-se na Figura 6.24 a base de dados do Exemplo 6.

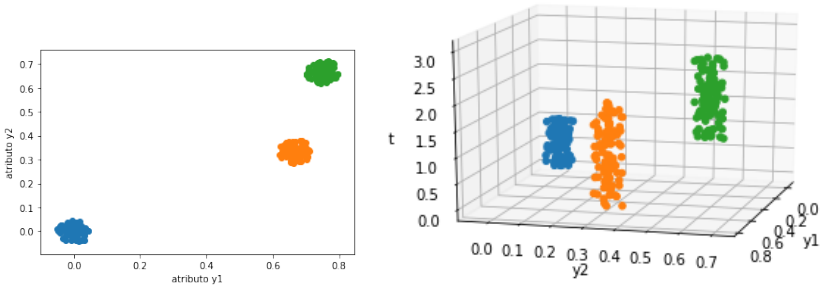


Figura 6.24: Elementos do exemplo 6.

Usou-se $\tau = 0.5$.

Single linkage

Apresentam-se os resultados na Tabela 6.19 e na Figura 6.25.

Tabela 6.19: Correspondência dos *clusters* (*single linkage*) — exemplo 6.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_2	C_2			
do	C_2	C_2	C_2	C_2							
Cluster						C_2	C_1	C_1	C_1	C_1	C_1

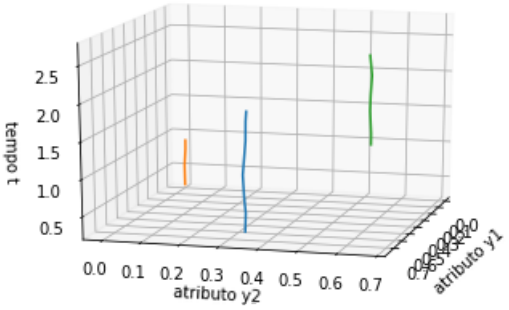


Figura 6.25: Trajetória dos *clusters* (*single linkage*) — exemplo 6.

Complete linkage

Apresentam-se os resultados na Tabela 6.20 e na Figura 6.26.

Tabela 6.20: Correspondência dos *clusters* (*complete linkage*) — exemplo 6.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_2	C_2			
do	C_2	C_2	C_2	C_2							
Cluster						C_2	C_1	C_1	C_1	C_1	C_1

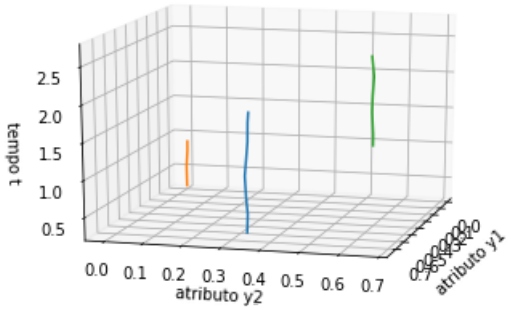


Figura 6.26: Trajetória dos *clusters* (*complete linkage*) — exemplo 6.

Average linkage

Apresentam-se os resultados na Tabela 6.21 e na Figura 6.27.

Tabela 6.21: Correspondência dos *clusters* (*average linkage*) — exemplo 6.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$
Índice	C_1	C_2	C_1	C_1	C_1	C_1	C_2	C_2			
do	C_2	C_1	C_2	C_2							
Cluster						C_2	C_1	C_1	C_1	C_1	C_1

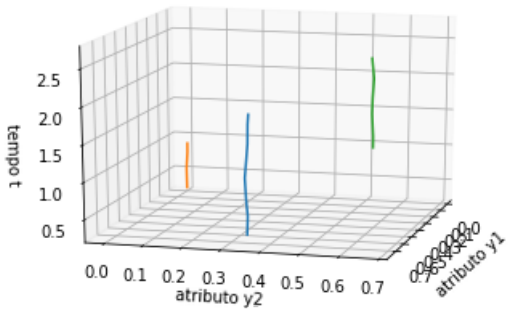


Figura 6.27: Trajetória dos *clusters* (*average linkage*) — exemplo 6.

Conclusão

Nesta base de dados há o aparecimento e desaparecimento de *clusters*, ou seja, existem *clusters* num determinado período do tempo que depois deixam de existir, ou o contrário, aparece um novo *cluster*. O objetivo da aplicação da abordagem das faixas era encontrar os três *clusters* e realmente foram encontrados.

6.8 Exemplo 7

A base de dados deste exemplo é constituída por 3 *clusters*, sendo que dois deles seguem a mesma reta, ou seja num certo momento no tempo um deles deixa de existir e mais tarde aparece. Cada um dos *clusters* tem 105, 105 e 108 elementos, respetivamente, e são da forma (y_1, y_2, t). Os pontos de cada *cluster* pertencem a cilindros e seguem as seguintes regras:

- *cluster* 1: $m_1 = 0$, $m_2 = 0$ e $\sigma = 0.05$
- *cluster* 2: $m_1 = 0$, $m_2 = 0$ e $\sigma = 0.05$
- *cluster* 3: $m_1 = \frac{3}{4}$, $m_2 = \frac{2}{3}$ e $\sigma = 0.05$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $r = \sigma \mathrm{rand}()$
- $\theta = 2\pi \mathrm{rand}()$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$
- *cluster* 1: $t = \mathrm{rand}(0, 1)$
- *cluster* 2: $t = \mathrm{rand}(2, 3)$

- *cluster* 3: $t = \text{rand}(1, 3)$

Representa-se na Figura 6.28 a base de dados do Exemplo 7.

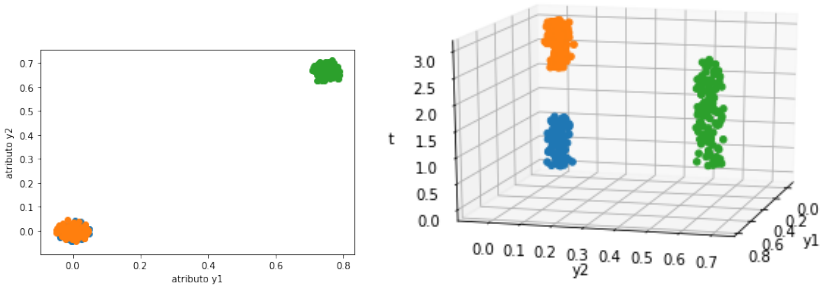


Figura 6.28: Elementos do exemplo 7.

Usou-se $\tau = 0.5$.

Single linkage

Apresentam-se os resultados na Tabela 6.22 e na Figura 6.29.

Tabela 6.22: Correspondência dos *clusters* (*single linkage*) — exemplo 7.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$
Índice	C_1	C_1	C_1	C_2							
do				C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
Cluster								C_2	C_2	C_2	C_2

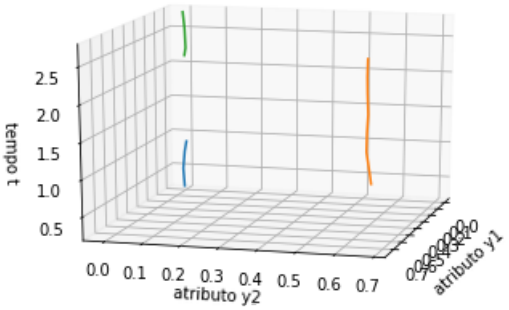


Figura 6.29: Trajetória dos *clusters* (*single linkage*) — exemplo 7.

Complete linkage

Apresentam-se os resultados na Tabela 6.23 e na Figura 6.30.

Tabela 6.23: Correspondência dos *clusters* (*complete linkage*) — exemplo 7.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$
Índice	C_1	C_1	C_1	C_1							
do				C_2	C_1	C_1	C_1	C_1	C_2	C_2	C_2
Cluster								C_2	C_1	C_1	C_1

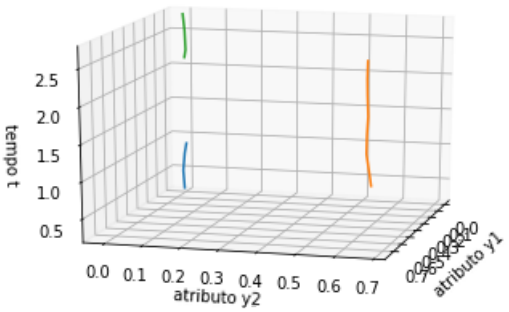


Figura 6.30: Trajetória dos *clusters* (*complete linkage*) — exemplo 7.

Average linkage

Apresentam-se os resultados na Tabela 6.24 e na Figura 6.31.

Tabela 6.24: Correspondência dos *clusters* (*average linkage*) — exemplo 7.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$
Índice	C_1	C_1	C_1	C_2							
do				C_1	C_1	C_1	C_1	C_1	C_1	C_2	C_2
Cluster								C_2	C_2	C_1	C_1

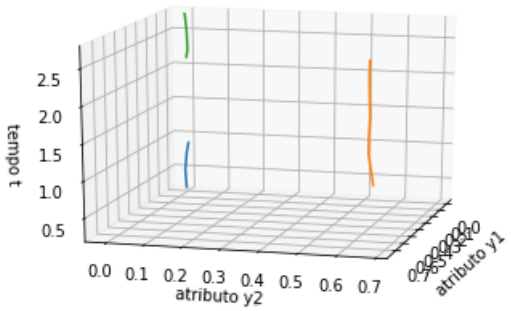


Figura 6.31: Trajetória dos *clusters* (*average linkage*) — exemplo 7.

Conclusão

Nesta base de dados também há o aparecimento e desaparecimento de *clusters*, sendo que um deles deixa de existir durante um certo período de tempo e depois volta a surgir com os mesmos atributos. Aplicando a abordagem das faixas, verificou-se exatamente isto.

6.9 Exemplo 8

A base de dados deste exemplo é constituída por 2 *clusters* que se cruzam num certo momento. Ambos têm 400 elementos e são da forma (y_1, y_2, t). Os pontos de cada

cluster pertencem a cilindros que seguem uma reta na diagonal, e seguem as seguintes regras:

- *cluster* 1: $m = (m_1, m_2) = (1.8 - 0.5t, 5.3 - 1.2t)$
- *cluster* 2: $m = (m_1, m_2) = (5.2 - 2.7t2, 0.7 + 1.55t2)$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $t = 3\text{rand}(0, 1)$
- $r = 0.5\text{rand}(0, 1)$
- $\theta = 2\pi\text{rand}(0, 1)$
- $y_1 = r\cos(\theta) + m_1$
- $y_2 = r\sin(\theta) + m_2$

Representa-se na Figura 6.32 a base de dados do Exemplo 8.

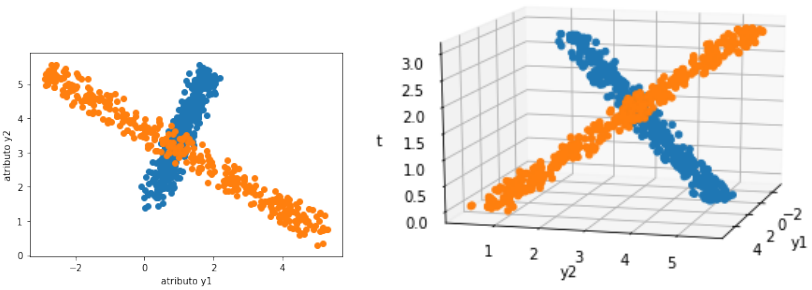


Figura 6.32: Elementos do exemplo 8.

Usou-se $\tau = 0.6$.

Single linkage

Apresentam-se os resultados na Tabela 6.25 e na Figura 6.33.

Tabela 6.25: Correspondência dos *clusters* (*single linkage*) — exemplo 8.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_2	C_1	C_1	C_1	C_1	C_1	C_1	C_2
do Cluster	C_2	C_1	C_2					C_2	C_1

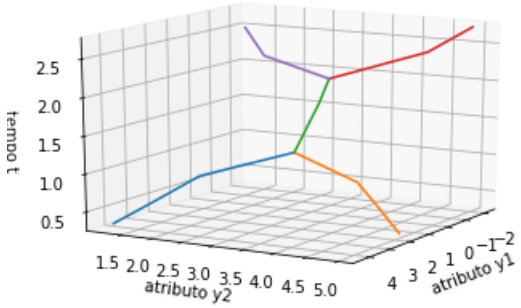


Figura 6.33: Trajetória dos *clusters* (*single linkage*) — exemplo 8.

Complete linkage

Apresentam-se os resultados na Tabela 6.26 e na Figura 6.34.

Tabela 6.26: Correspondência dos *clusters* (*complete linkage*) — exemplo 8.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_2
do Cluster	C_2	C_2	C_2				C_2	C_2	C_2

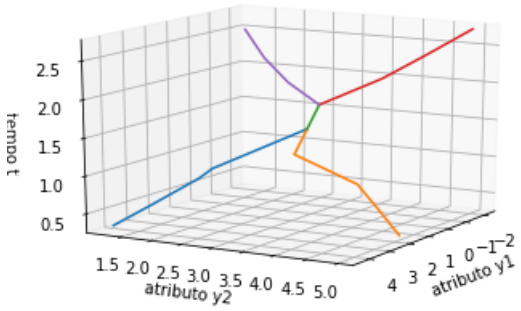


Figura 6.34: Trajetória dos *clusters* (*complete linkage*) — exemplo 8.

Average linkage

Apresentam-se os resultados na Tabela 6.27 e na Figura 6.35.

Tabela 6.27: Correspondência dos *clusters* (*average linkage*) — exemplo 8.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_2
do Cluster	C_2	C_2	C_2				C_2	C_2	C_2

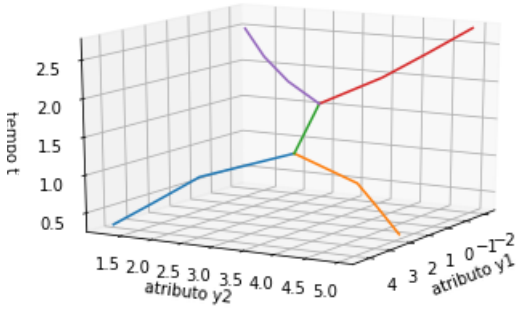


Figura 6.35: Trajetória dos *clusters* (*average linkage*) — exemplo 8.

Conclusão

Nesta base de dados existem dois *clusters* que se cruzam num certo momento, formando apenas um *cluster* nesse exato tempo. Mais tarde, voltam a separar-se em dois *clusters*. Também com a abordagem das faixas isto se verificou.

6.10 Exemplo 9

A base de dados deste exemplo é constituída por 3 *clusters* em forma de hélice, que não se intersetam. Os três têm 400 elementos cada e são da forma (y_1, y_2, t). Os pontos de cada *cluster* seguem as seguintes regras:

- *cluster* 1: $(m_1, m_2) = (0.2, 0.2)$ e $(r_1, r_2) = (0.05, 0.01)$
- *cluster* 2: $(m_1, m_2) = (0.8, 0.2)$ e $(r_1, r_2) = (0.05, 0.01)$
- *cluster* 3: $(m_1, m_2) = (0.5, 0.8)$ e $(r_1, r_2) = (0.1, 0.03)$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $t = \mathrm{rand}(0, 4\pi)$
- $r = r_1 + r_2 \mathrm{rand}(-1, 1)$
- $\theta = 2\pi \mathrm{rand}(0, 1)$
- $y_1 = r\cos(t) + m_1$
- $y_2 = r\sin(t) + m_2$

Representa-se na Figura 6.36 a base de dados do Exemplo 9.

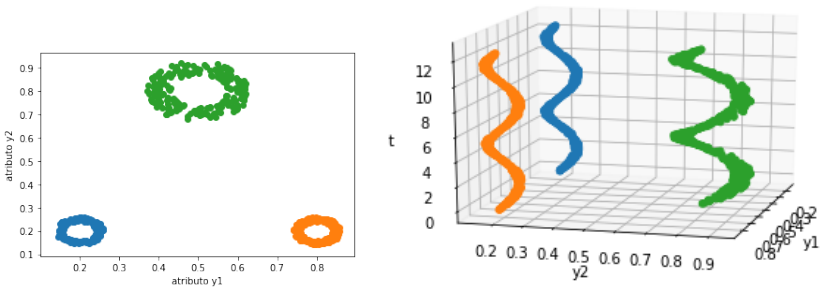


Figura 6.36: Elementos do exemplo 9.

Usou-se $\tau = 1$.

Single linkage

Apresentam-se os resultados na Tabela 6.28 e na Figura 6.37.

Tabela 6.28: Correspondência dos *clusters* (*single linkage*) – exemplo 9.

Índice da Faixa	$F^{\frac{1}{2}}$	F^1	$F^{\frac{3}{2}}$	F^2	$F^{\frac{5}{2}}$	F^3	$F^{\frac{7}{2}}$	F^4
Índice	C_1	C_2	C_2	C_2	C_2	C_2	C_1	C_2
do	C_2	C_1	C_3	C_3	C_3	C_3	C_2	C_3
Cluster	C_3	C_3	C_1	C_1	C_1	C_1	C_3	C_1

$F^{\frac{9}{2}}$	F^5	$F^{\frac{11}{2}}$	F^6	$F^{\frac{13}{2}}$	F^7	$F^{\frac{15}{2}}$	F^8
C_1	C_1	C_1	C_2	C_1	C_1	C_1	C_1
C_3	C_3	C_3	C_1	C_3	C_3	C_3	C_3
C_2	C_2	C_2	C_3	C_2	C_2	C_2	C_2

$F^{\frac{17}{2}}$	F^9	$F^{\frac{19}{2}}$	F^{10}	$F^{\frac{21}{2}}$	F^{11}	$F^{\frac{23}{2}}$	F^{12}	$F^{\frac{25}{2}}$
C_2	C_1	C_2	C_2	C_2	C_1	C_1	C_1	C_1
C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3
C_1	C_2	C_1	C_1	C_1	C_2	C_2	C_2	C_2

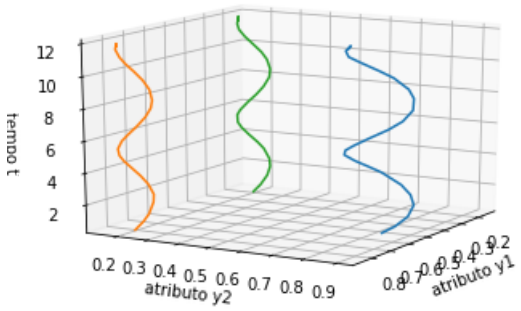


Figura 6.37: Trajetória dos *clusters* (*single linkage*) — exemplo 9.

Complete linkage

1ª versão

Apresentam-se os resultados na Tabela 6.29 e na Figura 6.38.

Tabela 6.29: Correspondência dos *clusters* (*complete linkage 1ª versão*) — exemplo 9.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}
Índice	C_1	C_2	C_1	C_1	C_2	C_1	C_1	C_1
do	C_2	C_3	C_2	C_3	C_3	C_2	C_2	C_2
Cluster	C_3	C_1	C_3	C_2	C_1	C_3	C_3	C_3

$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$	F^{6}	$F^{\frac{13}{2}}$	F^{7}	$F^{\frac{15}{2}}$	F^{8}
C_2	C_1	C_1	C_2	C_2	C_1	C_1	C_1
C_1	C_2	C_3	C_1	C_1	C_2	C_3	C_3
	C_3	C_2			C_3	C_2	C_2

$F^{\frac{17}{2}}$	F^{9}	$F^{\frac{19}{2}}$	F^{10}	$F^{\frac{21}{2}}$	F^{11}	$F^{\frac{23}{2}}$	F^{12}	$F^{\frac{25}{2}}$
C_2	C_1	C_2	C_2	C_2	C_2	C_2	C_2	C_3
C_1	C_2	C_3	C_1	C_1	C_1	C_1	C_1	C_1
	C_3	C_1	C_3		C_3			C_2

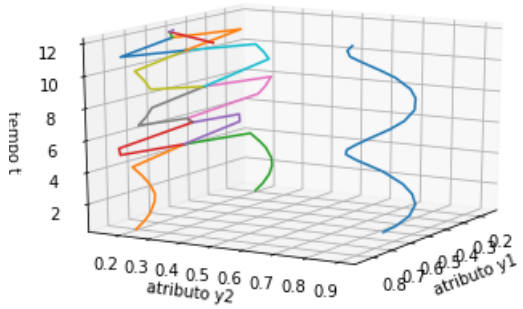


Figura 6.38: Trajetória dos *clusters* (*complete linkage 1ª versão*) — exemplo 9.

2ª versão

Apresentam-se os resultados na Tabela 6.30 e na Figura 6.39.

Tabela 6.30: Correspondência dos *clusters* (*complete linkage* 2ª versão) — exemplo 9.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}	$F^{\frac{9}{2}}$
Índice	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	$C1$
do Cluster	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2

F^{5}	$F^{\frac{11}{2}}$	F^{6}	$F^{\frac{13}{2}}$	F^{7}	$F^{\frac{15}{2}}$	F^{8}	$F^{\frac{17}{2}}$
C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2

F^{9}	$F^{\frac{19}{2}}$	$F^{1}0$	$F^{\frac{21}{2}}$	$F^{1}1$	$F^{\frac{23}{2}}$	$F^{1}2$	$F^{\frac{25}{2}}$
C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
C_2	C_2	C_2	C_2	C_2	C_2	C_2	C_2

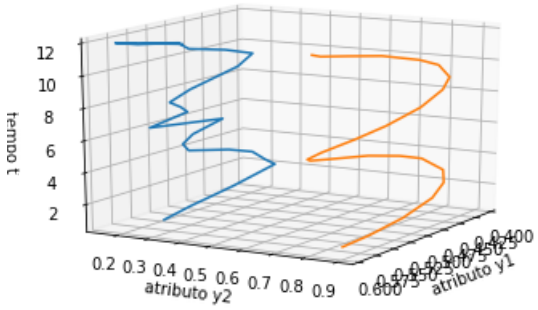


Figura 6.39: Trajetória dos *clusters* (*complete linkage 2ª versão*) — exemplo 9.

Average linkage

Apresentam-se os resultados na Tabela 6.31 e na Figura 6.40.

Tabela 6.31: Correspondência dos *clusters* (*average linkage*) — exemplo 9.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}
Índice	C_1	C_2	C_1	C_2	C_2	C_2	C_1	C_2
do	C_2	C_1	C_3	C_1	C_3	C_1	C_2	C_1
Cluster	C_3	C_3	C_2	C_3	C_1	C_3	C_3	C_3

$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$	F^{6}	$F^{\frac{13}{2}}$	F^{7}	$F^{\frac{15}{2}}$	F^{8}
C_1	C_1	C_2	C_1	C_2	C_1	C_1	C_1
C_2	C_3	C_3	C_3	C_1	C_3	C_3	C_2
C_3	C_2	C_1	C_2	C_3	C_2	C_2	C_3

$F^{\frac{17}{2}}$	F^{9}	$F^{\frac{19}{2}}$	F^{10}	$F^{\frac{21}{2}}$	F^{11}	$F^{\frac{23}{2}}$	F^{12}	$F^{\frac{25}{2}}$
C_2	C_1	C_2	C_1	C_1	C_2	C_1	C_2	C_2
C_1	C_3	C_1	C_3	C_3	C_1	C_3	C_3	C_3
C_3	C_2	C_3	C_2	C_2	C_3	C_2	C_1	C_1

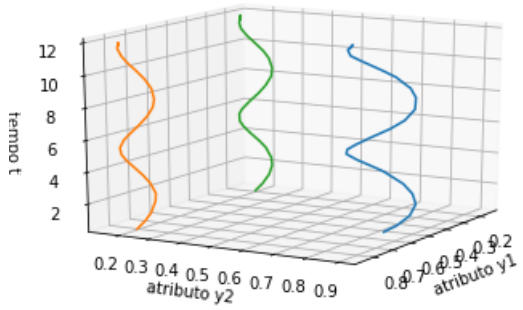


Figura 6.40: Trajetória dos *clusters* (*average linkage*) — exemplo 9.

Conclusão

Nesta base de dados existem três *clusters* bem separados, só que em forma de

hélice. Mais uma vez, a abordagem das faixas conseguiu distinguir bem os *clusters* tanto no *single linkage* como no *average*. No entanto, o *complete linkage* apenas encontrou 2 *clusters*, o que poderia ter resultado melhor se tivesse sido usado um valor de τ mais pequeno.

6.11 Exemplo 10

A base de dados deste exemplo é constituída por 3 clusters em forma de hélice com o mesmo eixo mas desfasadas. Os três têm 400 elementos cada e são da forma (y_1, y_2, t). Os pontos de cada *cluster* seguem as seguintes regras:

- *cluster* 1: $t_0 = 0$
- *cluster* 2: $t_0 = \frac{2\pi}{3}$
- *cluster* 3: $t_0 = \frac{4\pi}{3}$

E de seguida para cada elemento de cada *cluster*, calculou-se:

- $(m_1, m_2) = (0.5, 0.5)$
- $(r_1, r_2) = (0.05, 0.01)$
- $t_{final} = 4\pi$
- $t = \mathrm{rand}(0, t_{final})$
- $r = r_1 + r_2 \mathrm{rand}(-1, 1)$
- $y_1 = m_1 + r\cos(t - t_0)$
- $y_2 = m_2 + r\sin(t - t_0)$

Representa-se na Figura 6.41 a base de dados do Exemplo 10.

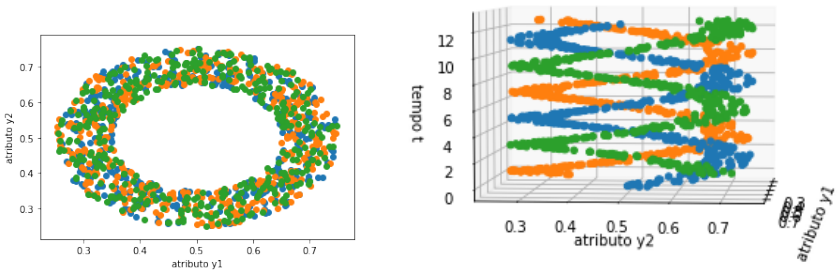


Figura 6.41: Elementos do exemplo 10.

Usou-se $\tau = 1$.

Single linkage

Apresentam-se os resultados na Tabela 6.32 e na Figura 6.42.

Tabela 6.32: Correspondência dos *clusters* (*single linkage*) — exemplo 10.

Índice da Faixa	$F^{\frac{1}{2}}$	F^{1}	$F^{\frac{3}{2}}$	F^{2}	$F^{\frac{5}{2}}$	F^{3}	$F^{\frac{7}{2}}$	F^{4}
Índice	C_1	C_1	C_3	C_2	C_2	C_1	C_2	C_1
do	C_2	C_3	C_2	C_3	C_3	C_3	C_3	C_2
Cluster	C_3	C_2	C_1	C_1	C_1	C_2	C_1	C_3

$F^{\frac{9}{2}}$	F^{5}	$F^{\frac{11}{2}}$	F^{6}	$F^{\frac{13}{2}}$	F^{7}	$F^{\frac{15}{2}}$	F^{8}
C_1	C_2	C_2	C_3	C_1	C_1	C_3	C_2
C_2	C_1	C_3	C_2	C_2	C_2	C_2	C_3
C_3	C_3	C_1	C_1	C_3	C_3	C_3	C_1

$F^{\frac{17}{2}}$	F^{9}	$F^{\frac{19}{2}}$	F^{10}	$F^{\frac{21}{2}}$	F^{11}	$F^{\frac{23}{2}}$	F^{12}	$F^{\frac{25}{2}}$
C_2	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3
C_3	C_2	C_2	C_2	C_2	C_1	C_1	C_1	C_1
C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2	C_2

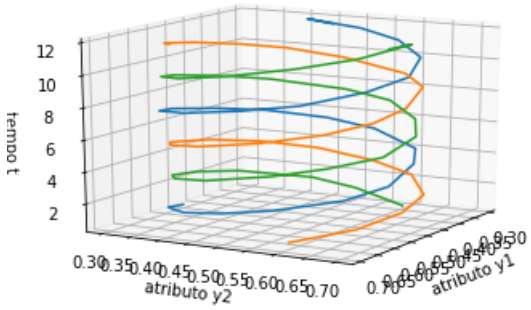


Figura 6.42: Trajetória dos *clusters* (*single linkage*) — exemplo 10.

Complete linkage

Neste método, como havia pouca relação entre as faixas, não fazia sentido fazer uma

tabela de correspondência. Apresentam-se os resultados na Figura 6.43.

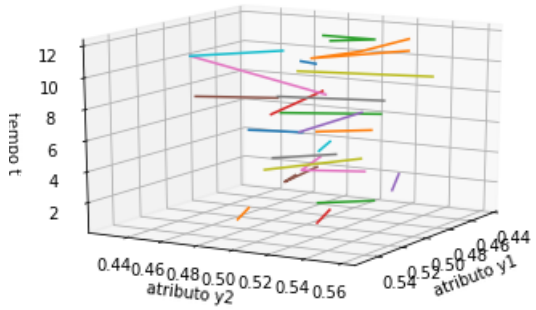


Figura 6.43: trajetória dos *clusters* (*complete linkage*) – exemplo 10.

Average linkage

Neste método, como havia pouca relação entre as faixas, não fazia sentido fazer uma tabela de correspondência. Apresentam-se os resultados na Figura 6.44.

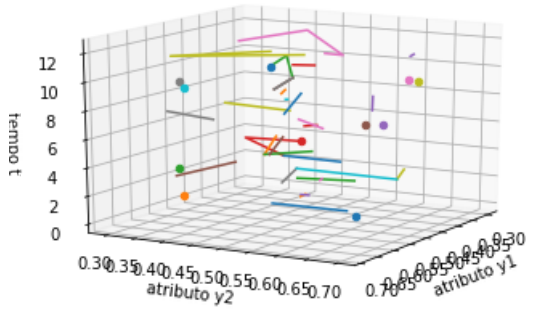


Figura 6.44: Trajetória dos *clusters* (*average linkage*) – exemplo 10.

Conclusão

Nesta base de dados também foram criados três *clusters* em forma de hélice, no entanto, havendo um entroncamento maior torna mais difícil a distinção dos *clusters*. O *single linkage* conseguiu distinguir os três *clusters* perfeitamente. Já nos outros *linkage* tal não se verificou, mostrando mesmo que os *clusters* das diferentes faixas não tinham

qualquer correspondência entre si. Isto poderia ser melhorado com um valor de τ menor.

6.12 Conclusões

Uma vez que todos os exemplos obtiveram resultados favoráveis, pode-se concluir que à partida este método funciona bem. Tanto para dados mais simples, os casos onde os *clusters* são visualmente bem definidos, tanto os casos mais complexos, onde há interseções de *clusters* em determinados momentos do tempo, onde pode haver aparecimento e desaparecimento de *clusters*, este método funcionou como esperado.

6.13 Automatização da escolha do número óptimo de clusters

Nos exemplos anteriores foi utilizado o algoritmo de uma maneira com intervenção humana, ou seja, sabia-se quais os resultados que se queriam obter (número de *clusters*) em cada faixa e por isso o algoritmo era executado, sempre que possível, de maneira a obter esses resultados, respeitando certas condições, como o número mínimo de elementos que cada *cluster* deveria ter, por exemplo. Ora, tal podia ser feito porque existiam os dados sintéticos e sabia-se o que se pretendia. No entanto, quando se estiver a trabalhar com dados reais não se irá saber como tem de ser o resultado e mediante isto, houve a necessidade de criar um *script* do algoritmo automatizado, ou seja, apenas é inserido o *linkage* que se pretende utilizar (*single, complete, average*) e a medida que as faixas deverão ter, e de seguida o algoritmo faz o seu trabalho, respeitando algumas condições, e assim encontra o número ideal de *clusters*.

6.13.1 Descrição do script

Primeiramente, todos os elementos são divididos por faixas, com uma certa medida τ que se refere ao atributo do tempo.

De seguida, para cada faixa, é executado o *Agglomerative Clustering* com todos os números de *clusters* possíveis. A cada partição com número de *clusters* diferentes está associada uma distância, por exemplo à partição com apenas dois *clusters* está associada a distância que corresponde a juntar os dois *clusters* mais próximos da partição que tinha três *clusters*. Seja $d = \{d_1, d_2, \ldots, d_n\}$ esta distância, onde n é o número de elementos da faixa em questão. Recorde-se que inicialmente todos os elementos são um *cluster*. Neste caso, inverteu-se o conjunto d, ou seja, d_n é a distância entre os dois primeiros elementos que foram unidos e formaram um novo *cluster*, e assim passa a haver $n - 1$ *clusters*, e assim sucessivamente, até d_1 que é distância associada à última junção que junta todos os elementos num só *cluster*.

Assim, após estarem calculadas todas as distâncias, vão ser calculados três critérios para escolher o número ideal de *clusters*. Sejam eles:

- **Rácio**: Para todos os pontos das distâncias calculam-se os declives entre eles. Depois de se ter uma lista com os declives ($e = \{e_1, e_2, \ldots, e_{n-1}\}$, onde e_n é o declive entre a distância d_n e d_{n+1}), calcula-se o rácio entre estes declives. Assim temos um conjunto $r = \{r_1, \ldots, r_{n-2}\}$ onde r_n é o rácio entre os declives e_n e e_{n+1} e vai corresponder ao número de *clusters* igual a $n + 1$. Por exemplo, r_1 é o rácio entre os declives e_1 e e_2 e vai corresponder ao número de *clusters* igual a 2, e assim sucessivamente. Quando se tiver a lista toda dos rácios, o valor que iremos escolher para o número de *clusters* é o que tiver o maior rácio.

- **Diferença**: Para todos os pontos das distâncias calculam-se os declives entre eles. Depois de se ter uma lista com os declives ($e = \{e_1, e_2, \ldots, e_{n-1}\}$, onde e_1 é o declive entre a distância d_1 e d_2), calcula-se a diferença entre estes declives. Assim

temos um conjunto $f = \{f_1, \ldots, f_{n-2}\}$ onde f_n é a diferença entre os declives e_n e e_{n+1} e vai corresponder ao número de *clusters* igual a $n+1$. Ou seja, f_1 é a diferença entre os declives e_1 e e_2 e vai corresponder ao número de *clusters* igual a 2, e assim sucessivamente. Quando se tiver a lista toda das diferenças, o valor que iremos escolher para o número de *clusters* é o que tiver a maior diferença.

- **Curvatura**: Calcula-se a lista das curvaturas para todos os valores de n, fazendo:

$$K(n) = \frac{d''(n)}{(1+d'(n)^2)^{\frac{3}{2}}}$$

Se n só tiver valores à direita então:

  - $d''(n) = 1d_n - 2d_{n+1} + 1d_{n+2}$
  - $d'(n) = -1d_n + 1d_{n+1}$

Se n tiver valores à direita e à esquerda então:

  - $d''(n) = 1d_{n-1} - 2d_n + 1d_{n+1}$
  - $d'(n) = -\frac{1}{2}d_{n-1} + \frac{1}{2}d_{n+1}$

Quando se tiver a lista toda das curvaturas, o valor que iremos escolher para o número de *clusters* é o que tiver a maior curvatura. [13]

Tanto o **Rácio** como a **Diferença** não incluem os casos do número de *clusters* ser igual a 1, então foi necessário criar outro critério para ser aplicado antes de aplicar estes dois. O que este novo critério diz é que se $d_1 < 1.2d_2$ então considera-se logo que o número de *clusters* ideal é 1.

Após estarem calculados todos os *clusters* de todas as faixas, é necessário estabelecer uma certa conectividade entre eles, sendo isto feito tendo em conta o maior número de elementos em comum, ou seja, se uma faixa tem 3 *clusters* (a, b e c) e a seguinte

também tem 3 (d, f e g), por exemplo a liga-se a d se o número de elementos da intersecção destes dois *clusters* for maior que o número de elementos da intersecção de a com f e de a com g. Depois de estabelecida esta conectividade, ligam-se os centróides dos *clusters* de cada faixa aos centróides correspondentes dos *clusters* da faixa seguinte, até à última faixa. Desta forma, obtém-se a trajetória dos *clusters* ao longo do tempo É de esperar que desta forma, os *clusters* obtidos não irão ser tão semelhantes aos que eram pretendidos e que se conseguiram nos exemplos anteriores. No entanto, verifica-se que os resultados também são bons. De seguida, serão aprofundados algumas partes de alguns exemplos.

Exemplo I (referente ao anterior exemplo 1)

Quando o *script* é executado para o exemplo 1, com *single linkage* e $\tau = 0.6$ verifica-se que nem todas as faixas encontram três *clusters* como pretendido. Observemos as Tabelas 6.33 e 6.34 de duas faixas seguidas, a $F^{\frac{5}{2}}$ e a F^3, respetivamente.

Tabela 6.33: Tabela de critérios para escolha do número de *clusters* da faixa $F^{\frac{5}{2}}$ (*single linkage*) — exemplo 1.

N° de clusters	Distâncias	Rácio	Diferença	Curvatura
1	0.91735	–	–	–
2	0.32168	3.23687	0.41856	0.33971
3	0.14456	12.6392	0.17231	0.1702
4	0.13976	0.28315	−0.0375	−0.0374
5	0.09747	4.12763	0.0396	0.03959
6	0.09480	0.82066	−0.0028	−0.0028

Tabela 6.34: Tabela de critérios para escolha do número de *clusters* da faixa F^3 (*single linkage*) — exemplo 1.

N° de clusters	Distâncias	Rácio	Diferença	Curvatura
1	0.91126	—	—	—
2	0.30347	3.65785	0.4489	0.36545
3	0.14456	3.11259	0.11463	0.11288
4	0.1003	5.0477	0.04351	0.04348
5	0.0995	1.04059	0.00042	0.000419
6	0.0992	0.87954	−0.0014	−0.001415

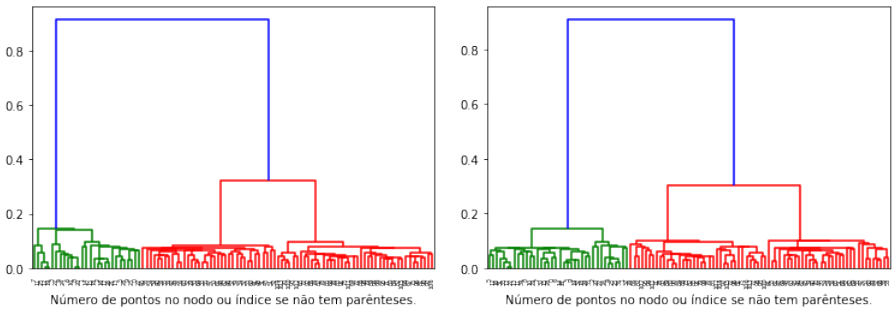


Figura 6.45: Dendrogramas das faixas $F^{\frac{5}{2}}$ e F^3 (*single linkage*) — exemplo 1.

Para o primeiro dendrograma, da Figura 6.45, que corresponde à faixa $F^{\frac{5}{2}}$, o número ideal de *clusters* que o algoritmo encontrou utilizando o critério do rácio foi 3, uma vez que toma o maior valor como se pode ver na Tabela 6.33 enquanto que no segundo dendrograma referente à faixa F^3 o número ideal de *clusters* é 4.

Após ser executado o *Agglomerative Clustering* para todas as faixas e fazendo a ligação dos centróides, obtêm-se as trajetórias. Esta comparação é feita na Figura 6.46.

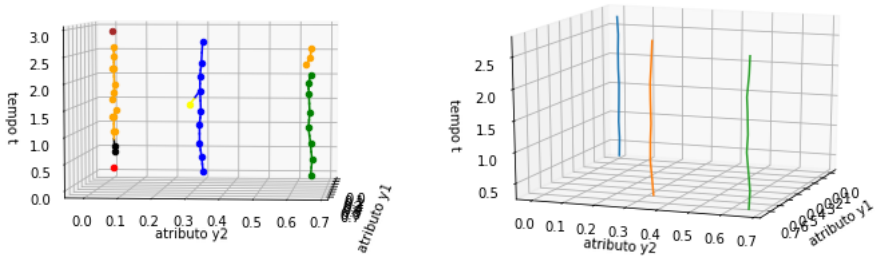


Figura 6.46: Trajetória dos *clusters* (*single linkage*) — exemplo 1.

A primeira imagem da Figura 6.46 é referente à trajetória calculada pelo *script* enquanto que a segunda é referente à trajetória *clusters* calculada anteriormente, que deu o mesmo resultado dos *clusters* criados artificialmente. Verifica-se que na maior parte das faixas existem 3 *clusters*, no entanto existem algumas exceções. Isto acontece porque em certas faixas os elementos estão mais afastados ou mais próximos e já não formam os *clusters* esperados, o que não significa que não esteja correto.

Exemplo II (referente ao anterior exemplo 5)

Também este exemplo, quando executado com o *script*, usando o *single linkage* e $\tau = 0.2$, os resultados não são iguais aos pretendidos. Mas, mais uma vez, isto não significa que o algortimo trabalhou mal, significa apenas que para o algortimo o número ideal de *clusters* não é o mesmo dos que foram criados artificialmente, usandos os critérios que este *script* acha mais relevantes.

Na Figura 6.47 podemos comparar ambas as trajetórias, a do *script* e a calculada anteriormente.

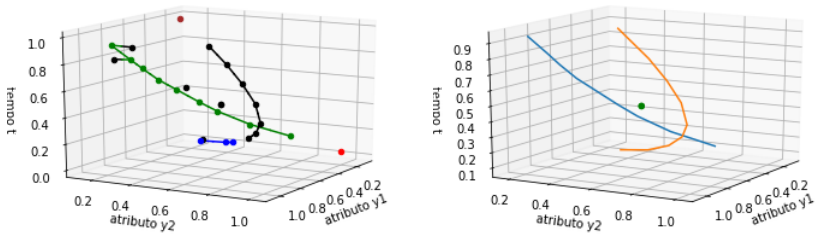


Figura 6.47: Trajetória dos *clusters* (*single linkage*) — exemplo 5.

Neste caso, o *script* encontrou apenas dois *clusters* para cada faixa, enquanto que nos dados artificiais havia duas faixas com três.

Exemplo III (referente ao anterior exemplo 6)

Neste exemplo, o *script* foi executado usando *single linkage* e $\tau = 0.5$, e mais uma vez os resultados não foram exatamente os esperados, mas foram bastante semelhantes, como se pode observar na Figura 6.48

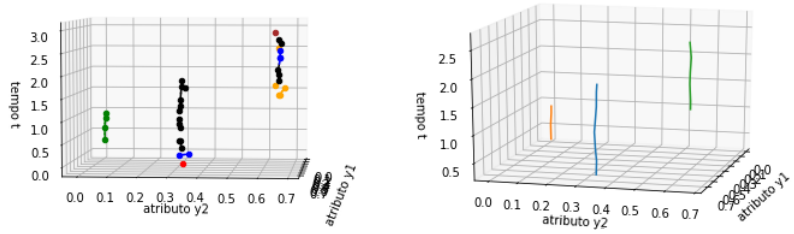


Figura 6.48: Trajetória dos *clusters* (*single linkage*) — exemplo 6.

6.13.2 Conclusões

Esta abordagem de dividir a base de dados em faixas obteve bons resultados na maioria dos casos, podendo considerar-se assim um bom método para utilizar nos dados reais.

Capítulo 7

Aplicação aos dados reais da empresa Pact Coffee

Neste capítulo vai-se trabalhar com dados reais da empresa *Pact Coffee* que vende café *online*.

7.1 A empresa Pact Coffee

A Pact Coffee é uma empresa fundada em 2012 no Reino Unido cujo objetivo é vender café *online*. Esta empresa compra o café diretamente aos produtores de mais de 9 origens diferentes. No *site* da empresa é possível fazer um plano e mudá-lo a qualquer altura. Neste plano é possível escolher desde sacos ou cápsulas até à moagem do café, passando por outras etapas seletivas.

7.2 Formato dos dados reais

A base de dados que vai ser estudada é referente a encomendas, ou seja, cada elemento refere-se a uma encomenda onde são considerados os seguintes atributos: o tempo

(há quantos dias foi feita a encomenda em relação à data da extracção dos dados), a moagem (classificada entre 1 e 6), o perfil de sabor (classificado de 1 a 4), o perfil de torrefação (classificado entre 1 e 7) e o plano (classificado entre 1 e 3). Note-se que apesar de numéricos, todos os atributos com exceção do tempo são puramente categóricos e foram inseridos nesta base de dados porque são atributos escolhidos pelos utilizadores sempre que iniciam um processo de escolha e compra de um plano de café no *site*.

7.3 Construção de uma nova métrica

O algoritmo criado no Capítulo 6 para fazer o *clustering* dos dados aplica-se a esta base de dados e mediante os resultados sofrerá alguns ajustes que melhor se adequam a esta base, uma vez que já não se tratam de dados sintéticos. Estes ajustes são necessários porque agora temos dados categóricos, e da maneira que o algoritmo está feito não tem isso em conta. Por exemplo no caso do atributo do perfil de torrefação, quando este tem valor 1 está à mesma distância do valor 2 do que do valor 5, não existe uma ordem. Para o *clustering* o atributo do tempo não vai ser necessário, uma vez que apenas é usado para a separação dos elementos em faixas.

O algoritmo começa por fazer uma divisão dos dados por faixas, mediante uma medida dada τ. Para ter uma ideia dos ajustes que devem ser feitos, fez-se uma estatística para cada faixa, calculando-se a frequência de cada valor de cada atributo nessa faixa. Esta estatística encontra-se na Tabela 7.1, com valores aproximados, apenas para as primeiras faixas.

Tabela 7.1: Percentagem aproximada de elementos de cada valor de cada atributo.

	A_1					
	1	2	3	4	5	6
$F^{\frac{1}{2}}$	55 %	0 %	12 %	18 %	10 %	5 %
F^{1}	48 %	0 %	15 %	17 %	16 %	5 %
$F^{\frac{3}{2}}$	48 %	0 %	14 %	18 %	14 %	5 %

	A_2			
	1	2	3	4
$F^{\frac{1}{2}}$	57 %	12 %	15 %	16 %
F^{1}	54 %	15 %	15 %	16 %
$F^{\frac{3}{2}}$	46 %	14 %	22 %	18 %

	A_3						
	1	2	3	4	5	6	7
$F^{\frac{1}{2}}$	26 %	17 %	16 %	6 %	9 %	10 %	15 %
F^{1}	22 %	15 %	16 %	3 %	8 %	10 %	25 %
$F^{\frac{3}{2}}$	19 %	18 %	18 %	0 %	12 %	9 %	22 %

	A_4		
	1	2	3
$F^{\frac{1}{2}}$	14 %	70 %	16 %
F^{1}	13 %	74 %	14 %
$F^{\frac{3}{2}}$	12 %	76 %	12 %

Assim se conclui que a frequência de cada valor de cada atributo em cada faixa é mais importante que a do número de vezes que esse valor aparece, uma vez que existem atributos com 8 valores e outros com 3, ou seja, é de esperar que os que têm apenas 3, cada um deles apareça em maior quantidade do que quando são 8.

7.3.1 Primeiro ajuste

A primeira tentativa foi criar uma função de similaridade para calcular a semelhança entre cada evento de encomendas. No entanto, no algoritmo de *clustering* aglomerativo ou entra a base de dados e o algoritmo usa as distâncias nele incorporadas para calcular a distância entre os eventos ou então em vez disso, insere-se uma matriz calculada anteriormente para a base de dados, sendo que esta tem de ser de dissimilaridade e não de similaridade. Então, o que foi feito foi criar uma função de similaridade e depois transformá-la numa distância, e aplicá-la à base de dados de modo a construir uma matriz de dissimilaridade.

Seja $F = (x^n)_{n=1}^N$ uma faixa qualquer onde N é o número de elementos dessa faixa e $A = \{A_1, A_2, A_3, A_4\}$ o conjunto formado pelos quatro atributos referidos anteriormente (moagem, perfil de sabor, perfil de torrefação e plano, respetivamente). Assim, temos $A_1 = \{a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, a_{16}\}$, constituído pelos seis valores que o atributo "moagem" pode assumir, $A_2 = \{a_{21}, a_{22}, a_{23}, a_{24}\}$, constituído pelos quatro valores que o atributo "perfil de sabor" pode assumir, $A_3 = \{a_{31}, a_{32}, a_{33}, a_{34}, a_{35}, a_{36}, a_{37}\}$, constituído pelos sete valores que o atributo "perfil de torrefação" pode assumir, e finalmente $A_4 = \{a_{41}, a_{42}, a_{43}\}$, constituído pelos três valores que o atributo "plano" pode assumir.

De seguida, calculam-se as frequência de cada valor de cada atributo na faixa, ou seja f_{ij} onde i se refere ao atributo e j ao índice do valor em causa. Por exemplo, f_{11} é referente à frequência do valor a_{11} do atributo A_1.

Então, tendo dois elementos x e x' da base de dados, a similaridade entre eles vai ser calculada da seguinte forma:

$$s(x, x') = \sum_{i=1}^{4} s_i(x, x')$$

onde

$$s_i(x, x') = \begin{cases} f_{ij}, & \text{se } x_i = x'_i, \text{ com } a_{ij} = x_i \\ 0, & \text{caso contrário.} \end{cases}$$

Depois de se ter a função de similaridade, transforma-se a mesma numa dissimilaridade através da expressão

$$d(x, x') = \frac{1}{s(x, x') + 10^{-6}}.$$

Assim, calcula-se a matriz de dessimilaridade que vai entrar no *clustering*. O *clustering* é feito para cada faixa e usa-se o critério do maior rácio para a escolha do número de *clusters*. Uma vez criados os *clusters* verifica-se a conexão destes entre as faixas, fazendo a intersecção de cada *cluster* de uma faixa com cada um da faixa seguinte e ligando os que têm maior número de elementos em comum.

Funcionou?

Não. Os resultados da aplicação desta métrica mostraram que na formação dos *clusters*, estes estavam a ser agrupados unidade a unidade, ou seja quando se tinha n *clusters*, a próxima junção, que daria $n-1$ *clusters* era feita agrupando apenas um elemento ao *cluster* "grande" que já existia, concluindo-se assim que não estavam a ser criados *clusters* bem agrupados. De facto, havia uma conexão entre as faixas, o problema é que quase todos os elementos se concentravam apenas num único *cluster*, ou seja numa faixa tínhamos por exemplo 3 *clusters*, sendo que dois deles tinham apenas 1 elemento cada, e o outro *cluster* concentrava os restantes elementos, e quando se fazia a conexão com a faixa seguinte, havia uma ligação porque mais uma vez, a maior parte dos elementos concentravam-se num *cluster* apenas.

7.3.2 Segundo ajuste

Como a primeira tentativa falhou, decidiu-se criar uma nova função distância. Mais uma vez, seja $F = (x^n)_{n=1}^N$ uma faixa qualquer onde N é o número de elementos dessa faixa, $A = \{A_1, A_2, \ldots, A_I\}$ o espaço de atributos, onde I é o número máximo de atributos da base de dados e sejam as frequências de cada valor de cada atributo numa faixa representadas por f_{ij} onde i se refere ao atributo e j ao índice do valor em causa. A nova proposta distância entre dois elementos x e x' é dada por

$$d(x, x') = \sum_{i=1}^{I} d_i(x, x') \tag{7.1}$$

onde

$$d_i(x, x') = \begin{cases} 0, & \text{se } x_i = x'_i \\ f_{ij} + f_{ik}, \text{ onde } a_{ij} = x_i \text{ e } a_{ik} = x'_i, & \text{caso contrário.} \end{cases}$$

Para esta função ser uma distância tem de obedecer aos critérios definidos no Capítulo 2. De seguida será apresentada a prova que esta nova função proposta é realmente uma distância.

- $d_i(x, x') = 0 \vee d_i(x, x') = f_{ij} + f_{ik}$, e $f_{ij} + f_{ik} \geq 0$, então $d_i(x, x') \geq 0$ e assim $d(x, x') \geq 0$. Está provado assim o 1° critério.

- $d_i(x, x') = 0 \vee d_i(x, x') = f_{ij} + f_{ik}$ e $d_i(x', x) = 0 \vee d_i(x', x) = f_{ik} + f_{ij}$. Uma vez que a adição é comutativa, $f_{ij} + f_{ik} = f_{ik} + f_{ij}$, logo $d_i(x, x') = d_i(x', x)$, então $d(x, x') = d(x', x)$, e assim está provado o 2° critério.

- $d_i(x, x'') = 0 \vee d_i(x, x') = f_{ij} + f_{im}$, $d_i(x, x') = 0 \vee d_i(x, x') = f_{ij} + f_{ik}$ e $d_i(x', x'') = 0 \vee d_i(x', x'') = f_{ik} + f_{im}$. Ora, $f_{ij} + f_{im} \leq f_{ij} + f_{ik} + f_{ik} + f_{im}$, logo $d_i(x, x'') \leq d_i(x, x') + d_i(x', x'')$ e assim $d(x, x'') \leq d(x, x') + d(x', x'')$. Está provado o 3° critério.

- $d(x, x') = 0$ se $d_i(x, x') = 0$ e isto acontece se $x_i = x'_i \ \forall i$, e assim tem de acontecer $x = x'$. Está provado o 4º critério.

Está provado assim que a função 7.1 é uma distância.

7.4 Análise dos dados

Para o segundo ajuste foram feitos dois *benchmarks*, o primeiro com apenas 10% dos dados (corresponde mais ao menos a 10 mil elementos), extraídos aleatoriamente, e o segundo com 30% dos dados. Para cada um destes *benchmarks* foram calculadas as faixas e de seguida aplicou-se a função de distância criada, para calcular a matriz de dissimilaridade que entrará no *clustering*. O número de *clusters* escolhido para cada faixa terá em conta o maior rácio. Para cada *cluster* será calculado um representante. Este representante é calculado através da moda de cada atributo num *cluster*, por exemplo num determinado *cluster* se para o atributo A_1 o valor 2 aparecer em mais elementos que todos os outros, para o atributo A_2 o valor 3 aparecer mais vezes, para o atributo A_3 o valor 2 aparecer mais vezes e para o atributo A_4 o valor 1 aparecer mais vezes, então significa que o representante deste *cluster* será $[2, 3, 3, 1]$. O próximo passo é verificar quantos elementos tem em comum os *clusters* de uma faixa com a seguinte, ou seja, ligar cada um dos *clusters* de uma faixa com o *cluster* da faixa seguinte que tiver maior número de elementos em comum. Para tal são feitas matrizes de intersecção entre as faixas.

Primeiro benchmark

Considerem-se as seguintes matrizes de intersecção:

$$F^{\frac{1}{2}} \cap F^1 = \begin{bmatrix} 663 & 0 & 0 & 0 \\ 0 & 226 & 0 & 0 \\ 0 & 0 & 224 & 0 \\ 0 & 0 & 0 & 456 \end{bmatrix} \quad F^1 \cap F^{\frac{3}{2}} = \begin{bmatrix} 230 & 0 & 0 & 0 \\ 0 & 75 & 0 & 0 \\ 0 & 0 & 91 & 0 \\ 0 & 0 & 0 & 132 \end{bmatrix}$$

$$F^{\frac{3}{2}} \cap F^2 = \begin{bmatrix} 514 & 0 & 0 & 0 \\ 0 & 220 & 0 & 0 \\ 0 & 0 & 0 & 232 \\ 0 & 0 & 531 & 0 \end{bmatrix} \quad F^2 \cap F^{\frac{5}{2}} = \begin{bmatrix} 0 & 0 & 0187 \\ 67 & 0 & 0 \\ 0 & 0 & 193 \\ 0 & 88 & 0 \end{bmatrix}$$

$$F^{\frac{5}{2}} \cap F^3 = \begin{bmatrix} 260 & 0 & 0 \\ 0 & 267 & 0 \\ 0 & 0 & 1211 \end{bmatrix} \quad F^3 \cap F^{\frac{7}{2}} = \begin{bmatrix} 0 & 0 & 61 & 0 \\ 0 & 0 & 0 & 41 \\ 148 & 120 & 0 & 0 \end{bmatrix}$$

$$F^{\frac{7}{2}} \cap F^4 = \begin{bmatrix} 0 & 0 & 495 & 0 \\ 0 & 583 & 0 & 0 \\ 0 & 0 & 0 & 222 \\ 198 & 0 & 0 & 0 \end{bmatrix} \quad F^4 \cap F^{\frac{9}{2}} = \begin{bmatrix} 79 & 0 & 0 & 0 \\ 0 & 205 & 0 & 0 \\ 0 & 0 & 188 & 0 \\ 0 & 0 & 0 & 78 \end{bmatrix}$$

Através destas matrizes é possível traçar uma trajetória desde os *clusters* da primeira faixa até à última. A cada *cluster* está associado o número de elementos que este contém. Observando estas matrizes também se conclui que a conexão entre os *clusters* de uma faixa para a seguinte é excelente, já que na maior parte dos casos se verifica que todos os elementos de um *cluster* estão apenas intersectados com um *cluster* da faixa seguinte, sendo 0 a intersecção com os restantes *clusters*. As exceções são em $F^2 \cap F^{\frac{5}{2}}$, onde se vê que dois *clusters* da faixa F^2 se juntam num só *cluster* na faixa $F^{\frac{5}{2}}$, e depois em $F^3 \cap F^{\frac{7}{2}}$, esse mesmo *cluster* volta a dividir-se em dois na faixa $F^{\frac{7}{2}}$.

Assim é possível observar na Figura 7.1 as trajetórias criadas e o número de elementos de cada *cluster* em cada uma das faixas.

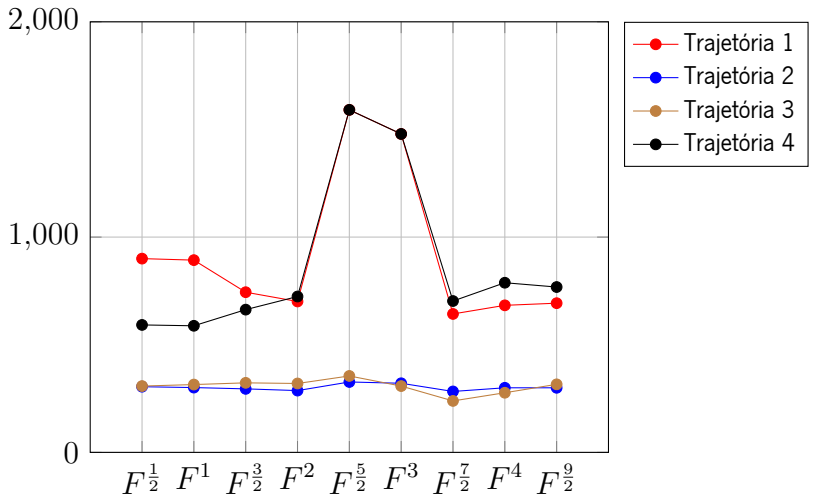


Figura 7.1: Trajetórias dos *clusters* ao longo das faixas — primeiro *benchmark*.

Pelo gráfico observa-se que todos os *clusters* mantiveram-se estáveis em relação ao número de elementos ao longo do tempo. No entanto é natural que, quando a trajetória 1 se juntou à trajetória 2, formando apenas um único *cluster*, este passou a ter um grande número de elementos, uma vez que foi a junção de dois *clusters*. Para cada trajetória está associada também uma lista de representantes de cada *cluster*, como se observa na Tabela 7.2.

Tabela 7.2: Representantes de cada trajetória — primeiro *benchmark*.

	Trajetória 1	Trajetória 2	Trajetória 3	Trajetória 4	junção das trajetórias 1 e 4
$F^{\frac{1}{2}}$	[1, 1, 1, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 2, 5, 2]	
F^{1}	[1, 1, 1, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 2, 5, 2]	
$F^{\frac{3}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 3, 2]	
F^{2}	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]	
$F^{\frac{5}{2}}$		[1, 1, 1, 1]	[1, 4, 3, 3]		[1, 1, 7, 2]
F^{3}		[1, 1, 1, 1]	[1, 4, 3, 3]		[1, 1, 7, 2]
$F^{\frac{7}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]	
F^{4}	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]	
$F^{\frac{9}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]	
Representante dos representantes	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]	[1, 1, 7, 2]

Note-se que duas das trajetórias, mais especificamente a 2 e a 3 têm sempre o mesmo representante. Isto significa que os *clusters* de uma faixa para a seguinte são coerentes. Nas restante isto não acontece mas ainda assim não variam muito. Observa-se também que quando a trajetória 1 foi unida à 4, os elementos tinham na maioria as características da trajetória 1.

Também na Tabela 7.2 estão representados cada representante dos representantes para cada trajetória, que foi calculado escolhendo o representante que aparecia mais vezes naquela trajetória. De cada um destes representantes concluímos para cada trajetória que os seus elementos:

- **Trajetória 1**: Antigamente escolhiam uma moagem de valor 1, um perfil de sabor

1, um perfil de torrefação 1 e tinham o plano 2, mas eventualmente à medida do tempo foram mudando o perfil de torrefação para 7, mantendo os restantes atributos.

- **Trajetória 2**: Escolheram o valor 1 para todos os atributos.
- **Trajetória 3**: Mantiveram-se sempre com as mesmas características, sendo elas 1 para a moagem, 4 para o perfil de sabor, 3 para o perfil de torrefação e usaram o plano 3.
- **Trajetória 4**: Inicialmente escolhiam uma moagem de valor 1, mantendo-se ao longo do tempo, o perfil de sabor foi variando de 2 a 3 ou 1 quando foram agrupados com os elementos da trajetória 1, e o perfil de torrefação de 5 a 3 ou 7 quando foram agrupados com os elementos da trajetória 1, mantendo sempre o plano 2.

Posteriormente, identificaram-se 5 clientes, que fizeram mais que três compras nesta base de dados, de maneira a se perceber em qual das trajetórias estes se encaixam. Sejam eles denominados por *cliente 1*, *cliente 2*, *cliente 3* , *cliente 4* e *cliente 5*, representados na Tabela 7.3:

Tabela 7.3: Compras de 5 clientes – primeiro *benchmark*.

cliente 1	*cliente 2*	*cliente 3*	*cliente 4*	*cliente 5*
[1, 2, 2, 2] t = 26	[1, 1, 2, 1] t = 29	[1, 4, 3, 3] t = 34	[4,3,5,2] t = 29	[5,1,7,2] t = 33
[1, 2, 2, 2] t = 20	[1, 1, 2, 1] t = 15	[1, 4, 3, 3] t = 25	[4,1, 1, 2] t = 21	[5,3,5,2] t = 25
[1, 2, 2, 2] t = 18	[1, 1, 2, 1] t = 8	[1, 4, 3, 3] t = 15	[4,1,7,2] t = 15	[5,1,7,2] t = 11
[1, 2, 2, 2] t = 18	[1, 1, 2, 1] t = 4	[1, 4, 3, 3] t = 13	[4,1,1, 2] t = 8	[5,2,7,2] t = 7
[1, 2, 2, 2] t = 8				
[1, 2, 2, 2] t = 8				
[1, 2, 2, 2] t = 8				
[1, 2, 2, 2] t = 1				
[1, 1, 1, 2] t = 1				
[1, 2, 2, 2] t = 1				

- *cliente 1*: Fez 10 compras, escolhendo maioritariamente a moagem 1, 2 tanto para o perfil de sabor como para o de torrefação, e sempre com o plano 2. Ora, as trajetórias mais parecidas com estes atributos são a trajetória 1 e a 4, pelo que este cliente deverá estar numa destas trajetórias ou variar entre elas.

- *cliente 2*: Fez 4 compras, escolhendo sempre a moagem 1, o perfil de sabor 1, o perfil de torrefação 1 o plano 1, sendo assim a trajetória a que pertence possivelmente a 2.

- *cliente 3*: Fez 4 compras, e claramente pertence à trajetória 3, uma vez que todas as suas compras têm as mesmas características dos elementos desta trajetória.

- *cliente 4*: Fez 4 compras. Manteve sempre a mesma moagem e o mesmo plano, no entanto começou com um perfil de sabor, mantendo depois a preferência noutro, e o perfil de torrefação foi experimentando vários. Também este cliente parece variar entre a trajetória 1 e 4.

- *cliente 5*: Fez 4 compras, e também este manteve a mesma moagem e o mesmo plano e foi variando os outros atributos. Deverá estar incluído entre a trajetória 1 e 4.

Segundo benchmark

Considerem-se as seguintes matrizes de intersecção:

$$F^{\frac{1}{2}} \cap F^{1} = \begin{bmatrix} 2022 & 0 & 0 & 0 \\ 0 & 684 & 0 & 0 \\ 0 & 0 & 698 & 0 \\ 0 & 0 & 0 & 1378 \end{bmatrix} \quad F^{1} \cap F^{\frac{3}{2}} = \begin{bmatrix} 709 & 0 & 0 & 0 \\ 0 & 227 & 0 & 0 \\ 0 & 0 & 244 & 0 \\ 0 & 0 & 0 & 422 \end{bmatrix}$$

$$F^{\frac{3}{2}} \cap F^{2} = \begin{bmatrix} 1499 & 0 & 0 & 0 \\ 0 & 607 & 0 & 0 \\ 0 & 0 & 0 & 662 \\ 0 & 0 & 1620 & 0 \end{bmatrix} \quad F^{2} \cap F^{\frac{5}{2}} = \begin{bmatrix} 559 & 0 & 0 & 0 \\ 0 & 215 & 0 & 0 \\ 0 & 0 & 594 & 0 \\ 0 & 0 & 0 & 270 \end{bmatrix}$$

$$F^{\frac{5}{2}} \cap F^{3} = \begin{bmatrix} 1685 & 0 & 0 & 0 \\ 0 & 741 & 0 & 0 \\ 0 & 0 & 1902 & 0 \\ 0 & 0 & 0 & 811 \end{bmatrix} \quad F^{3} \cap F^{\frac{7}{2}} = \begin{bmatrix} 0 & 434 & 0 & 0 \\ 0 & 0 & 190 & 0 \\ 402 & 0 & 0 & 0 \\ 0 & 0 & 0 & 169 \end{bmatrix}$$

$$F^{\frac{7}{2}} \cap F^{4} = \begin{bmatrix} 1744 & 0 & 0 & 0 \\ 0 & 1515 & 0 & 0 \\ 0 & 0 & 640 & 0 \\ 0 & 0 & 0 & 645 \end{bmatrix} \quad F^{4} \cap F^{\frac{9}{2}} = \begin{bmatrix} 591 & 0 & 0 & 0 \\ 0 & 538 & 0 & 0 \\ 0 & 0 & 241 & 0 \\ 0 & 0 & 0 & 242 \end{bmatrix}$$

Mais uma vez, é possível traçar uma trajetória desde os *clusters* da primeira faixa até à última através destas matrizes. Observando estas matrizes conclui-se que a conexão entre os *clusters* de uma faixa para a seguinte é excelente novamente, sendo neste caso ainda mais, porque para todas as faixas existem o mesmo número de *clusters* e estão todos bem diferenciados.

Assim é possível observar na Figura 7.2 as trajetórias criadas e o número de elementos de cada *cluster* em cada uma das faixas.

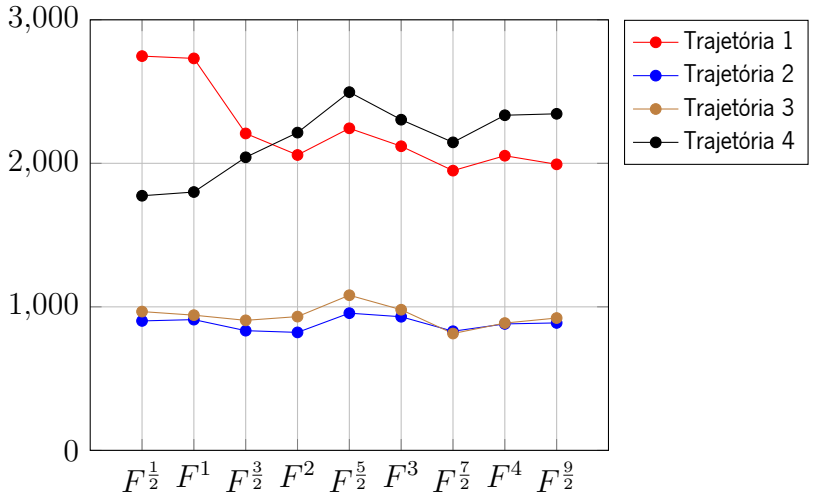


Figura 7.2: Trajetórias dos *clusters* ao longo das faixas — segundo *benchmark*.

Observa-se que tanto a trajetória 2 como a 3 mantêm uma certa estabilidade no número de elementos. Já a trajetória 1 tem uma ligeira descida, ao contrário da trajetória 2 que tem uma subida. Isto pode significar por exemplo que alguns clientes da trajetória 1 passaram para a trajetória 2.

Para cada trajetória está associada também uma lista de representantes de cada *cluster*, representados na Tabela 7.4.

Tabela 7.4: Representantes de cada trajetória — segundo *benchmark.*

	Trajetória 1	Trajetória 2	Trajetória 3	Trajetória 4
$F^{\frac{1}{2}}$	[1, 1, 1, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 2, 3, 2]
F^{1}	[1, 1, 1, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 2, 5, 2]
$F^{\frac{3}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
F^{2}	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
$F^{\frac{5}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
F^{3}	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
$F^{\frac{7}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
F^{4}	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
$F^{\frac{9}{2}}$	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]
Rep.	[1, 1, 7, 2]	[1, 1, 1, 1]	[1, 4, 3, 3]	[1, 3, 5, 2]

Nesta caso, as trajetórias da trajetória 2 e 3 têm sempre o mesmo representante, tal como aconteceu no primeiro *benchmark* com as trajetórias correspondentes. Significa então que os *clusters* desta trajetória são coerentes. Nas restantes isto não acontece, mais precisamente, as duas faixas mais antigas têm um representante diferente das restantes faixas. De qualquer das maneiras, os *clusters* são coerentes, sendo que a diferença é pequena. No final da Tabela 7.4 estão representados cada representante dos representantes para cada trajetória, que foi calculado, novamente, escolhendo o representante que aparecia mais vezes naquela trajetória, sendo então as características dos elementos de cada trajetória as seguintes:

- **Trajetória 1**: Inicialmente escolhiam uma moagem de valor 1, um perfil de sabor 1, um perfil de torrefação 1 e tinham o plano 2, e pouco tempo depois mudaram o perfil de torrefação para 7.
- **Trajetória 2**: Escolheram o valor 1 para todos os atributos.

- **Trajetória 3**: Mantiveram-se sempre com as mesmas características também, 1 para a moagem, 4 para o perfil de sabor, 3 para o perfil de torrefação e usaram o plano 3.

- **Trajetória 4**: Começaram com preferência na moagem de valor 1, 2 no perfil de sabor, 3 no perfil de torrefação e com o plano 2, de seguida mudaram o perfil de torrefação para 5. E daí em diante, mudaram apenas o perfil de sabor para 3, mantendo-se durante muito tempo assim.

Posteriormente, identificaram-se 5 clientes, que fizeram mais que cinco compras nesta base de dados, para mais uma vez se perceber em qual das trajetórias estes se encaixam. Sejam eles denominados por *cliente 6*, *cliente 7*, *cliente 8* , *cliente 9* e *cliente 10*, representados na Tabela 7.5.

Tabela 7.5: Compras de 5 clientes – segundo *benchmark*.

cliente 6		*cliente 7*	*cliente 8*	*cliente 9*	*cliente 10*
	[1, 2, 2, 2] t = 33	[1, 3, 4, 3] t = 29	[1, 3, 4, 3] t = 32	[3 ,4 ,3,3] t = 33	[1,2,7,2] t = 32
2	[1, 2, 2, 2] t = 29	[1, 4,3,3] t = 20	[3, 3, 6, 3] t = 29	[3,1,7,2] t = 20	[1,4,5,2] t = 28
2	[1, 2, 2, 2] t = 26	[1,4,3,3] t = 12	[3, 4, 3, 3] t = 25	[5,1,1,2] t = 20	[1,1,7,2] t = 21
2	[1, 2, 2, 2] t = 22	[1, 4,3,3] t = 8	[1,3,6, 3] t = 15	[3,4,3,3] t = 12	[1,3,5,2] t = 15
2	[1, 2, 2, 2] t = 20	[1, 4,3,3] t = 6	[1,3,6, 3] t = 15	[5,1,1, 2] t = 6	[1,1,1,2] t = 7
2	[1, 2, 2, 2] t = 18	[1, 4,3,3] t = 5	[3,3,6, 3] t = 13	[3,3,5, 2] t = 6	[1,1,1,2] t = 1
	[6, 1, 1, 2] t = 18	[1, 4,3,3] t = 4			
	[1, 2, 2, 2] t = 13	[1, 3,4,3] t = 1			
	[1, 2, 2, 2] t = 12				
4	[1, 2, 2, 2] t = 8				
2	[1, 2, 2, 2] t = 5				
4	[1, 2, 2, 2] t = 1				

- *cliente 6*: Fez 24 compras, algumas delas iguais no mesmo dia. Todas elas foram com as mesmas características, à exceção de apenas uma em que mudou o valor da moagem de 1 para 6. Tendo em conta que teve sempre o plano 2, poderá encaixar-se ou na trajetória 1 ou na 4.

- *cliente 7*: Fez 8 compras, e todas as suas compras parecem identificar-se com a trajetória 3.

- *cliente 8*: Fez 6 compras. As suas características parecem ser mais parecidas com as da trajetória 3, de salientar o plano.

- *cliente 9*: Fez 6 compras. Este cliente apresenta característica da trajetória 3 da 1 e 4, por isso é possível que não pertence a apenas uma, mas que se enquadre nas três.

- *cliente 10*: Fez 6 compras, e apresenta características da trajetória 1 e 4, podendo variar a sua trajetória entre estas duas.

Funcionou?

Sim. Podemos observar que tanto no primeiro *benchmark* como no segundo obtiveram-se *clusters* coerentes e bem conectados temporalmente. Assim como também se viu que na maioria dos casos, os clientes mantém-se sempre na mesma trajetória.

Capítulo 8

Conclusão

O objetivo do presente trabalho era prever o comportamentos de clientes de uma empresa que vende café *online*. O contexto, é assim, o de "*Streaming Data*" e consistiu na aplicação de algoritmos de *clustering* hierárquico aglomerativo, onde a escolha de uma métrica apropriada é um fator crítico para a obtenção de bons resultados.

Desde o início deste trabalho que se entendeu que o atributo do tempo teria de ser diferenciado de alguma forma dos restantes atributos, uma vez que o objetivo era trabalhar com uma base de dados em que cada evento tinha um tempo associado relevante para o estudo.

Assim, inicialmente, criou-se uma métrica que tinha em conta o tempo de cada evento e também um tempo de referência a partir do qual os dados foram analisados. Foi feito um grande estudo sobre esta métrica de modo a provar que poderia resultar para atingir o objetivo pretendido. Esta métrica foi criada para calcular uma matriz de dissimilaridade, e foi testada para várias bases de dados criadas artificialmente. Para cada um destes exemplos de bases de dados foi então aplicado o *clustering* hierárquico aglomerativo, e avaliou-se o comportamento da métrica. Para os primeiros exemplos, os resultados foram positivos e chegou-se a considerar que a métrica era realmente funcional. No entanto, face a situações de dados mais complexos, esta métrica não respondeu de uma maneira

adequada. Chegou-se à conclusão, que esta métrica correspondia a uma projeção dos dados, e depois calculava os *clusters* com uma métrica unicamente dependente dos atributos sem tomar em conta o tempo.

Na sequência deste resultado negativo, surgiu a necessidade de propor outra abordagem tendo em conta o atributo temporal na construção dos *clusters*. Foi então que surgiu o método que divide a base de dados em faixas temporais.

Este método consistiu em dividir a base de dados em várias faixas, tendo em conta uma medida de tempo (tempo característico) para cada faixa, e mediante o atributo de tempo associado a cada evento, esse evento era colocado na respetiva faixa. Também para testar este método foram usadas as bases de dados artificiais anteriores e criadas novas bases de dados mais sofisticadas,

- Após os dados estarem divididos por faixas, aplicou-se para cada uma o *clustering* hierárquico aglomerativo com a distância de *Manhattan*.
- Aplicando a metodologia da curva do cotovelo, verificou-se que os *clusters* identificados correspondiam aos *clusters* definidos na base de dados artificial.
- Determinou-se uma excelente correspondência entre as partições de cada faixa (continuidade dos *clusters*).
- Construíram-se assim trajetórias dos *clusters* temporais.

Esta abordagem baseada em faixas temporais permite lidar com dados sempre a chegar à base de dados, uma vez que não era necessário voltar a fazer o *clustering* para toda a base de dados, mas apenas para os últimos dados inseridos. Assim, decidiu-se aplicar aos dados da empresa *Pact Coffee* esta mesma abordagem.

Uma vez que as bases de dados sintéticas eram constituídas por atributos numéricos, houve a necessidade de adaptar a métrica a dados nominativos, pois a base de dados real que se pretendia tratar continha apenas atributos deste tipo. Assim, surgiu a necessidade

de construir uma métrica inovadora que tem em consideração as frequências dos valores dos atributos por forma a destacar a importância relativa entre os valores dos atributos.

Conseguiu-se então usando o método das faixas juntamente com a nova métrica criada um *clustering* eficiente, isto é, foram calculados *clusters* que realmente demonstram que esta abordagem produziu resultados de qualidade e com mais informação, sendo possível traçar uma trajetória destes mesmos *clusters* ao longo do tempo. Em particular, conseguiu-se identificar que os clientes de um determinado *cluster* se mantinham neste mesmo ao longo do tempo, o que era de esperar uma vez que a partir do momento que um cliente subscreve um plano espera-se que se mantenha nele.

Verifica-se então que esta abordagem pode ser aplicada noutras bases de dados que tenham a mesma particularidade de dados sempre a entrarem ("*Streaming Data*") com dados nominativos.

Bibliografia

[1] Introdução a clusterização e os diferentes métodos. URL https://portaldatascience.com/introducao-a-clusterizacao-e-os-diferentes-metodos/F.

[2] K. M. Cassiano. AnÁlise de sÉries temporais usando anÁlise espectral singular (ssa) e clusterizaÇÃo de suas componentes baseada em densidade. URL https://www.maxwell.vrac.puc-rio.br/24787/24787_5.PDF.

[3] S. L. Clain. *Machine Learning Theory and applications*.

[4] E. G. de Souza. Entendendo o que é matriz de confusão com python. URL https://medium.com/data-hackers/entendendo-o-que-%C3%A9-matriz-de-confus%C3%A3o-com-python-114e683ec509.

[5] U. Kokate, A. Deshpande, P. Mahalle, and P. Patil. Data stream clustering techniques, applications, and models: Comparative analysis and discussion. *Big Data and Cognitive Computing*, 2(4):32, Oct 2018. ISSN 2504-2289. doi: 10.3390/bdcc2040032. URL http://dx.doi.org/10.3390/bdcc2040032.

[6] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[7] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical*

Statistics and Probability, Volume 1: Statistics, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200512992.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] S. Priy. Clustering in machine learning. URL https://www.geeksforgeeks.org/clustering-in-machine-learning/.

[10] P. Sharma. The most comprehensive guide to k-means clustering you'll ever need. URL https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

[11] C. Silva and B. Ribeiro. *Aprendizagem Computacional em Engenharia*. Ensino. Imprensa Da Universidade de Coimbra / Coimbra Univ, 2018. ISBN 9789892615073. URL https://books.google.pt/books?id=oFhQDwAAQBAJ.

[12] J. Silva, E. Faria, R. Barros, E. Hruschka, A. de Carvalho, and J. Gama. Data stream clustering: A survey. *ACM Computing Surveys*, 46, 03 2014. doi: 10.1145/2522968.2522981.

[13] Wikipedia contributors. Finite difference coefficient — Wikipedia, the free encyclopedia, 2020. URL https://en.wikipedia.org/w/index.php?title=Finite_difference_coefficient&oldid=987174365. [Online; accessed 19-January-2021].