Monitorizando a Evolução de Clusters

por

Márcia Daniela Barbosa de Oliveira

Tese de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Orientada por

Professor Doutor João Manuel Portela da Gama

Faculdade de Economia

Universidade do Porto

2010

Aos meus pais

Nota Biográfica

Márcia Daniela Barbosa de Oliveira nasceu em Aveiro, no dia 3 de Dezembro de 1986. Foi na Universidade da sua cidade natal que se licenciou em Gestão, em Junho de 2008. Em Setembro do mesmo ano, decide ingressar no Mestrado em Análise de Dados e Sistemas de Apoio à Decisão, da Faculdade de Economia da Universidade do Porto. No decurso do Mestrado, descobriu o mundo do *Data Mining*, que lhe despertou o interesse desde o primeiro minuto. O gosto pela área materializou-se em duas publicações em conferências: *Bipartite Graphs for Monitoring Clusters Transitions* Proc. XVII Jornadas de Classificação e Análise de Dados, JOCLAD 2010 (pp. 195-198), Lisboa, Portugal, e IDA - *Intelligent Data Analysis* 2010, *Bipartite Graphs for Monitoring Clusters Transitions*, *Proceedings of the Ninth International Symposium on Intelligent Data Analysis*, IDA 2010, Tucson, Arizona, USA, vol.6065 *Lecture Notes in Computer Science*, Springer. Actualmente, é bolseira de investigação do LIAAD - INESC Porto LA.

Agradecimentos

Começo por expressar os meus sinceros agradecimentos ao meu orientador, o Professor Doutor João Gama, pelas constantes palavras de incentivo e pela motivação que me soube transmitir, pelas críticas e sugestões que me fez e que tanto me ensinaram, e pelo empenho, interesse e disponibilidade que desde a primeira hora colocou nesta orientação e que foram fulcrais para a concretização deste trabalho.

De uma forma especial agradeço aos meus pais e ao meu irmão por terem sempre acreditado em mim, pela força que me deram nos momentos mais difíceis, pelo carinho, pela preocupação e pelo enorme apoio no alcance dos meus objectivos.

Ao João expresso a minha gratidão pelos conselhos, pela paciência, pela compreensão e pelo facto de me ter ajudado a encontrar o equilíbrio entre o trabalho e os momentos de lazer.

Aos meus amigos, pelas sugestões, pelo apoio, pela força, pela amizade, pelos bons momentos de descontracção e, sobretudo, pela forma como compreenderam as minhas ausências.

À Silvana, minha inseparável colega de Mestrado, pela companhia, pela partilha de ideias e experiências, pela força que me deu nos momentos mais frágeis e por fomentar o meu espírito crítico.

Gostaria também de agradecer ao apoio do projecto *Knowledge Discovery from Ubiquitous Data Streams*, financiado pela FCT (PTDC/EIA-EIA/098355/2008).

Aos revisores anónimos dos artigos submetidos em conferências, agradeço as críticas úteis, fundamentais para a melhoria deste trabalho.

À Direcção Geral de Administração Interna um sincero agradecimento pela disponibilização dos dados do escrutínio.

Aos colegas do LIAAD, por se terem demonstrado sempre dispostos a ajudar.

Tabela de Símbolos Matemáticos

Notação Matemática	**Descrição**
$\emptyset \rightarrow C_u(t_j)$	Nascimento de um *cluster*
$C_m(t_i) \rightarrow \emptyset$	Morte de um *cluster*
$C_m(t_i) \overset{\subseteq}{\rightarrow} \{C_1(t_j), ..., C_r(t_j)\}$	Cisão de um *cluster* em r *clusters*
$\{C_1(t_i), ..., C_p(t_i)\} \overset{\subseteq}{\rightarrow} C_u(t_j)$	Fusão de p *clusters* num único *cluster*
$C_m(t_i) \rightarrow C_u(t_j)$	Sobrevivência de um *cluster*
$C_m(t_i) \nearrow C_u(t_j)$	Expansão de um *cluster*
$C_m(t_i) \searrow C_u(t_j)$	Contracção de um *cluster*
$C_m(t_i) \overset{\bullet}{\rightarrow} C_u(t_j)$	Compactação de um *cluster*
$C_m(t_i) \overset{*}{\rightarrow} C_u(t_j)$	Dispersão de um *cluster*

Resumo

O estudo da evolução tornou-se um tema relevante, principalmente na última década, devido a uma maior consciência da volatilidade do nosso mundo. O progresso registado no domínio da ciência e da tecnologia potenciou o surgimento de um mundo volátil e em constante evolução, que exige a adopção de novas perspectivas no âmbito da descoberta de conhecimento em dados como, por exemplo, perspectivas orientadas para o tempo. Como consequência, um novo paradigma surgiu para responder mais eficazmente a uma nova classe de problemas de *Data Mining*. Nesta dissertação abordamos o problema da monitorização da evolução de *clusters* e propomos a metodologia MEC, que foi desenvolvida segundo os príncipios do paradigma *Change Mining*. Neste trabalho, a evolução é traçada por via da detecção e categorização das transições experienciadas pelos *clusters*. Nós adoptamos duas estratégias principais para a caracterização dos *clusters* - representação em extensão e representação em compreensão -, de forma a alargar o domínio de aplicação do sistema de monitorização da evolução. A metodologia proposta engloba uma taxonomia dos vários tipos de transições de *clusters*, que podem ser endógenas ou exógenas, um mecanismo de acompanhamento que depende da representação dos *clusters*, e um algoritmo de detecção de transições. O mecanismo de acompanhamento pode ser subdividido em dois métodos que foram edificados para monitorizar a evolução das estruturas de *clusters*: um baseado nas transições de grafos bipartidos e em probabilidades condicionadas, e outro assente no grau de sobreposição de *clusters* no espaço de atributos. O *input* de ambos os métodos são as estruturas de *clusters* retornadas por um dado algoritmo de *Clustering*, e o *output* é o conjunto de transições a que os *clusters* foram submetidos no intervalo de tempo em análise (por exemplo, Cisão, Fusão, Nascimento, Contracção e Compressão). Para demonstrar a viabilidade e exequibilidade da metodologia MEC conduzimos experiências controladas com dados artificiais. Também demonstramos a aplicabilidade, bem como a capacidade da nossa abordagem em fornecer um diagnóstico eficaz das transições de *clusters* realizando pequenos casos de estudos, recorrendo para o efeito a conjuntos de dados provenientes de diferentes áreas do conhecimento, nomeadamente, Economia, Educação, Território e Política.

Palavras-Chave: Grafos Bipartidos, Change Mining, Clustering, Monitorização, Transições

Abstract

The study of evolution has become an important research issue, especially in the last decade, due to a greater awareness of our world's volatility. The rapid progress made in science and technology has contributed to the emergence of a fast pace evolving world, which demands new perspectives in knowledge discovery upon data, such as time-oriented perspectives. As a consequence, a new paradigm has emerged to respond more effectively to a class of new problems in Data Mining. In this thesis we address the problem of monitoring the evolution of clusters and propose the MEC framework, which was developed along the lines of this new Change Mining paradigm. In this work, the evolution is traced through the detection and categorization of transitions undergone by clusters. We adopt two main strategies for cluster characterization - representation by enumeration and representation by comprehension -, to improve the applicability of our evolution monitoring system. The proposed framework encompasses a taxonomy of various types of cluster transitions, that can be internal or external, a tracking mechanism that depends on the cluster representation, and a transition detection algorithm. Our tracking mechanism can be subdivided in two methods that were designed to monitor the evolution of clusters' structures: one based on bipartite graph's transitions and in conditional probabilities, and another based on the overlapping degree of clusters in the feature space. The input of both methods are clusters' structures obtained using a given Clustering algorithm and the output is a set of transitions experienced by each cluster (e.g. Split, Merge, Birth, Shrinkage and Compression). To demonstrate the feasibility of MEC framework we present controlled experiences with syntetic data. We also demonstrate the applicability and prove the ability of our framework in providing an efficient diagnosis of cluster's transitions conducting case studies using datasets from different knowledge areas, namely, Economy, Education, Territory and Politics.

Keywords: Bipartite Graphs, Change Mining, Clustering, Monitoring, Transitions

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Este Capítulo tem como intuito fornecer uma visão geral do trabalho desenvolvido e reportado nesta dissertação, através da descrição e exposição do tema e das principais motivações que conduziram à sua escolha, da enunciação dos objectivos que se pretende alcançar com este estudo e, ainda, através da apresentação das principais contribuições deste trabalho para o desenvolvimento científico na área.

1.1 Motivação

A celeridade a que se processa a evolução, tipicamente caracterizada por rupturas e alteração de paradigmas, tem aumentado exponencialmente nas últimas décadas. O progresso registado a nível científico e tecnológico potenciou o surgimento de um mundo volátil, onde as verdades são constantemente postas em causa e não passam de meros conceitos relativos. O crescimento ilimitado do conhecimento coloca o ser humano, e os agentes de decisão em particular, numa posição difícil em momentos de tomada de decisão que, tendencialmente, exigem processamento e análise de volumes substanciais de dados e selecção de informação relevante e útil, que possa ser facilmente transformada em acção.

O *Data Mining* surgiu como um grande aliado dos agentes de decisão, neste contexto de evolução, uma vez que oferece respostas rápidas a problemas que envolvem grandes quantidades de dados, através da aplicação de mecanismos que mimicam a capacidade de generalização e abstracção do cérebro humano, para criação de conhecimento. Uma das tarefas mais comuns de *Data Mining*, e aquela sobre a qual se debruçará o nosso estudo, é a aprendizagem não-supervisionada, cujo objectivo é classificar um conjunto de dados em classes desconhecidas *a priori*, por meio da sumarização e explicação das características-chave dos dados. O facto de incidir sobre dados não classificados e não exigir que sejam realizadas assumpções sobre os mesmos, torna esta tarefa extremamente útil e apelativa, alargando o seu âmbito de aplicação, dado que não é restringida pela inexistência de um atributo-classe.

Um dos métodos mais conhecidos e populares de aprendizagem não-supervisionada é o *Clustering*, que procede ao agrupamento de um conjunto de dados multidimensionais num conjunto de classes, vulgarmente denominadas *clusters*, com base no grau de semelhança das observações (Jain et al., 1999). Pretende-se que as observações afectas a um dado *cluster* apresentem um elevado grau de associação natural entre si - classes homogéneas - e que os *clusters* sejam relativamente distintos uns dos outros - classes bem separadas. Dadas as suas características, este método tem sido abordado em inúmeros contextos, sendo o processamento de imagem e o reconhecimento de padrões as aplicações com maior prevalência nesta área. Assim, um *cluster* pode ser encarado como um conceito que, devido à permanente evolução e consequentes impactos nos mecanismos de geração dos dados, pode sofrer mudanças (Kaur et al., 2009). A constatação da não estacionaridade dos conceitos ao longo do tempo, aliada à consciência de que é mais importante compreender a dinâmica e a natureza da evolução, do que simplesmente detectá-la (Spiliopoulou et al., 2006), foi o principal motivo que nos instigou à escolha do tema em estudo. O acompanhamento e compreensão das mudanças, que podem ocorrer nos conceitos, podem ser realizados com base na análise das características intrínsecas e extrínsecas dos *clusters*. Estas características servem de fundamento e alicerçam todo o processo de monitorização das transições intra-*clusters* e inter-*clusters*, que poderão emergir em diferentes instantes temporais.

A monitorização da dinâmica de estruturas de *clusters* contribui para sedimentar e/ou formar conhecimento sustentado sobre o fenómeno subjacente aos dados e, por conseguinte, para fomentar atitudes de pro-actividade. Por estes motivos, este estudo pode beneficiar inúmeras áreas, nomeadamente, o Marketing, a Detecção de Fraude, a Economia e a Sociologia. A compreensão da natureza das mudanças dos *clusters* tem bastante interesse do ponto de vista do Marketing, mais especificamente, em termos da análise da evolução dos segmentos de mercado e do comportamento do consumidor, ao permitir a detecção de alterações nas preferências e hábitos de consumo, e consequente acompanhamento e previsão de tendências que sustentem a redefinição das estratégias e políticas de Marketing e uma melhor gestão da relação com o cliente - CRM (*Customer Relationship Management*). De facto, a habilidade para detectar e acompanhar este tipo de informação pode funcionar como factor de diferenciação no mercado, acarretando vantagens competitivas para as empresas, uma vez que promove a adopção de atitudes pró-activas, em termos da antecipação e adaptação a potenciais consumidores, num futuro próximo. O âmbito deste estudo poderá, igualmente, favorecer a Detecção de Fraude, por exemplo, através da análise das transacções de crédito dos clientes de um Banco. No que diz respeito à Economia, poderá ter interesse efectuar o acompanhamento das tendências do tecido empresarial de um dado país, de modo a descobrir áreas de negócio emergentes e com elevado potencial de crescimento, com impactos ao nível do estudo da competitividade dos países. Em termos sociológicos, este tipo de estudo poderá ser direccionado para a descoberta e observação do modo como os grupos sociais

tendem a evoluir através de uma dimensão temporal, o que permite aprofundar o conhecimento sobre as sociedades e as respectivas tendências evolutivas. Este tipo de segmentação social poderá, igualmente, favorecer tarefas como a publicidade dirigida e a personalização de contéudos e serviços, adequando-os às necessidades e preferências dos consumidores.

A importância e os contributos de um estudo nesta área de conhecimento não se esgotam nestes pequenos exemplos de aplicação. O estudo das dinâmicas de *clusters* contribui para o alcance de um maior entendimento sobre os processos evolutivos de grupos de entidades, alargando os horizontes e abrindo novos caminhos na forma de abordar e pensar os problemas.

1.2 Objectivos

O presente trabalho tem como objectivo primordial estudar as dinâmicas de estruturas de *clusters*, sendo os nossos esforços direccionados, em termos genéricos, para a compreensão da evolução experienciada por *clusters*, ao longo do tempo. São inúmeras as questões que surgem, à medida que vamos mergulhando no tema da evolução dos *clusters*: quais os tipos de transições ou mudanças que podem ocorrer nos *clusters*, quer em termos internos, quer em termos externos? Como definir e construir métricas capazes de medir e monitorizar as transições tipificadas? Quais as características, intrínsecas e extrínsecas, dos *clusters* que melhor descrevem as suas peculiaridades e que devem ser consideradas na definição das métricas? Como analisar os resultados da aplicação das métricas, de tal forma que se consiga concluir sobre a significância das potenciais transições experienciadas por clusters construídos em instantes temporais distintos? Muitas outras perguntas poderiam ser colocadas, no entanto, por limitações de natureza operacional e de tempo, cingiremos a nossa investigação ao encontro de algumas possibilidades de resposta para as questões acima levantadas. Assim, neste trabalho pretende-se monitorizar o conhecimento existente sobre um dado domínio, representado sob a forma de estruturas de *clusters*, em diferentes instantes temporais e, sobretudo, perceber a natureza das transições que ocorrem, nessas estruturas, ao longo do tempo. Para alcançar estes objectivos propomos uma metodologia, que designámos de MEC, e que engloba uma taxonomia dos vários tipos de transições de *clusters*, um mecanismo de acompanhamento destas estruturas que depende do esquema de representação dos *clusters* e um algoritmo de detecção de transições. O mecanismo de acompanhamento é subdividido em dois novos métodos que foram concebidos para monitorizar esta evolução: um método baseado nas transições de grafos bipartidos e outro baseado no grau de sobreposição dos *clusters*. Pretendemos avaliar os métodos propostos com base em casos de estudo reais.

1.3 Contribuições

Neste trabalho apresenta-se um método que visa abordar o problema da monitorização e detecção de mudanças em estruturas de *clusters*. Para esse efeito definimos e propomos duas estratégias diferentes para representar *clusters*: **representação em extensão** e **representação em compreensão**. Também expomos dois métodos novos para detectar as transições experienciadas por *clusters*, cada um adaptado a um dos esquemas de representação. A inovação presente nestes métodos (ou mecanismos de acompanhamento) baseia-se no recurso a grafos bipartidos e probabilidades condicionadas, no caso de *clusters* representados em extensão, e na computação e avaliação do grau de semelhança dos *clusters* representados em compreensão.

Assim, a principal contribuição deste trabalho consiste na edificação de uma metodologia completa e eficiente para a monitorização da evolução de *clusters*. Esta metodologia não é restringida pelo forma como os *clusters* estão representados, o que alarga o âmbito da sua aplicação, e proporciona uma forma visualmente apelativa de apresentação e detecção das transições de *clusters* representados em extensão, o que permite compreender melhor o processo de monitorização subjacente.

1.4 Organização

Esta dissertação está dividida em três partes pincipais: revisão da literatura e estado da arte, descrição da metodologia MEC e avaliação experimental dos métodos propostos. A primeira parte integra o Capítulo 2 e o Capítulo 3. No Capítulo 2 expõem-se os principais conceitos teóricos que estão na base do *Clustering*, descrevem-se os passos necessários à obtenção de um agrupamento de qualidade e referem-se os avanços mais recentes da área. No Capítulo 3 fornece-se uma visão geral do Estado da Arte na evolução dos *clusters* através da exposição das principais taxonomias definidas para a classificação das transições de *clusters* e da explanação dos métodos e algoritmos considerados relevantes no âmbito da detecção e/ou monitorização das transições em estruturas de dados. Para cada método da literatura efectua-se um paralelismo entre o respectivo método e a nossa proposta, com o intuito de realçar as respectivas diferenças e/ou limitações.

Na segunda parte desta dissertação, referente ao Capítulo 4, apresentamos com detalhe a metodologia MEC. Este capítulo é encetado com a definição e distinção dos dois esquemas de representação de *clusters*. Depois, introduz-se formalmente a nossa metodologia de monitorização, mais especificamente, através da apresentação da taxonomia das transições exógenas e endógenas consideradas, da explicação do nosso mecanismo de acompanhamento de *clusters*, para cada tipo de representação, e da exposição dos alicerces do nosso algoritmo de detecção de transições.

No Capítulo 5, demonstramos e discutimos os resultados da aplicação da abordagem proposta a conjuntos de dados artificiais, procedemos à comparação dos dois

métodos de monitorização, efectuamos uma análise da sensibilidade dos respectivos limiares e apresentamos quatro casos de estudo, que abordam problemáticas provenientes de áreas de conhecimento distintas. Com estes conjuntos de dados reais pretendemos demonstrar a vasta aplicação da nossa metodologia e, ainda, realçar a sua utilidade no estudo da evolução e na detecção de transições associadas a mudanças significativas ocorridas no domínio de conhecimento subjacente.

Por fim, no Capítulo 6, apresentam-se as principais conclusões retidas e as perspectivas de desenvolvimento futuro.

Capítulo 2

Clustering

Como referido, o objectivo do presente trabalho é perceber as transições entre *clusters*, sendo o nosso objecto de estudo os próprios *clusters*. Neste sentido, e antes de iniciar o estudo propriamente dito, é essencial dispor de dados agrupados. Porém, e dada a diversidade de abordagens que existem no âmbito do agrupamento dos dados, torna-se necessário realizar a escolha dos tipos de *Clustering*, bem como das técnicas disponíveis, que serão aplicadas aos dados em bruto, com o intuito de obter o *input* do estudo. Com a finalidade de orientar esta escolha, considerou-se importante introduzir, em primeiro lugar, os principais conceitos que se encontram na base do *Clustering*, fornecer uma visão geral dos métodos e algoritmos de agrupamento, salientar os aspectos que são essenciais para a produção de um agrupamento de qualidade e, ainda, apresentar resumidamente os avanços mais recentes na área.

2.1 Breve introdução ao Clustering de dados

O *Clustering*, enquanto separação e classificação de objectos, é uma das actividades culturais mais básicas da humanidade (Hampel, 2002). Por este motivo, tem sido um recorrente alvo de interesse e um tema extremamente apelativo para uma alargada comunidade de investigadores e profissionais de diversas áreas, o que reflecte o carácter interdisciplinar desta técnica.

A Análise de *Clusters*, também denominada *Clustering*, Classificação Automática, *Q-analysis*, *Clumping*, Análise Tipológica e Taxonomia Numérica, consoante o campo de conhecimento onde é aplicada, pode ser definida como "o estudo formal dos métodos e algoritmos para agrupamento de objectos com base na semelhança ou nas características intrínsecas percebidas"(Jain, 2010). É uma técnica de aprendizagem não-supervisionada, uma vez que a classe a que pertence cada um dos objectos analisados não se encontra previamente atribuída (Halkidi et al., 2002), e de natureza exploratória ou descritiva, visto que tem como intuito compreender as características gerais e encontrar a estrutura subjacente aos dados através da descoberta de

padrões e da exploração das relações escondidas nos dados em bruto. O seu objectivo geral consiste, assim, na descoberta do agrupamento natural de um conjunto de objectos, e o seu alcance poderá satisfazer inúmeros propósitos. Segundo Jain (2010), existem três grandes propósitos para o recurso a técnicas de *Clustering*, por parte de investigadores e profissionais:

- Estrutura subjacente - para efeitos de compreensão dos dados, geração de hipóteses, detecção de anomalias e identificação de características salientes
- Classificação natural - para efeitos de identificação do grau de semelhança entre formas e organismos
- Compressão - como método de organização e sumarização dos dados

No que concerne à aplicabilidade desta técnica, as áreas de conhecimento que mais têm contribuído para a concepção de métodos e algoritmos de agrupamento e que mais têm usufruído do seu desenvolvimento são a Biologia, a Bioinformática, a Psicologia, a Sociologia, a Medicina, o Marketing, a Aprendizagem Automática, o *Data Mining*, a Matemática e a Estatística.

2.1.1 Definição Operacional de Clustering

Anil K. Jain, com base nas suas numerosas e extensas pesquisas na área do agrupamento de dados, propôs a seguinte definição operacional de *Clustering* (Jain, 2010): dada uma representação de n objectos, encontrar k clusters (ou grupos) com base numa medida de semelhança, de tal forma que as semelhanças entre os objectos pertencentes ao mesmo *cluster* sejam elevadas e as semelhanças entre objectos afectos a *clusters* diferentes sejam reduzidas. No fundo, o *Clustering* pretende descobrir grupos cuja inércia intra-grupo seja reduzida e cuja inércia inter-grupo seja elevada.

Para efeitos deste trabalho, o *Clustering* é definido da seguinte forma (Definição 1):

Definição 1 - CLUSTERING:
Dado o conjunto de dados D, *composto por* N *observações (pontos ou registos),* $D = \{\vec{x_1}, \vec{x_2}, ..., \vec{x_N}\}$ *e sendo a i-ésima observação* $\vec{x_i}$ $(i = 1, ..., N)$ *definida como um vector de atributos numéricos no espaço d-dimensional* $\vec{x_i} = (x_{i,1}, x_{i,2}, ..., x_{i,d})$ *um Clustering* ξ *é uma partição específica do conjunto de dados* D *em* K *partições, usualmente denominadas por* clusters, $\xi = \{C_1, ..., C_i, ..., C_K\}$, *onde cada* cluster *é um conjunto de observações* $C_k = \left\{\vec{x_{k_1}}, \vec{x_{k_2}}, ..., \vec{x_{k_{n_k}}}\right\}$, *com* $\sum n_k = N$ *e* $n_k \geq 1$, *para* $K = 1, 2, ..., k$.

O Clustering ξ *é definido de tal forma que:*

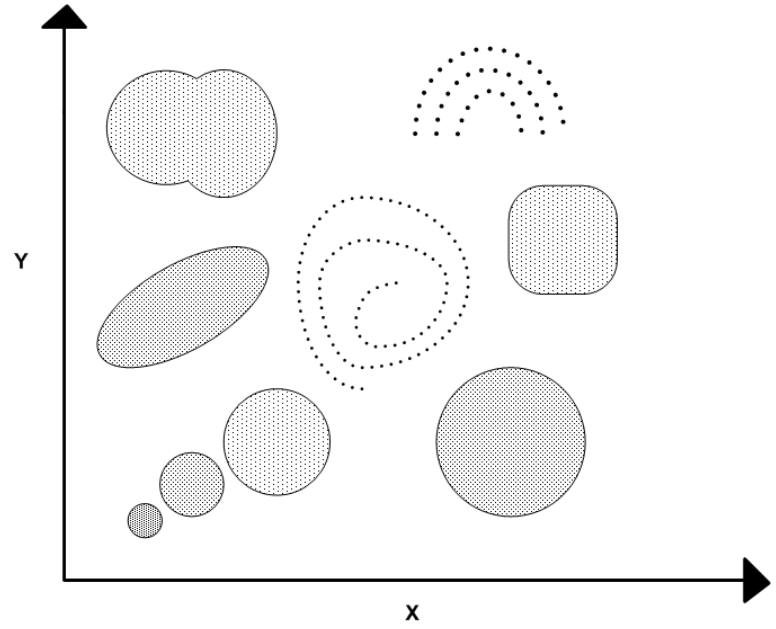


Figura 2.1: *Clusters* de várias formas, dimensões e densidades representados num espaço bidimensional

1. $C_i \cap C_j = \emptyset$, $\forall_{i \neq j}$ - *os* clusters *são disjuntos (ou mutuamente exclusivos);*
2. $\cup_{i=1}^{K} C_i = D$ - *os* clusters *são colectivamente exaustivos;*
3. *As observações afectas a um dado* cluster *são mais semelhantes entre si do que as observações afectas aos restantes* clusters *do* Clustering ξ.

2.1.2 Conceito de Cluster

O conceito de *cluster* é um conceito basilar na Teoria do *Clustering* e pode ser entendido como um "conjunto de pontos, objectos ou observações que é compacto e isolado"(Jain, 2010), como uma "região de elevada densidade, no espaço de atributos, separada de regiões de baixa densidade"(Yip et al., 2005), como "regiões não sobrepostas descritas por um conjunto de atributos - componente de estrutura - e correspondentes a um conjunto de dados em bruto - componente de medida"(Ganti et al., 1999) ou, analogamente, como "grupos de localizações espaciais que são mais densas que os seus arredores"(Sander et al., 1998). A opção por uma ou outra definição varia consoante a perspectiva do método utilizado.

Os *clusters* podem diferir em termos de forma, densidade e dimensão. Na Figura 2.1 encontram-se retratados alguns tipos de *clusters* num espaço de atributos bidimensional, para facilitar a compreensão do conceito.

2.2 Componentes da tarefa de Clustering

A tarefa de *Clustering*, para ser desempenhada com sucesso, exige o cumprimento de um conjunto de subtarefas, que seguem uma ordem lógica. Os passos envolvidos na actividade de agrupamento de dados e a respectiva importância serão explanados, com algum detalhe, nos pontos seguintes.

A divisão apresentada é baseada na proposta de Jain et al. (1999).

2.2.1 Representação dos dados

A forma adoptada para a representação dos dados, que irão ser submetidos a um processo de *Clustering*, é um factor determinante e decisivo na qualidade e significância dos resultados obtidos. A maioria das abordagens de *Clustering* costuma adoptar uma representação baseada num vector de atributos (ou variáveis), ie, cada objecto (ou indivíduo) é caracterizado por um vector multidimensional, em que cada dimensão corresponde a um único atributo (Duda and Hart, 1973). Neste tipo de representação, a escolha dos atributos é um aspecto fulcral, pelo que é altamente aconselhável uma investigação cuidadosa dos mesmos e, se necessário, a aplicação de técnicas de selecção e extracção de atributos. A **selecção de atributos** é o processo de identificação do subconjunto de atributos mais representativo do problema em estudo, de entre todos os atributos disponíveis. Por outro lado, a **extracção de atributos** consiste na criação de novos atributos a partir de uma ou várias transformações ao conjunto original de atributos (e.g., Análise em Componentes Principais ou PCA). O recurso a estas técnicas de redução congrui para o aperfeiçoamento do desempenho do *Clustering* e para a melhoria da eficiência computacional. A importância da representação dos dados é enfatizada por Jain et al. (1999), onde esclarecem que uma boa representação pode contribuir, frequentemente, para a obtenção de estruturas simples e fáceis de compreender, enquanto que uma representação pobre dos dados poderá originar agrupamentos complexos cuja verdadeira estrutura é difícil ou impossível de discernir. No entanto, é de salientar o facto de não existirem regras universais na selecção da representação dos dados, devendo esta escolha ser guiada pelo domínio do conhecimento onde se insere o problema e pelo propósito do agrupamento.

2.2.2 Definição de uma medida de proximidade ou semelhança

Um passo fundamental na grande maioria dos procedimentos de agrupamento de dados é a escolha de uma medida de proximidade, capaz de determinar a semelhança entre pares de observações. A importância desta escolha assenta no facto da semelhança ser o conceito base do *Clustering* e um dos seus principais alicerces. Além disso, é decisiva em termos de resultados, visto que influencia fortemente a forma dos

clusters obtidos. Desta forma, a sua escolha deve ser orientada pelo conhecimento detido pelo analista sobre a área de aplicação, de modo a adoptar-se uma medida apropriada ao domínio dos dados. A semelhança entre pares de observações é, usualmente, medida através de uma função distância. Algumas das mais comummente utilizadas são as seguintes:

- **Distância Euclidiana**
  - É uma das métricas mais populares no cálculo da distância entre os pontos (ou observações) e os centróides dos *clusters*, no espaço multidimensional definido por atributos contínuos, funcionando bem quando o conjunto de dados tem *clusters* compactos e isolados;
  - Tende a encontrar *clusters* de forma esférica.
- **Distância de Mahalanobis**
  - Esta métrica assume que as densidades dos atributos seguem uma Distribuição Normal multivariada e utiliza um esquema de ponderação dos atributos com base nas respectivas variâncias e correlações lineares entre pares de atributos (Jain et al., 1999);
  - Tende a encontrar *clusters* de forma hiper-elipsoidal (Mao and Jain, 1994).
- **Distância de Minkowski**
  - Estabelece uma medida genérica para o cálculo da distância entre dois pontos no espaço d-dimensional, de acordo com o valor do parâmetro r.
- **Distância de Manhattan**
  - Calcula a distância entre dois pontos como a soma das diferenças absolutas dos valores dos atributos;
  - Apresenta a vantagem de atribuir maiores pesos às diferenças em cada dimensão.
- **MND** (*Mutual Neighbor Distance*)
  - Medida de distância que tem em consideração o efeito do contexto ou das observações vizinhas (Gowda and Krishna, 1978);
  - Contrariamente às medidas anteriormente referidas, a MND não é uma métrica porque não satisfaz as propriedades da desigualdade triangular.

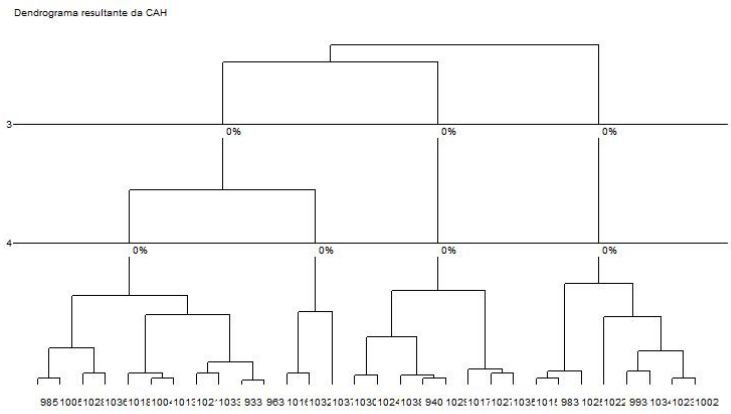


Figura 2.2: Dendrograma com partições em 3 e 4 *clusters*

2.2.3 Escolha do algoritmo de Clustering

Mais de 50 anos de investigação na área do agrupamento de dados deram origem a uma panóplia de métodos e algoritmos de *Clustering*. Dada a sua enorme diversidade, apenas serão referidos os mais preponderantes e utilizados nas tarefas de *Clustering*. Na literatura foram propostas diferentes taxonomias de técnicas de *Clustering*, embora estas apenas difiram em pequenos aspectos motivados por questões de perspectiva. A classificação de algoritmos proposta neste trabalho não pretende ser exaustiva, resultando antes de um esforço de sumarização das abordagens com maior prevalência na área, sendo seu intento fornecer uma visão geral e simplista da principal investigação realizada no âmbito do *Clustering*. Tendo por base esta filosofia, considerou-se a seguinte classificação: algoritmos hierárquicos, algoritmos particionais, algoritmos baseados na densidade, algoritmos baseados em modelos probabilísticos e algoritmos baseados na Teoria dos Grafos.

Algoritmos Hierárquicos

Os algoritmos hierárquicos, tal como o próprio nome sugere, criam uma hierarquia de *clusters* que, usualmente, é representada por meio de um dendrograma (Figura 2.2).

Ou seja, estes algoritmos produzem um conjunto de partições aninhadas, com recurso a critérios de união ou divisão de *clusters*, e com base na noção de semelhança (Jain et al., 1999). Existem duas grandes abordagens para a criação da hierarquia: a abordagem divisiva ou *top-down*, que inicializa o algoritmo considerando todos os dados como um único *cluster* e, recursivamente, procede à divisão dos *clusters* (Murtagh, 1983); e a abordagem aglomerativa ou *bottom-up*, que utiliza o proced-

imento inverso, ie, inicialmente considera cada observação como um *cluster* e, em cada etapa, reúne os dois *clusters* cuja dissemelhança é mínima. O algoritmo termina quando cada *cluster* é apenas constituído por uma observação, no caso da abordagem divisiva, ou quando todas as observações integram o mesmo *cluster*, na abordagem aglomerativa, o que, em termos gráficos, corresponde à base e ao topo do dendrograma, respectivamente. O *input* deste algoritmo é uma matriz de semelhança $n \times n$, em que n representa o número de objectos a serem agrupados, e o *output* é um conjunto de *clusters* organizados numa estrutura encaixada. Foram desenvolvidas inúmeras medidas para aferir a semelhança entre pares de observações, e índices ou critérios para a agregação dos *clusters*. A medida de semelhança mais comum é a Distância Euclidiana e, em termos de critérios de agregação, é frequente utilizar o índice do mínimo (*Single linkage*) (Sneath, 1957; Johnson, 1967) ou o índice do máximo (*Complete linkage*) (King, 1967), que definem a distância entre dois *clusters* como o mínimo ou o máximo entre todos os pares de observações retiradas de ambos os *clusters* (o par é formado por uma observação do primeiro *cluster* e outra observação do segundo *cluster*). No entanto, o índice de agregação de Ward (Ward, 1963), que se baseia no conceito de inércia e que maximiza, em cada etapa do algoritmo de *Clustering*, a inércia inter-*cluster* e minimiza a inércia intra-*cluster*, tende a alcançar melhores desempenhos no agrupamento dos dados. Contrariamente ao algoritmo das k-Médias, no método Hierárquico o número de *clusters* a reter é escolhido *a posteriori*, sendo esta decisão comummente fundamentada nas variações registadas para o índice de agregação, para diferentes partições. As principais vantagens do recurso a este método são o facto de não exigir a definição do número de *clusters* a reter *a priori*, serem mais versáteis que os algoritmos particionais e, ainda, serem extremamente intuitivos em termos de análise dos resultados, pois estes são apresentados graficamente por meio de dendrogramas. A desvantagem resume-se à falta de eficiência computacional, devido à complexidade quadrática, sendo apenas indicado na resolução de problemas relativamente pequenos.

Algoritmos Particionais

Os algoritmos particionais, em contraste com os algoritmos hierárquicos, encontram todos os *clusters* simultaneamente, como uma partição dos dados, não impondo uma estrutura hierárquica. Normalmente, este tipo de algoritmos procura identificar a partição que optimize localmente uma dada função critério (a função mais intuitiva e frequentemente utilizada é o critério do Erro Quadrado), produzindo apenas uma partição dos dados (Jain et al., 1999). O *input* destes algoritmos é uma matriz $n \times d$, onde n objectos estão incorporados num espaço de atributos d-dimensional, ou uma matriz de semelhança $n \times n$, e o *output* é um conjunto de *clusters* disjuntos cuja união cobre todo o universo de dados. O algoritmo particional mais popular e simples é o k-Médias, que foi desenvolvido e apresentado à comunidade científica pela primeira vez em 1955 (Steinhaus, 1956; Lloyd, 1982; MacQueen, 1967; Ball and Hall, 1965).

O k-Médias exige três parâmetros pré-definidos pelo utilizador antes de poder ser aplicado: o número de *clusters* k, que é a escolha mais crítica e com maiores impactos nos resultados finais (o artigo de Dubes (1987) pode funcionar como guia na tomada de decisão sobre o melhor número de *clusters*), uma partição inicial aleatória ou aproximada dos dados e uma métrica de distância. Com base nesta informação, o algoritmo é inicializado a partir da solução inicial definida (conjunto de centróides) afectando, em cada iteração, cada observação ao *cluster* mais próximo, com base na métrica de distância escolhida e de forma a minimizar o Erro Quadrado (diferença entre o centróide empírico do *cluster* e os pontos no *cluster*). Após cada afectação, o centróide do *cluster* é recalculado a partir da média das observações que, até ao momento, constituem aquele *cluster*. Ao longo do processo iterativo, estes centróides são sucessivamente melhorados até estabilizarem, ie, convergirem para uma solução. Os motivos que justificam a popularidade deste algoritmo são a sua simplicidade, facilidade de implementação, eficiência computacional, sucesso empírico e o facto de ser independente da ordem (gera a mesma partição independentemente da ordem em que os dados são apresentados) (Jain, 2010; Jain et al., 1999). Estas características tornam-no adequado para aplicações que envolvam conjuntos de dados de elevada dimensão. No entanto, também apresenta alguns inconvenientes, nomeadamente, o facto de ser bastante dependente da inicialização, quer no que respeita ao número k de *clusters*, quer em termos das soluções aleatórias iniciais; o facto de ser um algoritmo *greedy*, o que se reflecte na tendência para convergir para mínimos locais, não garantindo a melhor solução de agrupamento se a partição inicial não for bem escolhida (Lu and Huang, 2005); a sensibilidade a *outliers* e a pouca versatilidade do algoritmo que, no máximo, apenas consegue produzir *clusters* hiper-esféricos (Jain et al., 1999). Apesar das suas desvantagens, o algoritmo das k-Médias tem sido uma fonte de inspiração no desenvolvimento de novos algoritmos. Estes procuram ultrapassar as limitações do k-Médias, mas garantindo a manutenção, ou eventual reforço, das suas qualidades, nomeadamente, a eficiência e a simplicidade. Algumas variantes do k-Médias são seguidamente apresentadas:

- **Algoritmos ISODATA (Ball and Hall, 1965; Dubes, 1973) e FORGY (Forgy, 1965)** - estes algoritmos, à semelhança do k-Médias, implementam o método do Erro Quadrado, afectando cada observação a um único *cluster* (*hard assignment*)

- ***Fuzzy c-means* (Dunn, 1973)** - extensão do k-Médias, onde cada observação pode ser membro de múltiplos *clusters*, apresentando diferentes graus de pertença a cada *cluster* (*soft assignment*)

- ***Bisecting k-means* (Steinbach et al., 2000)** - versão hierárquica divisiva do k-Médias que efectua, recursivamente e em cada passo, a partição dos dados em dois *clusters*

- ***Kd-tree* (Bentley, 1975)**- criado para aumentar a eficiência na identificação dos centróides mais próximos, para todas as observações

- ***X-means* (Pelleg and Moore, 1999)** - algoritmo que encontra, de forma automática, o valor de k, por via da optimização de um critério como o AIC (*Akaike's Information Criterion*) ou o BIC (*Bayesian Information Criterion*)

- ***k-medoid* (Kaufman and Rousseeuw, 1990)**- a diferença em relação ao k-Médias é que este algoritmo utiliza a mediana dos dados em detrimento da média, por ser uma medida mais resistente (Zhang and Couloigner, 2005)

Efectuando um paralelismo com os algoritmos hierárquicos, é de sublinhar a possibilidade de transformar o *output* de um algoritmo desta natureza numa partição única dos dados, através da simples especificação de um limiar para a semelhança. Ou seja, um algoritmo hierárquico pode ser encarado como um algoritmo particional, mas com a diferença de oferecer um conjunto mais alargado de possibilidades na escolha da partição final dos dados, o que resulta do facto de não exigir uma definição *a priori* do número de *clusters*.

Algoritmos baseados na Densidade

Os algoritmos baseados na densidade foram desenvolvidos com o propósito de descobrirem *clusters* de formas arbitrárias e encaram os *clusters* como regiões de elevada densidade, procurando directamente regiões densas conectadas no espaço de atributos. O DBSCAN (Sander et al., 1998), OPTICS (Ankerst et al., 1999) e o algoritmo de Jarvis-Patrick são exemplos de algoritmos assentes nesta filosofia. A desvantagem associada aos algoritmos baseados na densidade é a dificuldade em lidar com dados multidimensionais, visto que, quando os dados são de elevada dimensionalidade, o espaço de atributos tende a ser esparso, dificultando a tarefa de distinção entre regiões de densidade alta e reduzida (Jain, 2010). Os algoritmos de *Subspace Clustering* foram criados para colmatar esta dificuldade, como é o caso, por exemplo, do algoritmo CLIQUE (Agrawal et al., 2005).

Algoritmos baseados em Modelos Probabilísticos

Estes algoritmos assumem que os dados são gerados a partir de uma distribuição mistura, onde cada *cluster* é descrito por uma ou mais componentes de mistura. O desafio consiste, assim, em descobrir as distribuições de proveniência dos dados e os respectivos parâmetros. A maioria da investigação realizada nesta área assumiu que os componentes individuais da densidade mistura são Gaussianos (Jain et al., 1999) e, nestes casos, os procedimentos focam-se na estimação dos parâmetros com base numa abordagem *data-driven*. Exemplos de algoritmos que têm por base esta ideia são o algoritmo EM (*Expectation Maximization*) (Dempster et al., 1977) para

dados em falta, o algoritmo LDA (*Latent Dirichlet Allocation*) (Blei et al., 2003) e o algoritmo *Pachinko Allocation Model* (Li and McCallum, 2006).

Algoritmos baseados na Teoria dos Grafos

Diversos algoritmos de *Clustering* foram concebidos com base em técnicas e ideias importadas da Teoria dos Grafos. Este tipo de algoritmos representa os pontos como nós num grafo pesado, sendo a ponderação atribuída às arestas que conectam os nós baseada na semelhança existente entre pares de nós. A ideia central consiste em efectuar a partição dos nós em dois subconjuntos A e B, de tal forma que a dimensão do corte, ie, a soma dos pesos das arestas que ligam os nós A e B, seja minimizado (Jain, 2010). Existem inúmeros algoritmos que implementam esta ideia, nomeadamente, os algoritmos *Minimum Cut* , *Ratio Cut* (Hagen and Kahng, 1992), *Normalized Cut* (Shi and Malik, 2000), *Modified Normalized Cut* (Meila and Shi, 2001) e *Laplacian Eigenmap* (Belkin and Niyogi, 2001).

Comparação de Algoritmos

Dada a diversidade de estruturas passíveis de serem encontradas em diferentes conjuntos de dados multidimensionais, e os inúmeros objectivos que induzem o analista a aplicar técnicas de *Clustering*, é difícil, ou mesmo impossível, qualificar um algoritmo como perfeito ou melhor que os outros. Isto porque, a melhor solução varia consoante os dados e o problema em estudo. Por outro lado, e dado que o *Clustering* é um processo não-supervisionado, os vários algoritmos existentes baseiam-se em determinadas assumpções de forma a serem capazes de obter um agrupamento dos dados. Por conseguinte, assumem diversos comportamentos e originam diferentes resultados consoante **(i)** os atributos ou variáveis considerados no conjunto de dados, que influenciam a geometria e distribuição de densidade dos *clusters*, e **(ii)** os valores dos parâmetros de entrada (Halkidi et al., 2002). Tendo consciência da relatividade do desempenho dos algoritmos neste contexto e da natureza subjectiva da tarefa de *Clustering*, é fundamental explorar e experimentar diversas abordagens para determinar o algoritmo de *Clustering* apropriado à tarefa e aos dados em causa.

2.2.4 Abstracção dos dados

No contexto do *Clustering*, a abstracção dos dados consiste na extracção de uma representação simples e compacta da estrutura de *clusters* obtida (Jain et al., 1999). A criação de descrições compactas, simples e intuitivas dos *clusters* finais é extremamente útil em aplicações em que se pretende descobrir o número de classes e em aplicações que envolvem a tomada de decisão, uma vez que aumenta a capacidade de compreensão humana dos resultados do agrupamento, ajuda a alcançar um maior

conhecimento sobre os dados, que pode posteriormente ser explorado por computador, e poderá, eventualmente, contribuir para o aumento da eficiência das tomadas de decisão. Os esquemas de representação de *clusters* mais comummente utilizados são os seguintes (Duran and Odell, 1974; Michalski et al., 1981; Diday and Simon, 1976):

- Centróide
- Pontos mais distantes ou Pontos fronteira
- Nós numa árvore de decisão
- Expressões conjuntivas lógicas

2.2.5 Avaliação do resultado final

A avaliação é um passo muito importante na tarefa de *Clustering* enquanto forma de validar e assegurar a significância dos *clusters* obtidos, no contexto do problema em estudo. Como enunciado no trabalho de Jain et al. (1999), a necessidade de avaliação surgiu, por um lado, devido à constatação de que a disponibilização de dados a qualquer algoritmo de *Clustering* tem como resultado um conjunto de *clusters*, independentemente dos dados possuírem, ou não, grupos naturais e, por outro lado, devido ao facto de alguns algoritmos se comportarem melhor que outros em determinados problemas. Desta forma, para realizar uma avaliação completa é fundamental analisar, em primeiro lugar, o domínio dos dados, para verificar se estes apresentam algum tipo de tendência de agrupamento; em segundo lugar, efectuar a validação dos *clusters* finais, para garantir que estes não são resultado directo do acaso nem um artefacto do algoritmo; e, por fim, estudar a estabilidade dos *clusters*, ie, a capacidade de generalização do algoritmo de *Clustering* adoptado, através da medição da quantidade de variação na solução de agrupamento sobre diferentes sub-amostras retiradas dos dados iniciais.

A validação dos *clusters* refere-se aos procedimentos formais que permitem avaliar os resultados da Análise de *Clusters* e seleccionar o esquema que melhor se ajusta aos dados, de uma forma quantitativa e objectiva (Halkidi et al., 2002), podendo assentar em critérios internos, externos ou relativos. Os critérios internos procuram determinar se a estrutura é intrinsecamente apropriada para os dados em estudo. Por sua vez, os critérios externos comparam a estrutura obtida com informação detida *a priori* (por exemplo, as verdadeiras classes a que pertence cada observação). Por fim, os critérios relativos comparam duas ou mais estruturas e medem o seu mérito relativo em determinado aspecto.

Kleinberg (2003) e Fisher and vanNess (1971) propuseram, também, um conjunto de critérios de admissibilidade para algoritmos de *Clustering*, que têm como objectivo testar a sensibilidade e o comportamento dos algoritmos a mudanças que

não alterem a estrutura dos dados. Um pequeno resumo do trabalho destes autores pode ser consultado na recente pesquisa de Jain (2010).

2.3 Investigação recente na área do Clustering

A investigação na área do *Clustering* tem vindo a evoluir, procurando explorar novas metodologias e esforçando-se por ultrapassar alguns obstáculos colocados pelas abordagens tradicionais. Neste ponto pretende-se desvendar, de forma superficial, algumas das tendências emergentes nesta área, nomeadamente os *Clustering Ensembles*, o *Clustering* semi-supervisionado e o *Clustering* de grande escala.

2.3.1 Clustering Ensembles

Os *Clustering Ensembles* procuram conjugar o conhecimento proporcionado por diversas partições dos mesmos dados, para obter uma partição de qualidade superior, construída com base nas sub-estruturas comuns a todas elas. Múltiplas partições - *Clustering Ensembles* - podem ser geradas por via da aplicação de diferentes algoritmos de *Clustering*, pela aplicação do mesmo algoritmo mas com diferentes valores de parâmetros ou inicializações ou, ainda, através da combinação de diferentes representações de dados e algoritmos de *Clustering* (Iam-on et al., 2008). Estas partições são posteriormente comparadas através de uma nova medida de semelhança: a co-ocorrência (número de vezes que dois pontos co-ocorrem no mesmo *cluster*) e, com base nesta medida, é criada uma nova partição composta pelos factores comuns a todas as partições. O objectivo é melhorar a qualidade das estruturas finais de *Clustering* e captar *clusters* de forma atípica, que não são compactos nem bem separados.

2.3.2 Clustering semi-supervisionado

A natureza *data-driven* do *Clustering* dificulta o desenvolvimento de algoritmos capazes de extrair a verdadeira estrutura dos dados. Assim, toda a informação adicional que esteja disponível, ou seja possível recolher, pode revelar-se decisiva e uma mais-valia na criação de boas partições dos dados. É o recurso a este tipo de informação que distingue os processos não-supervisionados dos processos semi-supervisionados. A especificação da informação adicional é realizada através de restrições entre pares de observações, que podem ser do tipo ***must-link*** (as observações em causa pertencem ao mesmo *cluster*) ou do tipo ***cannot-link*** (as observações pertencem a *clusters* distintos) (Chapelle et al., 2006). Normalmente, estas restrições são criadas e fornecidas pelo especialista no domínio de conhecimento onde se insere o problema em estudo.

2.3.3 Clustering de grande escala

Aplicações como a Segmentação de Imagem e o *Information Retrieval* criaram novos desafios à tarefa de *Clustering*, uma vez que exigem o agrupamento de enormes colecções de dados. As abordagens tradicionais não conseguem estar à altura das exigências colocadas pela elevada dimensão dos conjuntos de dados, estando restringidas a conjuntos relativamente pequenos. Neste sentido, surgiu a necessidade de conceber novos algoritmos e/ou aperfeiçoar algoritmos já existentes, para solucionar esta nova categoria de problemas.

Os estudos realizados neste âmbito são variados e podem ser classificados em cinco categorias principais (Jain, 2010):

- Sumarização dos dados - a ideia base consiste em melhorar a eficiência computacional através da sumarização de um conjunto de dados de elevada dimensão num subconjunto relativamente pequeno, o qual é posteriormente submetido a um algoritmo de *Clustering*;

- *Clustering* Incremental - os avanços registados no *Data Mining* fomentaram o desenvolvimento de algoritmos incrementais, que se baseiam na assumpção de que é possível considerar as observações, uma de cada vez, e atribuí-las sequencialmente aos *clusters* existentes; nestes algoritmos, a matriz original dos dados é armazenada numa memória secundária e as observações são transferidas, sequencialmente, para a memória principal, para serem submetidas ao processo de *Clustering*. Exemplos: COBWEB, BIRCH (Zhang et al., 1996), LSEARCH (O'Callaghan et al., 2002);

- Métodos baseados em amostragem - estes métodos escolhem, selectivamente, uma amostra representativa do conjunto de dados de elevada dimensão, aplicam o algoritmo de *Clustering* a essa amostra e, mais tarde, inferem a estrutura de *clusters* final com base na estrutura obtida para a amostra.

Os avanços registados nas tecnologias de armazenamento de dados, o recurso diário a ferramentas de pesquisa na Internet e a difusão, potenciada pela redução dos custos, da utilização de instrumentos como as câmaras de vídeo e os RFID foram responsáveis pela criação de volumes abismais de dados de elevada dimensionalidade. Aliado à dimensão, surge um novo desafio relacionado com o carácter dinâmico dos dados. Actualmente, o conceito de *Data Streams* ou Fluxo Contínuo de Dados tem vindo a ganhar força e é visível um novo rumo de investigação nesse sentido. Os Fluxos Contínuos de Dados são um tipo de dados dinâmicos, de natureza efémera, que não podem ser armazenados num disco rígido devido às suas características. Estas resumem-se ao acesso sequencial, elevado volume e potencial tamanho ilimitado, sendo dados dinamicamente evolutivos que, contrariamente aos dados estáticos, mudam ao longo do tempo. Estes factores impõem requisitos adicionais

aos algoritmos tradicionais de *Clustering*, que devem ser dotados da capacidade de processamento e sumarização de grandes quantidades de dados, que chegam continuamente. Além disso, o seu carácter dinâmico exige, também, a habilidade dos algoritmos se adaptarem às mudanças que, eventualmente, ocorram na distribuição subjacente aos dados, bem como a aptidão de detectar *clusters* emergentes, unir *clusters* antigos ou descartar *clusters* desactualizados (Barbará, 2002). Os algoritmos incrementais anteriormente mencionados (COBWEB, BIRCH, LSEARCH) foram desenvolvidos com o intuito de efectuarem o *Clustering* em ambientes de fluxos contínuos de dados.

Após a exposição e análise das abordagens com maior prevalência na área do *Clustering*, já dispomos de informação suficiente para fundamentar a escolha dos algoritmos que irão ser utilizados para obter o *input* deste estudo. Tendo por base o conhecimento adquirido, pretendemos realizar experiências recorrendo a, pelo menos, dois tipos de *Clustering*, com a finalidade de provar que o método a propor é independente dos algoritmos de *Clustering* adoptados. Para efeitos deste trabalho, iremos proceder à construção de *clusters* com base em algoritmos de partição e em algoritmos hierárquicos, mais especificamente, utilizando o algoritmo das k -médias, para diferentes valores de k, e o Método de *Clustering* Hierárquico Aglomerativo ou Ascendente (*bottom-up*), utilizando como índice de agregação o índice do máximo (*Complete Linkage*) que, em geral, produz hierarquias mais úteis do que o índice do mínimo (*Single Linkage*) (Jain and Dubes, 1988). As razões subjacentes a esta selecção assentam no facto de estes serem os algoritmos mais populares e utilizados, serem eficientes, uma vez que obtêm, na maioria dos casos, *clusters* de boa qualidade (Lu and Huang, 2005) e, ainda, por serem métodos pertencentes a diferentes classes de *Clustering* de dados, o que garante uma correcta comparação das experiências e a obtenção de conclusões válidas sobre a independência do método proposto.

Capítulo 3

Evolução de Clusters - Estado da Arte

Recentemente, devido à natureza dinâmica da maioria dos conjuntos de dados, emergiram novos métodos e técnicas para manter e actualizar conhecimento anteriormente descoberto (Baron et al., 2003). Os esforços de investigação na área do *Clustering* têm sido, sobretudo, direccionados para o problema da adaptação dos *clusters* às populações que sofreram mudanças na respectiva distribuição (Spiliopoulou et al., 2006). A motivação subjacente assenta numa alteração de paradigma, motivada pela evolução, e consequente constatação da obsolescência e do carácter efémero dos modelos tradicionais de *Data Mining*. A incessante evolução dos dados coloca, desta forma, novos desafios à descoberta de conhecimento nos dados, exigindo a adopção de novas perspectivas, nomeadamente, perspectivas orientadas para o tempo. O tempo surge, assim, como uma dimensão adicional através do qual as populações de dados podem evoluir e sofrer mudanças. O paradigma do *Change Mining* emerge como consequência deste novo rumo na investigação, englobando mecanismos de *Data Mining* que monitorizam modelos e padrões ao longo do tempo, comparamos, detectam, interpretam e quantificam mudanças de acordo com o seu interesse (Böttcher et al., 2008). O ênfase deste paradigma é colocado na modelação e compreensão da mudança, em detrimento do mero ajustamento de modelos ou padrões. De modo análogo, este trabalho preocupa-se com a compreensão da natureza das próprias mudanças, alcançada por via da tipificação e monitorização das transições de *clusters*.

A apresentação das metodologias e métodos propostos, até ao momento, pelos investigadores com interesse na área emergente da detecção e monitorização de mudanças ocorridas em estruturas de dados, é o *leitmotiv* do presente capítulo. Este será encetado com a exposição das principais taxonomias definidas para a classificação das transições passíveis de ocorrer em *clusters* sendo, posteriormente, explanados os métodos e algoritmos considerados relevantes no âmbito da detecção e/ou monitorização de transições em estruturas de dados. Os referidos métodos serão classificados

consoante o objectivo do respectivo desenvolvimento, procedendo-se a uma divisão assente nos ambientes para os quais foram projectados: ambientes de conjuntos de dados estáticos (comparação de *snapshots*) e ambientes de conjuntos de dados dinâmicos ou Fluxos Contínuos de Dados. Por fim, é de sublinhar que a exposição dos métodos e algoritmos será realizada de forma sucinta e tendo em mente apenas os aspectos considerados pertinentes no âmbito do tema em estudo.

3.1 Taxonomias das transições de clusters

Várias abordagens foram propostas pela comunidade científica neste contexto e existem, pelo menos, oito esquemas taxonómicos para a classificação de transições em *clusters*, padrões ou conceitos que evoluem ao longo do tempo.

Falkowski et al. (2006) previram quatro tipos de transições típicas, no âmbito do estudo sobre a evolução de redes sociais, nomeadamente, na abordagem proposta para analisar a evolução experienciada por subgrupos de indivíduos, pertencentes a comunidades com estruturas de participação relativamente estáveis. Neste contexto, mencionaram como possíveis transições o **Crescimento**, o **Declínio**, a **União** e a **Divisão** dos subgrupos ou *clusters*. Para observar estas transições, os autores efectuaram a monitorização e acompanhamento de cada subgrupo ao longo do tempo, através da medição da respectiva equivalência estrutural, avaliada com recurso a medidas como a distância Euclidiana e o coeficiente de correlação.

Chen and Liu (2006) conceberam um método para a detecção de mudança em estruturas de *Clustering* para Fluxos Contínuos de Dados categóricos, sendo a ocorrência de mudança indicada pela alteração do número óptimo de *clusters* k. Os autores consideram que estas mudanças podem ser impelidas pela **Emergência** de novos *clusters* ou, ainda, pelo **Desaparecimento** de *clusters* provocado pela convergência de *clusters* em crescimento.

Com o intuito de estudarem o problema do *Clustering* de objectos móveis, Li et al. (2004b) conceberam o algoritmo MMC (*Moving Micro-Cluster*), que encerra em si dois objectivos primordiais: agrupar eficientemente objectos espácio-temporais, garantindo a produção de *clusters* de elevada qualidade, e detectar mudanças interessantes ocorridas nos padrões durante o processo de movimento. No que respeita a este último objectivo, e tendo consciência de que os objectos contidos no interior dos micro-clusters, num dado instante temporal, apresentam forte tendência para se deslocarem em direcções distintas num futuro próximo, os autores definiram dois tipos de eventos: eventos de **Divisão** e eventos de **Colisão**. Para auxiliar a detecção destes acontecimentos, procedeu-se ao encapsulamento dos micro-clusters em rectângulos, que crescem com o tempo. Assim, os eventos de Divisão correspondem a situações em que a altura ou largura do rectângulo atinge um determinado limiar, tendo como consequência a divisão dos micro-clusters. Analogamente, os eventos de Colisão são caracterizados pelo "choque"de um par de micro-clusters e consequente

sobreposição dos rectângulos atinentes a cada um.

No estudo de Yang et al. (2005), que incide sobre a mineração de padrões espácio-temporais em dados científicos, foram discernidos três tipos de eventos na descrição do comportamento evolucionário dos SOAP's (*Spatial Objects Association Patterns*) ou *clusters*: **Formação**, **Dissipação** e **Continuação**. A Formação de SOAP's ocorre quando o número de instâncias de um dado SOAP passa de zero a um número positivo. Por sua vez, a Dissipação é caracterizada pela invalidez de todas as instâncias de um determinado SOAP, num dado instante temporal. O evento de Continuação é assinalado pela existência de, pelo menos, uma instância de um dado SOAP em dois instantes temporais adjacentes. Com base na definição destes três tipos de eventos é possível retirar ilações sobre a estabilidade dos SOAP's ao longo do tempo. Neste contexto, os autores sugerem, ainda, a construção de episódios espácio-temporais, caracterizados por pares de eventos de Formação e Dissipação, e utilizam-nos para inferir eventos críticos na estrutura dos dados.

Aggarwal (2005) considera a existência de três tipos diferentes de mudanças ou tendências: **Coagulação** dos dados, **Dissolução** dos dados e **Deslocação** dos dados. Esta taxonomia é construída com base no conceito de densidade da velocidade, que mede a taxa de mudança da concentração dos dados, numa dada localização espacial, e num horizonte temporal pré-definido. Assim, a Coagulação dos dados pode ser definida como uma região com elevações, que apresenta uma densidade de velocidade superior a um determinado limiar, a Dissolução dos dados refere-se a regiões com depressões e cuja densidade de velocidade é inferior a um dado limiar, e a Deslocação dos dados é detectada por meio da identificação de linhas, que efectuam a ligação entre os epicentros de dissolução e coagulação, fornecendo uma ideia dos movimentos e direcções dos dados, no espaço e ao longo do tempo.

Num trabalho de investigação posterior, Aggarwal (2003) estudaram e edificaram uma nova estrutura de *Clustering* centrada em aplicações de Fluxos Contínuos de Dados. O método *CluStream*, contrariamente a outras abordagens semelhantes, oferece uma grande flexibilidade ao analista na descoberta e na exploração dos *clusters* em horizontes temporais distintos. Desta forma, o utilizador pode analisar a evolução da estrutura de *Clustering* no período temporal que lhe suscita maior interesse, podendo detectar eventos de **Eliminação** de micro-clusters, **Adição** de micro-clusters ou **Manutenção** de micro-clusters.

Kaur et al. (2009) propuseram outra taxonomia, orientada para a detecção de mudanças de conceito em Fluxos Contínuos de Dados não classificados, que assenta na premissa de que a taxa de chegada dos dados a um conceito ou *cluster* é um bom indicador da sua evolução. Neste âmbito, foram definidos cinco tipos de conceitos: **conceito Novo**, **conceito Consistente**, **conceito Emergente**, **conceito Enfraquecido** e **conceito Aleatório**. Um conceito Novo é um conceito recentemente descoberto e suportado por poucas observações. Por sua vez, um conceito é Consistente se a taxa de chegada, no respectivo processo, não varia de modo significativo. Este tipo de conceito, que prevalece constante ao longo do tempo, raramente é alvo

de interesse. Por outro lado, conceitos Emergentes são caracterizados por taxas de chegada crescentes, ie, o número de observações que apoiam o conceito aumenta com o tempo, contribuindo para a respectiva consolidação. De modo análogo, um conceito Enfraquecido desvanece-se com o tempo, através de uma clara perda de suporte do número de observações que a ele pertencem. Este efeito degradativo manifesta-se por meio de taxas de chegada decrescentes. Por fim, um conceito é Aleatório se não puder ser categorizado como nenhum dos tipos mencionados, e se a respectiva amostra de taxas de chegada exibir, apenas, aleatoriedade. Estes investigadores prevêem, também, a possibilidade dos estados dos conceitos sofrerem alterações ao longo do respectivo tempo de vida.

Outra abordagem, mais complexa e detalhada, foi apresentada no contexto da metodologia MONIC, desenvolvida por Spiliopoulou et al. (2006). Os autores do MONIC propuseram uma tipificação das transições que podem ser experienciadas por um *cluster*, em termos das respectivas características intrínsecas e extrínsecas, e o seu desenvolvimento teve como intuito descrever a natureza da evolução de cada conceito. Com base no estudo destes investigadores, é possível depreender a existência de duas grandes classes de transições: as transições internas ou intra-*cluster*, que dizem respeito ao conteúdo e à forma do *cluster* e apenas são monitorizadas nos *clusters* que sobrevivem; e as transições externas ou inter-*cluster*, que se referem ao relacionamento do *cluster* com todo o *Clustering*, ie, analisam as alterações que ocorreram no *cluster* como membro integrante de um esquema mais geral que é o *Clustering*. Os tipos previstos de transições externas são os seguintes: **Sobrevivência**, **Absorção**, **Divisão**, **Emergência** ou **Desaparecimento**. Ou seja, um *cluster* pode ser absorvido por outro(s) *cluster(s)*, pode dividir-se em mais do que um *cluster*, pode simplesmente desaparecer ou, ainda, não sofrer qualquer uma destas mudanças e sobreviver. Também é possível detectar o surgimento de novos *clusters*. A definição destas mudanças assenta nos conceitos de sobreposição e combinação de *clusters*. Por sua vez, as transições internas podem ser desencadeadas por mudanças na dimensão, na compactação e/ou na localização dos *clusters* que sobrevivem, ou seja, os *clusters* podem expandir-se ou contrair-se, tornarem-se mais compactos ou mais dilatados e, ainda, sofrer alterações em termos do respectivo centróide ou distribuição. Também se podem verificar situações em que nenhuma mudança é registada e, nesses casos, o *cluster* do período temporal t_i é exactamente o mesmo que o *cluster* do período temporal seguinte t_{i+1}. À semelhança do modelo desenvolvido por Kaur et al. (2009), a metodologia MONIC considera a possibilidade dos *clusters* experienciarem tipos distintos de transições, ao longo do seu tempo de vida. Porém, não se limita, apenas, à constatação deste facto, preocupando-se, também, em explorar esta informação de modo a ser capaz de extrair conclusões sobre a mutabilidade, estabilidade e tempo de vida dos *clusters*.

3.2 Algoritmos de detecção de transições

O carácter dinâmico da maioria dos dados estimulou a tomada de novos rumos na investigação, sendo visível esse esforço nas áreas de Aprendizagem Automática e Análise de Dados. Actualmente, existem diversos algoritmos que, directa ou indirectamente, almejam captar e, sobretudo, compreender a natureza dinâmica destes conjuntos de dados, particularmente susceptíveis à ocorrência de mudanças na estrutura subjacente. A última década tem sido especialmente profusa em matéria de concepção de algoritmos para a detecção de transições, sendo estes pautados pela diversidade. Com base no estudo realizado, foi possível depreender uma classificação preliminar e simples para os algoritmos construídos neste contexto. De um modo geral, existem algoritmos projectados para operar em ambientes relativamente estáticos (*snapshots*) ou altamente dinâmicos (*Data Streams*). No que respeita aos primeiros, estes podem ser focados nas transições experienciadas por padrões genéricos, *clusters* ou regras de associação. No âmbito dos *Data Streams* ou Fluxos Contínuos de Dados, são bastante comuns as abordagens com enfoque em dados não classificados, emergindo, inclusive, novas técnicas para o agrupamento de objectos móveis ou espácio-temporais. Estas abordagens elegem como principal estrutura de dados os *clusters* e preocupam-se com a eficiência e escalabilidade dos algoritmos.

De seguida, será realizada uma exposição sucinta dos algoritmos considerados mais pertinentes no âmbito deste estudo, e concedendo-se especial destaque ao funcionamento geral do algoritmo e às medidas eventualmente utilizadas para aferir a natureza das mudanças.

3.2.1 Conjuntos de dados estáticos

A detecção de mudanças através da comparação de instantes temporais (*snapshots*) de dados é uma função importante em inúmeras aplicações (Chawathe and Garcia-Molina, 1997), nomeadamente no Marketing e na Detecção de Fraude (Spiliopoulou et al., 2006). Nesta secção são apresentados os principais algoritmos desenvolvidos com o intuito de comparar instantes temporais distintos (por exemplo, comparar a estrutura de dados obtida em 2008 com a estrutura de dados obtida em 2009). Estes encontram-se classificados de acordo com a estrutura de dados utilizada.

Padrões genéricos

Metodologia FOCUS: O FOCUS é uma metodologia de detecção de mudanças, concebida por Ganti et al. (1999), vocacionada para a quantificação da diferença ou desvio entre dois conjuntos de dados. A ideia central do método resulta da constatação de que uma grande parte dos modelos de *Data Mining* (por exemplo, árvores de decisão, itens frequentes ou *clusters*) podem ser descritos por uma Componente de Estrutura e por uma Componente de Medida. Com base nestas componentes,

obtidas a partir dos modelos, é possível calcular o desvio entre dois conjuntos de dados e quantificar as diferenças entre eles, sendo o grau de significância da diferença aferido com recurso a técnicas estatísticas, nomeadamente, testes de hipóteses *standard*.

O inconveniente do FOCUS é o facto de ser muito superficial e genérico, não permitindo o alcance de um conhecimento mais profundo sobre as diferenças entre dois conjuntos de dados e a respectiva natureza.

Metodologia PANDA: Bartolini et al. (2004, 2009) propuseram uma metodologia genérica e flexível, em termos de tipo de padrões e de critérios de semelhança, para comparação de padrões simples e padrões complexos, denominada PANDA. Neste contexto, um padrão complexo deve ser entendido como um padrão construído com base noutros padrões. A semelhança entre padrões é calculada por meio de um operador de semelhança que tem em conta, quer a semelhança entre as estruturas dos padrões, quer a semelhança entre as medidas. Esta semelhança pode ser calculada por meio de uma função de agregação. Os padrões são caracterizados, assim, por duas componentes: a Componente de Estrutura, que define o espaço do padrão, e a Componente de Medida, que descreve as medidas que quantificam a qualidade da representação da fonte dos dados alcançada por cada padrão. Por exemplo, no padrão simples, a estrutura pode ser representada pelo centro e pelo raio do *cluster* e as medidas podem incluir a Distância Média Intra-*cluster* e o Suporte do *Cluster* (fracção de dados representado por aquele *cluster*).

Quando os padrões são complexos, a semelhança entre as respectivas estruturas depende, em parte, da semelhança entre os padrões simples que os compõem. Nestes casos, a semelhança entre as estruturas dos padrões é avaliada conceptualmente através de duas abstracções fundamentais: Tipo de emparelhamento - *coupling type* -, utilizado para estabelecer a forma como os padrões componentes podem ser combinados; e Lógica de agregação - *aggregation logic* -, utilizada para combinar os scores de semelhança obtidos para padrões componentes emparelhados num único score, que representa a semelhança entre padrões complexos.

O PANDA apresenta a desvantagem de apenas se concentrar na realização de comparações genéricas e eficientes de padrões, em detrimento da detecção e interpretação das transições.

Algoritmo MH-DIFF: O algoritmo MH-DIFF é um algoritmo eficiente, desenvolvido por Chawathe and Garcia-Molina (1997), para detectar mudanças significativas em dados hierarquicamente estruturados em forma de árvore. O problema de detecção de mudanças entre estruturas construídas em diferentes instantes temporais, é encarado pelos autores como o problema de encontrar a melhor forma de editar a representação da árvore criada em t_1, para obter a representação de outra árvore, obtida em t_2, com $t_1 < t_2$. Para que esta transformação possa ser operada

é necessário definir operações de edição, susceptíveis de serem aplicadas a dados estruturados. As operações de edição previstas pelos autores são as seguintes:

- **Inserção** - uma operação de inserção cria um novo nó, com uma dada etiqueta, e coloca-o num dada posição na árvore;
- **Eliminação** - esta operação elimina um nó da árvore;
- **Actualização** - a operação de actualização altera a etiqueta do nó da árvore;
- **Mover** - esta operação move uma sub-árvore, com raíz num dado nó, para outra posição na árvore;
- **Cópia** - a operação de cópia, tal como o nome indica, copia a sub-árvore, com raiz num dado nó, para outra posição;
- **Colagem** - esta operação é o inverso de uma operação de cópia;

Uma sequência de operações de edição é denominada *script* de edição. O *script* de edição mínimo é a sequência de operações estritamente necessárias para transformar uma árvore noutra árvore, sendo obtido a partir do Modelo de Custos. Este modelo foi concebido pelos autores como uma forma de alcançar a sequência óptima de operações e, concomitantemente, eliminar a ambiguidade inerente à escolha do *script* de edição.

Este algoritmo, apesar de não ser totalmente genérico, uma vez que é restringido a dados estruturados, poderá ser considerado como tal, dado que pode ser aplicado a todo o tipo de dados que apresentem alguma estrutura. Além disso, a ideia base do método apresenta um contributo importante para este estudo, ao permitir deduzir que duas estruturas de *Clustering* diferentes, obtidas em instantes temporais distintos, podem ser transformadas através de pequenas mudanças estruturais.

Regras de associação

Modelo GRM: O GRM (*Generic Rule Model*) foi inicialmente apresentado por Steffan Baron e Myra Spiliopoulou no trabalho intitulado *Monitoring Change in Mining Results* (Baron and Spiliopoulou, 2001) tendo, posteriormente, sido adoptado pelos autores na criação de estruturas de actualização e monitorização de conhecimento mais complexas. Este modelo foi motivado pela necessidade de construir conhecimento sustentável, reforçando a importância e indispensabilidade da monitorização e acompanhamento da evolução do conhecimento, ao longo de uma dimensão temporal. Segundo os autores, a mera actualização do conhecimento não é, por si só, suficiente para criar sustentabilidade. Tendo presente esta ideia, edificaram o GRM, que permite representar temporalmente padrões, ao modelizar e efectuar o tratamento integrado do conteúdo e das propriedades estatísticas das regras de associação. Com base nestas duas componentes (conteúdo e propriedades estatísticas),

definiram-se, formalmente, dois tipos diferentes de evolução de padrões: a Emergência e a Extinção, com o intuito de observar a evolução das regras de associação ao longo do tempo.

$$R = (ID, query, timestamp, statistics, body, head)$$

A representação temporal das regras de associação, supra retratada, tem uma assinatura composta pelos seguintes elementos:

- **ID** - identificador único do padrão
- ***Q*uery** - conjunto de restrições que limitam o nº de regras descobertas como, por exemplo, limite inferior de suporte e confiança e limite superior para o tamanho da regra. Para os padrões serem comparáveis têm de ser provenientes de uma mesma query.
- **Timestamp** - instante de tempo em que a regra de associação foi produzida
- **Statistics** - armazena as estatísticas atinentes à regra de associação como um todo, e.g., o valor do suporte e da confiança da regra de associação.
- **Body** - antecedente da regra de associação
- **Head** - consequente da regra de associação

O GRM, apesar de ter sido inicialmente desenvolvido para modelizar regras de associação, cobre resultados baseados em diferentes paradigmas de mineração, como sejam, sequências frequentes e *clusters*, o que o torna bastante apelativo no âmbito deste trabalho.

Monitor PAM: O PAM (*Automated Pattern Monitor*) é uma metodologia genérica para monitorização de padrões e detecção de mudanças interessantes, que apresenta a vantagem de ser eficiente em termos computacionais (Baron and Spiliopoulou, 2004). A eficiência resulta do facto do PAM ser dotado da capacidade de actualizar o conteúdo dos padrões sem exigir a reaplicação completa do algoritmo de *Data Mining* a todo o conjunto de dados. A actualização é efectuada com base nas estatísticas dos padrões considerados interessantes e a monitorização recai, apenas, sobre um subconjunto dos padrões descobertos, sendo a detecção de mudanças na população assente na análise do impacto nas estatísticas. Desta forma, consegue-se reduzir significativamente o esforço em termos de *Data Mining*, sem comprometer a fiabilidade da informação contida nos dados.

A metodologia utilizada é baseada na representação temporal de padrões GRM e decompõe-se em duas fases: na primeira fase, o conjunto de dados é submetido

a um algoritmo de *Data Mining*, para efeitos de descoberta das regras de associação que são, posteriormente, importadas para o monitor e armazenadas de acordo com o GRM; na segunda fase, procede-se à selecção das regras que irão ser monitorizadas, com base na respectiva representatividade da estrutura da população, estabelecem-se grupos de regras claramente relacionadas em termos de conteúdo, especificam-se as estatísticas a observar e os correspondentes limiares de alerta e, por fim, monitorizam-se, regularmente, os grupos de regras especificados, através da comparação das estatísticas actualizadas com os limiares pré-definidos. A condução deste processo possibilita a detecção de mudanças significativas, cuja eventual ocorrência exige uma reaplicação do algoritmo de *Data Mining*, para efeitos de perscrutação das causas subjacentes. Assim, é possível comprovar que a frequência de mineração é drasticamente reduzida, visto que passa a ser realizada apenas nos instantes temporais que justifiquem uma inspecção mais minuciosa do conjunto de dados.

Porém, o GRM apresenta a desvantagem de não possibilitar a detecção directa de novos padrões, uma vez que apenas as estatísticas das regras conhecidas são calculadas. Outro inconveniente resume-se ao facto de ter sido especialmente projectado para o entendimento das mudanças ocorridas em bases de regras de associação.

Clusters

Metodologia MONIC: A metodologia MONIC (Spiliopoulou et al., 2006) é a abordagem da literatura com as características mais próximas às que se pretendem impregnar no presente estudo, debruçando-se sobre a modelação e monitorização das transições de clusters. É constituída por duas componentes: taxonomia ou tipificação das transições passíveis de serem experienciadas por *clusters*, quer a nível interno, quer a nível externo; e algoritmo de detecção das transições. Para as transições detectadas entre duas estruturas de *Clustering*, os autores construíram métricas para concluir sobre o tempo de vida e estabilidade dos *clusters* e do *Clustering*, ao longo da dimensão temporal.

A ideia base que norteou a concepção e desenvolvimento desta metodologia resultou da consciencialização de que, mais importante que detectar a mudança, é compreender a sua natureza. Com este intuito, foi concebida uma taxonomia capaz de descrever a natureza das mudanças e que consagra transições de dois níveis: transições internas, experienciadas por cada *cluster*, como entidade isolada; e transições externas, que captam a evolução da estrutura de *Clustering* como um todo. Na primeira categoria, estão previstas mudanças na dimensão, localização e compactação/difusão, ie, mudanças atinentes ao conteúdo e à forma do *cluster*, sendo apenas monitorizadas em *clusters* que sobrevivem de um instante temporal para outro. A segunda categoria, por sua vez, analisa as alterações que ocorreram no *cluster* como membro integrante de um esquema mais geral que é o *Clustering*, prevendo a Sobrevivência, Absorção, Divisão, Emergência e Desaparecimento de *clus-*

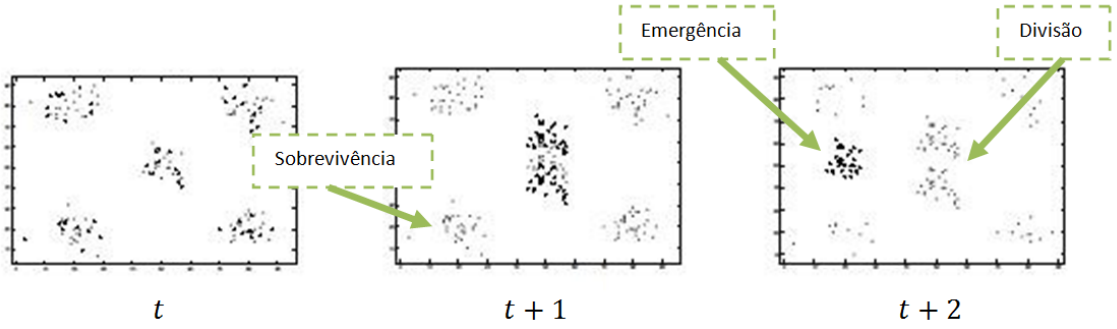


Figura 3.1: Estruturas de *Clustering* obtidas em três instantes temporais diferentes e respectivas transições externas. Os pontos mais escuros correspondem a observações mais recentes (Spiliopoulou et al., 2006)

ters. É de salientar que o MONIC apenas permite detectar a emergência de novos *clusters* devido ao facto de assumir re-*Clustering*, em detrimento de adaptação dos *clusters* em cada período temporal. Na Figura 3.1 encontram-se retratadas algumas destas transições.

Tendo por base esta taxonomia, foi desenvolvido o algoritmo de detecção de transições, que se encontra alicerçado em dois conceitos fundamentais: *cluster overlap*, ou sobreposição de *clusters*, e *cluster match*, ou combinação de *clusters*. O objectivo subjacente à criação destes conceitos é, sobretudo, perceber quais os *clusters* do instante temporal t_2 que correspondem aos *clusters* do instante temporal t_1, com $t_2 > t_1$, para depois poder concluir sobre as mudanças ocorridas na passagem de uma estrutura para outra. Após a detecção das transições, calcula-se, para cada entidade isolada, o respectivo tempo de vida, definido como o número de períodos temporais em que o *cluster* sobreviveu, possivelmente com transições internas. Para avaliar a estabilidade da população analisa-se a sobrevivência do *Clustering*, que se reflecte na quantidade de *clusters* que sobreviveu no *Clustering* construído no período temporal seguinte. O estudo do tempo de vida permite concluir sobre a existência de populações estáveis, ie, com poucas variações e elevada percentagem de sobrevivências, ou de populações voláteis, caracterizadas pela frequência das transições.

Esta metodologia apresenta algumas peculiaridades, nomeadamente, o facto de incorporar uma Função de Envelhecimento dos dados, que especifica os pesos que devem ser atribuídos às observações processadas pelo algoritmo de *Clustering*, com o objectivo de diminuir o impacto das observações antigas na construção da estrutura de agrupamento e, ainda, o facto de não exigir que o espaço de atributos seja invariante ao longo do tempo, tornando-o particularmente adequado para o *Text Stream Mining*. O principal problema da metodologia proposta radica na ineficiência do algoritmo, em virtude do recurso a todas as observações para efectuar o *matching* dos *clusters*. Para contornar esta dificuldade os autores sugerem a utilização futura

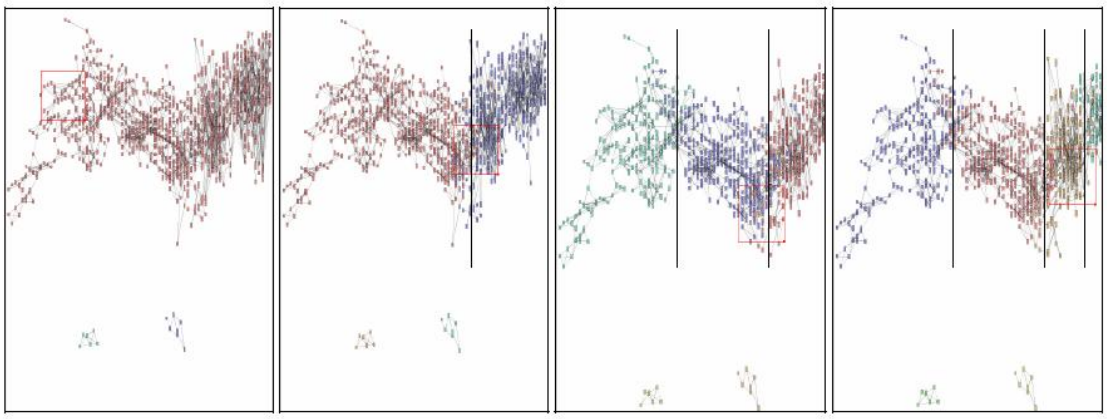


Figura 3.2: Visualização dos *clusters* e das respectivas rupturas estruturais representadas por linhas verticais (Falkowski et al., 2006)

de sumários de dados, que apresentam vantagens em termos de confidencialidade dos dados e cumprimento de exigências de memória computacional.

Monitor de Comunidades: O Monitor de Comunidades, proposto por Falkowski et al. (2006), visa essencialmente analisar a evolução de dois tipos de comunidades distintos e detectar rupturas na respectiva evolução. O interesse por um estudo deste género foi despoletado pela importância crescente das redes sociais e pela consequente necessidade de determinação dos factores, externos ou internos, que constituem a causa das respectivas dinâmicas e impelem um certo desenvolvimento destas redes. Para tratar este problema, os autores propuseram duas abordagens para analisar a evolução de dois tipos diferentes de comunidades, ao nível dos subgrupos: o primeiro método foca-se em comunidades que apresentam uma estrutura de participação de membros relativamente estável, e assenta em visualizações e análises estatísticas que permitem uma análise interactiva das evoluções experienciadas pelos subgrupos; o segundo método foi concebido para a detecção de comunidades em ambientes de elevada flutuação de membros, conseguida por meio do agrupamento de instâncias de comunidade semelhantes e posterior visualização temporal, para efeitos de detecção de mudanças e rupturas no padrão evolutivo. Note-se que, neste contexto, os subgrupos ou instâncias de comunidade são grupos de indivíduos que interagem entre si como membros de comunidades online, daí terem sido consagrados na subsecção referente aos *clusters*.

Cada abordagem comporta uma sequência de etapas que irão ser sucintamente explanadas. Na primeira abordagem, o eixo temporal é particionado em janelas temporais deslizantes sobrepostas, para permitir uma análise temporal da rede. Seguidamente, constrói-se um grafo pesado de interacções entre indivíduos para cada janela

temporal, o qual será submetido a um algoritmo de *Clustering* Hierárquico Divisivo para encontrar subgrupos de indivíduos densamente conectados. Por fim, procede-se à monitorização de cada subgrupo detectado, para efeitos de análise do respectivo desenvolvimento temporal. Para suportar esta análise e, consequentemente, possibilitar a detecção de transições (crescimento, declínio, união ou divisão de subgrupos), os autores recorrem à medição e interpretação de um vasto conjunto de medidas estatísticas, nomeadamente, a estabilidade, a densidade, a coesão, a distância euclidiana entre dois subgrupos, o coeficiente de correlação e a actividade do grupo, que permitem aferir o padrão evolutivo dos subgrupos. A segunda abordagem, por incidir sobre comunidades com elevada flutuação de membros, exige um tratamento especial e distinto. O carácter oscilante dos membros obriga à introdução de técnicas para apuramento da semelhança entre subgrupos. Neste sentido, os autores definem a semelhança entre instâncias de comunidade, descobertas em períodos temporais distintos, como o grau de sobreposição de membros entre as duas instâncias. Esse constitui o primeiro passo deste segundo método. Posteriormente, a noção de semelhança é utilizada para criar um grafo de instâncias de comunidades, o qual é submetido a um algoritmo de *Clustering* Hierárquico Divisivo, para efeitos de detecção de grupos de instâncias de comunidades semelhantes. Por fim, procede-se à visualização da evolução dos *clusters* ou subgrupos, utilizando-se cores diferentes para cada *cluster* de modo a facilitar a detecção das rupturas indicativas das mudanças e transições ocorridas nas comunidades (Figura 3.2).

A vantagem deste monitor assenta no facto de ter em consideração diferentes graus de dinâmica de indivíduos, o que se revela extremamente útil em termos de aplicação. A apresentação de métodos alternativos aumenta a flexibilidade prática, ao permitir abordar o problema de um modo mais adaptado às respectivas peculiaridades. À semelhança da metodologia MONIC, apresenta o inconveniente de não trabalhar com informação sumária, nomeadamente, no apuramento da semelhança entre indivíduos de diferentes períodos temporais.

Metodologia para conjuntos de dados científicos: Yang et al, no trabalho apresentado em (Yang et al., 2005), dirigiram os seus esforços na concepção de uma metodologia genérica, vocacionada para a descoberta de diferentes padrões de interacção entre objectos espaciais e para a posterior derivação de episódios espácio-temporais, com base na incorporação de informação temporal na análise geral. Esta metodologia é especialmente adequada para o tratamento e análise de conjuntos de dados científicos.

Na metodologia é possível discernir duas fases: identificação de vários tipos de interacção entre os atributos, ou SOAP's (*Spatial Object Association Patterns*) e construção de episódios espácio-temporais. Os SOAP's podem ser encarados como *clusters* que, ao longo da dimensão temporal, podem dissipar-se, manter-se ou formar-se. Os episódios espácio-temporais estão sustentados nesta taxonomia (Formação,

Dissipação e Continuação) e são caracterizados por pares de eventos de formação e dissipação de SOAP's. Com base nestes episódios os autores inferem eventos críticos e, através da combinação dos episódios associados aos diferentes SOAP's, conseguem modelar o comportamento evolucionário das interacções entre os atributos.

3.2.2 Conjuntos de dados dinâmicos ou *Data Streams*

A relevância da análise eficiente de Fluxos Contínuos de Dados tem sido fomentada pelo incremento assinalável do número de transacções realizadas e armazenadas, de forma automática, nas empresas, e consequente crescimento ilimitado das Bases de Dados (Aggarwal, 2005). O aumento exponencial do volume de dados e as restrições computacionais, ao nível da memória e processamento, induziram os investigadores à criação de métodos e algoritmos que actualizam os resultados de mineração sem examinar novamente todo o conjunto de dados - actualização incremental (Baron and Spiliopoulou, 2001). O desenvolvimento de algoritmos dotados de incrementalidade é uma actividade cada vez mais comum na investigação sendo que, nos últimos anos, se tem concedido especial destaque à criação de algoritmos eficientes para a execução de tarefas de aprendizagem não-supervisionada, nomeadamente, tarefas de *Clustering*. Alguns métodos, além da eficiência, oferecem mecanismos adicionais para monitorização das mudanças em tempo real. Isto porque, o *Clustering* de *Data Streams* pode fornecer importantes pistas sobre a emergência de novos padrões de dados, úteis para os agentes de decisão poderem prever os eventos que irão ocorrer e agirem em conformidade e atempadamente (Chen and Liu, 2006). A construção de métodos para agrupamento das observações em ambientes de *Data Streams* é um trabalho com algumas exigências. De acordo com Lu e Huang (Lu and Huang, 2005), os algoritmos de *Clustering* para Fluxos Contínuos de Dados devem ser incrementais, eficientes, e dotados da capacidade de análise em espaços multidimensionais, bem como da capacidade de geração de *clusters* de elevada qualidade, sem necessidade de recorrer às observações mais antigas.

Seguidamente, serão apresentadas as linhas gerais e as características mais relevantes de alguns métodos desenvolvidos neste âmbito.

Algoritmo EDS

O método proposto por Kaur et al. (2009) foi criado com o intuito de compreender a evolução de um Fluxo Contínuo de Dados multidimensional, uniforme e não classificado, ao longo do tempo, através da detecção de mudanças de conceito. Os autores apresentam uma abordagem focada no estudo dos padrões de chegadas, que implicou a definição de uma taxonomia de vários tipos de mudanças de conceito e respectivas relações entre eles, bem como a modelação das transições que ocorrem no ciclo de vida de um conceito, e a criação de um algoritmo incremental de detecção de mudança para Fluxos Contínuos de Dados não classificados, denominado EDS

(*Evolution in Data Stream*). A premissa básica do algoritmo proposto é que a taxa de chegada dos dados num conceito (*cluster*) é um bom indicador da sua evolução.

A metodologia consiste na criação e desenvolvimento de uma componente *online* capaz de processar os dados e capturar a taxa de chegada para cada conceito (*cluster*) em intervalos de tempo aleatórios. O processamento do fluxo é efectuado com recurso a grelhas, ou colecções de hiper-cubóides. A amostragem aleatória de conceitos no fluxo contínuo de dados garante a obtenção de amostras i.i.d. de observações de taxas de chegada, com distribuição desconhecida. Após se ter acumulado uma amostra aleatória de dimensão razoável, testes estatísticos não-paramétricos são aplicados à amostra, com o objectivo de categorizar os conceitos de acordo com a taxonomia definida: conceito Novo, conceito Consistente, conceito Emergente, conceito Enfraquecido e conceito Aleatório.

Os autores do EDS, além de proporem uma taxonomia bastante completa, prevêem a possibilidade dos *clusters* experienciarem diferentes estados, ao longo do seu tempo de vida. Assim, consideram que um *cluster* inicia a sua vida como um Conceito Novo, que corresponde à fase em que é descoberto pela primeira vez mas em que ainda não existe uma amostra completa que permita testar a natureza da sua evolução. Posteriormente, pode tornar-se um conceito Emergente, Enfraquecido ou Consistente, dependendo da taxa a que os novos dados se juntam ao conceito. Um conceito Emergente pode manter o seu estado ou mudar para Consistente ou Enfraquecido. Um conceito Consistente persiste enquanto a taxa de chegada dos dados a este conceito se mantiver relativamente constante. De outra forma, o conceito sofre mutação e é categorizado noutro estado.

Metodologia CluStream

A metodologia CluStream foi desenvolvida por Aggarwal (2003), no âmbito do estudo do problema do *Clustering* para aplicações de *Data Streams*. A principal inovação deste trabalho consiste na introdução de flexibilidade e interactividade no algoritmo de *Clustering*, uma vez que oferece ao utilizador a possibilidade de este decidir qual o horizonte temporal no qual pretende realizar o agrupamento dos dados e, posteriormente, incidir a sua análise. Esta filosofia é, assim, centrada na aplicação, preocupando-se com as necessidades reais do utilizador. A ideia base assenta na divisão do processo de *Clustering* em duas componentes (ou camadas): a componente *online* e a componente *offline*.

A fase inicial corresponde à manutenção *online* dos micro-*clusters* e é a fase em que o algoritmo procede à recolha dos dados estatísticos, em tempo real. A escolha de micro-*clusters* é justificada, sobretudo, pela sua propriedade aditiva, o que os torna bastante apelativos em ambientes de *Data Streams*. Além disso, os micro-*clusters* permitem manter as estatísticas a um nível suficientemente elevado de granularidade, quer a nível temporal, quer a nível espacial, sem comprometerem a eficiência do algoritmo. Esta fase inclui, também, o armazenamento periódico

das estatísticas sumárias detalhadas, relativas aos micro-*clusters*, na memória do computador. Esta periodicidade segue um padrão piramidal que assegura que os *clusters*, em qualquer horizonte temporal pré-definido pelo utilizador, podem ser aproximados.

A segunda fase corresponde à criação dos macro-*clusters*, que podem ser entendidos como os *clusters* tipicamente encontrados por algoritmos de agrupamento. Nesta fase, assume-se que o utilizador especifica o número de *clusters* que pretende obter e o horizonte temporal a considerar na operação. Com base nestes parâmetros, e no sumário estatístico compacto obtido na operação de armazenamento periódico da fase *online*, é possível criar um *Clustering* adequado às necessidades de informação do utilizador. Adicionalmente, o CluStream permite a exploração da natureza da evolução dos *clusters* em diferentes períodos temporais. Para tal, basta o utilizador repetir o processo, para um horizonte temporal diferente, e o algoritmo devolve novos dados, nomeadamente, o número de micro-*clusters* adicionados, o número de micro-*clusters* eliminados e o número de micro-*clusters* mantidos, de um período temporal para outro.

Os autores conseguiram provar que o CluStream é efectivo para fluxos que evoluem e para fluxos estáveis e, também, que é um algoritmo mais eficiente, fiável e preciso que o algoritmo STREAM (O'Callaghan et al., 2002), que se baseia em todo o histórico do *Data Stream*, não permitindo aproximações de horizontes temporais.

Metodologia para visualização e diagnóstico de mudanças em Fluxos Contínuos de Dados

Com o intuito de auxiliar os utilizadores a adquirir capacidades de compreensão e entendimento profundo das tendências emergentes nos fenómenos, Aggarwal estudou o problema da evolução dos dados e apresentou uma ferramenta genérica para compreensão, visualização e diagnóstico da natureza das mudanças ocorridas nas características dos dados (Aggarwal, 2005). Neste estudo, Aggarwal propôs uma metodologia para realização de diagnósticos de Fluxos Contínuos de Dados multidimensionais, recorrendo, para o efeito, ao conceito de estimação da densidade da velocidade. Esta densidade pode ser utilizada para criar dois tipos de perfis visuais da evolução dos dados - *perfis da velocidade temporal* e *perfis da velocidade espacial* - que fornecem diferentes perspectivas da natureza da mudança subjacente ao fenómeno. Enquanto o *perfil de velocidade temporal* oferece uma perspectiva global da taxa de mudança das densidades, ao longo de um dado período temporal, numa localização espacial fixa, o *perfil de velocidade espacial* oferece uma perspectiva global das reorganizações na densidade relativa dos dados, em diferentes pontos e num período temporal fixo, permitindo obter uma compreensão profunda de como os dados estão a mudar e que direcções estão a seguir no espaço, ao longo do tempo. Estes conceitos são a base de todo o método, sendo que a sua análise possibilita a detecção de mudanças e a respectiva classificação, de acordo com a taxonomia

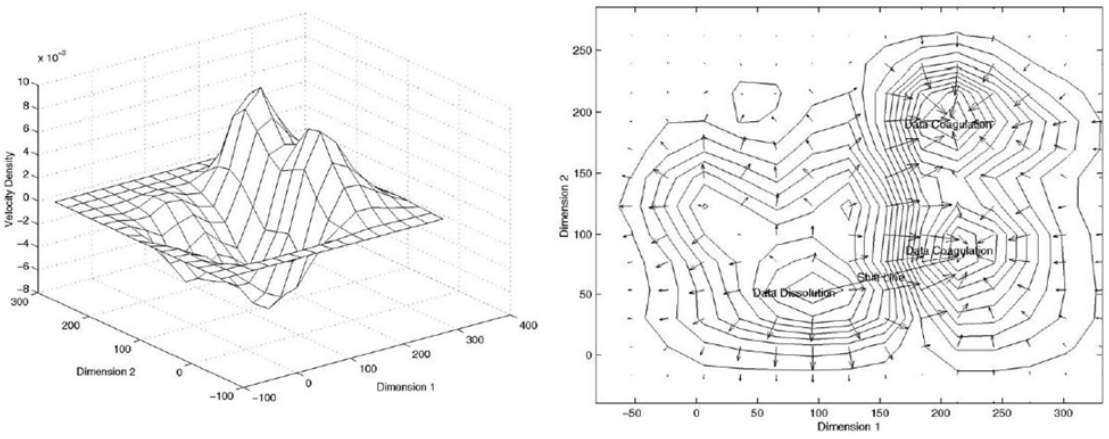


Figura 3.3: Perfil de velocidade temporal e Perfil de velocidade espacial em espaços bidimensionais (Aggarwal, 2005)

definida pelo autor: Coagulação, Dissolução e Deslocação dos dados.

No que concerne à componente visual do método, é importante salientar alguns aspectos, nomeadamente, o facto da visualização dos perfis ser inerentemente bidimensional. Esta constatação acarreta alguns problemas, uma vez que implica a selecção das localizações espaciais e dos horizontes temporais onde ocorrem as mudanças mais significativas nos dados, dada a impossibilidade de visualizar os dados em mais do que duas ou três dimensões. O que acontece mais frequentemente é a inexistência de mudança em todos os conjuntos de dimensões, mas a ocorrência de mudança apenas em combinações particulares das mesmas. Tendo consciência desta situação, o autor descarta a hipótese de medir, simultaneamente, a mudança em todo o conjunto de dimensões e sugere o cálculo da densidade de velocidade apenas para cada combinação de dimensões, de forma isolada. A combinação de dimensões que apresentar a evolução mais considerável é seleccionada, automaticamente, pelo método, para visualização bidimensional. Ou seja, nas combinações de atributos mais significativas, constroem-se gráficos para visualização dos *perfis temporais* e dos *perfis espaciais*. Os *perfis temporais* são visualizados através de gráficos de densidade tridimensionais, e os *perfis espaciais* através de gráficos de contorno, como é possível comprovar na Figura 3.3.

A metodologia dispõe, também, de uma componente dotada da habilidade de realizar diagnósticos, concisos e resumidos, de tendências específicas dos dados, em dadas localizações espaciais. Esta característica do método emergiu da necessidade dos analistas em conhecerem, num dado horizonte temporal, as localizações espaciais nas quais os dados estão a reduzir - Dissolução dos dados -, nas quais estão a aumentar - Coagulação dos dados - ou, ainda, nas quais os dados se estão deslocar

para outras localizações - Deslocação dos dados. Através desta taxonomia é possível caracterizar a natureza da evolução dos dados por meio da detecção de tendências emergentes.

Esta técnica, baseada na estimação da densidade da velocidade, possui as vantagens de ser escalável, eficiente e permitir uma boa perspectiva visual da natureza da evolução dos dados.

Metodologia de Clustering para Fluxos Contínuos de Dados categóricos

A metodologia proposta por Chen e Liu foi desenvolvida para resolver o problema do *Clustering* de *Data Streams*, quando os atributos que caracterizam o fluxo de dados são categóricos. O carácter desafiante e interessante deste problema tem origem na constatação de que a maioria das aplicações, com interesse nesta área (por exemplo, monitorização de desastres, anti-terrorismo e detecção de intrusos na rede), incluem grandes quantidades de dados categóricos (Chen and Liu, 2006). Adicionalmente, a metodologia procura contemplar uma das fases mais decisivas da tarefa de *Clustering*: a avaliação e validação *online* do *Clustering* obtido, que resulta na descoberta do número óptimo de clusters k. Esta preocupação advém do facto da maioria dos algoritmos de *Clustering*, desenvolvidos para Fluxos Contínuos de Dados, simplificar bastante o problema, ao assumir que o número óptimo k é dado (Barbará et al., 2002).

Tendo por base estes objectivos essenciais, os autores apresentam uma metodologia para a detecção da mudança do melhor número de k, que é um factor indicativo da ocorrência de mudança nas estruturas de *Clustering*, em Fluxos Contínuos de Dados categóricos. Esta metodologia estende o trabalho realizado na determinação do melhor número de k, para conjuntos de dados estáticos - método BkPlot (Chen and Liu, 2005), a Fluxos Contínuos de Dados categóricos, com a ajuda de uma estrutura de sumarização em forma de árvore, denominada HE-Tree (*Hierarchical Entropy Tree*). Para apurar a semelhança entre pares de observações, os autores recorrem a medidas de semelhança baseadas na pureza do conjunto de dados. Devido à inexistência de uma definição intuitiva de distância, para valores categóricos, recentemente a Entropia, que é um conceito oriundo da Teoria da Informação, tem sido aplicada no *Clustering* de dados categóricos (Barbará et al., 2002; Li et al., 2004a; Chakrabarti et al., 2004; Dhillon et al., 2003). Os autores adoptaram, assim, o critério da Entropia Esperada para efectuar o agrupamento dos dados categóricos.

A ideia chave da metodologia proposta é a combinação do método BkPlot com a estrutura HE-Tree. A HE-Tree foi concebida como uma metodologia eficiente, que utiliza uma pequena quantidade de memória para sumarizar a propriedade de entropia dos Fluxos Contínuos de Dados, e agrupa os registos de dados num conjunto de sub-*clusters* localizados nas folhas da HE-Tree. O algoritmo ACE (Chen and Liu, 2005) estendido consegue lidar com os sub-*clusters* e gerar um instante aproximado do BkPlot para a identificação do melhor k num dado intervalo de tempo. A infor-

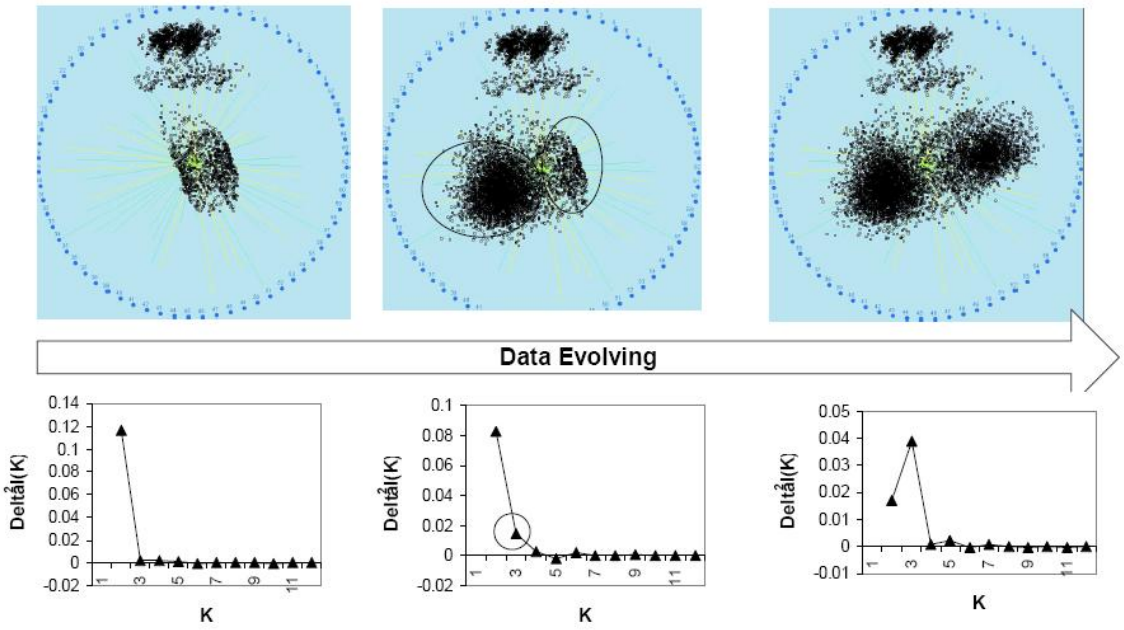


Figura 3.4: Monitorização de um conjunto de dados (Chen and Liu, 2006)

mação gerada pelo BkPlot permite identificar as mudanças ocorridas na estrutura de *Clustering*, que podem ser convenientemente detectadas através da comparação de pontos distintivos nos BkPlots respeitantes a diferentes instantes temporais. A alteração do número de k pode resultar da emergência de novos *clusters* ou do desaparecimento de *clusters* provocado pela convergência de *clusters* em crescimento. Na Figura 3.4 é ilustrado o processo de monitorização realizado com recurso a esta metodologia.

A mais-valia da metodologia de Chen e Liu é o facto de possibilitar a detecção célere da mudança (ou número k de *clusters*) em estruturas de *Clustering*, o que contribui para a eliminação do trabalho custoso de avaliação de *clusters* e, ainda, o facto de possibilitar a monitorização eficiente dos *Data Streams* categóricos, uma vez que apenas se procede à análise individual dos clusters quando é detectada alguma mudança na estrutura de agrupamento. No entanto, a metodologia também possui algumas desvantagens, nomeadamente, o facto de não gerar, automaticamente, informação sobre o tipo de transições ocorridas nos *Clusterings* e nos *clusters*, de um instante temporal para outro.

3.2.3 Algoritmos de Clustering de objectos móveis

A mineração de objectos móveis, ou objectos espácio-temporais, tornou-se uma área de interesse na investigação. Este interesse foi fomentado pelos constantes avanços

em tecnologias como o GPS, computadores portáteis e dispositivos de comunicação sem fio (Elnekave et al., 2007). Actualmente, a investigação tem-se focado na concepção de algoritmos incrementais de *Clustering*, para agrupamento deste tipo específico de objectos, e consequente detecção de padrões de movimentos semelhantes ao longo de uma dimensão temporal. A descoberta destes padrões pode influenciar significativamente diversos campos de conhecimento, nomeadamente a Ecologia, através do estudo da evolução da migração de grupos de animais, os Sistemas de vigilância do trânsito, por via da identificação de áreas densas de tráfego automóvel, a Previsão Meteorológica, entre outros (Kalnis et al., 2005).

Os trabalhos realizados neste âmbito, apesar de distintos, apresentam algumas semelhanças, nomeadamente, o facto de recorrem a históricos de trajectórias espácio-temporais dos objectos (cada objecto móvel é descrito por uma trajectória) e terem como objectivo encontrar *clusters* móveis de boa qualidade, ie, "grupos ou conjuntos de objectos, que se movem próximos uns dos outros por um longo período de tempo"(Kalnis et al., 2005).

Li et al. (2004b) apresentaram uma solução que emprega micro-*clusters* móveis e que é especialmente adequada para conjuntos de dados de elevada dimensão. Nanni and Pedreschi (2006) desenvolveram um algoritmo de *Clustering* baseado na densidade, que agrupa trajectórias de objectos, com base na respectiva distância. Kalnis et al. (2005) dirigiram os esforços do seu estudo para a descoberta de *clusters* móveis, numa Base de Dados de trajectórias de objectos, mas adoptando uma abordagem focada na identificação e monitorização de *clusters* móveis que podem ver modificado o seu conteúdo e localização, ie, não exigem que os objectos do *cluster* sejam sempre os mesmos num dado horizonte temporal, atribuindo maior importância às características do *cluster* como um todo, como, por exemplo, a densidade. Por sua vez, Elnekave et al. (2007) procuraram inovar na representação das trajectórias dos objectos, ao torná-la mais compacta, e na definição de uma nova medida de semelhança entre trajectórias. No entanto, pouco trabalho foi dedicado à monitorização das transições ocorridas nos *clusters* móveis. Com base nesta pesquisa, apenas o trabalho de Li et al prevê a possibilidade de ocorrência de eventos de Colisão ou Divisão de micro-*clusters* móveis (Li et al., 2004b).

3.3 Análise de Dados em Painel

Além dos algoritmos apresentados, existem na literatura métodos estatísticos clássicos focados no estudo da dinâmica dos dados. Neste âmbito, a Análise de Dados em Painel (ADP) é das técnicas mais utilizadas (um bom *survey* pode ser encontrado em (Urga, 1992)).

A ADP insere-se no campo da Análise de Dados longitudinais, que procede à combinação de técnicas de Regressão com técnicas de Análise de Séries Temporais,

preocupando-se com o estudo da dinâmica das relações e com a modelação das diferenças, ou heterogeneidade, entre os indivíduos/objectos. Os dados em painel podem ser entendidos como um conjunto de objectos cujas características (ou atributos) são repetidamente observadas ao longo do tempo. A sua análise é realizada por meio de regressões e a forma mais popular de estimar modelos de dados em painel é conhecida como *modelo de componentes do erro*.

À semelhança dos métodos que são propostos neste trabalho, a ADP também tem como objectivo estudar a evolução e modelar a dinâmica dos dados. Além disso, quando os painéis de dados são balanceados, esta análise é realizada sobre o mesmo conjunto de objectos, ou indivíduos, tal como na metodologia MEC. Porém, o foco do estudo são os objectos, em detrimento dos *clusters*; recorre-se à Regressão que não fornece muita informação sobre a natureza das mudanças ocorridas; e, por fim, os seus resultados não são intuitivos e fáceis de analisar.

Capítulo 4

Monitorização da Evolução de Clusters

Neste capítulo é apresentada a metodologia MEC (*Monitorização da Evolução de Clusters*), que foi desenvolvida segundo as principais linhas de orientação do paradigma *Change Mining*, e que surge como uma proposta para resolver o problema da monitorização das transições de *clusters* ao longo do tempo, através da identificação das relações temporais entre estas estruturas. É assumido que os *clusters* podem ser representados de duas formas distintas - em extensão e em compreensão - e são adoptadas duas estratégias principais para monitorizar e classificar mudanças experienciadas pelos *clusters*, para cada representação considerada. Desta forma, a nossa metodologia comporta uma taxonomia dos vários tipos de transições de *clusters*, que podem ser endógenas ou exógenas, um mecanismo de acompanhamento que depende da forma como os *clusters* são representados e, ainda, um algoritmo de detecção de transições.

4.1 Esquemas de Representação dos Clusters

A metodologia MEC assume que os *clusters* podem ser representados recorrendo a duas grandes estratégias ou esquemas de representação: representação em extensão e representação em compreensão. Na **representação em extensão** (Definição 2), um *cluster* é caracterizado pelos seus membros, ie, pelas observações que lhe foram atribuídas por um dado algoritmo de *Clustering*.

Definição 2 - Representação em Extensão:
Seja $\vec{x_i}$*, a i-ésima observação* $(i = 1, ..., N)$*, definida como um vector de atributos numéricos no espaço d-dimensional* $(j = 1, ..., d)$*,* $\vec{x_i} = (x_{i,1}, x_{i,2}, ..., x_{i,j}, ..., x_{i,d})$*, uma possível representação temporal de* clusters *pode ser definida da seguinte maneira:*

$$C_i(t) = \{\vec{x_1}, ..., \vec{x_m}\}$$

onde m representa o número de observações afectas ao cluster $C_i(t)$, $i = (1, ..., k)$ *e* $t = (1, ..., T)$.

Este tipo de representação não envolve perda de informação e possibilita o acompanhamento do percurso de cada observação ao longo do tempo, congruindo para o alcance de resultados de transição mais fiáveis e precisos. Porém, nem sempre é possível definir os *clusters* com base neste método de representação, e.g., devido a requisitos de memória e armazenamento ou, ainda, por questões motivadas pela confidencialidade dos dados.

Alternativamente, um *cluster* pode ser caracterizado através de sumários, ie, com base em estatísticas que sumarizem as suas características internas. Esta é a ideia subjacente à **representação em compreensão** (Definição 3).

Definição 3 - Representação em Compreensão:
De acordo com esta representação, um cluster C_i *é um objecto temporal caracterizado pelas seguintes estatísticas:*

$$C_i = \{ID, t, m, sup, r, \rho, \vec{c}\}$$

onde o ID *é o identificador único de* C_i $(i = 1, ..., k)$, t *é o instante temporal onde o* cluster *aparece pela primeira vez* $(t = 1, ..., T)$, m *é o número de observações afectas a* C_i, *sup é o suporte do* cluster, *útil para avaliar a sua importância relativa,* r *é o raio do* cluster, ρ *representa a densidade do* cluster *e* $\vec{c}$ *representa o centróide do* cluster.

Para definir este tipo de representação considerou-se uma medida de *centralidade*, uma medida de *dispersão* e uma medida de *densidade*. Em favor da simplicidade adoptou-se o *centróide do cluster* como a medida de centralidade, o *raio do cluster* como a medida de dispersão e utilizou-se uma definição de densidade que assume objectos esféricos d-dimensionais. Outras medidas mais complexas poderiam ter sido utilizadas como, por exemplo, a distância de Mahalanobis. No entanto, este tipo de medidas implicam um maior custo e complexidade computacional.

Este tipo de representação compacta tem-se revelado muito apelativa numa panóplia de aplicações do mundo real, em especial nas aplicações cujo acesso a toda a informação útil e necessária é restringida, e.g. por motivos de confidencialidade dos dados. No entanto, este esquema de caracterização assume *clusters* esféricos ou em forma de *bola* (por utilizar o conceito de raio e centróide) e, além disso, implica perda de informação, o que poderá comprometer a precisão do processo de mapeamento dos *clusters*.

Este esquema de representação de *clusters* em compreensão (Definição 3) estende o *Generic Rule Model* de (Baron et al., 2003), que foi desenvolvido para modelar eficientemente o conteúdo de uma regra de associação como um objecto temporal, ao caso das estruturas de *clusters*.

As definições de **Centróide** $\vec{c}$, **Raio** r, **Suporte** sup e **Densidade** ρ, previstos na Definição 3, são os apresentados nas equações 4.1, 4.2, 4.3 e 4.4. Neste trabalho, consideraram-se as definições de Centróide e Raio presentes em Zhang et al. (1996), pelo que se assume que o Raio é a distância média das observações afectas a um dado *cluster* ao respectivo Centróide. No que concerne à Densidade, esta é calculada como a massa m (neste contexto, m corresponde ao número de objectos afectos ao *cluster*) por unidade de volume V, sendo o volume calculado de forma distinta consoante o número de variáveis, ou atributos, em causa.

$$\vec{c} = \frac{\sum_{j=1}^{n_k} \vec{x_j}}{n_k} \tag{4.1}$$

$$r = \sqrt{\frac{\sum_{j=1}^{n_k} (\vec{x_j} - \vec{c})^2}{n_k}} \tag{4.2}$$

$$sup = \frac{n_k}{N} \tag{4.3}$$

$$\rho = \frac{m}{V} \tag{4.4}$$

O cálculo do Volume V para $n = 3$ variáveis (ou atributos) é realizado utilizando a seguinte fórmula (se $n = 2$ calcula-se a área):

$$V = \frac{4}{3}\pi r^3 \tag{4.5}$$

Para $n > 3$ variáveis (ou atributos), o Volume V de uma esfera de raio r é calculado utilizando uma fórmula alternativa (Rennie, 2005):

$$V_n(r) = \begin{cases} \frac{2^{\frac{(n+1)}{2}} \pi^{\frac{(n-1)}{2}} r^n}{n(n-2)!!} & \text{se } n \text{ ímpar} \\ \frac{2\pi^{\frac{n}{2}} r^n}{n(\frac{n}{2}-1)!} & \text{se } n \text{ par} \end{cases}$$

4.2 Metodologia MEC

A metodologia MEC foi edificada com o propósito de monitorizar a evolução das estruturas de *clusters* obtidas numa sequência de instantes temporais (*snapshots*)

t, $t+1$, $t+2$,.... Neste contexto, o conceito de evolução refere-se às transições experienciadas pelos *clusters* no intervalo de tempo sob observação $[t_i, t_{i+1}]$.

Uma vez que existem, pelo menos, duas estratégias para representar *clusters*, nós desenhámos um mecanismo de acompanhamento flexível capaz de detectar, eficientemente, as transições experienciadas pelos *clusters*, independentemente da forma como estes estão representados (em extensão ou em compreensão). Deste modo, o nosso mecanismo de acompanhamento pode ser subdividido em dois métodos diferentes, cada um deles adaptado a um esquema de representação de *clusters* distinto. Nas secções seguintes apresentamos a nossa taxonomia das transições e os dois métodos desenvolvidos.

4.2.1 Taxonomia das Transições de Clusters

Como referido no Capítulo 3, na literatura existem, pelo menos, oito esquemas taxonómicos para a tipificação e classificação das transições de *clusters*, padrões ou conceitos que evoluem ao longo do tempo (Falkowski et al., 2006; Aggarwal, 2005, 2003; Chen and Liu, 2006; Kaur et al., 2009; Li et al., 2004b; Spiliopoulou et al., 2006; Yang et al., 2005). Com a finalidade de detectar as mudanças susceptíveis de ocorrer em estruturas de *clusters*, considerou-se a seguinte taxonomia: **Nascimento**, **Morte**, **Cisão**, **Fusão** e **Sobrevivência** de *clusters*. As transições previamente mencionadas são **Exógenas** (ou **Externas**), uma vez que se referem a mudanças ocorridas em todo o *Clustering*. O conceito chave na detecção e avaliação destas transições é o conceito de *mapping*, que pode ser definido como o processo de descoberta das correspondências exactas entre os *clusters* do instante temporal t_i e os *clusters* do instante temporal posterior t_{i+1}, no caso de ainda existirem. As correspondências podem ser realizadas em termos de objectos, no caso dos *clusters* representados em extensão, ou em termos das características sumárias (por exemplo,através do raio e do centróide), para representações de *clusters* em compreensão.

Por outro lado, também é possível categorizar transições **Endógenas** (ou **Internas**), ie, mudanças ocorridas e relacionadas com o conteúdo de cada *cluster*, isoladamente. Este grupo de transições apenas é monitorizado para *clusters* que experienciaram uma transição externa específica: a Sobrevivência. Com este propósito, considerámos dois tipos de transições, consoante seja afectada a cardinalidade ou a densidade do *cluster*: transições de Dimensão e transições de Compressão (Spiliopoulou et al., 2006). As primeiras prevêem a **Expansão** ou **Contracção** do *cluster*, dependendo do aumento ou diminuição da cardinalidade do *cluster* sobrevivente. Por sua vez, transições de Compressão incluem **Compactação** ou **Dispersão** do *cluster*, estando a opção por uma ou por outra dependente do incremento ou decréscimo registado na densidade do *cluster* sobrevivente. Os tipos de transições referidas podem ser facilmente detectados por via da monitorização dos dados sumários, como sejam a cardinalidade e a densidade dos *clusters* sobreviventes. Poderá, igualmente, verificar-se o caso em que nenhuma transição endógena é detectada. Nestes casos,

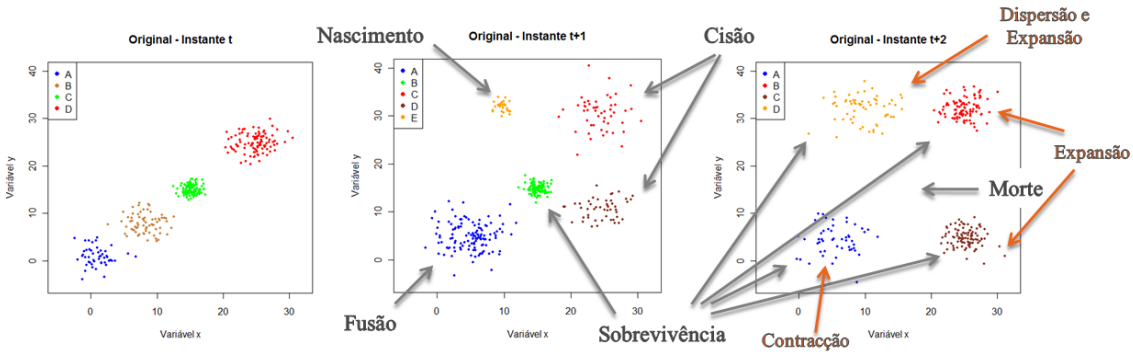


Figura 4.1: Representação bidimensional de uma sequência temporal de estruturas de *clusters* e exemplificação de alguns tipos de transições exógenas e endógenas

assume-se que o estado do *cluster* sobrevivente não sofreu qualquer tipo de modificações, ie, que se mantém constante no intervalo de tempo sob análise.

A Figura 4.1 tem como intuito ilustrar o problema abordado neste trabalho e exemplificar, de uma forma visualmente apelativa, as transições, endógenas e exógenas, previstas na taxonomia anteriormente exposta. Nestas imagens apresenta-se uma sequência de estruturas de *clusters* obtidas em instantes temporais seguidos - t, $t+1$ e $t+2$ - e representadas num espaço bidimensional. Como se pode observar, do instante de tempo t para o instante $t+1$ houve um *cluster* que nasceu (*cluster* amarelo de $t+1$), dois *clusters* que se fundiram num só *cluster* (*clusters* azul e bege de t), um *cluster* que sofreu uma cisão, ie, uma divisão em dois *clusters* (*cluster* vermelho de t) e, por fim, verifica-se a sobrevivência do *cluster* verde. No intervalo de tempo seguinte $[t+1, t+2]$, constata-se a sobrevivência de todos os *clusters*, com excepção do *cluster* verde, que morre. No que concerne às transições endógenas, verificam-se mutações dos *clusters* sobreviventes ao nível interno. Por exemplo, o *cluster* amarelo expande e torna-se mais disperso, sofrendo modificações na respectiva cardinalidade (que aumenta) e densidade (que diminui); os *clusters* vermelho e castanho aumentam o seu suporte, através do incremento da cardinalidade, e o *cluster* azul contrai-se, devido a uma ligeira redução do número de observações que o compõem.

No fundo, o escopo deste capítulo é expor e explanar os métodos concebidos para capturar este tipo de evolução de *clusters*, que se revelam extremamente úteis no aprofundamento do conhecimento sobre os fenómenos, por meio da compreensão da sua componente dinâmica.

4.2.2 Método para Clusters representados em Extensão

Como previamente mencionado, foram desenvolvidos dois métodos distintos para lidar com dois tipos diferentes de representação ou caracterização de *clusters*. Também foi sublinhada a importância do mapeamento, ou *mapping*, no processo de detecção e avaliação da natureza das transições experienciadas por um *cluster* ao longo do tempo. A sua importância deriva, essencialmente, do facto deste processo permitir descobrir quais os *clusters* do instante temporal t_{i+1} que correspondem aos *clusters* previamente encontrados no instante temporal t. Porém, as estratégias utilizadas no processo de *mapping* têm necessariamente de variar com o tipo de representação de *clusters* adoptado, dado que a natureza da informação disponível em cada um dos casos difere significativamente. Desta forma, procurámos conceber um mecanismo de acompanhamento que consagrasse processos de mapeamento adaptados a cada um dos esquemas de representação de *clusters* considerados. Neste sentido, e encetando com a representação em extensão de *clusters*, assumimos que a tarefa de *mapping* seria bem desempenhada explorando o conceito de *Probabilidade Condicionada*. A ideia consiste no cálculo das probabilidades condicionadas para todos as combinações possíveis de *clusters* pertencentes a agrupamentos, ou *Clusterings*, distintos e separados no tempo. O valor desta probabilidade fornece indicações importantes sobre a relação temporal existente entre pares de *clusters*, uma vez que indica a probabilidade do conjunto de objectos que formam o *cluster* C_i do instante t pertencer ao *cluster* C_j do instante temporal posterior $(t+1)$. Quanto maior esse valor, ie, quanto mais próximo de 1, maior o número de objectos transferidos de um *cluster* para o *outro*. Dito de outra forma, mais provável é a hipótese do *cluster* de t sobreviver no *cluster* de $t+1$. Para esclarecer as situações que devem, ou não, ser consideradas uma correspondência aproximada, introduziu-se um limiar pré-definido pelo utilizador, que denominámos **Limiar de Sobrevivência** τ, e que assume o valor mínimo de $\tau = 0.5$ (o peso da ligação entre dois *clusters* de momentos temporais diferentes tem de ser, no mínimo, 0.5 para estes poderem ser considerados uma correspondência aproximada um do outro).

O facto de se adoptar este conceito como o elemento basilar do processo de *mapping* apresenta vantagens, mas também inconvenientes. A mais-valia assenta na facilidade de implementar um processo desta natureza e na simplicidade de compreensão dos resultados. O inconveniente resulta do facto de requerer conjuntos de dados estruturalmente idênticos, ie, compostos por observações dos mesmos objectos.

Com a finalidade de melhorar o processo de *mapping* em termos visuais e, consequentemente, facilitar a posterior categorização das transições, recorreu-se ao auxílio de *Grafos Bipartidos*. A opção por este tipo de *instrumento* relaciona-se com a sua utilidade na modelação de problemas de correspondência (*matching*) e, ainda, com o facto das representações baseadas em grafos serem visualmente atractivas, explorando o poder do olhar e da intuição humana (Iam-on et al., 2008). Os alicerces do

nosso método de monitorização para *clusters* representados em extensão baseiam-se nesta ideia e podem ser definidos da seguinte maneira (Definição 4):

Definição 4 - MEC para Clusters representados em Extensão: *Dada uma sequência de estruturas de* clusters ξ_i, ξ_j $(i < j)$, *obtidas nos instantes temporais* t_i, t_j, *um grafo* $G = (U, V, E)$ *pode ser construído, onde* U *é o primeiro subconjunto de vértices, ou nós, representando os* clusters *descobertos em* t_i, V *é o segundo subconjunto de vértices, representando os* clusters *descobertos em* t_j, *e* E *denota o conjunto de arestas pesadas entre qualquer par de* clusters *pertencentes a* ξ_i *e* ξ_j. *Formalmente, o peso atribuído à aresta que conecta os* clusters $C(t_i)$ *e* $C(t_j)$ *é estimado de acordo com a seguinte probabilidade condicionada:*

$$peso(C_m(t_i), C_u(t_j)) = P(X \in C_u(t_j) | X \in C_m(t_i)) = \\ = \frac{\sum P(x \in C_m(t_i) \cap C_u(t_j))}{\sum P(x \in C_m(t_i))}$$

onde X *é o conjunto de objectos atribuídos ao* cluster $C_m(t_i)$ $(m = 1, ..., p)$ *e* $P(X \in C_u(t_j) | X \in C_m(t_i))$ *representa a probabilidade de* X *pertencer ao* cluster C_u *de* t_j *sabendo que* X *pertence ao* cluster C_m *obtido no instante temporal anterior* t_i
$C(t_i) = \{C_1, ..., C_m, ..., C_p\}$
$C(t_j) = \{C_1, ..., C_u, ..., C_r\}$.

Sublinhe-se que o grafo bipartido G não tem necessariamente de ser **balanceado**, ie, $\|U\| \neq \|V\|$, pois o número óptimo de *clusters* para cada instante de tempo (correspondente a um subconjunto de vértices) pode ser diferente. Porém, o método requer que a soma do número de observações afectas a cada *cluster* (ou vértice) do subconjunto U iguale o número de observações atribuídas aos *clusters* de V, ie, $\sum_{i=1}^{\|U\|} \|C_i\| : C_i \in U$.

Na Figura 4.2 é ilustrado um exemplo de grafo bipartido, em que o subconjunto U é formado pelos *clusters* resultantes da aplicação de um dado algoritmo de *Clustering* ao conjunto de dados do instante de tempo t e o subconjunto V é composto pelos *clusters* descobertos no instante de tempo $t+1$. Os pesos das ligações entre os vários pares de *clusters* são calculados com base na fórmula apresentada na Definição 4.

Com o intuito de detectar as mudanças, definiram-se formalmente as transições que um *cluster* $C \in \xi_i$ pode experienciar, com respeito a ξ_j, $(i < j)$. Um novo limiar foi introduzido para ajudar a definição destas transições: o **Limiar de Cisão** ρ. Este limiar revela-se extremamente útil no esclarecimento das situações que devem, ou não, ser consideradas como uma Cisão de *clusters*, assumindo-se, por defeito, o valor de $\rho = 0.2$. O estabelecimento de limiares - τ e ρ - cujos valores são

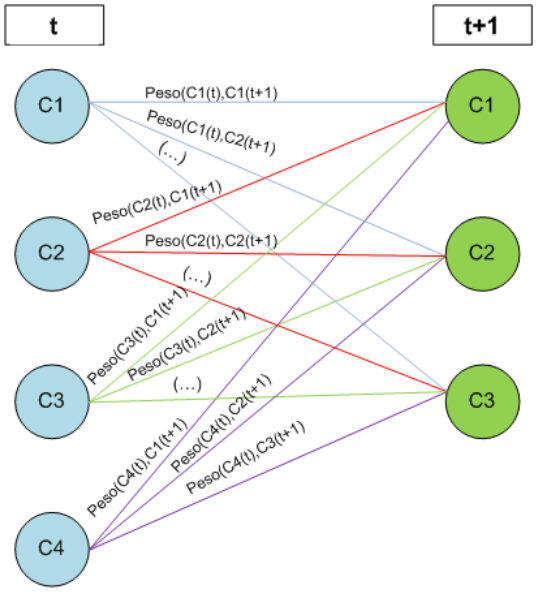


Figura 4.2: Grafo bipartido cujos vértices representam os *clusters* e cujas arestas representam a força da ligação entre *clusters* pertencentes a agrupamentos separados no tempo

passíveis de serem decididos pelo utilizador efectuou-se com o objectivo de introduzir uma maior flexibilidade no método. Este desenho formal é baseado nas transições externas da metodologia MONIC (Spiliopoulou et al., 2006) e encontra-se retratado na Tabela 4.1. O incoveniente deste método, desenhado para *clusters* representados em extensão, resume-se ao facto da monitorização baseada em transições de grafos apenas permitir a detecção de transições exógenas.

Na Tabela 4.1 é apresentada a definição formal de cada uma das transições externas previstas na taxonomia, bem como a notação a utilizar em cada um dos casos. O **Nascimento** de um *cluster* ocorre quando os pesos associados a todas as arestas conectadas a este *cluster* são inferiores ao limiar de Sobrevivência τ, ie, não é possível encontrar no universo de *clusters* do instante temporal anterior um único *cluster* que possa ser considerado uma correspondência aproximada. De modo análogo, um *cluster* **Morre** quando as suas ligações com os *clusters* do instante temporal posterior apresentam pesos inferiores ao limiar de Sobrevivência, com excepção dos casos previstos na Cisão de *clusters*. Relativamente à **Cisão** de *clusters*, assume-se que um *cluster* de t_i se divide em, pelo menos, dois *clusters* no instante t_j, se existirem, pelo menos, dois *clusters* $C_u(t_j)$ e $C_v(t_j)$ de t_j cujo peso associado às respectivas arestas seja igual ou superior ao limiar de Cisão ρ e o somatório dos respectivos pesos seja igual ou superior ao limiar de Sobrevivência. Por sua vez, a **Fusão** de *clusters* ocorre quando existem, pelo menos, dois *clusters* diferentes de t_i

Tabela 4.1: Definição formal das transições exógenas de um *cluster* representado em extensão

Taxonomia das Transições	**Notação**	**Definição Formal**
Nascimento	$\emptyset \rightarrow C_u(t_j)$	$0 < peso(C_m(t_i), C_u(t_j)) < \tau \forall_m$
Morte	$C_m(t_i) \rightarrow \emptyset$	$peso(C_m(t_i), C_u(t_j)) < \rho \forall_u$
Cisão	$C_m(t_i) \overset{\subseteq}{\rightarrow} \{C_1(t_j), ..., C_r(t_j)\}$	$(\exists_u \exists_v : peso(C_m(t_i), C_u(t_j)) \geq \rho \wedge$ $peso(C_m(t_i), C_v(t_j)) \geq \rho) \wedge$ $\sum_{u=1}^{r} peso(C_m(t_i), C_u(t_j)) \geq \tau$
Fusão	$\{C_1(t_i), ..., C_p(t_i)\} \overset{\subseteq}{\rightarrow} C_u(t_j)$	$(peso(C_m(t_i), C_u(t_j)) \geq \tau) \wedge$ $\exists C_p \in \xi_i \setminus \{C_m\} : peso(C_p(t_i), C_u(t_j)) \geq \tau$
Sobrevivência	$C_m(t_i) \rightarrow C_u(t_j)$	$(peso(C_m(t_i), C_u(t_j)) \geq \tau) \wedge$ $\nexists C_p \in \xi_i \setminus \{C_m\} : peso(C_p(t_i), C_u(t_j)) \geq \tau$

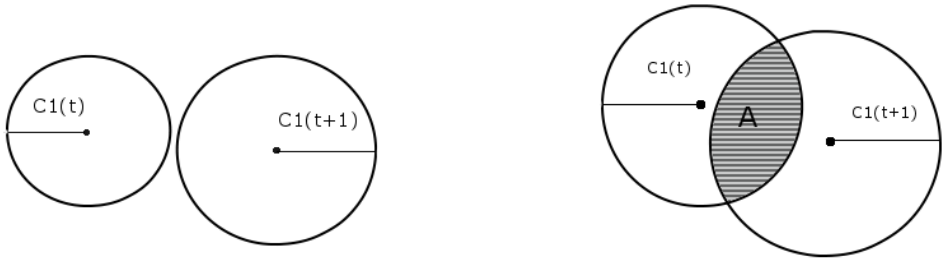
cujos pesos das ligações a um dado *cluster* de t_j são iguais ou superiores ao limiar de Sobrevivência. Por fim, assume-se que um *cluster* **Sobrevive** quando a ligação entre dois *clusters* pertencentes a agrupamentos diferentes e separados no tempo apresenta peso igual ou superior ao limiar de Sobrevivência, e esse par de *clusters* apresenta uma correspondência única (ou seja, não existem mais arestas com pesos iguais ou superiores ao limiar referido em cada um dos *clusters*).

4.2.3 Método para Clusters representados em Compreensão

Para representações compactas de *clusters*, desenvolvemos uma metodologia para avaliar a semelhança entre *clusters* e, sobretudo, o grau de semelhança a partir do qual é possível assumir, com uma confiança razoável, que o *cluster* $C_m(t_i)$ do *Clustering* ξ_i é a correspondência aproximada do *cluster* $C_u(t_j)$ no *Clustering* posterior ξ_j $(i < j)$. Para desempenhar esta tarefa calculamos a *Distância Euclidiana* entre os centróides $\vec{c}$ dos *clusters* (Definição 5) para todos os pares de *clusters* pertencentes a agrupamentos gerados em diferentes momentos.

Posteriormente, com o objectivo de averiguar se a distância, ou semelhança, entre os *clusters* é significativa, comparamo-la com a soma dos raios dos *clusters* em questão (Spinosa et al., 2007). Se a distância entre os centróides for igual ou superior à soma dos raios dos *clusters* - Equação 4.6 -, então podemos deduzir que os *clusters* não se intersectam e, consequentemente, o *cluster* de t_i não poderá ser considerado uma correspondência aproximada do *cluster* de t_j. Caso contrário - Equação 4.7 -, podemos assumir que o grau de sobreposição entre este par de *clusters* é significativo e concluir que eles formam uma correspondência aproximada.

Definição 5 - Distância Euclidiana entre Centróides:
Sejam C_m e C_u dois clusters *obtidos em t_i e t_j $(i < j)$, respectivamente, e definidos no espaço d-dimensional e sejam $\vec{c_m} = (c_{m,1}, c_{m,2}, ..., c_{m,d})$ e $\vec{c_u} = (c_{u,1}, c_{u,2}, ..., c_{u,d})$ os centróides correspondentes. A Distância Euclidiana entre Centróides é utilizada*



(a) (b)

Figura 4.3: Par de *clusters* com diferentes etiquetas temporais, definidos num espaço bidimensional: (a) que não se sobrepõem; (b) que se sobrepõem (a sobreposição é indicada pela região de intersecção A)

para medir a semelhança entre $C_m(t_i)$ *e* $C_u(t_j)$*, e pode ser definida como:*

$$d(\vec{c_m}, \vec{c_u}) = \sqrt{\sum_{i=1}^{d}(c_{m,i} - c_{u,i})^2}$$

$$d(C_m(t_i), C_u(t_j)) \geq r_{C_m(t_i)} + r_{C_u(t_j)} \tag{4.6}$$

$$d(C_m(t_i), C_u(t_j)) < r_{C_m(t_i)} + r_{C_u(t_j)} \tag{4.7}$$

Basicamente, a ideia consiste em averiguar se existe sobreposição de *clusters* gerados em instantes de tempo distintos, ou seja, se é possível encontrar uma região de intersecção, num dado intervalo de tempo $[t, t+1]$, formada por pares de *clusters*. No caso de ser possível detectar essa região, assume-se que se trata do mesmo *cluster*. Caso contrário, assumem-se *clusters* diferentes. O processo criado para efectuar este mapeamento herda algumas das limitações inerentes à representação em compreensão de *clusters*, nomeadamente, o facto de assumir que os *clusters* em questão são esféricos (é possível ter representações não esféricas mas isso requer, e.g. matrizes de covariância).

Por outro lado, a assumpção de correspondência quando a região de intersecção é muito reduzida poderá ser pouco rigorosa. Num cenário extremo, um par de *clusters* apenas deveria ser considerado um *match* se os respectivos centróides e raios fossem os mesmos. Porém, estas situações são muito raras e não detêm interesse prático.

Tabela 4.2: Definição formal das transições exógenas de um *cluster* representado em compreensão

Taxonomia das Transições	**Notação**	**Definição Formal**
Nascimento	$\emptyset \rightarrow C_u(t_j)$	$d(C_m(t_i), C_u(t_j)) \geq r_{C_m(t_i)} + r_{C_u(t_j)} \forall_m$
Morte	$C_m(t_i) \rightarrow \emptyset$	$d(C_m(t_i), C_u(t_j)) \geq r_{C_m(t_i)} + r_{C_u(t_j)} \forall_u$
Cisão	$C_m(t_i) \overset{\subseteq}{\rightarrow} \{C_1(t_j), ..., C_r(t_j)\}$	$(d(C_m(t_i), C_u(t_j)) < r_{C_m(t_i)} + r_{C_u(t_j)}) \wedge$ $\exists C_r \in \xi_j \setminus \{C_u\} : d(C_m(t_i), C_r(t_j)) < r_{C_m(t_i)} + r_{C_r(t_j)}$
Fusão	$\{C_1(t_i), ..., C_p(t_i)\} \overset{\subseteq}{\rightarrow} C_u(t_j)$	$(d(C_m(t_i), C_u(t_j)) < r_{C_m(t_i)} + r_{C_u(t_j)}) \wedge$ $\exists C_p \in \xi_i \setminus \{C_m\} : d(C_p(t_i), C_u(t_j)) < r_{C_p(t_i)} + r_{C_u(t_j)}$
Sobrevivência	$C_m(t_i) \rightarrow C_u(t_j)$	$(d(C_m(t_i), C_u(t_j)) < r_{C_m(t_i)} + r_{C_u(t_j)}) \wedge$ $\nexists C_p \in \xi_i \setminus \{C_m\} : d(C_p(t_i), C_u(t_j)) < r_{C_p(t_i)} + r_{C_u(t_j)}$

Tabela 4.3: Definição formal das transições endógenas de um *cluster* representado em compreensão

Taxonomia das Transições	**Notação**	**Definição Formal**
Expansão	$C_m(t_i) \nearrow C_u(t_j)$	$\#C_m(t_i) < \#C_u(t_j)$
Contracção	$C_m(t_i) \searrow C_u(t_j)$	$\#C_m(t_i) \geq \#C_u(t_j)$
Compactação	$C_m(t_i) \overset{\bullet}{\rightarrow} C_u(t_j)$	$\rho_{C_m(t_i)} < \rho_{C_u(t_j)}$
Dispersão	$C_m(t_i) \overset{*}{\rightarrow} C_u(t_j)$	$\rho_{C_m(t_i)} > \rho_{C_u(t_j)}$

Para melhor compreender esta ideia e o seu fundamento, na Figura 4.3, ilustra-se cada uma das situações previstas. Na Figura 4.3(a) é retratada a situação em que os *clusters* não são uma correspondência aproximada, ie, não existe qualquer intersecção entre eles. Isto é bem captado pelo método que concebemos pois, nestes casos, a soma dos raios dos *clusters* é inferior à respectiva distância entre os centróides. A situação complementar é vísível na Figura 4.3(b), onde é possível constatar a existência de uma região de intersecção - Região A - indicativa da sobreposição dos *clusters*. Nestes casos, é fácil deduzir que a soma dos respectivos raios é superior à distância entre os centróides.

À semelhança do que foi efectuado para o método anterior e com o intuito de detectar as transições exógenas e endógenas e, consequentemente, descobrir as transições experienciadas pelos *clusters* num dado intervalo de tempo, nós definimos formalmente as transições com base nos conceitos previamente expostos - *Distância Euclidiana entre Centróides*, *Raio*, *Cardinalidade* e *Densidade*. Este desenho formal encontra-se retratado nas Tabelas 4.2 e 4.3. Mais uma vez, alerta-se para o facto das transições internas, ou endógenas, apenas serem monitorizadas para *clusters* que sobrevivem de um instante temporal para outro. Acrescente-se, também, que a utilização da Distância Euclidiana implica que as variáveis, ou atributos, estejam estandardizadas, ie, expressas na mesma escala.

Nas Tabelas 4.2 e 4.3 apresentam-se as definições formais das transições externas e internas, para *clusters* representados em compreensão, bem como a notação a utilizar em cada um dos casos. No que concerne às transições externas, assume-se o **Nascimento** de um *cluster* quando não é possível encontrar nenhuma correspondência aproximada para um dado *cluster* de $t+1$. Ou seja, para todas as combinações de *clusters* do instante t com o *cluster* de $t+1$ a distância euclidiana entre os centróides de cada par de *clusters* é sempre superior, ou igual, à soma dos respectivos raios. A **Morte** de um *cluster* é definida de forma análoga, com a diferença de se referir a um dado *cluster* de t, em detrimento de um dado *cluster* de $t+1$. Por sua vez, a **Cisão** manifesta-se pela existência de, pelo menos, duas correspondências aproximadas para o mesmo *cluster* de t, em $t+1$. Dito de outra forma, existem pelo menos dois *clusters* de $t+1$, cujo emparelhamento com o *cluster* de t permite concluir que a distância entre os respectivos centróides é inferior à soma dos raios. A **Fusão** de *clusters* define-se de forma semelhante à Cisão, no entanto, na Fusão as correspondências exactas de dois ou mais *clusters* de t convergem para um único *cluster* de $t+1$. Por fim, a **Sobrevivência** de um *cluster* em $t+1$ caracteriza-se pela existência de uma correspondência unívoca entre dois *clusters* de instantes temporais distintos, ie, o *cluster* de t só se sobrepõe ao *cluster* de $t+1$ e o *cluster* de $t+1$ só se intersecta com o *cluster* de t.
Como anteriormente mencionado, para os *clusters* sobreviventes, também interessa monitorizar as modificações internas a que, eventualmente, foram submetidos no intervalo de tempo sob análise. Neste contexto interno, o *cluster* pode sofrer um aumento (ou diminuição) do número de observações que o compõem, ie, **Expandir-se** (ou **Contrair-se**, respectivamente) ou, ainda, ver alterada a sua densidade, que pode incrementar e originar um *cluster* mais **Compacto** ou decrescer e dar lugar a um *cluster* mais **Disperso**.

No Capítulo seguinte será realizada a avaliação experimental da metodologia MEC exposta neste Capítulo, e demonstrada a sua capacidade de diagnosticar transições entre dois ou mais *Clusterings* obtidos em tempos diferentes.

Capítulo 5

Avaliação Experimental

O objectivo primordial deste capítulo assenta na realização de uma avaliação experimental da nossa abordagem ao problema da monitorização da evolução de *clusters*. Inicialmente, é apresentada a metodologia adoptada na execução das experiências, procedendo-se, posteriormente, à calibração do MEC e à apresentação de pequenos casos de estudo. Com o intuito de testar o nosso sistema de monitorização da evolução, discutido no capítulo anterior, recorreu-se a uma combinação de conjuntos de dados artificiais e reais. Nos conjuntos de dados artificiais, os *clusters* naturais e as respectivas transições são conhecidas *a priori*, mas esta informação não é explicitamente utilizada pelo algoritmo de detecção de transições. No que concerne aos conjuntos de dados reais, procurou-se obter dados provenientes de diversas fontes e de áreas de conhecimento distintas, nomeadamente, Economia, Educação, Território e Política, para mostrar a versatilidade de aplicação da metodologia MEC, bem como a sua exequibilidade e utilidade prática. As experiências foram conduzidas no software R 2.10.0 e incidiram quer sobre o método desenvolvido para *clusters* representados em extensão, quer sobre o método concebido para *clusters* representados em compreensão, de modo a ser possível efectuar uma comparação dos resultados sugeridos pelo algoritmo para ambos os casos. Para efeitos de geração dos *clusters*, utilizaram-se as funções `hclust` e `kmeans` da *package* `stats`, que executam o algoritmo hierárquico aglomerativo e o algoritmo das k-médias, respectivamente. Para reduzir a instabilidade característica do k-médias, foi introduzida uma melhoria no algoritmo. Esta melhoria consistiu em correr várias vezes o k-médias e efectuar a assignação das observações aos *clusters* com base na frequência do grau de pertença a um dado *cluster* (e.g., se a observação 1 em, digamos, 5 aplicações do algoritmo, foi afectada 4 vezes ao *cluster* C_2, então, no resultado final esta observação deverá estar atribuída a esse *cluster*). Saliente-se que esta avaliação experimental não tem como propósito avaliar a eficiência e a escalabilidade do algoritmo, uma vez que a metodologia foi projectada para ambientes estáticos que, normalmente, não acarretam grandes problemas ao nível do processamento, armazenagem e tempo de execução dos algoritmos. O foco deste estudo consiste, assim, na avaliação da

capacidade do algoritmo em diagnosticar as transições e, por conseguinte, em fomentar a compreensão dos fenómenos e dos factores que sustentam a evolução detectada.

5.1 Metodologia Experimental

Antes de apresentar os resultados das experiências será exposta, de forma sucinta, a metodologia adoptada para fins de obtenção dos referidos resultados. A metodologia engloba uma fase de pré-processamento, comum a ambos os métodos de monitorização, e uma fase de aplicação do MEC, cujas etapas variam consoante a estratégia de representação de *clusters* eleita. A primeira fase refere-se à tarefa de *Clustering* propriamente dita, ie, ao processo de geração do *input* deste estudo, e a segunda fase relaciona-se com decisões prévias exigidas ao utilizador da metodologia MEC para aplicação dos mecanismos de monitorização. A arquitectura do processo de avaliação experimental encontra-se ilustrado na Figura 5.1.

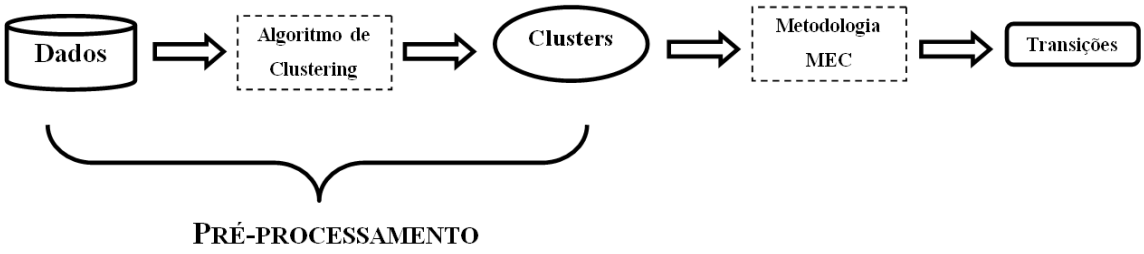


Figura 5.1: Arquitectura do processo de avaliação experimental

5.1.1 Fase de Pré-processamento

No âmbito da nossa investigação, os *clusters* são encarados como um ingrediente necessário à construção de uma estrutura de conhecimento, capaz de captar as mudanças ocorridas nos *clusters*, em diferentes instantes temporais. Ou seja, os *clusters* constituem o nosso objecto de estudo, em detrimento dos dados em bruto, pelo que a aplicação de técnicas de *Clustering* resume-se, apenas, a uma fase preliminar. A fase de pré-processamento da metodologia experimental reflecte os passos necessários à obtenção deste *input*, e é comum a ambos os métodos de monitorização.

Etapas:

1. Selecção do algoritmo de Clustering - em primeiro lugar deve-se escolher um, ou vários, algoritmos de agrupamento para obter as estruturas de *clusters* exigidas na aplicação do MEC; se não existir preferência por nenhum algoritmo de *Clustering* específico, é aconselhável a utilização de algoritmos pertencentes a classes diferentes (algoritmos hierárquicos, algoritmos particionais, algoritmos baseados na densidade, etc.). Nós optámos por apresentar experiências recorrendo a dois algoritmos comummente utilizados na tarefa de *Clustering*, nomeadamente, o algoritmo hierárquico aglomerativo, com o critério de Ward, e o algoritmo particional das k-médias, apesar de termos conduzido experiências usando algoritmos de todas as classes referidas no Capítulo 2.

2. Escolha do número óptimo de clusters - a optimalidade de uma estrutura de *clusters* é encarada, neste contexto, como a proximidade da partição obtida por um dado algoritmo à partição *natural* dos dados (Halkidi et al., 2002). Para efeitos de escolha do número óptimo de *clusters*, que é um dos parâmetros de inicialização de alguns algoritmos, podem ser explorados diferentes critérios de validação; a escolha do número de *clusters* também pode ser guiada pela análise do dendrograma, quando disponível, ou pelo domínio do conhecimento.

3. Obtenção de uma partição dos dados - tendo disponível a informação sobre o número óptimo de *clusters* k, basta aplicar o algoritmo de agrupamento seleccionado para obter as estruturas de *clusters*.

5.1.2 Fase de Aplicação do MEC

Cada representação de *clusters* - em extensão ou em compreensão - é caracterizada por etapas específicas. Esta especificidade advém do facto de se terem desenvolvido métodos de monitorização diferentes e adaptados a cada uma das estratégias de representação.

Etapas Específicas de cada Método:

1. ***C*lusters representados em extensão**

   (a) Estabelecimento dos valores do limiar de Sobrevivência e do limiar de Cisão - no presente estudo, apresentam-se experiências para um limiar de Sobrevivência $\tau = 0.5$ e para um limiar de Cisão $\rho = 0.2$, de modo a não tornar muito exigente o processo de mapeamento; também foram realizadas experiências com valores mais extremos, mas considerou-se que os valores referidos eram os mais razoáveis.

   (b) Aplicação do algoritmo de detecção de transições MEC

(c) Desenho dos grafos bipartidos recorrendo a um software apropriado - para executar este passo nós optámos pela utilização do Microsoft Visio 2007 ®.

2. ***C*lusters representados em compreensão**

   (a) Normalização das variáveis - em todas as experiências realizadas aplicando este método normalizaram-se as variáveis, de modo a não deturpar os resultados decorrentes do cálculo da distância Euclidiana entre os centróides dos *clusters*; a técnica de normalização adoptada foi o *Z*-scores, por ser uma das mais utilizadas.

   (b) Aplicação do algoritmo de detecção de transições MEC

5.1.3 Conceitos importantes

Critérios de Validação e o Coeficiente de Silhueta Médio

Como foi mencionado no Capítulo 2, segundo a *leges artis* existem três grandes tipos de medidas de validação dos resultados de agrupamento: medidas ou critérios *internos*, *externos* e *relativos*. Estes critérios revelam-se extremamente úteis na fase de validação e avaliação da qualidade dos resultados obtidos por um dado algoritmo de *Clustering*. Dentro de cada uma das classes referidas, é possível encontrar uma enorme diversidade de critérios e medidas para efectuar a validação. Neste trabalho, optou-se por um método de validação interna *standard*, que apesar de não ser necessariamente o melhor do conjunto existente, é dos mais utilizados: o coeficiente de silhueta médio (ou *Silhouette Width*) (Rousseeuw, 1987). Handl et al. fornecem uma boa visão geral da literatura em termos de medidas de validação interna (Handl et al., 2005). A adopção deste critério teve como intuito principal a orientação na escolha do número óptimo de *clusters*.

O coeficiente de silhueta médio é uma medida de validação interna que reflecte a coesão e a separação das partições de *clusters*. A **coesão** avalia a homogeneidade dos *clusters*, por meio da análise da inércia intra-*cluster*, e a **separação**, tal como o nome indica, quantifica o grau de separação entre os *clusters*, usualmente, através do cálculo da distância entre os respectivos centróides. O coeficiente de silhueta médio é um método popular que procede à combinação não-linear da coesão e da separação, permitindo medir o grau de confiança na afectação de uma observação a um dado *cluster* através da comparação destas duas características.

Para obter o valor global do coeficiente de silhueta médio calcula-se o valor de *silhueta* para cada observação do conjunto de dados e, seguidamente, calcula-se a média de todos os valores de *silhueta*.

O valor de *silhueta* para cada observação i é dado pela seguinte fórmula:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)} \tag{5.1}$$

onde a_i representa a distância média entre i e todas as observações do mesmo *cluster* e b_i é a distância média entre i e todas as observações afectas ao *cluster* vizinho mais próximo, ie

$$b_i = \min_{C_k \in \xi \setminus C_i} \sum_{j \in C_k} \frac{dist(i,j)}{n_k} \tag{5.2}$$

onde C_i é o *cluster* que contém a observação i, $dist(i,j)$ representa a distância (e.g. Euclidiana) entre as observações i e j, e n_k representa a cardinalidade do *cluster* C_k.

Os valores assumidos pelo valor de *silhueta* situam-se no intervalo $[-1, 1]$ e devem ser maximizados. Isto porque, se o valor de *silhueta* for igual, ou estiver próximo, de 1 significa que a observação foi bem agrupada e atribuída ao *cluster* apropriado; se o valor for aproximadamente zero, indica que a observação também poderia ter sido afectada ao *cluster* mais próximo; por fim, se o valor do coeficiente é igual, ou próximo, de -1 a probabilidade da observação ter sido mal atribuída ao *cluster* é muito elevada, pelo que o *Clustering* resultante é fraco.

A média dos valores de *silhueta* pode funcionar como um bom indicador do número óptimo de *clusters*. Para isso, basta calcular este coeficiente para várias alternativas de agrupamento, ie, vários números de *clusters*, desenhar um gráfico que permita visualizar melhor a qualidade das soluções testadas e optar pelo número de *clusters* ao qual corresponde uma espécie de "joelho"no gráfico, ie, deve-se encontrar o valor do coeficiente cujos valores posteriores sejam inferiores. Quando o "joelho"não é facilmente detectável ou muito evidente, deve-se usar o bom senso na escolha do k. Com base na Figura 5.2 pretende-se exemplificar este processo. Nesta Figura é fácil comprovar que o k óptimo é 4, dado que corresponde à solução com um valor mais elevado do coeficiente de silhueta médio, sendo os valores imediatamente anterior e posterior inferiores.

Normalização das variáveis

Dado que, com frequência, a maioria das variáveis que caracteriza um dado conjunto de observações se encontra expressa em diferentes escalas e possui dispersões bastantes diferentes, urge a necessidade de submetê-las a um processo de normalização. Para efectuar a normalização, optou-se pela técnica Z-scores, também denominada "standardização", que submete as variáveis a um processo de centragem e redução (Equação 5.3). A standardização reduz todas as variáveis à mesma escala,

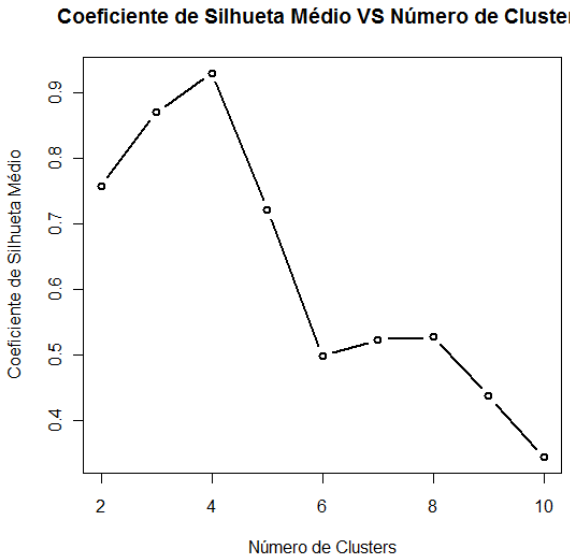


Figura 5.2: Valores do coeficiente de silhueta médio para diferentes soluções de agrupamento (a gama de valores considerada razoável para o número de *clusters* foi de $[2, 10]$)

tornando-as independentes das unidades de medida em que foram originalmente expressas melhorando, por conseguinte, a sua interpretabilidade. Consequentemente, as variáveis tornam-se equivalentes, em termos de importância, o que, por sua vez, anula qualquer enviesamento passível de ocorrer nos algoritmos de *Clustering* ou no cálculo das distâncias com recurso à distância Euclidiana, em virtude da magnitude das variáveis. No método de monitorização para representações compactas de *clusters*, são calculadas distâncias Euclidianas, pelo que se torna necessário standardizar as variáveis.

$$x_{i,j} = \frac{x_{i,j} - \overline{x}_j}{s_j} \tag{5.3}$$

onde, $x_{i,j}$ representa a i-ésima observação da j-ésima variável, $\overline{x}_j$ corresponde à média da j-ésima variável e s_j indica o desvio-padrão da j-ésima variável.

5.2 Calibração do MEC

Com o intuito de avaliar a capacidade do MEC em detectar eficazmente as transições dos *clusters*, e proceder à afinação dos valores dos limiares do modelo, conduziram-se

Tabela 5.1: Características gerais dos três conjuntos de dados artificialmente gerados

Conjunto de dados Artificiais	Número de objectos	Número de Clusters
Dados t	300	4
Dados t+1	300	5
Dados t+2	300	4

experiências controladas com base em conjuntos de dados artificiais. Estes conjuntos foram criados a partir de um gerador de *clusters* desenvolvido para o efeito. As experiências controladas apresentam a vantagem de permitirem simular cenários, e.g. através da imposição de determinadas transições aos *clusters*, possibilitando a avaliação e o estudo do comportamento dos métodos e algoritmos, sob condições pré-definidas.

Nesta secção procede-se, em primeiro lugar, à descrição dos conjuntos de dados artificialmente gerados. Em segundo lugar, testa-se a metodologia MEC, por via da sua aplicação aos referidos dados, e avalia-se a precisão dos resultados através da comparação das transições detectadas com as transições previamente impostas. Por fim, efectua-se a análise de sensibilidade do algoritmo aos valores assumidos pelos limiares de *Sobrevivência* e *Cisão*, para efeitos de definição e afinação dos limiares.

5.2.1 Descrição dos Dados Artificiais

Geraram-se conjuntos de dados artificiais para os quais se conhece *a priori* a distribuição, a disposição e o número de *clusters*, bem como a natureza das respectivas transições. Cada conjunto de dados corresponde a um determinado instante temporal e é composto por k *clusters* bidimensionais. A opção pela bidimensionalidade prende-se, sobretudo, com questões de visualização no espaço. Os pontos, ou objectos, que constituem cada *cluster* são, assim, definidos por duas dimensões e seguem uma Distribuição Gaussiana com média, variância e número de pontos pré-especificados pelo utilizador. Ou seja, cada *cluster* é caracterizado por quatro parâmetros: o número de objectos - n_k -, a média da variável x - μ_x -, a média da variável y - μ_y - e a variância - σ. Com o propósito de testar o algoritmo, foram gerados três conjuntos de dados artificiais, correspondentes aos instantes temporais t, $t+1$ e $t+2$.

A caracterização pormenorizada destes conjuntos pode ser consultada nas Tabelas 5.1 e 5.2 e o aspecto gráfico dos *clusters* pode ser visualizado na Figura 5.3.

Tabela 5.2: Características detalhadas dos três conjuntos de dados artificialmente gerados

Clusters Artificiais					
Instante temporal	**Cluster**	n_k	μ_x	μ_y	σ
t	A	50	5	5	1
	B	50	20	20	1
	C	100	45	45	1
	D	100	65	65	1
$t+1$	A	100	10	10	2
	B	100	45	45	1
	C	35	65	80	1
	D	35	65	25	1
	E	30	20	75	1
$t+2$	A	90	10	10	1
	B	75	65	80	1
	C	75	65	25	1
	D	60	20	75	2

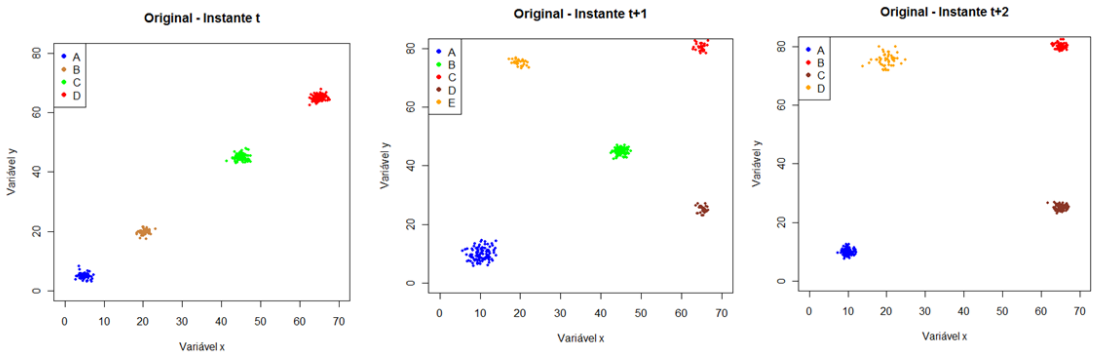


Figura 5.3: *Clusters* gerados artificialmente para três instantes temporais distintos

Tabela 5.3: Transições exógenas impostas aos conjuntos de dados artificiais

Transicões Exógenas		
Intervalo de Tempo	**Tipo de transição**	**Notação**
$[t, t+1]$	Nascimento	$\emptyset \rightarrow E(t+1)$
	Sobrevivência	$C(t) \rightarrow B(t+1)$
	Fusão	$\{A(t), B(t)\} \overset{\subseteq}{\rightarrow} A(t+1)$
	Cisão	$D(t) \overset{\subseteq}{\rightarrow} \{C(t+1), D(t+1)\}$
$[t+1, t+2]$	Sobrevivência	$A(t+1) \rightarrow A(t+2)$
		$C(t+1) \rightarrow B(t+2)$
		$D(t+1) \rightarrow C(t+2)$
		$E(t+1) \rightarrow D(t+2)$
	Morte	$B(t+1) \rightarrow \emptyset$

Tabela 5.4: Transições endógenas impostas aos conjuntos de dados artificiais

Transicões Endógenas		
Intervalo de Tempo	**Tipo de transição**	**Notação**
$[t, t+1]$	Nenhuma mudança	
$[t+1, t+2]$	Expansão	$E(t+1) \nearrow D(t+2)$
		$C(t+1) \nearrow B(t+2)$
		$D(t+1) \nearrow C(t+2)$
	Contracção	$A(t+1) \searrow A(t+2)$
	Compactação	$E(t+1) \overset{\bullet}{\rightarrow} D(t+2)$
	Dispersão	$A(t+1) \overset{*}{\rightarrow} A(t+2)$

5.2.2 Avaliação do MEC usando Conjuntos de Dados Artificiais

Para efeitos de avaliação do MEC conduziram-se experiências controladas, realizadas com recurso a três conjuntos de dados artificialmente gerados. O facto de serem controladas significa que se conhece *a priori* a partição natural dos dados e a natureza das transições, endógenas e exógenas, experienciadas pelos *clusters*. Na Tabela 5.3 e na Tabela 5.4 resume-se esta informação que irá ser, posteriormente, confrontada com as soluções do algoritmo de detecção de transições.

Aplicação do MEC para clusters representados em extensão

Inicialmente, assumiu-se que os *clusters* estavam representados em extensão, ie, que cada *cluster* era caracterizado pelas observações que lhe tinham sido atribuídas. Com base nesta assumpção, aplicou-se o mecanismo de monitorização da metodologia MEC mais apropriado - Método baseado nas Probabilidades Condicionadas -, para obter informação sobre a evolução das estruturas de *clusters* ao longo dos três instantes temporais. O processo foi replicado para dois algoritmos de *Clustering* distintos: o algoritmo hierárquico aglomerativo, usando o índice de Ward, e o al-

goritmo particional das k-médias. Os valores dos limiares de Sobrevivência e Cisão definidos são os mencionados na Metodologia - $\tau = 0.5$ e $\rho = 0.2$.

Algoritmo Hierárquico Aglomerativo (índice de Ward). Para conduzir a primeira experiência, foi seleccionado o algoritmo hierárquico aglomerativo para efeitos de geração das estruturas de *clusters*. A escolha do número óptimo de *clusters* foi guiada pela análise dos dendrogramas e do coeficiente de silhueta médio. Os dendrogramas representados na Figura 5.4 sugerem a existência de quatro, cinco e, novamente, quatro *clusters*, para t, $t+1$ e $t+2$, respectivamente. A escolha destas partições pode ser comprovada pela análise dos valores da medida de Silhueta Média para várias alternativas de agrupamento (Figura 5.5). Estes gráficos vêm corroborar as conclusões obtidas com os dendrogramas, ie, que em t o número óptimo de *clusters* é quatro, em $t+1$ é cinco e em $t+2$ é novamente quatro. Assumindo estas partições, obtiveram-se as representações gráficas, apresentadas na Figura 5.6, das estruturas de *clusters* no espaço bidimensional formado pela projecção dos dados nas duas componentes principais, que explicam a maior parte da variância contida nos dados (neste caso em concreto, como os dados são descritos por dois atributos, a variância é explicada quase na totalidade pelas duas componentes principais). Na Figura 5.7 apresenta-se uma representação análoga dos *Clusterings*, mas que permite fazer o contraste entre os *clusters naturais* da Figura 5.3 e os *clusters* descobertos pelo algoritmo hierárquico. Note-se que as etiquetas dos *clusters* da Figura 5.7 são meramente ilustrativas, servindo apenas para discriminar os diferentes *clusters*. Da observação atenta das figuras facilmente se comprova que o método hierárquico foi bem sucedido na descoberta da verdadeira estrutura presente nos dados artificiais, uma vez que conseguiu encontrar o respectivo agrupamento *natural*.

Após a geração dos *clusters*, aplicou-se o método MEC apropriado. Com base no *output* do algoritmo que implementa este método desenharam-se os grafos bipartidos, representados na Figura 5.8. Nos grafos eliminaram-se as arestas com pouca ou nenhuma relevância na detecção das transições, nomeadamente, as ligações com peso inferior ao limiar de Cisão $\rho = 0.2$; e reforçou-se a espessura das arestas cujo peso era igual ou superior ao limiar de Sobrevivência $\tau = 0.5$, de modo a melhorar o processo visual de categorização das transições (embora as transições ocorridas em cada intervalo de tempo sejam previamente devolvidas pelo algoritmo). Seguidamente, procedeu-se à análise das ligações entre os nós dos grafos para concluir sobre as transições externas dos *clusters*.

No intervalo de tempo $[t, t+1]$, detectou-se:

1. Fusão de dois *clusters* - $\{C_1(t), C_2(t)\} \overset{\subseteq}{\rightarrow} C_1(t+1)$;
2. Cisão de um *cluster* em três *clusters* - $C_4(t) \overset{\subseteq}{\rightarrow} \{C_3(t+1), C_4(t+1), C_5(t+1)\}$;
3. Sobrevivência de um único *cluster* - $C_3(t) \rightarrow C_2(t+1)$.

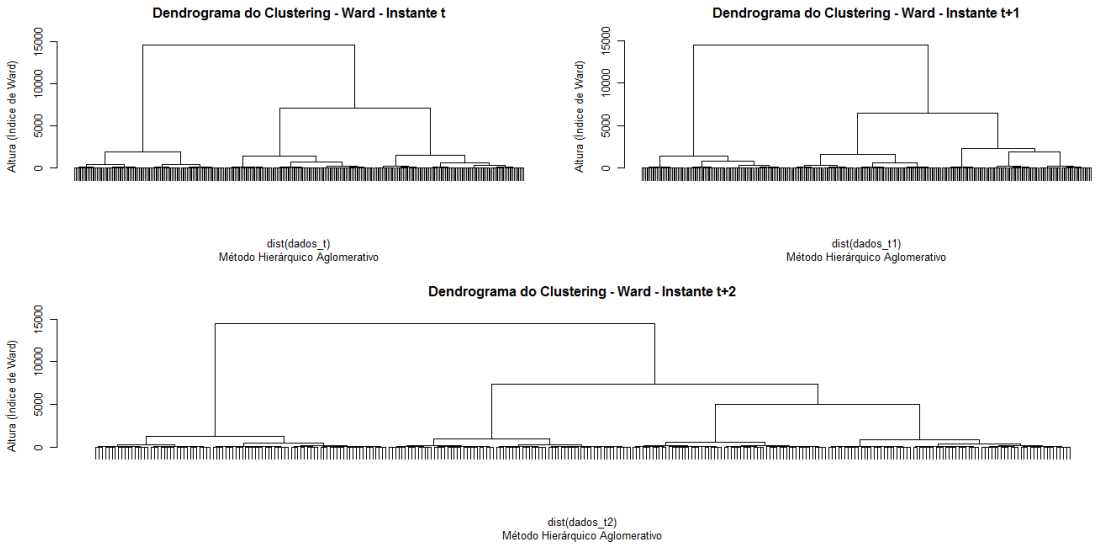


Figura 5.4: Dendrogramas resultantes da aplicação do algoritmo hierárquico aglomerativo (índice de Ward) aos conjuntos de dados artificiais, para diferentes instantes de tempo - t, $t+1$ e $t+2$

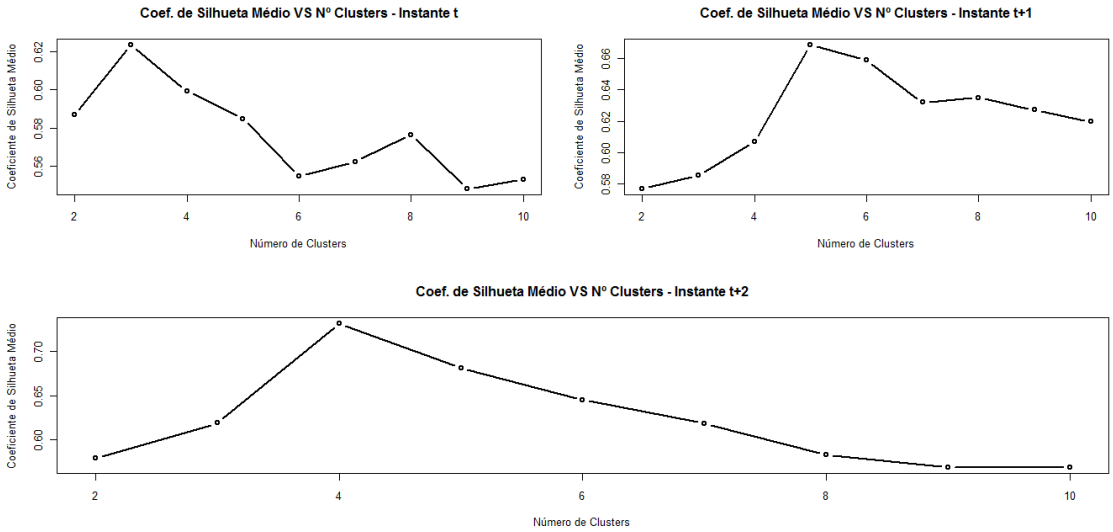


Figura 5.5: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo hierárquico aglomerativo (índice de Ward) e para diferentes instantes temporais

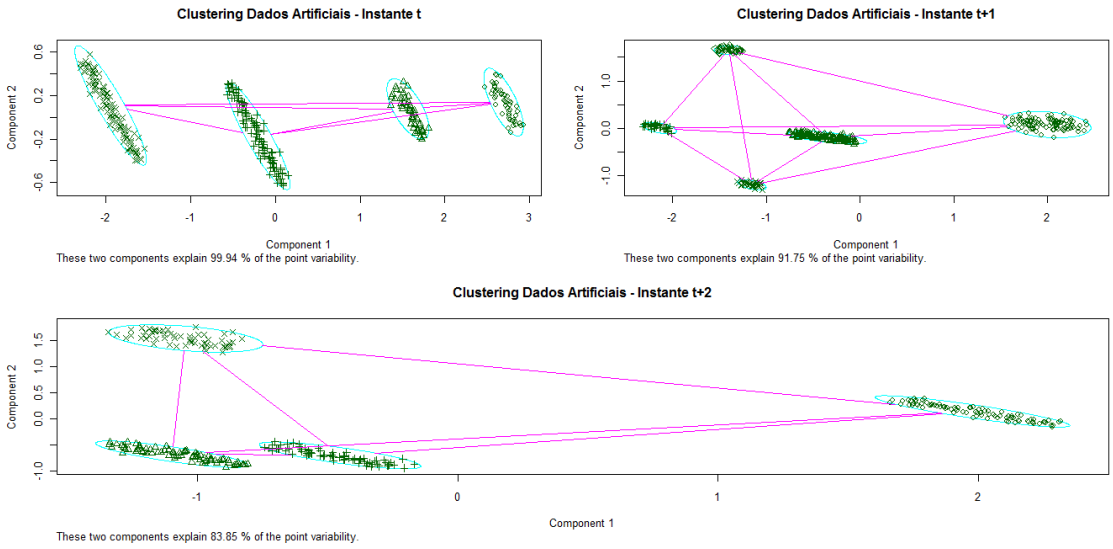


Figura 5.6: Representação gráfica dos *clusters* obtidos com recurso ao algoritmo hierárquico aglomerativo, com índice de Ward, no espaço formado pela projecção dos dados nas duas componentes principais, nos instantes de tempo t, $t+1$ e $t+2$

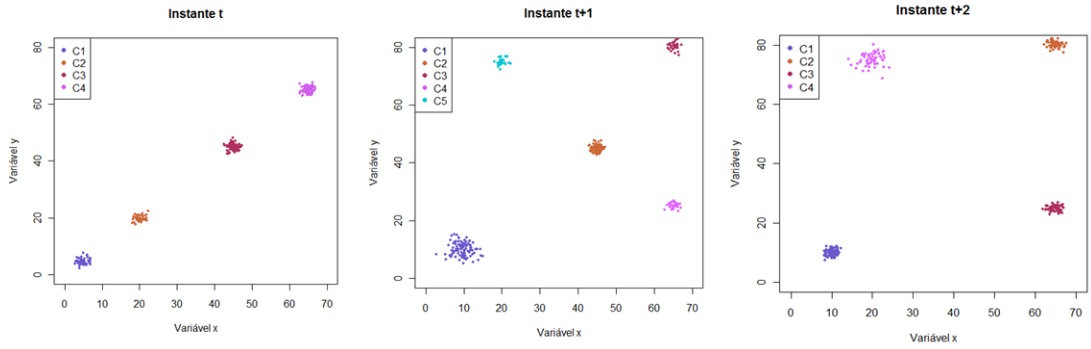


Figura 5.7: *Clusters* obtidos pelo algoritmo hierárquico aglomerativo, para três instantes temporais distintos, e com a partição dos dados sugerida pela análise dos dendrogramas e do coeficiente de silhueta médio

No intervalo de tempo seguinte, ocorreram as seguintes transições:

1. Sobrevivência de três *clusters* - $C_1(t+1) \rightarrow C_1(t+2)$, $C_2(t+1) \rightarrow C_2(t+2)$ e $C_3(t+1) \rightarrow C_3(t+2)$;

2. Fusão de dois *clusters* - $\{C_4(t+1), C_5(t+1)\} \overset{\subseteq}{\rightarrow} C_4(t+2)$.

Comparando estas transições com as transições exógenas impostas artificialmente aos *clusters* e descritas na Tabela 5.3, verifica-se que existem algumas discordâncias, nomeadamente, o algoritmo não detecta o nascimento de um *cluster* no instante $t+1$ nem a morte de um *cluster* em $t+1$. Por outro lado, considera a existência de uma fusão de *clusters* que não foi artificialmente imposta neste período temporal. Estas diferenças foram, sobretudo, despoletadas pelos valores escolhidos para os limiares e pelo facto deste algoritmo monitorizar os mesmos objectos ao longo do tempo. Assim, se se definirem valores diferentes para os limiares, em particular: $\tau = 0.6$ e $\rho = 0.35$; consegue-se facilmente detectar as transições correctas, como é evidente na Figura 5.9. Com estas alterações, o algoritmo passa a detectar, no intervalo $[t, t+1]$:

1. Nascimento de um *cluster* - $\emptyset \rightarrow C_5(t+1)$;

2. Cisão de um *cluster* em dois *clusters* - $C_4(t) \overset{\subseteq}{\rightarrow} \{C_3(t+1), C_4(t+1)\}$.

No intervalo de tempo subsequente, constata-se que a introdução de limiares mais exigentes permite assegurar a detecção da sobrevivência de todos os *clusters*, nomeadamente, C_1, C_2, C_3 e C_4. Contudo, o algoritmo continua a não capturar a morte do *cluster* $C_4(t)$, assumindo a sua divisão ($C_4(t+1) \overset{\subseteq}{\rightarrow} \{C_3(t+2), C_4(t+2)\}$). Esta diferença deve ser alvo de uma análise mais detalhada. Como já foi referido, o mecanismo de monitorização desenvolvido para *clusters* representados em extensão é restringido a conjuntos de dados recolhidos em instantes temporais distintos e compostos pelos mesmos objectos. Tendo este aspecto em consideração, facilmente se deduz que o mecanismo tem alguma dificuldade em detectar mortes e nascimentos pois, de acordo com a definição formal das transições, a morte/nascimento corresponde a situações em que a probabilidade condicionada, ou peso de todas as ligações de um dado *cluster*, é inferior ao limiar de Cisão. Assim, assumindo um limiar $\rho = 0.2$, dificilmente se consegue detectar a morte/nascimento do *cluster* excepto se o número n_k de *clusters* do *Clustering* de $t/t+1$ for $n_k > 5$. Ao obrigar a estrutura a ter um elevado número n_k de *clusters* e assumindo que as ligações de um dado *cluster* são equiprováveis, ie, $P(C_1(t+1)|C_1(t)) = P(C_2(t+1)|C_1(t)) = ... = P(C_p(t+1)|C_r(t)) = \frac{1}{n_k}$,, consegue-se reduzir o peso associado a cada ligação, (por via do aumento do denominador da fracção $\frac{1}{n_k}$) e, consequentemente, detectar mais mortes e nascimentos. Neste caso em concreto, como n_k é reduzido para os três instantes temporais, em particular, para $t+2$, o algoritmo não é capaz de detectar

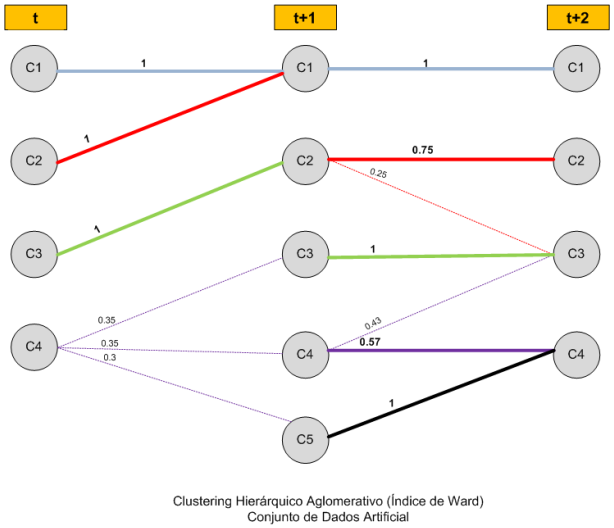


Figura 5.8: Grafos bipartidos, correspondentes aos intervalos de tempo $[t, t+1]$ e $[t+1, t+2]$, dos conjuntos de dados artificiais (algoritmo hierárquico aglomerativo), com a espessura das arestas a indicar os pesos superiores ou iguais ao limiar de Sobrevivência $\tau = 0.5$ e superiores ou iguais ao limiar de Cisão $\rho = 0.2$

a morte de $C_4(t+1)$ pois, mesmo que os pesos das ligações fossem equiprováveis, cada ligação teria um peso associado de $\frac{1}{4} = 0.25$ o que, para um limiar $\rho = 0.2$, se traduz numa cisão. Por outro lado, o facto de se monitorizarem os mesmos objectos impede a detecção de mortes e nascimentos *puros*, ie, que resultem de ligações com $peso = 0$, uma vez que o universo de objectos é sempre o mesmo, variando apenas a forma como estes são agrupados. Este também é um factor que explica a Cisão de $[t+1, t+2]$: a suposta morte do *cluster* $C_4(t+1)$ não é mais do que a transferência dos seus pontos/objectos para os *clusters* $C_3(t+2)$ e $C_4(t+2)$.

Tendo por base esta análise podemos, assim, concluir que o método MEC baseado em probabilidades condicionadas, conseguiu diagnosticar correctamente as transições exógenas impostas artificialmente aos dados.

Algoritmo Particional k-médias. Na segunda experiência, geraram-se os *Clusterings* referentes aos três instantes temporais, com recurso ao algoritmo particional das k-médias. De acordo com o coeficiente de silhueta médio - Figura A.1 -, a partição óptima dos dados, para os vários instantes de tempo é, respectivamente, quatro, cinco e quatro *clusters*, embora em $t+1$ e em $t+2$, esta optimilidade não seja tão evidente, pois existem vários "joelhos"nos gráficos. Esta análise confirma as soluções de agrupamento conhecidas *a priori*. As representações gráficas das estruturas de *clusters*, no espaço bidimensional, tendo por base estas partições, podem ser visualizadas nas Figuras A.2 e A.3. Atente-se ao facto de ter sido necessário

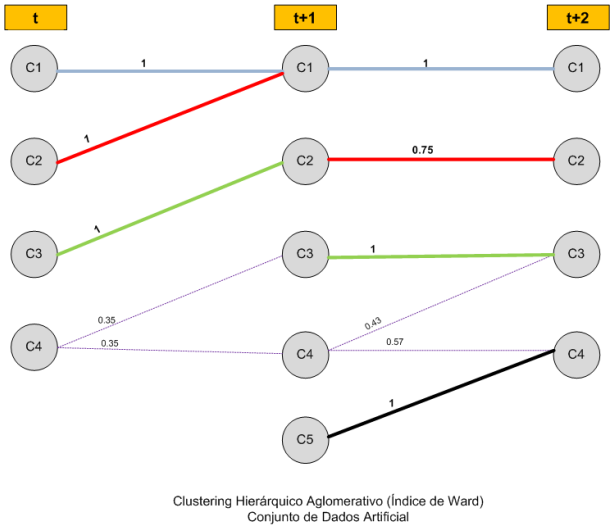


Figura 5.9: Grafos bipartidos, correspondentes aos intervalos de tempo $[t, t+1]$ e $[t+1, t+2]$, dos conjuntos de dados artificiais (algoritmo hierárquico aglomerativo), mas para um limiar de Sobrevivência $\tau = 0.6$ e um limiar de Cisão $\rho = 0.35$

proceder a alterações no posicionamento dos *clusters* no espaço formado pelos dois atributos, para facilitar a descoberta dos *clusters* pelo algoritmo particional, que é, por norma, mais instável que o algoritmo hierárquico. A comparação destas imagens com a estrutura original, presente na Figura 5.3 permite concluir que o algoritmo das k-médias, não obstante o facto de ter demonstrado mais dificuldades nesta tarefa do que o algoritmo hierárquico, conseguiu efectuar o correcto agrupamento dos dados.

Tendo por base esta partição dos dados, o algoritmo de detecção de transições obteve os seguintes resultados, passíveis de serem visualizados no par de grafos bipartidos da Figura 5.10: no primeiro intervalo de tempo $[t, t+1]$ verificou-se,

1. Cisão de um *cluster* em três *clusters* - $C_1(t) \overset{\subset}{\rightarrow} \{C_1(t+1), C_2(t+1), C_4(t+1)\}$;
2. Fusão de dois *clusters* - $\{C_2(t), C_3(t)\} \overset{\subset}{\rightarrow} C_5(t+1)$;
3. Sobrevivência de um *cluster* - $C_4(t) \rightarrow C_3(t)$.

Em $[t+1, t+2]$, detectaram-se:

1. Três sobrevivências - $C_1(t+1) \rightarrow C_3(t+2)$, $C_3(t+1) \rightarrow C_4(t+2)$ e $C_5(t+1) \rightarrow C_1(t+2)$;
2. Fusão de dois *clusters* - $\{C_2(t+1), C_4(t+1)\} \overset{\subset}{\rightarrow} C_2(t+2)$.

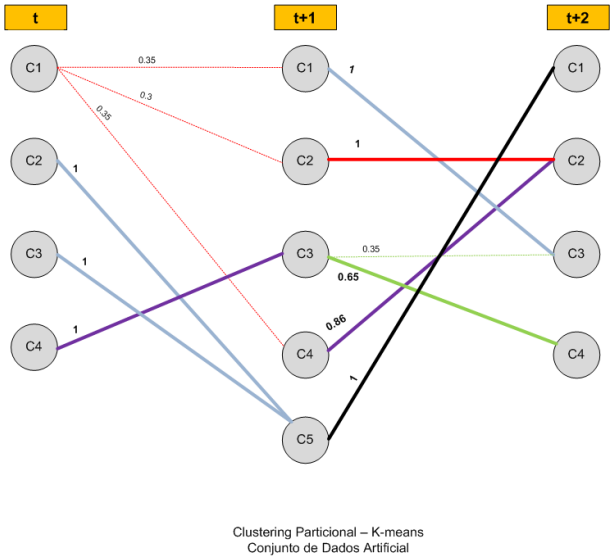


Figura 5.10: Grafos bipartidos, correspondentes aos intervalos de tempo $[t, t+1]$ e $[t+1, t+2]$, dos conjuntos de dados artificiais (algoritmo particional das k-médias), com a espessura das arestas a indicar os pesos superiores ou iguais ao limiar de Sobrevivência $\tau = 0.5$ e superiores ou iguais ao limiar de Cisão $\rho = 0.2$

O confronto destes resultados com as transições definidas na Tabela 5.3 permite constatar a existência de algumas discordâncias. A origem destas diferenças assenta nos limiares seleccionados, tal como já foi explanado e justificado na experiência anterior. Por outro lado, se se comparar as transições detectadas para o algoritmo particional com as transições encontradas para o algoritmo hierárquico, conclui-se que estas são bastante idênticas, o que sugere características de robustez e resiliência do método MEC ao algoritmo de *Clustering* adoptado. Ou seja, o MEC é independente do algoritmo de *Clustering* utilizado para obter as partições dos dados.

Aplicação do MEC para clusters representados em compreensão

Para efeitos da aplicação do método MEC a *clusters* representados em compreensão, em primeiro lugar foi necessário proceder à transformação dos conjuntos de dados originais, que possuem informação detalhada de cada objecto, numa representação baseada nas estatísticas sumárias dos *clusters* descobertos (Figura 5.4 e 5.5). Esta transformação foi operada com base numa função criada para o efeito - `RepComp`. O *output* desta função é o ingrediente necessário à aplicação do método MEC baseado na sobreposição de *clusters*.

Algoritmo Hierárquico Aglomerativo (índice de Ward). Com base no resultado deste método, para o agrupamento gerado com base no algoritmo hierárquico aglomerativo, conclui-se que, no período $[t, t+1]$, houve:

1. Sobrevivência de um *cluster* - $C_3(t) \to C_2(t+1)$;
2. Morte de três *clusters* - $C_1(t) \to \emptyset$, $C_2(t) \to \emptyset$ e $C_4(t) \to \emptyset$;
3. Nascimento de quatro *clusters* - $\emptyset \to C_1(t+1)$, $\emptyset \to C_3(t+1)$, $\emptyset \to C_4(t+1)$ e $\emptyset \to C_5(t+1)$.

O *cluster* sobrevivente não sofreu quaisquer transições internas. Comparando estes resultados com o conhecimento detido *a priori*, verifica-se que o algoritmo detectou bem a sobrevivência de $C_3(t)$ e a inexistência das respectivas modificações internas. No que respeita às restantes transições exógenas, mais especificamente, às mortes e nascimentos, optou-se, em primeiro lugar, por distinguir as *verdadeiras* mortes e nascimentos. Isto efectuou-se verificando se o mesmo *cluster* tinha sofrido, simultaneamente, uma morte e um nascimento. Se isto aconteceu, assume-se que a morte/nascimento é *falsa*. Caso contrário, a morte/nascimento é *verdadeira*. Assim, apenas $C_2(t)$ morreu verdadeiramente e apenas $C_3(t+1)$ e $C_5(t+1)$ tiveram um verdadeiro nascimento. O nascimento de $C_5(t+1)$ corresponde ao nascimento artificialmente imposto. Porém, existem cisões e fusões que, na realidade aconteceram, e que não foram capturadas pelo algoritmo. Esta situação é facilmente justificada pela análise da Figura 5.7. Se atentarmos às duas primeiras imagens, correspondentes a t e a $t+1$, verificamos que a sobreposição dos *clusters* $C_1(t)$ e $C_2(t)$ com o *cluster* $C_1(t+1)$ não acontece (para compreender melhor a explicação o leitor deve imaginar que cada imagem é uma folha de papel e que se coloca a segunda imagem sobre a primeira), principalmente porque $C_2(t)$ está relativamente afastado de $C_1(t+1)$. Na melhor das situações, o algoritmo apenas conseguiria detectar a sobrevivência $C_1(t) \to C_1(t+1)$, porque os pontos estão mais próximos uns dos outros. O mesmo acontece com a cisão de $C_4(t)$ em $C_3(t+1)$ e $C_4(t+1)$. Neste caso, a tarefa de detecção usando técnicas de sobreposição ainda é mais difícil, uma vez que $C_3(t+1)$ e $C_4(t+1)$ estão bastante afastados um do outro e em localizações espaciais bastante diferentes da localização inicial de $C_4(t)$. Assim, é fácil deduzir o porquê das discordâncias do algoritmo MEC com as transições artificiais.
Foram realizadas mais experiências mas impondo outras localizações espaciais dos *clusters* que sofreram cisões e fusões, de forma a que fosse possível ocorrer sobreposição, e os resultados do algoritmo foram extremamente satisfatórios, uma vez que detectou correctamente todas as transições.
Na fase seguinte, aplicou-se o algoritmo de detecção de transições ao intervalo de tempo $[t+1, t+2]$. Os resultados alcançados para este período coincidem exactamente com as transições impostas previamente, uma vez que se detectaram:

1. Quatro sobrevivências - $C_1(t+1) \rightarrow C_1(t+2)$, $C_3(t+1) \rightarrow C_2(t+2)$, $C_4(t+1) \rightarrow C_3(t+2)$ e $C_5(t+1) \rightarrow C_4(t+2)$;
2. Uma morte - $C_2(t+1) \rightarrow \emptyset$.

Para os *clusters* sobreviventes examinámos a eventual existência de transições internas. As transições internas encontradas relacionam-se com alterações ao nível da dimensão e da densidade dos *clusters* e são as seguintes:

1. Contracção e dispersão de $C_1(t+2)$ - $C_1(t+1) \searrow C_1(t+2)$ e $C_1(t+1) \xrightarrow{*} C_1(t+2)$;
2. Expansão de $C_3(t+1)$, $C_4(t+1)$ e $C_5(t+1)$ - $C_3(t+1) \nearrow C_2(t+2)$, $C_4(t+1) \nearrow C_3(t+2)$ e $C_5(t+1) \nearrow C_4(t+2)$;
3. Compactação de $C_4(t+2)$ - $C_5(t+1) \xrightarrow{\bullet} C_4(t+2)$.

Estas transições eram expectáveis e reflectem a verdadeira evolução dos *clusters* ao longo do tempo, uma vez que coincidem com as transições a que os *clusters* foram submetidos artificialmente.

Algoritmo Particional k-médias. Para o agrupamento gerado com recurso ao algoritmo das k-médias, o método retornou os resultados que seguidamente se apresentam: no intervalo de tempo $[t, t+1]$ detectaram-se,

1. Uma sobrevivência, sem transições internas - $C_4(t) \rightarrow C_2(t+1)$;
2. Morte de C_1, C_2 e C_3 em t - $C_1(t) \rightarrow \emptyset$, $C_2(t) \rightarrow \emptyset$ e $C_3(t) \rightarrow \emptyset$;
3. Nascimento de C_1, C_4, C_3 e C_5, em $t+1$ - $\emptyset \rightarrow C_1(t+1)$, $\emptyset \rightarrow C_4(t+1)$, $\emptyset \rightarrow C_3(t+1)$ e $\emptyset \rightarrow C_5(t+1)$.

Analisando estas transições externas, verifica-se que o único nascimento *verdadeiro* foi o do *cluster* C_5, em $t+1$. As restantes mortes/nascimentos resultaram da dificuldade de efectuar a sobreposição dos *clusters* que se fundiram e que se dividiram, uma vez que estes *clusters* estavam distantes uns dos outros no espaço de atributos referente a cada um dos instantes de tempo sob observação (ver Figura A.3). Assim, neste intervalo de tempo, o algoritmo apenas detectou correctamente a sobrevivência de $C_4(t)$ e o nascimento de $C_5(t+1)$. O motivo por detrás da dificuldade do algoritmo em detectar a fusão e a cisão de *clusters* já foi referido na experiência anterior. No que concerne ao intervalo de tempo subsequente $[t+1, t+2]$, o método detectou correctamente todas as transições, mais especificamente:

1. Quatro sobrevivências - $C_1(t+1) \rightarrow C_4(t+2)$, $C_2(t+1) \rightarrow C_2(t+2)$, $C_4(t+1) \rightarrow C_3(t+2)$ e $C_5(t+1) \rightarrow C_1(t+2)$;
2. Morte de $C_3(t+1)$ - $C_3(t+1) \rightarrow \emptyset$.

O mesmo se constata quando se analisam as transições internas:

1. Expansão e compactação de $C_1(t+1)$ - $C_1(t+1) \nearrow C_4(t+2)$ e $C_1(t+1) \overset{\bullet}{\rightarrow} C_4(t+2)$;
2. Contracção e dispersão de $C_2(t+1)$ - $C_2(t+1) \searrow C_2(t+2)$ e $C_2(t+1) \overset{*}{\rightarrow} C_2(t+2)$;
3. Expansão dos restantes *clusters* - $C_4(t+1) \nearrow C_3(t+2)$ e $C_5(t+1) \nearrow C_1(t+2)$.

Desta forma, demonstra-se a capacidade deste método em detectar, com precisão, as transições a que os *clusters* foram submetidos. As pequenas incoerências detectadas prendem-se com as limitações do método e com o facto de não se dispor de toda a informação sobre os dados.

Com base na análise efectuada conclui-se, assim, que a metodologia MEC é exequível e capaz de efectuar o diagnóstico eficiente das transições, exógenas e endógenas, dos *clusters*.

5.2.3 Análise da Sensibilidade dos Limiares

Nas experiências anteriores constatou-se que, uma pequena variação nos valores assumidos pelos limiares τ e ρ, originava resultados diferentes. Esta evidência não deve ser negligenciada e deve ser alvo de uma análise mais profunda, que permita retirar conclusões sobre o impacto destas pequenas variações no resultado final do algoritmo de detecção de transições. Neste sentido, considerou-se relevante efectuar uma análise da sensibilidade do algoritmo aos valores pré-definidos dos limiares. Esta análise incide sobre os dados artificiais, mais especificamente, sobre os *clusters* gerados com o algoritmo hierárquico aglomerativo, e é dividida de acordo com o intervalo de tempo e de acordo com o limiar. Conduziram-se experiências para todos os valores do limiar de Sobrevivência compreendidos no intervalo $[0.5, 1]$, uma vez que se impôs um valor mínimo; e para os valores do limiar de Cisão compreendidos no intervalo $[0, 0.4]$. A análise foi realizada separadamente, fazendo variar os valores de um dos limiares e mantendo tudo o resto constante, ou seja, quando se observa o comportamento de τ, assume-se um valor constante de $\rho = 0.2$, e quando se analisa ρ, considera-se um valor fixo de sobrevivência de $\tau = 0.5$.

Limiar de Sobrevivência

Na Figura 5.11, referente ao intervalo de tempo $[t, t+1]$, é possível observar a relação entre os valores de τ e o número de ocorrências de cada uma das transições exógenas. Em termos lógicos, seria de esperar que, quanto maior o valor do limiar de Sobrevivência, menor o número de sobrevivências detectadas e, consequentemente, de fusões. Seria igualmente expectável que, o aumento de τ gerasse maior número de nascimentos e/ou cisões. Porém, a análise da figura contraria esta hipótese inicial, dado que o número de transições é invariante com o limiar ($sobrevivencias = fusoes = cisoes = 1$ e $nascimentos = mortes = 0$). Isto é justificado pelos pesos atribuídos às arestas do grafo bipartido (ver Figura 5.8), que assumem valor máximo de sobrevivência ($peso = 1$). Este caso reflecte uma situação de **transições permanentes**, ie, de transições robustas a variações do τ. Este tipo de transições não gera dúvidas sobre a evolução sofrida pelos *clusters*, revelando as mudanças mais pertinentes no domínio de conhecimento subjacente. Por sua vez, no intervalo de tempo posterior $[t+1, t+2]$, verifica-se alguma instabilidade nos resultados do algoritmo com a alteração do limiar, o que pode ser observado na Figura 5.12. À medida que τ se aproxima do seu valor máximo, o número de mortes e de cisões aumenta, diminui o número de sobrevivências e de fusões (o pico de sobrevivência atingido para $\tau = 0.9$ deve-se à extinção de uma fusão e, por conseguinte, à transformação desta fusão numa sobrevivência, visto que uma das ligações é podada) e o número de nascimentos mantém-se constante. Neste caso, estamos perante um comportamento mais previsível que corrobora a hipótese inicial. Este tipo de transições são mais voláteis e sensíveis a pequenas variações de τ - **transições relativas** -, revelando mudanças pouco consolidadas no contexto em estudo.

Em suma, quanto mais exigente o limiar de Sobrevivência, menor o número de sobrevivências e de fusões e maior o número de mortes e de cisões.

Limiar de Cisão

Em relação ao limiar de Cisão, espera-se que, quanto maior o seu valor, maior o número de mortes, de nascimentos e menor o número de cisões. Preve-se, igualmente, que o número de fusões e de sobrevivências não seja afectado pela variação de ρ, tendo por base a respectiva definição formal. Analisando a variação do limiar para o período $[t, t+1]$ (Figura 5.13) comprova-se esta hipótese. Contudo, no intervalo seguinte (Figura 5.14), o número de transições exógenas mantém-se inalterado para diferentes valores de ρ. O motivo subjacente a este acontecimento pode ser facilmente deduzido com base na análise do grafo da Figura 5.8, em que se verifica que a poda das ligações com $peso \leq 0.4$, não interfere nos resultados finais. Analogamente ao que se constatou na análise de τ, estas tratam-se de **transições permanentes**, com respeito ao limiar de Cisão e, por isso, o seu impacto nos resultados é nulo.

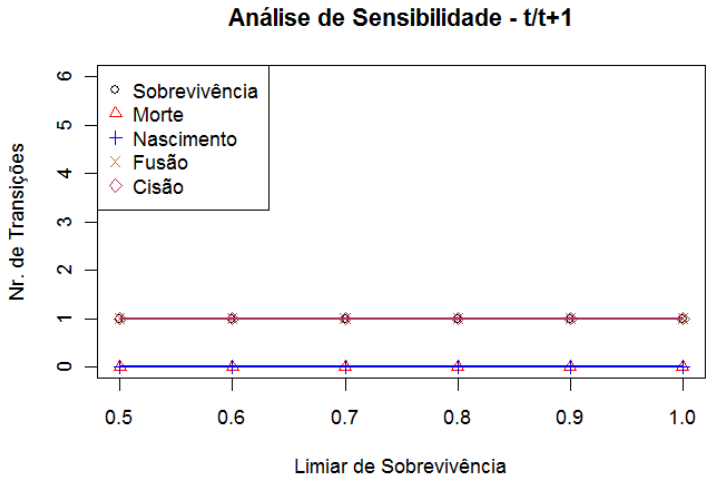


Figura 5.11: Impacto no número de transições exógenas motivado pela variação do limiar de Sobrevivência τ, para o período $[t, t+1]$

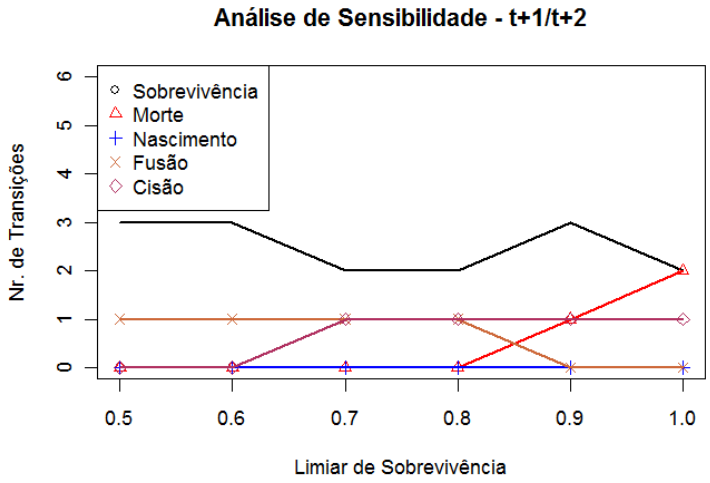


Figura 5.12: Impacto no número de transições exógenas motivado pela variação do limiar de Sobrevivência τ, para o período $[t+1, t+2]$

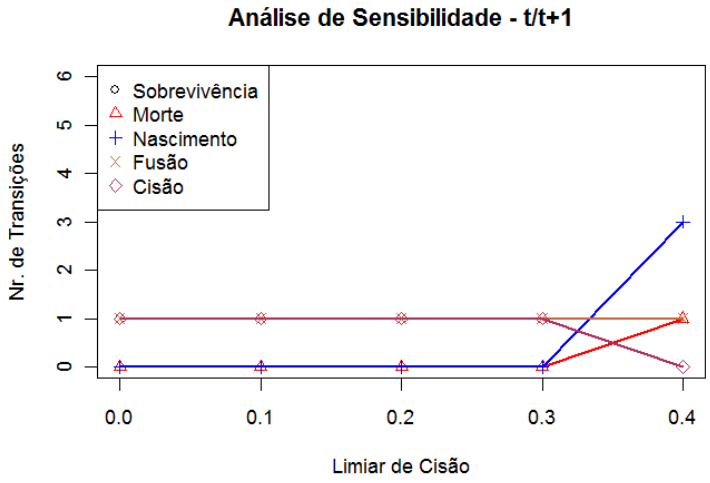


Figura 5.13: Impacto no número de transições exógenas motivado pela variação do limiar de Cisão ρ, para o período $[t, t+1]$

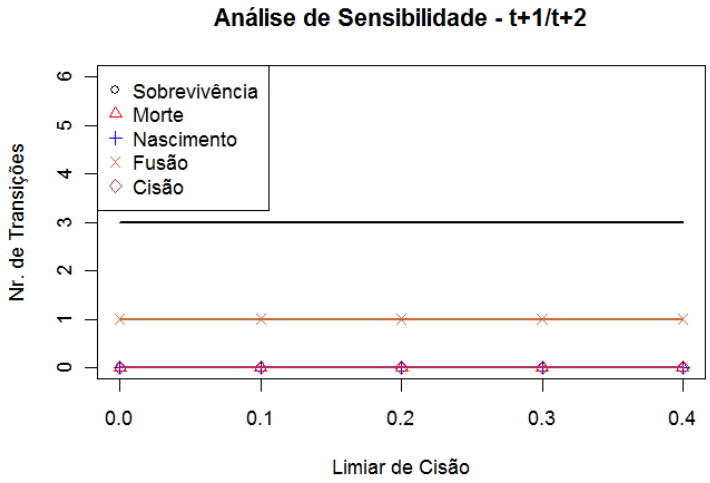


Figura 5.14: Impacto no número de transições exógenas motivado pela variação do limiar de Cisão ρ, para o período $[t+1, t+2]$

Assim, conclui-se que quanto mais exigente o limiar de Cisão, maior o número de mortes/nascimentos e menor o número de cisões consideradas. Destacam-se, igualmente, os conceitos de **transições permanentes** e de **transições relativas**, que revelam a estabilidade das transições ocorridas nas estruturas de *clusters* e, por conseguinte, indicam as mudanças mais, ou menos, proeminentes no domínio de conhecimento.

Definição dos limiares

Tendo por base as conclusões da análise de sensibilidade efectuada, assume-se que os valores $\tau = 0.5$, para o limiar de *Sobrevivência*, e $\rho = 0.2$, para o limiar de *Cisão*, são razoáveis e que, apesar de mais relaxados, têm a vantagem de permitir detectar uma maior variedade de transições. Estes serão, assim, os valores adoptados para a condução dos casos de estudo da secção seguinte.

5.3 Aplicação do MEC a Conjuntos de Dados Reais

Após uma procura exaustiva de conjuntos de dados que se enquadrassem no problema em estudo, optou-se pela utilização de quatro conjuntos de dados provenientes de fontes como o Banco de Portugal, o Instituto Nacional de Estatística (INE) e a Direcção Geral de Administração Interna (DGAI). Os requisitos que estão na base da escolha destes dados são, nomeadamente: a diversidade, o interesse e a pertinência dos problemas que em si encerram; o tipo de variáveis em causa, uma vez que a Análise de *Clusters* é uma técnica comummente aplicada a dados quantitativos ou numéricos; e, ainda, a dimensão do conjunto de dados, importante para assegurar a razoabilidade e a fiabilidade dos resultados do estudo e garantir a inexistência de problemas significativos ao nível da análise destes mesmos resultados.

As experiências que seguidamente se apresentam foram realizadas com base nestes conjuntos de dados e devem ser encaradas como pequenos casos de estudo, que têm como escopo a mera ilustração da aplicabilidade dos métodos, uma vez que nenhuma hipótese formulada *a priori* é testada.

5.3.1 Banco de Portugal - Sectores de Actividade Económica

O primeiro caso de estudo insere-se na área da Economia e consiste numa experiência realizada com dados da Central de Balanços do Banco de Portugal. A Central de Balanços é uma Base de Dados, construída pelo Banco de Portugal, que tem por base a informação reportada pelas empresas não financeiras em Inquéritos Trimestrais e Anuais, realizados em parceria pelo Banco de Portugal e pelo INE. A informação extraída dos Inquéritos tem índole económica e financeira e é de base contabilística, não consolidada. A divulgação desta informação pelo Banco de Portugal "visa con-

tribuir para um melhor conhecimento da situação económica e financeira do sector das sociedades não financeiras portuguesas"[1].

A Central de Balanços do Banco de Portugal origina duas publicações relacionadas: os *Quadros da Empresa e do Sector* e os *Quadros do Sector*. Os dados que servirão de base à avaliação empírica da metodologia genérica proposta no presente trabalho foram extraídos da aplicação *Quadros do Sector*, que pode ser carregada no *website* do Banco de Portugal [2]. Os *Quadros do Sector* são uma Base de Dados que contém diversos indicadores e rácios agregados (cerca de 239 variáveis) sobre os sectores de actividade em que se inserem as empresas não financeiras, classificadas de acordo com a Classificação Portuguesa das Actividades Económicas (CAE). Os indicadores são de base anual e a Base de Dados dispõe de informação para o período temporal compreendido entre 1991 e 2007. No âmbito do estudo da monitorização da evolução de *clusters* optou-se pela exploração dos últimos anos: 2005, 2006 e 2007, o que corresponde a um horizonte temporal de três anos.

Os conjuntos de dados disponibilizados, para cada ano, na Base de Dados *Quadros do Sector* foram submetidos a técnicas de pré-processamento de dados, com o objectivo de orientar a escolha e a selecção das variáveis e o nível de granularidade dos Sectores. Procurou-se seleccionar variáveis ou atributos não redundantes, que apresentassem pouca correlação entre si, mas capazes de caracterizar de forma adequada e completa os sectores de actividade. No que concerne aos objectos, optou-se pelo maior nível de granularidade, que se reflecte na utilização do código de 5 dígitos da CAE, de forma a obter um número razoável de observações [3]. Os conjuntos de dados finais, obtidos após a materialização destas escolhas, e respeitantes aos períodos temporais de 2005, 2006 e 2007, são constituídos por 439 sectores de actividade e por 12 variáveis ou atributos. É de salientar o facto dos sectores de actividade, e respectivos indicadores, serem exactamente os mesmos, para os diferentes períodos em análise. Das 12 variáveis quantitativas, 2 apenas servem o propósito de identificação dos sectores, sendo as restantes 10 focadas na descrição e caracterização dos mesmos. Para melhor compreender o significado de cada uma das variáveis, no Anexo B encontra-se uma breve descrição das mesmas. Como as variáveis estão expressas em diferentes escalas e apresentam dispersões bastante diferentes surgiu, também, a necessidade de submetê-las a um processo de normalização para anular o efeito das unidades de medida.

O objectivo deste pequeno estudo é investigar se existiram mudanças relevantes

[1] Banco de Portugal, Estatísticas das Empresas Não Financeiras da Central de Balanços, Suplemento 5|2005 ao Boletim Estatístico|Dezembro 2005

[2] http://www.bportugal.pt

[3] O sistema de codificação da CAE assenta em cinco níveis: Secção, Divisão, Grupo, Classe e Subclasse. A utilização do maior nível de granularidade, a que correspondem códigos CAE de 5 dígitos, permite uma análise mais detalhada dos sectores, bem como a obtenção de um maior número de objectos.

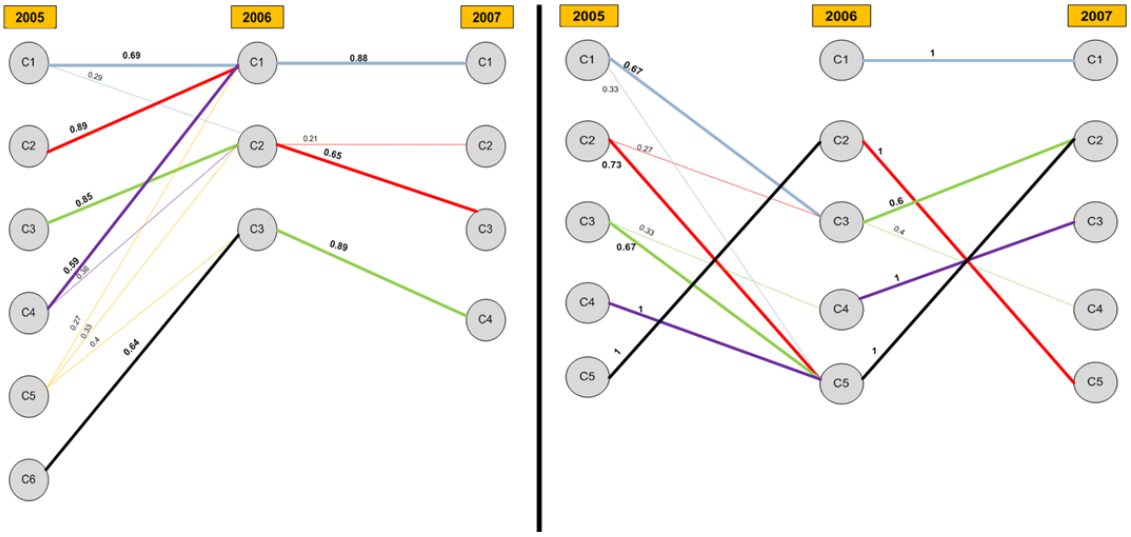


Figura 5.15: Grafos bipartidos, correspondentes aos intervalos de tempo [2005, 2006] e [2006, 2007], dos conjuntos de dados da Central de Balanços do Banco de Portugal. O lado direito do grafo tem os resultados do algoritmo hierárquico e o lado esquerdo os resultados do algoritmo particional.

na estrutura económica do país, por exemplo, se houve crescimento ou declínio de alguns *clusters* de sectores ou se surgiram novos grupos de sectores, representativos de áreas emergentes na Economia.

Aplicação do MEC para clusters representados em extensão

Seguindo as etapas de pré-processamento previstas na Metodologia, obtiveram-se os dendrogramas e gráficos do coeficiente de silhueta médio representados nas Figuras A.4, A.5 e A.6. As primeiras duas figuras referem-se ao número óptimo de *clusters* do algoritmo hierárquico que, nestes dados, parece sugerir uma partição em seis, três e quatro *clusters*, para os anos de 2005, 2006 e 2007, respectivamente. A terceira figura alude ao algoritmo das k-médias, para o qual uma partição em cinco *clusters*, para os três anos em análise, parece ser a melhor opção. Daqui podemos depreender que os resultados do MEC em termos de transições de *clusters* vão ser diferentes para cada um dos algoritmos de agrupamento, uma vez que as estruturas de *clusters* subjacentes são, também elas, diferentes (o número óptimo de *clusters* difere). A fase seguinte consistiu na aplicação do método MEC adequado a *clusters* representados em extensão. Este método retornou os resultados visíveis na Figura 5.15.

Analisando em primeiro lugar os grafos bipartidos do algoritmo hierárquico conclui-se que, de 2005 para 2006, houveram as seguintes transições:

1. Fusão de três *clusters* num só - $\{C_1(2005), C_2(2005), C_4(2005)\} \hookrightarrow C_1(2006)$;

2. Duas sobrevivências - $C_3(2005) \rightarrow C_2(2006)$ e $C_6(2005) \rightarrow C_3(2006)$;

3. Cisão de um *cluster* em três *clusters* - $C_5(2005) \overset{\subsetneq}{\rightarrow} \{C_1(2006), C_2(2006), C_3(2006)\}$.

No período de tempo $[2006, 2007]$, verifica-se:

1. Sobrevivência de todos os *clusters* - $C_1(2006) \rightarrow C_1(2007)$, $C_2(2006) \rightarrow C_3(2007)$ e $C_3(2006) \rightarrow C_4(2007)$;

2. Nascimento de um *cluster* - $\emptyset \rightarrow C_2(2007)$.

O lado esquerdo da Figura 5.15 corresponde aos grafos bipartidos do algoritmo das k-médias. A observação dos grafos indica que, no intervalo de tempo $[2005, 2006]$, ocorreram as seguintes transições externas:

1. Duas sobrevivências - $C_1(2005) \rightarrow C_3(2006)$ e $C_5(2005) \rightarrow C_2(2006)$;

2. Fusão de três *clusters* - $\{C_2(2005), C_3(2005), C_4(2005)\} \overset{\subsetneq}{\rightarrow} C_5(2006)$;

3. Dois nascimentos - $\emptyset \rightarrow C_1(2006)$ e $\emptyset \rightarrow C_4(2006)$.

No período subsequente $[2006, 2007]$:

1. Três *clusters* sobreviveram - $C_1(2006) \rightarrow C_1(2007)$, $C_2(2006) \rightarrow C_5(2007)$ e $C_4(2006) \rightarrow C_3(2007)$;

2. Dois *clusters* fundiram-se - $\{C_3(2006), C_5(2006)\} \overset{\subsetneq}{\rightarrow} C_2(2007)$;

3. Um *cluster* emergiu - $\emptyset \rightarrow C_4(2007)$.

Após a análise dos grafos podemos facilmente constatar que as transições detectadas são diferentes consoante o algoritmo de *Clustering* utilizado. Isto era de prever, uma vez que os agrupamentos obtidos também foram diferentes. No entanto, isto não se verifica quando se utiliza o mesmo algoritmo. Por exemplo, nós testámos se para diferentes inicializações do k-médias as transições se mantinham inalteradas, e constatámos, com base em várias experiências, que isto de facto acontecia. Ou seja, a metodologia MEC é estável para estruturas de *clusters* diferentes, mas obtidas pelo mesmo algoritmo de agrupamento, mantendo tudo o resto constante (número óptimo de *clusters* e limiares τ e ρ).

No entanto, não obstante o facto das transições detectadas diferirem, a fusão de três *clusters* no intervalo de tempo $[2005, 2006]$ foi detectada por ambos os algoritmos. Assim, assumiu-se que esta se tratava de uma transição importante, que deveria ser analisada. A inspecção dos dados permitiu concluir que esta fusão se deveu ao facto dos sectores de actividade afectos a estes *clusters* terem piorado o

Tabela 5.5: Evolução do número de empresas que reportam os respectivos dados económicos e financeiros ao Banco de Portugal, no período compreendido entre 2005 e 2007

Ano	2005	2006	2007
Número de empresas	12068	244595	229129

seu desempenho económico e financeiro, o que se reflectiu na mitigação das suas diferenças iniciais. Mas porque é que isto aconteceu? Para responder a esta questão pesquisámos por informação relevante acerca do tópico e descobrimos que, apesar da Central de Balanços disponibilizar dados agregados sobre as empresas desde o início da década de 90, mais especificamente, desde 1991, em 2006 assistiu-se a um processo de homogeneização da informação, no âmbito do Programa Simplex, que teve como objectivo incorporar, num único documento - a IES (Informação Empresarial Simplificada) -, e numa única operação de entrega, a informação que as empresas são obrigadas a prestar, anualmente, a quatro entidades públicas distintas: o Banco de Portugal, o INE, o Ministério da Justiça e a Administração Fiscal. Além disso, a prestação desta informação adquiriu carácter de obrigatoriedade, o que contribuiu para o aumento substancial do número de empresas a procederem ao preenchimento da IES (ver Tabela 5.5), com consequências positivas no grau de cobertura da Central de Balanços e na fiabilidade das estatísticas produzidas. Este processo poderá ser a razão por detrás desta fusão de *clusters*, uma vez que os dados passaram a reflectir uma imagem mais realística do país.

Outra transição importante, e detectada por ambos os algoritmos, foi o nascimento de um *cluster* em 2007 (no algoritmo hierárquico este *cluster* corresponde a C_2 e no algoritmo particional corresponde a C_4). A prescrutação dos dados permitiu concluir que, em ambos os casos, o *cluster* emergente agrupa os sectores com maiores taxas de investimento e com capital humano intensivo, representativos da Indústria Extractiva (e.g. sector de Extracção de Saibro, areia e pedra britada e sector de Extracção de gesso) e da Indústria da Produção Animal, Caça e Pesca (e.g. sector do Abate de gado, aves e coelhos e sector de Reparação e Congelação de produtos da pesca e da aquacultura). Este aglomerado de sectores é, também, dos que gera maiores lucros. A emergência deste *cluster* poderá, assim, ser indicativa de uma aproximação dos sectores primários na economia portuguesa.

Aplicação do MEC para clusters representados em compreensão

Com o intuito de testar o mecanismo de monitorização para representações compactas de *clusters*, e dado que temos informação detalhada sobre todas as observações, transformámos a representação original - representação em extensão -, num esquema de representação em compreensão, utilizando a função `RepComp`, imple-

mentada em R e criada para o efeito. Após obter uma caracterização dos *clusters* com base nas estatísticas sumárias, e assumindo o número de *clusters* sugerido pelos dendrogramas e pelo coeficiente de silhueta médio, submeteram-se os dados ao método MEC para *clusters* representados em compreensão. Os resultados deste método, para o agrupamento gerado pelo algoritmo hierárquico, foram os seguintes: no período $[2005, 2006]$ detectaram-se,

1. Três fusões de *clusters* - $\{C_1(2005), C_2(2005)\} \overset{\subseteq}{\to} C_1(2006)$, $\{C_1(2005), C_3(2005)\} \overset{\subseteq}{\to} C_2(2006)$ e $\{C_5(2005), C_6(2005)\} \overset{\subseteq}{\to} C_3(2006)$;
2. Uma cisão - $C_1(2005) \overset{\subseteq}{\to} \{C_1(2006), C_2(2006)\}$;
3. Uma morte - $C_4(2005) \to \emptyset$.

No intervalo de tempo subsequente, observam-se:

1. Duas cisões - $C_1(2006) \overset{\subseteq}{\to} \{C_1(2007), C_4(2007)\}$ e $C_2(2006) \overset{\subseteq}{\to} \{C_1(2007), C_2(2007), C_3(2007), C_4(2007)\}$;
2. Duas fusões - $\{C_1(2006), C_2(2006)\} \overset{\subseteq}{\to} C_1(2007)$ e $\{C_1(2006), C_2(2006), C_3(2006)\} \overset{\subseteq}{\to} C_4(2007)$.

No que concerne ao algoritmo das k-médias, no intervalo de tempo $[2005, 2006]$, detectaram-se:

1. Duas sobrevivências - $C_1(2005) \to C_3(2006)$ e $C_4(2005) \to C_5(2006)$;
2. Três mortes - $C_2(2005) \to \emptyset$, $C_3(2005) \to \emptyset$ e $C_5(2005) \to \emptyset$;
3. Três nascimentos - $\emptyset \to C_1(2005)$, $\emptyset \to C_2(2005)$ e $\emptyset \to C_4(2005)$.

Para os *clusters* sobreviventes investigou-se a possibilidade de ocorrência de transições endógenas e constatou-se que ambos os *clusters* se expandem e ficam mais dispersos ($C_1(2005) \nearrow C_3(2006)$, $C_4(2005) \nearrow C_5(2006)$, $C_1(2005) \overset{*}{\to} C_3(2006)$ e $C_4(2005) \overset{*}{\to} C_5(2006)$). Por sua vez, no decurso de $[2006, 2007]$, verificam-se transições muito idênticas às do período anterior:

1. Duas sobrevivências, com modificações ao nível da cardinalidade e da densidade - $C_3(2006) \to C_4(2007)$ e $C_5(2006) \to C_2(2007)$, $C_3(2006) \searrow C_4(2007)$, $C_3(2006) \overset{\bullet}{\to} C_4(2007)$, $C_5(2006) \nearrow C_2(2007)$ e $C_5(2006) \overset{*}{\to} C_2(2007)$;
2. Três mortes - $C_1(2006) \to \emptyset$, $C_2(2006) \to \emptyset$ e $C_4(2006) \to \emptyset$;
3. Três nascimentos - $\emptyset \to C_1(2007)$, $\emptyset \to C_3(2007)$ e $\emptyset \to C_5(2007)$.

A comparação dos resultados deste método com o anterior torna óbvia a diferença nas transições detectadas. Esta diferença é justificada pela concepção distinta dos dois métodos (um baseia-se em probabilidades condicionadas e outro baseia-se na aferição do grau de sobreposição dos *clusters* no espaço de atributos), como já foi previamente explicado nas experiências com dados artificiais.

Para finalizar a análise dos dados do Banco de Portugal, efectuou-se uma experiência que consistiu no seguinte: primeiro, agregaram-se os três conjuntos de dados num só conjunto, mas mantendo as etiquetas temporais (cada objecto aparece três vezes); em segundo lugar, aplicou-se um algoritmo de *Clustering* escolhendo a partição dos dados sugerida pela análise do coeficiente de silhueta médio; seguidamente, averiguou-se se as observações correspondentes ao mesmo sector, mas para as etiquetas temporais 2005, 2006 e 2007, tinham sido afectas ao mesmo *cluster*. A hipótese por detrás desta experiência é a seguinte: sectores que sofreram mudanças significativas no triénio, devem ter as respectivas observações dispersas por vários *clusters* (e.g. o sector x sofreu mudanças significativas se, cumulativamente, a observação $sector_x2005$ pertencer ao *cluster* C_1, a observação $sector_x2006$ estiver atribuída ao *cluster* C_2 e a observação $sector_x2007$ for afecta ao *cluster* C_3). Adoptando um número óptimo de *clusters* igual a seis, para o algoritmo hierárquico, e assumindo uma partição em cinco *clusters*, para o algoritmo particional, concluiu-se que as observações associadas aos sectores atribuídos a *clusters* que tinham sofrido fusões, cisões e nascimentos, tinham sido, nesta experiência, afectos a *clusters* diferentes e que as observações associadas aos sectores dos *clusters* sobreviventes estavam concentrados num só *cluster*. Deste modo, confirma-se a hipótese inicial.

5.3.2 INE - Estudantes matriculados no Ensino Não-Superior

O segundo caso de estudo foi realizado com os dados do INE, disponíveis para o triénio 2001, 2002 e 2003, e incide sobre a área da Educação. Cada conjunto de dados é composto por 30 observações, correspondentes às 30 unidades de análise da NUTS III (Nomenclatura de Unidades Territoriais para Fins Estatísticos III), e por cinco atributos quantitativos discretos, expressos em termos do número de estudantes matriculados em cada nível de ensino não-superior ministrado em Portugal - Pré-escolar, Primeiro Ciclo, Segundo Ciclo, Terceiro Ciclo e Secundário. É de sublinhar o facto de não se discriminar o número de estudantes matriculados por natureza institucional (pública ou privada).

Este conjunto de dados poderá servir vários propósitos mas, neste contexto, será utilizado com o objectivo de perceber a evolução do acesso à Educação não-superior em Portugal e, consoante os resultados, inferir a existência de reformas assinaláveis na Educação do país.

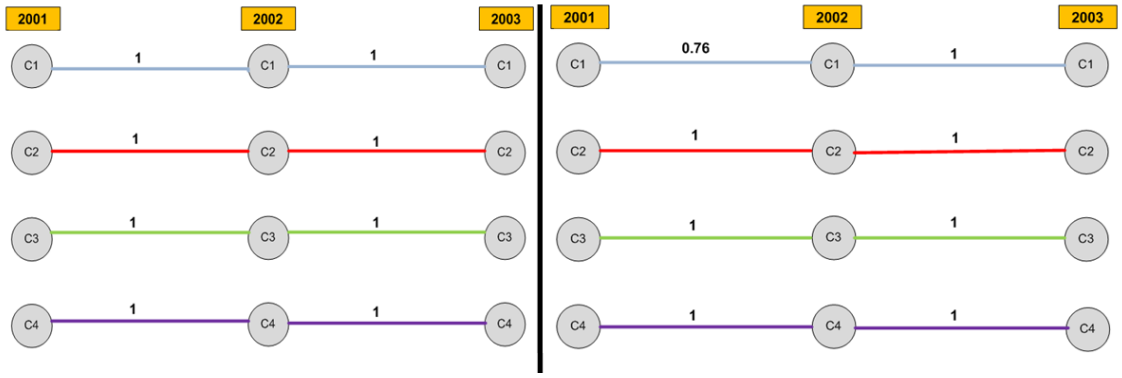


Figura 5.16: Grafos bipartidos, correspondentes aos intervalos de tempo [2001, 2002] e [2002, 2003], dos conjuntos de dados do INE (Educação), com a espessura das arestas a indicar os pesos superiores ou iguais ao limiar de Sobrevivência $\tau = 0.5$. As ligações com pesos inferiores ao limiar de Cisão foram removidos dos grafos, devido à sua insignificância. O lado direito do grafo tem os resultados do algoritmo hierárquico e o lado esquerdo os resultados do algoritmo particional.

Aplicação do MEC para clusters representados em extensão

Este caso de estudo foi encetado assumindo que os *clusters* estavam representados em extensão. Os gráficos referentes ao pré-processamento podem ser visualizados nas Figuras A.7, A.8 e A.9. A partir dos gráficos deduz-se que, com base no algoritmo hierárquico, o número óptimo de *clusters* é quatro, para todos os anos em estudo. No que respeita ao algoritmo das k-médias, a partição óptima sugerida é de cinco *clusters* para o triénio 2001, 2002 e 2003. Contudo, observou-se que a aplicação do algoritmo nestas condições originava uma partição em que um dos *clusters* não possuía nenhum objecto membro. Esta constatação motivou a escolha de um menor número de *clusters* tendo-se seleccionado o mesmo número de *clusters* que o algoritmo hierárquico, ie, quatro grupos.

Numa fase posterior, submeteram-se as estruturas de *clusters* devolvidas por ambos os algoritmos de agrupamento, ao método MEC para *clusters* representados em extensão. Os grafos resultantes, que ilustram as transições detectadas, estão representados na Figura 5.16.

A análise da Figura 5.16 permite efectuar um diagnóstico das mudanças ocorridas em dois intervalos de tempo: [2001, 2002] e [2002, 2003]. Como se pode concluir observando os grafos, todos os *clusters* sobreviveram. Esta conclusão estende-se às duas alternativas de agrupamento. De facto, ambos os algoritmos de *Clustering* concordam que não houve nenhuma mudança estrutural ao longo dos três anos. Isto indica, claramente, que no período sob análise não se registaram alterações significativas no número de estudantes matriculados no ensino não-superior. A situação

observada permite também inferir, embora com alguma incerteza associada que, de 2001 a 2003, Portugal não implementou reformas educacionais assinaláveis no contexto estudado.

Aplicação do MEC para clusters representados em compreensão

Supondo que a informação sobre todas as sub-regiões, relativas ao número de alunos matriculados em cada etapa do ensino não-superior, não estava disponível de forma detalhada e que apenas os dados sumários de cada um dos quatro *clusters* eram disponibilizados, o mecanismo mais adequado para monitorizar a evolução dos grupos, nestes casos, é o método MEC para *clusters* representados em compreensão. O emprego deste método produziu os seguintes resultados: recorrendo às estruturas geradas pelo algoritmo hierárquico constata-se que todos os *clusters* sobrevivem, como seria de prever através da análise dos dendrogramas que são muito idênticos entre si, em ambos os intervalos temporais. Porém, detectam-se algumas modificações internas, respeitantes à respectiva densidade, nomeadamente:

1. Dispersão - $C_1(2001) \xrightarrow{*} C_1(2002)$, $C_1(2002) \xrightarrow{*} C_1(2003)$, $C_2(2002) \xrightarrow{*} C_2(2003)$ e $C_3(2002) \xrightarrow{*} C_3(2003)$;

2. Compactação - $C_2(2001) \xrightarrow{\bullet} C_2(2002)$, $C_3(2001) \xrightarrow{\bullet} C_3(2002)$, $C_4(2001) \xrightarrow{\bullet} C_4(2002)$ e $C_4(2002) \xrightarrow{\bullet} C_4(2003)$.

Analogamente, usando o algoritmo das k-médias, o método captou, no período $[2001, 2002]$:

1. Duas sobrevivências, com aumento da densidade - $C_1(2001) \rightarrow C_1(2002)$, $C_2(2001) \rightarrow C_1(2002)$, $C_1(2001) \xrightarrow{\bullet} C_1(2002)$ e $C_2(2001) \xrightarrow{\bullet} C_2(2002)$;

2. Duas mortes - $C_3(2001) \rightarrow \emptyset$ e $C_4(2002) \rightarrow \emptyset$;

3. Dois nascimentos - $\emptyset \rightarrow C_3(2002)$ e $\emptyset \rightarrow C_4(2002)$.

No decurso do intervalo $[2002, 2003]$, observam-se as mesmas transições exógenas, porém os *clusters* sobreviventes sofrem uma redução da densidade, e não um aumento, como verificado no período anterior ($C_1(2002) \xrightarrow{*} C_1(2003)$ e $C_2(2002) \xrightarrow{*} C_2(2003)$).

Comparação dos resultados

Comparando os resultados obtidos para ambos os esquemas de representação, concluímos que o seu comportamento difere e que os mecanismos subjacentes a cada método analisam a mesma realidade segundo diferentes perspectivas. Em termos gerais, as ilações que podemos retirar são as mesmas, isto porque, a morte de um

cluster específico seguido de um nascimento desse mesmo *cluster* poderá ser entendido como uma sobrevivência. Neste caso concreto, esta conclusão só pode ser alcançada porque se detém conhecimento sobre os resultados atingidos por cada um dos métodos e estes podem, de certa forma, ser combinados e comparados. Para comprovar a teoria de que a morte/nascimento de um *cluster* corresponde, neste caso em particular, a uma sobrevivência, apresentamos a representação gráfica dos grupos num espaço bidimensional (Figura A.10). Nesta figura verificamos que dois, dos quatro *clusters*, são formados por apenas uma região. Estas regiões são *outliers* e correspondem às cidades mais populosas de Portugal: Grande Porto e Grande Lisboa. A análise temporal da representação bidimensional destes *clusters* permite constatar que apenas houveram sobrevivências e que, estruturalmente, não se registaram mudanças significativas. Não obstante esta evidência, o método para *clusters* representados em compreensão detectou mortes e nascimentos. Esta situação é justificada pelo processo utilizado no mapeamento dos *clusters*: o facto de se explorar o conceito de raio implica, necessariamente, que o *cluster* seja formado por mais do que uma observação; caso contrário, o raio é zero e, consequentemente, a distância entre os centróides será sempre superior à soma dos raios. Por conseguinte, mesmo que exista uma correspondência de *clusters*, como acontece neste caso, esta hipótese é imediatamente descartada. Este caso de estudo permitiu, assim, detectar uma das limitações do método, que se resume à assumpção de que todos os *clusters* possuem cardinalidade superior a 1 ($n_k > 1$).

5.3.3 INE - Índice Sintético de Desenvolvimento Regional

O terceiro caso de estudo foi elaborado com base em dois conjuntos de dados extraídos do INE, um referente ao ano de 2004 e outro ao ano de 2006. Contudo, estes dados abordam uma problemática distinta da anterior, uma vez que se focam nas questões do Território e do desenvolvimento regional nacional, cujo estudo se "afigura útil para apoiar a análise de contexto das políticas públicas territorializadas ou com impactos diferenciados no território, bem como servir de base de trabalho para múltiplos agentes interessados nas questões do território"[4]. Cada conjunto de dados contém 30 objectos, correspondentes às unidades de análise da NUTS III, caracterizados por três atributos contínuos (índices normalizados) que sumarizam o desenvolvimento das sub-regiões em todos os aspectos considerados relevantes, nomeadamente, *Coesão*, *Competitividade* e *Qualidade ambiental*.

Com a monitorização destes dados no período temporal [2004, 2006], pretende-se apurar a existência de convergência ou divergência das 30 sub-regiões portuguesas, em termos de desenvolvimento territorial, ao longo do tempo.

[4]INE, Destaque: Índice Sintético de Desenvolvimento Regional, publicação de 26 Maio de 2006

Aplicação do MEC para clusters representados em extensão

As Figuras A.11, A.12 e A.13 revelam que os dados devem ser agrupados em oito e quatro *clusters*, no caso do algoritmo hierárquico e para os anos 2004 e 2006; e em sete *clusters*, para ambos os anos, no caso do algoritmo das k-médias.

A nossa abordagem para *clusters* em extensão retornou os resultados demonstrados na Figura 5.17.

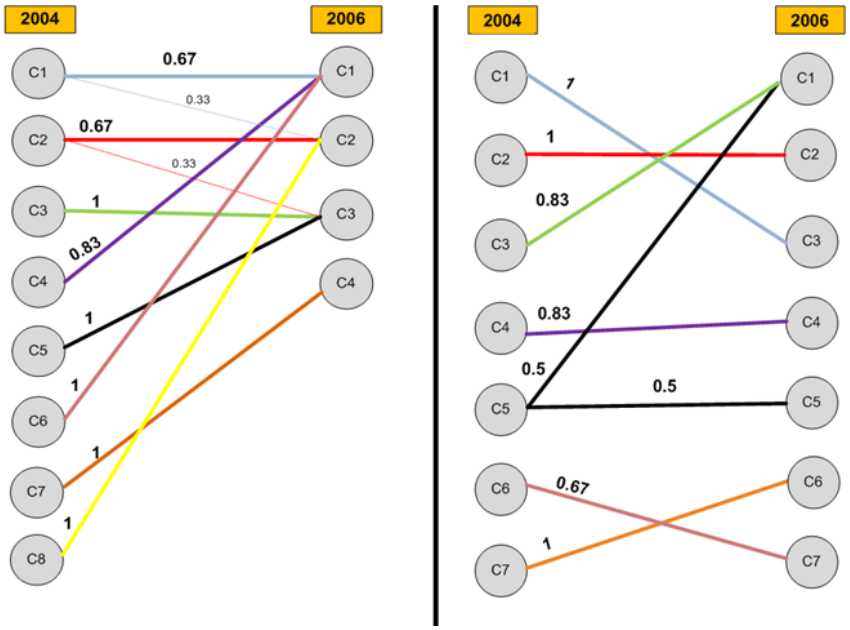


Figura 5.17: Grafos bipartidos, correspondentes ao intervalo de tempo $[2004, 2006]$, dos conjuntos de dados do INE (Território). O lado direito do grafo tem os resultados do algoritmo hierárquico e o lado esquerdo os resultados do algoritmo particional.

A análise do grafo bipartido obtido com recurso ao algoritmo hierárquico, para o intervalo de tempo $[2004, 2006]$, e que se encontra ilustrado no lado esquerdo da figura mencionada, sugere a existência de:

1. Três fusões - $\{C_1(2004), C_4(2004), C_6(2004)\} \subsetneq C_1(2006)$,
   $\{C_2(2004), C_8(2004)\} \subsetneq C_2(2006)$
   e $\{C_3(2004), C_5(2004)\} \subsetneq C_3(2006)$;
2. Uma sobrevivência - $C_7(2004) \rightarrow C_4(2006)$.

O elevado número de fusões era, de certo modo, expectável uma vez que em 2004 as observações estavam distribuídas por um número muito superior de *clusters*. Esta

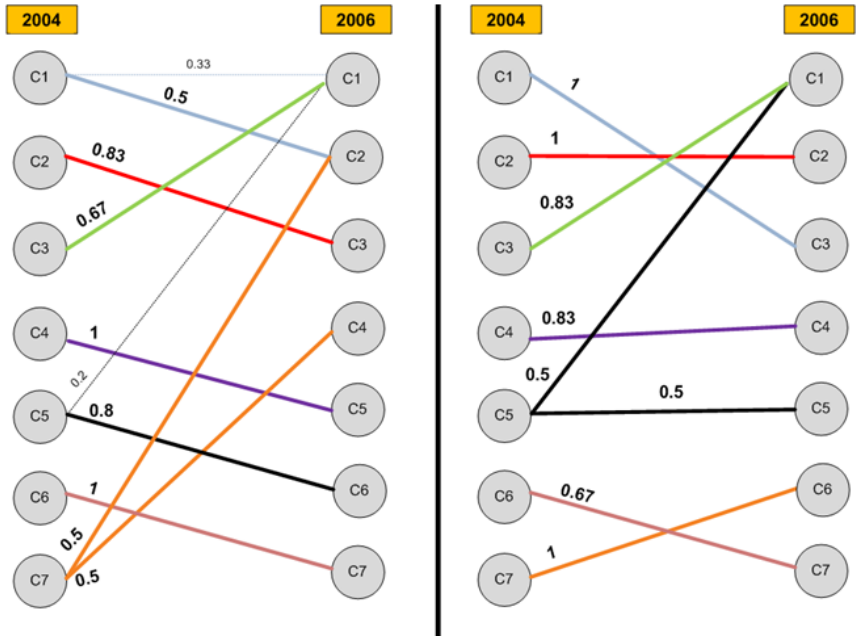


Figura 5.18: Grafos bipartidos, correspondentes ao intervalo de tempo $[2004, 2006]$, dos conjuntos de dados do INE (Território) e assumindo o mesmo número de *clusters* para ambos os algoritmos de agrupamento. O lado direito do grafo tem os resultados do algoritmo hierárquico e o lado esquerdo os resultados do algoritmo particional.

situação obriga a que, no instante posterior, os *clusters* mais próximos se agrupem num só *cluster*, congruindo para a ocorrência de um maior número de fusões. O segundo grafo bipartido respeita ao agrupamento gerado com base no algoritmo das k-médias, e a sua análise revela a existência das seguintes transições exógenas:

1. Seis sobrevivências - $C_2(2004) \rightarrow C_2(2006)$, $C_1(2004) \rightarrow C_3(2006)$, $C_4(2004) \rightarrow C_4(2006)$, $C_5(2004) \rightarrow C_5(2006)$, $C_6(2004) \rightarrow C_7(2006)$ e $C_7(2004) \rightarrow C_6(2006)$;

2. Fusão de dois *clusters* - $\{C_3(2004), C_5(2004)\} \subsetneq C_1(2006)$.

Da análise deste último grafo deve-se conceder especial destaque à transição do *cluster* $C_5(2004)$, uma vez que contempla uma situação pouco vulgar e que consiste no facto de possuir duas ligações com o mesmo peso. O curioso é que, apesar do grafo parecer sugerir uma cisão do *cluster* em dois *clusters* ($C_1(2006)$ e $C_5(2006)$, respectivamente), de acordo com a definição formal das transições exógenas, constante na Tabela 4.1, esta suposta divisão deve ser considerada uma sobrevivência, uma vez que o peso iguala o Limiar de Sobrevivência definido *a priori* $\tau = 0.5$. Assim, e

apesar de na prática, esta situação revelar nitidamente uma divisão de *clusters*, em que metade das observações do *cluster* $C_5(2004)$ são transferidas para $C_1(2006)$ e a outra metade para $C_5(2006)$, iremos assumir como uma sobrevivência do *cluster* em dois outros *clusters*.

O comportamento distinto dos dois algoritmos de *Clustering*, em termos de transições exógenas, é outro dos aspectos que merece alguma atenção. Este comportamento é justificado pelo facto de cada algoritmo de agrupamento estar sustentado em métodos diferentes, como referido no Capítulo 3, pelo que as estruturas descobertas, por cada um deles, tendem também a ser diferentes. Com o intuito de validar esta hipótese realizou-se uma experiência assumindo o mesmo número de *clusters* para ambos os algoritmos (sete *clusters* para os dois anos em análise). Os resultados da aplicação do algoritmo de detecção de transições podem ser visualizados na Figura 5.18. Nesta imagem podemos comprovar que, impondo o mesmo número de *clusters* a ambos os algoritmos, os resultados do nosso sistema de monitorização são os mesmos (uma fusão e seis sobrevivências), ie, a hipótese de que os resultados diferem devido ao facto das estruturas descobertas pelos algoritmos de *Clustering* serem diferentes tem sustentação empírica.

Para compreender as mudanças sugeridas pelas transições detectadas, mais especificamente, pela fusão de dois *clusters* no período $[2004, 2006]$, procurámos descobrir os motivos que poderão estar na base destas diferenças inter-anuais. A inspecção dos dados sugere que a causa da fusão de *clusters* foi a convergência das sub-regiões do Norte do país ($C_1(2004)$ - Cávado, Ave e Entre Douro e Vouga) com a região da Madeira ($C_7(2004)$), despoletada por uma melhoria do nível de qualidade ambiental destas regiões. Contudo, o elevado número de sobrevivências é um factor indicativo de que, em geral, não houveram diferenças significativas no desenvolvimento regional de Portugal, no período compreendido entre 2004 e 2006.

Aplicação do MEC para clusters representados em compreensão

No período de tempo em análise, o método MEC baseado no grau de sobreposição detectou:

1. Cinco mortes - $C_1(2004) \rightarrow \emptyset$, $C_5(2004) \rightarrow \emptyset$, $C_6(2004) \rightarrow \emptyset$, $C_7(2004) \rightarrow \emptyset$ e $C_8(2004) \rightarrow \emptyset$;
2. Nascimento de C_1 em 2006 - $\emptyset \rightarrow C_1(2006)$;
3. Sobrevivência dos restantes *clusters* - $C_2(2004) \rightarrow C_2(2006)$, $C_3(2004) \rightarrow C_3(2006)$ e $C_4(2004) \rightarrow C_4(2006)$.

Os *clusters* sobreviventes sofreram mutações internas, nomeadamente, ao nível da cardinalidade, que aumentou para C_2 e C_3 ($C_2(2004) \nearrow C_2(2006)$ e $C_3(2004) \nearrow$

$C_3(2006)$), e ao nível da densidade, verificando-se a existência quer de dispersões ($C_2(2004) \xrightarrow{*} C_2(2006)$ e ($C_3(2004) \xrightarrow{*} C_3(2006)$), quer de compactações ($C_4(2004) \xrightarrow{\bullet} C_4(2006)$). Por outro lado, no agrupamento gerado pelo algoritmo das k-médias, detectam-se:

1. Quatro mortes - $C_1(2004) \rightarrow \emptyset$, $C_3(2004) \rightarrow \emptyset$, $C_5(2004) \rightarrow \emptyset$ e $C_7(2004) \rightarrow \emptyset$;
2. Três nascimentos - $\emptyset \rightarrow C_1(2006)$, $\emptyset \rightarrow C_3(2006)$, $\emptyset \rightarrow C_5(2006)$ e $\emptyset \rightarrow C_6(2006)$;
3. Três sobrevivências - $C_2(2004) \rightarrow C_2(2006)$, $C_4(2004) \rightarrow C_4(2006)$ e $C_6(2004) \rightarrow C_7(2006)$.

Os *clusters* sobreviventes sofrem, no entanto, mutações internas. De facto, verifica-se que todos os *clusters* se dispersam ao longo do tempo ($C_2(2004) \xrightarrow{*} C_2(2006)$, $C_4(2004) \xrightarrow{*} C_4(2006)$ e $C_6(2004) \xrightarrow{*} C_6(2006)$). Adicionalmente, os *clusters* C_2 e C_4 sujeitam-se a uma redução do respectivo número de membros ($C_2(2004) \searrow C_2(2006)$ e $C_4(2004) \searrow C_4(2006)$).

Comparação dos resultados

A examinação das transições detectadas por cada uma das abordagens conduz-nos, mais uma vez, a resultados diferentes. Para o caso do algoritmo hierárquico, enquanto que no método baseado nas probabilidades condicionadas se detectam três fusões e uma sobrevivência, no método baseado na sobreposição dos *clusters*, detectam-se três sobrevivências e cinco mortes. O mesmo tipo de diferenças podem ser encontradas nos *Clusterings* gerados pelo algoritmo das k-médias em que, no primeiro caso, se detecta uma fusão e seis sobrevivências, e no segundo caso detectam-se quatro mortes, três nascimentos e três sobrevivências. Se analisarmos com atenção estas diferenças, verificamos que existe algo que é comum e recorrente: o método para *clusters* em extensão detecta com mais frequência fusões e cisões, e o método para *clusters* em compreensão assinala com maior regularidade a existência de mortes e nascimentos de *clusters*. As causas que poderão justificar esta situação relacionam-se com os valores definidos para o limiar de Sobrevivência e de Cisão. Até ao momento, temos utilizado os valores *default* $\tau = 0.5$ e $\rho = 0.2$, que são valores relaxados. Se, porventura, o método baseado nas probabilidades condicionadas fosse aplicado utilizando limiares mais exigentes como, por exemplo, $\tau = 0.9$ e $\rho = 0.4$, o número de sobrevivências, fusões e cisões seria substancialmente menor. Assim, repetindo as experiências com estes limiares, obtiveram-se resultados muito idênticos aos alcançados pelo método baseado na sobreposição de *clusters*, comprovando-se a teoria exposta.

5.3.4 DGAI - Resultados das Eleições Legislativas

O último caso de estudo debruça-se sobre a área temática da Política, incidindo sobre os dados referentes aos resultados eleitorais para a Assembleia da República portuguesa, dos últimos anos (2002, 2005 e 2009). Cada conjunto de dados é composto por 308 objectos (concelhos de Portugal) e por seis variáveis quantitativas, que expressam a frequência absoluta de votos válidos nos cinco maiores partidos políticos - BE, CDS, PCP, PSD e PS -, e nos partidos minoritários - Outros [5].

Com esta análise pretende-se investigar a eventual estabilidade ou alteração da ideologia política dominante no país.

Aplicação do MEC para clusters representados em extensão

A evidência do número óptimo de *clusters* é retirada da informação contida nas Figuras A.14, A.15 e A.16. Ambos os algoritmos de *Clustering* concordam que, para os três anos em estudo, o número ideal de grupos é quatro, embora três grupos também fosse uma boa opção.

Assumindo esta partição dos dados, o método MEC para *clusters* representados em extensão obteve os resultados demonstrados na Figura 5.19. À partida, e efectuando uma analogia com o caso de estudo da Educação, o facto do número de *clusters* sugerido pelo algoritmo de agrupamento ser o mesmo para os três anos, cria indícios de que transições como cisão, fusão, morte ou nascimento de *clusters* foram muito pouco significativas ou mesmo inexistentes. A análise dos pares de grafos da Figura 5.19, para cada algoritmo de *Clustering*, vem corroborar esta hipótese inicial. De facto, em ambos os casos, apenas se registam sobrevivências de *clusters*. A inexistência de mudanças revela que, no horizonte temporal e espacial considerado, os concelhos de Portugal mantiveram as suas opções políticas relativamente estáveis. Porém, a estabilidade não foi total, uma vez que os pesos associados a algumas ligações (e.g., $C_3(2002) \rightarrow C_3(2005)$, no algoritmo hierárquico, e $C_1(2005) \rightarrow C_1(2009)$, no algoritmo das k-médias) estão um pouco afastados da probabilidade máxima $peso = 1$, indicativa da total estabilidade dos *clusters*. Contextualizando, poderá concluir-se que, apesar da ideologia política dominante se situar no centro do espectro político (centro-direita, representado pelo PSD e centro-esquerda representado pelo PS), poderão ter existido oscilações nas opções políticas de determinados concelhos que determinaram o partido vencedor das eleições legislativas, em cada um dos períodos. Isto verificou-se na prática pois, enquanto que em 2002 o PSD venceu as eleições legislativas, em 2005 o poder político foi transferido para o PS (o que se manteve em 2009).

[5] Os votos respeitantes a cada um dos partidos minoritários foram agrupados numa única variável, que designámos por "Outros".

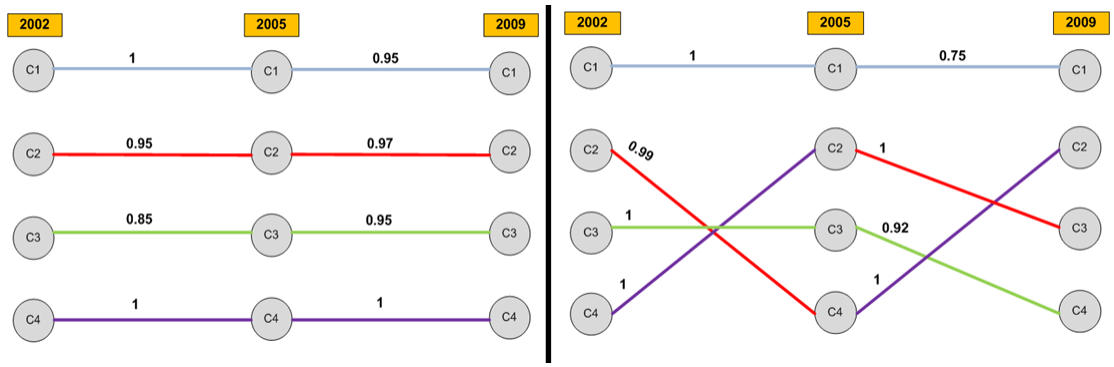


Figura 5.19: Grafos bipartidos, correspondentes aos intervalos de tempo $[2002, 2005]$ e $[2005, 2009]$, dos conjuntos de dados do DGAI. O lado direito do grafo tem os resultados do algoritmo hierárquico e o lado esquerdo os resultados do algoritmo particional.

Aplicação do MEC para clusters representados em compreensão

O *output* do algoritmo de detecção de transições para *clusters* representados em compreensão detectou, para o agrupamento do algoritmo hierárquico e no período $[2002, 2005]$:

1. Sobrevivência de três *clusters* - $C_1(2002) \rightarrow C_1(2005)$, $C_2(2002) \rightarrow C_2(2005)$ e $C_3(2002) \rightarrow C_3(2005)$;
2. Morte e nascimento do *cluster* C_4 - $C_4(2002) \rightarrow \emptyset$ e $\emptyset \rightarrow C_4(2005)$.

Os grupos de concelhos "sobreviventes"sofreram, também, transições de carácter interno:

1. Expansão e dispersão de C_1 - $C_1(2002) \nearrow C_1(2005)$ e $C_1(2002) \overset{*}{\rightarrow} C_1(2005)$;
2. Contracção e compactação de C_2 e C_3 - $C_2(2002) \searrow C_2(2005)$, $C_3(2002) \searrow C_3(2005)$, $C_2(2002) \overset{\bullet}{\rightarrow} C_2(2005)$ e $C_3(2002) \overset{\bullet}{\rightarrow} C_3(2005)$.

No mesmo período, o algoritmo das k-médias captura exactamente as mesmas transições externas, diferindo apenas nas transições internas do *cluster* C_1, que apenas se dispersa ($C_1(2002) \overset{*}{\rightarrow} C_1(2005)$), e do *cluster* C_3, que em vez de se contrair, expande-se ($C_3(2002) \searrow C_3(2005)$). No intervalo de tempo seguinte $[2005, 2009]$, o algoritmo hierárquico apenas detecta sobrevivências, mas com modificações ao nível da dimensão e densidade, nomeadamente:

1. Expansão e dispersão de C_1 - $C_1(2005) \nearrow C_1(2009)$ e $C_1(2005) \overset{*}{\rightarrow} C_1(2009)$;

2. Contracção e dispersão de C_2 - $C_2(2005) \searrow C_2(2009)$ e $C_2(2005) \xrightarrow{*} C_2(2009)$;

3. Compactação de C_3 e C_4 - $C_3(2005) \xrightarrow{\bullet} C_3(2009)$ e $C_4(2005) \xrightarrow{\bullet} C_4(2009)$.

Por sua vez, o algoritmo das k-médias captura o mesmo número e tipo de transições exógenas que no período antecedente (neste caso, o *cluster* que nasce e morre é C_2 ($C_2(2005) \rightarrow \emptyset$ e $\emptyset \rightarrow C_3(2009)$, em vez de C_4), discordando apenas nas transições endógenas. Assim, ao nível interno verifica-se:

1. Contracção e compactação de C_1 - $C_1(2005) \searrow C_1(2009)$
   e $C_1(2005) \xrightarrow{\bullet} C_1(2009)$;

2. Expansão e dispersão de C_4 - $C_4(2005) \nearrow C_2(2009)$ e $C_4(2005) \xrightarrow{*} C_2(2009)$;

3. Dispersão de C_3 - $C_3(2005) \xrightarrow{*} C_4(2009)$.

É importante relembrar que as expansões (ou contracções) se referem, neste contexto, a um aumento (ou diminuição) do número de concelhos pertencente a um dado *cluster*, e que as dispersões (ou compactações) dizem respeito ao afastamento (ou aproximação) dos concelhos em termos de votos nos 6 partidos considerados.

Comparação dos resultados

No que concerne aos resultados das eleições legislativas, constatou-se que as transições exógenas detectadas por ambos os algoritmos de *Clustering* são as mesmas. Existem diferenças subtis num ou noutro *cluster*, cuja transição é categorizada de morte/nascimento, em detrimento de sobrevivência. No entanto, e como já foi referido no caso de estudo da Educação, esta desigualdade deve-se à existência de *clusters* compostos por um único concelho (Figura A.17). Neste caso, o *outlier* é o concelho de Lisboa, o que é justificado pela elevada discrepância do número de habitantes em relação aos restantes concelhos, o que, por sua vez, se reflecte em frequências absolutas significativamente superiores. Deste modo, pode-se concluir que ambas as abordagens da estrutura de monitorização proposta neste trabalho alcançam os mesmos resultados, e que as diferenças são originadas por *outliers* ou por limiares mais relaxados. De facto, se se remover as observações anormais ou se se ajustarem os parâmetros do modelo (τ e ρ) os resultados obtidos por ambos os métodos convergem.

Capítulo 6

Conclusões

Neste capítulo apresentamos as principais conclusões retidas, bem como as perspectivas de desenvolvimento futuro do presente trabalho.

6.1 Resultados

Nesta dissertação desenvolvemos uma metodologia completa para a monitorização da evolução de *clusters*. Definimos exaustivamente as suas componentes, nomeadamente, a taxonomia das transições endógenas e exógenas, os dois métodos de monitorização adaptados a cada esquema de representação de *clusters* e os alicerces do algoritmo de detecção de transições. Posteriormente, procedemos à avaliação experimental da nossa abordagem e à realização de pequenos casos de estudo.

As experiências com os conjuntos de dados, quer artificiais, quer reais, permitiram concluir que a metodologia proposta é exequível e capaz de fornecer um bom diagnóstico dos vários tipos de transições que reflectem a evolução dos *clusters* ao longo do tempo. Com os casos de estudo provámos a vasta aplicação desta metodologia genérica a dados passíveis de serem organizados em *clusters*, e demonstrámos como a análise das transições pode ajudar a detectar eventos importantes do domínio de conhecimento e a inferir tendências evolutivas dos fenómenos.

No que concerne às diferenças de resultados retornados pelos diferentes métodos, concluiu-se que estas eram, sobretudo, despoletadas pelos valores escolhidos para os limiares do método baseado nas probabilidades condicionadas e comprovou-se que, a realização de pequenos ajustes nestes limiares conduzia, na maioria dos casos, à convergência dos resultados de ambos os métodos. Por outro lado, observou-se uma certa dificuldade do método para esquemas de representação em extensão, em detectar mortes e nascimentos de *clusters*. As análises permitiram concluir que esta situação era motivada pelo facto deste método ser orientado para a monitorização

dos mesmos objectos ao longo do tempo, o que impede a detecção de mortes e nascimentos *puros* (ie, correspondentes a ligações com $peso = 0$). Em relação ao método para *clusters* representados em compreensão, verificou-se que, o facto deste método assentar na aferição do grau de sobreposição dos *clusters* no espaço de atributos, gerava alguns problemas na detecção de transições como sejam as cisões e as fusões. Estes problemas eram provocados pela deslocação abrupta de alguns *clusters* sobreviventes no espaço de atributos, de um instante temporal para o outro, o que dificultava o mapeamento destes *clusters*.

Por outro lado, a comparação das transições sugeridas por diferentes algoritmos de *Clustering* permitiu concluir que, quando se impõe a mesma estrutura aos dados (ie, quando se assume o mesmo número de *clusters* para ambos os algoritmos), as transições detectadas são iguais, ou muito idênticas, o que revela características de robustez e resiliência do método MEC ao algoritmo de *Clustering* adoptado. Também se comprovou experimentalmente que os métodos MEC são estáveis para *Clusterings* diferentes, mas obtidos pelo mesmo algoritmo de agrupamento. Ou seja, o impacto de diferentes inicializações ou valores de parâmetros de algoritmos como o k-médias, é muito reduzido ou mesmo nulo nos resultados finais dos métodos de monitorização, o que torna a metodologia MEC independente do algoritmo de *Clustering* adoptado.

Por fim, com a análise de sensibilidade dos limiares do método baseado nas probabilidades condicionadas concluiu-se que, quando existem transições de índole permanente, o impacto da variação dos valores assumidos pelos limiares de **sobrevivência** e de **cisão** é nulo. Porém, quando as transições são relativas, uma pequena variação dos limiares causa um aumento ou diminuição do número de alguns tipos de transições exógenas, consoante se trate do limiar de **cisão** ou do limiar de **sobrevivência**.

6.2 Limitações e Trabalho Futuro

O trabalho reportado nesta dissertação, por se tratar de um tema emergente e pouco estudado pela comunidade científica, ainda pode ser bastante explorado. Várias melhorias podem ser introduzidas na metodologia proposta, de modo a torná-la mais abrangente. Por outro lado, novos métodos podem ser desenvolvidos para complementar os resultados da nossa abordagem, ou para colmatar algumas das respectivas limitações. Assim, no futuro tenciona-se estender o domínio de aplicabilidade da metodologia MEC a variáveis qualitativas, através da sua consideração no processo de *Clustering*. Pretende-se, igualmente, explorar outras medidas (e.g, índice de Rand) capazes de efectuar o mapeamento dos *clusters* representados em extensão, ou avaliar a semelhança entre *Clusterings*, e que não exijam o mesmo

número de observações na monitorização da evolução. Seria também interessante construir esquemas alternativos para a representação em compreensão de *clusters*, que não estejam baseados na assumpção de que os *clusters* são esféricos. Por outro lado, considera-se a possibilidade de conduzir experiências para um maior número de períodos temporais (e.g. dez anos), de modo a detectar os períodos de maior instabilidade e os períodos que não foram sujeitos a mudanças significativas. Este tipo de estudo permitiria aprofundar o conhecimento sobre o domínio em causa e traçar um perfil de evolução mais completo sobre o fenómeno. No que concerne a novos métodos, seria importante desenvolver um método de avaliação objectiva da qualidade dos resultados, em termos de transições, sugeridos por vários algoritmos de *Clustering*, de preferência pertencentes a classes diferentes. Este método teria como objectivo eliminar a eventual ambiguidade associada à utilização de diferentes algoritmos de *Clustering*. Neste contexto, poderiam também ser utilizados *Clustering ensembles* para geração do *input* dos métodos propostos, uma vez que obtêm partições dos dados de qualidade superior por meio da consideração das sub-estruturas comuns a todas as partições. Um objectivo ambicioso consistiria na extrapolação desta metodologia para um ambiente de fluxos contínuos de dados (*Data Streams*), em que seria possível monitorizar continuamente os dados, prever e acompanhar, em tempo real, a tendência de evolução dos *clusters*. No futuro, tencionamos também desenvolver uma metodologia semelhante para monitorizar a evolução de redes sociais.

Bibliografia

Aggarwal, C. C. (2003). A framework for change diagnosis of data streams. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, June 9-12, 2003, San Diego, California, USA*, pages 575–586. ACM.

Aggarwal, C. C. (2005). On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):587–600.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 11(1):5–33.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*, pages 49–60. ACM Press.

Ball, G. H. and Hall, D. J. (1965). Isodata: a novel method of data analysis and classification. Technical report, Standford University, CA, USA.

Barbará, D. (2002). Requirements for clustering data streams. *SIGKDD Explorations*, 3(2):23–27.

Barbará, D., Li, Y., and Couto, J. (2002). Coolcat: an entropy-based algorithm for categorical clustering. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 582–589. ACM.

Baron, S. and Spiliopoulou, M. (2001). Monitoring change in mining results. In *Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery, Munich, Germany, September 5-7, 2001*, volume 2114 of *Lecture Notes in Computer Science*, pages 51–60. Springer.

Baron, S. and Spiliopoulou, M. (2004). Monitoring the evolution of web usage patterns. In *Proceedings of the 1st European Web Mining Forum, Cavtat-Dubrovnik, Croatia, September 22, 2003*, volume 3209 of *Lecture Notes in Computer Science*, pages 181–200. Springer.

Baron, S., Spiliopoulou, M., and Günther, O. (2003). Efficient monitoring of patterns in data mining environments. In *Proceedings of the 7th East European Conference on Advances in Databases and Information Systems, Dresden, Germany, September 3-6, 2003*, volume 2798 of *Lecture Notes in Computer Science*, pages 253–265. Springer.

Bartolini, I., Ciaccia, P., Ntoutsi, I., Patella, M., and Theodoridis, Y. (2004). A unified and flexible framework for comparing simple and complex patterns. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, 2004*, volume 3202 of *Lecture Notes in Computer Science*, pages 496–499. Springer.

Bartolini, I., Ciaccia, P., Ntoutsi, I., Patella, M., and Theodoridis, Y. (2009). The panda framework for comparing patterns. *Data and Knowledge Engineering*, 68(2):244–260.

Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 15th Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-8, 2001*, pages 585–591. MIT Press.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Böttcher, M., Höppner, F., and Spiliopoulou, M. (2008). On exploiting the power of time in data mining. *SIGKDD Explorations*, 10(2):3–11.

Chakrabarti, D., Papadimitriou, S., Modha, D. S., and Faloutsos, C. (2004). Fully automatic cross-associations. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 79–88. ACM.

Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge,MA,USA.

Chawathe, S. S. and Garcia-Molina, H. (1997). Meaningful change detection in structured data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, May 13-15, 1997*, pages 26–37. ACM Press.

Chen, K. and Liu, L. (2005). The "best k"for entropy-based categorical data clustering. In *Proceedings of the 17th international conference on Scientific and statistical database management, Santa Barbara, CA, USA, 27-29 June, 2005*, pages 253–262. Lawrence Berkeley Laboratory.

Chen, K. and Liu, L. (2006). Detecting the change of clustering structure in categorical data streams. In *Proceedings of the 6th SIAM International Conference on Data Mining, Bethesda, MD, USA, April 20-22, 2006*. SIAM.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Dhillon, I. S., Mallela, S., and Modha, D. S. (2003). Information-theoretic coclustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 89–98. ACM.

Diday, E. and Simon, J. C. (1976). Clustering analysis. In Fu, K. S., editor, *Digital Pattern Recognition*, pages 47–49. Springer-Verlag, Secaucus, NJ.

Dubes, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics and Systems*, 3:32–57.

Dubes, R. C. (1987). How many clusters are best? - an experiment. *Pattern Recognition*, 20(6):645–663.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, USA.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics and Systems*, 3(3):32–57.

Duran, B. S. and Odell, P. L. (1974). *Cluster Analysis: a survey*. Springer-Verlag, New York, USA.

Elnekave, S., Last, M., and Maimon, O. (2007). Incremental clustering of mobile objects. In *ICDE Workshops*, pages 585–592.

Falkowski, T., Bartelheimer, J., and Spiliopoulou, M. (2006). Mining and visualizing the evolution of subgroups in social networks. In *Proceedings of the 2006 IEEE / WIC / ACM International Conference on Web Intelligence, Hong Kong, China, 18-22 December, 2006*, pages 52–58. IEEE Computer Society.

Fisher, L. and vanNess, J. (1971). Admissible clustering procedures. *Biometrika*, 58:91–104.

Forgy (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). A framework for measuring changes in data characteristics. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 126–137, New York, NY, USA. ACM.

Gowda, K. C. and Krishna, G. (1978). Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, pages 105–112.

Hagen, L. W. and Kahng, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11(9):1074–1085.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Cluster validity methods: Part i. *SIGMOD Record*, 31(2):40–45.

Hampel, F. (2002). Some thoughts about classification. In *Proceedings of the 8th Conference of the International Federation of Classification Societies, Cracow, Poland, July 16-19, 2002*, pages 1–19. Springer.

Handl, J., Knowles, J. D., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Journal of Bioinformatics*, 21(15):3201–3212.

Iam-on, N., Boongoen, T., and Garrett, S. (2008). Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. In *Proceedings of the 11th International Conference, Budapest, Hungary, October 13-16, 2008*, volume 5255 of *Lecture Notes in Computer Science*, pages 222–233. Springer.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Kalnis, P., Mamoulis, N., and Bakiras, S. (2005). On discovering moving clusters in spatio-temporal data. In *Proceedings of the 9th International Symposium on Advances in Spatial and Temporal Databases, Angra dos Reis, Brazil, August 22-24,2005*, volume 3633 of *Lecture Notes in Computer Science*, pages 364–381. Springer.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley.

Kaur, S., Bhatnagar, V., Mehta, S., and Kapoor, S. (2009). Concept drift in unlabeled data stream. Technical report, University of Delhi.

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, pages 86–101.

Kleinberg, J. M. (2003). An impossibility theorem for clustering. In *Proceedings of the 2002 conference on Advances in Neural Information Processing Systems,Vancouver, British Columbia, Canada, December 9-14, 2002*, pages 446–453. MIT Press.

Li, T., Ma, S., and Ogihara, M. (2004a). Entropy-based criterion in categorical clustering. In *Proceedings of the 21th international conference on Machine learning, Banff, Alberta, Canada*, pages 68–75. ACM.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23th International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 577–584. ACM.

Li, Y., Han, J., and Yang, J. (2004b). Clustering moving objects. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 617–622. ACM.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–136.

Lu, Y.-H. and Huang, Y. (2005). Mining data streams using clustering. In *Proceedings of the 4th International Conference on Machine Learning and Cybernetics, Guangzhou, China, August 18-21, 2005*, pages 2079–2083. IEEE Computer Society.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability, Statistical Laboratory of the University of California, Berkeley, USA*, volume 1, pages 281–297. University of California Press.

Mao, J. and Jain, A. K. (1994). A self-organizing network for hyperellipsoidal clustering (hec). In *Proceedings of the 1994 IEEE World Congress on Computational Intelligence, 27 June - 2 July, 1994*, volume 5, pages 2967–2972. IEEE.

Meila, M. and Shi, J. (2001). A random walks view of spectral segmentation. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics, Hyatt Hotel, Key West, Florida, January 4-7, 2001*.

Michalski, R. S., Stepp, R. E., and Diday, E. (1981). A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts. In Kanal, L. and Rosenfeld, A., editors, *Progress in Pattern Recognition*, volume 1. North-Holland Publishing Co., Amsterdam, The Netherlands.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, pages 354–359.

Nanni, M. and Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3):267–289.

O'Callaghan, L., Meyerson, A., Motwani, R., Mishra, N., and Guha, S. (2002). Streaming-data algorithms for high-quality clustering. In *Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, 26 February - 1 March 2002*, page 685. IEEE Computer Society.

Pelleg, D. and Moore, A. (1999). Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, United States*, pages 277–281. ACM.

Rennie, J. (2005). Volume of the n-sphere. `http://people.csail.mit.edu/jrennie/writing/sphereVolume.pdf`.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.

Sander, J., Ester, M., Kriegel, H.-P., and Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 22(8):888–905.

Sneath, P. H. (1957). The applications of computers to taxonomy. *Journal of General Microbiology*, 17:201–226.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., and Schult, R. (2006). Monic: modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 706–711. ACM.

Spinosa, E. J., de Leon Ferreira de Carvalho, A. C. P., and Gama, J. (2007). Olindda: a cluster-based approach for detecting novelty and concept drift in data streams. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC), Seoul, Korea, March 11-15, 2007*, pages 448–452. ACM.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. In *Workshop on Knowledge Discovery and Data Mining*.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bulletin of the Polish Academy of Sciences*, pages 801–804.

Urga, G. (1992). The econometrics of panel data: A selective introduction. Economics Series Working Papers 99151, University of Oxford, Department of Economics.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistics Association*, 58(301):236–244.

Yang, H., Parthasarathy, S., and Mehta, S. (2005). A generalized framework for mining spatio-temporal patterns in scientific data. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*, pages 716–721. ACM.

Yip, A. M., Ding, C. H. Q., and Chan, T. F. (2005). Dynamic cluster formation using level set methods. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Hanoi, Vietnam, May 18-20, 2005*, volume 3518 of *Lecture Notes in Computer Science*, pages 388–398. Springer.

Zhang, Q. and Couloigner, I. (2005). A new and efficient k-medoid algorithm for spatial clustering. In *Proceedings of the 2005 International Conference on Computational Science and Its Applications,International Conference, Singapore, May 9-12, 2005*, volume 3482 of *Lecture Notes in Computer Science*, pages 181–189. Springer.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 103–114. ACM Press.

Anexo A

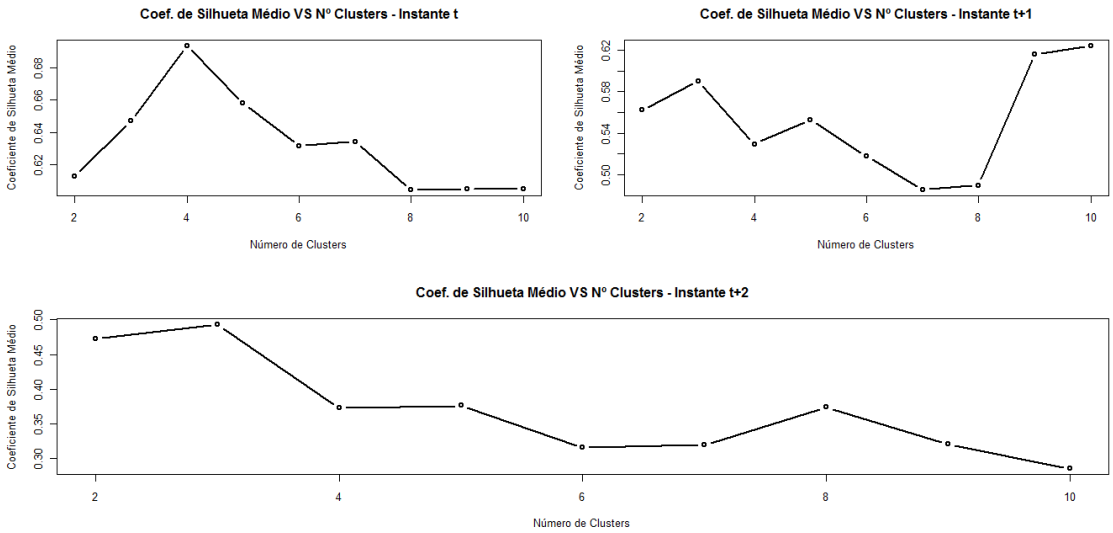


Figura A.1: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo particional das k-médias e para diferentes instantes temporais

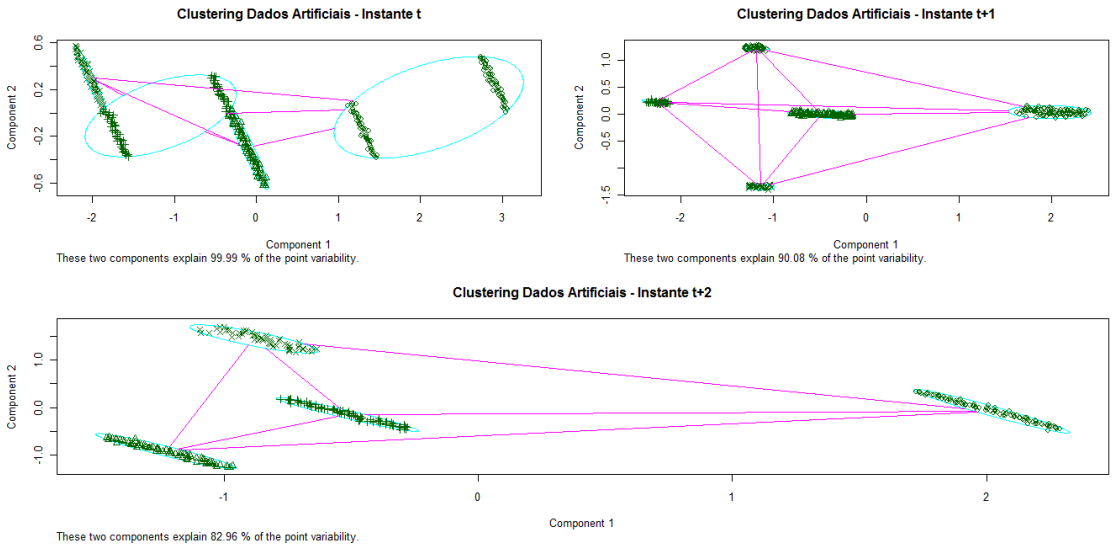


Figura A.2: Representação gráfica dos *clusters* obtidos com recurso ao algoritmo particional das k-médias no espaço formado pela projecção dos dados nas duas componentes principais, nos instantes de tempo t, $t+1$ e $t+2$

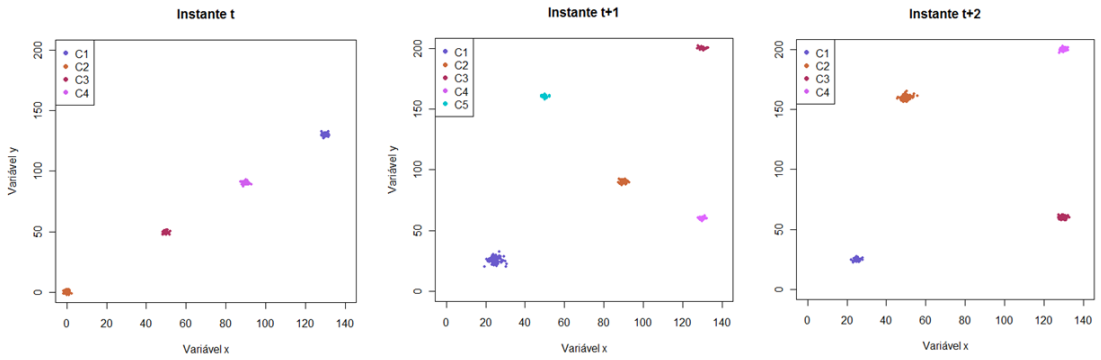


Figura A.3: *Clusters* obtidos pelo algoritmo particional das k-médias, para três instantes temporais distintos, e com a partição dos dados sugerida pela análise do coeficiente de silhueta médio

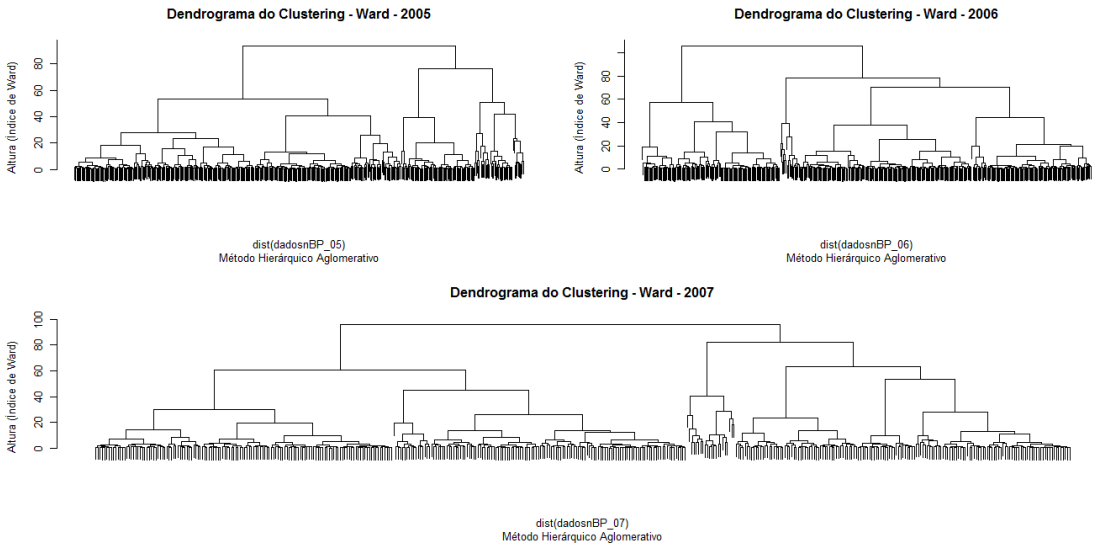


Figura A.4: Dendrogramas resultantes da aplicação do algoritmo hierárquico aglomerativo (índice de Ward) aos dados da Central de Balanços do Banco de Portugal, para os anos de 2005, 2006 e 2007

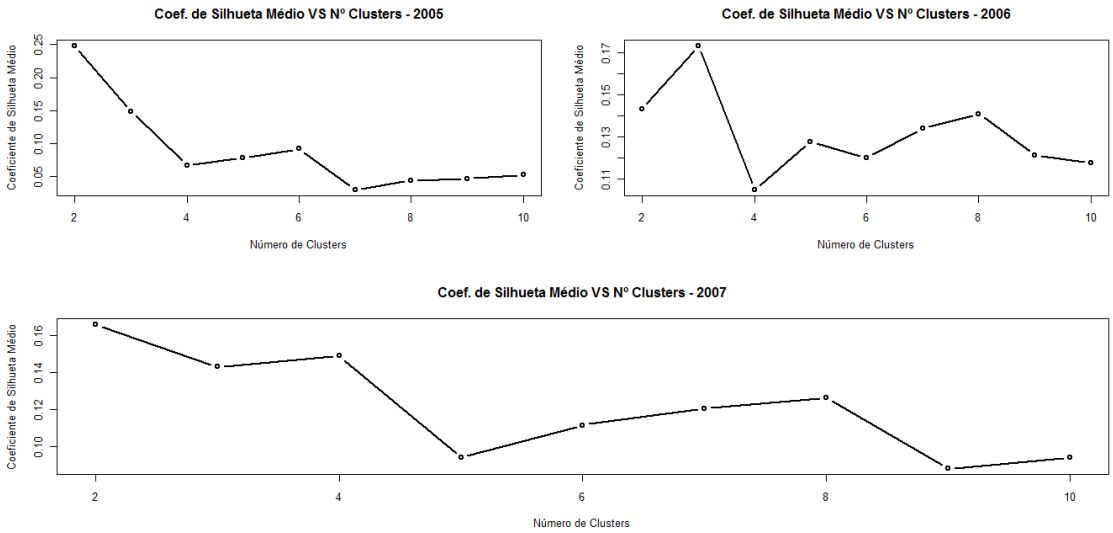


Figura A.5: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo hierárquico aglomerativo (índice de Ward). Cada gráfico corresponde a um determinado ano do conjunto de dados da Central de Balanços do Banco de Portugal: 2005, 2006 e 2007, respectivamente.

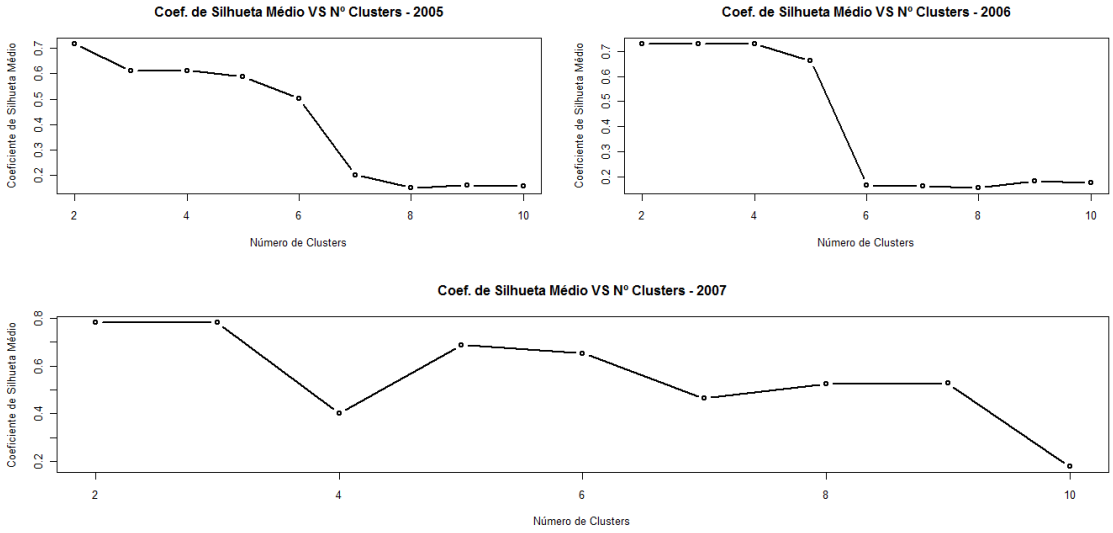


Figura A.6: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo particional das k-médias. Cada gráfico corresponde a um determinado ano do conjunto de dados da Central de Balanços do Banco de Portugal: 2005, 2006 e 2007, respectivamente.

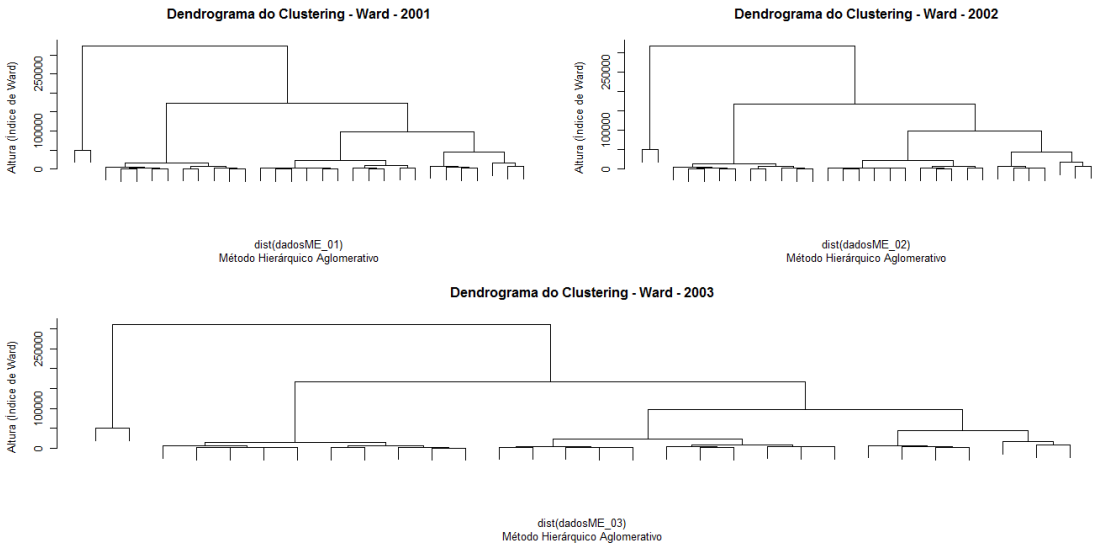


Figura A.7: Dendrogramas resultantes da aplicação do algoritmo hierárquico aglomerativo (índice de Ward) aos dados do INE referentes ao número de estudantes matriculados no ensino não-superior, para diferentes anos - 2001, 2002 e 2003

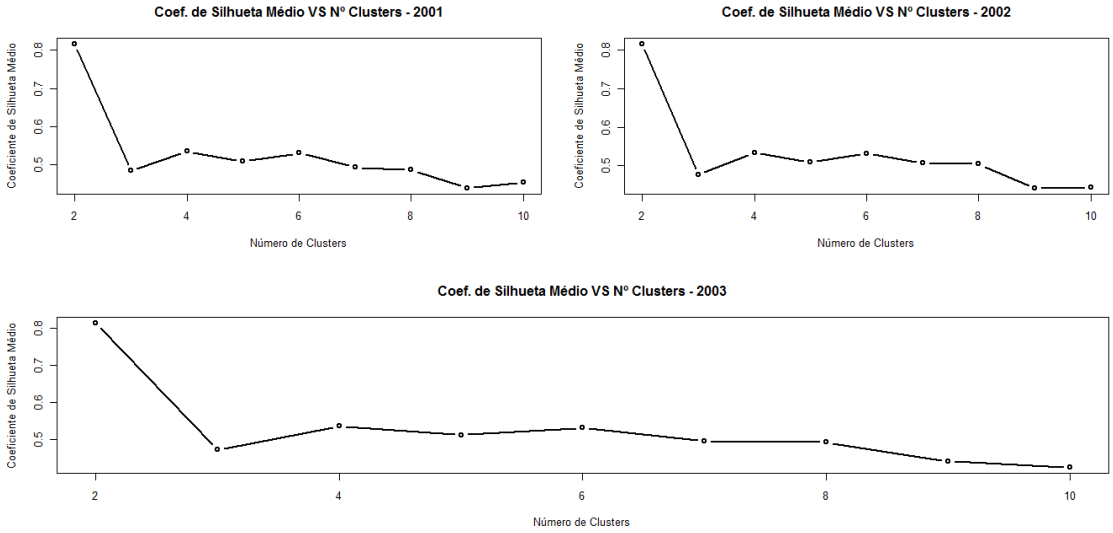


Figura A.8: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo hierárquico aglomerativo (índice de Ward). Cada gráfico corresponde a um determinado ano do conjunto de dados do INE (Educação): 2001, 2002 e 2003, respectivamente.

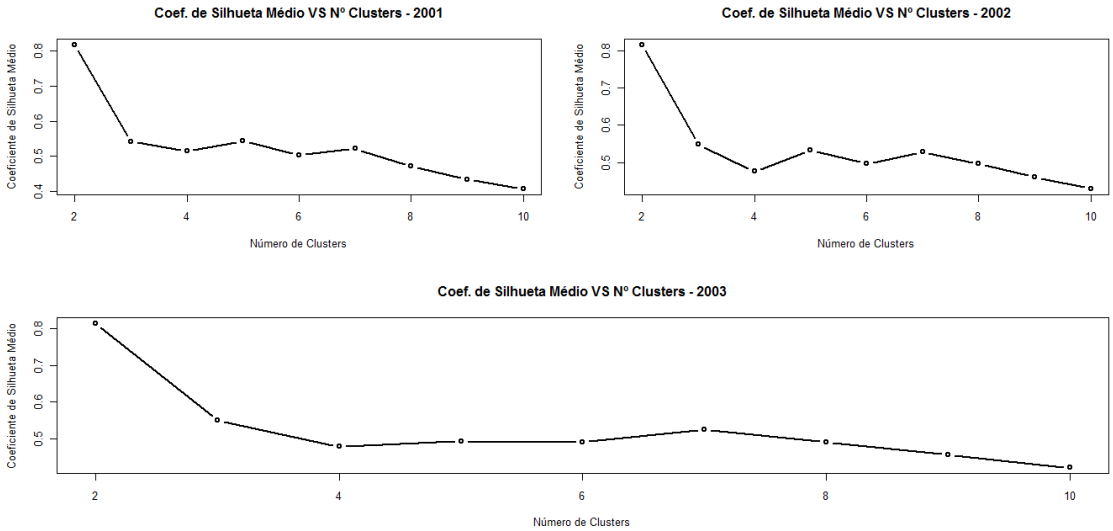


Figura A.9: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo particional das k-médias. Cada gráfico corresponde a um determinado ano do conjunto de dados do INE (Educação): 2001, 2002 e 2003.

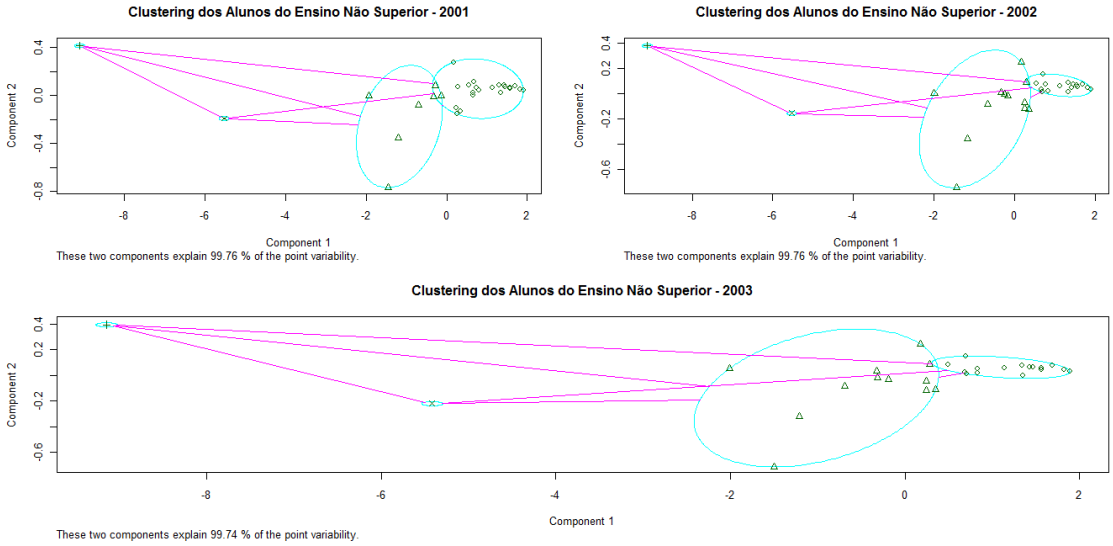


Figura A.10: Representação gráfica dos *clusters*, no espaço formado pela projecção dos dados do INE (Educação) nas duas componentes principais, no triénio 2001, 2002 e 2003

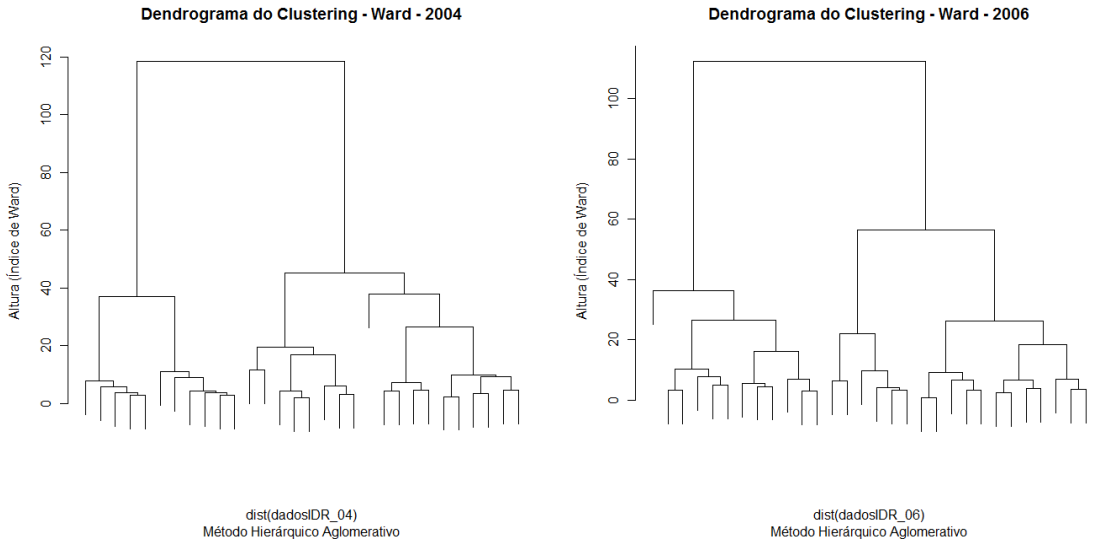


Figura A.11: Dendrogramas resultantes da aplicação do algoritmo hierárquico aglomerativo (índice de Ward) aos dados do INE referentes ao índice de desenvolvimento regional, para os anos de 2004 e 2006

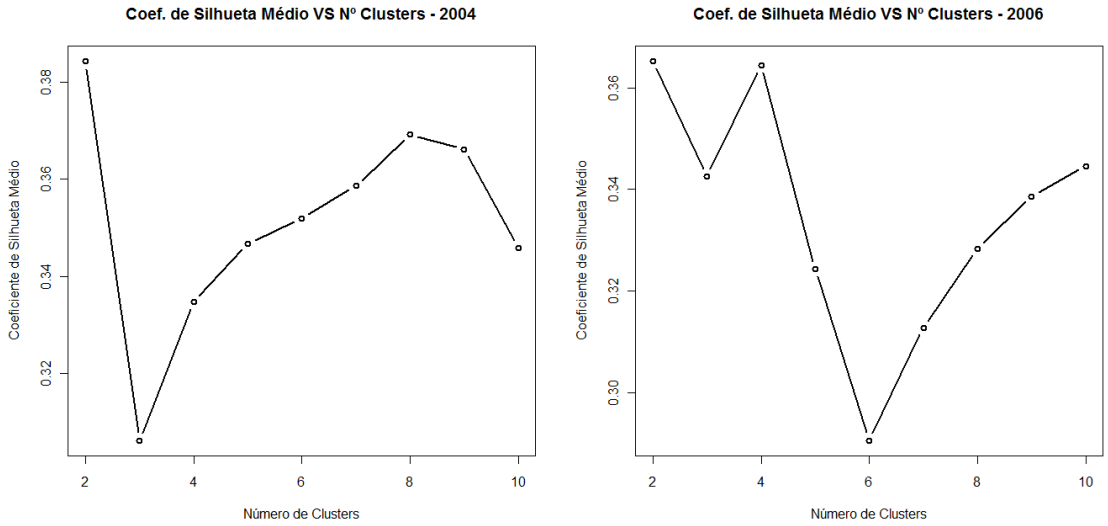


Figura A.12: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo hierárquico aglomerativo (índice de Ward). Cada gráfico corresponde a um determinado ano do conjunto de dados do INE (Território): 2004 e 2006, respectivamente.

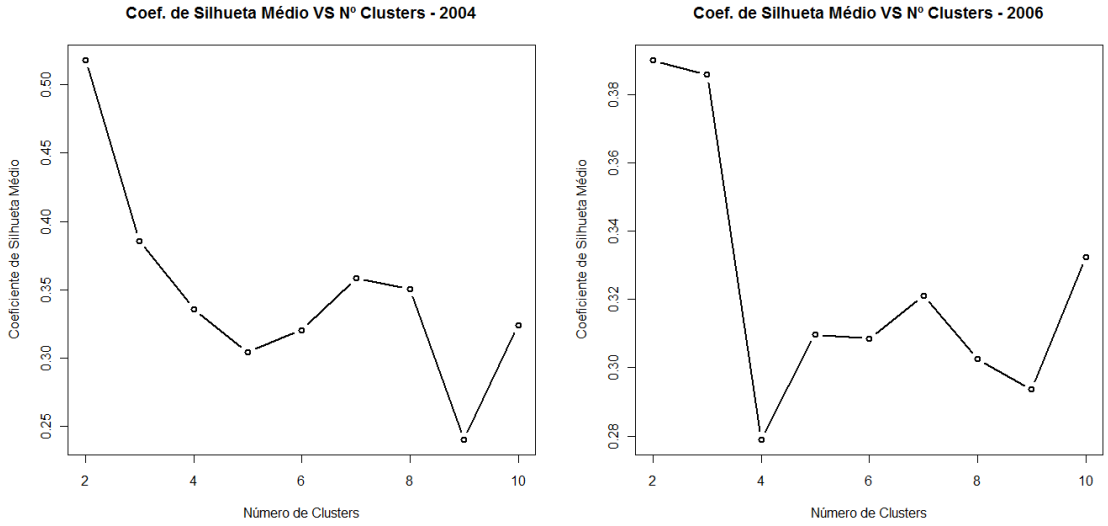


Figura A.13: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo particional das k-médias. Cada gráfico corresponde a um determinado ano do conjunto de dados do INE (Território): 2004 e 2006, respectivamente.

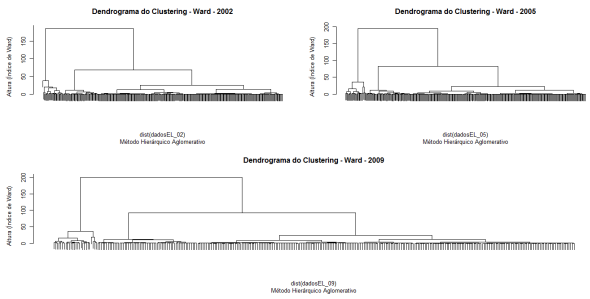


Figura A.14: Dendrogramas resultantes da aplicação do algoritmo hierárquico aglomerativo (índice de Ward) aos dados do DGAI referentes aos resultados das eleições legislativas, para os anos de 2002, 2005 e 2009

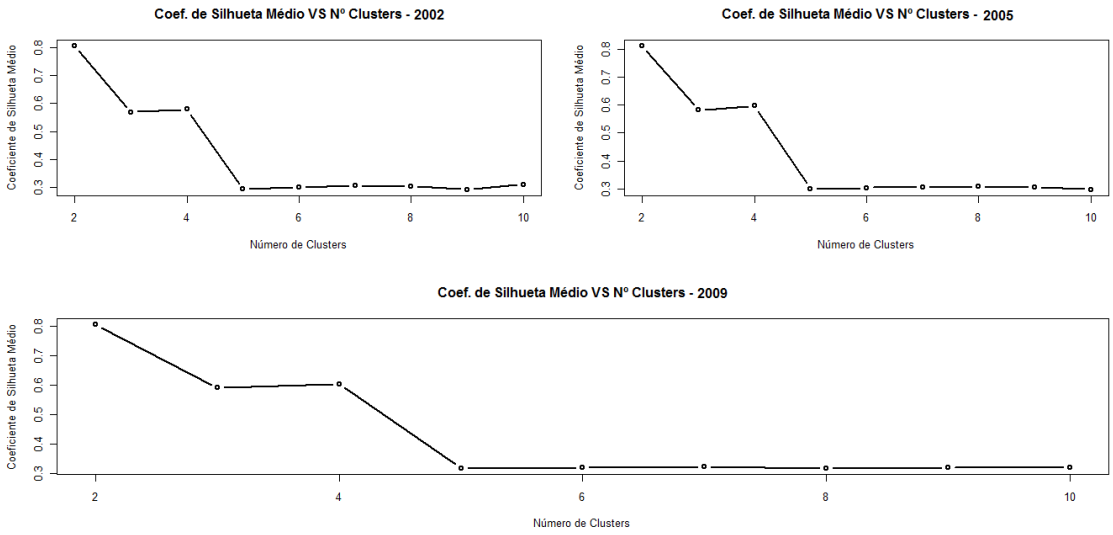


Figura A.15: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo hierárquico aglomerativo (índice de Ward). Cada gráfico corresponde a um determinado ano do conjunto de dados do DGAI: 2002, 2005 e 2009, respectivamente.

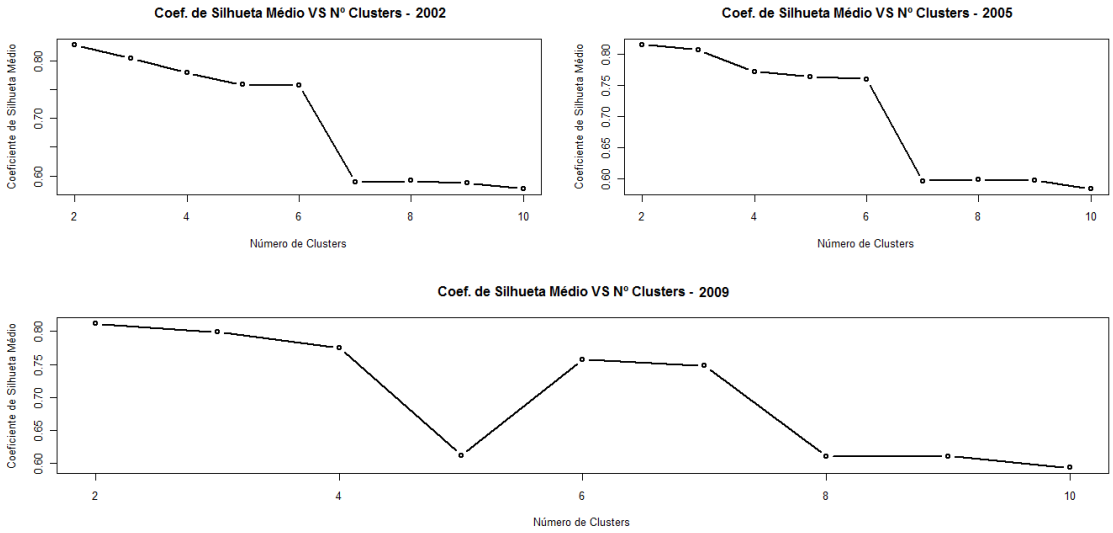


Figura A.16: Valores do coeficiente de silhueta médio para $k \in [2, 10]$, em agrupamentos gerados com recurso ao algoritmo particional das k-médias. Cada gráfico corresponde a um determinado ano do conjunto de dados do DGAI: 2002, 2005 e 2009, respectivamente.

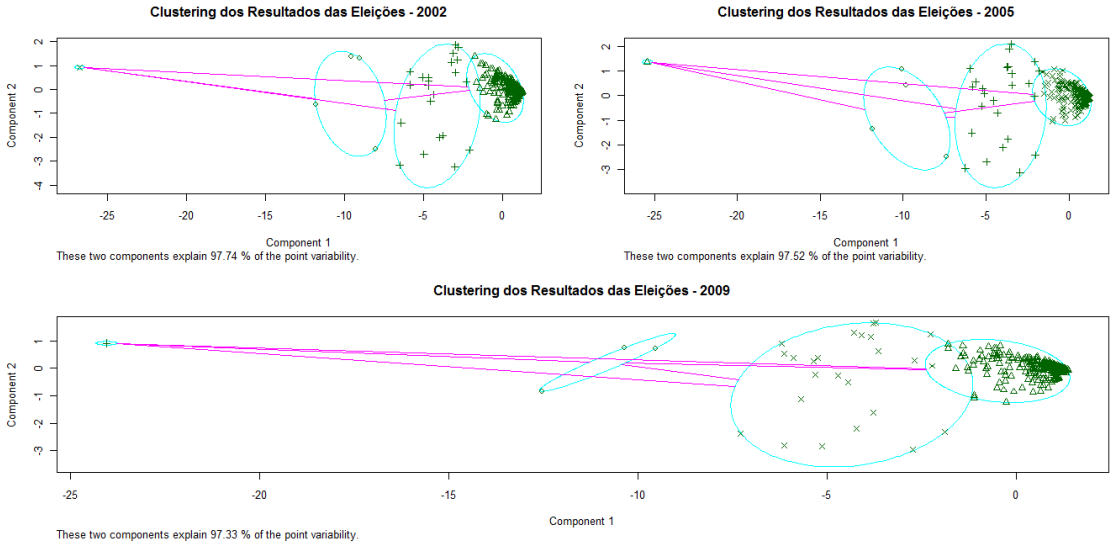


Figura A.17: Representação gráfica dos *clusters*, no espaço formado pela projecção dos dados do DGAI nas duas componentes principais, no triénio 2002, 2005 e 2009

Anexo B

Variáveis de Identificação:

- **Ano** - instante temporal a que se referem os dados de um determinado sector
- **CAE** - código de 5 dígitos que permite identificar, de forma unívoca, cada sector de actividade

Variáveis de Descrição e Caracterização:

- **Número de Empresas** - indicador da representatividade de um determinado sector; número de empresas de um determinado sector de actividade que, num dado ano, preencheu os Inquéritos do Banco de Portugal; variável expressa em valor absoluto.
- **Resultado Líquido do Exercício** - valor líquido de impostos, positivo (lucro) ou negativo (prejuízo), gerado pelas empresas pertencentes a um dado sector de actividade no decurso do respectivo exercício económico; esta variável encontra-se expressa em euros.
- **Taxa de Investimento** - rácio obtido através da divisão do total de investimentos realizados em Imobilizações, pelo total de rendimentos das empresas afectas a um dado sector; variável expressa em percentagem.
- **Rendibilidade do Capital Próprio** - rácio que se obtém através da divisão do Resultado Líquido do Exercício pelo Capital Próprio das empresas de um determinado sector, reflectindo a capacidade desse sector em gerar resultados a partir dos Capitais Próprios investidos nas empresas; variável expressa em percentagem.
- **Rotação do Activo Líquido** - quociente entre as Vendas e Prestações de Serviços de um dado sector e o respectivo Activo Líquido; esta variável mede o grau de eficácia na utilização dos Activos e está expressa em "número de vezes".

- **Taxa de Valor Acrescentado** - rácio que relaciona o Valor Acrescentado Bruto do sector com os respectivos Proveitos de Exploração; variável expressa em percentagem.
- **Taxa de Endividamento** - rácio que resulta da divisão dos Capitais Alheios pelos Recursos Próprios das empresas afectas a um determinado sector, reflectindo o grau de dependência de um sector, face a capitais alheios, para fazer face aos seus compromissos; variável expressa em percentagem.
- **Produtividade do Equipamento** - quociente entre o Valor Acrescentado Bruto do sector e as respectivas Imobilizações Corpóreas.
- **Produtividade do Trabalho** - quociente entre o Valor Acrescentado Bruto do sector e o respectivo Volume de Emprego.
- **Emprego** - número médio de pessoas ao serviço de todas as empresas que integram um determinado sector de actividade, num dado período de tempo; variável expressa em valor absoluto.