

***Data Mining* Espacial: um Estudo de Caso**

Maribel Yasmina Santos

Universidade do Minho, Departamento de Sistemas de Informação, Guimarães, Portugal
maribel@dsi.uminho.pt

Luís Alfredo Amaral

Universidade do Minho, Departamento de Sistemas de Informação, Guimarães, Portugal
amaral@dsi.uminho.pt

Resumo

A descoberta de conhecimento em bases de dados está associada à identificação de relacionamentos implícitos existentes nos dados analisados. O processo global de descoberta de conhecimento, que passa por várias etapas, inclui a gestão dos algoritmos de *Data Mining*, utilizados para extrair padrões dos dados, e a interpretação dos padrões encontrados pelos mesmos.

Um caso particular da descoberta de conhecimento em bases de dados diz respeito à exploração de dados referenciados espacialmente, isto é, dados que incluem referências a objectos geográficos, localizações ou partes de uma divisão territorial. A análise destes dados impõe a verificação da componente espacial associada aos mesmos (direcções, adjacências, distâncias, ...), assim como a sua influência nos restantes dados explorados, já que um objecto geográfico pode ser afectado por acontecimentos verificados em objectos vizinhos.

A análise de dados espaciais, com o objectivo de descoberta de conhecimento, requer a utilização de técnicas específicas, que permitam a incorporação da semântica espacial, implícita na posição e dimensão dos objectos geográficos referenciados, no referido processo. Este documento descreve uma nova abordagem na análise de dados geo-referenciados, concretizada no sistema PADRÃO, baseada em mecanismos de posicionamento indirecto e estratégias de raciocínio espacial qualitativo.

A apresentação de um estudo de caso, com a análise de uma Base de Dados que integra o Sistema de Informação de Administração do Pessoal do Exército, permitiu validar os princípios conceptuais estabelecidos para o sistema PADRÃO, assim como constatar a sua utilidade na exploração de bases de dados geo-referenciadas de grande dimensão. Este estudo de caso evidencia a identificação de relacionamentos implícitos existentes entre os dados não geográficos e os dados geográficos analisados.

Palavras chave: descoberta de conhecimento em bases de dados, dados geográficos, raciocínio espacial qualitativo.

1 Introdução

Actualmente, as instituições produzem e armazenam grandes quantidades de dados, resultantes da sua actividade diária. O facto destes dados reflectir o comportamento e evolução das organizações ao longo do tempo, chama a atenção para os benefícios que podem decorrer da compreensão do conhecimento implícito nos mesmos, e da sua consequente utilização na tomada de decisão.

A investigação na área da Descoberta de Conhecimento em Bases de Dados (DCBD) [Fayyad, et al. 1996] [Han e Kamber 2001] tem evoluído consideravelmente, permitindo a implementação de algoritmos que automatizam o processo de análise de dados, com vista à descoberta de conhecimento implícito nos mesmos. A análise de dados referenciados espacialmente representa uma sub-área da DCBD [Koperski, et al. 1996], cujos principais desenvolvimentos estão associados à implementação de novos algoritmos de *Data Mining* (DM), ou à adaptação de algoritmos já existentes, por forma a estes incluírem a semântica associada à localização dos factos, no processo de descoberta de conhecimento.

A concepção, implementação e validação do PADRÃO [Santos 2001], permitiu a construção de um sistema de descoberta de conhecimento para Bases de Dados (BD) geo-referenciadas, baseado em mecanismos de posicionamento indirecto e estratégias de raciocínio espacial qualitativo. Os princípios estabelecidos para o sistema PADRÃO representam uma nova abordagem na análise de dados espaciais, que suprime a necessidade de desenvolvimento ou de adaptação de algoritmos. A utilização de mecanismos de posicionamento indirecto, referência espacial através de identificadores geográficos, evita ainda, a necessidade de definição geométrica das entidades geográficas referenciadas. Esta definição geométrica é requerida em abordagens quantitativas, nas quais o posicionamento directo dos objectos é especificado através de coordenadas de pontos, que indicam a localização espacial dos mesmos.

A abordagem qualitativa à referência espacial, considerada pelo sistema PADRÃO, permite que as BD organizacionais sejam analisadas sobre uma perspectiva espacial, independentemente da disponibilidade da geometria das diversas entidades geográficas referenciadas. A componente espacial associada aos dados geo-referenciados, e assim incluída no processo de descoberta de conhecimento, é manipulada através de mecanismos qualitativos de raciocínio espacial [Freksa 1991] que permitem a inferência de informação geográfica desconhecida.

A arquitectura do sistema PADRÃO [Santos e Amaral 2000a] integra três componentes: o Repositório de Dados e Conhecimento, a Análise de Dados e a Visualização de Resultados. O Repositório de Dados e Conhecimento é o responsável pelo armazenamento dos dados utilizados no sistema, e das regras que permitem a implementação dos mecanismos de inferência utilizados no processo de raciocínio espacial qualitativo. O componente de Análise de Dados integra as seis fases do processo de descoberta de conhecimento consideradas pelo PADRÃO, as quais permitem que dados geo-espaciais [Santos e Amaral 2000b] e dados não geográficos sejam analisados simultaneamente, possibilitando a identificação de padrões implícitos nos mesmos. Os resultados do processo de descoberta de conhecimento podem ser armazenados no componente de Visualização de Resultados, permitindo que os mesmos sejam explorados em mapas das regiões analisadas, facilitando o processo de interpretação das regras encontradas.

O PADRÃO foi integralmente implementado no Clementine [SPSS 1999], uma ferramenta de descoberta de conhecimento para bases de dados relacionais. O sistema de informação geográfica GeoMedia Professional [Intergraph 1999] foi utilizado na implementação do VisualPadrão, uma aplicação desenvolvida em Visual Basic, que depois de integrada no ambiente de trabalho do sistema Clementine, permite a visualização de resultados em mapas das regiões analisadas.

Neste artigo é apresentado um estudo de caso no qual é analisada uma BD organizacional de grande dimensão, permitindo constatar a utilidade do sistema PADRÃO na análise de BD geo-referenciadas, com o objectivo de descoberta de conhecimento. A BD explorada integra o Sistema de Informação de Administração do Pessoal do Exército (SIAPE), e na mesma foi possível proceder à identificação de características espaciais e tendências espaciais nos dados explorados.

Este documento encontra-se organizado da seguinte forma. A secção 2 apresenta a BD em estudo, detalhando as tabelas da mesma posteriormente exploradas com técnicas de DM. A secção 3 introduz os principais conceitos associados ao DM espacial, descrevendo ainda, as

tarefas de DM apresentadas neste artigo. Na secção 4 são descritas as fases do processo de descoberta de conhecimento, consideradas pelo sistema PADRÃO, que permitem a concretização das tarefas de DM definidas. Por último, a secção 5 sistematiza o trabalho descrito neste documento.

2 O Estudo de Caso: Caracterização da BD

A BD a analisar integra o SIAPE. Nesta BD é armazenado o denominado *Cadastro Geral*, no qual o Comando de Pessoal do Exército armazena os dados pessoais dos indivíduos que fizeram/fazem parte do quadro permanente do Exército e ainda, os dados pessoais dos mancebos inspeccionados por esta instituição.

Antes de proceder com as diversas fases do processo de descoberta de conhecimento, é necessário explorar a BD a analisar, assim como identificar a relevância de cada uma das suas tabelas para o processo de descoberta de conhecimento. Pela análise da Figura 1 constata-se que a entidade central da BD é representada pela classe *Indivíduo*, à qual estão relacionadas as diversas classes que permitem a sua caracterização.

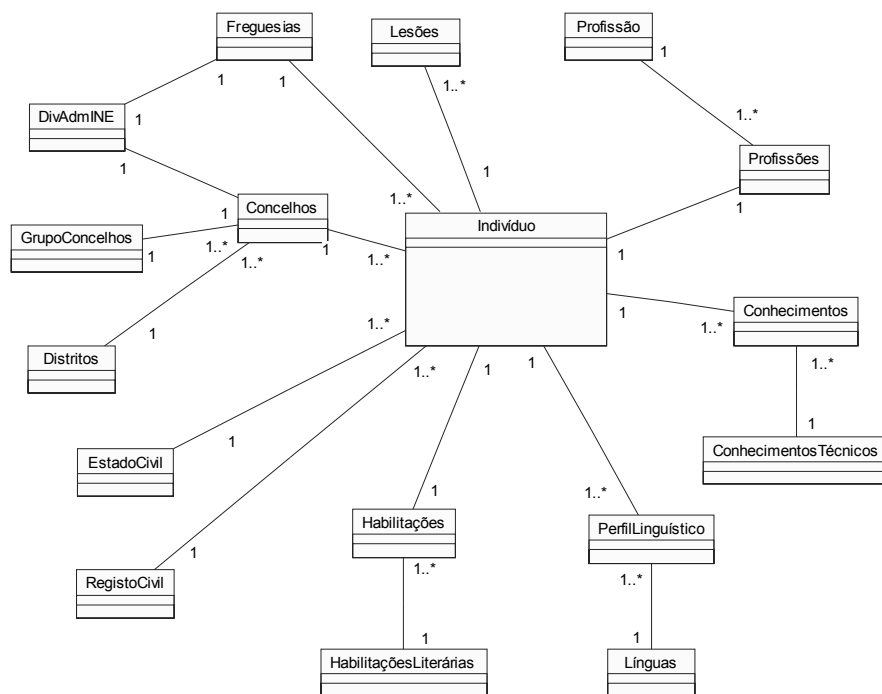


Figura 1 - Estrutura Lógica da BD

Para as tabelas identificadas como relevantes, e que são posteriormente exploradas nos exercícios de DM, é de seguida listado o seu conteúdo, referindo os seus atributos, assim como o significado de cada um dos mesmos. É ainda destacado o número total de registos que integram cada uma das tabelas.

Tabela Indivíduos: 1.328.573 registos

Número	Número de Identificação Militar
DataNascimento	Data de Nascimento
Sexo	Sexo
Concelho	Código do Concelho
Freguesia	Código da Freguesia
ConservRC	Código da Conservatória do Registo Civil
EstadoCivil	Código do Estado Civil
GrauAcadémico	Grau Académico
GrupoSanguíneo	Grupo Sanguíneo

Tabela Freguesias: 4.269 registos

Concelho	Código do Concelho
Freguesia	Código da Freguesia
DesigFreguesia	Designação da Freguesia
DivAdmINE	Divisão Administrativa INE

Tabela Concelhos: 313 registos

Concelho	Código do Concelho
DesigConcelho	Designação do Concelho
DistritoAdm	Identificação Distrito Administrativo
DivAdmINE	Divisão Administrativa INE
GrupoConcelhos	Grupo de Concelhos

Tabela PerfilLinguístico: 362.301 registos

Número	Número de Identificação Militar
Língua	Código da Língua
GrauConhecimento	Grau conhecimento
GrauFala	Grau fala
GrauEscreve	Grau escrita
GrauTraduz	Grau tradução
GrauCompreensão	Grau compreensão
GrauLeitura	Grau leitura

Tabela Línguas: 20 registos

Língua	Código da Língua
DesigLíngua	Designação da Língua

Tabela Lesões: 492.204 registos

Número	Número de Identificação Militar
Lesão	Código da Lesão
SIVAGE	Factor SIVAGE
GrauLesão	Grau da Lesão Detectada

As próximas subsecções sintetizam o conteúdo das tabelas Indivíduo, PerfilLinguístico e Lesões, uma vez que estas são utilizadas no processo de descoberta de conhecimento que será posteriormente apresentado.

2.1 Tabela Indivíduo

A tabela Indivíduo integra o *Cadastro Geral* do SIAPE, armazenando parte dos dados pessoais utilizados na caracterização dos indivíduos. Ao nível das datas de nascimento, refere-se que a BD inclui indivíduos nascidos entre o ano de 1858 e o ano de 1983. Até ao ano de 1981, inclusive, apenas eram introduzidos nesta BD dados referentes a indivíduos do quadro permanente. A partir de 1982, esta tabela passou também a armazenar os dados dos indivíduos inspeccionados pelo Exército. Refere-se que a data de nascimento destes indivíduos ronda o ano de 1964. No ano 2000, ano em que esta BD foi cedida para análise, registaram-se os mancebos nascidos 17/18 anos antes, pelo que as datas de nascimento registadas nesta tabela atingem, no máximo, o ano de 1983.

Os atributos GrauAcadémico e GrupoSanguíneo que integram a tabela Indivíduo não se encontram preenchidos. A Figura 2 apresenta a distribuição por Concelho (uma vez que devido a elevada quantidade de registos, as freguesias são generalizadas até este nível ou até ao nível dos distritos), Estado Civil e Sexo, dos diversos registos que integram esta tabela.

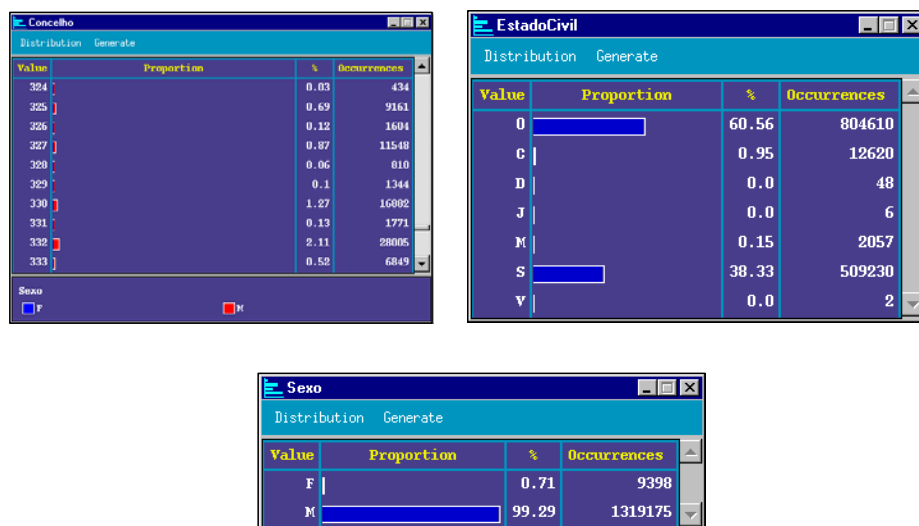


Figura 2 - Distribuição dos indivíduos pelos atributos Concelho, EstadoCivil e Sexo

Pela análise da Figura 2 verifica-se que a maioria dos indivíduos apresentam o estado civil *desconhecido* (código 0), seguido do estado civil *solteiro* (código S). Os restantes casos encontram-se distribuídos pelos estados *casado* (código C), *divorciado* (código D), *viúvo* (código V), *separado judicialmente* (código J) e *vive maritalmente sem ser casado* (código M). Uma vez que para a maioria dos registos não é conhecido o estado civil dos indivíduos, não é possível considerar este atributo nas tarefas de DM que serão posteriormente definidas. O mesmo acontece com o atributo Sexo, uma vez que dos 1.328.573 registos que integram esta tabela, 1.319.175 dizem respeito a indivíduos do sexo *masculino*, sendo apenas de 9.398 o número de indivíduos do sexo *feminino*.

Uma análise detalhada, das data de nascimento armazenadas nesta tabela, permitiu constatar que existem 215 registos para os quais não foi declarada a data de nascimento dos indivíduos. Quatro registos apresentam valores errados nas datas, mais especificamente as datas 12/06/979, 08/12/2886, 03/01/1989 e 14/03/1994, podendo as mesmas ser rectificadas para os valores correctos (identificados pelo Comando de Pessoal do Exército), que são 12/06/1979, 08/12/1886, 03/01/1889 e 14/03/1884, respectivamente. As datas em falta serão, posteriormente, devidamente assinaladas com a etiqueta 'Desconhecida'.

2.2 Tabela PerfilLinguístico

O perfil linguístico caracteriza o grau de conhecimento de uma determinada língua. Apesar de terem sido definidos diversos atributos para permitir esta caracterização, em factores como a compreensão ou a tradução, apenas a coluna respeitante ao grau de conhecimento (GrauConhecimento) se encontra preenchida. Este atributo é classificado recorrendo aos indicadores: D (desconhecido), M (muito), R (regular) e P (pouco). A Figura 3 apresenta um pequeno extracto dos registos armazenados na tabela PerfilLinguístico (à qual foi integrado o atributo DesigLíngua, da tabela Línguas), assim como a distribuição dos atributos DesigLíngua e GrauConhecimento, pelos diversos valores possíveis.

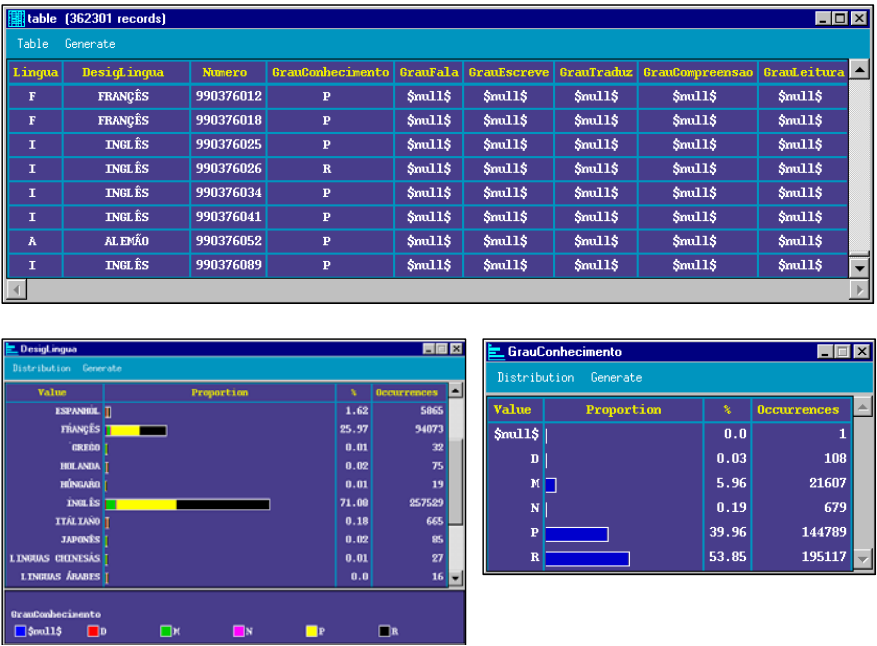


Figura 3 - Caracterização da tabela PerfilLinguístico

Pela análise da Figura 3 verifica-se que o atributo GrauConhecimento apresenta um registo com código omissor, devendo o mesmo ser substituído pelo código D. Consta-se, ainda, que 679 registos apresentam o código N, o qual não consta do conjunto de valores possíveis definido para este atributo. Duas situações podem ter ocorrido, um erro na digitação do código M, uma vez que a tecla N está imediatamente à esquerda desta, ou o N foi introduzido para caracterizar um grau de conhecimento *normal*, situação que implicaria a sua substituição pelo código R. Estas duas hipóteses foram averiguadas junto do Comando de Pessoal, permitindo verificar que se está perante a primeira situação, pelo que os códigos N serão substituídos pelo código M.

Em relação à distribuição dos indivíduos pelas diversas línguas (Figura 3), verifica-se que existe uma predominância do conhecimento do Inglês, Francês, Espanhol e Alemão, existindo contudo, conhecimentos em outras línguas.

2.3 Tabela Lesões

A tabela Lesões armazena o resultado dos exames médicos efectuados, o qual é transformado no factor SIVAGE. Este factor integra diversas características, são elas:

S	Membros Superiores
I	Membros Inferiores
V	Visão
A	Audição
G	Estado Geral
E	Estado Emocional

A cada um destes factores é atribuído um *grau de lesão* (GrauLesão), que pode assumir os seguintes valores: 1 (muito mau), 2 (mau), 3 (normal), 4 (bom) e 5 (muito bom).

A Figura 4 apresenta uma amostra dos dados armazenados na tabela Lesões, a distribuição dos dados pelo atributo SIVAGE e ainda, o histograma com a distribuição dos registos pelos diversos graus de lesão. Pela análise da figura constata-se que, ao nível do atributo SIVAGE, existem alguns registos com valores que não correspondem aos permitidos (apenas os códigos S, I, V, A, G e E). Na impossibilidade de identificação do valor correcto do atributo SIVAGE nestes registos, os mesmos serão removidos do conjunto de dados a analisar.

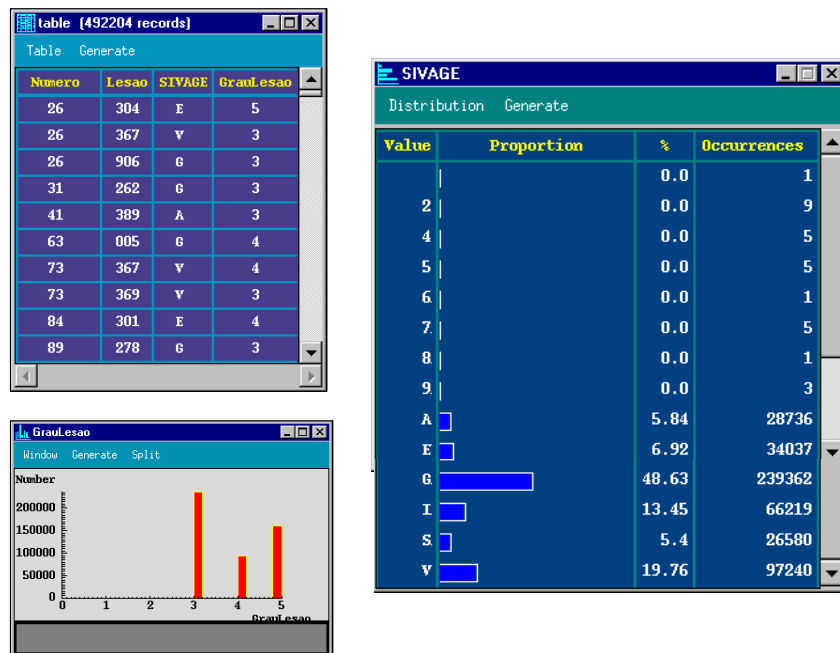


Figura 4 - Distribuição dos indivíduos pelos atributos SIVAGE e *grau de lesão*

Verificam-se, ainda, casos de duplicação no armazenamento de um determinado factor SIVAGE, para um dado indivíduo (como pode ser constatado na Figura 4, para o indivíduo com o número 73). Estas duplicações são ocasionadas pela verificação de diferentes *lesões*, que na sua codificação para um resultado, originam o mesmo código SIVAGE, mas por vezes com *graus* diferentes. Para estes casos, e como sugerido pelo Comando de Pessoal, deverá adoptar-se, na análise, o registo que apresenta o maior *grau*, permitindo uma *aproximação por excesso*, e nunca por defeito, à realidade.

3 O Data Mining Espacial e os seus objectivos

A Descoberta de Conhecimento em Bases de Dados Espaciais (DCBDE) refere-se ao processo de extracção de padrões ou regularidades espaciais nos dados, relacionamentos existentes entre dados espaciais e dados não espaciais, ou outras características implícitas em BD espaciais [Lu, et al. 1993]. Este processo desempenha um papel fundamental na percepção das características não espaciais associadas aos dados espaciais, e principalmente, na captura dos relacionamentos implícitos que existem entre estes dois conjuntos de dados.

As tarefas tradicionalmente associadas aos processo de DCBDE [Ester, et al. 1998] [Han e Kamber 2001] incluem:

- a descrição de distribuições espaciais nos dados não espaciais: **caracterização espacial**. A caracterização espacial de um conjunto de objectos consiste na descrição das propriedades espaciais e não espaciais comuns aos objectos analisados. Nesta caracterização, não são consideradas apenas as propriedades dos objectos alvo do estudo, mas também as propriedades dos seus vizinhos. Esta tarefa permite determinar o conjunto de registos (atributo, valor) e o conjunto de objectos para os quais a frequência relativa de incidência nesse conjunto, e nos seus vizinhos, é diferente da frequência relativa verificada nos restantes registos da BD.
- a verificação de distribuições espaciais nos dados não espaciais: **análise espacial discriminante**. A análise espacial discriminante permite contrastar padrões espaciais de dados não espaciais, comparando a variação dos atributos não espaciais em diversas regiões geográficas.
- o estabelecimento de relações entre dados espaciais, e entre dados espaciais e dados não espaciais: **associação espacial**. A associação espacial permite identificar a relação que existe entre um conjunto de objectos espaciais e um conjunto de dados não espaciais, ou entre dois conjuntos de dados espaciais, definindo a associação (implicação) que existe entre os mesmos. Uma regra de associação espacial deve integrar pelo menos um predicado espacial, que pode estar associado a relações do tipo direcção, distância ou topologia [Frank 1992] [Egenhofer 1994].
- a verificação de alterações regulares de um ou mais atributos não espaciais, associados a um dado objecto espacial: **detecção de tendências espaciais**. Uma tendência espacial consiste numa alteração regular de um ou mais atributos não espaciais, verificada no sucessivo afastamento de um dado objecto espacial. O conhecimento da vizinhança existente entre os objectos permite a movimentação entre os mesmos, sendo o afastamento em relação ao objecto inicial medido recorrendo à distância existente entre eles. As sucessivas alterações nas distâncias e os diferentes valores verificados pelos atributos permitem determinar tendências espaciais nos dados.

Neste estudo de caso, e atendendo aos dados disponíveis, foram definidos quatro objectivos para o DM, cada um dos quais com o intuito de concretizar cada uma das tarefas tradicionalmente associadas ao DM espacial. As tarefas incluem [Santos 2001]:

- Tarefa 1.** Caracterização espacial do *perfil linguístico* dos indivíduos. Na realização desta tarefa pretende-se verificar a distribuição espacial associada ao conhecimento de uma dada língua, mais especificamente o Alemão e o Francês. Este exercício permite identificar as zonas geográficas que apresentam uma maior incidência de determinado *grau de conhecimento*. No que diz respeito à componente geográfica, a análise será efectuada ao nível dos distritos que integram Portugal Continental.
- Tarefa 2.** Análise espacial discriminante do *perfil linguístico*, para os diversos distritos que integram Portugal Continental. A realização desta tarefa permite identificar um conjunto de regras que identificam a(s) língua(s) mais conhecida(s) em determinada região e o respectivo *grau de conhecimento*.
- Tarefa 3.** Detecção de associações espaciais entre as *habilitações literárias*, obtidas pelos indivíduos, e as *profissões* exercidas pelos mesmos. Na realização desta tarefa, e no sentido de efectuar a análise ao nível dos concelhos, a identificação de regras de associação espacial é realizada para o distrito de Braga.
- Tarefa 4.** Detecção de tendências espaciais no factor SIVAGE, que permitam identificar alterações regulares nos *graus de lesão* verificados pelos indivíduos. A realização desta tarefa permite a identificação de alterações regulares dos graus de lesão, associados a regiões que sucessivamente se afastam de uma dada entidade geográfica. Ao nível geográfico, esta tarefa é realizada analisando os concelhos que integram o distrito de Braga.

Neste artigo é apresentada a concretização de duas destas tarefas, descritas na próxima secção, tendo-se seleccionado a identificação de *características espaciais* (Tarefa 1) e a detecção de *tendências espaciais* nos dados (Tarefa 4).

4 O Processo de Descoberta de Conhecimento no sistema PADRÃO

As diversas fases do processo de descoberta de conhecimento, necessárias à satisfação dos objectivos do DM descritos anteriormente, e consideradas pelo sistema PADRÃO [Santos e Amaral 2000b] [Santos 2001], são apresentadas nas próximas subsecções. Refere-se que às cinco fases que tradicionalmente integram o processo de DCBD [Fayyad, et al. 1996], foi adicionada mais uma etapa, associada ao processamento da informação geo-espacial.

4.1 Selecção, tratamento e pré-processamento dos dados

O processo de compreensão dos dados apresentado anteriormente permitiu detectar a existência de atributos irrelevantes, atributos não preenchidos e registos com valores errados. Para cada uma das tabelas analisadas, identificaram-se as situações de erro e ainda, os atributos/registos a remover de cada uma das mesmas.

As etapas de selecção, tratamento e pré-processamento dos dados foram efectuadas em simultâneo (isto é, recorrendo a uma única *stream* Clementine para cada uma das tarefas), permitindo otimizar o tempo gasto nas mesmas, já que a quantidade de dados a analisar é extremamente elevada. Ainda com o objectivo de minorar o tempo consumido no processamento e análise dos dados, os registos resultantes da execução destas três etapas foram armazenados em *cache files*, ficheiros de armazenamento de dados com um formato proprietário do Clementine, que apresentam como vantagem uma maior velocidade de acesso aos dados.

No que diz respeito à tabela *Indivíduo*, executaram-se as seguintes tarefas:

- Ao nível da *selecção*, foram excluídos os atributos *Sexo*, *GrupoSanguíneo*, *GrauAcadémico* e *EstadoCivil*.
- Na fase de *tratamento* dos dados:
 - as datas de nascimento 08-12-2886, 14-03-1994, 12-06-979 e 03-01-1989 foram rectificadas para 08-12-1886, 14-03-1894, 12-06-1979 e 03-01-1889, respectivamente.
 - os registos com data de nascimento desconhecida, foram etiquetados com a marca "Desc".

A tabela *PerfilLinguístico* apresenta diversos atributos não preenchidos, conduzindo à remoção das colunas *GrauFala*, *GrauEscreve*, *GrauTraduz*, *GrauCompreensão* e *GrauLeitura*. No que diz respeito ao *tratamento*, e ao atributo *GrauConhecimento*, os valores "\$null\$" foram substituídos por "D", enquanto que os valores "N" foram substituídos por "M".

Na tabela *Lesões*, ao nível da *selecção* foi excluído o atributo *Lesão*, já que o mesmo está implícito no atributo *SIVAGE*. No que diz respeito ao *tratamento* dos dados:

- foram eliminados todos os registos cujo factor *SIVAGE* não coincide com os códigos S, I, V, A, G ou E.
- e para os indivíduos que apresentem duplicação de algum factor *SIVAGE*, foi removido aquele que apresenta o *grau de lesão* menor, como sugerido pelo Comando de Pessoal.

Para a satisfação da primeira tarefa, definida nos objectivos do DM, foi necessário proceder à integração de dados da tabela *Indivíduo*, com dados da tabela *PerfilLinguístico*. A geo-referenciação dos indivíduos, disponibilizada ao nível das freguesias, foi generalizada até ao nível dos distritos, uma vez que foi o determinado na definição da tarefa, e também, porque a quantidade de registos a analisar (mais de 360 mil registos) é extremamente elevada. A generalização até ao nível dos concelhos conduz à existência de mais de 275 casos distintos (entidades geográficas), interferindo com o desempenho dos algoritmos de DM. Ao nível dos distritos, e considerando a análise geográfica restrita ao continente, é possível agregar a informação a analisar em 18 classes distintas, que representam os 18 distritos que integram Portugal continental. O conjunto de dados de treino e de teste foram gerados a partir de todos os dados disponíveis para esta tarefa. Tal permite ao utilizador especificar, na fase de modelação, a *língua* alvo do estudo. Esta abordagem possibilita a análise de outras línguas, além do Alemão e do Francês, sem ter de construir novos conjuntos de dados de treino e de teste.

A divisão do conjunto inicial de dados, no conjunto de dados de treino e de teste, foi de 1/3 e 2/3 respectivamente, como pode ser constatado na Figura 5. No final deste processo, o conjunto de dados de treino agrega 120.767 registos, enquanto que 241.534 registos podem ser utilizados para verificar o desempenho dos modelos obtidos.

Para a satisfação da quarta tarefa de DM estabelecida, foi necessário proceder à integração de dados armazenados na tabela *Indivíduo*, com dados da tabela *Lesões*. A geo-referenciação dos indivíduos, disponibilizada ao nível das freguesias, foi generalizada até ao nível dos concelhos, permitindo a selecção dos registos associados ao distrito de Braga. No total existem 36.761 registos, que serão divididos pelos dois conjuntos de dados necessários, treino e teste. A Figura 6 evidencia a *stream* que permitiu a divisão do conjunto de dados disponível para análise, no conjunto de dados de treino (*Tar4Treino.cdf*) e no conjunto de dados de teste (*Tar4Teste.cdf*).

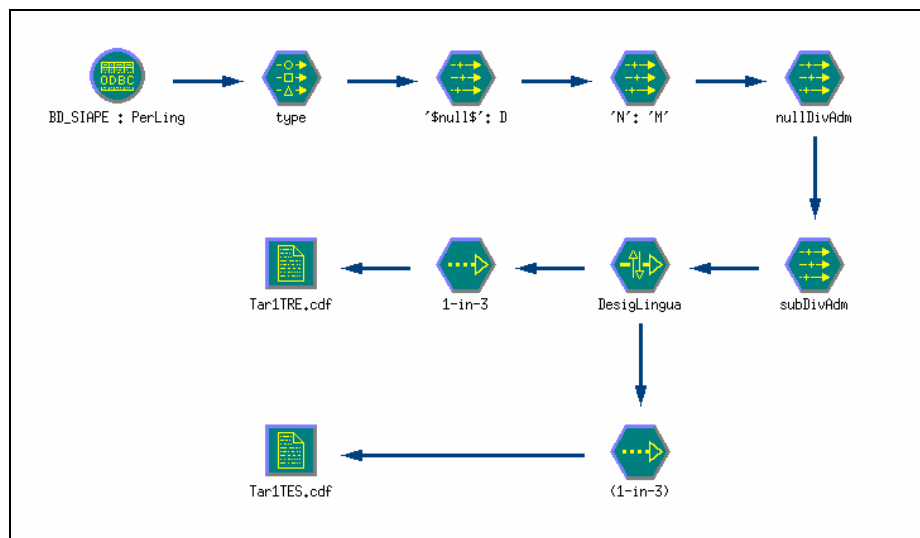


Figura 5 - Conjunto de dados de treino e de teste para a caracterização do *perfil linguístico*

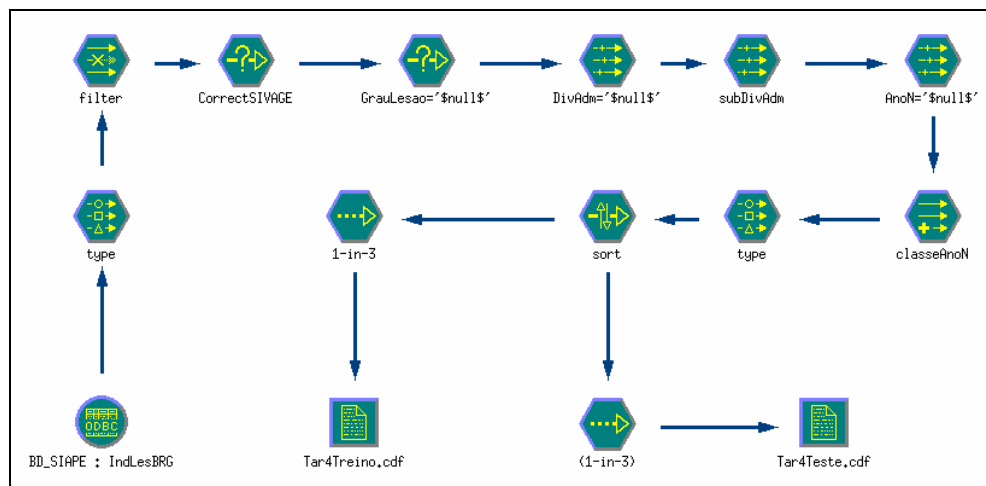


Figura 6 - Conjunto de dados de treino e de teste para a identificação de tendências espaciais

4.2 Processamento da informação geo-espacial

A abordagem ao processamento da informação geo-espacial considerada pelo sistema PADRÃO inclui, habitualmente, a inferência de relações espaciais desconhecidas, e a sua posterior utilização na construção de modelos, que permitam a inclusão da componente espacial no processo de descoberta de conhecimento.

No caso da realização da primeira tarefa, e dado que a componente geográfica foi generalizada até ao nível dos distritos, não é necessário proceder à inferência de informação geográfica. Nesta tarefa, a informação geográfica necessária diz respeito ao conhecimento da relação topológica existente entre distritos. Para a sua obtenção, apenas é necessário generalizar a

informação topológica existente para o nível dos concelhos. Esta informação está disponível na tabela Faces da Base de Dados Geográfica (BDG), que integra o componente de Repositório de Dados e Conhecimento. Esta tabela armazena as relações espaciais do tipo direcção, distância e topologia, existentes entre concelhos adjacentes. A generalização desta informação, até ao nível dos distritos, permite conhecer os distritos que são adjacentes. Todas as restantes relações topológicas existentes entre distritos, e não explícitas no conjunto inicial, dizem respeito a entidades não adjacentes, isto é, com relação topológica *deslocado* (*desl*). Esta generalização, e consequente explicitação da relação topológica existente entre distritos, permite que a informação geográfica necessária, à satisfação desta tarefa, seja integrada na fase de DM (apresentada na próxima subsecção).

No que diz respeito à quarta tarefa, a selecção da componente geográfica pode ser efectuada recorrendo aos nodos de manipulação de dados disponibilizados no Clementine. Os dados seleccionados são posteriormente analisados e processados, com vista à inferência de relações espaciais desconhecidas, necessárias na fase de DM.

O conhecimento explícito na Base de Conhecimento Espacial (BCE), do Repositório de Dados e Conhecimento, e que integra as regras de raciocínio que permitem inferir relações espaciais do tipo direcção, distância e topologia [Santos e Amaral 2000c], foi utilizado na construção de três modelos, *infDir*, *infDis* e *infTop*, que se complementam no processo de raciocínio qualitativo.

Após a selecção das relações espaciais armazenadas na BDG, para o distrito de Braga, os registos resultantes são manipulados e utilizados na inferência de informação geográfica desconhecida. Para inferir as relações espaciais que podem existir entre todos os concelhos do distrito em análise, foi construída uma *stream* (Figura 7) que é executada ciclicamente, até que não existam mais relações a inferir, isto é, relações entre concelhos desconhecidas. As novas relações espaciais existentes entre concelhos não adjacentes, e inferidas através deste processo cíclico, constituem conhecimento que passa a estar disponível na BDG (BDG:geoBraga).

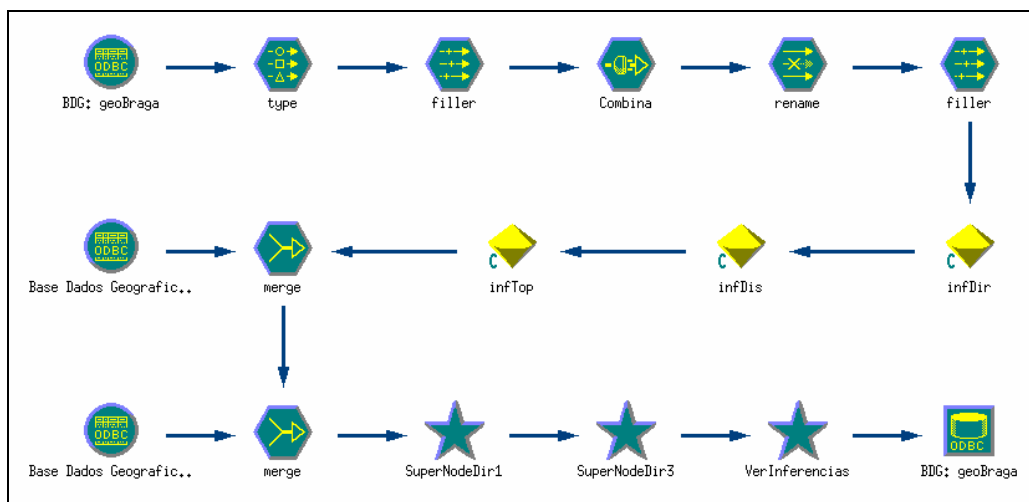


Figura 7 - Processo de inferência de relações espaciais desconhecidas

4.3 Data Mining

Na fase de DM apenas é necessário seleccionar o algoritmo apropriado para a execução de uma dada tarefa. No caso da primeira tarefa, a caracterização espacial será efectuada recorrendo ao

algoritmo C5.0. A árvore de decisão resultante identificará a região, ou regiões, em que se verifica uma maior incidência no conhecimento de uma dada língua.

A Figura 8 apresenta a *stream* construída para a caracterização geográfica do perfil linguístico, nomeadamente no conhecimento do Alemão e do Francês. Na referida figura é possível constatar que, ao ficheiro com os dados de treino (Tar1Tre.cdf) é integrada a tabela que armazena a hierarquia conceptual definida para o domínio geográfico (BDG: Hierarquias). Posteriormente, a relação topológica existente entre distritos (BDG: RelDistritos) é integrada com a generalização realizada, permitindo ao algoritmo de DM utilizado conhecer a relação topológica existente entre distritos.

Na Figura 8 é ainda possível verificar que foram construídos dois modelos, o primeiro, GrauConhAle, caracteriza a distribuição geográfica do conhecimento do Alemão, enquanto que o modelo GrauConhFran, caracteriza a distribuição geográfica do conhecimento do Francês. As regras explícitas em cada um destes modelos podem ser visualizadas na mesma figura (à esquerda, as respeitantes ao conhecimento do Alemão, e à direita, as respeitantes ao conhecimento do Francês), a qual apresenta ainda, o suporte e a confiança associada às mesmas. A confiança não se revela elevada, variando entre 34% e 67% no caso do Alemão, e 47% e 56% no caso do Francês. Na fase seguinte, *interpretação de resultados*, será avaliado o desempenho destes modelos na classificação do conjunto de dados de teste.

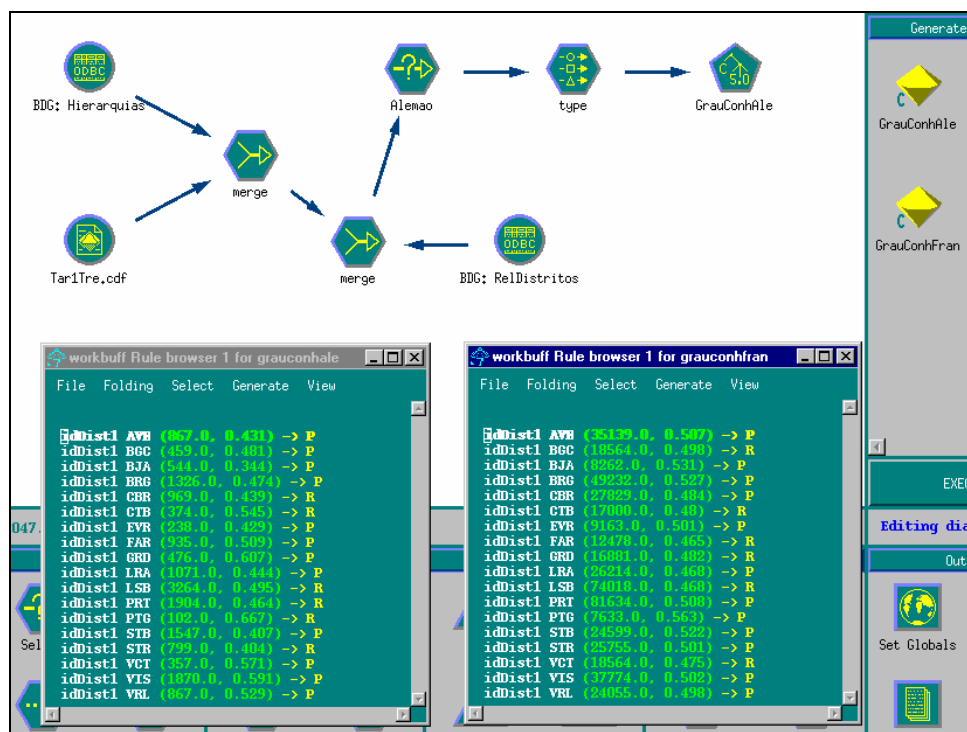


Figura 8 - Caracterização do conhecimento do Alemão e do Francês

A detecção de tendências espaciais, no factor SIVAGE, tem como objectivo identificar alterações regulares nos graus de lesão verificados pelos indivíduos. Basicamente, as tendências espaciais correspondem a alterações sucessivas de um ou mais atributos não espaciais, à medida que a análise se afasta de determinado objecto espacial. Para a realização desta tarefa, é necessário proceder à integração da componente geográfica do distrito em análise, já disponível

na BDG (geoBraga), com o conjunto de dados de treino. Este processo é evidenciado na Figura 9, na qual é ainda possível verificar que o algoritmo C5.0 é utilizado na identificação de regras que descrevem tendências nos dados. No caso particular do código V (*visão*), verifica-se uma distribuição, dos diferentes graus de lesão, por todos os concelhos do distrito. A análise da distribuição obtida permitiu seleccionar o concelho de Guimarães (código 308), para um estudo mais restrito (uma vez que o mesmo apresenta grau de lesão 5, e se encontra localizado na periferia do distrito). As distâncias qualitativas existentes entre os restantes concelhos do distrito e o concelho 308 foram analisadas pelo algoritmo C5.0, permitindo identificar um número bastante reduzido de regras que explicitamente exprimem a alteração do *grau de lesão*, à medida que este se afasta do concelho em análise. O modelo construído, GrauLesãoGUIM, é evidenciado na parte inferior esquerda da Figura 9. Mais adiante, na Visualização de Resultados, serão visualizados graficamente estes achados.

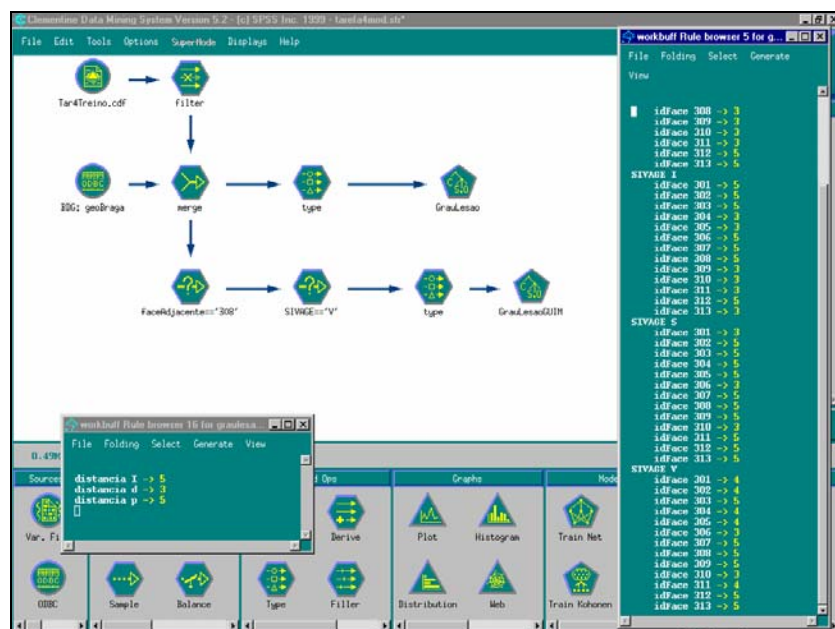


Figura 9 - Detecção de tendências espaciais no factor SIVAGE

4.4 Interpretação de Resultados

Os modelos GrauConhAle e GrauConhFran, construídos para a caracterização geográfica do conhecimento das línguas Alemão e Francês, respectivamente, são nesta fase utilizados para verificar como é que os mesmos se comportam na classificação de um conjunto de dados desconhecido, nomeadamente o conjunto de dados de teste. A Figura 10 apresenta a *stream* construída para o efeito. Nesta *stream*, os modelos atrás referidos são integrados ao conjunto de dados de teste (Tar1Tes.cdf), sendo o desempenho dos mesmos avaliado através de dois nodos analysis (analysisAle e analysisFran). Estes dois nodos certificam que a percentagem de concordância, entre a realidade explícita nos dados e os modelos, é no caso do Alemão de 46.19% e no caso do Francês de 49.28%.

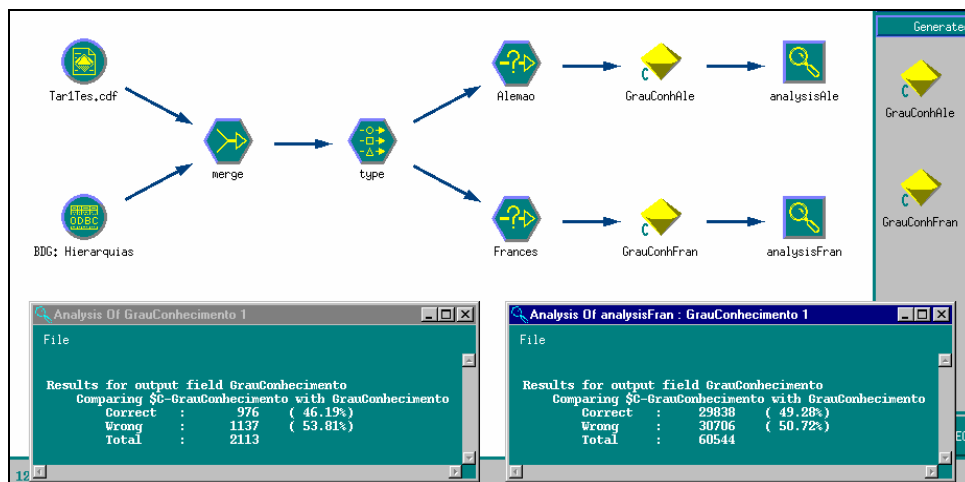


Figura 10 - Análise dos modelos construídos para a caracterização do *perfil linguístico*

A descoberta de conhecimento é um processo iterativo, pelo que nesta fase poderia ser equacionado o retrocesso a etapas anteriores, para, por exemplo, melhorar a confiança dos modelos encontrados. O aumento do tamanho do conjunto de dados de treino poderia conduzir a esta melhoria. Esta hipótese não é aqui explorada, por constituir uma repetição do processo anteriormente apresentado.

A árvore de decisão construída na satisfação da quarta tarefa, e que permitiu a identificação de tendências espaciais nos dados, foi utilizada para classificar dados desconhecidos. Nesta avaliação (Figura 11), utilizando o conjunto de dados de teste, o modelo apresentou uma percentagem de acerto de 47.17%.

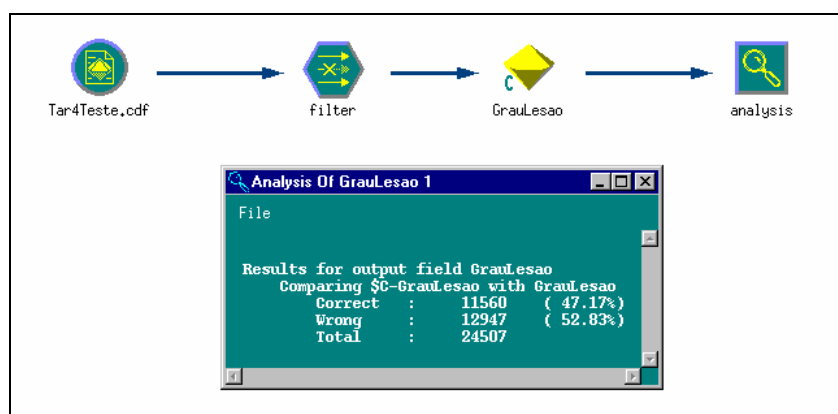


Figura 11 - Desempenho da árvore de decisão que identifica tendências espaciais

Os desempenhos apresentados nesta subsecção, e que sintetizam a confiança dos modelos construídos na fase de DM, quando utilizados na classificação de dados desconhecidos, são influenciados pela distribuição pouco homogênea dos dados analisados. Esta distribuição

dificulta o processo de aprendizagem, e pode condicionar a utilização dos modelos obtidos em tarefas de previsão.

4.5 A Visualização de Resultados

No componente de Visualização de Resultados, que integra o sistema PADRÃO, é possível armazenar na Base de Dados de Padrões (BDP) os modelos encontrados na fase de DM. Este procedimento permite a visualização das regras em mapas das regiões analisadas.

A Figura 12 apresenta o processo de armazenamento das regras, que descrevem a caracterização geográfica do conhecimento do Alemão. Na *stream* apresentada é possível verificar que é construída uma nova tabela (BDP:PerfLingAle) na BDP, a qual possibilita que o conhecimento encontrado durante a fase de DM seja visualizado num mapa. Para tal, recorre-se à utilização da aplicação VisualPadrão, executada a partir de um nodo user input, no qual se identifica a tabela da BDP que armazena as regras a visualizar.

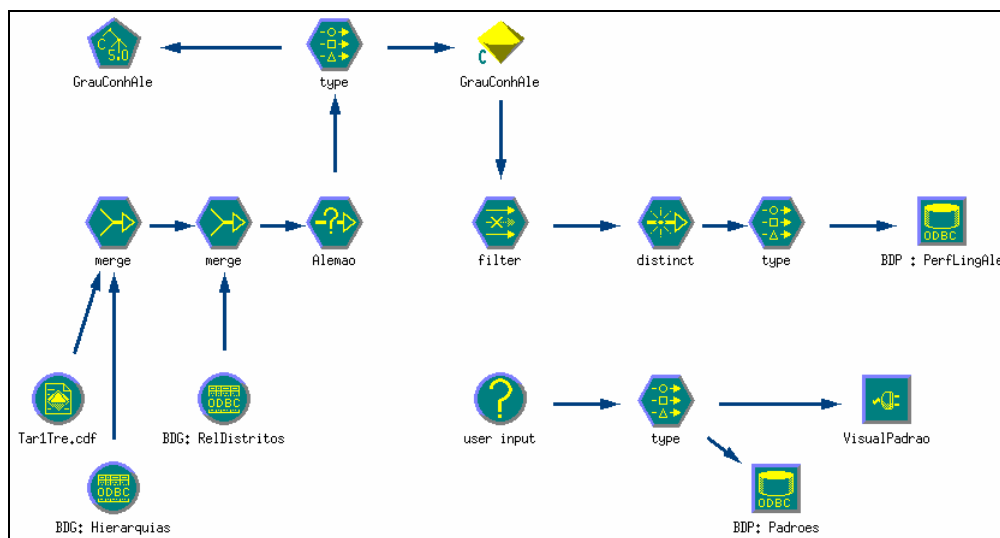


Figura 12 - Transferência das regras que caracterizam o *perfil linguístico* para a BDP

A utilização do VisualPadrão na visualização gráfica dos modelos GrauConhAle e GrauConhFran, conduziu à construção de dois mapas (Figura 13, à esquerda o mapa respeitante ao modelo GrauConhAle e à direita, o mapa que retrata o conhecimento explícito no modelo GrauConhFran), nos quais é possível verificar o grau de conhecimento de cada uma das línguas analisadas. Pela análise dos referidos mapas constata-se que as regiões que apresentam maior grau de conhecimento numa dada língua, verificam um grau de conhecimento menor na outra língua analisada. Os distritos de Lisboa e de Castelo Branco representam as duas exceções desta verificação.

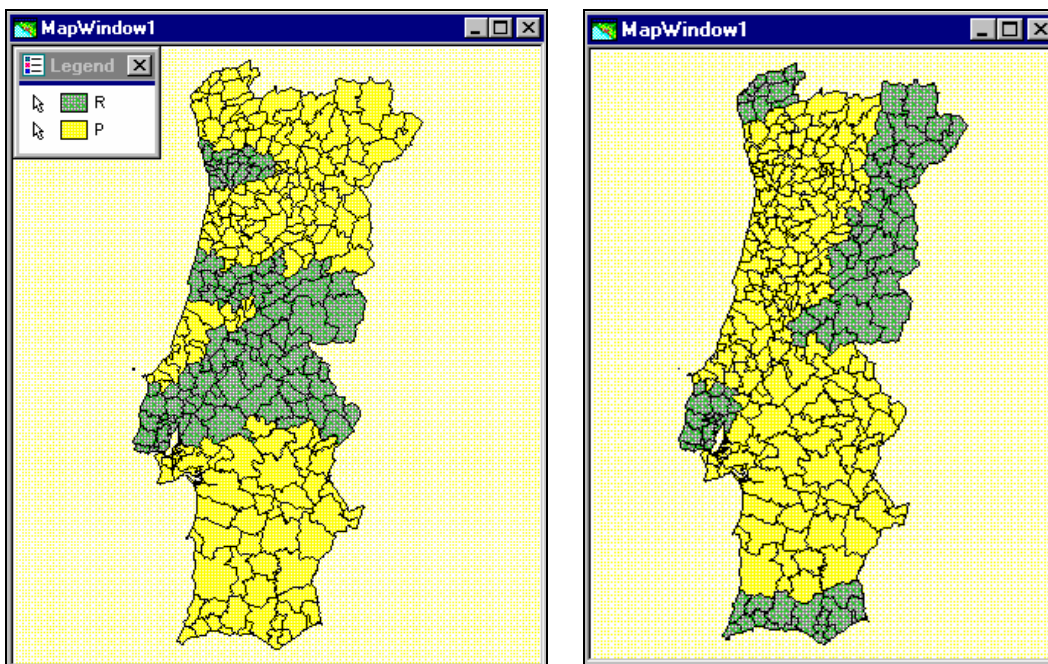


Figura 13 - Caracterização geográfica do conhecimento do Alemão e do Francês

A detecção de tendências espaciais no factor SIVAGE, explícita no modelo GrauLesão construído anteriormente, é nesta subsecção apresentada recorrendo à cartografia da região analisada. A Figura 14 apresenta a *stream* que permitiu a transferência das regras para a BDP, nomeadamente para a tabela TendEsp. A Figura 15 evidencia o mapa da região, na qual é possível constatar que, em relação ao grau de lesão 5, para o factor SIVAGE com o código V, existe uma diminuição do grau de lesão, à medida que aumenta a distância em relação ao concelho 308.

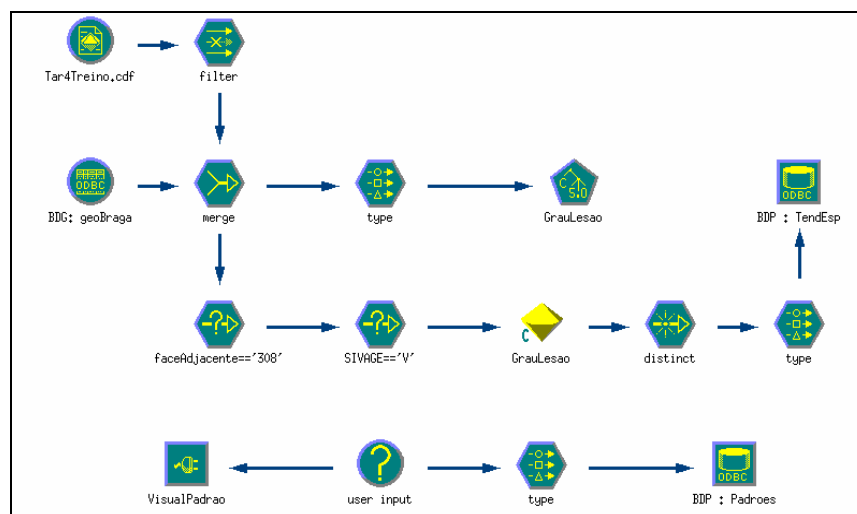


Figura 14 - Transferência das regras que explicitam as tendências espaciais para a BDP

6 Referências

- Egenhofer, M. J., "Deriving the Composition of Binary Topological Relations", *Journal of Visual Languages and Computing*, 5, 2 (1994), 133-149.
- Ester, M., A. Frommelt, H.-P. Kriegel, e J. Sander, *Algorithms for Characterization and Trend Detection in Spatial Databases*, 4th International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1998.
- Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, The MIT Press, Massachusetts, 1996.
- Frank, A. U., "Qualitative Spatial Reasoning about Distances and Directions in Geographic Space", *Journal of Visual Languages and Computing*, 3 (1992), 343-371.
- Freksa, C., *Qualitative Spatial Reasoning*, in Mark, D. M. e A. U. Frank (Eds.), *Cognitive and Linguistic Aspects of Geographic Space*, Kluwer Academic Publishers, 1991.
- Han, J., e M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
- Intergraph, *Geomedia Professional v3, Reference Manual*, Intergraph Corporation, 1999.
- Koperski, K., J. Adhikary, e J. Han, *Spatial Data Mining: Progress and Challenges*, Proc. of the 1996 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, 1996.
- Lu, W., J. Han, e B. C. Ooi, *Discovery of General Knowledge in Large Spatial Databases*, Proc. of the 1993 Far East Workshop on Geographic Information Systems, Singapura, 1993.
- Santos, M., e L. Amaral, "Knowledge Discovery in Spatial Databases: the Padrão's qualitative approach", *Cities and Regions*, GIS special issue, November (2000a), 33-49.
- Santos, M., e L. Amaral, *O Padrão na Descoberta de Conhecimento em Bases de Dados Demográficas*, 1ra. Conferência da Associação Portuguesa de Sistemas de Informação, Guimarães, 25-27 Outubro, Edição em CD-ROM, 2000b.
- Santos, M., e L. Amaral, *A Qualitative Spatial Reasoning Approach in Knowledge Discovery in Spatial Databases*, Data Mining 2000: Data Mining Methods and Databases for Engineering, Finance and Others Fields, Cambridge University, WIT Press, 5-7 July, 2000c.
- Santos, M. Y., *PADRÃO: Um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas*, Tese de Doutorado, Universidade do Minho, 2001.
- SPSS, *Clementine, User Guide, Version 5.2*, SPSS Inc., 1999.