Tiago Dias Silva Leão
Licenciado em Engenharia Informática

Caracterização Espacial utilizando Indução Orientada aos Atributos no SOLAP+

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática

Orientador: Doutor João Moura Pires,
Prof. Auxiliar, Faculdade de Ciências e Tecnologia

Júri:

Presidente:	Prof. Doutor Henrique João Lopes Domingos
Arguente:	Prof. Doutora Maribel Yasmina Santos
Vogal:	Prof. Doutor João Carlos Gomes Moura Pires

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Novembro, 2011

Agradecimentos

Os meus agradecimentos vão para todos aqueles que de alguma forma, directa ou indirectamente, tornaram a realização desta dissertação possível.

Em primeiro lugar, gostaria de agradecer a toda a minha família, com destaque merecido para a minha mãe, Alice Fernanda Dias Silva, que sempre me apoiou e me possibilitou chegar onde hoje estou. Sem ela não estaria aqui. Não esquecer também os meus avós António Santos Silva e Palmira Silva que me ajudaram a criar e a transmitir muitos dos valores que tenho hoje.

Um agradecimento também ao meu irmão e cunhada que por esta altura esperam pela vinda do novo membro da família, a minha sobrinha. Para eles o meu obrigado.

Para o Professor Doutor João Moura Pires por ter aceite o meu pedido de realização desta dissertação continuando a sua linha de trabalho e pela sua orientação, que através de todas as reuniões, discussões e críticas realizadas, me conduziram sempre a uma melhor compreensão dos factos, bem como a um melhoramento nas ideias que iam surgindo.

Para o Ricardo Silva, que me ajudou muito no processo inicial da dissertação com toda a sua paciência e disponibilidade.

Não poderia esquecer a minha outra "família", um grupo de 3 amigos muito especiais: Carlos D'Almeida, André Ribeiro e João Silva. Obrigado não só pelo apoio mas por toda a diversão que muitas vezes foi necessária para abstrair das horas passadas a realizar esta dissertação. A todos os outros meus amigos também um grande e sincero obrigado.

Para finalizar, gostava de agradecer a uma pessoa que infelizmente já não está entre nós mas que foi um pilar na minha vida, o meu pai, José Manuel Fangueiro Leão. Hoje, sei que estou a concluir um sonho, não só meu mas também dele. Onde quer que estejas, isto também é para ti. Obrigado.

Resumo

O protótipo SOLAP+, desenvolvido sobre a orientação de João Moura Pires, segue o conceito SOLAP apresentado por *Bédard et al.* combinando as funcionalidades dos sistemas de apoio à decisão OLAP com as capacidades dos sistemas de informação geográfica (SIG).

Com os trabalhos realizados por Rosa Martins (2006), Marlene Vitorino e Rodolfo Caldeira (2008), Ruben Jorge (2009) e Ricardo Silva (2010) foi desenvolvida uma abordagem genérica SOLAP que culminou com a implementação de um protótipo que, presentemente, já apresenta bastantes funcionalidades do ponto de vista da análise de dados e representação dos mesmos através de mapas temáticos e/ou gráficos.

Para ajudar nessa tarefa de análise, integramos mecanismos de descoberta de relações entre os dados e apresentamos no mapa essas relações. Esta forma de análise permite ao utilizador retirar conclusões através da extracção de informações que não estão explícitas nos dados presentes na base de dados.

Assim, o objectivo desta dissertação consiste em incorporar, respeitando o modelo multidimensional seguido, duas técnicas de indução: generalização de dominância espacial e generalização de dominância não espacial. Com os resultados da aplicação dessas técnicas, é feita uma caracterização semântica dos objectos espaciais no mapa.

Palavras-chave: SOLAP; *data mining*; indução; generalização; hierarquias; caracterização espacial.

Abstract

The prototype SOLAP+, developed on the direction of João Moura Pires, follows the concept SOLAP presented by *Bédard et al.* combining the functions of support systems for decision OLAP with the capacities of geographical information systems (GIS).

With the work undertaken by Rose Martins (2006), Marlene Vitorino and Rodolfo Caldeira (2008), Ruben Jorge (2009) and Ricardo Silva (2010) was developed a generic approach SOLAP that culminated with the implementation of a prototype that, currently, already presents a few features from the point of view of analysis of data and representation of that data using thematic maps and/or graphics.

To assist in this task of analysis, we integrate mechanisms to discover relationships between the data and present these relations on the map. This analysis will enable the user to take conclusions by means of information that is not explicit in the data present in the database.

Thus, the aim of this dissertation is to incorporate, with the multidimensional model followed, two induction techniques: spatial dominant generalization and non-spatial dominant generalization. With the results of applying these techniques, we semantically characterize the spatial objects in the map.

Keywords: SOLAP; data mining; induction; generalization; hierarchies; spatial characterization.

Índice

Índice de Figuras

1. Introdução

Este capítulo apresenta o contexto em que está inserida esta dissertação, incluindo uma breve descrição sobre os sistemas OLAP e SOLAP e sobre o que é o *data mining*. Posteriormente serão descritos os objectivos e contribuições desta dissertação e no final uma descrição da forma como o documento está estruturado.

1.1. Contexto

1.1.1. Sistemas SOLAP

Um *Data Warehouse* (DW) é um repositório de dados referentes ao histórico da actividade de uma organização provenientes, quer de fontes internas quer externas, que tem como objectivo dar uma visão mais compacta e interpretável da informação [1].

O conteúdo de um DW pode ser analisado por várias ferramentas. Uma delas são os sistemas OLAP (*On-Line Analytical Processing*) que surgiram para possibilitar uma forma simples e sobretudo rápida de exploração e análise dos dados através de operações de selecção e agregação/desagregação dos dados. A operação de *drill-down* consiste em aumentar o nível de detalhe dos dados e a de *drill-up* na agregação dos dados, sumarizando-os.

Os sistemas OLAP assentam num modelo multidimensional. Este é um modelo conceptual e não necessariamente físico, que descreve os dados através de um cubo cujas células correspondem a eventos ocorridos, onde estão presentes as métricas da actividade da organização. A informação neste cubo é indexada por dimensões, formadas por um conjunto de atributos. Estas dimensões fazem a categorização da informação em entidades consideradas relevantes para a organização detentora do modelo.

Existem três grandes tipos de sistemas OLAP: *Multidimensional OLAP* (MOLAP) que consiste numa estrutura específica para descrever cubos multidimensionais, *Relational OLAP* (ROLAP) que assenta sobre tecnologia relacional SGDB, utilizando o SQL *standard* ou extensões SQL com a componente OLAP, e *Hybrid OLAP* (HOLAP) que se refere a tecnologias que combinam MOLAP e ROLAP. Existem diferentes estruturas de organização do modelo multidimensional, como por exemplo, o *star schema* e o *snowflake schema*, ilustrados nas Figura 1.1 e Figura 1.2, respectivamente. Enquanto o *star schema* tem apenas ligações entre a tabela de factos e um conjunto de dimensões, no *snowflake schema* as dimensões estão associadas a uma série de tabelas normalizadas resultantes da expansão das hierarquias presentes na respectiva dimensão.

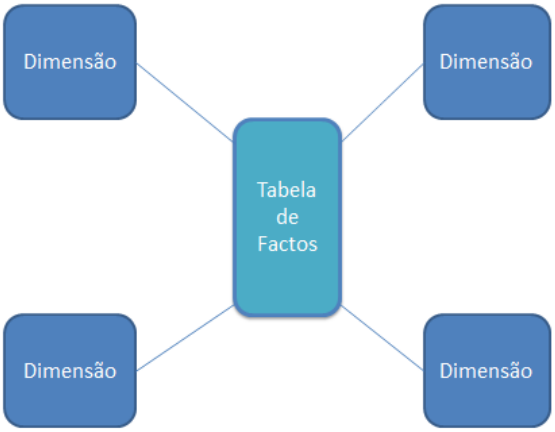


Figura 1.1 - *Star schema.*

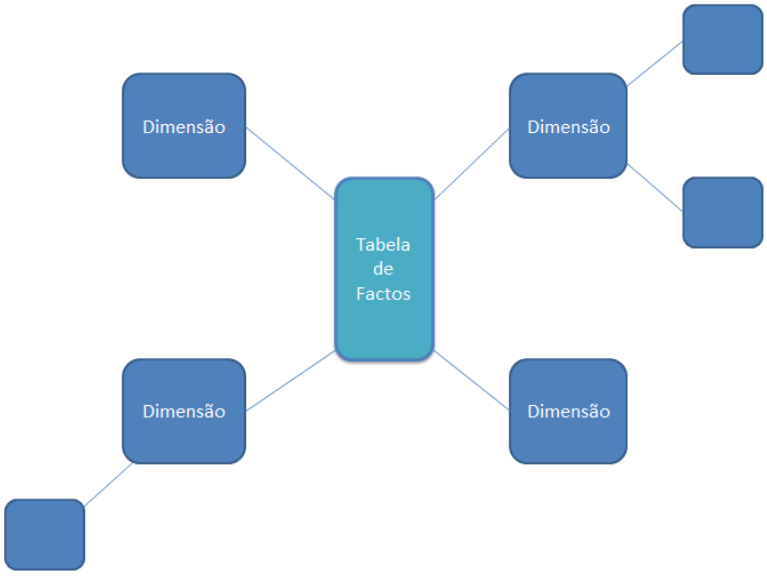


Figura 1.2 - *Snowflake schema.*

Se combinarmos os sistemas OLAP com as capacidades disponibilizadas pelos sistemas de informação geográfica (SIG), os quais gerem a informação espacial e possibilitam a representação do espaço e dos fenómenos que nele acontecem, obtemos os sistemas *Spatial On-Line Analytical Processing* (SOLAP).

Em [2] *Bédard et al.* definiram uma ferramenta SOLAP como "*a visual platform built especially to support rapid and easy spatio-temporal analysis and exploration of data following a multidimensional approach comprised of aggregation levels available in cartographic displays as well as in tabular and diagram displays*", ou seja, com os sistemas SOLAP acrescentamos aos sistemas OLAP a capacidade de visualização e análise dos dados espaciais.

Neste momento já existe um protótipo designado por SOLAP+ desenvolvido segundo a orientação de Moura Pires. Este projecto começou em 2006 com o trabalho de Rosa Matias [3], sendo continuado por Marlene Vitorino e Rodolfo Caldeira [4] em 2008 e por Ruben Jorge [5] em 2009. O último a trabalhar neste projecto foi Ricardo Silva em 2010 culminando num protótipo que apresenta já muitas das características desejadas em sistemas SOLAP: apresenta um modelo de integração genérico que inclui informação espacial nas dimensões, onde dados espaciais e não espaciais são utilizados em concordância; verifica a sincronização entre a visualização tabular e o mapa; permite filtragem pelos atributos das dimensões; utilização de gráficos para efeitos de análise; permite também o trabalho com informação espacial em mais do que uma dimensão, suportando diversos casos de interacção que são a base em qualquer análise.

1.1.2. Data mining

Para dar resposta à grande quantidade de informação gerada e guardada em bases de dados operacionais e científicas, surgiram os processos de *Knowledge Discovery from Databases* (KDD). KDD é definido em [6] como "*the discovery of interesting implicit, and preciously unknown knowledge from large databases*", ou seja, tem como objectivo extrair informação que não está explicita nos dados e encontrar relações e propriedades entre eles. Uma das formas de extrair essas informações é através de técnicas de *data mining* [6]. Se, para os dados analisados, for considerada a componente geográfica dos mesmos, então entramos num contexto de *data mining* espacial.

As técnicas de *data mining* podem ser classificadas consoante o tipo de informação que se pretende extrair dos dados analisados. Em seguida, e com base em [7], apresentamos alguns dos principais objectivos das técnicas de *data mining*:

- Sumarização – O principal objectivo é encontrar formas de descrever os dados de uma maneira mais geral. Por exemplo, tendo dados referentes a temperaturas durante um certo período de tempo, generaliza-se o atributo temperatura para conceitos de quente, frio, ou moderado.
- Agrupamento – Realiza agrupamentos de dados de acordo com as suas propriedades (espaciais e não espaciais). Um exemplo da utilidade desta técnica é realizar análises de dados referentes a preços das casas, agrupando pontos referentes a casas luxuosas. Esse agrupamento pode ser feito segundo diferentes factores, como por exemplo, o tipo de casa e/ou a sua proximidade. Assim, os objectos que foram considerados agrupáveis formam um conjunto, dando origem a um novo objecto representativo desse grupo.
- Identificação de classes – Tenta encontrar descrições para os dados de forma a realizar um bom particionamento da informação em diferentes classes. Um exemplo da aplicação desta técnica será realizar uma classificação das áreas de um país baseando-se no seu poder económico. Ao contrário do que acontece nas técnicas de agrupamento, aqui os objectos sobre os quais estamos a trabalhar permanecem os mesmos, sendo classificados de determinada forma.

- Associação – Tenta extrair regras de associação que reflictam dependências entre os dados. Desta forma, podemos tentar inferir a influência da existência de um determinado tipo de objecto espacial (lago, montanha, etc.) na zona que se caracteriza por ter um determinado valor para um atributo (elevada taxa de reformados, por exemplo).
- Tendências e desvios – Identificação de padrões que revelem tendências para os dados apresentarem determinadas características e, dessa forma, se identificarem desvios à normalidade. Um exemplo da obtenção de uma tendência num conjunto de dados analisados seria de que à medida que nos afastamos de determinado ponto verifica-se um decréscimo do salário médio da população.

Em virtude do rápido desenvolvimento nas áreas de *data mining* e de *data warehousing* nos últimos anos, têm surgido alguns sistemas com vista à exploração destas duas áreas. Um exemplo de sistemas que surgiram nesta área é o *DBMiner* [8]. O *DBMiner* é um sistema que têm vindo a ser desenvolvido para desempenhar funções de *data mining* sobre grandes bases de dados. Contudo, rapidamente surgiu a necessidade de incorporar a componente espacial surgindo o *GeoMiner* [8].

A incorporação de formas de *data mining* num protótipo como o SOLAP+ traria um maior número de formas de análise dos dados possibilitando a descoberta de informação que não está explicita nos dados e não conseguiria ser obtida através de operações OLAP. Uma vez que estamos perante um ambiente de análise de dados espaciais como é o SOLAP, a visualização dos dados obtidos sobre a forma de mapas temáticos pode ser relevante para o utilizador, na medida em que facilita a interpretação e consequente extracção de conhecimento útil.

Das técnicas de *data mining* analisadas, a escolha recaiu sobre as actividades de sumarização. Esta escolha deve-se ao facto desta ser uma actividade que lida com um dos principais problemas dos analistas e do ser humano em geral, que é a fraca capacidade de analisar grandes quantidades de informação e extrair informação relevante a partir de uma enorme quantidade de dados. Para aplicar estas técnicas de sumarização, utilizou-se uma técnica denominada de indução orientada aos atributos que consiste em subir nas cadeias conceptuais dos atributos. Desta forma, generaliza os valores destes atributos para valores mais gerais e tenta retirar conclusões sobre essa informação mais generalizada. Uma vez que se pretende utilizar dados geográficos na informação a generalizar, um dos tipos de conclusões que se pode retirar é ao nível das características mais evidenciadas em determinados locais. Para realizar estas tarefas de caracterização espacial através da indução orientada aos atributos, existe dois métodos: generalização de dominância espacial e generalização de dominância não espacial.

1.2. Objectivos e contribuições

O objectivo desta dissertação é estender o protótipo SOLAP+ [9], possibilitando um outro nível de análise dos dados por parte dos utilizadores. Para isso, o propósito desta dissertação é a caracterização espacial através da utilização de técnicas de indução orientada aos atributos, mais concretamente a generalização de dominância espacial e não espacial. A introdução deste tipo de técnicas trará uma mais-valia ao protótipo, possibilitando a extracção de informações que não se encontram explicitamente na base de dados mas que se podem encontrar de forma implícita no conjunto de dados guardados, representando essas conclusões através da criação de mapas temáticos caracterizadores dos diferentes objectos espaciais.

Assim, para dar solução ao objectivo definido, propomos:

- Inclusão de tarefas de sumarização dos dados através da generalização dos mesmos, utilizando processos de indução com foco tanto nos atributos espaciais como não espaciais. Para complementar, utilizamos técnicas para construção automática de hierarquias para os atributos numéricos;
- Utilização dos resultados da sumarização para realizar uma caracterização do espaço. Esta caracterização poderá ser baseada em dois factores: número de ocorrências de determinada característica, ou valor da métrica, no caso de esta existir.

Qualquer uma destas acções que se pretendem efectuar, devem ter em conta a interacção com o utilizador e a sua performance. Assim, deve ser possível o utilizador ter algum controlo sobre o processo e suas actividades.

1.3. Estrutura da dissertação

A estrutura deste documento, referente à dissertação, apresenta a seguinte organização e conteúdo. No capítulo 2 (Trabalho Relacionado), é apresentado o estado da arte, onde é realizada a análise de alguns sistemas SOLAP existentes e referindo determinados conceitos relevantes, nomeadamente técnicas de sumarização da informação em bases de dados, construção/alteração de hierarquias de conceitos e formas de agrupamento de pontos de polígonos. No capítulo 3, apresenta-se a extensão ao protótipo SOLAP+ existente através da incorporação do processo de sumarização com a finalidade de realizar caracterização espacial. Em seguida, no capítulo Arquitectura, é feita referência à arquitectura do sistema, aos diferentes componentes existentes e à forma de comunicação entre eles. O capítulo 5 (Implementação) descreve a implementação dos novos componentes adicionados e as alterações aos já existentes. No capítulo 6, através de casos de estudo, são validadas as propostas e exemplificadas formas de interacção com o sistema. Por último, apresenta-se um capítulo de conclusões e de trabalho que poderá ser realizado futuramente, o qual corresponde ao capítulo 7 deste documento.

2. Trabalho Relacionado

Existem várias tarefas de *data mining* e para cada uma dessas tarefas existem várias técnicas que podem ser aplicadas para extracção de informação. Tendo em conta o contexto em que o protótipo SOLAP+ está inserido, foram analisados vários tipos de tarefas de *data mining* recaindo a escolha sobre as tarefas de sumarização com recurso à técnica de indução orientada aos atributos. Uma vez que a incorporação iria ser realizada com recurso ao modelo multidimensional sobre o qual se baseia o SOLAP+, e um dos conceitos chave nos processos de sumarização são as hierarquias, a existência de hierarquias formadas pelos diferentes níveis das dimensões foi um dos factores a favor da escolha deste tipo de técnica. Por outro lado, uma vez que no protótipo SOLAP+ são criados mapas temáticos que, através do recurso ao elemento cor, permitem ao utilizador visualizar as características de cada objecto espacial representado, a utilização dos resultados provenientes da sumarização para realizar uma caracterização semântica incorpora-se perfeitamente no protótipo.

Quando comparado com a indução realizada sobre tuplos (tipo de indução que analisa cada um dos tuplos da relação sem processar qualquer sumarização), a indução orientada aos atributos apresenta um melhor desempenho uma vez que explora primeiro a generalização dos valores dos atributos e só depois olha para os diferentes tuplos formados. Esta análise dos tuplos feita a níveis conceptuais mais elevados torna mais fácil a extracção de regras caracterizadoras da informação, pois o espaço de tuplos da *prime relation* é menor e encontram-se mais valores em comum devido à generalização dos conceitos que foi realizada [10].

Por outro lado, este processo permite a paralelização da indução ao nível das várias partições de dados. Possibilita ainda, ser executada apenas sobre uma amostra dos dados da relação inicial, o que permite retirar conclusões usando apenas uma pequena quantidade da informação disponível. Esta técnica tem também a seu favor o facto de lidar bem com o ruído. Para isso, incorpora informação estatística através do cálculo das ocorrências de tuplos duplicados, sendo assim fácil a detecção de casos excepcionais que têm um impacto mínimo no resultado final.

Assim sendo, a inclusão do processo de indução orientada aos atributos permitirá fazer uma caracterização semântica dos diferentes objectos espaciais (pontos e polígonos), dando hipótese ao utilizador de reflectir o seu conhecimento através das hierarquias de conceitos, tendo uma forma de controlo sobre a profundidade da análise a ser realizada, tal como se pretende num ambiente de SOLAP+.

Neste capítulo começa-se por fazer uma análise a vários sistemas SOLAP existentes e ao protótipo SOLAP+. Em seguida, é feita uma descrição do processo de indução orientada aos atributos e dos conceitos envolventes. No seguimento do processo de indução são apresentadas técnicas de suporte utilizando hierarquias. Por último, é feita uma apresentação de formas de realizar agrupamento espacial, tanto sobre pontos como sobre polígonos.

2.1. Sistemas SOLAP

Nos últimos anos têm surgido algumas ferramentas que seguem os conceitos SOLAP, cada uma com características diferenciadas. Em [11] é feito um resumo de algumas das ferramentas SOLAP desenvolvidas e é realizada uma comparação entre elas de acordo com as suas capacidades de visualização. Na Figura 2.1 está representada a tabela retirada de [11] onde são apresentados vários sistemas SOLAP. O objectivo desta tabela não é apenas dar a conhecer algumas das ferramentas SOLAP disponíveis mas também apresentar as suas capacidades de geovisualização e a forma como é apresentada a informação (2D ou 3D) e, dessa forma, chegar a conclusões do que existe e do que se deve aproveitar de cada um deles.

Authors	Tool name	Platform	Geobrowser	Geovisualization techniques
Han et al. (1997)	GeoMiner	Desktop	2D	Simple
Shekhar et al. (2000)	Map Cube	Desktop	2D	Simple, choropleth and multimap
Stolte et al. (2002)	Rivet	Desktop	2D	Simple and choropleth
Fidalgo et al. (2004)	GMLA WS Client	Web	2D	Simple
Scotch and Parmanto (2005)	SOVAT	Desktop	2D	Simple and choropleth
Rivest et al. (2005)	JMap® Spatial OLAP Extension	Desktop	2D	Simple, choropleth, proportional symbol, multimap
Colonese et al. (2005)	PostGeoOlap	Desktop	2D	Simple
Bimonte et al. (2007)	GeWOlap	Web	2D	Simple, choropleth, bar, pie
Compieta et al. (2007)	-	Desktop	3D	Simple, remote sensing and complex vectors
Silva et al. (2008)	WebGeoOlap	Web	2D	Simple and remote sensing
Di Martino et al. (2009)	GooLAP	Web	3D	Simple, remote sensing, 3D-bar, pie
Pentaho	Pentaho Google Maps Dashboard	Web	2D	Simple, colored point
ESRI	OLAP for ArcGIS	Desktop	2D	Simple, bars
JRubik	JRubik	Web	2D	Simple, bars
SpatialAnalytics	SOLAPLayers	Web	2D	Simple, choropleth
	GlobeOLAP	Web	3D	Simple, remote sensing, choropleth, prism, bar and custom proportional symbols

Figura 2.1 - Tabela comparativa de várias ferramentas SOLAP.

A partir da análise da Figura 2.1, percebe-se que a principal característica deste tipo de sistemas é a sua simplicidade de representação dos dados de forma a facilitar a interpretação e análise da informação. Por outro lado, verifica-se que em muitos dos casos recorre-se à utilização de mapas temáticos, onde os objectos espaciais representam locais de onde existem dados registado e utiliza-se a cor, brilho e tamanho dos objectos para representar visualmente os valores que estão associados a cada um dos objectos. Assim sendo, isto é um indicativo das características principais que o protótipo como o SOLAP+ deve evidenciar.

Em seguida, é feita uma apresentação de dois destes sistemas, descrevendo para cada um deles as suas características, para desta forma apresentar uma visão global do estado deste tipo de ferramentas. Escolhemos para tal o SOVAT e o JMap®. O primeiro por se tratar de uma ferramenta criada para um domínio específico e o JMap® por ser genérico e ter sido desenvolvido no seguimento das investigações de dois elementos presentes na literatura, Rivest e Bédard. No final é discutido o estado actual do protótipo SOLAP+.

2.1.1. SOVAT

A aplicação *Spatial OLAP Visualization and Analysis Tool* (SOVAT) [12] foi desenvolvida para ser aplicada junto da comunidade da área da saúde e integra num só sistema várias características necessárias para suportar a tomada de decisões desta comunidade.

Na Figura 2.2 é apresentada a interface do sistema SOVAT que prima por quatro zonas distintas: um painel para construir a interrogação, uma secção para executar interrogações espaciais, uma zona para visualização de gráficos e outra para o mapa.

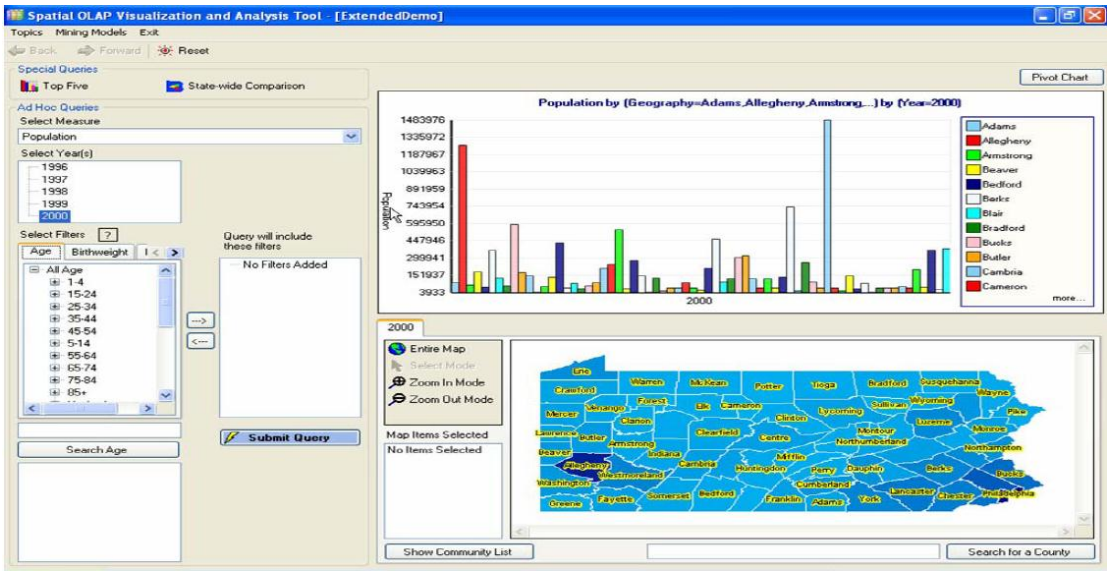


Figura 2.2 - Interface da aplicação SOVAT, retirada de [13].

Para realizar uma interrogação, primeiro é escolhida uma métrica através de uma *combo box* onde se encontram as métricas existentes. Em baixo dessa *combo box*, é definida a dimensão temporal que pretendemos analisar (no caso do exemplo apresentado trata-se do ano). As diferentes dimensões estão divididas por *tabs* e, através da escolha de elementos dessas dimensões, é possível criar filtros para a interrogação a ser executada. Esta aplicação apresenta ainda mecanismo de *drill-up* e *drill-down* sobre os atributos espaciais. O SOVAT não suporta o carregamento de outros conjuntos de dados pois, como já foi referido, trata-se de uma aplicação para um domínio específico.

2.1.2. JMap®

O projecto JMap apareceu em 2005 através do grupo de investigação de Bédard, passando depois a ser desenvolvida pela empresa *KHEOPS Technologies*, que passou a designar-se por *K2 Geospatial* [14] desde Outubro de 2009.

Na Figura 2.3 está esquematizada a arquitectura da aplicação. Suporta vários tipos de clientes, todos eles comunicando por Web Services com um servidor de dados espaciais. Como não foi construída para nenhum cenário aplicacional pré-definido (ao contrário do que acontecia com o sistema SOVAT, apresentado na secção 2.1.1), apresenta um módulo de administração com várias ferramentas que possibilitam a criação de bases de dados relacionais ou geo-espaciais e a configuração dos servidores. Esta aplicação oferece ainda extensões do tipo *"plug-and-play"* providenciando barras de ferramentas e funções especializadas ligadas tanto ao cliente como ao servidor.

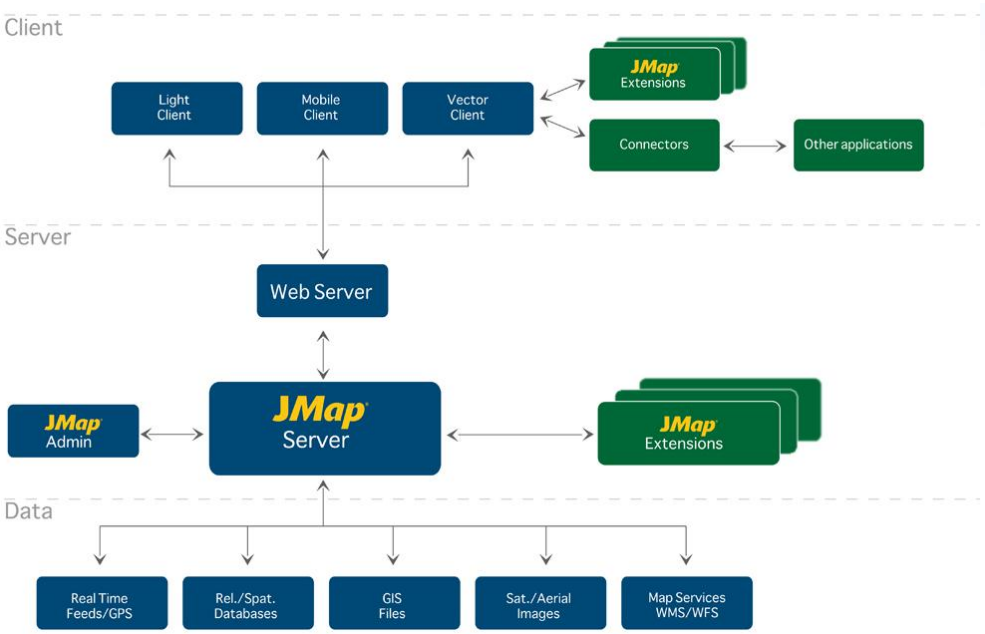


Figura 2.3 - Arquitectura da aplicação JMap.

Em seguida, na Figura 2.4, apresenta-se a interface JMap® onde se vê que a aplicação em causa apresenta uma secção de descrição do modelo, outra para visualização dos dados e, por fim, uma secção com as opções de visualização no topo (mapa, tabela ou gráfico).

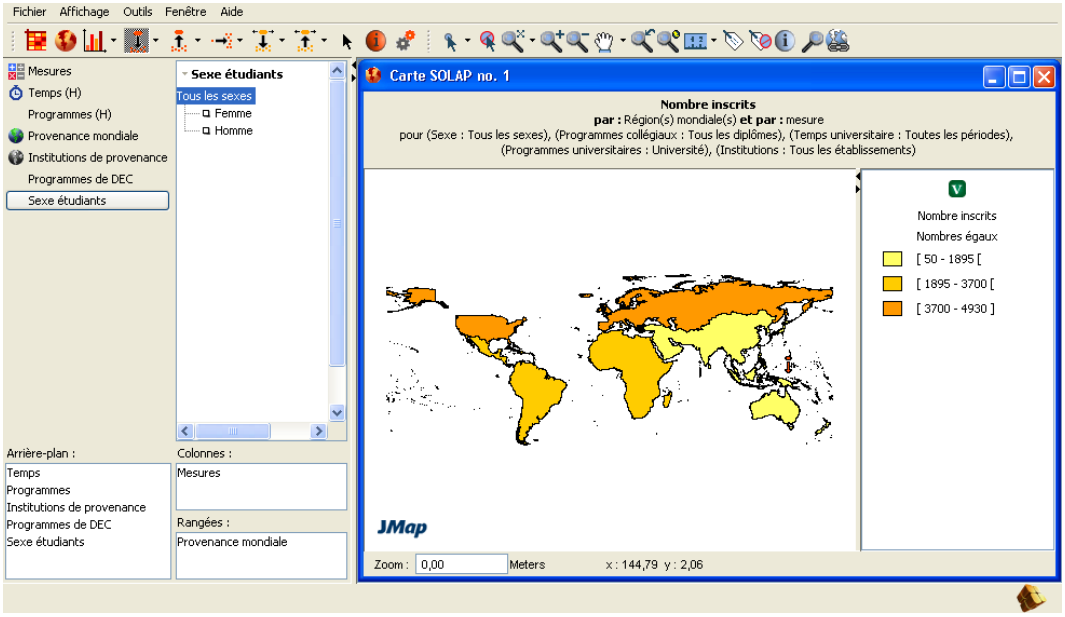


Figura 2.4 - Interface JMap Spatial OLAP.

2.1.3. SOLAP+

O SOLAP+ é um protótipo desenvolvido no Departamento de Informática da FCT-UNL, segundo a orientação de Moura Pires, J. [5] [9].

Este protótipo assenta sobre a ideia de um modelo de interacção genérico. A sua interface é apresentada na Figura 2.5. Esta interface está dividida em 5 áreas: *(i)* mapa, *(ii)* tabela de suporte, *(iii)* tabela de detalhe, *(iv)* descrição do modelo multidimensional e *(v)* área de controlo de aspecto do mapa e painéis de *slices*.

Um aspecto fundamental do SOLAP+ é que impõe e garante uma relação de 1:1 entre as linhas da tabela de suporte (Figura 2.5 *ii*) e os objectos espaciais que são representados no mapa (Figura 2.5 *i*). Não se verificando esta propriedade, ou seja, havendo múltiplas linhas da tabela de suporte associadas a um único objecto no mapa, este poderia torna-se extremamente confuso de se analisar. Por outro lado, havendo vários objectos do mapa associados a uma só linha da tabela de suporte, seria difícil diferenciar o contributo de cada um desses objectos para a informação contida na linha referenciada. Já a tabela de detalhe (Figura 2.5 *iii*) é utilizada para disponibilizar informação sobre os dados a um nível de granularidade inferior aos encontrados na tabela de suporte, apresentando assim uma relação de 1:N.

Na componente de descrição do modelo multidimensional (Figura 2.5 *iv*) são apresentadas as diferentes dimensões organizadas por níveis, bem como as métricas numéricas com os respectivos operadores de agregação possíveis de serem utilizados. Nesta componente existe ainda uma secção referente às *layers* que possibilitam a realização de *slices* espaciais.

Na secção *v* da Figura 2.5, encontram-se os controlos sobre o mapa, através dos quais é possível definir algumas das suas propriedades, como por exemplo, a capacidade de *zoom* e legenda tanto relativa à informação contextual como aos objectos espaciais. Apresenta também uma secção de *slices* que permite *slices* semânticos, espaciais e definir filtros para as métricas.

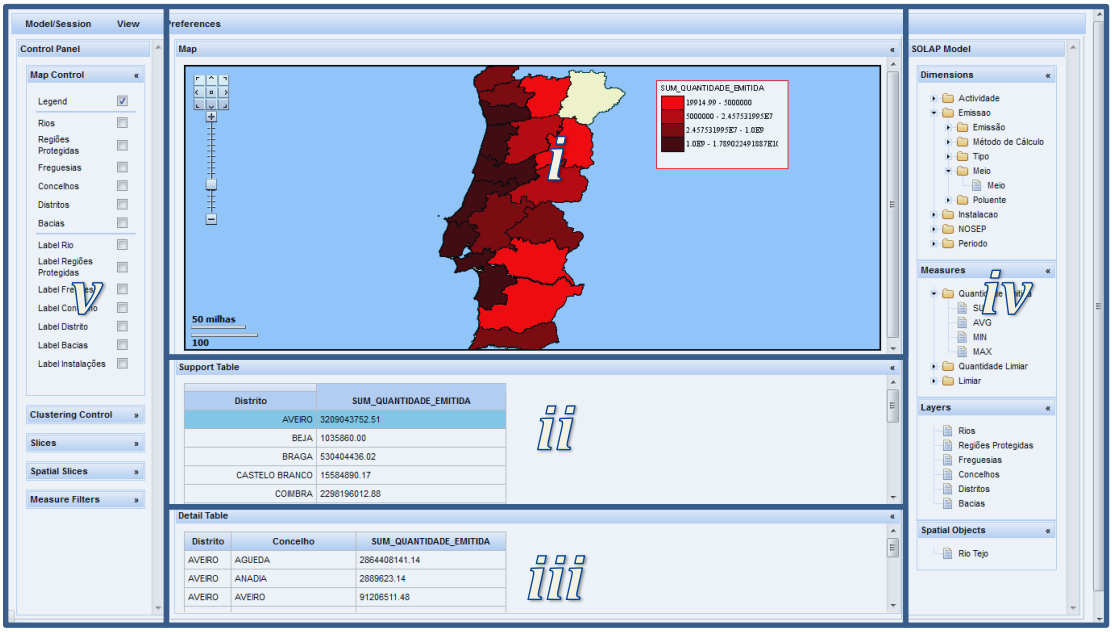


Figura 2.5 - Interface do modelo genérico SOLAP+.

Actualmente, no SOLAP+ já é feita uma utilização de algoritmos de *data mining*, mais concretamente algoritmos de agrupamento espacial. Contudo, a sua utilização tem apenas como propósito o aperfeiçoamento visual da apresentação dos dados no mapa e não fins analíticos. No que diz respeito à representação da informação no mapa, o protótipo contempla já a utilização de gráficos (barras ou circulares).

No que diz respeito à arquitectura utilizada neste protótipo, essa é descrita no capítulo 4 (Arquitectura) deste documento.

2.2. Indução orientada aos atributos

A indução orientada aos atributos é uma técnica de descoberta de informação em bases de dados que trabalha sobre uma relação inicial que é formada por um conjunto de dados relevantes. Esta técnica generaliza a informação contida nessa relação inicial atributo a atributo. Posteriormente, vai comprimir essa informação numa relação generalizada, removendo tuplos duplicados, fazendo a contagem dessas ocorrências, e por fim, extrai regras da informação generalizada. Desta forma, permite descobrir relações interessantes a um nível conceptual mais elevado.

De forma a ilustrar o processo que envolve esta técnica de indução apresenta-se o seguinte exemplo. Imagine-se que temos as hierarquias demonstradas na Figura 2.6, e define-se que o *attribute threshold* para os atributos X e Y é de 2.

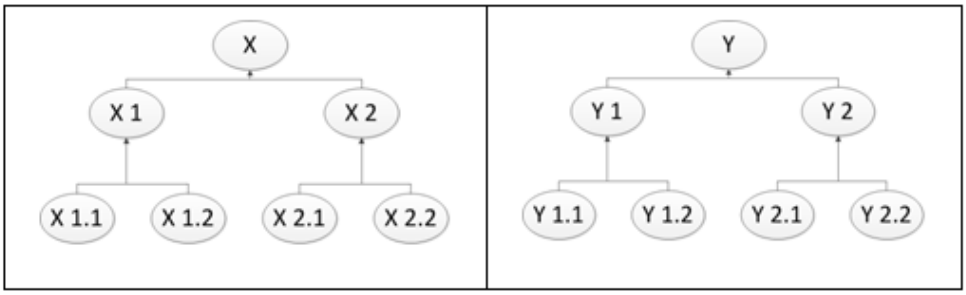


Figura 2.6 - Hierarquias de conceitos referentes a X e Y.

A relação inicial *R*, extraída da base de dados, corresponde à informação apresentada na primeira tabela do diagrama descrito na Figura 2.7. Numa fase posterior é feita a generalização dos conceitos presentes em *R*, baseando-se nas hierarquias de conceitos fornecidas (Figura 2.6). Na passagem para a fase seguinte é feita a remoção de duplicados e feita a contagem das ocorrências, para possibilitar a extracção de regras.

Attr. X	Attr. Y
X 2.1	Y 1.1
X 1.2	Y 1.1
X 1.1	Y 1.2
X 1.1	Y 1.2
X 1.2	Y 1.2
X 2.1	Y 2.2
X 2.1	Y 1.2
X 1.1	Y 1.2
X 2.2	Y 2.2
X 1.2	Y 1.2
X 2.1	Y 2.1
X 2.1	Y 1.1
X 1.1	Y 1.2
X 2.2	Y 1.1

Attr. X	Attr. Y
X 2	Y 1
X 1	Y 1
X 1	Y 1
X 1	Y 1
X 1	Y 1
X 2	Y 2
X 2	Y 1
X 1	Y 1
X 2	Y 2
X 1	Y 1
X 2	Y 2
X 2	Y 1
X 1	Y 1
X 2	Y 1

Attr. X	Attr. Y	Count
X 1	Y 1	7
X 2	Y 1	4
X 2	Y 2	3

Figura 2.7 - Processo de indução.

As regras extraídas podem ser de caracterização, discriminantes, de agrupamento e de associação [10]. As de caracterização pretendem indicar as propriedades que mais sobressaem. Por exemplo, indicar que uma determinada região se caracteriza pela venda de produtos do tipo dietético. As regras do tipo discriminante têm como objectivo evidenciar características que estão presentes numa classe e não estão presentes noutra, isto é, procura por conceitos que as distingam. Um exemplo será a distinção de uma doença *A* de uma doença *B*, através dos sintomas que os pacientes que sofrem de *A* evidenciam e os de *B* não.

A extracção de regras de agrupamento é fundamental para a criação de grupos com características idênticas, como por exemplo, criar um grupo formado por alunos que se localizam "próximos" uns dos outros e que têm algo em comum, como por exemplo, as notas superiores a 18. Neste caso, está a ser feito não só um agrupamento espacial, devido à proximidade das ocorrências analisadas, mas também semântico, uma vez que estamos a agrupar somente objectos que têm em comum uma determinada característica, formando um único objecto com essa característica mas com uma representação espacial diferente. Por último, uma regra de associação representa uma relação de associação entre um conjunto de valores na base de dados. Uma regra de associação descoberta ao analisar as transacções de uma loja, pode ser, por exemplo, o facto de 95% dos clientes que compram leite da marca *X* também compram café da marca *Y*.

2.2.1. Hierarquia de conceitos

A hierarquia de conceitos organiza a informação e os conceitos de uma forma hierárquica ou parcialmente ordenada, ajudando dessa forma a expressar conhecimento e relações entre os dados de uma forma concisa [15]. Mais concretamente, a hierarquia de conceitos faz um conjunto de associações entre os termos conceptuais mais baixos, isto é, mais detalhados, e os seus correspondentes de mais alto nível.

A hierarquia está parcialmente ordenada do mais geral para o mais específico. O nó que representa o conceito mais geral é descrito pela palavra-chave *"ANY"* e os nós da hierarquia que representam conceitos mais específicos correspondem aos valores dos atributos na base de dados.

Na Figura 2.8 é apresentada uma hierarquia de conceitos referente às áreas administrativas do *Canadá*. Assim, nas folhas da árvore (nós que não têm descendentes) temos as zonas mais pequenas em termos de área administrativa, e à medida que se sobe na hierarquia, apresentam-se nós que correspondem a zonas maiores que englobam as zonas dos seus descendentes.

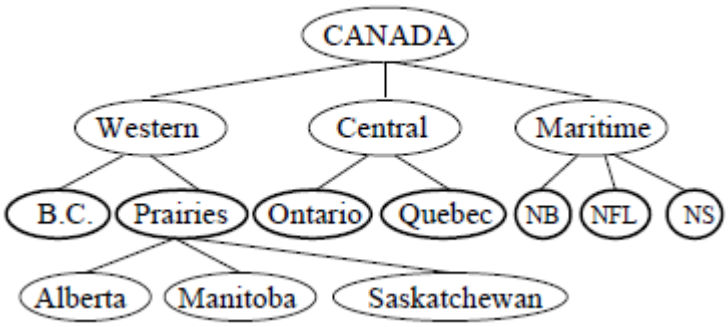


Figura 2.8 - Hierarquia de conceitos, retirada de [10].

Estas hierarquias podem ser obtidas de diversas maneiras. Em alguns casos, a própria hierarquia está definida implicitamente na base de dados. Nestes casos, basta serem indicadas as regras de mapeamento entre os atributos que formam a hierarquia. Noutros casos, as hierarquias são definidas por especialistas no domínio em que se está a trabalhar. Em casos mais excepcionais, hierarquias podem mesmo ser definidas e deduzidas computacionalmente. Isto acontece, por exemplo, quando se tratam de hierarquias para descrição de atributos numéricos.

Sem estas hierarquias, não haveria forma de identificar quais os conceitos que englobavam os valores referidos nas relações extraídas da base de dados, quais os conceitos que englobavam esses outros conceitos, e assim sucessivamente, não sendo por isso possível realizar a generalização dos dados. Desta forma, as hierarquias de conceitos desempenham um papel fundamental durante o processo de indução orientada aos atributos.

2.2.2. Extracção de regras de caracterização

Como já foi referido, um dos tipos de regras que podem ser extraídos utilizando a indução orientada aos atributos são regras de caracterização. Estas são regras que fazem sobressair e identificam factos (características) existentes nos dados das relações analisadas.

Antes de descrever a forma como essas características podem ser obtidas [10], é necessário ter presentes alguns conceitos importantes:

- *Attribute threshold* – Este é o indicador do nível de generalização que se pretende atingir em relação a um atributo. No final do processo de indução não deverá haver mais valores distintos de um atributo do que o seu *attribute threshold*. No preciso momento em que um atributo passa a respeitar este limite, está no seu nível mínimo desejado.
- *Prime relation* – Uma relação é uma *prime relation* se todos os seus atributos estão no nível mínimo desejado.

Definidos estes conceitos, o processo de indução orientado aos atributos para extracção de regras de caracterização [10] [16] segue os seguintes passos:

1) Obtenção da relação inicial *R*. Nesta relação devem apenas constar atributos que tenham uma hierarquia de conceitos associada, ou cuja hierarquia possa ser gerada (processo descrito na secção 2.3.1). Na relação *R* deve ainda ser evitado colocar atributos diferentes associados à mesma hierarquia, onde a diferença reside apenas no nível conceptual em que se encontram os valores desses atributos. Por exemplo, devemos evitar ter na relação inicial um atributo relativo ao conselho e outro relativo ao distrito, uma vez que estes são conceitos pertencentes à mesma hierarquia mas que se encontram a níveis diferentes. Estas regras devem ser seguidas de forma a evitar a remoção desses atributos na fase seguinte deste processo.

2) Execução do processo de generalização para cada um dos atributos presentes em *R*. Este processo envolve determinar os conceitos que tornam o número de valores distintos do atributo em análise dentro do limite definido pelo seu *attribute threshold*, e em seguida, fazer a associação desses conceitos com os dados da relação *R*.
3) Por último, realiza-se a substituição dos conceitos do atributo na relação *R* baseados nas associações definidas no passo 2. É feita ainda a remoção de tuplos repetidos que surjam e faz-se a contagem das ocorrências desses tuplos. Este processo culmina com a criação de uma *prime relation*.

Seguindo a análise de complexidade do processo feita em [10], verificamos que o segundo passo obriga a uma pesquisa completa de todos os *n* tuplos em *R* para que os seus atributos possam ser generalizados. No terceiro e último passo é feita uma ordenação dos tuplos e os tuplos duplicados são removidos levando a uma complexidade final de $O(n \log n)$. Quanto à complexidade espacial é $O(n)$, uma vez que todos os *n* tuplos da relação inicial estão em memória.

Através da generalização dos atributos e com base no número de ocorrências, é possível extrair regras caracterizadoras. No caso do exemplo da Figura 2.7, se o atributo *X* for o atributo espacial, como por exemplo o concelho, e o *Y* um atributo semântico, como o tipo de produto vendido, é possível concluir que ambos os concelhos *X1* e *X2* caracterizam-se por vender produtos do tipo *Y1*, uma vez que o número de ocorrência é maior.

Com base no processo anteriormente descrito, podemos retirar também outro tipo de regras [10]. Por exemplo, podemos utilizar o processo de indução para revelar regras discriminantes, isto é, regras que identificam conceitos existentes numa dada relação que não estão presentes noutra. Por exemplo, no caso de doenças, permite a identificação de sintomas que se manifestam numa determinada doença e noutra não. Isso é conseguido através da aplicação do processo de generalização ao mesmo nível nas duas relações que queremos distinguir. Depois é feita uma análise da *prime relation* obtida e através do número de ocorrências das várias características identificadas concluiu-se quais seriam bons elementos discriminantes.

2.2.3. Generalização de dominância espacial e não espacial

A descoberta de relações entre dados espaciais e não espaciais pode ser realizada utilizando a indução orientada aos atributos de duas maneiras diferentes enunciadas em [6]: *spatial dominant generalization* e *non-spatial dominant generalization*.

A diferença entre estes dois métodos está na prioridade que é dada aos atributos a serem generalizados. A generalização de dominância espacial primeiro realiza as tarefas de generalização nos dados espaciais, utilizando as hierarquias de conceitos, estruturas de dados hierárquicas ou algoritmos de agrupamento espaciais. Só no final desta generalização para o nível desejado dos atributos espaciais é que é feita a indução nos dados não espaciais. Por outro lado, na generalização de dominância não espacial numa primeira fase é realizada a indução orientada aos atributos nos dados não espaciais e, só posteriormente, é feito o agrupamento dos tuplos que têm referências espaciais próximas e apresentam, após a generalização, valores iguais. Esta proximidade no caso de pontos corresponde a uma distância que não deve ser excedida para que dois pontos sejam considerados "vizinhos", e no caso de polígonos, são considerados próximos polígonos que sejam adjacentes. Esta união dos objectos espaciais pode então ser realizada através de técnicas de agrupamento espacial, como as que são referidas na secção 2.4.

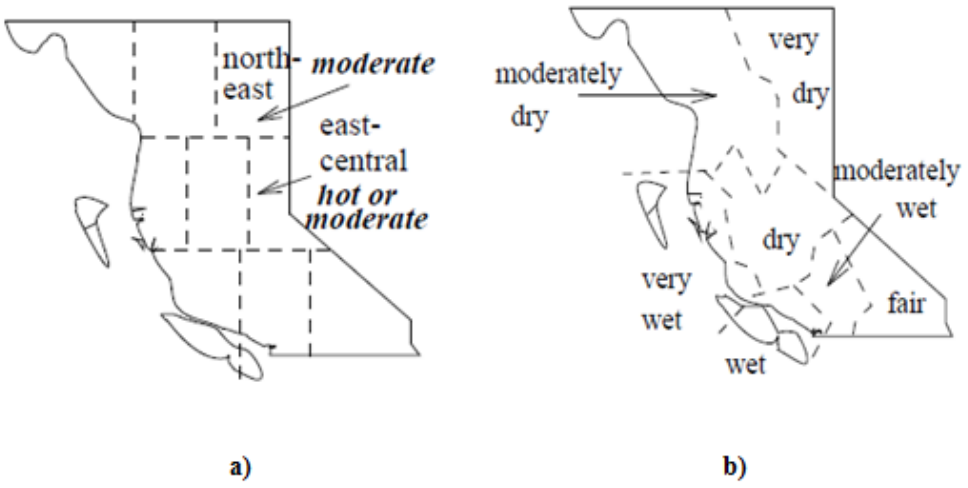


Figura 2.9 – Generalização de dominância espacial (a) e generalização de dominância não espacial (b).

Imaginemos o caso que queremos analisar os padrões climáticos das regiões de *British Columbia*. Com a possibilidade de utilizar qualquer um dos tipos de generalização, os resultados seriam parecidos aos apresentados na Figura 2.9 com as imagens retiradas de [6]. Como é ilustrado pela Figura 2.9 a), as partições são definidas pelas regiões definidas na hierarquia e cada partição é caracterizada de acordo com os dados não espaciais. O resultado da operação é diferente se for utilizado o método de generalização de dominância não espacial, ilustrado na Figura 2.9 b), pois os dados sobre o clima são generalizados primeiro e só depois é feito o agrupamento espacial dos objectos espaciais próximos que apresentam as mesmas características.

2.3. Hierarquias

Como já foi mencionado, as hierarquias têm um papel fundamental para a realização do processo de indução. São estas que demonstram a forma como se organizam os conceitos e como se pode ir ascendendo nos conceitos de acordo com o conhecimento de quem criou o modelo. Nesta secção serão apresentadas formas de gerar hierarquias para os atributos numéricos e depois demonstrada uma técnica de equilibrar uma árvore hierárquica de acordo com os parâmetros definidos pelo utilizador.

2.3.1. Geração automática de hierarquias para atributos numéricos

No caso dos atributos numéricos, as hierarquias de conceitos podem ser geradas autonomamente, não havendo necessidade destas serem indicadas por um especialista na área em causa. Para fazer a construção da hierarquia é realizada uma análise das características dos próprios dados.

Para a geração automática das hierarquias para atributos numéricos são tidas em conta duas normas: *completude* e *uniformidade* [15]. A completude implica que os valores abrangidos pela hierarquia de um atributo numérico cubram todos os valores existentes no conjunto de dados utilizados para a tarefa. Já a uniformidade consiste nos intervalos presentes na *prime relation* terem uma distribuição baseada na frequência dos valores dos atributos presentes no conjunto de dados da relação inicial.

Em [15] é descrito um algoritmo para criação de hierarquias de conceitos para atributos numéricos, tendo em consideração a distribuição dos dados na relação inicial. Assim sendo, dada a relação inicial e um *atribute threshold* (que, neste contexto, significa o número máximo de segmentos desejado para o atributo em causa), o algoritmo comporta-se da seguinte forma:

1) Percorre-se os dados da relação inicial para encontrar o intervalo no qual estão compreendidos os valores do atributo numérico, sendo *mínimo* e *máximo* o valor mais baixo e o mais alto da relação, respectivamente.
2) Determina-se o valor do intervalo como
$$intervalo = \frac{máximo - mínimo}{k \times T},$$
onde *k* é uma constante que geralmente varia entre 5 a 10 e representa a espessura da segmentação, e T o valor de *threshold*. Os valores referentes ao *intervalo*, *mínimo* e *máximo* podem ser arredondados e tornados personalizáveis pelo utilizador.
3) Constrói-se os segmentos baseados no intervalo, desde o valor referido por *mínimo* até ao *máximo*. Assim ficamos com,
$$[mínimo, mínimo + intervalo],$$
$$[mínimo + intervalo, mínimo + 2 \times intervalo],$$
$$\ldots,$$
$$[mínimo + (k \times T - 2) \times intervalo, mínimo + (k \times T - 1) \times intervalo].$$

4) É criado um histograma baseado nos valores do atributo da relação inicial (Figura 2.10). Para cada valor do atributo presente na relação inicial, é analisado em qual dos segmentos se insere e é feita a contagem das ocorrências para cada segmento.
5) Por último, para se obter uma distribuição equilibrada das ocorrências pelos segmentos, os segmentos têm que ser analisados. Organiza-se os segmentos por ordem crescente do seu valor de intervalo e, começando numas das extremidades, vão se juntando num mesmo segmento a sequência de segmentos contíguos até que a soma das suas ocorrências se aproxime o mais possível de $\frac{total\ de\ ocorrencias}{T}$. A extremidade por onde se começa esta fase do processo é relevante para o resultado final da hierarquia, uma vez que as somas das ocorrências nos diferentes segmentos podem levar à criação de segmentos finais diferentes. Desta forma, o intervalo de cada um dos segmentos resultantes é definido com base no valor do menor e maior intervalo que lhe deram origem, ou seja,

   $$[minimo\ do\ primeiro\ segmento, máximo\ do\ último\ segmento].$$

   O processo deve ser repetido até não haver mais segmentos a processar. Todo este processo culmina com a criação de uma hierarquia como a que está presente na Figura 2.10 que se baseou no histograma apresentado.

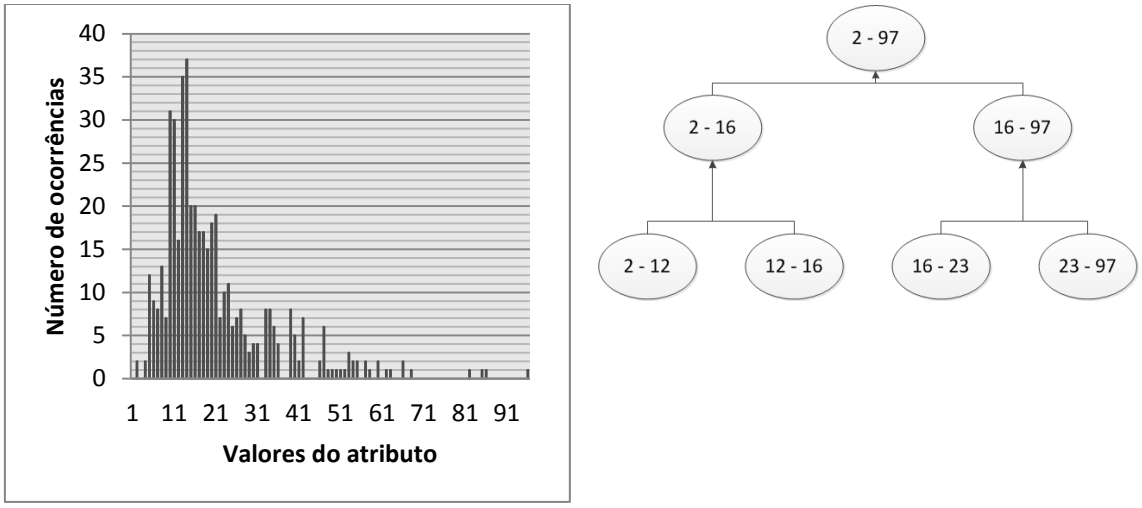


Figura 2.10 - Hierarquia criada a partir do histograma para $T = 4$.

Seguindo a análise de complexidade do processo feita em [15] , no primeiro passo temos complexidade $O(n)$ pois, percorremos a relação inicial para descobrir o *mínimo* e o *máximo*. Os passos 2 e 3 estão relacionados com a criação dos segmentos o que é muito inferior a *n*. No passo 4, a computação é $O(n)$ uma vez que necessita de passar por todos os tuplos da relação inicial para fazer a contagem das ocorrências em cada segmento. Por último, para união dos segmentos percorremos todos os segmentos, mas o número de segmentos criados é muito inferior a *n*. Assim, concluímos que no pior caso é $O(n)$, sendo *n* o número de tuplos da relação inicial.

Outro método de criação de segmentos pretende que os segmentos sejam criados já de forma a ter um número de ocorrências semelhante. A primeira operação a ser realizada é colocar os valores numa estrutura ordenada, e em simultâneo, guardar informação do número de ocorrências de cada um desses valores. Feito este pré-processamento, os valores são percorridos por ordem crescente e é somado o seu número de ocorrências. Quando o total dessa soma ultrapassa $\frac{Numero\ de\ ocorrencias}{threshold}$, um segmento é criado, sendo o seu intervalo definido por

$$[primeiro\ valor\ contabilizado,\ último\ valor\ contabilizado],$$

sendo estes valores, o número de ocorrências que foram contabilizados para o segmento em questão. Assim que um segmento é criado um outro começa a ser avaliado partindo do valor seguinte até ao final dos valores, ou até o limite ser novamente ultrapassado.

Analisando os dois métodos e utilizando os mesmos argumentos, reparamos que existe uma diferença em termos dos segmentos gerados. No primeiro método os segmentos finais construídos são resultantes da concatenação de outros segmentos com um tamanho definido pelo $intervalo$ calculado. No outro método estes segmentos finais são gerados à medida que estão a ser analisados os valores. Para além disso, o tamanho dos segmentos no primeiro método é sempre um múltiplo do $intervalo$ enquanto no segundo os segmentos podem surgir com tamanhos completamente aleatórios, uma vez que são criados directamente com base na distribuição dos valores.

2.3.2. Alteração dinâmica de hierarquias de conceitos

Em algumas situações, as hierarquias de conceitos fornecidas podem não ser as que melhor se encaixam para uma determinada tarefa de generalização quando desejamos que os conceitos tenham um número de ocorrências semelhante. Assim, quando verificamos que existe uma distribuição das ocorrências desigual entre os conceitos, podemos optar por alterar dinamicamente a hierarquia para que essa distribuição seja o mais uniforme possível.

As alterações a fazer devem ter em conta um conjunto de regras descritas em [15]. Estas alterações devem preservar o máximo possível a estrutura inicial da hierarquia. As alterações devem ser feitas a cada nova tarefa, uma vez que o conjunto de dados e a distribuição destes é provavelmente diferente. Desta forma, as alterações à hierarquia não devem ser permanentes. Por outro lado, deve ser feito um balanceamento da distribuição dentro dos níveis para que não haja dentro do mesmo nível de abstracção nós de grande cardinalidade e outros com uma cardinalidade muito pequena. Também deve ser tido em conta a quantidade de alterações que precisamos fazer. Na maioria dos casos não precisamos de alterar toda a hierarquia mas sim apenas os conceitos perto ou de nível superior àqueles em que o utilizador está interessado. Em seguida é apresentado um algoritmo para alteração dinâmica de uma hierarquia de conceitos.

A Figura 2.11 mostra uma hierarquia de conceitos, no seu formato inicial, com a contagem de ocorrências.

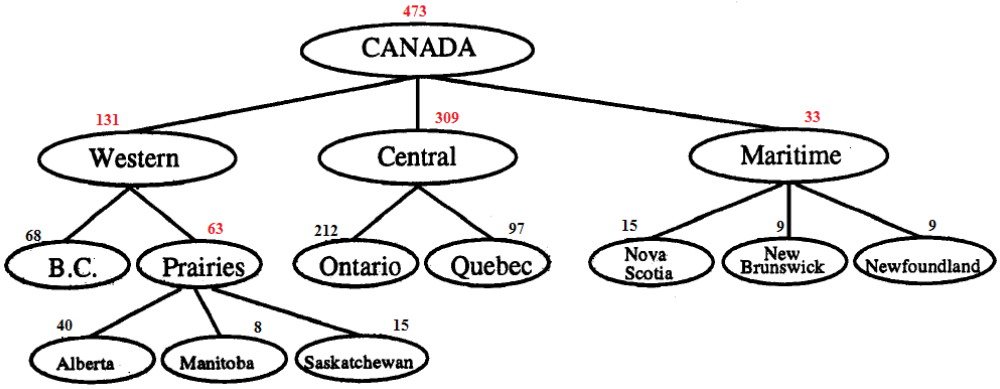


Figura 2.11 - Hierarquia original, retirada de [15].

Dada uma hierarquia de conceitos *H*, uma relação inicial *R* com um atributo *A* ao qual se refere a *H* e um *atribute threshold T*,

1) A primeira fase é uma fase de inicialização dos dados para preparação do algoritmo.
    a. Associa-se a cada nó (*n*) o seu nível na hierarquia *H*.
    b. Percorrendo toda a relação *R*, calcula-se o número de ocorrências em *R* de cada conceito representado pelas folhas de *H*. Esse cálculo é propagado aos nós superiores, através da soma do número de ocorrências dos seus descendentes directos. Defina-se *n.count* como o número de ocorrências do conceito referente a *n* em *R*. O número total de ocorrências (*total*) é dado pelo $\sum n.count$ das folhas da hierarquia.
2) Realizam-se as alterações na hierarquia *H* utilizando uma abordagem *top-down*.
    a. Preparam-se dois conjuntos. Um, a que damos o nome de *buff*, que contém inicialmente os nós no topo de *H*, e outro inicialmente vazio, a que chamamos *prime*. Para cada nó *n*, calcula-se o seu peso *weight* através da formula, $n.weight = n.count/total$. Os nós são depois classificados de acordo com o seu peso. Se $n.weight < r$, onde *r* é o limite de peso (*weight threshold*) definido por $r = \frac{1}{T}$, então o nó é considerado como um nó grande, caso contrário é um nó pequeno. A classificação é a seguinte:
        - Um nó grande que seja folha é classificado como *B*
        - Um nó grande que não seja folha é classificado como *B'*
        - Um nó pequeno que seja folha é classificado como *S*
        - Um nó pequeno que não seja folha é classificado como *S'*

b. Em seguida, são analisados os nós presentes em *buff*. Todos os nós *B* que estejam em *buff* são passados para o conjunto *prime*, e os *B'* são substituídos pelos seus descendentes directos. Este processo é repetido até *buff* ser composto apenas por nós *S* e *S'*.

c. Se a soma do número de elementos em *buff* e *prime* não for superior a *T* o processo termina aqui. Caso contrário volta-se ao início do passo 2) efectuado as seguintes alterações e relação aos nós em *buff*:

$$T' = T - |prime|$$

$$total' = \sum n.count$$

$$n.weight = \frac{n.count}{total'}$$

$$r' = \frac{1}{T'}$$

A classificação é agora baseada nos valores de *weight'* e *r'* e o processo de reclassificação e cálculo de novos pesos dos nós é realizado até que não se verifiquem alterações.

3) Caso ainda persistam nós em *buff*, é feito uma junção desses nós de baixo para cima. Começando pelos nós de nível inferior:

a. Primeiro juntam-se os nós que partilham o mesmo antecessor e é calculado o *weight'* desse novo nó formado. Caso seja maior ou igual que *r'*, então é passado para o conjunto *prime* e decrementa-se *T'*.

b. Verifica-se se $|buff| \leq T'$. Em caso afirmativo, todos os elementos em *buff* são passados para o conjunto *prime* e o processo termina. Caso contrário, realiza-se o processo de recalculo dos pesos, passa-se para o nível acima e volta-se ao passo 3) a.

c. No caso de não haver mais níveis, juntam-se os nós em *buff* em *T'* grupos e passam-se estes *T'* nós formados para o conjunto *prime*.

4) É feita a associação entre os nós presentes em *prime* e os valores dos atributos de *A* presentes na relação inicial *R*.

Na Figura 2.12 mostra-se a associação feita entre os nós da hierarquia original e os nós da nova hierarquia. Desta forma, é possível obter uma hierarquia em que os vários conceitos estão equilibrados em termos de número de ocorrências.

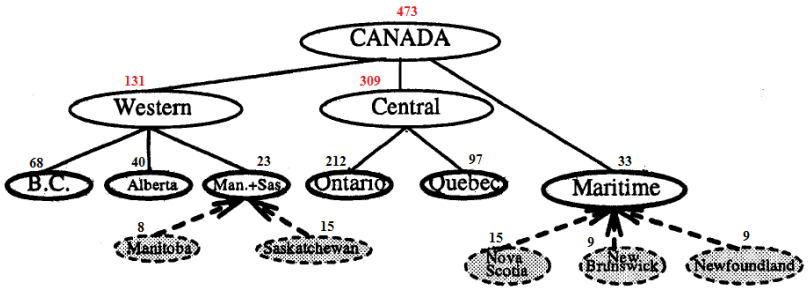


Figura 2.12 - Hierarquia remodelada.

2.4. Agrupamento espacial

Quando estamos perante um processo de generalização de dominância não espacial é necessário utilizar técnicas de agrupamento espacial para fazer a união dos objectos que sejam considerados "próximos" e que contenham a mesma característica dominante em comum. Em seguida, são apresentadas duas formas de fazer esse agrupamento, tanto em relação a pontos com em relação a polígonos.

2.4.1. DBSCAN

Em [9] foi realizado um estudo que fazia a avaliação de diferentes algoritmos de agrupamento espacial com vista à sua integração num contexto SOLAP. Contudo, no âmbito desta dissertação estamos interessados em algoritmos que possibilitem esse agrupamento não só com base na componente espacial dos objectos mas também na componente semântica. Uma vez que o *DBSCAN* possibilita o agrupamento com base nas características semânticas dos objectos e já se encontra integrado no protótipo SOLAP+, foi considerado a escolha mais acertada para realizar as tarefas de agrupamento que serão aplicadas sobre os resultados de uma generalização.

O *DBSCAN (Density Based Spatial Clustering of Applications with Noise*) é um algoritmo baseado na densidade [17]. Tem como ideia chave a identificação de grupos como regiões densas de objectos e considera como ruído regiões com baixa densidade de objectos. Uma vez que queremos também introduzir a semântica dos objectos como factor de agrupamento, a ideia chave direcciona-se no sentido de identificar grupos como regiões densas de objectos com as mesmas características.

Para descrever o algoritmo é necessário introduzir um conjunto de definições em que assenta o *DBSCAN*:

- ***ε-neighborhood***, $N_{Eps}(p)$, representa o conjunto de objectos que se encontram a uma distância $\leq Eps$ de um dado ponto p e que apresentam as mesmas características. Denominemos $char(p)$ como as características de um dado ponto p e $dist(p,q)$ pela distância entre os pontos p e q. Assim, $\forall\, q \in N_{Eps}(p), char(q) = char(p) \cap dist(p,q) < Eps$.
    - Se $|N_{Eps}(p)| \geq MinPts$, então p é um ***core point***.
- Dizemos que p é ***directly density-reachable*** de q se $p \in N_{Eps}(q)$ e se q é um *core point.*
- Um objecto p é ***density-reachable*** de q se existe uma cadeira de objectos $p_1, \ldots, p_n$, $p_1 = q$ e $p_n = p$ tal que p_{i+1}é *directly density-reachable* de p_i.
- De entre um conjunto de objectos D, p é ***density-connected*** a q se existir um objecto $o \in D$ em que p e q sejam *density-reachable* de o.

Assim sendo, todo o ***cluster*** (grupo) G identificado pelo DBSCAN verifica as seguintes propriedades:

- $\forall\, p, q$, se $p \in G$ e q é *density-reachable* de p então $q \in G$.
- $\forall\, p, q \in G$, p é *density-connected* a q.

Todos os pontos que não pertençam a qualquer grupo são considerados como ***noise*** (ruído).

Na Figura 2.13, retirada de [18], são demonstrados os conceitos descritos anteriormente. Como se pode ver p é *density-reachable* de q e s *density-connected* a r. Se consideramos $MinPts = 3$ então m, p, o e r são *core points*.

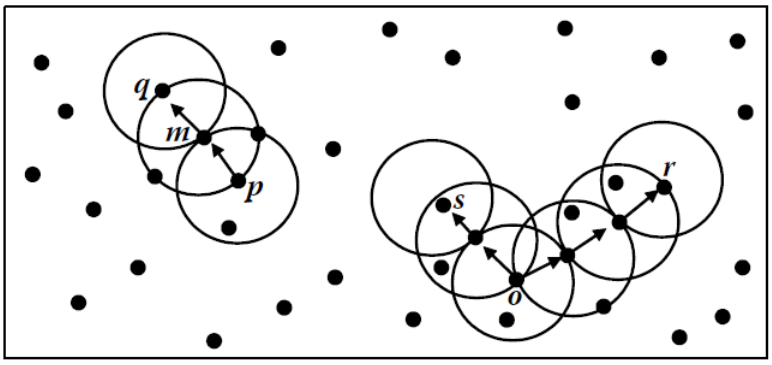


Figura 2.13 - Conceitos de density-reachable e density-connected no agrupamento baseado em densidade.

Esta técnica procura por *clusters* verificando para todos os pontos da base de dados o seu ε-*neighborhood*. Assim, dando um ponto arbitrário p não visitado, obtém todos os pontos *density-reachable* de p. Caso seja um *core point*, é criado um grupo e o ponto marcado como visitado. Caso contrário apenas é feito o registo de visita do ponto não havendo criação de qualquer grupo. O processo termina quando todos os pontos forem visitados, não existindo objectos que possam ser adicionados a qualquer grupo.

A complexidade temporal deste processo é $O(n^2)$ onde n é o número de objectos, a menos que sejam utilizados índices espaciais por parte da base de dados passando a operação de consulta a $O(\log(n))$ e, por isso, a complexidade total a $O(n \log(n))$. Tal como acontece em muitos outros algoritmos de agrupamento, também este tem o problema de necessitar de parâmetros de entrada. No caso do *DBSCAN* são Eps e $MinPts$. Contudo, estes parâmetros podem ser definidos através de heurísticas (estratégia já adoptada anteriormente no protótipo SOLAP+ com sucesso), de forma a tornar o processo o mais automático possível.

2.4.2. Regionalização

A regionalização [19] é um conceito frequentemente utilizado quando se fala em regiões sendo aplicado em vários cenários. Este consiste na formação de novas regiões a partir de um espaço inicial, quer seja agrupando diferentes locais, quer seja subdividindo-os.

Tal como é referido em [20], esta operação tornou-se fundamental quer para as aplicações SOLAP quer para as aplicações de *data mining* espacial. No âmbito desta dissertação, existe interesse em formar regiões que partilhem a mesma característica dominante. A este processo dá-se o nome de *polygon amalgamation* [20] que dado um conjunto de polígonos, leva à criação de um novo polígono formado pelo limite da união dos polígonos fornecidos.

Em seguida, na Figura 2.14, temos um esquema exemplificativo do que se pretende com este processo. Imagine-se que temos um conjunto de polígonos que representam uma série de países. Após aplicar o processo *Polygon Amalgamation,* agrupamos os polígonos e ficamos com a área representativa da região que engloba todos os países. No caso do exemplo, consegue-se definir a região dos Balcãs.

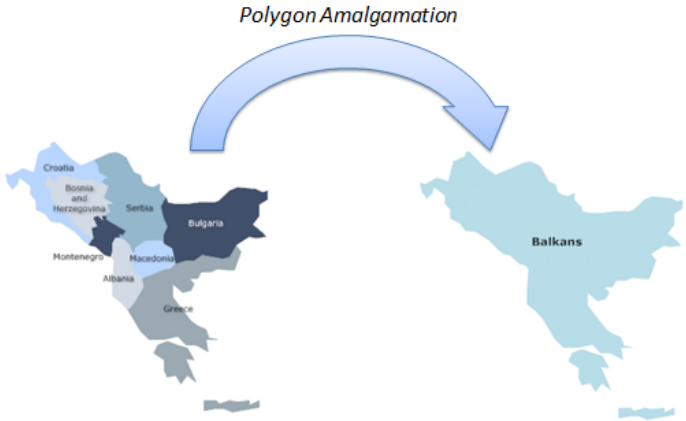


Figura 2.14 - *Polygon amalgamation*

Em [19] é descrito um método para realizar esta regionalização, o qual se baseia numa restrição de contiguidade espacial. Esta restrição de contiguidade pode ser definida de diferentes formas: como sendo a adjacência dos polígonos, o comprimento da fronteira partilhada entre polígonos, ou até a distância entre os seus centróides, por exemplo. Na Figura 2.15 está ilustrado o princípio do algoritmo.

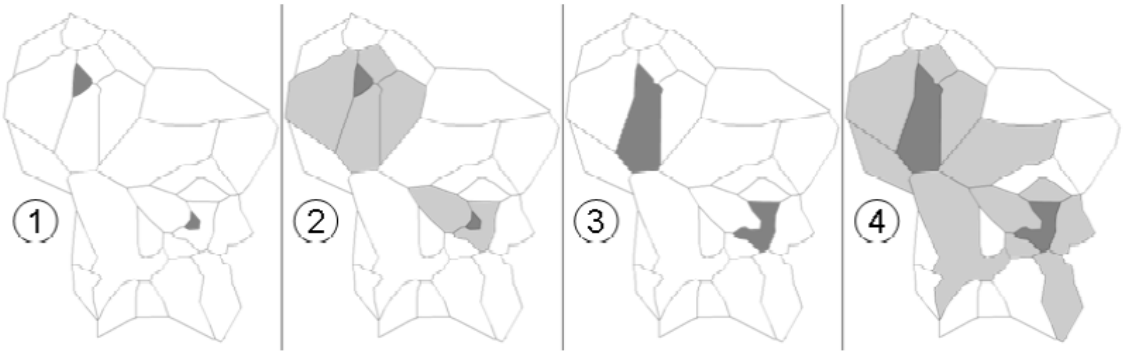


Figura 2.15 - Processo de regionalização

1. São escolhidos os polígonos de início do processo.
2. Seleccionados os candidatos de acordo com a restrição de contiguidade.
3. Juntam-se os polígonos que são considerados similares de acordo com as regras definidas.
4. Passamos para o próximo passo da iteração.

Como verificamos, este é um algoritmo de crescimento regional, prosseguindo passo a passo com uma estratégia de agrupamento, o que, em termos de filosofia do algoritmo, se assemelha à seguida pelo *DBSCAN*, sendo o processo de cálculo da complexidade semelhante.

O *P-DBSCAN*, variante do algoritmo *DBSCAN* destinada aos polígonos, não foi utilizada apesar de estar já presente no protótipo SOLAP+ porque a função de distancia utilizada baseia-se nos valores já calculados da distância de *hausdorff* ajustada, apresentada em [9]. Uma vez que a filosofia não era o agrupamento de objectos espaciais possivelmente espaçados entre si, mas sim somente a união de zonas que tivessem fronteiras em comum formando regiões com características iguais, optou-se pela separação dos processos.

3. Extensão ao SOLAP+

Ao longo deste capítulo será apresentada a proposta para inclusão no protótipo SOLAP+ de um mecanismo de sumarização de informação através da indução orientada aos atributos com a finalidade de caracterizar as regiões através de informação semântica e com base em métricas. Primeiro serão apresentados alguns detalhes do SOLAP+ que ajudam a compreender algumas das opções tomadas. Posteriormente, serão descritas várias formas de interacção possíveis, e todos os passos desde que uma relação é escolhida para análise até à obtenção de uma outra com a informação sumarizada.

3.1. Conceitos base do SOLAP+

O sistema SOLAP+ foi desenvolvido nos últimos dois anos por Jorge, R. [5] e Silva, R. [9]. Seguindo a nomenclatura utilizada em [9], temos o seguinte conjunto de definições:

- *Atributo Semântico (aS)* – Representa um atributo do tipo alfanumérico presente numa dimensão;
- *Atributo Espacial (aEP)* – Atributo de uma dimensão que está associado a informação espacial (conjunto de coordenadas). Existe sempre um atributo semântico associado a um atributo espacial, o qual é representado por *aS(aEP)*.
- *Métrica Numérica (mN)* – Designa uma métrica numérica associada a uma tabela de factos.

No modelo existem dois tipos de dimensões: semânticas e espaciais. As dimensões espaciais são todas aquelas que apresentem pelo menos um atributo espacial. No caso de uma dimensão não conter nenhum atributo espacial, é considerada como semântica.

Escolhidas as dimensões pretendidas para análise, estas farão parte de uma interrogação responsável por obter os dados pretendidos pelo utilizador. O resultado proveniente dessa irá ser apresentado sobre a forma tabular na tabela de suporte. Posto isto, um dos factores que se tem de ter em consideração é a estrutura das tabelas. Depois de processada a interrogação a tabela pivô que formará a tabela de suporte está sujeita a determinadas restrições de organização. Desta forma, o cabeçalho encontra-se organizado da forma apresentada na Figura 3.1, baseada num exemplo retirado do protótipo quando se analisava a quantidade emitida pelas instalações por diferentes meios. De notar que, no exemplo, a segunda linha da tabela não apresenta valores no campo referente à água, uma vez que não existem dados registados de emissões realizados por essa instalação por via desse meio.

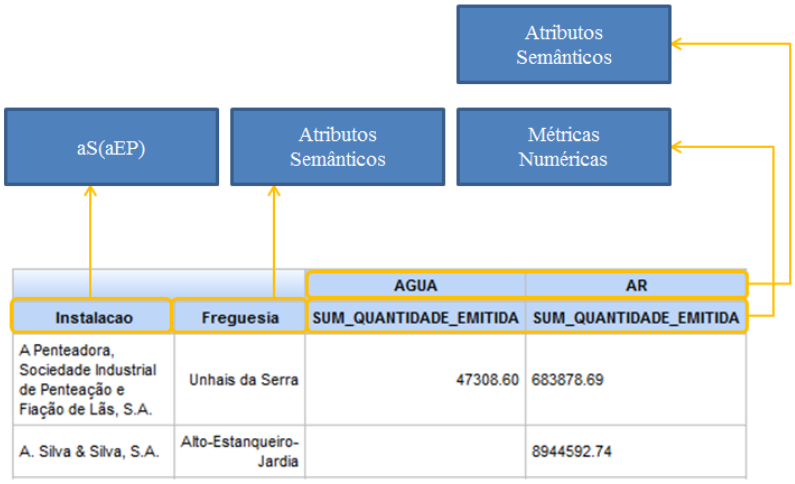


Figura 3.1 - Estrutura da tabela de suporte.

A primeira secção destina-se aos atributos semânticos que representam os objectos espaciais. Existem duas secções que se destinam aos atributos semânticos, uma vez que em alguns casos necessitamos que estes estejam presentes como cabeçalhos, mais precisamente quando provêm de dimensões semânticas ou quando provêm de dimensões espaciais de granularidade inferior quando comparadas com *aS(aEP)*. Só desta forma conseguimos, no caso das situações mencionadas, manter uma relação de 1:1 entre os elementos da tabela de suporte e os objectos representados no mapa. Por último, existe também uma secção destinada às métricas numéricas.

3.2. Integração do processo de sumarização

A integração de um processo de sumarização passa por introduzir um conjunto de acções a realizar numa fase de pré-processamento, no caso de ser requisitado pelo utilizador a sumarização da informação presente numa dada relação inicial.

Para a integração deste processo é preciso ter em consideração 3 grandes fases:

1) *Obtenção da relação inicial*
   É necessário definir as formas como a relação inicial é fornecida ao algoritmo de sumarização, bem como a maneira como o utilizador indica as suas pretensões no que diz respeito à informação a sumarizar, através de que conceitos sumarizar e o nível de sumarização pretendido.
2) *Execução do processo de sumarização*
   Fornecidos os dados recolhidos na fase anterior, será realizada a sumarização da informação através do processo de indução orientada aos atributos generalizando os dados de uma de duas maneiras: dando enfâse à informação espacial ou aplicando uma dominância não espacial. Qualquer uma delas culminará com a obtenção de dados para a formação de uma *prime relation*.

3) *Formas de apresentação da relação final*
   Por último, a relação obtida deverá ser trabalhada de forma a obter-se uma boa representação da informação para o utilizador para efeitos de análise. Para isso, é preciso transformar a relação numa tabela apresentável na zona destinada à tabela de suporte e representar essa informação no mapa.

3.2.1. Obtenção da relação inicial

A primeira situação a resolver é a obtenção da relação inicial sobre a qual vai ser aplicado o processo de sumarização e solicitar os parâmetros necessários à execução do processo.

3.2.1.1. Indicações necessárias por parte do utilizador

Para começar a execução do processo de sumarização será necessário fornecer informações sobre:

- Os atributos a sumarizar, os quais vão corresponder aos níveis de granularidade base do processo.
- Para cada um dos atributos, a cadeia conceptual que se pretende utilizar.
- O nível de sumarização pretendido para cada um dos atributos, isto é, o seu *threshold*.

Para a obtenção dos dados que se pretende sumarizar propomos duas formas distintas:

- Indicação do *foco da sumarização*, ou seja, o utilizador com recurso ao modelo dá as informações necessárias para que se possa obter a relação inicial;
- Partindo do resultado de uma interrogação *SOLAP*.

Um dos casos de interacção para definição da relação inicial é através da realização de uma interrogação SOLAP tal e qual como já se encontra definido no protótipo. Contudo, ao contrário do que acontece com os resultados das típicas interrogações SOLAP que apresentam um número de tuplos muito baixo, esta interrogação deverá resultar numa relação com elevada cardinalidade, para que o processo de sumarização seja mais relevante. No caso do SOLAP+, o resultado desta interrogação encontra-se presente na tabela de suporte. Em seguida, segue-se um exemplo de uma interrogação deste tipo.

```
SELECT c.nome_cliente, p.nome_produto, sum(c.compras)
FROM cliente c, produto p, compras cp, periodo t
WHERE c.num_cliente = cp.num_cliente AND
   p.cod_produto = cp.cod_produto AND
   cp.cod_periodo = t.cod_periodo AND
   t.mes = ”12”
GROUP BY c.nome_cliente, p.nome_produto
ORDER BY c.nome_cliente;
```

Quando estamos perante a situação de indicação do foco de sumarização, temos várias situações a analisar: (*i*) indicação apenas de atributos de uma dimensão, (*ii*) escolha de atributos de várias dimensões e (*iii*) atributos de várias dimensões e métricas numéricas.

No primeiro caso, de indicação apenas de atributos provenientes da mesma dimensão, pretendemos apenas extrair informação e encontrar padrões dentro daquela dimensão. Do ponto de vista analítico, nos casos que envolvem a escolha de atributos somente da mesma dimensão, o que se pretende é sumarizar as relações entre eles a um nível conceptual que se considere mais adequado, aproveitando para isso a relação entre níveis de granularidade existentes dentro da própria dimensão. Por exemplo, analise-se a dimensão *produto* representada na Figura 3.2.

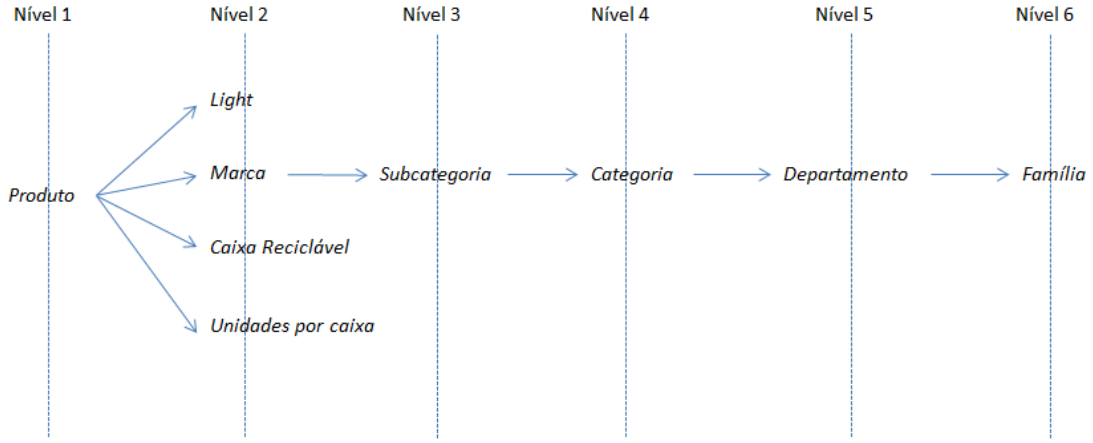


Figura 3.2 - Representação dos atributos da dimensão *produto*.

Durante a escolha dos atributos temos que ter consciência de que o processo é de sumarização e, por isso, é mais relevante quando os atributos escolhidos estão a níveis de granularidade mais baixos. Assim sendo, deve haver sempre pelo menos um atributo que não esteja no topo de uma cadeia hierárquica, sendo recomendado que seja o atributo de nível mais baixo dessa cadeia. Contudo, não é de descartar a situação de escolha de atributos que estejam no nível mais alto de granularidade da sua cadeia hierárquica, como, por exemplo, o atributo *Light*. Estes atributos não serão generalizados por já estarem no topo da sua cadeia mas são aceitáveis para a distinção a realizar entre tuplos. Por exemplo, sem a presença do atributo *Light*, produtos da mesma *Marca* estariam juntos independentemente se eram *Light* ou não.

Quando são escolhidos atributos de diferentes dimensões, terá que ser indicada uma tabela de factos. Essa tabela de factos indicará a forma como esses atributos se relacionam. Nesta situação, ao recolhermos os registos dessa tabela de factos é muito provável que surjam tuplos repetidos. Por exemplo, se estamos perante uma tabela de factos que guarda dados sobre vendas, o mesmo produto pode ser vendido várias vezes ao longo de um determinado período de tempo e, assim, surgem vários registos desse produto no mesmo dia. Caberá ao utilizador indicar se pretende que sejam mantidas as repetições ou que a relação inicial não contenha tuplos repetidos. Com esta decisão, o utilizador estará a indicar se pretende que a existência de vários tuplos referenciando a mesma situação se reflicta ou não nas contagens a efectuar durante o processo de indução. Utilizando um exemplo para analisar o impacto desta escolha, imagine-se que estamos a analisar as compras de diferentes produtos efectuadas por clientes em lojas diferentes. Ao serem utilizadas repetições, resultantes de compras que foram realizadas pela mesma pessoa na mesma loja várias vezes, estamos a indicar que o nosso interesse está nos actos de compra. Por outro lado, se eliminarmos as repetições o nosso interesse passa a recair sobre os produtos comprados. Desta forma, concluímos que a escolha pela presença ou ausência de repetições leva a que a semântica resultante no final do processo de generalização seja diferente.

Ao haver atributos de várias dimensões, teremos que colocar a hipótese de mais do que um desses atributos possuir informação espacial associada. O caso da informação ser referente a um atributo de uma determinada dimensão está plenamente comtemplado e incorporado no âmbito desta dissertação. Já o caso que envolve a existência de mais de um atributo com informação espacial é mais delicado, uma vez que teria que envolver a extracção de outro tipo de regras para além das caracterizadoras, como seria o caso das regras de associação para se conseguir correlacionar os dois espaços. Desta forma, e tendo conhecimento da existência destes dois casos, são consideradas apenas relações iniciais que apresentem apenas um atributo espacial de forma a esse ser caracterizado individualmente.

Poderemos ter ainda atributos de várias dimensões aos quais estão associadas métricas de uma tabela de factos. O utilizador indicará explicitamente os atributos e as métricas sobre quais deseja que o processo de sumarização seja aplicado.

No caso das métricas, poderá ou não ser indicado um operador de agregação (*SUM*, *AVG*, entre outros) que será utilizado sobre essa métrica a quando da detecção de tuplos idênticos. Quando são indicados os operadores de agregação, estes são indicados e aplicados por métrica. No caso de não haver indicação do operador de agregação, terá que ser criada uma hierarquia para essa métrica. Também aqui, existem as duas variantes respeitantes à inclusão de tuplos repetidos ou não, uma vez que esta decisão terá influência na construção dos segmentos que constituirão a hierarquia, dado que a sua definição é baseada no número de ocorrências, de forma a manter os segmentos balanceados em termos de número de casos registados.

Também aqui, a escolha pela utilização ou não de operadores de agregação levará a significados diferentes da informação no final do processo de sumarização, uma vez que o significado das hierarquias criadas será diferente. Por exemplo, no caso de estarmos a utilizar um operador de agregação como o AVG para o valor das compras, quando a hierarquia for construída serão obtidos intervalos referentes aos valores médios de compras. No caso de utilizarmos o operador SUM, os intervalos seriam referentes ao somatório do valor das vendas. Estes intervalos têm um significado completamente diferente daqueles que são criados quando não é feita qualquer agregação dos valores.

Depois dos vários casos indicados é importante frisar que não poderão ser utilizados atributos espaciais de dimensões diferentes. Sem esta restrição, poderíamos cair na situação indicada anteriormente de haver referências para objectos espaciais de dimensões diferentes e seria complicado fazer a sua representação.

Assim sendo, no caso geral obtém-se uma estrutura da relação inicial formada pelos elementos indicados na Figura 3.3. Cada linha da relação apresenta um elemento referente ao atributo semântico que descreve o atributo espacial e elementos referentes os atributos semânticos e métricas escolhidas. O atributo espacial será o objecto que será caracterizado e os atributos semânticos serão a base de caracterização, sendo os valores de cada registo concatenados dando origem a formação de uma característica (podendo ou não ser nova para o objecto em causa). Por último, temos as métricas cuja informação poderá fazer parte da característica, caso hajam hierarquias numéricas, ou, caso contrário, o seu valor será agregado para cada característica com base no operador de agregação escolhido.

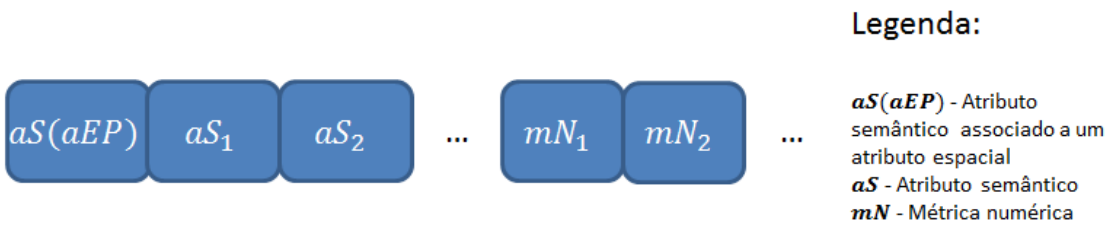


Figura 3.3 - Estrutura do conjunto dos atributos escolhidos.

Outra informação que tem que ser obrigatoriamente conhecida é a cadeia conceptual a utilizar para cada um dos atributos. As cadeias conceptuais correspondem à sequência de níveis definidos para cada dimensão no meta-modelo que é carregado no início de uma sessão. Um atributo de uma dada dimensão pode ter associado uma série de outros atributos num nível superior, formando uma cadeia de conceitos, tal como está esquematizado na Figura 3.2, anteriormente referida.

No caso do *threshold* existem duas formas de este ser indicado: através da indicação do limite pretendido ou com base nos níveis conceptuais presentes na hierarquia escolhida, indicando um desses níveis para limite do processo de generalização do atributo em causa. Na Figura 3.4 estão ilustradas as duas formas de indicação do nível de generalização máximo pretendido. Uma através da escolha do nível referente à *Categoria* e outra através da indicação directa do *threshold*, no caso do exemplo foi escolhido três, sendo o atributo *Produto* generalizado até ao nível *Família*, uma vez que do nível *Família* fazem parte somente dois conceitos e do nível *Departamento* fazem parte quatro e dessa forma só o nível *Família* respeita as condições do *threshold* indicado. Desta forma, o utilizador pode fazer uma escolha baseada no seu nível de conhecimento do modelo em que está a trabalhar ou então indicar um valor que deseja de acordo com as suas pretensões, sem ter necessariamente que ter conhecimento para qual dos níveis irá ser generalizado o atributo em causa.

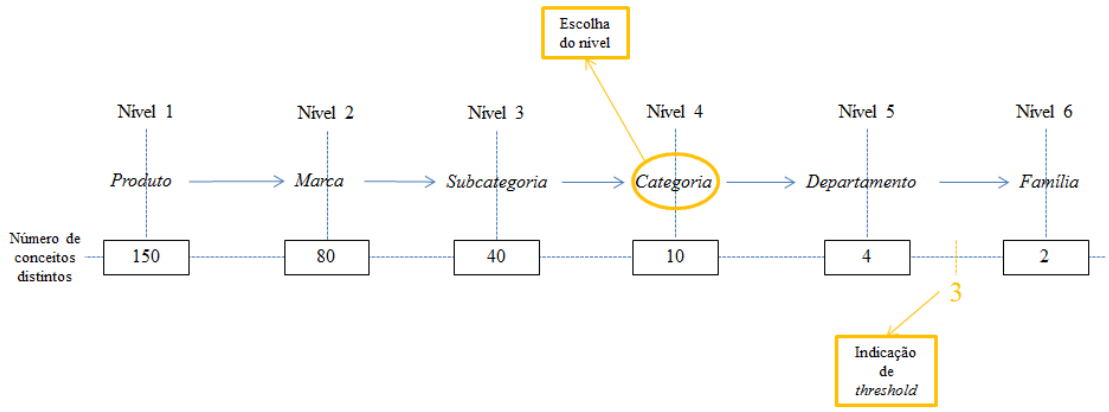


Figura 3.4 - Ilustração da forma de indicação dos limites para a generalização dos atributos.

3.2.2. Execução do processo de sumarização

Uma vez definidos os atributos, as suas cadeias conceptuais e os seus valores de *threshold*, pode-se então passar para o processo de indução orientada aos atributos. Aqui, existem duas fases distintas: uma de análise e construção de hierarquias e outra de aplicação da generalização.

3.2.2.1. Análise e construção de hierarquias

Nesta fase definimos os conceitos que serão utilizados para generalizar a informação existente na relação inicial. Aqui, temos que ter atenção a dois tipos de atributos: atributos semânticos (espaciais ou não espaciais) ou atributos numéricos.

Quando estamos a tratar um atributo do tipo alfanumérico, entramos em conta com a hierarquia escolhida e o valor de *threshold*. Através de uma interrogação à base de dados obtemos a informação da relação entre os valores do atributo em análise presentes na relação inicial e os valores dos atributos de nível de granularidade superior presentes na cadeia de conceitos escolhida. Desta forma, conseguimos criar uma representação da hierarquia formada pelos valores existentes nos diferentes níveis de granularidade da cadeia escolhida. Seguindo o exemplo do *Produto*, e escolhida a cadeia formada por $Produto \rightarrow Marca \rightarrow Subcategoria \rightarrow Categoria \rightarrow Departamento \rightarrow Família$, conseguimos construir uma hierarquia como a demonstrada na Figura 3.5.

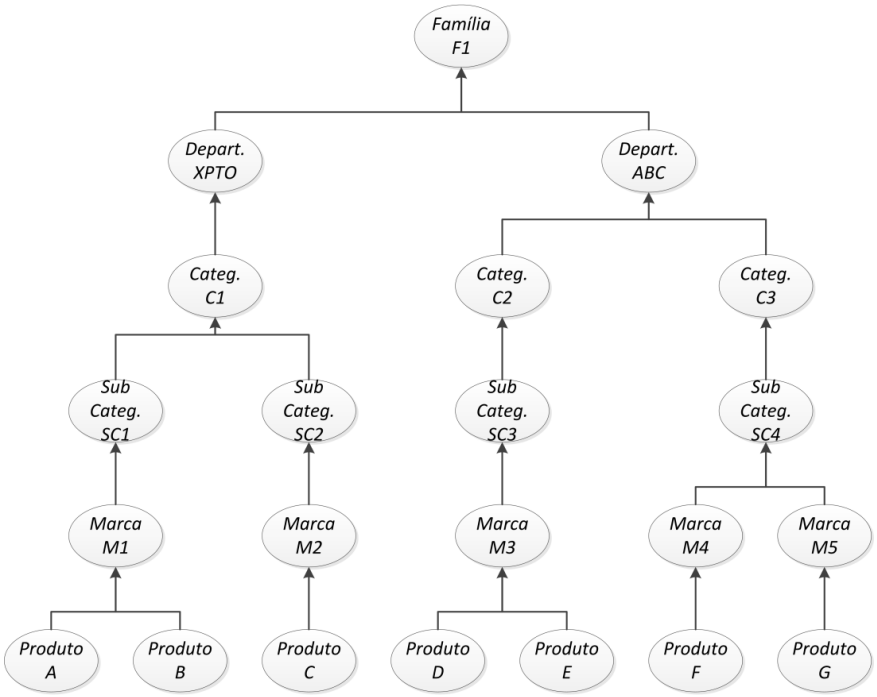


Figura 3.5 - Exemplo de uma hierarquia criada com base nos dados extraídos para efeitos de generalização.

Analisando o número de valores distintos existentes a cada nível da hierarquia, começando pelo nível mais baixo, vamos comparando esse valor com o limite definido (*threshold*) e o primeiro que ficar a igual ou a baixo desse limite corresponde ao nível mínimo da hierarquia a utilizar.

O utilizador poderá indicar que pretende que as hierarquias sejam refinadas de acordo com o número de elementos de cada conceito. Neste caso, os conceitos são revistos e, caso seja necessário, fundem-se conceitos evitando que haja uma grande discrepância entre o número de elementos presentes em cada conceito. Para isso, será utilizado o método apresentado na secção 2.3.2 deste documento.

Quando se fala nos atributos numéricos escolhidos, está-se a fazer referência às métricas numéricas. Uma vez que estas não possuem qualquer hierarquia definida, terá que ser criada uma. Para isso, será utilizado o processo abordado na secção 2.3.1.

3.2.2.2. Generalização

Uma vez terminado o processo de definição das hierarquias e definição dos conceitos a utilizar, podemos passar então ao processo de generalização dos dados presentes na relação inicial fornecida.

No caso da generalização de dominância espacial, primeiro vai ser aplicado o processo de indução aos dados espaciais, isto é, subir na hierarquia de conceitos substituindo os valores da relação inicial pelos conceitos definidos no passo anterior. Com isto, estamos a fazer uma generalização dos objectos espaciais presentes na relação inicial, formando regiões que são definidas pelos atributos de granularidade superior. Para cada região formada, os atributos não espaciais são então generalizados. No caso de terem sido construídas hierarquias para as métricas, estas são tratadas como atributos semânticos e generalizadas para os conceitos definidos. No final fazemos a junção de tuplos idênticos, utilizando contadores para cada tuplo, indicando o seu número de ocorrências.

Em seguida teremos que decidir qual a característica que predomina em cada um dos objectos espaciais. Uma característica é definida pelo conjunto formado pelos atributos semânticos não associadas a um atributo espacial escolhidos. Entrará também para a formação da característica o segmento da hierarquia numérica em que o valor da métrica se encaixar. Esta parte somente é valida se tiver sido solicitada pelo utilizador a criação da hierarquia para essa métrica.

É de esperar que cada objecto contenha evidências de vários tipos de características e por isso é preciso decidir qual delas vai ser a caracterizadora do objecto espacial em causa. A solução proposta para estes casos é realizar uma análise ao "peso" de cada uma das características no objecto. Através das contagens efectuadas, é feita a identificação da tem um maior "peso" em termos de número de ocorrências e o objecto espacial será designado como um local onde aquela característica predomina em relação às outras. No caso do número de ocorrências ser exactamente igual, e tendo sempre que haver uma classificação para o objecto em análise, classificamos como *indefinido* a caracterização de objectos nessas situações.

Quando é aplicada a generalização de dominância não espacial, o processo de indução começa por ser aplicado aos atributos não espaciais, substituindo-os pelos conceitos do nível mínimo que mantém o número de elementos distintos igual ou abaixo do *threshold* definido para cada atributo, ou simplesmente pelo seu correspondente do nível hierárquico indicado como limite de generalização desejado. Posteriormente é feita a caracterização dos objectos espaciais, não havendo generalização do atributo representante desse objecto. Por último, é realizada a junção de objectos espaciais que satisfaçam a condição de proximidade e que apresentem características iguais (evidenciadas através do valores generalizados dos atributos).

Para realizar esse agrupamento temos que ter em consideração se estamos perante pontos ou polígonos. Para realizar o agrupamento de pontos, foi utilizado o *DBSCAN*, apresentado na secção **Erro! A origem da referência não foi encontrada.**, em que a função distância nvolvida neste algoritmo entra em consideração, para além do factor proximidade, também com as propriedades semânticas dos objectos. Estas propriedades semânticas são formadas pelos valores dos atributos semânticos que formam a característica predominante nesse objecto espacial. No caso de estarmos perante objectos que são representados sobre a forma de polígonos, realiza-se um processo de regionalização em que são unidos os polígonos contíguos e com características consideradas iguais, tal como foi explicado na secção 2.4.2.

Em ambos os tipos de generalização as métricas têm um tratamento especial, dependendo de terem sido criadas hierarquias para elas ou não. No caso de haver hierarquias, o valor das métricas é generalizado para um dos conceitos criados (segmentos que representam intervalos entre os valores) e, a partir daí, a métrica passa a ser tratada como um atributo semântico. Ao ser tratada como tal, é tida em conta para a identificação de tuplos idênticos, entrando na formação da característica. No caso de não ter sido criada uma hierarquia, permanecendo o campo respeitante à métrica com valores numéricos, o campo relativo a métrica não é tido em conta no momento da comparação entre tuplos. Quando são encontrados tuplos idênticos é aplicado o operador de agregação definido pelo utilizador na fase inicial para fazer a agregação dos valores. Nestes casos, e dada a possibilidade que existe de agregar o valor das métricas para as características iguais, cria-se uma nova base de caracterização: escolher a característica predominante no objecto espacial, não de acordo com o número de ocorrências, mas sim, com base no valor final apresentado pela métrica para as várias características. Ao utilizador é então dada a possibilidade de escolher qual dos métodos pretende usar, uma vez que o resultado final poderá acabar por ser diferente, dado que uma característica que tenha um maior número de ocorrências não é obrigatoriamente aquela que mais contribui no âmbito da métrica numérica que estamos a analisar.

3.2.3. Apresentação da relação final

Apresentação da relação final terá em conta os conceitos base do SOLAP+. Dessa forma, tem que se considerar os três elementos base de representação da informação presentes: mapa, tabela de suporte e tabela de detalhe. Terá também que se ter sempre em consideração uma das condições chave do SOLAP+ que é a de manter em todos os casos uma relação de 1:1 entre os elementos presentes no mapa e as linhas apresentadas na tabela de suporte. Já a tabela de detalhe mantem uma relação de 1:N com a tabela de suporte.

No resultado final da generalização de dominância espacial temos um atributo semântico com associação espacial, podendo este estar associado a várias combinações dos restantes atributos semânticos. Uma vez que temos as contagens realizadas das ocorrências de cada um desses tuplos com as diferentes combinações que surgiram, podemos então inferir qual o que tem maior protagonismo para aquele objecto espacial. Assim sendo, a combinação que apresentar um maior números de ocorrências será a escolhida para ser apresentada juntamente com esse objecto espacial na tabela de suporte. O mesmo é dizer que naquele objecto espacial predomina aquela característica. Quando não existe predominância de nenhuma característica, então os campos referentes aos atributos desse objecto espacial são marcados com *indefinido*, e dessa forma o utilizador tem a noção de que não foi encontrada nenhuma predominância. Na tabela de detalhe aparece então a descriminação dos resultados onde será possível ver o número de ocorrências de cada um dos casos para o objecto seleccionado. Assim, na tabela de suporte mantemos a importante relação de 1:1 com os objectos presentes no mapa e, na tabela de detalhe, os valores que fundamentam a caracterização do objecto.

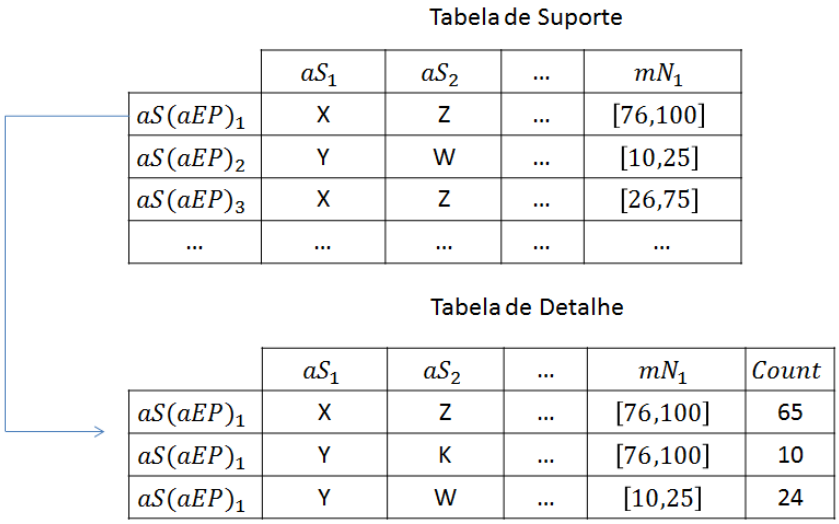


Figura 3.6 - Resultado a apresentar nas tabelas para a generalização de dominância espacial.

O exemplo da Figura 3.6 ilustra que para cada objecto espacial, representado na tabela pelo seu atributo semântico, foi escolhido o conjunto de valores dos atributos semânticos que maior número de ocorrências evidenciou, sendo tal informação apresentada na tabela de detalhe. De notar que no exemplo apresentado foi criada uma hierarquia numérica para a métrica e por isso esta é considerada como um atributo semântico e os seus valores serem o intervalo dos segmentos criados.

No que diz respeito à representação do mapa, apresenta-se na Figura 3.7 a forma como é visualmente reflectida a informação presente na tabela de suporte. Por exemplo, a Figura 3.7 poderia referir-se a um processo de generalização de dominância espacial em que o atributo espacial foi generalizado para o nível *País* da hierarquia espacial.

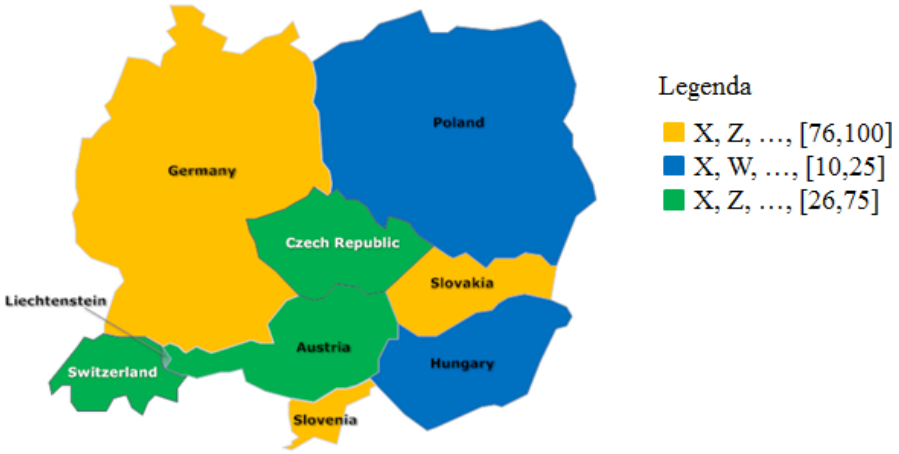


Figura 3.7 - Exemplo de representação no mapa dum processo de generalização de dominância espacial.

No caso da generalização de dominância não espacial, o objecto espacial representado tanto pode corresponder a grupos formados por agrupamento de objectos ou por elementos que não entraram na formação de nenhum dos grupos formados, quer seja por uma questão de proximidade, quer seja por questões de diferenças semânticas. Assim sendo, e como ilustrado na Figura 3.8, na tabela de suporte será apresentado o grupo juntamente com os atributos semânticos e métricas que o caracterizam, e na tabela de detalhe poderemos ver quais os objectos que compõem o grupo. No caso dos objectos espaciais que não entraram na formação dos grupos a sua caracterização também é apresentada na tabela de suporte.

Tabela de Suporte

	aS_1	aS_2	…	mN_1
G_1	X	Z	…	89
G_2	Y	W	…	15
G_3	X	Z	…	57
…	…	…	…	…

Tabela de Detalhe

	aS_1	aS_2	…	mN_1
$aS(aEP)_1$	X	Z	…	30
$aS(aEP)_3$	X	Z	…	24
$aS(aEP)_7$	X	Z	…	35

Figura 3.8 - Resultado a apresentar nas tabelas para a generalização de dominância não espacial.

Na Figura 3.9 pode-se ver a forma de representação no mapa dos dados apresentados nas tabelas da figura anterior. Aqui, os pontos representam os elementos que não se encaixaram em nenhum grupo e os polígonos desenhados reflectem as zonas compostas por objectos com as mesmas características.

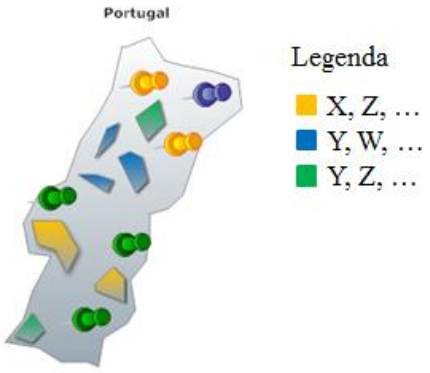


Figura 3.9 - Exemplo de representação no mapa dum processo de generalização de dominância não espacial.

4. Arquitectura

Este capítulo apresenta a arquitectura do protótipo SOLAP+. Primeiro é feita uma apresentação da arquitectura conceptual do sistema, depois é descrito o meta-modelo utilizado para descrever o modelo multidimensional e, em seguida, vai-se ao detalhe de cada um dos grandes componentes do protótipo: servidor e cliente. Por último, demonstra-se como é realizada a comunicação entre as diferentes partes. Nesta secção, sempre que for apropriado, serão apresentadas as alterações realizadas no âmbito deste trabalho para incorporar no SOLAP+ os mecanismos de indução anteriormente discutidos.

4.1. Arquitectura geral

A arquitectura baseia-se na inicialmente definida por Jorge, R. [5] e seguida por Silva, R. [9]. Todos os princípios por eles seguidos foram tidos em consideração ao longo da implementação realizada no âmbito desta dissertação.

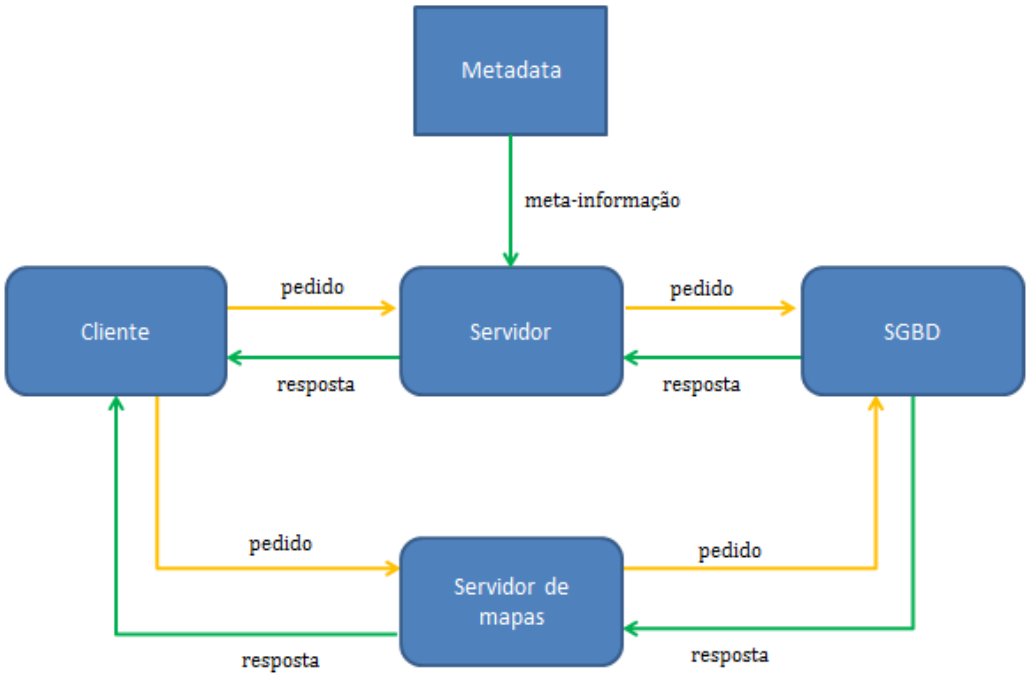


Figura 4.1 - Arquitectura geral e interacções.

Como é ilustrado na Figura 4.1, em termos gerais o sistema baseia-se em 5 componentes: um **cliente** onde é realizada a interacção com o utilizador; um **servidor** que recebe os pedidos e que os processa criando uma resposta; os **metadados** que vão conter a descrição do modelo multidimensional e informação das tabelas que o suportam; um **SGBD** com o qual o servidor comunica para obter os dados sobre os quais se realizam os processos de análise; e um **servidor de mapas** que permite a geração dinâmica de mapas temáticos.

4.2. Meta-modelo

A estrutura do meta-modelo não sofreu grandes alterações desde [9] uma vez que não havia necessidade, e um dos grandes propósitos desta dissertação era incorporar as técnicas apresentadas no modelo multidimensional, tal e qual como já existia e era descrito. Este meta-modelo é descrito através de um ficheiro XML.

O meta-modelo continua dividido em 4 secções, cada uma correspondente a um elemento presente na raiz do documento: *(i) databases*, *(ii) mapservers*, *(iii) styles* e *(iv) multidimensional*. Em seguida faremos apenas uma breve descrição do propósito de cada um dos elementos sendo que em [5] temos uma descrição mais pormenorizada de *(i)* e *(ii)*, e em [9] é feita a descrição de *(iii)* e *(iv)*.

O elemento *databases* contém a descrição das tabelas presentes em base de dados relacionais que formam o modelo multidimensional. O *mapservers* apresenta a descrição do servidor de mapas, contendo informação como os dados necessários para a ligação ao servidor, descrição de elementos contextuais (*layers* ou objectos), e descrição do mapa base sobre o qual serão representados os vários objectos espaciais. A componente *styles* serve para suportar a gestão de estilos de representação no mapa, contendo a descrição de vários estilos e os contextos em que podem ser utilizados. Por último, o elemento *multidimensional* descreve o modelo multidimensional (*star schema* ou *snowflake schema*) através da descrição das dimensões (elemento *dimension*) e dos cubos (elemento *cube*). Foi no elemento *dimension* que foi feita a única alteração à estrutura do meta-modelo, sendo essa descrita na subsecção seguinte.

4.2.1. *Shared borders*

O elemento *dimension* tem como seu descendente um elemento *levels*, formado por elementos *level*. O elemento *levels* contém a descrição dos diversos níveis e referências para o nível base. Cada *level* informa sobre os atributos que o formam, referencia os níveis de granularidade superior, e pode ainda apresentar uma secção chamada *preComputing* introduzida em [9], no caso de ser um nível espacial. Este elemento *preComputing* tem como seus descendentes elementos que descrevem as tabelas que contêm informação pré-calculada, com por exemplo, qual a tabela que guarda o cálculo de distâncias entre os objectos espaciais. Um exemplo do elemento *level* é apresentado em seguida:

```
<level id="14" primaryAttribute="38" displayAttribute="39"
sortAttribute="39" spatialAttribute="40" tableRef="8" name="…">
    <attribute id="38" columnRef="71" name="…"/>
    <attribute id="39" columnRef="72" name="…"/>
    <attribute id="40" columnRef="73" name="… " spatial="true"/>
    <upperLevels>
        <upperLevel levelRef="15"/>
    </upperLevels>
    <preComputing>
        <distances tableRef="20">
            <from columnRef="502" />
            <to columnRef="503"/>
            <distanceValue columnRef="504"/>
        </distances>
        <shared_borders tableRef="22">
            <from columnRef="702" />
            <to columnRef="703"/>
            <shared_border columnRef="704"/>
        </shared_borders>
    </preComputing>
</level>
```

A este elemento *preComputing* foi introduzido o elemento *shared_borders* que tem uma função idêntica ao *distances*. Este elemento descreve a tabela que guarda o cálculo da fronteira partilhada entre os atributos espaciais. O elemento *from* e *to* indicam as colunas que contêm a informação sobre os espaços que estão ser comparados e o elemento shared_border a coluna que contém o valor com informação da fronteira partilhada entre esses dois locais. No caso do valor contido nessa coluna da tabela ser *0* significa que os dois locais em causa não são adjacentes. No caso de não serem adjacentes, não serão agrupados quando forem processados pelo algoritmo de regionalização, mesmo que tenham a mesma característica dominante.

4.3. Servidor

A estrutura base da arquitectura do servidor não sofreu alterações, havendo apenas a inclusão de alguns módulos adicionais para tratar dos pedidos de generalização. A arquitectura do servidor está ilustrada na Figura 4.2.

Quando um pedido é submetido ao servidor é recebido pelo módulo de comunicação e validado através de um XML *schema*. Assumindo que o pedido está correctamente formulado, este é então passado para o módulo de processamento de parâmetros. Este módulo analisa o pedido e retira deste aos parâmetros associados, que serão necessários para o desencadeamento das acções.

Consoante o tipo de pedido, este segue para um de dois módulos: para o módulo de processamento de metadados ou para o módulo de processamento de dados. O primeiro trata dos pedidos que não necessitam de aceder à base de dados utilizando apenas o conteúdo dos metadados (quer seja sobre o modelo, cubos ou estilos) e produzindo por si mesmo uma resposta. O módulo de processamento de dados é responsável por atender todos os outros pedidos (entre eles os de generalização) onde existe a necessidade de interagir com o SGDB. Em qualquer um destes módulos, assim que uma resposta é produzida é passada para o módulo de processamento de parâmetros, sendo por fim enviada à aplicação cliente a resposta por parte do módulo de comunicação.

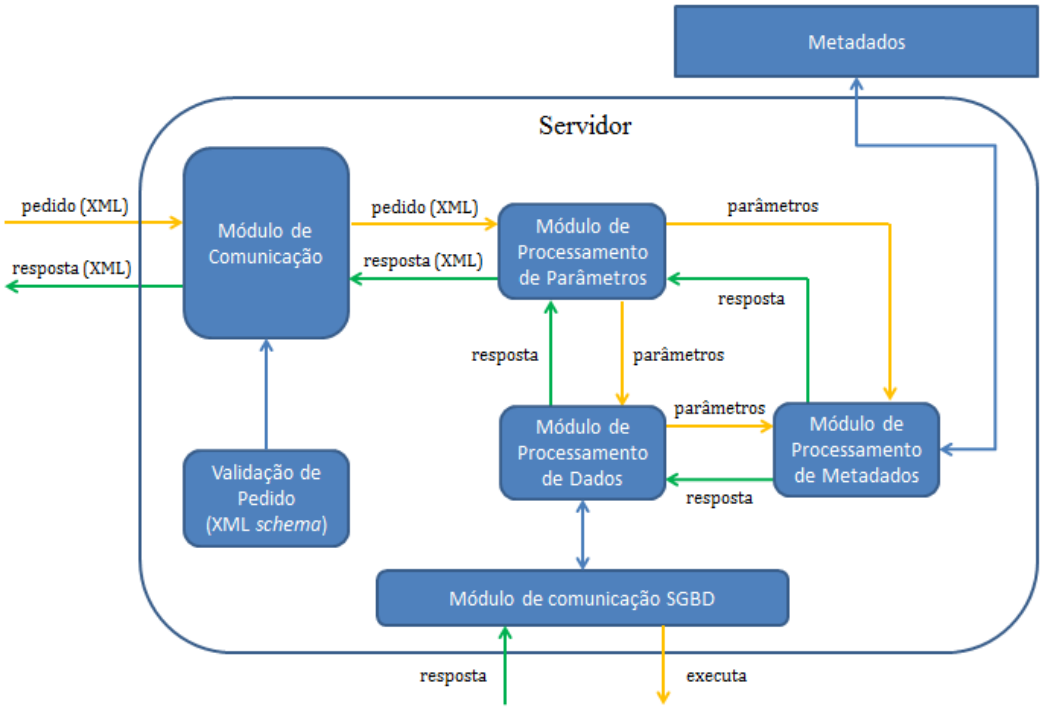


Figura 4.2 - Arquitectura do servidor.

No módulo de processamento de parâmetros houve as necessárias adaptações para incorporar as técnicas de indução, sendo acrescentada a capacidade de processamento e extracção de informação de um novo tipo de pedido com uma estrutura totalmente nova. O módulo de processamento de dados foi o que sofreu alterações mais significativas, sendo por isso descrito em maior pormenor.

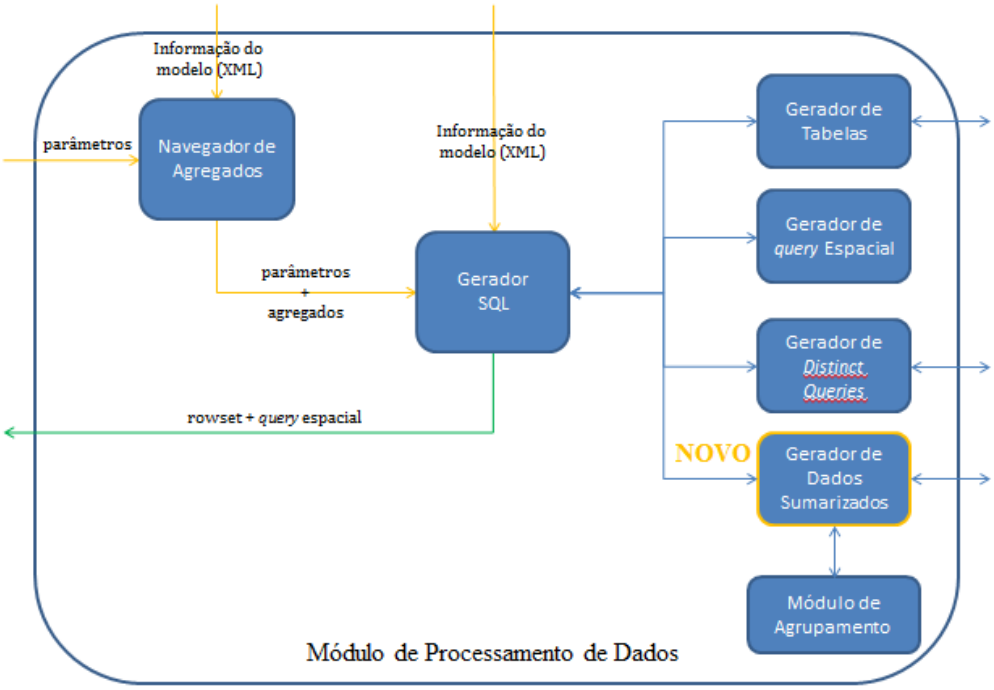


Figura 4.3 - Arquitectura do módulo de processamento de dados.

Na Figura 4.3 apresenta-se a arquitectura do módulo de processamento de dados. O navegador de agregados tem a função de escolher o agregado apropriado e a sua descrição pode ser vista em mais pormenor em [5], uma vez que não sofreu nenhuma alteração. Assim a informação dos parâmetros (mais os agregados dependendo do tipo de pedido) é passada ao gerador SQL. Este tem a finalidade de gerar uma interrogação que será anexada à resposta, e que será utilizada pelo servidor de mapas, de modo a ser construído um mapa com a mesma informação que está na tabela de suporte.

Este gerador SQL está interligado a outros componentes. Foi aqui que foi adicionado um novo módulo, responsável por realizar os pedidos relacionados com a sumarização da informação, denominado de Gerador de Dados Sumarizados. Este é responsável por analisar um *rowset*, realizar os processos de generalização de dominância espacial e não espacial, extrair as características para cada objecto espacial e criar novos *rowsets* para serem incorporados na resposta final. Nos casos, de generalização não espacial este módulo comunica com o módulo de agrupamento de forma a realizar as operações de agrupamento de pontos ou regionalização sobre polígonos.

4.4. Cliente

A aplicação cliente realiza a interacção com o utilizador e envia os pedidos solicitados por este ao servidor. Posteriormente, interpreta as respostas expondo-as quer seja nos menus, quer seja nas tabelas ou mapa. Na Figura 4.4 é apresentada a arquitectura base do cliente.

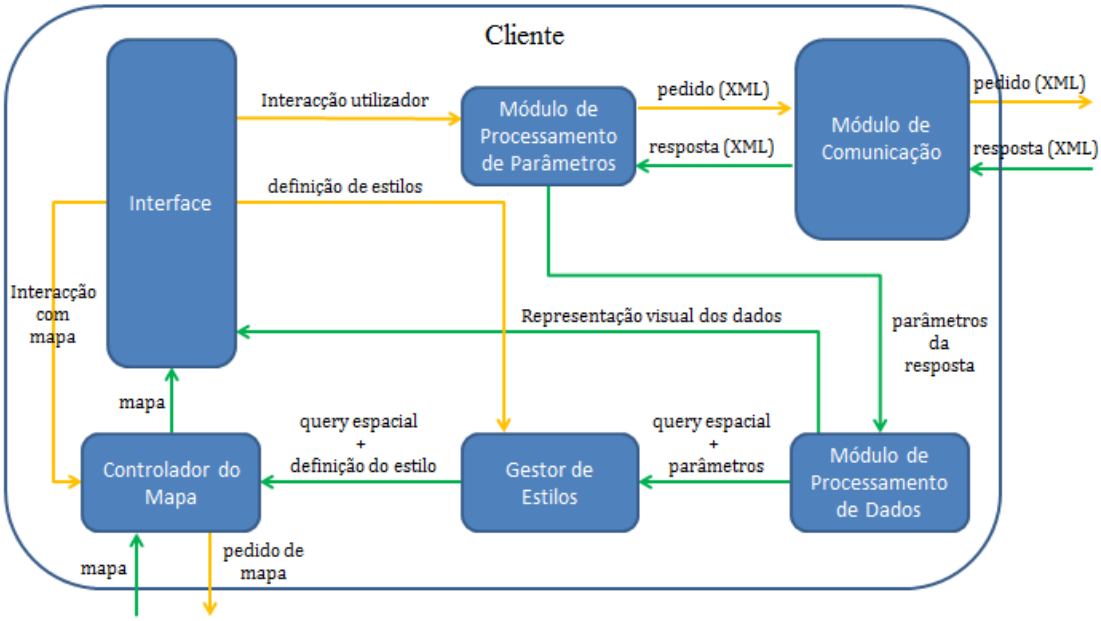


Figura 4.4 - Arquitectura do cliente.

Também aqui existem os módulos de processamento de parâmetros, de dados e de comunicação. O primeiro é responsável tanto por recolher as informações passadas pelo utilizador através da interface do sistema e produzir um pedido XML para ser enviado pelo módulo de comunicação ao servidor, como extrair os parâmetros necessários das respostas que chegam desse mesmo servidor. Esses dados extraídos serão depois encaminhados para o módulo de processamento de dados. Este é responsável pelo tratamento dos dados de forma a serem representados visualmente, quer seja em estruturas de interacção na interface, quer seja nas tabelas de suporte e detalhe. O gestor de estilos define um estilo apropriado para a representação visual dos dados. Esse será enviado ao controlador do mapa e juntamente com a *query* espacial (interrogação responsável pela obtenção dos objectos espaciais a serem representados no mapa), construir um pedido ao servidor de mapas para que seja gerado um novo mapa temático. O controlador do mapa é ainda responsável por tratar a interacção que ocorre com o mapa resultante da interacção do utilizador com a interface como por exemplo, aumentar e diminuir o *zoom*, visualização de legenda, visualização de *labels*, entre outras. Mais pormenores podem ser encontrados em [5] e [9].

Assim sendo, o funcionamento base e composição dos módulos continua a ser o mesmo, tendo sido acrescentadas apenas funcionalidades para gerar e interpretar pedidos de generalização (módulo de processamento de parâmetros), criação das tabelas a partir das respostas a esses pedidos (módulo de processamento de dados) e criação dos estilos visuais a apresentar no mapa para os pedidos de sumarização da informação (gestor de estilos).

4.5. Protocolo de comunicação

Todas as comunicações realizadas entre a aplicação cliente e o servidor seguem o paradigma pedido-resposta. Todos os pedidos seguem a seguinte estrutura:

```
<solapplus>
    <request call="tipo_do_pedido">
    …
    </request>
</solapplus>
```

Os pedidos podem ser de vários tipos: *(i)* obtenção da lista de cubos existentes; *(ii)* selecção de um cubo e carregamento da meta informação associada; *(iii)* obtenção de dados a partir de uma interrogação SOLAP; *(iv)* obtenção de valores distintos existentes para um determinado atributo e *(v)* generalização dos dados provenientes de uma interrogação com os atributos seleccionados. Os dois primeiros são apenas para obter informação que está presente nos metadados, não sendo necessária qualquer interrogação à base de dados. O pedido para obtenção de dados refere-se à realização de uma interrogação com os elementos seleccionados e com as restrições impostas podendo os dados provenientes serem depois alvo de um processo de agrupamento. Já o pedido *(iv)* é utilizado para obtenção de dados relativos aos elementos distintos de um dado atributo para que possam ser escolhidos para a realização de *slices* semânticos.

Os tipos *(i)*, *(ii)*, *(iii)* e *(iv)* não sofreram qualquer alteração vindo do trabalho elaborado por Ruben, J. [5] e Silva, R. [9]. Já o tipo *(v)* foi totalmente criado de raiz, sendo criada uma nova estrutura para os pedidos de generalização, como se demonstra no exemplo seguinte relativo a um pedido de generalização de dominância espacial.

```
<params cubeId="1" filename="emissao.xml"/>
<generalize type="SPATIAL" refine="No" from="All" defineLabels="No"
charMeasures="No" distinct="No" zoomLevel="4">
    <attribute id="13" name="Poluente" dimensionId="2" levelId="7"
threshold="null" >
        <upperLevelAttribute id="13" />
        <upperLevelAttribute id="11" />
        <generalizeAttributeTo id="11" type="semantic" />
    </attribute>
    <attribute id="18" name="Instalacao" dimensionId="3" levelId="8"
threshold="null" >
        <upperLevelAttribute id="18" />
        <upperLevelAttribute id="36" />
        <upperLevelAttribute id="39" />
        <upperLevelAttribute id="42" />
        <generalizeAttributeTo id="42" type="geometric" />
    </attribute>
    <measure id="1" name="Quantidade Emitida" threshold="2" agg="SUM"
createHierarchy="true">
    </measure>
    <facttable id="6" name="emissao_fact"/>
</generalize>
```

A *tag params* tem as indicações gerais sobre o cubo que está a ser utilizado e indicação do nome do ficheiro que contém a meta informação sobre esse mesmo cubo. A *tag generalize* contém a informação referente às opções de generalização e caracterização:

- *Type* – Corresponde ao tipo de generalização. Os valores possíveis são *SPATIAL* ou *NON_SPATIAL*.
- *Refine* – Indicador se é para realizar um refinamento das hierarquias.
- *From* – Indica se a informação que utilizaremos para criar as hierarquias provêem de toda a base de dados ou apenas dos dados da interrogação realizada. Os valores possíveis são *ALL* ou *QUERY*.
- *DefineLabels* – Informa sobre a pretensão do utilizador querer ou não dar nomes aos segmentos criados para as hierarquias geradas.
- *CharMeasure* – Indica se a caracterização é para ser realizada segundo o valor das métricas.
- *Distinct* – Indicação se é para realizar uma interrogação em que o resultado contenha apenas valores distintos.
- *ZoomLevel* – Informação refente ao nível de *zoom* em que o mapa se encontra. Esta informação é utilizada pelo algoritmo de agrupamento de forma a ajustar a formação dos grupos.
- *GroupsParam* – Presente apenas no caso de estarmos perante uma generalização do tipo não espacial, dá-nos a indicação sobre o valor do parâmetro referente à quantidade de grupos criados, o qual será utilizado pelo algoritmo de agrupamento. Os valores (-2, -1, 0, 1 ou 2) são assim indicadores do desejo do utilizador de haver mais ou menos grupos formados. Quanto maior for o valor deste indicador mais grupos são formados sendo estes, contudo, de menor dimensão em termos de elementos que os compõem.

Em seguida vem a informação sobre os atributos, com informação dos níveis que se lhe seguem na hierarquia e para qual será generalizado (no caso de pertencer a alguma hierarquia). No caso de se estar perante um pedido de generalização de dominância não espacial, é incluído um elemento extra no elemento *atribute*, chamado *preProcessing*, que indica através do atributo *sharedBordersPreComputed* se está feita a análise da existência de fronteiras partilhadas entre os elementos daquele atributo e, caso exista, qual a referência da tabela que contém essa informação.

```
<generalize  ... >
    ....
    <attribute id="42" name="Distrito" dimensionId="3" levelId="15"
threshold="null" >
        <preProcessing  sharedBordersPreComputed="true"
tableRef="21"/>
    </attribute>
    ...
</generalize>
```

Quanto às métricas temos a indicação do *threshold* que indica o número de classes que conterá a hierarquia criada, isto no caso de ter sido pedida a sua criação (informação presente na *tag createHierarchy*), ou null no caso de não se pretender geração de hierarquia. Nessa situação, é utilizado para agregação de valores o operador indicado na *tag agg*.

Existem alguns campos e atributos que não estão a ser utilizados devido à forma como todo o algoritmo está implementado mas foram deixados para o caso de mais tarde serem necessários em futuras implementações de novas funcionalidades. Por exemplo, o atributo *threshold* na *tag attribute* é colocado sempre a *null* e não é utilizado uma vez que a escolha do utilizador é auxiliada por intermédio de um *slider* onde se apresentam os vários níveis para os quais se pode generalizar o atributo escolhido, estando a indicação do nível escolhido na *tag generalizeAttributeTo*.

A resposta a um pedido de generalização segue o mesmo esquema que as respostas aos pedidos de obtenção de dados. A diferença está em que aqui é logo enviada a informação referente à tabela de detalhe para que quando for solicitada alguma informação sobre esta não ser necessário realizar novamente todo o processo de generalização para extrair apenas informação complementar.

```
<table count="" nMeasures="">
    <rowset>...</rowset>
    <associatedAttributes>...</associatedAttributes>
    <attributesLevels>...</attributesLevels>
</table>
<table count="" nMeasures="">
    <rowset>...</rowset>
    <associatedAttributes>...</associatedAttributes>
    <attributesLevels>...</attributesLevels>
</table>
   <query sql="..." geometryType="point" numDistict="..."
createdGroups="..."/>
```

O primeiro elemento *table* corresponde à informação para a tabela de suporte, e o segundo à informação a ser utilizada para construir a tabela de detalhe. Cada uma delas tem indicação do número de linhas presentes e do número de métricas presentes. O elemento *rowset* é composto por um conjunto de elementos *row* onde cada um corresponde a uma linha a ser mostrada na tabela. O elemento *associatedAttributes* contém informação sobre os atributos espaciais e o *attributeLevels* sobre os atributos que compõem as linhas do *rowset*, descrevendo também a relação entre estes atributos e os atributos espaciais. Já na parte relativa query espacial, a *tag numDistinct* informa sobre o número de características distintas presentes no documento de resposta para ajudar em efeitos de criação do mapa temático e da legenda: No caso de terem sido criados grupos, a *tag createdGroups* indica o número de grupos criados.

5. Implementação

Este capítulo apresenta os pormenores da implementação dos novos componentes e das alterações aos já existentes, realizadas no protótipo SOLAP+ ao longo desta dissertação.

5.1. Tecnologias

A Figura 5.1 mostra as interacções entre os diferentes componentes e as tecnologias utilizadas por cada um.

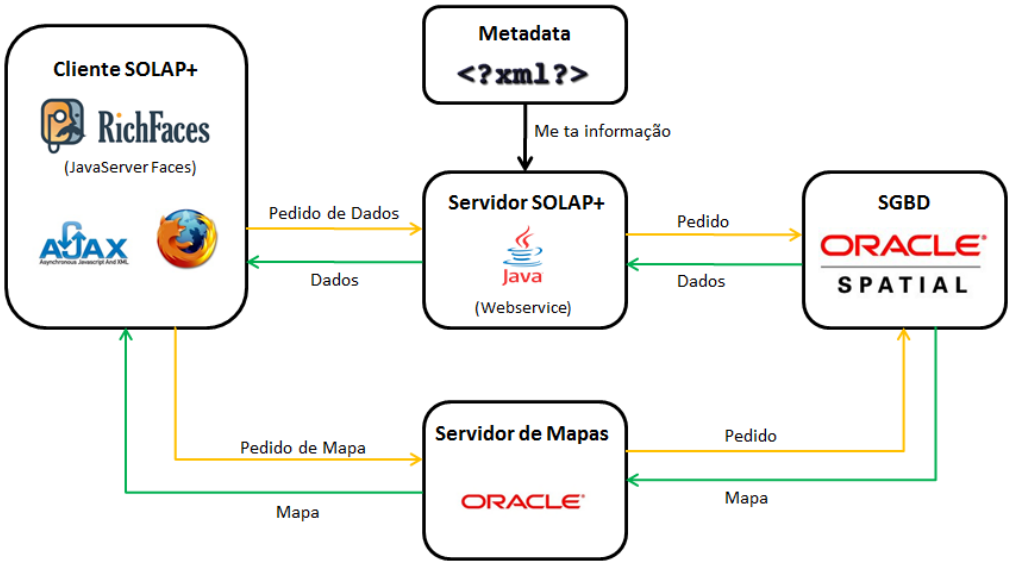


Figura 5.1 - Tecnologias utilizadas nos diferentes componentes do protótipo SOLAP+.

5.2. Servidor

Como foi referido na secção 4.3, foi introduzido um Gerador de Dados Sumarizados à arquitectura do servidor. Em seguida, é feita a descrição do *rowset* tido como parâmetro de entrada e depois realizada a descrição do módulo introduzido em termos do seu funcionamento e implementação.

Antes de se poder fazer qualquer sumarização dos dados temos que ter um *rowset* sobre o qual trabalhar. O *rowset* obtido com a execução de uma interrogação proveniente de um pedido de sumarização contém a estrutura indicada na Figura 5.2. O exemplo ilustrado na Figura 5.2, demonstra os vários atributos presentes no resultado da interrogação, verificando-se a presença tanto do atributo base como do atributo de nível superior na hierarquia para o qual devem ser generalizados os valores. Por exemplo, no caso dos atributos semânticos que representam os objectos espaciais, $aS(aEP)_x$ é o atributo base e $aS(aEP)_X$ é o atributo para o qual deve ser generalizado. O mesmo acontece com $aS_{1.1}$ e aS_{1}, sendo o primeiro o atributo base e o segundo o atributo para o qual deve ser generalizado. Não é obrigatório que o atributo tenha que ser generalizado, podendo estar presente apenas para fins de descriminação, como acontece com aS_2.

	$aS(aEP)_x$	$aS(aEP)_X$	$aS_{1.1}$	aS_1	aS_2	aS_3	$aS_{3.1}$	...	mN
1	...	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...	...
3	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...

Figura 5.2 - Estrutura do rowset inicial.

Dado um *rowset* correspondente a nossa relação inicial e o conjunto de parâmetros indicados pelo utilizador, o processo de indução com vista a extracção de características espaciais segue um conjunto de fases, tal como é ilustrado na Figura 5.3.

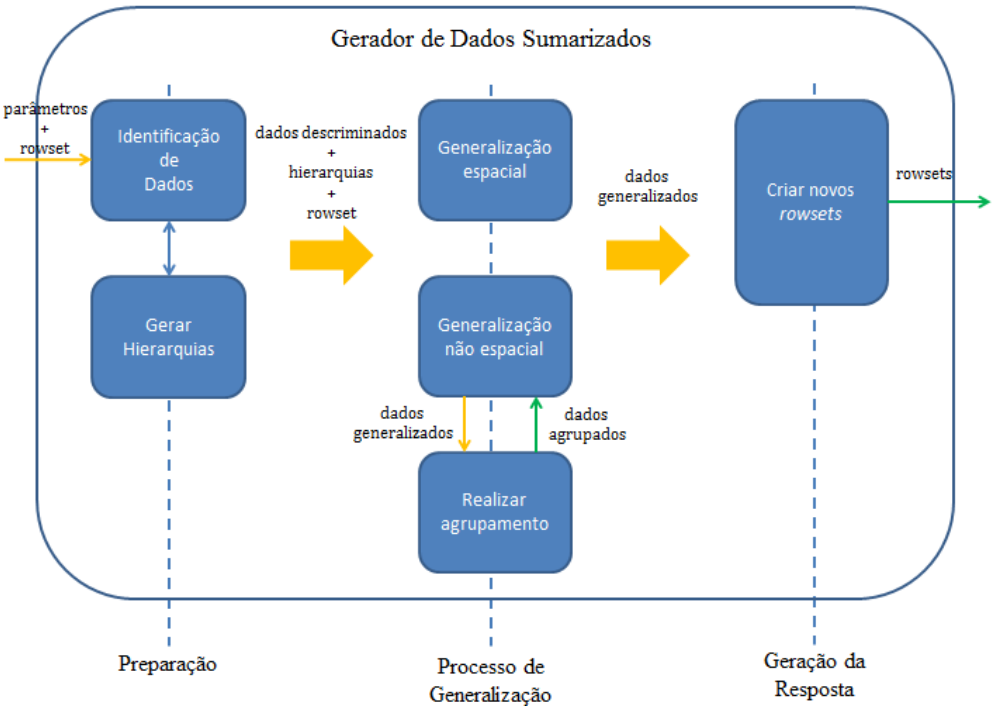


Figura 5.3- Fases que compõem o processamento de dados para sumarização e caracterização.

A primeira fase corresponde a uma preparação dos dados e das estruturas que serão utilizadas durante o processo. Aqui são analisados os parâmetros passados e identificados quais são os atributos espaciais e semânticos, quais os que são para generalizar e para que atributos generalizar. Após obter essa informação, poderá ser ou não necessário gerar hierarquias numéricas, responsáveis para criar as hierarquias referentes às métricas. Para a criação dessas hierarquias foram implementados os dois métodos indicados na secção 2.3.1.

No final desta fase tem-se a informação sobre qual o tipo de cada um dos atributos presentes no *rowset* e, caso necessário, as hierarquias numéricas para enquadrar os valores das métricas. Tendo esta informação presente podemos então iniciar o processo de indução de qualquer tipo (dominância espacial ou não espacial). Existe um ponto em comum em ambos os processos de generalização, que corresponde à contagem das diferentes características evidenciadas em cada um dos objectos espaciais identificados pelo atributo semântico correspondente ao atributo espacial.

No caso de estarmos perante uma generalização de dominância não espacial, após terem sido recolhidas todas as ocorrências evidenciadas para cada um dos objectos espaciais, os dados são fornecidos ao módulo de agrupamento que aplica o algoritmo apropriado para tentar formar grupos que partilhem a mesma característica dominante. Os algoritmos aplicados podem ser o *DBSCAN*, no caso dos objectos espaciais que estamos a caracterizar corresponderem a pontos, ou o algoritmo de regionalização para quando se tratam de polígonos.

Existem várias formas de decidir qual é a característica dominante para determinado local. Uma das possíveis, e que se encontra actualmente implementada no protótipo SOLAP+, é analisar as contagens e a característica que tiver maior número de ocorrências será a considerada dominante. Em caso de empate classifica-se o local como indefinido. Contudo, outras soluções existem para determinar se uma característica é dominadora em relação às outras. Por exemplo, poderia só ser considerada dominante, uma característica com uma percentagem de ocorrência naquela local superior a um valor pré-definido ou até imposto pelo utilizador. Por outro lado, poderia também levar-se em consideração a proximidade em termos de número de ocorrências entre as diversas características encontradas para o local em causa e somente considerar dominante a que se superiorizasse às outras por um determinado valor, que também poderia ser definido pelo utilizador.

A última fase deste processo de sumarização e caracterização dos objectos espaciais corresponde à criação dos *rowsets* com os novos dados que irão integrar a resposta final a ser enviada para o cliente. São gerados logo dois *rowsets*: um com a informação para tabela de suporte e outro para a tabela de detalhe. O primeiro *rowset*, correspondente à tabela de suporte, vai conter para cada objecto espacial o valor dos atributos que formam a característica dominante naquele objecto. Já a informação contida no segundo *rowset*, que engloba os dados para colocar na tabela de detalhe, varia conforme o tipo de generalização. No caso da dominância espacial, cada objecto espacial terá tantas linhas presentes quantas o número de características que se evidenciaram naquele local, com informação do número de ocorrências de cada uma delas. Quando estamos perante uma generalização de dominância não espacial, este *rowset* contém a informação dos elementos que formam cada um dos grupos criados.

5.3. Cliente

No que diz respeito à aplicação cliente, não houve nenhum módulo construído de raiz, uma vez que houve sempre a preocupação de tentar incorporar nas estruturas já existentes os dados referentes a estes novos pedidos. Nas subsecções seguintes explica-se as alterações efectuadas tanto a nível da interface como a nível do processamento das respostas enviadas pelo servidor aos pedidos de sumarização.

5.3.1. Interface

À interface que foi apresentada anteriormente na Figura 2.5, foi adicionado um painel de controlo para os processos de sumarização. Através da interface produzida, o utilizador escolhe os atributos e métricas que pretende aplicar para realizar o processo de generalização e com base neles caracterizar o atributo espacial escolhido. A sequência de passos está ilustrada na Figura 5.4.

No painel *i* da Figura 5.4 estão os atributos que foram escolhidos para fazerem parte do processo de sumarização. Estes atributos foram escolhidos a partir do painel correspondente ao modelo multidimensional, arrastando os atributos ou métricas pretendidas, mantendo a mesma filosofia já presente na interface, por exemplo, para a definição de *slices*.

Escolhidos os atributos, e ainda antes de poder fazer o pedido ao servidor, é necessário o utilizador fornecer mais algumas informações relativamente aos elementos seleccionados. Essas indicações são dadas no painel *ii* da Figura 5.4. Nesse painel temos a definição das hierarquias a utilizar para os atributos *Poluente* e *Instalação*, sendo que no primeiro não existe escolha a efectuar e, por isso, é automaticamente seleccionada a única hierarquia existente. Já o atributo *Tipo* não está inserido em nenhuma hierarquia sendo essa informação transmitida ao utilizador. No caso da métrica é dada a opção de criar uma hierarquia numérica para a métrica em causa caso o utilizador assim o deseje.

Por último, no painel *iii* apresentado na Figura 5.4, é feita a selecção do nível de generalização se pretende para cada um dos atributos que tem hierarquias. Essa indicação é fornecida pelo utilizador com ajuda de um *slider* que possibilita navegar nos níveis da hierarquia desde o nível do atributo que foi seleccionado (elemento mais a esquerda no *slider*) até ao topo da hierarquia (elemento mais à direita). No exemplo presente na Figura 5.4, ainda se ilustra a indicação do número de classes que se pretende criar para a métrica escolhida. No final destes passos, é gerado o pedido XML, como apresentado na secção 4.5.

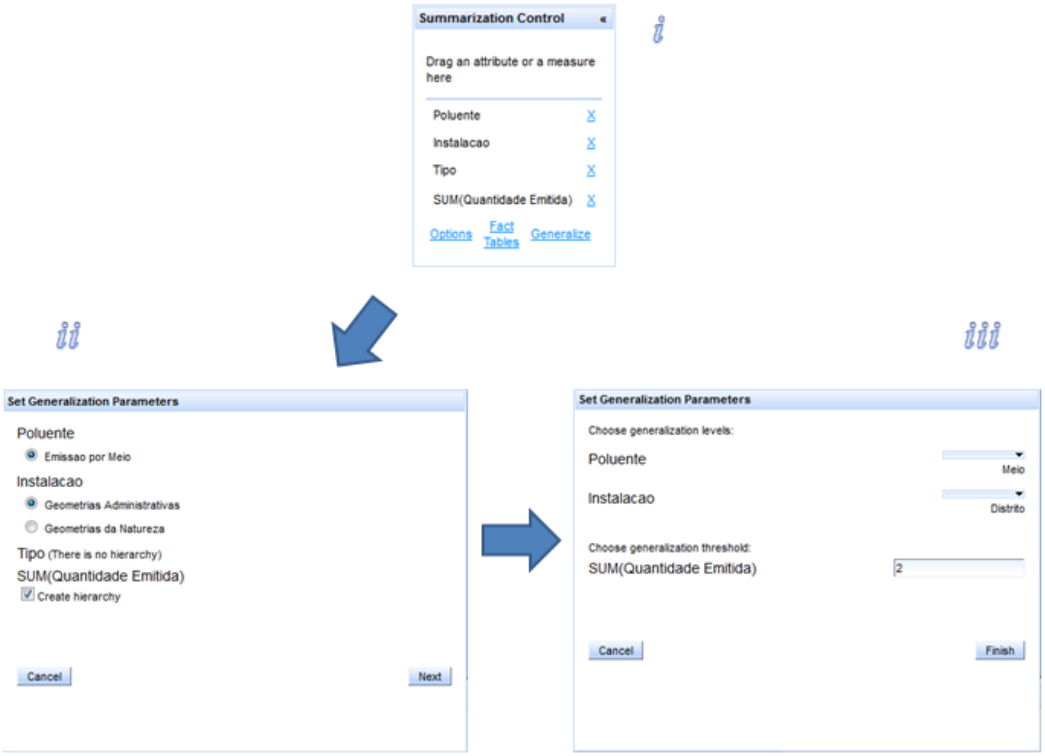


Figura 5.4 - Sequência de painéis para realizar um processo de sumarização.

Quando são escolhidos atributos de diferentes dimensões, é necessário indicar qual a tabela de factos que faz a relação entre essas dimensões. Nessas condições, surge uma ligação no painel *i* presente na Figura 5.4 com a designação de *Fact Tables* que quando activada abre um novo painel onde é possível realizar a escolha da tabela de factos a utilizar. No caso de haver só uma tabela de facto em todo o modelo, essa tabela é automaticamente seleccionada como sendo a tabela a utilizar, tal como acontece no exemplo ilustrado na Figura 5.5.

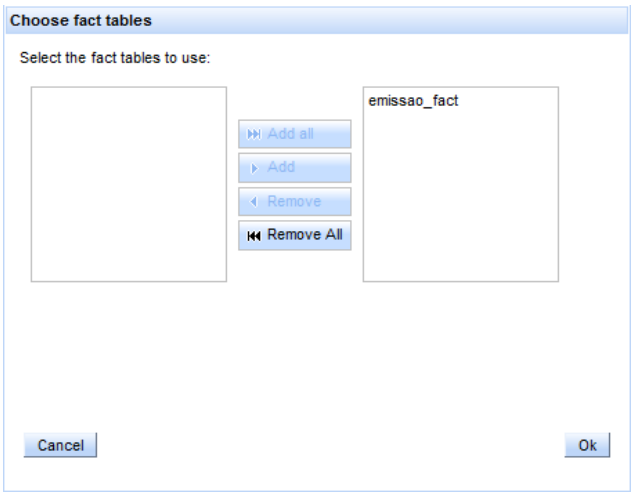


Figura 5.5 - Painel de escolha de tabelas de factos.

Existe ainda o painel de opções (Figura 5.6) onde são definidos os parâmetros dos processos de generalização e as bases para a caracterização espacial. De notar que aqui se inclui a definição do parâmetro relativa ao algoritmo de agrupamento no caso de ser utilizada a generalização de dominância não espacial. Também aqui é adoptado um *slider*, navegando da esquerda para a direita como forma de indicação que se pretende um menor ou maior número de grupos, respectivamente.

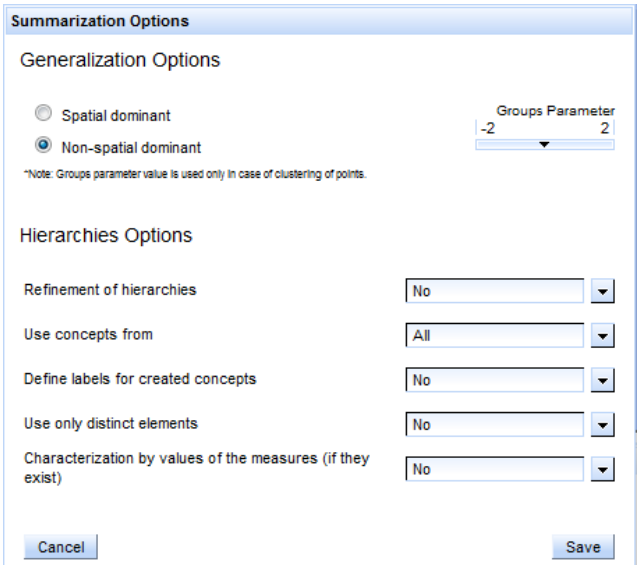


Figura 5.6 - Painel de opções de sumarização.

5.3.2. Processamento das respostas aos pedidos de sumarização

Quando uma resposta a um pedido de sumarização chega ao cliente, a primeira coisa que é processada são os dados que se destinam à tabela de suporte. Uma vez que nas respostas provenientes de pedidos de generalização vêm tanto os dados para a tabela de suporte como os dados para a tabela de detalhe, é extraída apenas a parte referente à tabela de suporte para ser processada. Contudo, todo o conteúdo da resposta é guardado para o caso de ser solicitada a apresentação dos detalhes. Quando essa solicitação acontece, extrai-se apenas essa secção da resposta. Dessa parte, são depois retirados apenas os dados que dizem respeito ao objecto espacial que foi seleccionado. Uma vez passados esses dados para o *extractor* (elemento responsável pelo processamento e extracção dos dados das respostas), o processo é realizado como qualquer outro pedido no que diz respeito à apresentação dos dados na tabela respectiva.

Uma vez realizada esta etapa referente ao preenchimento da tabela de suporte, passamos à parte de criação do XML referente à definição do estilo para o mapa temático a ser criado. Este processo é todo desempenhado pelo módulo de gestor de estilos. Aqui é construído um contexto que levará à criação do estilo (definição da forma como os objectos espaciais serão representados no mapa, cor com que serão identificados, entre outros).

Nesta fase de definição de estilos intervém um novo elemento criado, uma cache de estilos. Anteriormente, em duas análises seguidas utilizando os mesmos atributos, poderia acontecer que objectos espaciais definidos pela mesma característica dominante tivessem cores diferentes. Isto acontecia devido ao facto de não haver controlo sobre a ordem de definição de cores para as diferentes características que eram identificadas. Para resolver esse problema, foi criada esta cache que guarda a informação referente à associação realizada durante o último processo de sumarização entre cores e características. A cada processo esta cache é utilizada durante a fase de definição de estilos e, no final, actualizada com os novos valores. Esta cache é importante para facilitar a análise dos mapas temáticos na medida em que torna mais perceptíveis as diferenças existentes entre essas análises no que diz respeito à caracterização espacial.

6. Casos de estudo

Este capítulo apresenta os resultados obtidos com os dois casos de estudo com o objectivo de validar as propostas realizadas nesta dissertação. Esta avaliação limita-se a considerar os dados reais de cada um dos casos de estudo e a ilustrar a utilização do SOLAP+ no que diz respeito aos aspectos de generalização desenvolvidos neste trabalho.

6.1. Emissão de poluentes

Neste caso de estudo o objectivo era analisar quais seriam as características evidenciadas por diferentes locais em relação à emissão de poluentes em Portugal e dessa forma testar os dois tipos de generalização.

Os dados analisados dizem respeito a instalações, cuja localização é definida por coordenadas. Estas instalações realizam várias actividades industriais (sendo uma dessas actividades considerada principal) através de um ou mais processos industriais. Como consequência da sua actividade, emitem poluentes para o ambiente tanto por ar como por água (directa ou indirectamente), estando a quantidade emitida por cada emissão registada.

As informações relevantes sobre o modelo de dados para este caso de estudo são as seguintes:

- O modelo contém uma dimensão espacial designada *Instalação* que contém 5 atributos espaciais: *localização*, *bacia hidrográfica*, *freguesia*, *concelhos* e *distrito.*
- Existem duas hierarquias espaciais na dimensão *Instalação*:
    - *Instalação -> Bacia Hidrográfica*;
    - *Instalação -> Freguesia -> Concelhos -> Distrito.*

De modo a analisar a generalização de dominância espacial e avaliar a caracterização espacial que se consegue obter depois de realizado esse processo, nos exemplos que se seguem a base de análise será formada pelos seguintes dois atributos: *Instalação* (atributo espacial com hierarquia) e *Poluente* (atributo incluído numa hierarquia semântica).

No exemplo da Figura 6.1 é apresentado o resultado obtido utilizando somente estes dois atributos, generalizando o atributo *Instalação* para *Distrito* e mantendo o atributo *Poluente*. Foi efectuado um *slice* semântico escolhendo apenas alguns dos poluentes dos quais existe registo, de forma a não misturar emissões de diferentes tipos de poluente que não estão relacionados.

Na figura em causa estão presentes os três elementos base do protótipo: mapa (onde se visualiza a caracterização espacial com base nos atributos semânticos), tabela de suporte (apresenta os dados que dão suporte à informação visual presente no mapa), e tabela de detalhe (onde se mostra os detalhes que levaram a tal caracterização).

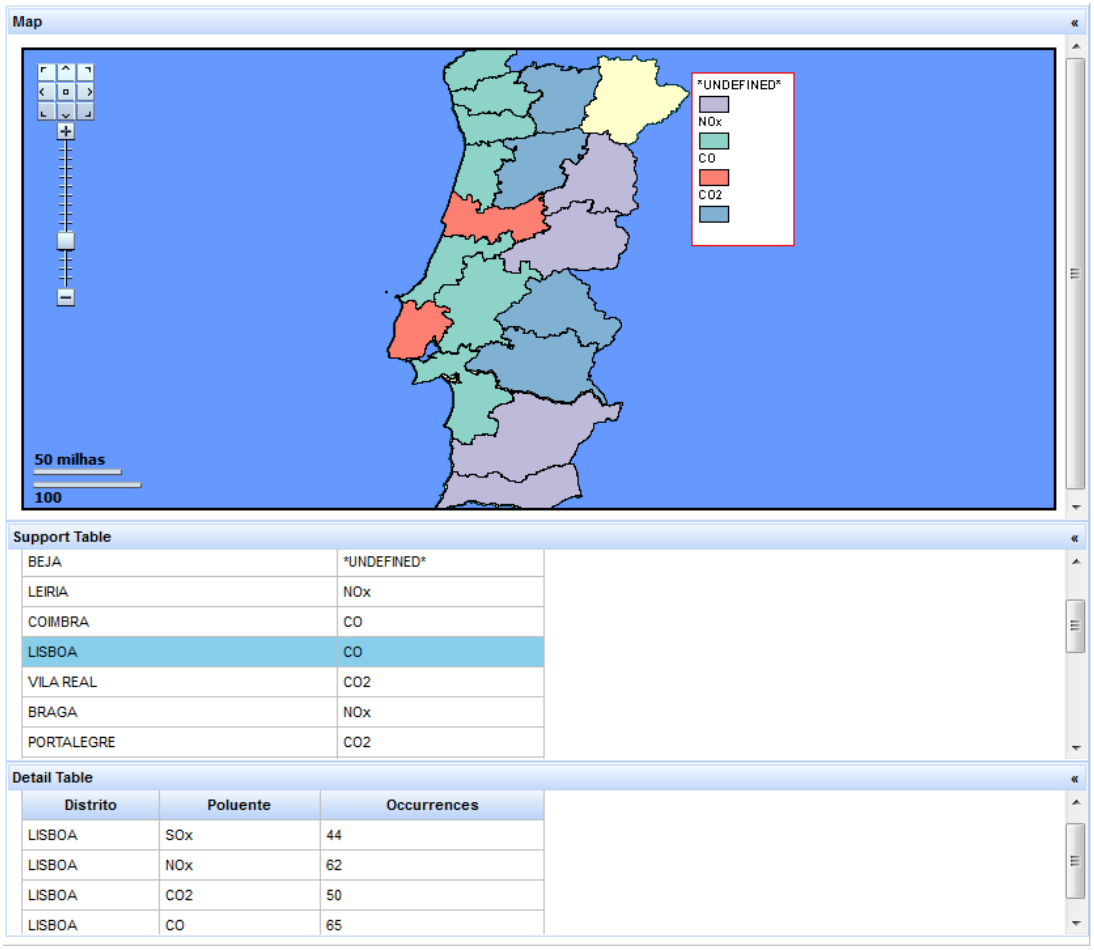


Figura 6.1 - Resultado obtido com a generalizaçao dos atributos *Instalação* e *Poluente*.

Este é um caso em que o utilizador está interessado em saber qual o poluente que é mais vezes emitido por distrito (não em termos de quantidade mas sim somente baseado no número de emissões registadas). Como demonstra a Figura 6.1, o protótipo neste momento tem capacidade de responder a este tipo de questões indicando, por exemplo, que no distrito de *Lisboa* predominam as emissões de *Monóxido de Carbono*. O distrito de *Bragança* aparece sem nenhuma das cores presente na legenda pois não existem dados para aquela região e, dessa forma, permanece com a cor original que apresentava o mapa. Ainda se verifica que como resultado deste processo surgiram elementos, como por exemplo *Beja*, que têm uma classificação de *Undefined* para os atributos semânticos Estes são objectos espaciais onde nenhuma característica se superiorizou a todas as outras quanto ao número de ocorrências.

Este tipo de resultados não era possível obter utilizando o protótipo SOLAP+ como se encontrava anteriormente à realização desta dissertação. Agora o sistema tem a capacidade de definir que em dada região predomina determinado facto, tendo noção que para ter esse tipo de resultado se perde detalhe.

Em seguida serão demonstrados os resultados obtidos com a inclusão de métricas nos elementos que formam a relação inicial. Foi adicionada a métrica *Quantidade Limiar*, que corresponde à quantidade emitida expressa em limiares, com o operador *SUM*, sendo o resultado obtido apresentado na Figura 6.2. O mapa apresenta uma caracterização espacial igual à da Figura 6.1 uma vez que a decisão da característica dominadora para cada espaço continua a ser realizada através do número de ocorrências. Nesta análise pretende-se visualizar, para além do poluente que mais vezes é emitido para cada região, também a sua quantidade expressa em limiares. De notar ainda, que na linha referente ao objecto considerado como indefinido, não é apresentado qualquer valor no campo relativo à métrica.

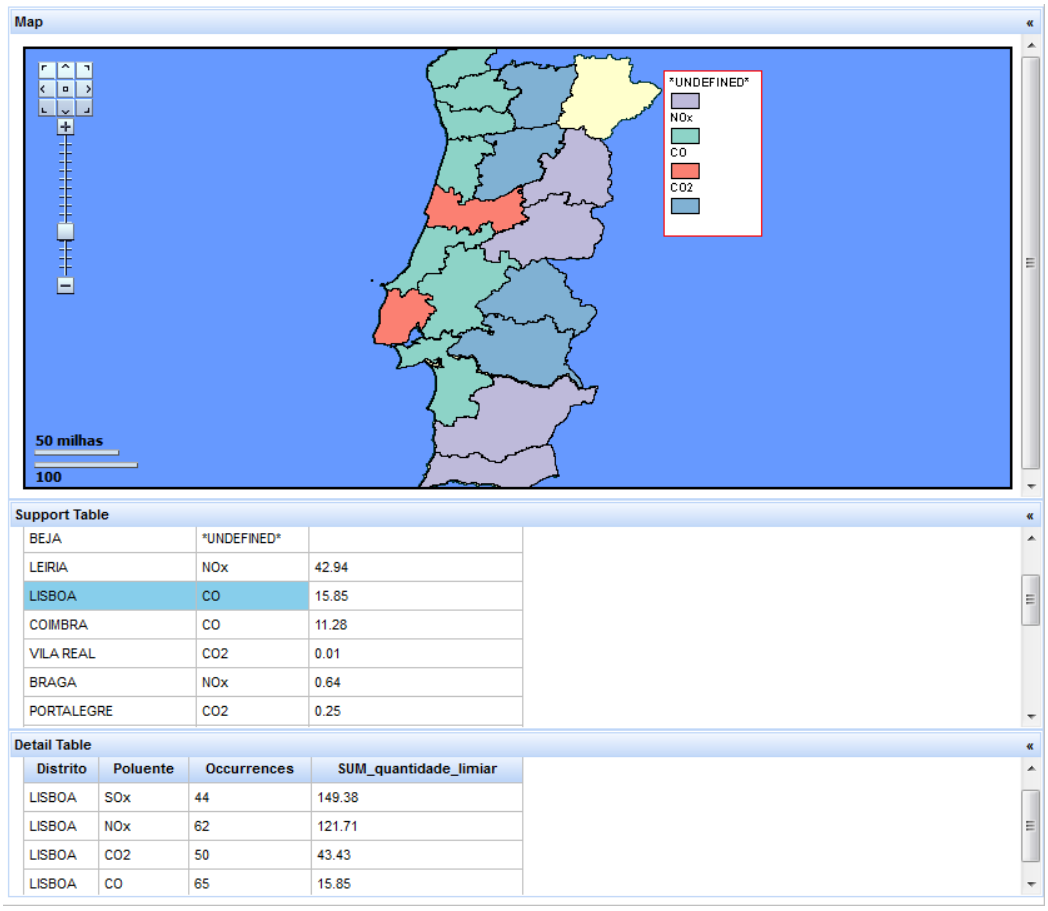


Figura 6.2 - Generalização de dominância espacial com os atributos *Instalação*, *Poluente* e a *Quantidade Limiar* como métrica.

Ao analisar-se os vários locais com recurso a informação da tabela de detalhe verifica-se a existência de casos como o de *Lisboa* onde apesar de haver um maior número de ocorrências de emissões de *CO*, essas emissões são numa quantidade expressa em limiares muito menor do que as de outros poluentes. Indicando que se pretende que a caracterização seja feita com base no valor da métrica, obtemos resultados iguais em termos de detalhe mas a escolha da característica dominante é alterada, sendo esse facto reflectido no mapa como pode ser visualizado na Figura 6.3. Aqui, a pergunta a que se pretende responder é qual o poluente que é emitido em maior quantidade (expressa em limiares) para cada distrito.

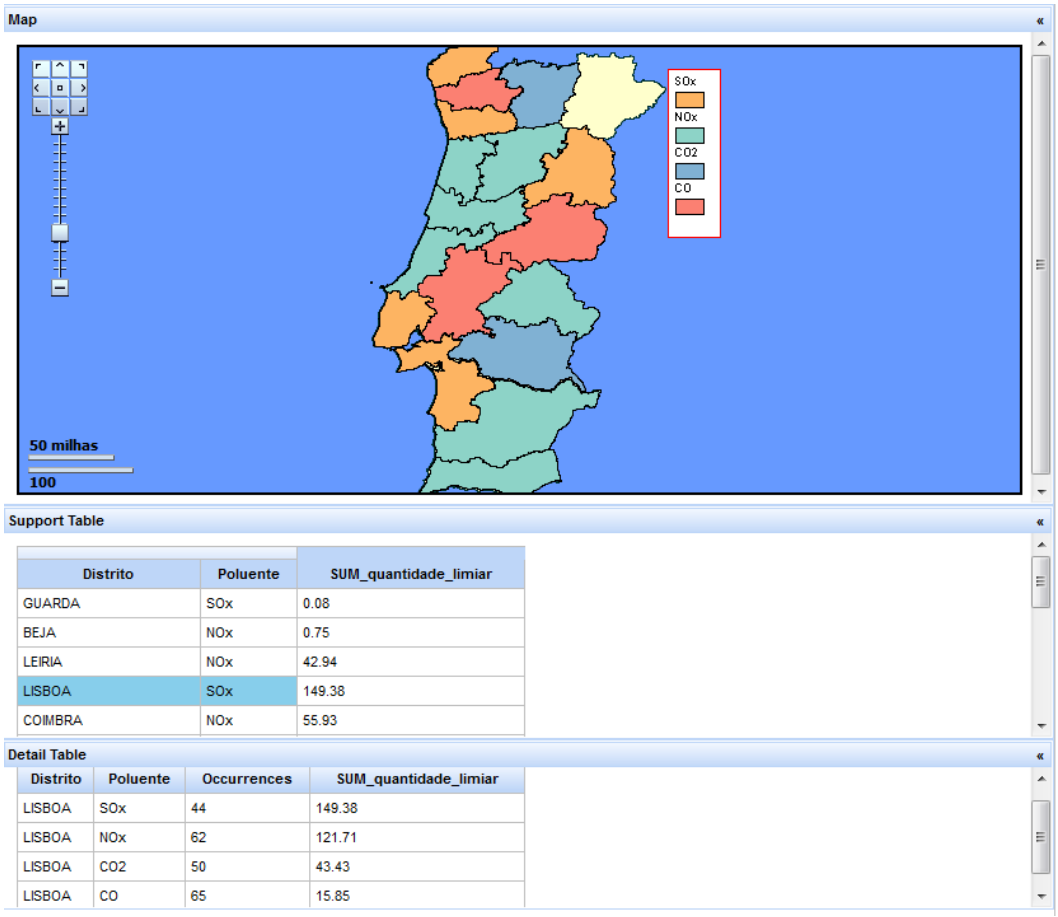
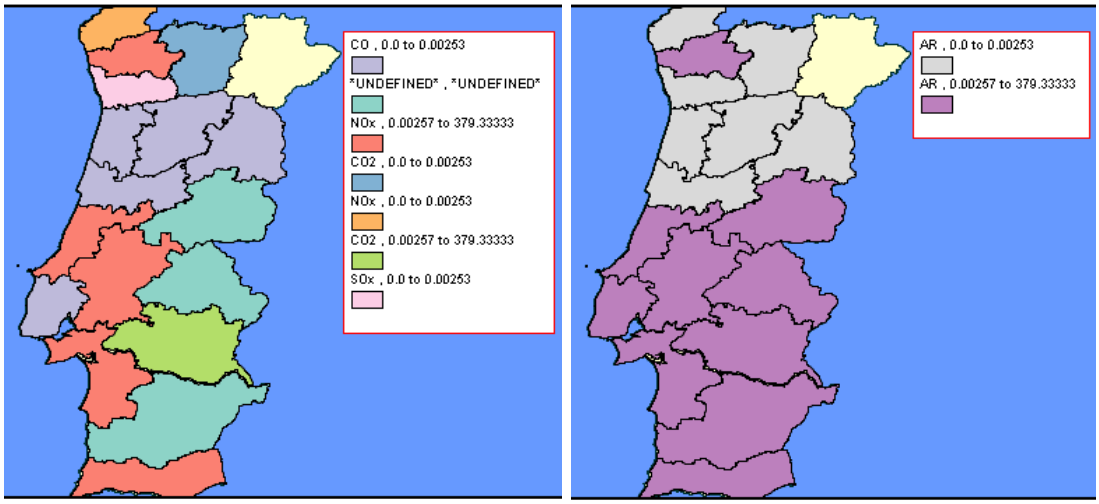


Figura 6.3 - Mapa resultante da caracterização com base no valor da métrica.

No caso de ser solicitada a criação de uma hierarquia numérica para a métrica escolhida, esta vai ser tratada com um atributo semântico entrando para constituição das características, como pode ser visto na Figura 6.4 a) que corresponde a uma análise efectuada exactamente com os mesmos atributos que deram origem ao mapa anterior: *Instalação*, *Poluente* e *Quantidade Limiar*. O *threshold* definido para a hierarquia foi de 2, sendo assim criadas duas classes para encaixar os valores. Como se vê, pela análise do mapa a generalização não foi muito eficiente para a extracção de características para a maior parte dos locais. Assim sendo, decidimos fazer uma generalização mais profunda, isto é, generalizar também o atributo *Poluente* para o seu *Meio*. Desta forma estamos a mudar o sentido da análise para saber qual o meio segundo o qual são realizadas mais emissões de poluentes e classificando as *Quantidade Limiar* em duas classes distintas. Como se pode ver pela Figura 6.4 b), que representa o resultado do mapa para a análise indicada, agora já existe uma classificação de um maior número de *Distritos* comprovando assim que quanto mais se sobe nas hierarquias de conceitos mais fácil é a extracção de características dominantes para os locais analisados.



a) b)

Figura 6.4 - Resultados obtidos com a geração de hierarquia para métrica, a) sem generalização do atributo *Poluente* e b) com generalização do mesmo para o *Meio*.

Em seguida, é ilustrada a partir da Figura 6.5 a generalização de dominância não espacial, utilizando como atributos iniciais a *Instalação* e o *Poluente*, optando-se agora pela generalização do atributo semântico *Poluente* para o *Meio*. Aqui, não há generalização do atributo espacial e o parâmetro referente ao algoritmo de agrupamento *DBSCAN* foi definido com valor *0*, indicando que pretendemos um número equilibrado de grupos, isto é, nem muitos grupos com poucos elementos, nem poucos grupos que englobem um número elevado de objectos, neste caso, de instalações. Verifica-se que a criação de grupos gera regiões que não estão pré-definidas, ao contrário do que acontecia na generalização de dominância espacial onde o atributo espacial era generalizado para um outro atributo espacial de nível superior, como era o caso dos distritos. Verifica-se ainda que os grupos se podem interceptar e criar zonas de indefinição de dominância e podem depois ser alvo de análise por parte do utilizador através da verificação individual de cada um dos elementos dos grupos e suas características.

Figura 6.5 – Mapa obtido através do processo de generalização de dominância não espacial sobre pontos.

Por último, apresentamos na Figura 6.6 o mapa obtido com a generalização de dominância não espacial, onde foram escolhidos também o atributo *Poluente* e a métrica *Quantidade Emitida* mas, desta vez, actuando sobre objectos que são representados por polígonos, como é o caso do *Distrito*, ao invés da *Instalação*. Verifica-se que a base do mapa é idêntica à da Figura 6.4 b) com a diferença de que os polígonos foram alvo do algoritmo de regionalização e assim formam polígonos únicos que contêm as mesmas características dominantes, cada um deles representando um grupo de distritos. Isto encontra-se ilustrado na Figura 6.6 através do grupo que esta identificado como *Grupo 1*, do qual fazem parte vários distritos, como por exemplo, *Guarda* e *Aveiro*.

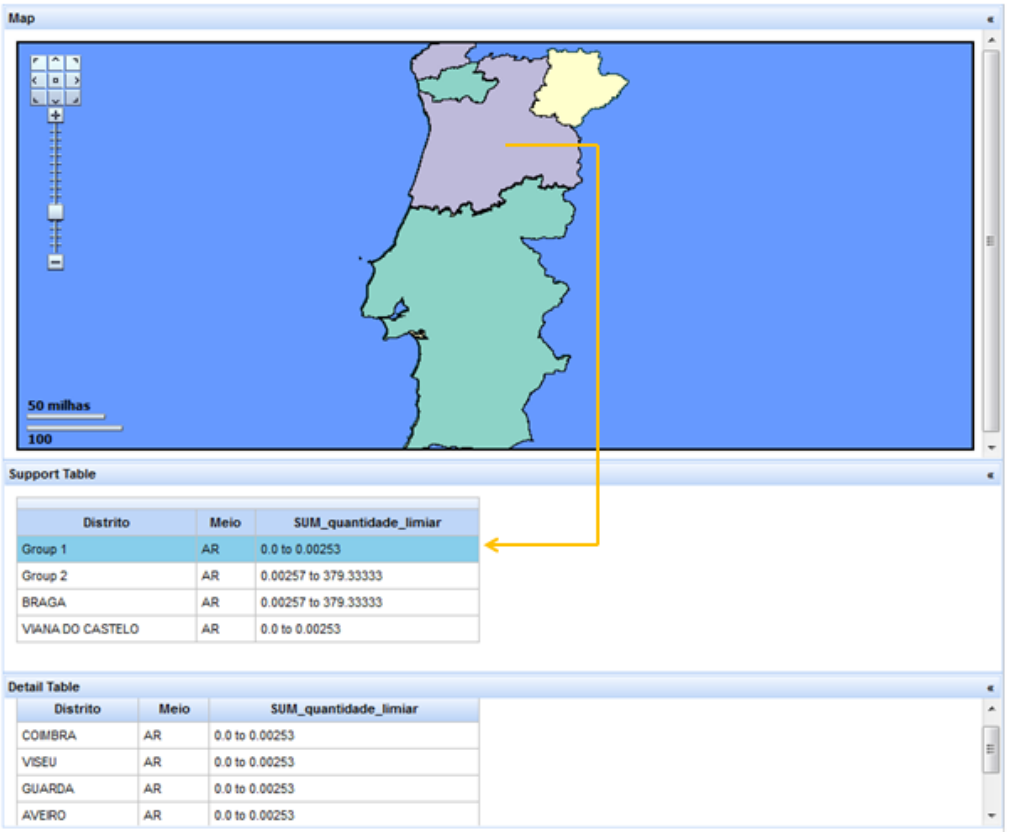


Figura 6.6 – Dados obtidos por generalização de dominância não espacial sobre polígonos.

6.2. Inquérito da Fundação Portuguesa do Pulmão

Neste caso de estudo, as análises são realizadas sobre dados fornecidos pela Fundação Portuguesa do Pulmão, uma organização sem fins lucrativos que realiza várias actividades para recolher, guardar e analisar dados referentes a diversas doenças. Com este caso de estudo pretende-se demonstrar que o protótipo SOLAP+ permanece genérico e capaz de realizar os novos tipos de análise implementados sobre dados de vários domínios.

Os dados sobre os quais se trabalha neste caso de estudo são referentes a um inquérito de âmbito nacional realizado pela Fundação Portuguesa do Pulmão em 2007. Deste inquérito faziam parte 27 questões, todas relacionadas com sintomas e doenças respiratórias, sendo inquiridas 1898 pessoas durante o processo.

As informações relevantes sobre o modelo de dados para este caso de estudo são as seguintes:

- O modelo contém uma dimensão espacial designada *Paciente* que contém 3 atributos espaciais: *freguesia*, *concelhos* e *distrito*. Existe uma hierarquia que engloba esses atributos espaciais:
  - *Freguesia -> Concelhos -> Distrito*
- O resto das dimensões são caracterizações do estado do paciente em termos de alergias, cansaço, actividade de fumador, tosse e doenças pulmonares. De referir que em algumas destas dimensões existem hierarquias semânticas envolvendo alguns dos seus atributos.

Um dos casos que podemos analisar é por exemplo a incidência dos casos de *Tosse Perlongada* pelos diferentes *Distritos*. Para isso, realizou-se um *slice* com a finalidade de analisar somente os casos que apresentam *Tosse Seca* e, dessa forma, determinar quais os distritos onde existe dominância de fenómenos de *Tosse Perlongada*. A resposta a esta pergunta está ilustrada na Figura 6.7, através da qual se concluí que, na maioria dos distritos onde se verificam ocorrências de tosse, essa tosse não é perlongada. Conclui-se também que os fenómenos de tosse se verificam mais no litoral e sul de Portugal, uma vez que não existem fenómenos de tosse registados nas regiões do interior do norte e centro do país e por isso não puderam ser caracterizadas como regiões de tosse predominantemente perlongada ou de curta duração.

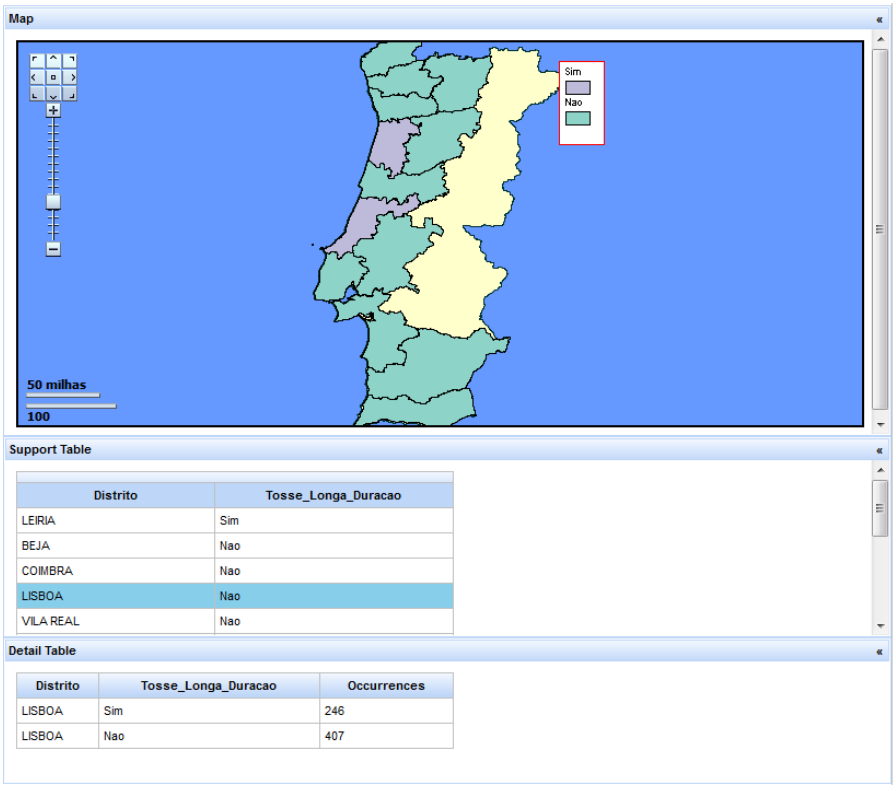


Figura 6.7 - Análise de fenómenos de *Tosse Perlongada* no país.

Um outra situação que pode ser analisada é a faixa etária das pessoas que são fumadoras e das quais existe registo de alguma doença pulmonar nos dados recolhidos. Para isso são escolhidos como atributos iniciais da generalização a *Freguesia* (atributo espacial mais baixo da hierarquia, sendo generalizado para o *Distrito*), a *Faixa Etária* correspondente ao paciente e o atributo *Fumador* relativo a caracterização do paciente a nível tabagístico. A resposta está apresentada na Figura 6.8, correspondente ao mapa temático gerado pelo sistema SOLAP+ como resposta à pergunta inicial. Conclui-se então que, com base nos dados recolhidos com o inquérito, a maior parte dos pacientes fumadores encontram-se entre os trinta e os cinquenta anos. Mais uma vez no mapa aparecem zonas preenchidas com a cor base do mapa, pois novamente não existiam dados para essas regiões.

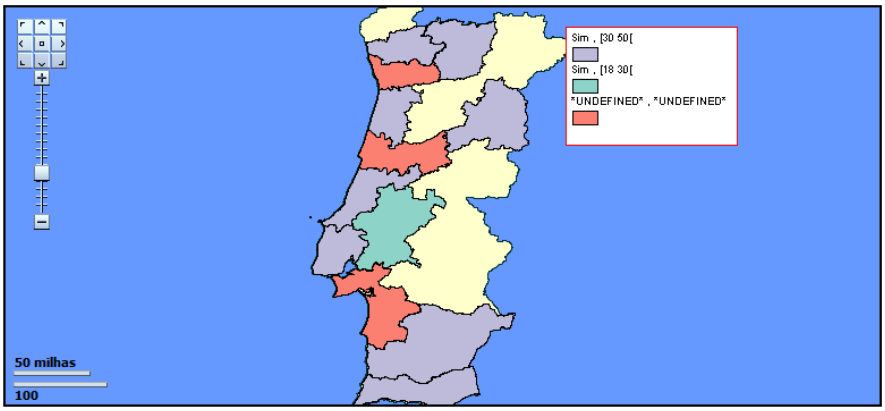
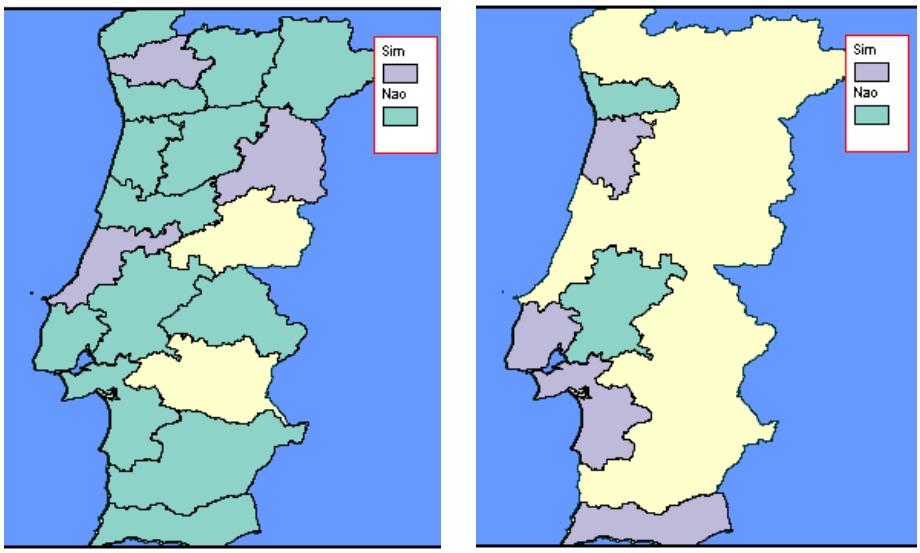


Figura 6.8 - Análise à faixa etária dos fumadores.

A última análise que se demonstra para este caso de estudo, relaciona-se com o maior cansaço evidenciado pelos pacientes relativamente à idade. A Figura 6.9 a) representa a caracterização espacial por distritos das evidências de maior cansaço apresentada pelos pacientes, com base em todos os registos recolhidos. Como é visível existe uma maior predominância de pessoas que não evidenciam maior cansaço, a excepção de três distritos cuja caracterização com valor positivo deve-se ao facto de haver poucos registos nessas regiões e desses na sua maioria serem positivos. Na Figura 6.9 b) apresenta-se o mapa obtido realizando um *slice* semântico para analisar somente os indivíduos que apresentem *DPOC* (Doença Pulmonar Obstrutiva Crónica). Neste caso já se vê maior predominância de regiões onde a característica dominante é haver um maior cansaço relativamente à idade que os pacientes apresentam. Mais uma vez, os distritos que fogem à regra, neste caso *Porto* e *Santarém*, foram analisados com recurso a tabela de detalhe e concluiu-se que apresentam apenas um registo de pacientes com DPOC, o que explica o facto de fugirem à norma.



a) b)

Figura 6.9 - Mapas obtidos com finalidade de analisar da influência da DPOC nos pacientes.

Desta forma, avaliou-se o efeito da generalização e o poder de análise que traz ao protótipo SOLAP+ em domínios diferentes, como é este, referente a um inquérito, onde os dados são quase na sua totalidade semânticos e não existe utilização das métricas para efectuar as análises pretendidas.

7. Conclusões e trabalho futuro

Nesta secção são apresentadas as conclusões finais a que se chegou com a elaboração desta dissertação e apresentadas futuras contribuições que podem ser dadas ao protótipo SOLAP+ no âmbito da caracterização espacial com base na generalização dos atributos, quer a nível de novas funcionalidades e representação dos dados, quer ao nível da avaliação do protótipo.

7.1. Conclusões

O propósito desta dissertação era implementar técnicas de *data mining* no protótipo SOLAP+. Esta incorporação deveria ter em consideração a existência de um modelo multidimensional, com a presença de tabelas de factos e métricas numéricas, e tanto quanto possível utilizar a informação que este fornece. Esta integração devia também ter sempre presente a componente espacial, através do mapa transmitir a maior quantidade de informação relevante possível.

Das diferentes técnicas existentes, adoptou-se a que se considerou que melhor se encaixava nas características do protótipo, recaindo a escolha sobre a técnica de indução orientada aos atributos como forma de sumarizar informação e, através dessa, conseguir extrair regras que permitissem caracterizar semanticamente os objectos espaciais.

A inclusão do processo de generalização no SOLAP+ foi realizada com sucesso e houve algumas características do protótipo que facilitaram essa integração. A existência de hierarquias definidas no modelo foi fundamental para a realização do processo permitindo assim subir no nível conceptual dos atributos. Por ouro lado, a utilização dos níveis que compõem as hierarquias definidas no modelo multidimensional, permitiu definir uma nova forma de definição do limite de generalização de cada atributo. Utilizaram-se os níveis para construção de *sliders* que permitem a indicação do nível para o qual se pretende generalizar os atributos ao invés da indicação de um valor de *threshold* com indicado na literatura, obrigando a computações extra para definir o limite da generalização.

A caracterização espacial foi realizada de duas maneiras distintas: através da análise do número de ocorrências de cada característica, ou em função do valor apresentado pela métrica numérica. Esta última forma de definição da característica dominante foi uma novidade introduzida que só foi possível devido à abordagem seguida pelo SOLAP+, fazendo uso da presença das métricas nas tabelas de facto e utilizando os operadores de agregação definidos para fazer a junção dos valores.

No final da preparação desta dissertação foram definidas várias tarefas a desempenhar. No quadro seguinte estão presentes as funcionalidades que tinham sido identificadas e a indicação das que foram implementadas (+) e as que ficaram por implementar (−) devido à falta de tempo.

Pedido de sumarização por indicação dos atributos que compõem a relação inicial	+
Pedido de sumarização utilizando uma interrogação SOLAP como relação inicial.	–
Generalização de dominância espacial	+
Generalização de dominância não espacial	+
Refinamento de hierarquias	–
Criação de hierarquias para os atributos numéricos	+
Alteração do algoritmo de agrupamento DBSCAN	+
Implementação do algoritmo de regionalização	+
Gravar hierarquias criadas para serem usadas em análises posteriores	–
Caracterização espacial com base no número de ocorrências	+
Caracterização espacial com base no valor da métrica	+

De salientar ainda, que o protótipo foi utilizado em dois casos de estudo de forma a verificar a aplicabilidade das soluções implementadas e validar os objectivos desta dissertação.

7.2. Trabalho futuro

Para além de terminar a implementação das funcionalidades que ficaram por realizar e que tinham sido inicialmente planeadas, existe uma serie de acções e ideias que podem ser realizadas futuramente.

Uma das actividades a ser realizada seria analisar o comportamento do SOLAP+ em casos reais. Com esta análise seria possível verificar as capacidades actuais do sistema e encontrar limitações que persistam, com o objectivo de as contornar e assim melhorar o protótipo. Através desta utilização em casos reais, seria possível retirar conclusões relativas à importância de combinar as técnicas de indução com o SOLAP+ e as vantagens que isso traria do ponto de vista do analista.

Uma funcionalidade interessante de ser incorporada está relacionada com a escolha da característica dominante para cada objecto espacial. Havendo locais que evidenciam diferentes características e optando por uma delas para representar aquele local, temos que ter presente qual o nível de diferença necessário entre as diferentes características para realizar tal escolha. Esta funcionalidade permitiria ao utilizador expressar o seu conhecimento da realidade sobre o que realmente é considerado relevante para o caso em estudo e faria com que os resultados provenientes das decisões tomadas pelo protótipo fossem mais fortes e de encontro com os interesses do utilizador. Assim, algumas das abordagens possíveis poderiam ser permitir ao utilizador, por exemplo:

- Escolher o nível a partir do qual as características seriam consideradas relevantes e dessa forma entrariam na composição da resposta final como existentes nos locais. Este nível poderia ser, por exemplo, uma percentagem relacionada com o número de ocorrências,
- Definir o grau de diferença necessário para se considerar uma característica dominante sobre as restantes. Desta forma, evitar-se-ia os casos em que uma característica é considerada dominante mesmo apresentado um diferença mínima em relação a outras que também se evidenciaram nesse local.

No seguimento do problema anterior, poderia surgir um nova abordagem de representação dos resultados obtidos para facilitar a análise e retirar conclusões mais claras sobre os mesmos. Seria interessante fazer a distinção dos vários casos através de aplicação de diferentes estilos, abordando assim os casos onde surgiria, para um dado objecto espacial, mais do que uma característica com valores considerados relevantes, mas não sendo nenhuma dominante sobre as outras.

8. Bibliografia

1. Wrembel, R.: Data Warehouses and OLAP: Concepts, Architectures and Solutions. IRM Press (2006)

2. Rivest, S., Bédard, Y., Marchand, P.: Toward Better Support for Spatial Decision Making: Defining the Characteristics of Spatial On-Line Analytical Processing (SOLAP). Geomatica, 539--555 (2001)

3. Matias, R.: Integração de Informação Geográfica em Sistemas OLAP. Dissertação de Mestrado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Caparica (2006)

4. Vitorino, M., Caldeira, R.: The Spatial One. Dissertação de Mestrado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Caparica (2008)

5. Jorge, R.: SOLAP+: Extending the Interaction Model. Dissertação de Mestrado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Caparica (2009)

6. Koperski, K., Adhikary, J., Han, J.: Spatial Data Mining: Progress and Challenges. In : Sigmod Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD), pp.1-10 (1996)

7. Zeitouni, K.: A Survey on Spatial Data Mining Methods Databases and Statistics. In : Point of Views, Information Resources Management Association International Conference (IRMA'2000), Data Warehousing and Mining Track (2000)

8. Han, J., Koperski, K., Stefanovic, N.: GeoMiner: A System Propotype for Spatial Data Mining. (1997)

9. Silva, R.: SOLAP+. Dissertação de Mestrado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Caparica (2010)

10. Han, J., Fu, Y.: Exploration of the power of attribute-oriented induction in data mining. (1996)

11. Ferraz, V., Santos, M.: GlobeOLAP: Improving the Geospatial Realism in Multidimensional Analysis Environment. ICEIS, 99-107 (2010)

12. Scotch, M., Parmanto, B.: SOVAT: Spatial OLAP Visualization and Analysis Tool. In : 38th Annual Hawaii International Conference on System Sciences (HICSS'05), vol. VI (2005)

13. Scotch, , Parmanto, B., Monaco, V.: Usability Evaluation of the Spatial OLAP Visualization and Analysis Tool (SOVAT). Journal of Usability Studies 2 (2007)

14. In: Online geographic information - K2 Geospatial. Available at: http://www.k2geospatial.com/

15. Han, , Fu, : Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. AAAI-94 Workshop on Knowledge Discovery in Databases, 157-168 (1994)

16. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann (2006)

17. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 2nd International Conference on Knowledge Discovery and Data Mining, 226-231 (1996)

18. Harvey, M., Han, J.: Geographic Data Mining and Knowledge Discovery 2nd edn. CRC Press (2009)

19. Tiede, D., Strobl, J.: Polygon-based Regionalisation in GIS Environment. (2006)

20. Zhou, X., Truffet, D., Han, J.: Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining. In : 6th International Symposium on Large Spatial Databases (SSD'99), Hong Kong, pp.167--187 (1999)