Maribel Yasmina Campos Alves Santos

Padrão

Um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas

Universidade do Minho
2001

Maribel Yasmina Campos Alves Santos

Padrão

Um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas

Tese submetida à Universidade do Minho para a obtenção do grau de Doutor em Tecnologias e Sistemas de Informação, área de conhecimento Engenharia da Programação e dos Sistemas Informáticos.

Universidade do Minho
Departamento de Sistemas de Informação
Escola de Engenharia
Junho de 2001

Projecto parcialmente …nanciado por uma bolsa do PRODEP II, acção 5.2, concurso n.º3/98, Doutoramentos.

A designação PADRÃO, adoptada para o sistema proposto neste trabalho, além de estar associada ao resultado do processo de descoberta de conhecimento, materializado nos modelos ou regras que descrevem padrões (*patterns*) nos dados, traduz uma homenagem aos descobridores portugueses, que em tempos passados assinalaram as suas descobertas através da colocação do *Padrão dos Descobrimentos* nos territórios descobertos.

"Padrões eram marcos em forma de coluna de pedra, terminando na parte superior em forma de cubo, rematado por uma cruz, numa das faces do qual tinha esculpido o escudo de Portugal e, na outra, uma inscrição com a indicação do descobridor e a data do descobrimento. Destinavam-se a garantir a prioridade do descobrimento e da posse dos territórios por parte de Portugal"

(Portugal no Mundo, Séculos XII-XV, Selecções do Reader's Digest, Volume I, p. 519)

Ao Carlos,

ao Bruno e ao Pedro.

”O homem nunca faria nada se esperasse até ser tão perfeito que ninguém encontrasse falhas na sua obra”

Cardeal Newman

Agradecimentos

A realização deste trabalho, apesar da sua natureza individual, foi conseguida com o apoio de diversas pessoas e instituições. A todos manifesto a minha gratidão, deixando aqui expresso um especial agradecimento:

Ao Prof. Doutor Luís Amaral, orientador deste trabalho, pelas sucessivas palavras de incentivo e pelas sabias sugestões sempre proferidas.

Ao Departamento de Sistemas de Informação, e em particular ao saudoso Prof. Doutor Altamiro Barbosa Machado, pela sua visão e constante preocupação com a formação dos seus assistentes. Ao Prof. Doutor João Álvaro Carvalho, por continuar a obra iniciada.

À Universidade do Minho, e em particular à Escola de Engenharia, por terem acolhido e apoiado este trabalho.

Ao Ministério da Defesa Nacional, por ter cedido parte dos dados necessários à realização deste trabalho, e muito especialmente ao Tenente Coronel Rui Afonso do Comando de Pessoal, pelo seu empenho na preparação dos mesmos e pela sua disponibilidade no esclarecimento das dúvidas que surgiram.

Ao Núcleo de Estudos da População e Sociedade, e em particular à Prof. Doutora Norberta Amorim, pela disponibilização de diversas bases de dados para estudo.

Aos meus colegas e amigos do Departamento de Sistemas de Informação, em particular à Ana Alice pelas nossas frequentes discussões e divagações, e ao Miguel Brito e Carlos Sousa Pinto pela amizade sempre demonstrada.

Ao Prof. Doutor Pedro Henriques, cujo apoio incondicional está sempre presente.

À Fátima Rodrigues, pelo seu exemplo de determinação e pelas nossas frequentes conversas.

À minha família, em particular aos meus pais e irmãos, pelo apoio sempre presente.

À D. Felicidade, ao Sr. Rodrigues, à Lena e ao Joca, pelo enorme carinho e atenção que sempre dedicaram ao Bruno e ao Pedro, permitindo que nestes últimos anos, os meus dias fossem dedicados ao doutoramento.

Ao Carlos, pelo seu constante encorajamento, e pela sua dedicação nesta fase tão difícil e importante da minha carreira.

Aos meus …lhos, Bruno e Pedro, por constituírem duas grandes fontes de estímulo, dando-me sempre motivos para continuar.

Resumo

A Descoberta de Conhecimento em Bases de Dados está associada à identi…cação de relacionamentos implícitos existentes nos dados analisados. O processo global de descoberta de conhecimento, que se desenrola em várias fases, 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 geo-referenciados, isto é, dados que incluem referências a objectos geográ…cos, localizações ou partes de uma divisão territorial. A análise destes dados impõe a veri…cação da componente espacial associada aos mesmos (distâncias, direcções, adjacências, ...), e a sua in‡uência nos restantes dados explorados, já que um objecto geográ…co pode ser afectado por acontecimentos veri…cados em objectos vizinhos.

Os algoritmos de Data Mining disponíveis em ferramentas de descoberta de conhecimento tradicionais, que permitem a exploração de dados armazenados em bases de dados relacionais, não estão preparados para a análise desta componente, motivando: i) o desenvolvimento de novos algoritmos; ii) a adaptação de algoritmos já existentes; iii) a utilização de sistemas gestores de bases de dados espaciais ou sistemas de informação geográ…ca, que permitam a incorporação da componente espacial dos dados no processo de descoberta de conhecimento.

A existência nas bases de dados organizacionais de identi…cadores geográ…cos qualitativos, como moradas, os quais possibilitam a geo-referenciação da informação através de sistemas de posicionamento indirecto, conduziu à identi…cação de uma abordagem alternativa à análise de dados espaciais, utilizada neste trabalho, que permite a integração da componente espacial dos dados, no processo de descoberta de conhecimento, através da utilização de estratégias de raciocínio espacial qualitativo.

Os princípios estabelecidos para o Padrão, o sistema proposto nesta tese, representam uma nova abordagem na análise de dados espaciais, que apresenta como vantagens: o facto de permitir utilizar uma diversidade de técnicas de Data Mining, já disponíveis para dados não espaciais; o suprimir a necessidade de caracterização geométrica das entidades geográ…cas referenciadas; e o permitir aos algoritmos de Data Mining analisar simultaneamente dados geo-espaciais e dados não espaciais, não condicionando ou limitando os resultados que podem ser obtidos.

A apresentação de um estudo de caso, com a análise de uma base de dados de grande dimensão, permitiu constatar a utilidade do sistema Padrão na exploração de bases de dados geo-referenciadas, nomeadamente, na identi…cação de relacionamentos implícitos existentes entre os dados geo-espaciais e os dados não espaciais analisados.

Abstract

Knowledge Discovery in Databases is a process that aims the discovery of associations within data sets. Data Mining is the central step of this process. It corresponds to the application of algorithms for identifying patterns within data. Other steps are related to incorporating prior domain knowledge and interpretation of results.

Geo-referenced data sets constitute a special case that demands a particular approach within the knowledge discovery process. Geo-referenced data sets include allusion to geographic objects, locations or administrative sub-divisions of a region. The geographic location and extension of those objects have implicit relationships of spatial neighbourhood. The Data Mining algorithms have to take this spatial neighbourhood into account when looking for associations among data.

Data Mining algorithms available in traditional knowledge discovery tools, developed for the analysis of relational databases, are not prepared for the analysis of this spatial component. This situation led to: i) the development of new algorithms capable of dealing with spatial relationships; ii) the adaptation of existing algorithms in order to enable them no deal with those spatial relationships; iii) the integration of the capabilities for spatial analysis of spatial database management systems or geographic information systems with the tools normally used in the knowledge discovery process.

Most of the geographic attributes normally found out in organisational databases (e.g., addresses) correspond to a type of spatial information that can be described using indirect positioning systems.

This work proposes a new approach - the Padrão system - to the analysis of spatial data based on qualitative spatial reasoning strategies that allow the integration of the spatial component in the knowledge discovery process. The main advantages of this approach include: the use of already existing Data Mining algorithms applied to the analysis of non-spatial data; avoid the geometric characterisation of spatial objects; and enable that Data Mining algorithms deal with geo-spatial and non-spatial data simultaneously thus imposing no limits and constraints to the results achieved.

The e¢cacy and usefulness of Padrão has been tested with a case study where a large database has been subject to a knowledge discovery process. The results con…rm that Padrão enables the identi…cation of implicit relationships among geo-spatial and non-spatial data.

Índice

Índice de Figuras

Índice de Tabelas

Siglas

Neste documento são adoptadas diversas siglas, que representam abreviaturas de designações utilizadas. As siglas empregues são:

BCE ..... Base de Conhecimento Espacial
BD ..... Base de Dados
BDE ..... Base de Dados Espacial
BDG ..... Base de Dados Geográ...ca
BDnG ..... Base de Dados não Geográ...ca
BDP ..... Base de Dados de Padrões
CEN ..... Comité Europeu de Normalização
CLARANS ..... Clustering Large Applications based on RANdomized Search
CLEM ..... Clementine Language for Expression Manipulation
CNIG ..... Centro Nacional de Informação Geográ...ca
CRISP-DM ..... CRoss Industry Standard Process for Data Mining
DCBD ..... Descoberta de Conhecimento em Bases de Dados
DCBDE ..... Descoberta de Conhecimento em Bases de Dados Espaciais
DGA ..... Direcção Geral do Ambiente
DM ..... Data Mining
DME ..... Data Mining Espacial
DSI ..... Departamento de Sistemas de Informação
E-R ..... Entidades e Relacionamentos
GMQL ..... Geo-Mining Query Language
GPL ..... Graphical Presentation Language
GPS ..... Sistema Global de Posicionamento (Global Positioning System)
HYPERGEO ..... Easy and friendly access to geographic information for mobile users
ILP ..... Programação Lógica Indutiva (Inductive Logic Programing)
INE ..... Instituto Nacional de Estatística

IPQ ..... Instituto Português da Qualidade
ISO ..... Organização Internacional de Normalização (International Standard Organization)
IST ..... Information Society Tecnologies
KNOMAD . KNOwledge Management And Discovery for Distributed Geographic Information Systems
MBR ..... Minimum Bounding Rectangle
NSD ..... Non-Spatial Dominant
ODBC ..... Open Database Connectivity
OGIS ..... Open Geodata Interoperability Speci…cation
OO ..... Orientado aos Objectos (Object Oriented)
OpenGIS ..... Open GIS Consortium
SAND ..... Spatial And Non-spatial Data
SD ..... Spatial Dominant
SDC ..... Sistema de Descoberta de Conhecimento
SEED ..... Sistema de Estudo para a Evolução Demográ…ca
SGBD ..... Sistema Gestor de Bases de Dados
SGBDE ..... Sistema Gestor de Bases de Dados Espaciais
SIAPE ..... Sistema de Informação de Administração do Pessoal do Exército
SIG ..... Sistema de Informação Geográ…ca
SNIG ..... Sistema Nacional de Informação Geográ…ca
SQL ..... Structured Query Language
TC ..... Comissão Técnica (Technical Comission)
TI ..... Tecnologias da Informação
UML ..... Uni…ed Modeling Language
VB ..... Visual Basic
WG ..... Grupo de Trabalho (Working Group)
WWW ..... World Wide Web
XML ..... eXtensible Markup Language

Capítulo 1

Introdução

O volume de dados armazenados e manipulados pela maioria das organizações cresce diariamente a uma taxa que ultrapassa a nossa capacidade de analisar, sintetizar e extrair conhecimento a partir desses dados. Apesar dos Sistemas Gestores de Bases de Dados (SGBD) fornecerem mecanismos capazes de armazenar e utilizar grandes quantidades de dados, o uso de ferramentas especí…cas, concebidas e implementadas com o objectivo de automatizar o processo de análise de grandes quantidades de dados, é cada vez mais necessário.

Este contexto justi…ca a existência de uma área de investigação, a descoberta de conhecimento em bases de dados, a qual está intimamente ligada à inteligência arti…cial através do desenvolvimento de algoritmos inteligentes que automatizem o processo de análise dos dados.

A Descoberta de Conhecimento em Bases de Dados[1] (DCBD) é genericamente de…nida como "o processo não trivial de identi…cação de padrões válidos e potencialmente úteis, perceptíveis a partir dos dados" ([Fayyad et al., 1996b] p.6).

O processo global de descoberta de conhecimento, que se desenrola em várias fases, inclui a gestão dos algoritmos de Data Mining[2] (DM), utilizados para extrair padrões dos dados, e a interpretação dos padrões encontrados pelos mesmos. As ferramentas de DCBD utilizam uma diversidade de algoritmos para identi…car relacionamentos e padrões que estão implícitos nos dados. Estes representam conhecimento acerca da Base de Dados (BD) explorada e das entidades nela contidas. Decidir se os achados re‡ectem ou não conhecimento útil, é uma das fases do processo na qual a participação do utilizador é requerida [Fayyad et al., 1996a].

Os grandes progressos conseguidos até ao momento na área da DCBD têm-se restringido quase que exclusivamente [Koperski et al., 1996] à exploração de dados armazenados em BD relacionais. Existe, contudo, na maioria das BD organizacionais uma dimensão dos dados, a espacial (normalmente explícita através de uma morada, código postal, etc.), cuja semântica não é interpretada pelos algoritmos de DM tradicionais[3].

Bases de Dados Espaciais (BDE) são, comummente, BD relacionais que integram no seu modelo de dados estruturas especí…cas, que permitem o armazenamento da localização espacial

[1] Em inglês, Knowledge Discovery in Databases.

[2] Ao longo deste documento é adoptada a designação em inglês deste conceito, por se considerar que as traduções por vezes adoptadas, Mineração de Dados ou Minagem de Dados, não conseguem transmitir a semântica associada à designação original.

[3] Desenvolvidos para a exploração de dados armazenados em BD relacionais.

e da dimensão de entidades geográficas [Ester et al., 1997]. Os atributos que caracterizam a componente espacial das entidades definem relações de vizinhança, já que uma entidade pode ser afectada pelo comportamento de entidades vizinhas. Os algoritmos de DM disponíveis em ferramentas de DCBD tradicionais não estão preparados para a análise desta componente, motivando o desenvolvimento de novos algoritmos capazes de integrar a componente espacial dos dados, no processo de descoberta de conhecimento.

O DM Espacial (DME) é considerado uma sub-área do DM, dedicado à "descoberta de conhecimento implícito existente entre dados espaciais e dados não espaciais, relacionamentos espaciais, ou outros padrões não explícitos em BDE" [Han et al., 1997].

A análise de dados espaciais com o objectivo de descoberta de conhecimento requer a utilização de técnicas específicas, que permitam a inclusão da semântica espacial, implícita na posição e dimensão dos objectos geográficos referenciados, no referido processo. As técnicas existentes [Koperski et al., 1996] baseiam-se, essencialmente, no desenvolvimento de novos algoritmos de DM capazes de incluir a semântica espacial no processo de descoberta de conhecimento, ou, na integração de Sistemas Gestores de BDE (SGBDE) com ferramentas de descoberta de conhecimento, os quais permitem a manipulação dos dados espaciais e consequente junção dos resultados com os restantes dados não espaciais (para análise na ferramenta de descoberta de conhecimento).

Estas abordagens requerem a descrição geométrica dos diversos objectos geográficos referenciados, uma vez que se baseiam em estratégias quantitativas de raciocínio espacial, que manipulam as coordenadas dos pontos que descrevem as diversas entidades geográficas.

Inúmeras organizações possuem BD nas quais a dimensão espacial é referenciada recorrendo a identificadores geográficos qualitativos, como moradas, para os quais as mesmas não necessitam (e como tal não possuem) no seu funcionamento diário, da descrição geométrica dos objectos geográficos referenciados.

A existência destes identificadores qualitativos, nas BD, permite a utilização de sistemas indirectos de posicionamento geográfico, que suprimem a necessidade de caracterização geométrica das diversas entidades geográficas referenciadas, e ainda, evitam o desenvolvimento de novos algoritmos de DM para análise dos dados, ou a utilização de SGBDE para manipulação dos dados espaciais.

A incorporação da componente espacial no processo de DCBD, passa neste trabalho pela estruturação de uma BD Geográfica (BDG), construída recorrendo às pré-normas europeias para informação geográfica [CEN/TC-287, 1996c], que armazena parte[4] dos relacionamentos espaciais[5] existentes entre as entidades geográficas utilizadas na geo-referenciação da informação.

Estas entidades encontram-se estruturadas num sistema de posicionamento indirecto[6] [CEN/TC-287, 1998h]. Nestes sistemas, uma posição é indexada a uma localização no espaço, através da identificação de um objecto real. Os identificadores utilizados para referenciar estes

[4] Como será constatado no Capítulo 2, subsecção 2.4.2, apenas é possível especificar os relacionamentos espaciais existentes entre entidades geográficas adjacentes. Os restantes relacionamentos podem ser inferidos, recorrendo aos princípios do raciocínio espacial qualitativo utilizados neste trabalho.

[5] Definidos recorrendo a identificadores qualitativos, como Norte, próximo, etc.

[6] Um sistema de posicionamento indirecto não recorre a utilização de coordenadas de pontos, na geo-referenciação da informação.

objectos são denominados de identi...cadores geográ...cos[7].

A utilização de identi...cadores geográ...cos, na geo-referenciação da informação, é uma constante nas BD organizacionais. Tal facto permite, na abordagem proposta neste trabalho (Figura 1.1), a incorporação dos relacionamentos espaciais existentes entre as entidades geográ-...cas endereçadas, no processo de descoberta de conhecimento.

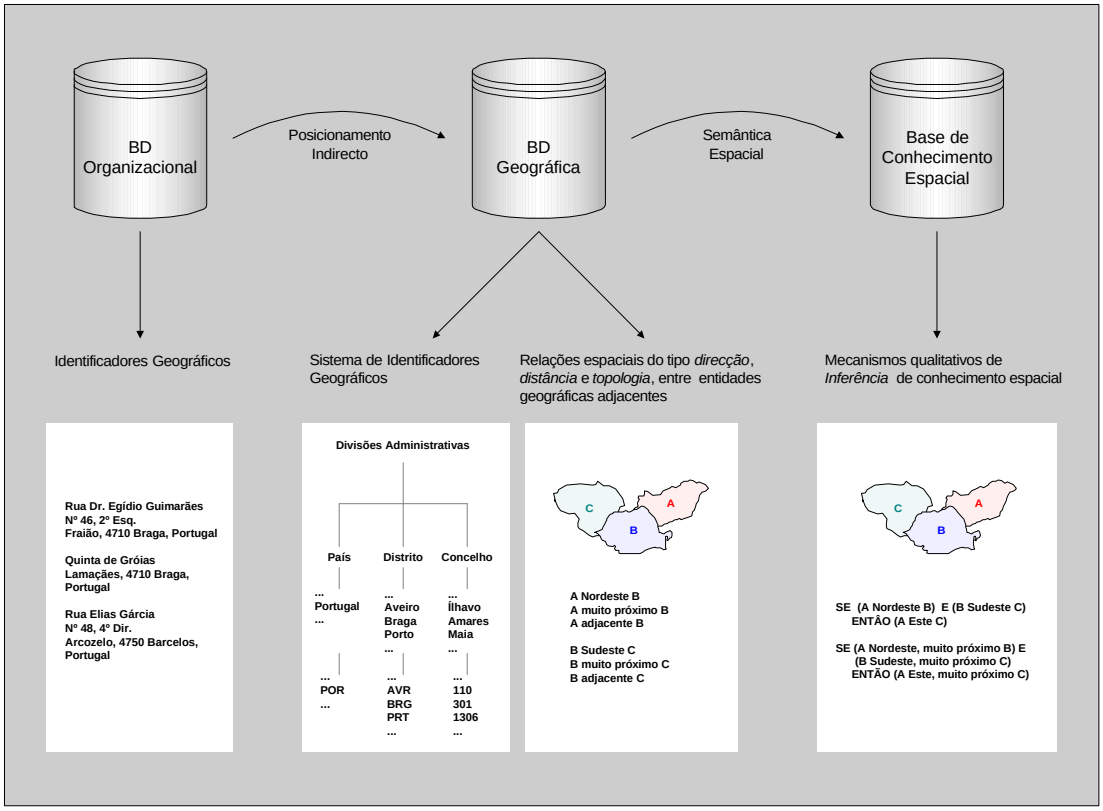


Figura 1.1: Integração da componente espacial no processo de descoberta de conhecimento

Para que o raciocínio espacial seja possível, isto é, para que a semântica espacial associada aos identi...cadores geográ...cos seja efectivamente integrada no processo de descoberta de conhecimento, é necessário utilizar estratégias de raciocínio espacial qualitativo que permitam raciocinar com informação geográ...ca incompleta ou imprecisa.

O raciocínio espacial qualitativo constitui um mecanismo automático de derivação de relações espaciais não explícitas em BDG [Abdelmoty e El-Geresy, 1995]. É baseado na manipulação de identi...cadores qualitativos como Norte, Sul, próximo, distante, adjacente, etc., evitando a utilização de informação quantitativa, como coordenadas ou distâncias entre pontos. Neste projecto, relações espaciais do tipo direcção, distância e topologia são integradas segundo os princípios do raciocínio espacial qualitativo, permitindo a construção de uma Base de Conhecimento Espacial (BCE). Esta BCE é utilizada, no processo de descoberta de conhecimento, na inferência de relacionamentos espaciais desconhecidos, necessários aos algoritmos de DM.

[7] Neste contexto inserem-se nomes de ruas, monumentos, cidades, etc. [CEN/TC-287, 1998h].

Defende-se neste trabalho que é possível a utilização de algoritmos de DM tradicionais, na exploração da componente espacial associada às entidades geográficas referenciadas nas BD organizacionais. Esta componente é integrada, no processo de descoberta de conhecimento, através da BDG construída. Esta BDG, que caracteriza o sistema de posicionamento indirecto utilizado, é manipulada recorrendo aos princípios do raciocínio espacial qualitativo.

A concepção, implementação[8] e validação de um sistema de descoberta de conhecimento em BD geo-referenciadas, baseado em mecanismos de referenciação indirectos e de raciocínio espacial qualitativo, torna-se a principal finalidade deste trabalho.

O Padrão, o sistema proposto, não inclui o desenvolvimento de novos algoritmos de DM adaptados à componente espacial dos dados. Inclui, sim, o aproveitamento das capacidades de análise exploratória de dados conseguidas até ao momento, pelas ferramentas de DCBD já desenvolvidas para BD relacionais[9]. O Padrão foi implementado no Clementine[10] [SPSS, 1999b] [SPSS, 1999a], uma ferramenta que permitiu a assimilação dos princípios qualitativos utilizados no raciocínio espacial. Estes foram posteriormente utilizados na identificação de relacionamentos espaciais, existentes entre os dados geo-espaciais e os dados não espaciais analisados no processo de descoberta de conhecimento.

As opções estruturais que caracterizam o sistema Padrão, mecanismos de referenciação indirectos e estratégias de raciocínio espacial qualitativo, representam uma nova abordagem na análise de dados geo-espaciais e dados não espaciais com o objectivo de descoberta de conhecimento, constituindo esta a principal contribuição deste trabalho. A estratégia utilizada permite que dados organizacionais sejam analisados, independentemente da disponibilidade de algoritmos de DM específicos (isto é, capazes de integrar a componente espacial no processo de descoberta de conhecimento) e da geometria dos objectos geográficos referenciados. Outras vantagens desta aproximação estão associadas ao facto de esta abordagem permitir utilizar uma diversidade de algoritmos de DM (já disponíveis em ferramentas de descoberta de conhecimento tradicionais), possibilitando o uso de um vasto conjunto de técnicas na análise dos dados, e principalmente, permitir aos algoritmos de DM analisar simultaneamente dados geo-espaciais e dados não espaciais, não condicionando ou limitando os resultados que podem ser obtidos.

1.1 Motivações, objectivos e contribuições fundamentais

Apesar da evolução tecnológica ocorrida ao nível da capacidade de armazenamento de dados (máquinas cada vez mais acessíveis economicamente) e na velocidade de acesso aos mesmos (processadores cada vez mais rápidos), as organizações continuam a não ter capacidade de ana-

[8] Implementação é o termo utilizado ao longo deste trabalho para designar o conjunto de tarefas que permitem a execução ou realização física de algo, neste caso, a construção de sistemas informáticos.

[9] Goebel e Gruenwald [Goebel e Gruenwald, 1999] comparam diversas ferramentas de descoberta de conhecimento, verificando entre outras características, a quantidade máxima de dados que conseguem manipular, o tipo de tarefas que permitem executar e as técnicas de DM disponíveis nas mesmas.

[10] Na avaliação efectuada por Abbott et al. [Abbott et al., 1998] a 5 ferramentas de descoberta de conhecimento (o Intelligent Miner da IBM, o Darwin da Thinking Machines, o Enterprise Miner do SAS Institute, o Pattern Recognition Workbench da Unica Technologies e o Clementine da SPSS), refere-se que o Clementine obteve a melhor classificação no parâmetro compreensão dos modelos, sendo a segunda classificada nos restantes três parâmetros analisados, carregamento e manipulação dos dados, construção dos modelos e suporte técnico. Estes quatro parâmetros permitiram aos autores avaliar a facilidade de utilização das ferramentas exploradas.

lisar, em tempo útil, tais dados. Tal deve-se à enorme quantidade de dados armazenados e ao facto deste volume crescer diariamente, a uma taxa que ultrapassa a capacidade humana de análise e síntese dos mesmos, para suporte à tomada de decisão.

A procura de uma solução para este problema conduziu à constatação de que a investigação na área da DCBD veri…ca um considerável progresso [Fayyad et al., 1996c], o qual não é tão acentuado no que diz respeito à exploração de BDG [Koperski et al., 1996]. É neste contexto que deriva a motivação deste trabalho, no qual se pretende conceber, implementar e validar um sistema de descoberta de conhecimento, que permita a análise de BD geo-referenciadas. A referenciação geográ…ca da informação é conseguida através de mecanismos de posicionamento indirecto, pelo que o Padrão recorre aos princípios do raciocínio espacial qualitativo para a inferência de relações espaciais requeridas no processo de descoberta de conhecimento. Estes princípios, depois de devidamente assimilados pelo Clementine, são utilizados na descoberta de padrões que evidenciam os relacionamentos implícitos existentes entre os dados[11] analisados.

A integração da componente espacial, associada aos identi…cadores geográ…cos utilizados na geo-referenciação da informação, no processo de descoberta de conhecimento, constitui a principal contribuição deste trabalho. A abordagem proposta permite a inclusão da semântica espacial no referido processo, através de estratégias de raciocínio qualitativo, e suprime a necessidade de desenvolvimento de novos algoritmos de DM, capazes de lidar com esta componente.

De acordo com a …nalidade deste trabalho, é possível formular um conjunto de sete objectivos a atingir, bem como os principais resultados e contributos associados à realização de cada um dos mesmos. A Figura 1.2 sistematiza o conjunto dos objectivos, os quais são de seguida descritos.

O primeiro objectivo consiste em enquadrar e clari…car os conceitos associados ao domínio geo-espacial, frequentemente utilizados ao longo deste documento. Além das várias de…nições apresentadas (informação geográ…ca, sistema de informação geográ…ca, ...), sistematizam-se os principais modelos de dados utilizados na modelação de aplicações geográ…cas e ainda os mecanismos de armazenamento e manipulação de informação geográ…ca. Ao nível das iniciativas em curso no âmbito da normalização da informação geográ…ca, descrevem-se os grupos de trabalho CEN TC 287[12] e ISO TC 211[13], dando particular ênfase ao CEN TC 287 por ser o adoptado neste projecto. Como principais resultados associados à realização deste objectivo, destaca-se o enquadramento conceptual efectuado ao domínio geo-espacial, e ainda o papel das pré-normas CEN TC 287 na estruturação da BDG utilizada neste trabalho.

O segundo objectivo é o de rever os fundamentos teóricos subjacentes ao raciocínio espacial qualitativo. Neste contexto, e além da análise dos seus princípios, veri…cam-se as diferentes formas de representação da informação, os tipos de relações espaciais existentes e as diversas estratégias de raciocínio que podem ser adoptadas. Da revisão efectuada à literatura, destacam-se como principais resultados da execução deste objectivo:

- o enquadramento conceptual dos princípios do raciocínio espacial qualitativo. Este enquadramento conjuga a descrição das relações espaciais utilizadas neste trabalho (direcção, distância e topologia), com as estratégias de raciocínio homogénea, heterogénea e integra-

[11] Geográ…cos e não geográ…cos.
[12] Comité Europeu de Normalização (CEN), Comissão Técnica 287.
[13] Organização Internacional de Normalização (ISO), Comissão Técnica 211.

Problema: As bases de dados organizacionais armazenam identificadores geográficos, cuja componente espacial, implícita na localização e dimensão das entidades geográficas referenciadas, não é analisada por algoritmos de DM tradicionais.

Tese: É possível integrar, no processo de descoberta de conhecimento, a componente espacial dos dados através de estratégias de raciocínio espacial qualitativo.

Finalidade: Concepção, implementação e validação de um sistema de descoberta de conhecimento em bases de dados geo-referenciadas, baseado em mecanismos de posicionamento indirecto e estratégias de raciocínio espacial qualitativo.

Objectivos	**Tarefas**	**Resultados e Contribuições**
Revisão dos conceitos associados ao domínio geo-espacial.	Revisão teórica/bibliográfica	▪ Enquadramento conceptual do domínio geo-espacial: - Tecnologia de bases de dados espaciais; - Modelação da informação geográfica. ▪ Iniciativas de normalização em curso na área da Informação Geográfica. ▪ Aplicabilidade do esquema espacial e do esquema de identificadores geográficos, das pré-normas CEN TC 287.
Revisão dos princípios subjacentes ao raciocínio espacial qualitativo.	Revisão teórica/bibliográfica	▪ Enquadramento conceptual dos princípios subjacentes ao raciocínio espacial qualitativo: - Representação qualitativa da informação; - Tipos de relações espaciais; - Estratégias de raciocínio. ▪ Sistema de inferências que integra a *direcção*, *distância* e *topologia*, segundo os princípios do raciocínio espacial qualitativo.
Revisão da literatura associada à descoberta de conhecimento em bases de dados.	Revisão teórica/bibliográfica	▪ Enquadramento conceptual da descoberta de conhecimento em bases de dados: - Fases do processo; - O conhecimento do domínio e as principais dificuldades do processo. - Tarefas e técnicas de DM. ▪ Principais abordagens na descoberta de conhecimento em bases de dados geo-espaciais.
Concepção da arquitectura do sistema	Construção do PADRÃO	▪ Arquitectura do PADRÃO
Implementação do sistema		▪ Aplicação PADRÃO
Validação do sistema		▪ Sistema PADRÃO
Promoção da utilização do PADRÃO e da sua evolução conceptual.	Formulação e proposta de trabalho futuro	▪ Projectos de trabalho futuro

Figura 1.2: Objectivos, resultados e contribuições fundamentais

da que podem ser aplicadas sobre as mesmas. Para a abordagem integrada, analisaram-se com particular detalhe dois sistemas de raciocínio. O primeiro integra relações espaciais do tipo direcção e distância; e o segundo, relações espaciais do tipo direcção e topologia. Ambos permitiram:

2 a construção de um sistema de raciocínio qualitativo que integra relações espaciais do tipo direcção, distância e topologia, na inferência de relações espaciais desconhecidas.

O terceiro objectivo consiste em rever a literatura associada ao processo de DCBD. Como principais resultados descrevem-se os fundamentos associados ao referido processo, tais como: as fases que o caracterizam; a importância do conhecimento do domínio de aplicação no processo de descoberta de conhecimento; as tarefas levadas a cabo pelos algoritmos de DM e as principais técnicas utilizadas pelos mesmos. Sistematizam-se, ainda, as principais abordagens à descoberta

de conhecimento em BD geo-espaciais, assim como os principais resultados alcançados pelas mesmas. Esta sistematização permite diferenciar a abordagem proposta nesta tese, relativamente às outras iniciativas em curso nesta área.

O quarto objectivo consiste em conceber a arquitectura do Padrão. A concepção deste sistema de descoberta de conhecimento é baseada na utilização de mecanismos de referenciação indirectos, os quais são complementados com estratégias de raciocínio qualitativo, que permitem a inferência de informação espacial desconhecida e necessária aos algoritmos de DM.

O quinto objectivo prende-se com a implementação do sistema Padrão. Na sua implementação utilizou-se o Clementine como ferramenta de assimilação dos princípios qualitativos adoptados, e ainda como "fonte" de algoritmos de DM utilizados na exploração dos dados. Ao nível das BD, todos os dados, espaciais e não espaciais, foram armazenados em BD relacionais, estando acessíveis via ligações ODBC (Open DataBase Connectivity). Ao nível geográ…co, e essencialmente para a visualização de resultados, utilizou-se o Sistema de Informação Geográ…ca (SIG) Geomedia Professional.

O sexto objectivo consiste em validar o sistema proposto, veri…cando o desempenho do Padrão na descoberta de conhecimento em BD geo-referenciadas. A avaliação efectuada permitiu veri…car a qualidade das inferências obtidas ao nível geográ…co e a aptidão para detectar relacionamentos implícitos nos dados. Esta validação preliminar do sistema foi posteriormente complementada com um estudo de caso, no qual foi analisada uma BD organizacional de grande dimensão. Este estudo de caso permitiu validar a utilidade do sistema na análise de BD geo-referenciadas e, ainda, sistematizar o tipo de conhecimento que o sistema proposto permite identi…car.

O sétimo objectivo visa promover a utilização e evolução conceptual e tecnológica do Padrão, através da formulação e proposta de projectos de trabalho futuro. Apesar dos resultados deste objectivo não poderem ser avaliados no âmbito desta tese, a convicção de que a utilização do Padrão é de extrema importância para as organizações, na identi…cação dos relacionamentos implícitos existentes nos dados armazenados, é por si só um dos resultados mais esperados deste trabalho.

1.2 Metodologia de investigação

A adopção de uma determinada metodologia de investigação torna-se fundamental num projecto de doutoramento, uma vez que estas permitem auxiliar o desenvolvimento do trabalho, ao mesmo tempo que fornecem directivas para a sua correcta execução e validação.

O Padrão, o sistema proposto neste trabalho, assenta na integração de diversas tecnologias e conceitos, cuja utilização conjunta permite a construção de uma solução para a resolução de um dado problema. A validação do sistema passa essencialmente pela veri…cação da qualidade das inferências obtidas através dos mecanismos qualitativos de raciocínio espacial, pela averiguação da capacidade de identi…cação de relacionamentos implícitos nos dados e pela con…rmação da utilidade do sistema na exploração de BD organizacionais.

Esta tese de doutoramento está inserida na área de conhecimento da Engenharia da Programação e dos Sistemas Informáticos. A adopção de determinado método cientí…co (baseado na experimentação) deverá permitir validar o sistema proposto, à luz das experiências levadas a

cabo com o mesmo. Na validação tecnológica, e apesar da experimentação apenas comprovar a existência de erros, e nunca a sua ausência, esta constitui a única forma possível de validação, quando não existem disponíveis mecanismos dedutivos, que permitam a previsão de fenómenos a partir de uma teoria. No caso da validação tecnológica, a experimentação permite a indução de teorias, já que estas são formuladas a partir da observação [Tichy, 1998].

As ciências tradicionais utilizam a experimentação como mecanismo de validação de teorias e métodos, uma vez que esta disponibiliza um ciclo de interacção, que permite aos cientistas validar, modi…car e melhorar determinada teoria [Zelkowitz e Wallace, 1997]. Ao nível da validação de tecnologia, Zelkowitz e Wallace [Zelkowitz e Wallace, 1998] agrupam os doze métodos de validação que podem ser utilizados, project monitoring, case study, assertion, …eld study, literature search, legacy, lessons learned, static analysis, replicated, synthetic, dynamic analysis e simulation, em três categorias, observação, histórica e controlada.

Dos doze métodos referidos anteriormente, seleccionaram-se a pesquisa bibliográ…ca (literature search), a asserção (assertion) e o estudo de caso (case study), como os métodos cientí…cos utilizados na investigação apresentada nesta tese de doutoramento.

A pesquisa bibliográ…ca, na qual a recolha de dados é baseada na análise de artigos e outros documentos públicos, é neste trabalho utilizada na construção do enquadramento conceptual que permite aos leitores um entendimento comum sobre os temas e conceitos aqui utilizados, ao mesmo tempo que conduz à tradicional revisão do estado da arte, essencial numa tese de doutoramento. Este método está inserido na categoria histórica. Ao nível da observação, utiliza-se a asserção e o estudo de caso.

A asserção constitui uma forma simples de validação (que não requer a satisfação de normas cientí…cas rigorosas), que permite julgar a e…cácia da experimentação efectuada pelo investigador. Constitui o método de validação mais utilizado na avaliação de tecnologia, uma vez que o investigador assume o papel de avaliador, sendo também o responsável pelo desenvolvimento da tecnologia (tornando-se, também, alvo de avaliação). A asserção representa uma avaliação preliminar até que uma validação mais rigorosa da e…cácia da tecnologia seja efectuada. Quando a tecnologia desenvolvida é utilizada num projecto real, no qual o investigador não tem o mesmo grau de controlo sobre as condições experimentais, o método de validação é designado de estudo de caso.

O estudo de caso é caracterizado neste projecto pela análise de uma BD organizacional de grande dimensão, permitindo avaliar a utilidade do Padrão para as organizações.

1.3 Enquadramento institucional

A investigação na área da DCBD e na área da informação geográ…ca, integradas neste trabalho, permitiu a consolidação de conhecimentos nestes domínios e a sua posterior transferência para projectos de Investigação e Desenvolvimento, nos quais participa o Departamento de Sistemas de Informação (DSI) da Universidade do Minho, departamento que acolheu a realização do trabalho descrito nesta tese de doutoramento.

Estas duas áreas de interesse, e con…rmando os benefícios da sua integração, relacionam-se em diversos projectos em curso neste departamento. A referência[14] aos mesmos é, de seguida,

[14] São descritos os projectos em que estas duas áreas de interesse surgem integradas, e ainda, os projectos em

realizada pela sequência cronológica em que os mesmos surgiram.

Se é um facto que os princípios subjacentes à DCBD bene…ciam inúmeros domínios de aplicação, caracterizados por armazenarem grandes quantidades de dados, a Demogra…a Histórica foi um dos primeiros a ser trabalhado na Universidade do Minho, e consequentemente a bene…ciar da tecnologia disponibilizada por esta área do saber. O projecto SEED (Sistema de Estudo para a Evolução Demográ…ca) conduziu à concepção de uma arquitectura heterogénea para a extracção de conhecimento a partir de dados [Rodrigues, 2000], na qual foi incluído um SIG para auxiliar os historiadores demógrafos na análise espacial dos dados armazenados.

Este domínio de aplicação con…rmou a importância da componente geográ…ca na descoberta de conhecimento e consequente automatização do processo de análise espacial, através da utilização de algoritmos de aprendizagem automática.

O Núcleo de Estudos da População e Sociedade, da Universidade do Minho, parceiro neste trabalho, dedica-se à recolha, organização e análise dos registos demográ…cos existentes em diversas paróquias [Amorim, 1992] [Amorim e Correia, 1999]. Uma das BD construídas, integrando diversas paróquias do distrito de Aveiro, foi analisada pelo Padrão [Santos e Amaral, 2000a] [Santos e Amaral, 2000d] [Santos e Amaral, 2000c] [Santos e Amaral, 2000b], permitindo a identi…cação de relações implícitas, existentes entre os dados demográ…cos e os dados geoespaciais analisados.

Refere-se que o SEED surgiu em …nais de 1997, e que a sua concepção e implementação é apresentada por Rodrigues [Rodrigues, 2000] na sua tese de doutoramento. O trabalho desenvolvido permitiu a formulação de novas áreas de actuação para este trabalho [Amorim et al., 2001], sistematizadas num projecto submetido em Fevereiro de 2000 ao programa SAPIENS, do Ministério da Ciência e Tecnologia, o qual obteve a sua aprovação em Janeiro de 2001.

Em Junho de 1999 o DSI passa a integrar o conjunto de parceiros do projecto HYPERGEO (Easy and friendly access to geographic information for mobile users), submetendo, ao programa IST (Information Society Technologies) da Comunidade Europeia, o projecto para respectiva apreciação. Os parceiros que integram o consórcio, além do DSI, são: a Matra Systemes & Information S.A. (França), a Nouvelles Frontieres España S.A. (Espanha), o Institut Jozef Stefan (Eslovénia), a University of Education in Hradec Králové (República Checa), a CAS Software AG (Alemanha), a Aristotle University of Thessaloniki (Grécia) e a City University (Inglaterra).

O HYPERGEO tem como principal …nalidade a disponibilização, a utilizadores móveis, de informação sobre turismo, contextualizada geogra…camente. A localização dos indivíduos é obtida recorrendo a um sistema GPS (Global Positioning System), permitindo a identi…cação de determinado contexto geográ…co. Para este contexto, e atendendo às preferências do utilizador de…nidas previamente, é enviada uma resposta ao terminal do utilizador, contendo as informações solicitadas pelo mesmo (que podem ser sobre hotéis, restaurantes, percursos, etc.).

O registo do per…l dos utilizadores, das informações solicitadas anteriormente, assim como das escolhas realizadas pelos mesmos nas respostas anteriormente formuladas pelo HYPERGEO, permite a um sistema de DM melhorar os mecanismos de pesquisa, fornecendo ao utilizador uma resposta cada vez mais limitada e próxima daquilo que pretende (uma vez que é baseada no comportamento anterior veri…cado pelo mesmo). Este projecto iniciou-se em Janeiro de 2000, e

---

que as mesmas não se relacionam, mas que apresentam uma forte componente de uma destas áreas, informação geográ…ca ou DCBD.

a sua conclusão está prevista para Dezembro de 2001.

Uma outra oportunidade para aplicar os conceitos subjacentes à DCBD é estabelecida com o Departamento de Engenharia Têxtil da Universidade do Minho. A identi…cação da relação existente entre as propriedades físicas e as propriedades químicas dos …os de algodão representa um enorme desa…o para os investigadores desta área, não só pela di…culdade da tarefa, como também pela quantidade de informação envolvida. Assim surge o projecto Propriedades do Algodão: inferência através de técnicas de DM, submetido em Fevereiro de 2000 ao programa SAPIENS. O início dos trabalhos ocorreu em Outubro de 2000, estando o projecto …nanciado até Setembro de 2003.

O projecto KNOMAD (Knowledge Management and Discovery for Distributed Geographic Information Systems) tem como principal …nalidade a concepção e implementação de um ambiente distribuído para a gestão e descoberta de conhecimento em SIG. Este trabalho visa suportar as necessidades de organizações estatísticas, quer estas operem ao nível internacional, nacional ou mesmo local. A harmonização e integração de dados estatísticos provenientes de diversas fontes representa um dos graves problemas destas instituições, pelo que o KNOMAD se propõe construir um protótipo que permita a integração de dados de diversas proveniências, os quais serão analisados com técnicas de análise espacial, como o DME, facilitando a exploração de dados estatísticos, e diminuindo o tempo e o custo associados a este processo.

Os parceiros que integram o projecto KNOMAD são a University of Ulster (Irlanda do Norte), a Regional Research Laboratories Network (Inglaterra), o GMD (Alemanha), a Universidade do Minho (Portugal), a University of Palermo (Itália), a City of Helsinki Urban Facts (Finlândia) e a MINEIT Software Limited (Irlanda do Norte). O projecto foi submetido em Maio de 2000 ao programa IST da Comunidade Económica Europeia, estando ainda em fase de apreciação.

Em Setembro de 2000 é estabelecido um protocolo de cooperação entre o Centro Nacional de Informação Geográ…ca (CNIG) e a Universidade do Minho, ao abrigo do qual foi celebrado um contrato de prestação de serviços ao exterior. Neste âmbito, o CNIG solicita à Universidade do Minho o desenvolvimento de soluções informáticas no domínio dos sistemas de meta-informação distribuídos, que permitam a descentralização do repositório de dados do Sistema Nacional de Informação Geográ…ca (SNIG). A estrutura do repositório de dados do SNIG, cobrindo actualmente as directivas do CEN TC 287 no que diz respeito aos metadados para catalogação de informação geográ…ca, deverá ser actualizada por forma a contemplar as directivas das prénormas para os metadados, do ISO TC 211 [Gouveia et al., 2001]. Este trabalho inclui ainda a construção de documentos XML (eXtensible Markup Language[15]) para transferência de dados entre instituições e/ou serviços, e uma interface para a World Wide Web (WWW) que permita a gestão dos metadados através da manipulação de documentos XML.

1.4 Organização da tese

A estrutura deste documento re‡ecte a sequência de trabalhos realizados ao abrigo dos objectivos propostos. O seu conteúdo inclui três capítulos de revisão teórica/bibliográ…ca (Capítulo 2,

[15] A linguagem XML é o novo standard proposto pelo W3C (World Wide Web Consortium) que visa a representação e permuta de dados na WWW. Constitui uma linguagem de marcação de documentos que permite o armazenamento e estruturação de dados, e ainda, a interligação de BD heterogéneas [Sousa et al., 2000].

Capítulo 3 e Capítulo 4), que permitiram a sistematização dos conceitos necessários à concepção, implementação e validação do Padrão, o sistema proposto nesta tese. Além da concepção e implementação do sistema, tarefas descritas no Capítulo 5, a validação apresentada no Capítulo 6 permitiu a utilização do sistema na análise de uma BD organizacional de grande dimensão, cujos resultados são apresentados no Capítulo 7.

Este primeiro capítulo é iniciado com uma breve síntese de todo este projecto, destacando as principais características do sistema proposto nesta tese. Na primeira secção é apresentada a principal …nalidade deste trabalho, apresentando as motivações que conduziram à sua formulação, e ainda a sequência de objectivos necessários à sua realização. Para cada um dos objectivos são referidos os resultados e contribuições esperadas. Este capítulo prossegue com a apresentação da metodologia de investigação adoptada para a execução dos trabalhos. A terceira secção resume alguns dos projectos em que o DSI participa, e que estão relacionados com as duas áreas temáticas integradas neste projecto, a DCBD e a informação geográ…ca. Este capítulo culmina com a descrição da organização da tese.

No segundo capítulo é elaborado um enquadramento teórico aos conceitos associados ao domínio geo-espacial, frequentemente utilizados ao longo deste trabalho. Este enquadramento é iniciado com uma breve caracterização do domínio geográ…co, apresentando o signi…cado da designação geo-espacial adoptada ao longo deste documento. A segunda secção apresenta a tecnologia de BDE, referindo os tipos e representações adoptadas para os dados espaciais. Destacam-se, ainda, as características que as linguagens de manipulação de dados espaciais devem possuir, mecanismos de integração de dados espaciais e dados não espaciais, e o papel dos SIG e da análise espacial na exploração deste tipo de dados. Este capítulo prossegue, terceira secção, com a apresentação de técnicas de modelação para informação geográ…ca. A última secção deste capítulo descreve as iniciativas em curso no âmbito da normalização da informação geográ…ca, dando particular ênfase aos grupos de trabalho do ISO e do CEN. Para este último, analisaram-se em detalhe o esquema espacial e o esquema de identi…cadores geográ…cos propostos nas pré-normas, e que permitiram de…nir o conteúdo da BDG utilizada neste trabalho.

O terceiro capítulo descreve os princípios nos quais se baseia o raciocínio espacial qualitativo, destacando os tipos de relações espaciais existentes, as diferentes abordagens ao raciocínio e ainda, mecanismos de representação de conhecimento espacial qualitativo. A terceira secção apresenta o raciocínio temporal qualitativo, cujos princípios foram adaptados ao domínio espacial, permitindo raciocinar com informação geográ…ca incompleta ou imprecisa. Este capítulo prossegue com uma descrição detalhada dos três tipos de relações espaciais, direcção, distância e topologia, adoptados neste trabalho. Para cada um deles, apresentam-se as tabelas de composição que permitem a inferência de informação espacial desconhecida, assim como os princípios utilizados na construção das regras explícitas nestas tabelas. A quinta secção apresenta dois sistemas de raciocínio construídos segundo uma abordagem integrada, os quais permitiram a construção do sistema de inferências utilizado neste trabalho. Este sistema integra relações espaciais do tipo direcção, distância e topologia, segundo os princípios do raciocínio espacial qualitativo. O capítulo culmina com uma breve referência à importância do tamanho dos objectos geográ…cos no processo de raciocínio, descrevendo ainda uma forma de representação desta característica e de mecanismos de raciocínio qualitativos para a mesma.

No quarto capítulo é efectuado um enquadramento conceptual da área da DCBD, destacando os seus princípios, as fases do processo, a importância do conhecimento do domínio de aplicação neste processo, e ainda, as principais di…culdades encontradas na execução do mesmo.

A segunda secção introduz o conceito de DM, destacando as tarefas que podem ser executadas através desta tecnologia, e ainda, as técnicas utilizadas para a satisfação das mesmas. A última secção é dedicada à exploração de BD geo-espaciais, destacando diversas iniciativas em curso nesta área. Esta secção culmina com uma breve síntese das várias abordagens descritas, salientando as principais características associadas às mesmas.

O quinto capítulo começa por apresentar o enquadramento estrutural que conduziu à de…nição da arquitectura do sistema Padrão, a qual integra três componentes: o Repositório de Dados e Conhecimento, a Análise de Dados e a Visualização de Resultados. Cada um destes componentes é documentado recorrendo a diagramas de caso de uso e a diagramas de classes. Os primeiros são utilizados para especi…car o modo de funcionamento do sistema e a sua interacção com o exterior. Os diagramas de classes são utilizados para de…nir a estrutura lógica dos diversos repositórios de dados utilizados pelo Padrão, nomeadamente nos componentes de Repositório de Dados e Conhecimento e Visualização de Resultados. Este capítulo prossegue com a implementação do Padrão, referindo as opções tecnológicas adoptadas para a sua concretização.

O sexto capítulo descreve as diversas tarefas levadas a cabo para validar o sistema proposto. Como já referido anteriormente, esta validação pretende veri…car a qualidade das inferências obtidas através dos mecanismos qualitativos de raciocínio espacial e averiguar a capacidade do sistema de identi…cação de relacionamentos implícitos nos dados. No primeiro caso, avaliou-se o desempenho do sistema de inferências, veri…cando os desvios que decorrem da sua utilização. No que diz respeito ao processo de descoberta de conhecimento, a análise de um conjunto de dados manipulado para o efeito, permitiu constatar que as regras inseridas propositadamente na BD explorada, são efectivamente encontradas.

No sétimo capítulo apresenta-se o estudo de caso que complementa a validação do sistema Padrão apresentada no capítulo anterior. A exploração de uma componente do Sistema de Administração do Pessoal do Exército permitiu detectar relacionamentos implícitos nos dados, con…rmando a utilidade do sistema na análise de BD organizacionais de grande dimensão. Neste capítulo, a análise da BD é precedida da descrição dos dados a analisar e da qualidade dos mesmos. A de…nição de diversas tarefas para o exercício de DM e das hierarquias conceptuais a utilizar, permitiu a execução das diversas fases do processo de descoberta de conhecimento consideradas pelo Padrão. Através deste processo foi possível identi…car um conjunto de relacionamentos que caracterizam os dados analisados.

No oitavo e último capítulo deste documento elabora-se uma síntese de todo o projecto, dando particular ênfase aos resultados obtidos com a realização do mesmo e à satisfação dos objectivos inicialmente impostos ao trabalho. Apresentam-se, ainda, algumas propostas de trabalho futuro, que visam essencialmente a evolução do sistema Padrão. Este capítulo culmina com algumas considerações …nais acerca do trabalho realizado.

Além dos oito capítulos acima referidos, este documento agrega um conjunto de quatro apêndices, cujo conteúdo complementa alguns dos capítulos, integrando informações e descrições associadas aos mesmos.

Capítulo 2

O domínio geo-espacial

O espaço é normalmente utilizado para de…nir relações entre objectos [Gatrell, 1991]. Estes objectos são denominados geográ…cos se implícita ou explicitamente dizem respeito a uma posição relativa à superfície da Terra.

Este capítulo visa contextualizar os conceitos associados à informação geográ…ca, salientando os tipos de dados espaciais existentes, os diferentes tipos de representação da informação disponíveis no domínio espacial, as linguagens de manipulação dos dados, e ainda os mecanismos de integração de dados espaciais e dados não espaciais. Descrevem-se técnicas de modelação para informação geográ…ca, e ainda, os SIG, dando particular ênfase à sua capacidade de análise espacial.

Este capítulo culmina com uma breve descrição das principais iniciativas em curso na área da normalização da informação geográ…ca, destacando os principais grupos de trabalho, e a importância das especi…cações produzidas pelos mesmos para este projecto.

2.1 Caracterização

Os mapas constituem a forma mais familiar de representação de dados geográ…cos. Integram um conjunto de pontos, linhas e polígonos, posicionados recorrendo a determinado sistema de coordenadas[1] [Painho, 1997]. São normalmente representados bi-dimensionalmente. As suas legendas permitem a integração de dados não espaciais, como nomes de locais, símbolos, cores, etc., com dados espaciais, isto é, com a localização dos elementos representados no mapa [Arono¤, 1989]. A informação associada a um elemento geográ…co apresenta quatro componentes fundamentais: a sua posição, os seus atributos, os seus relacionamentos espaciais e o tempo. Basicamente, um elemento é caracterizado pelas questões: Onde está? O que é? Que relações possui com outros elementos? Em que período temporal existiu?

Ao nível das iniciativas internacionais de normalização, descritas na secção 2.4, existe um entendimento comum [CEN/TC-287, 1998i] [ISO/TC-211, 1999d] no que diz respeito à de…nição de dados geográ…cos e informação geográ…ca, sendo em ambos os casos de…nidos como:

[1] Os sistemas de coordenadas permitem descrever matematicamente a posição de um ponto, em relação a outros pontos distribuídos no espaço [CEN/TC-287, 1998g].

Dados geográficos - dados respeitantes a fenómenos associados implícita ou explicitamente a uma localização relativa à Terra, sendo os mesmos representações, tratáveis automaticamente, de informação geográfica.

Informação geográfica - informação relacionada a fenómenos associados directa ou indirectamente (isto é, implícita ou explicitamente) a uma localização relativa à Terra.

Um conjunto de dados geográficos inclui entre os dados que o constituem, pelo menos um aspecto espacial [ISO/TC-211, 1999d], permitindo a definição das características geométricas e topológicas associadas ao mesmo [CEN/TC-287, 1998i].

As primitivas topológicas permitem a descrição, total ou parcial, dos aspectos topológicos de um objecto. Os aspectos topológicos permitem descrever a conectividade existente entre as entidades geográficas, propriedades estas que permanecem invariantes à transformações do espaço, como sejam mudanças de escala ou de sistema de referenciação quantitativo. As primitivas geométricas permitem a descrição, total ou parcial, de um objecto recorrendo a coordenadas e funções matemáticas. Esta descrição quantitativa das entidades inclui a definição de características como dimensão, posição, forma e orientação, as quais variam dependendo do sistema de coordenadas utilizado na referenciação geográfica da informação [CEN/TC-287, 1998i].

O termo geográfico está assim associado à localização, contextualizando determinado objecto ou acontecimento no espaço, enquanto que o termo espacial é utilizado para referir as características dessa localização, como por exemplo a sua geometria e topologia. Informação geográfica tem então associada informação espacial, pelo que ao longo deste trabalho o termo geo-espacial é utilizado para fazer referência explícita a estes dois conjuntos de informação.

Dados espaciais são objectos espaciais representados por pontos, linhas, áreas ou volumes, podendo ainda apresentar outras dimensões como o tempo [Samet, 1995]. Dados espaciais e dados não espaciais são frequentemente integrados, por forma a complementar a informação espacial com as características não espaciais associadas à mesma. BDE constituem repositórios de dados espaciais e não espaciais, os quais podem ser manipulados utilizando operadores e funções específicas.

2.2 Tecnologia de bases de dados espaciais

A informação é um dos recursos mais importantes de uma organização, contribuindo decisivamente para a sua competitividade. Para que os dados armazenados possam ser transformados em informação, estes têm de ser relacionados ou interpretados de alguma forma. De entre as Tecnologias da Informação[2] (TI), a tecnologia de BD fornece meios e ferramentas para extracção de informação relevante a partir dos dados armazenados [Pereira, 1997]. Num sistema de BD, os dados encontram-se armazenados num conjunto de ficheiros, organizados de forma transparente aos utilizadores, sendo o acesso aos mesmos efectuado através de um SGBD[3], que centraliza em

[2] O termo Tecnologias da Informação integra o conjunto de equipamentos (hardware) e suportes lógicos (software) que permitem executar tarefas como aquisição, transmissão, armazenamento, recuperação e visualização de dados ([Alter, 1992] p. 9).

[3] Um SGBD é constituído por um conjunto de aplicações destinadas a gerir todo o armazenamento e manipulação dos dados do sistema, fazendo o interface entre o nível aplicacional e a BD. Permite a partilha concorrente dos dados, incorporando mecanismos que asseguram a validade, a segurança e a recuperação dos mesmos, em caso

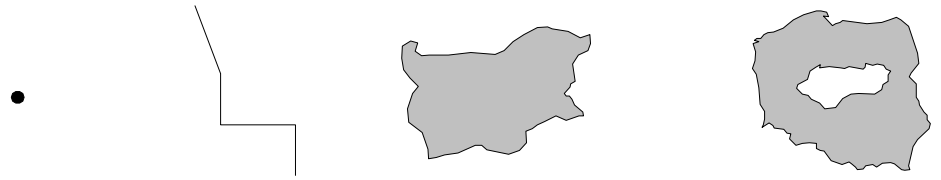
si o acesso à BD.

Um SGBDE é sistema de BD que disponibiliza, no seu modelo de dados, tipos de dados espaciais e uma linguagem que permita a manipulação dos dados armazenados na mesma. Entre os tipos de dados básicos para o espaço bi-dimensional, encontra-se o ponto, a linha e o polígono. Entre as operações mais utilizadas destaca-se o cálculo de distâncias ou intersecções entre regiões.

Os SGBDE disponibilizam a tecnologia de BD necessária a implementação de SIG (descritos na subsecção 2.2.5), fornecendo tipos de dados, linguagens de pesquisa e mecanismos de integração de dados espaciais e dados não espaciais [Samet e Aref, 1995], os quais são descritos nas próximas subsecções.

2.2.1 Tipos de dados espaciais

A modelação de objectos espaciais passa pela sua abstracção recorrendo a pontos, linhas ou polígonos (Figura 2.1). Um ponto representa o aspecto geométrico de um objecto para o qual apenas a sua localização no espaço é relevante. Uma linha (constituída por um conjunto de segmentos de recta) permite a representação de objectos através dos quais é possível a movimentação no espaço. Um polígono, com ou sem buracos, é uma abstracção de um objecto cuja localização e extensão são relevantes.



a) Ponto b) Linha c) Polígono d) Polígono com buraco

Figura 2.1: Abstracções básicas utilizadas na representação de dados espaciais

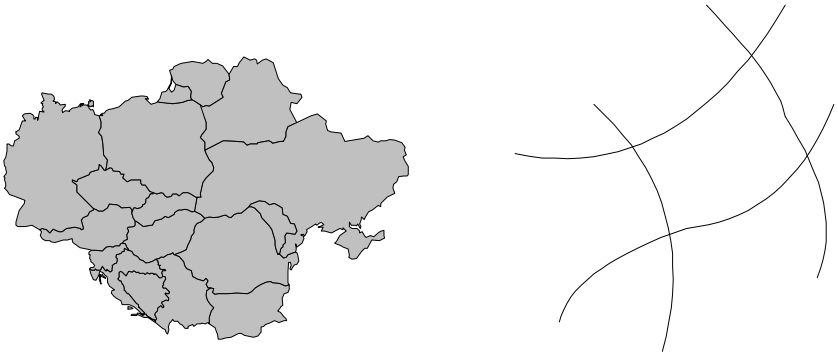
As duas principais colecções de objectos referenciados espacialmente são as partições e as redes (Figura 2.2). Uma partição[4] do plano representa um conjunto de regiões não sobrepostas, que partilham limites (linhas) e para as quais a adjacência constitui uma relação espacial de particular interesse. Uma rede[5] representa um grafo embebido no plano, constituído por um conjunto de pontos (que formam os nodos da rede) e por um conjunto de linhas (que representam as arestas da rede) [Güting, 1994].

Tipos de dados espaciais permitem representar entidades geométricas, pontos, linhas e polígonos, assim como os relacionamentos existentes entre as mesmas (*l intersecta r*). Estas entidades podem ser manipuladas veri...cando determinadas propriedades (*área(r)>1000*), ou ainda executando operações especí...cas (*intersecção(l,r)*) sobre as mesmas [Güting, 1994].

de falhas ou acidentes [Pereira, 1997].

[4] As partições são normalmente utilizadas para representar mapas temáticos, divisões administrativas, etc.

[5] As redes têm presença obrigatória em aplicações geográ...cas, na representação de estradas, redes de transportes públicos, rios, etc.



a) Partição b) Rede

Figura 2.2: Colecção de objectos espaciais: partições e redes

De entre o vasto conjunto de operações que podem ser disponibilizadas para manusear dados espaciais, as mais importantes são as que manipulam relações espaciais. Estas têm sido classi…cadas em três grandes grupos: topológicas (tais como adjacência, intersecção, ...), de direcção (tais como norte, sul, ...) e de distância (próximo, distante, ...). Estas relações são abordadas em detalhe no Capítulo 3.

A integração de tipos de dados espaciais, no modelo de dados de um SGBD, permite a representação de objectos espaciais por entidades que possuem pelo menos um atributo do tipo espacial. Assim, são ampliados os tipos de dados normalmente disponíveis (INTEGER, REAL, STRING, ...) num SGBD, disponibilizando tipos de dados espaciais (POINT, LINE, REGION).

2.2.2 Representação de dados espaciais

Os modelos de dados espaciais constituem abstracções dos dados, que escondem os detalhes associados ao armazenamento dos mesmos, segundo determinada representação [Shekhar et al., 1999].

A representação de dados espaciais pode ser efectuada através de dois métodos: representação baseada em células ou baseada em objectos. Na representação baseada em células, o espaço geográ…co é dividido num conjunto de células independentes que cobrem toda a região geográ…ca em análise. Os objectos geográ…cos encontram-se embebidos no espaço, sendo representados através do conteúdo das células. Esta representação matricial da informação é baseada em áreas, uma vez que os objectos são representados pela área que os constitui, e não pelo seu limite [Adam e Gangopadhyay, 1997]. Os modelos matriciais podem ser classi…cados em dois grupos: os de resolução espacial …xa e os de resolução espacial variável. Os modelos de resolução …xa, utilizam uma estrutura de dados denominada raster, que é caracterizada por ser constituída por um conjunto de células de tamanho …xo, as quais são utilizadas para descrever as entidades geográ…cas. Cada célula tem associados dois valores: a posição, que representa a sua localização, e o valor da área geográ…ca que representa. Este valor é uniforme ao longo de todas as células que constituem uma mesma entidade geográ…ca.

Na resolução espacial variável, a matriz é constituída por um conjunto de células de tamanho variável, obtidas por divisão do espaço geográ…co, com vista a optimização da representação de determinado objecto geográ…co. Inicialmente, o espaço geográ…co é decomposto em quatro quadrantes, estrutura quadtree, permitindo que cada um dos mesmos possa ser posteriormente subdividido, de forma a aumentar a resolução das células que o constituem. Uma quadtree[6] é uma estrutura de dados hierárquica, na qual cada nível pode ser sucessivamente decomposto em quatro quadrantes. O modelo é apelidado de resolução variável uma vez que determinado quadrante, num dado nível de informação, apenas é decomposto se existir alguma variação nos valores das células que o constituem [Adam e Gangopadhyay, 1997].

A Figura 2.3 apresenta um exemplo de representação de um objecto geográ…co recorrendo às duas estruturas de representação apresentadas. Para o modelo de resolução espacial variável, é ainda apresentada a quadtree correspondente. Pela análise da referida …gura veri…ca-se que a representação recorrendo a estrutura raster pode ser optimizada através de uma representação por blocos, Figura 2.3 c), que agrega células com valor idêntico. Esta optimização permite representar o mesmo objecto geográ…co recorrendo apenas a 46 células, ao invés das 256 utilizadas pela estrutura raster (Figura 2.3 b)).

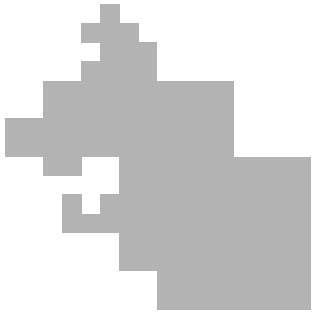
Na representação baseada em objectos, modelo vectorial, os objectos são representados pelas entidades geográ…cas que os constituem. Pontos, linhas e polígonos são as três unidades geométricas fundamentais utilizadas na representação de objectos geográ…cos. Estas unidades podem ser combinadas por forma a representarem objectos compostos (agregação de várias unidades geométricas do mesmo tipo) ou objectos complexos (constituídos pela agregação de unidades geométricas de diferentes tipos) [Adam e Gangopadhyay, 1997].

Um ponto é descrito por um único par de coordenadas, uma linha por uma sequência de pares de coordenadas que de…nem segmentos de recta, e um polígono, representado por uma área fechada, é descrito por uma sequência de pares de coordenadas, na qual o primeiro e último par de coordenadas, coincidem [Burrough, 1986]. Este modelo constitui uma estrutura de armazenamento da informação e…ciente, uma vez que apenas são representadas as coordenadas que de…nem um acontecimento. A Figura 2.4 apresenta a representação vectorial do objecto geográ…co descrito na Figura 2.3 através duma estrutura baseada em células.

A representação dos dados espaciais seguindo o modelo vectorial, pode utilizar duas estruturas de dados distintas: o modelo de dados esparguete ou o modelo de dados topológico. O modelo em esparguete permite que os elementos geográ…cos sejam representados como uma lista de coordenadas XY. Um ponto é codi…cado recorrendo a um único par de coordenadas; uma linha a um conjunto de pares de coordenadas; e uma área por um polígono que de…ne os seus limites através de um conjunto de pares de coordenadas, no qual o primeiro e último par coincidem. Um limite comum entre dois polígonos é armazenado duas vezes, um para cada polígono. O armazenamento físico desta estrutura de dados (Figura 2.5) consiste num …cheiro sem uma estrutura especí…ca, no qual os elementos geográ…cos são representados por uma colecção de caracteres [Arono¤, 1989].

Este modelo apresenta-se bastante ine…ciente para a maior parte das tarefas de análise

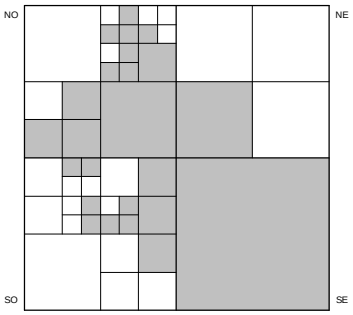
[6] A estrutura quadtree forma uma árvore cujos nodos representam áreas heterogéneas, enquanto que as folhas representam áreas homogéneas dentro de um mesmo quadrante. Reduzem signi…cativamente o espaço necessário para armazenar uma imagem raster, uma vez que agregam células vizinhas com o mesmo valor [Egenhofer e Herring, 1991].



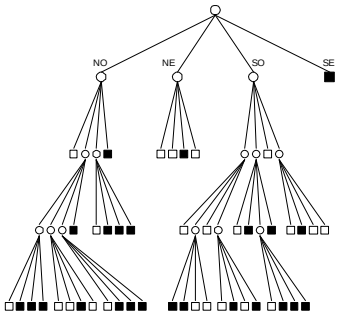
a) Objecto geográfico

0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0
0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1

b) Representação *raster*

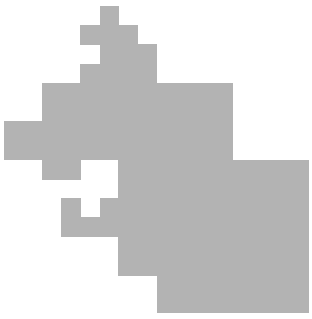


c) Representação *variável* (por blocos)

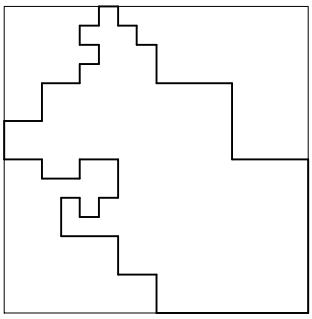


d) Representação *quadtree* dos blocos

Figura 2.3: Representação por células: as estruturas de dados ...xa, variável e quadtree (Adaptado de: [Gatrell, 1991] p. 125)

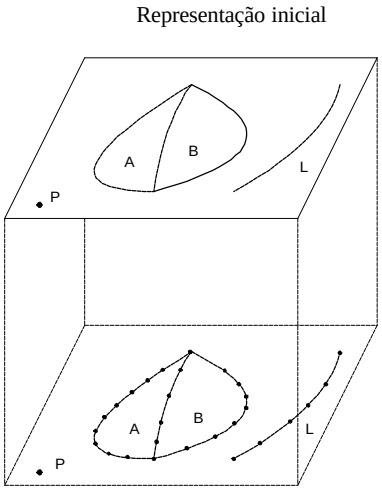


a) Objecto geográfico



b) Representação *vectorial*

Figura 2.4: Representação vectorial da informação



Estrutura de dados

Tipo	ID	Localização
Ponto	P	(x, y)
Linha	L	$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
Polígono	A	$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), (x_1, y_1)$
	B	$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), (x_1, y_1)$

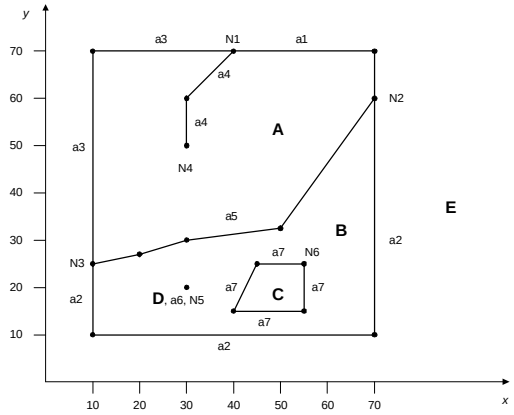
Figura 2.5: Modelo de dados esparguete (Adaptado de: [Arono¤, 1989] p. 174)

espacial, já que têm de ser derivados todos os relacionamentos espaciais existentes entre as entidades geográ…cas analisadas. É frequentemente utilizado em aplicações de reprodução de mapas, nas quais não é necessária informação adicional, para além das coordenadas dos elementos geográ…cos representados.

O modelo de dados topológico é o mais utilizado, uma vez que permite que os relacionamentos espaciais existentes entre entidades geográ…cas sejam expressos através da topologia. O elemento básico utilizado na representação é a aresta, a qual integra um conjunto de pontos, que iniciam e terminam num nodo. Um nodo representa um ponto de intersecção entre duas ou mais arestas. Uma face é representada recorrendo a uma cadeia de arestas, que delimitam a área representada. A topologia é neste modelo armazenada recorrendo a quatro tabelas (Figura 2.6), uma para cada tipo de elemento espacial (nodo, aresta ou face), e a quarta para as coordenadas das arestas [Arono¤, 1989].

Na tabela correspondente às faces, armazenam-se as arestas que as delimitam, seguindo o sentido dos ponteiros do relógio. Sempre que uma face contiver outros elementos, a identi…cação dos mesmos é precedida de um 0. Na tabela de nodos identi…cam-se as arestas a que cada nodo está associado, enquanto que na tabela de arestas estabelece-se a relação existente entre estas e os nodos e as faces. Estas três tabelas permitem a identi…cação da posição relativa dos elementos representados. É possível identi…car todas as faces adjacentes a uma determinada face, pesquisando a tabela das arestas e veri…cando as arestas comuns.

A principal vantagem deste modelo é que análises espaciais podem ser efectuadas sem recorrer às coordenadas que identi…cam os vários elementos geográ…cos. Esta aproximação aumenta a e…ciência do sistema, evitando a derivação dos vários relacionamentos espaciais a partir das coordenadas.



Faces	
Face	Arestas
A	a1, a5, a3
B	a2, a5, 0, a6, 0, a7
C	a7
D	a6
E	área exterior do mapa

Nodos	
Nodo	Arestas
N1	a1, a3, a4
N2	a1, a2, a5
N3	a2, a3, a5
N4	a4
N5	a6
N6	a7

Arestas				
Aresta	Nodo inicial	Nodo final	Face esquerda	Face direita
a1	N1	N2	E	A
a2	N2	N3	E	B
a3	N3	N1	E	A
a4	N4	N1	A	A
a5	N3	N2	A	B
a6	N5	N5	B	B
a7	N6	N6	B	C

Coordenadas das Arestas			
Aresta	(x, y) inicial	(x, y) intermédios	(x, y) final
a1	(40,70)	(70,70)	(70,60)
a2	(70,60)	(70,10), (10,10)	(10,25)
a3	(10,25)	(10,70)	(40,70)
a4	(40,70)	(30,60)	(30,50)
a5	(10,25)	(20,27), (30,30), (50,32)	(70,60)
a6	(30,20)		(30,20)
a7	(55,25)	(55,15), (40,15), (45,25)	(55,25)

Figura 2.6: Modelo de dados topológico (Adaptado de: [Arono¤, 1989] p. 175)

2.2.3 Linguagem de manipulação de dados espaciais

Ao nível das linguagens para a manipulação de dados, destaca-se o esforço de desenvolvimento de extensões do SQL (Structured Query Language) [Egenhofer, 1994a] [Samet e Aref, 1995] [Ravada e Sharma, 1999] [OGC, 1999b], que permitam a manipulação de dados espaciais. Estas extensões têm como objectivo acrescentar à tradicional sintaxe `select ... from ... where`, operadores e funções espaciais capazes de manusear este tipo de dados.

Güting [Güting, 1994] considera que as operações necessárias à manipulação de objectos com atributos espaciais, podem ser agrupadas em três grandes grupos: selecção espacial, junção espacial e funções espaciais.

A selecção espacial permite a escolha de um conjunto de objectos que veri…cam determinada condição, construída recorrendo a pelo menos um predicado espacial. A questão "seleccionar todos os municípios de Braga", assume que Braga existe na BD como uma região e que o predicado `inside`, implícito na questão, está disponível no conjunto de operadores espaciais disponibilizados pela linguagem, para a manipulação dos dados.

Utilizando a sintaxe proposta por Samet e Aref [Samet e Aref, 1995], e seguindo o exemplo anterior, a de…nição das tabelas necessárias ao armazenamento da informação, pode ser efectuada através de:

```
CREATE TABLE concelhos
(id_concelho INTEGER,
 nome_conc CHAR(30),
 população INTEGER,
 concelho REGION);

CREATE TABLE distritos
(id_distrito INTEGER,
 nome_dist CHAR(30),
 distrito REGION);
```

A selecção propriamente dita, manipulação dos dados espaciais e não espaciais, é conseguida com a seguinte instrução[7]:

```
SELECT nome_conc
FROM concelhos, distritos
WHERE nome_dist=''Braga'' AND inside(concelho, distrito)
```

A junção espacial permite a integração de duas tabelas, cujos atributos espaciais são utilizados na construção de uma condição, recorrendo a operadores ou funções espaciais, que guia o processo de integração. A questão "Quais os concelhos adjacentes aqueles que possuem uma densidade populacional superior a 100 000 habitantes" pode ser especi…cada através de:

```
SELECT all
FROM l concelhos, k concelhos
WHERE adjacent_to(l.concelho, k.concelho) AND l.população>100000
```

O processo de selecção dos dados armazenados numa BD é habitualmente conduzido por determinada condição. Esta pode incluir a avaliação dos resultados obtidos pela execução de uma função espacial. Entre as diversas funções espaciais disponíveis [Samet e Aref, 1995], encontram-se:

```
area (região) > valor
centroid (região)
lenght (linha) > valor
...
```

[7] Refere-se que o Spatial SQL [Egenhofer, 1994a] utiliza uma sintaxe muito semelhante à proposta por Samet e Aref [Samet e Aref, 1995], mas com diferentes níveis de abstracção na de…nição dos dados espaciais. Apresenta o conceito de domínio espacial, disponibilizando o spatial_0, spatial_1, spatial_2 e spatial_3 como os tipos de dados disponíveis para a de…nição de objectos espaciais com zero, uma, duas e três dimensões, respectivamente. Egenhofer [Egenhofer, 1994a] apresenta ainda a GPL, Graphical Presentation Language, uma linguagem de representação de resultados, que permite a manipulação grá…ca dos resultados obtidos com a satisfação de determinada questão.

Operador	Funcionalidade
Equals(anotherGeometry:Geometry):Integer	Verifica se os dois objectos são espacialmente iguais.
Disjoint(anotherGeometry:Geometry):Integer	Verifica se os dois objectos não se tocam.
Intersects(anotherGeometry:Geometry):Integer	Verifica se os dois objectos se intersectam.
Touches(anotherGeometry:Geometry):Integer	Verifica se os dois objectos são adjacentes.
Crosses(anotherGeometry:Geometry):Integer	Verifica se os dois objectos se cruzam.
Within(anotherGeometry:Geometry):Integer	Verifica se um dos objectos está dentro do outro objecto.
Contains(anotherGeometry:Geometry):Integer	Verifica se um dos objectos contém o outro objecto.
Overlaps(anotherGeometry:Geometry):Integer	Verifica se os dois objectos se sobrepõem.

Tabela 2.1: Operadores Espaciais

No que diz respeito à componente topológica dos objectos geográficos, a Tabela 2.1 sintetiza o conjunto de operadores que, segundo o Open GIS Consortium (OpenGIS) [OGC, 1999b], deve integrar uma linguagem de manipulação de dados espaciais. Estes operadores permitem verificar as relações espaciais existentes entre a geometria de dois objectos.

2.2.4 Integração de dados espaciais e dados não espaciais

A construção de um SGBD para dados espaciais inclui a definição do processo de integração dos dados espaciais e não espaciais, que serão geridos pelo mesmo. A arquitectura dual proposta por Aref e Samet [Samet e Aref, 1995], denominada de SAND (Spatial And Non-spatial Data) permite que dados espaciais e não espaciais estejam associados bi-direccionalmente.

Nesta arquitectura, cada registo duma relação não espacial está associado a um objecto espacial. Entre os dados espaciais e não espaciais relacionados são mantidas duas ligações lógicas: ligação para a frente e ligação para trás (Figura 2.7). As instâncias relacionadas, juntamente com as ligações, formam um relacionamento espacial. Os apontadores para a frente são utilizados para seleccionar a informação espacial associada a um objecto, dada a sua correspondente informação não espacial. Da mesma forma, os apontadores para trás são utilizados para seleccionar a informação não espacial de um objecto, dada a sua correspondente informação espacial.

A manutenção destes dois grupos de apontadores entre os dados espaciais e não espaciais, possibilita a navegação entre as duas partes, facilitando ainda o processamento de questões.

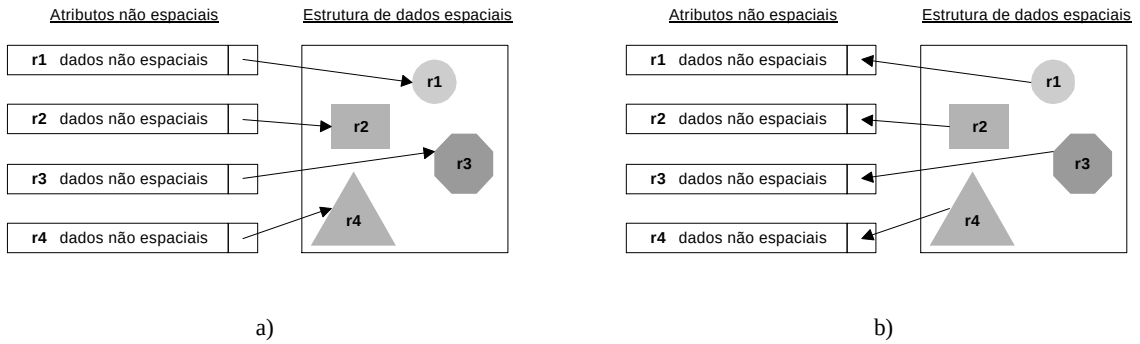


Figura 2.7: Arquitectura SAND: a) apontadores para a frente b) apontadores para trás (Adaptado de: [Samet e Aref, 1995])

2.2.5 Os sistemas de informação geográ…ca e a análise espacial

Os SIG podem ser de…nidos como sistemas de apoio à decisão, que permitem a integração de dados espaciais e dados não espaciais. Sobre os mesmos é possível a execução de operações de pesquisa, análise e visualização [Maguire, 1991]. A tecnologia utilizada pelos SIG integra operações comuns às BD, possuindo ainda so…sticadas técnicas de visualização e análise espacial, que auxiliam as organizações na explicação de acontecimentos, previsão de resultados e planeamento de estratégias [Painho, 1996].

Um SIG integra diversos componentes, os quais lhe proporcionam as funcionalidades necessárias à execução da sua principal …nalidade, a análise espacial. Entre eles, destacam-se as aplicações, o equipamento, os dados e os utilizadores.

As aplicações fornecem as ferramentas e funções necessárias ao armazenamento, pesquisa, análise e visualização de informação geográ…ca. Integram mecanismos de aquisição de dados espaciais e não espaciais, assim como incluem um SGBD para armazenamento e gestão dos dados. Disponibilizam, ainda, técnicas de visualização da informação, mecanismos de pesquisa e análise espacial, e uma interface grá…ca com o utilizador.

O equipamento onde o SIG opera depende das características da aplicação utilizada, e pode incluir máquinas com arquitecturas cliente-servidor, até computadores pessoais utilizados isoladamente ou integrados em determinada rede. Ao nível dos periféricos de entrada/saída, destaca-se a necessidade de periféricos especí…cos, como scanners para aquisição de imagens, ou plotters para impressão de resultados.

Os dados e os utilizadores constituem os componentes de maior importância num SIG. Os primeiros podem ser recolhidos ou adquiridos a entidades competentes, sendo posteriormente organizados por forma a poderem ser armazenados e geridos pelo SGBD. Os utilizadores constituem a chave do sistema, desenvolvendo estratégias de aplicação. Devem ainda, desenhar e manter as funcionalidades do sistema, que permitam o suporte à tomada de decisão.

Apesar dos SIG terem sido inicialmente concebidos como ferramentas de armazenamento, pesquisa e visualização de dados geográ…cos [Fotheringham e Rogerson, 1994], a integração de mecanismos de análise espacial para estes sistemas, rapidamente se tornou uma prioridade no

Função	Funcionalidade
Distance(anotherGeometry:Geometry):Double	Calcula, dentro de um sistema de referenciação espacial, a distância mais curta entre dois objectos.
Buffer(distance:Double):Geometry	Determina o(s) objecto(s) cuja geometria está completamente inserida na distância especi…cada.
ConvexHull():Geometry	Determina a forma geométrica convexa de determinado objecto.
Intersection(anotherGeometry:Geometry):Geometry	Determina a forma geométrica resultante da intersecção de dois objectos.
Union(anotherGeometry:Geometry):Geometry	Determina a forma geométrica resultante da união de dois objectos.
Difference(anotherGeometry:Geometry):Geometry	Determina a forma geométrica resultante da diferença de dois objectos.
SymDifference(anotherGeometry:Geometry):Geometry	Determina a forma geométrica resultante da diferença simétrica entre dois objectos.

Tabela 2.2: Funções requeridas na análise espacial

desenvolvimento destas aplicações[8].

Genericamente, a análise espacial é de…nida por Bailey ([Bailey, 1994] p. 15) como a "capacidade de manipular dados espaciais de diferentes formas e extrair conhecimento como resultado". Esta capacidade integra um conjunto de técnicas para a manipulação de entidades geográ…cas, sendo o resultado da análise condicionado pela disposição espacial das mesmas. As entidades geográ…cas são representadas por pontos, linhas ou polígonos, sendo os seus dados não espaciais armazenados como um conjunto de atributos numa tabela [Haining, 1994].

De entre as funções de análise espacial que deverão estar disponíveis num SIG, o OpenGIS [OGC, 1999b] destaca as evidenciadas na Tabela 2.2.

Num SIG, a execução de tarefas de análise espacial requer a estruturação de uma questão, na qual o utilizador especi…ca os atributos de interesse, a área geográ…ca a analisar e ainda os operadores e funções espaciais necessários à análise dos dados.

A obrigatoriedade de estruturação das questões que conduzem à pesquisa, introduz uma di…culdade adicional na execução de análises espaciais. Em primeiro lugar, é necessário que o utilizador identi…que todos os atributos relevantes e posteriormente, que possua os conhecimen-

[8] Em grande parte motivado pelo notável crescimento da quantidade de informação geo-referenciada armazenada pela maioria das organizações, e ainda do número de utilizadores interessados na análise da mesma [O'Kelly, 1994].

tos necessários à estruturação da questão. Em segundo lugar, o facto de existir uma selecção prévia dos atributos a considerar na análise, condiciona a partida os resultados que podem ser encontrados.

Estas preocupações, referidas por Openshaw [Openshaw, 1991], conduziram o autor à enumeração de um conjunto de caminhos, pelos quais sugere que a análise espacial evolua. Entre eles, destaca-se a utilização de técnicas provenientes da inteligência arti…cial, nomeadamente as redes neuronais[9], na detecção de padrões espaciais nos dados. Este tipo de abordagem permite a detecção de padrões nos dados, não condicionados por qualquer hipótese formulada à partida.

A sugestão acima expressa con…rma a necessidade de utilização de mecanismos automáticos na análise de dados espaciais, revalidando a motivação que conduziu à de…nição da …nalidade do trabalho apresentado neste documento: a integração de mecanismos automáticos de exploração de dados não espaciais, nomeadamente através de técnicas de DM, com dados espaciais, que permitam detectar padrões implícitos existentes nos dados analisados.

2.3 Modelação da informação geográ…ca

Um *modelo de dados* integra um conjunto de conceitos, utilizados para descrever a estrutura de uma BD, assim como o conjunto de operações que podem ser executadas sobre a mesma [Navathe, 1992]. Por estrutura entende-se tipos de dados, relacionamentos e restrições, que de…nem como a BD se encontra organizada. Um modelo de dados é utilizado para de…nir modelos de aplicação, que constituem a descrição duma BD para um dado domínio de aplicação.

A modelação de dados é a actividade que, "debruçando-se sobre a totalidade dos requisitos de informação de um sistema de informação[10], tenta encontrar um modelo que traduza a estrutura lógica dos dados que satisfaz esses requisitos" ([Pereira, 1997] p. 16).

A modelação de dados vista como o conjunto de actividades (Figura 2.8) que conduz ao desenho da BD, passa por três etapas [Adam e Gangopadhyay, 1997] [Navathe, 1992]:

Desenho conceptual. Consiste na construção do modelo conceptual de dados, o qual re‡ecte a percepção que os utilizadores têm dos dados, sendo independente de qualquer implementação física. Uma das abordagens mais utilizadas na modelação conceptual de dados são os diagramas E-R[11] (Entidades e Relacionamentos) [Chen, 1976].

Desenho lógico. Corresponde à transformação do modelo conceptual em estruturas de dados que são implementáveis no SGBD seleccionado. O modelo lógico de dados permite espe-

[9] As características destas redes são descritas no Capítulo 4.

[10] Um sistema de informação é um sistema que reúne, processa e faculta informação relevente para a organização (ou sociedade), de modo a que a informação seja acessível e útil para aqueles que a querem utilizar ([Buckinghan et al., 1987] p. 18).

[11] O modelo E-R é utilizado para descrever a estrutura dos dados especí…ca de uma dada aplicação. Os componentes básicos utilizados na modelação são a entidade, o relacionamento e o atributo. Entidades constituem objectos ou conceitos de interesse no domínio de aplicação em causa. Relacionamentos representam associações entre entidades, nos quais o grau da relação determina o número de instâncias de uma entidade que participam, ou podem participar, na mesma. Os atributos de uma entidade/relacionamento de…nem propriedades, podendo assumir valores dentro do contexto de…nido para o domínio de aplicação modelado.

ci…car detalhes físicos a considerar na implementação. O modelo relacional[12] tem sido o modelo mais utilizado na modelação lógica de dados [Codd, 1970].

Desenho físico. Passa pela de…nição dos detalhes físicos que serão considerados na implementação do modelo lógico. Permite de…nir os métodos de acesso aos dados e os detalhes associados à organização física dos …cheiros, e que são especí…cos do SGBD adoptado.

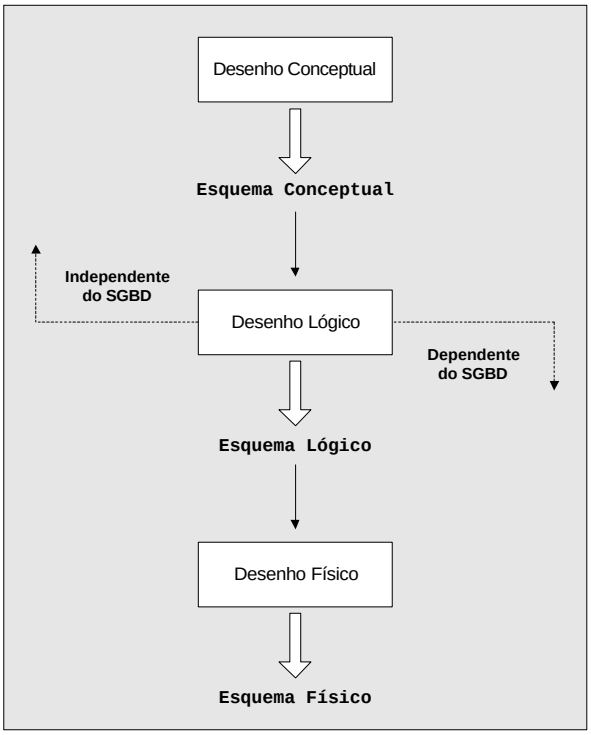


Figura 2.8: O processo de modelação: desenho conceptual, lógico e físico

O desenho de uma BD, para utilização numa determinada aplicação, tem sido caracterizado por um desenvolvimento top-down, que passa por três grandes passos. O primeiro diz respeito ao modelo conceptual de dados, caracterizado por ser um modelo de dados de alto nível, que determina os requisitos da BD na fase de análise de requisitos. O modelo conceptual de dados é independente do SGBD utilizado na implementação. O segundo diz respeito ao modelo lógico de dados e corresponde à transformação do modelo conceptual em estruturas de

[12] A estrutura de dados utilizada no modelo relacional é a relação, podendo ser de…nida como uma tabela constituída por linhas e colunas, na qual as colunas representam os atributos e as linhas os registos ou instâncias da relação. Para cada atributo de…ne-se o domínio de valores que o mesmo pode tomar. O modelo relacional permite a manipulação dos dados através da álgebra e cálculo relacional. Na álgebra relacional, e dado que uma relação é um conjunto, todas as operações da teoria de conjuntos (união, intersecção, diferença e produto cartesiano), podem ser aplicadas à relação. Outras operações especí…cas do modelo relacional (selecção, projecção, junção e divisão) podem ainda ser utilizadas. No cálculo relacional o utilizador pode de…nir o que pretende através de uma forma declarativa, sendo o cálculo baseado em predicados de 1 ^{o}ordem. Para além dos operadores lógicos e condicionais, podem ainda ser utilizados o quanti…cador existencial (9) e o quanti…cador universal (8) [Pereira, 1997].

dados consistentes com o SGBD seleccionado para a implementação. Ao nível mais baixo temos os detalhes da organização dos …cheiros e métodos de acesso aos dados, característicos de um determinado SGBD. Este passo diz respeito ao modelo físico de dados [Adam e Gangopadhyay, 1997].

As limitações dos modelos de dados ditos convencionais, tais como o modelo relacional, começaram a evidenciar-se quando surgiram novas áreas de aplicação, tais como os SIG, com diferentes necessidades ao nível do armazenamento e processamento dos dados. Grande parte destas aplicações são construídas sobre sistemas de …cheiros dedicados [Pereira, 1997], nos quais os dados são armazenados em formatos próprios, cuja manutenção e portabilidade são difíceis de conseguir [Hadzilacos e Tryfona, 1996]. Hadzilacos e Tryfona [Hadzilacos e Tryfona, 1996] defendem que as aplicações geográ…cas devem ser desenvolvidas recorrendo à modelação de dados, para o qual é possível estender as metodologias habituais, por forma a estas suportarem as particularidades da informação espacial. Aplicações geográ…cas manuseiam objectos cuja posição no espaço é relevante. O facto de ocuparem determinada posição faz com que os objectos se encontrem relacionados uns com os outros, pelo que a modelação deverá permitir de…nir os relacionamentos topológicos existentes entre os mesmos.

A proposta de um novo modelo (o modelo geo-relacional, apresentado na subsecção 2.3.2) para a fase de desenho lógico de aplicações geográ…cas, surge como resposta ao descuido a que esta fase tem sido sujeita, em favor do desenho físico. O modelo integra princípios do paradigma OO[13] (Object-oriented) com o modelo relacional.

A modelação de dados geográ…cos, ao nível do desenho lógico, deverá ser realizada recorrendo a ferramentas que permitam [Hadzilacos e Tryfona, 1996]:

- De…nir atributos espaciais, incluindo os seus tipos geométricos (ponto, linha, polígono);
- Organizar atributos espaciais por níveis;
- De…nir atributos não espaciais e seus respectivos relacionamentos com os atributos espaciais;
- Especi…car restrições topológicas ou de integridade dos dados espaciais;
- Etiquetar as versões das diversas entidades, introduzindo a componente temporal na modelação.

Nas próximas subsecções apresentam-se técnicas de modelação para o domínio geográ…co. Ao nível conceptual descreve-se uma extensão do modelo E-R para dados geográ…cos. Ao nível lógico apresenta-se o modelo geo-relacional, e ainda uma abordagem formal à modelação de aplicações geográ…cas. Esta última representa uma abordagem apropriada à modelação da informação geográ…ca se se pretender utilizar técnicas de programação lógica indutiva (Inductive Logic Programing, ILP), na inclusão da componente espacial no processo de descoberta de conhecimento. Nesta secção não é apresentado o UML (Uni…ed Modeling Language), apesar de ser esta a linguagem de modelação adoptada neste trabalho, por este não necessitar de qualquer extensão ou alteração que permita a sua utilização na modelação de informação geográ…ca.

[13] O paradigma OO introduz conceitos como abstracção e encapsulamento no processo de modelação [Adam e Gangopadhyay, 1997].

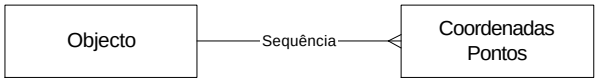
Destaca-se que pela sua ampla utilização, constituindo um standard de facto, é a linguagem utilizada no Capítulo 5 para documentar a arquitectura do sistema Padrão.

2.3.1 Extensão do modelo E-R para dados geográficos

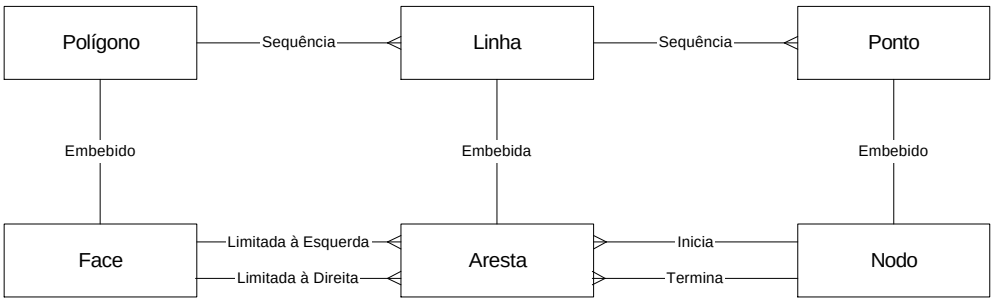
A extensão do modelo E-R para dados geográficos [Laurini e Thompson, 1992] permite representar objectos espaciais como entidades, cujas propriedades estruturais e topológicas com outros objectos espaciais, são representadas através de relacionamentos entre entidades.

No caso de uma representação vectorial[14] da informação, os objectos espaciais são descritos recorrendo a agrupamentos de objectos do tipo ponto, linha ou polígono. Cada objecto espacial, representado como uma entidade, é relacionado com os objectos que o constituem, através de relacionamentos estruturais de diferentes cardinalidades.

A Figura 2.9 apresenta os diagramas E-R para o modelo de dados esparguete e para o modelo de dados topológico, descritos anteriormente na subsecção 2.2.2.



Modelo Esparguete



Modelo Topológico

Figura 2.9: Diagramas E-R na representação de dados espaciais (Adaptado de: [Shekhar et al., 1999] p. 50)

[14] Apenas se descreve a extensão do modelo E-R para este tipo de representação, por ser o modelo de representação de dados espaciais utilizado neste trabalho. Para o caso de uma representação por células, o processo de modelação é semelhante ao aqui apresentado.

2.3.2 Extensão do modelo relacional para dados geográ…cos: O modelo geo-relacional

O modelo geo-relacional estende o modelo relacional por forma a este permitir modelar dados geográ…cos. A incorporação no modelo de componentes necessários ao domínio geográ…co, tais como níveis, relações, níveis virtuais, classes de objectos e restrições de integridade, permitem a construção de modelos lógicos de dados[15], assim como vistas[16] sobre os mesmos [Hadzilacos e Tryfona, 1996].

O modelo geo-relacional (Figura 2.10) permite que a informação seja armazenada em diversas tabelas, relacionadas através de atributos comuns. Estas ligações permitem que a pesquisa de informação possa ser iniciada nos dados espaciais ou nos dados não espaciais. Diferentes conjuntos de atributos são armazenados em diferentes tabelas, através das quais é possível, veri…cando as relações existentes entre as mesmas, seleccionar todos os atributos associados a determinada entidade espacial [Shepherd, 1991].

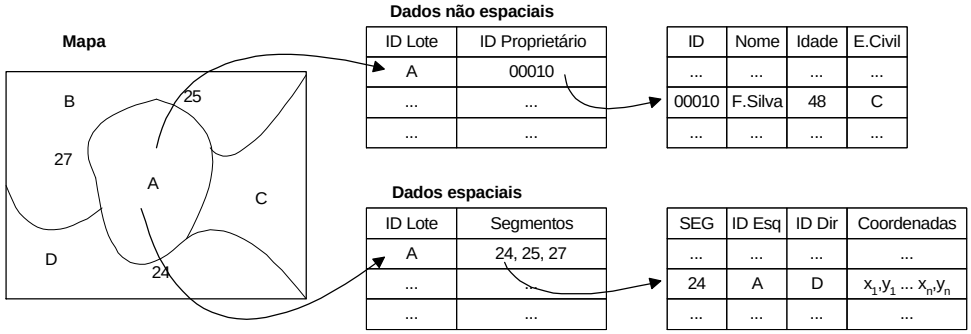


Figura 2.10: O modelo geo-relacional (Adaptado de: [Shepherd, 1991] p. 340)

Segundo Shepherd [Shepherd, 1991], a maioria dos SIG do tipo vectorial adoptam o modelo geo-relacional na conceptualização das ligações existentes entre dados espaciais e dados não espaciais. Esta aproximação passa por armazenar separadamente dados espaciais e dados não espaciais, estando o SGBD utilizado encarregado de assegurar a ligação entre os dois conjuntos de dados.

No modelo geo-relacional, um nível é utilizado para descrever propriedades geográ…cas, que de…nem associações entre o espaço geográ…co e um dado conjunto de atributos. Cada nível é caracterizado por representar um tipo de entidade geométrica (ponto, linha, polígono), declarado através das palavras reservadas GEOMETRIC TYPE, devendo ser identi…cado por um número, e opcionalmente, por um identi…cador:

DEFINE LAYER n <identi…cador> <TEMPORAL>

ATTR (at_1, Domain, <UNIQUE>), ..., (at_n, Domain, <UNIQUE>)

[15] Doravante designados esquemas lógicos, os quais são construídos recorrendo a componentes do tipo níveis e relações.

[16] Doravante designados esquemas externos, construídos recorrendo a componentes do tipo níveis virtuais, classes de objectos e restrições de integridade.

GEOMETRIC TYPE Geometric_type
<POSITIONING sistema_coordenadas>
<CONSTRAINT restrição>

Domain = {INT, REAL, STRING}
Geometric_type ={POINT, LINE, REGION}

As **relações** são utilizadas para modelar entidades não geográficas. A chave de uma relação resulta da concatenação dos atributos caracterizados com a palavra reservada KEY:

DEFINE RELATION nome_relação
ATTR (at_1, Domain, <KEY>), ..., (at_n, Domain, <KEY>)
<TIME_POINT at_j>

Os **níveis virtuais** são utilizados para descrever informação que é obtida (calculada) a partir da informação espacial definida nos diversos níveis do esquema lógico. Entre as operações que podem ser realizadas encontra-se a sobreposição, intersecção, união, etc., permitindo que novos níveis de informação sejam gerados a partir dos níveis existentes. O modelo geo-relacional disponibiliza quatro tipos de operações:

- derivação de atributos, que permite adicionar novos atributos a um nível, sem alterar as características geométricas do mesmo (construção de um novo atributo, baseado num ou mais atributos já existentes no nível);

  DEFINE V_LAYER n <nome> <TEMPORAL>
  AS COMPUTE ATTRS (m, novo_at_1 = $f_1(at_1, ..., at_n)$, ..., novo_at_k = $f_k(at_1, ..., at_n)$)

- computação geométrica, permite construir um novo nível, caracterizado por representar entidades com determinada característica geométrica, que foi calculada a partir das entidades geográficas existentes no nível que lhe deu origem;

  DEFINE V_LAYER n <nome> <TEMPORAL>
  AS COMPUTE GEOMETRIC (m, novo_at, $f(at_1, ..., at_n)$)

- sobreposição de dois níveis, permitindo criar um novo nível que é o resultado da intersecção das entidades geométricas contidas nos níveis que lhe dão origem;

  DEFINE V_LAYER n <nome> <TEMPORAL>
  AS OVERLAY (m, k)

- agregação, que permite integrar entidades geométricas adjacentes, que apresentem o mesmo valor para determinado atributo.

DEFINE V_LAYER n <nome> <TEMPORAL>

AS RECLASS OF (m, at_1, ..., at_n)

As aplicações geográ…cas requerem frequentemente a integração de atributos pertencentes a diversos níveis e relações. Classes de objectos permitem a agregação de entidades geométricas de diferentes tipos, mas que estão inseridas no mesmo espaço geográ…co. A de…nição de uma classe de objectos pode ser efectuada: i) agregando todas as entidades geométricas que se encontram nos níveis especi…cados em ON LAYERS; ii) limitando as entidades geométricas às discriminadas em GEOMETRIC TYPES; ou iii) restringindo os objectos intervenientes aqueles que possuem como atributos o conjunto mencionado em WITH ATTRIBUTES.

As opções ON LAYERS e GEOMETRIC TYPES poderão ser omitidas se a classe de objectos que está a ser de…nida constitui um subtipo, SUBTYPE OF, de outra. Neste caso é assumido que a classe a criar possui o mesmo nível e tipo geométrico.

DEFINE OBJECT CLASS nome_classe_objecto

<GEOMETRIC TYPE Geometric_type>

<SUBTYPE OF classe>

<ON LAYERS $id\_nível_1$, ..., $id\_nível_k$>

<WITH ATTRIBUTES at_1, ..., at_n>

<CONSTRAINT restrição>

Uma restrição, explícita através de predicados em lógica de 1ª ordem[17], indica o domínio de valores que as entidades da BD podem tomar. Ao nível dos predicados, estes podem ser construídos recorrendo a operadores aritméticos, operadores lógicos, funções espaciais (distância, área, ...) e operadores topológicos.

DEFINE CONSTRAINT restrição

ON {nome_classe_objecto j id_nível}

AMONG {$classe\_obj_1$, ..., $classe\_obj_k$ j $id\_nível_1$, ..., $id\_nível_k$}

AS condição

De seguida é apresentado um pequeno exemplo, que visa evidenciar a utilização do modelo geo-relacional na de…nição das estruturas de dados e restrições necessárias à execução de determinada questão.

Exemplo de utilização do modelo geo-relacional

Formulada a questão "Determinar todos os distritos que possuem pelo menos três escolas", as estruturas de dados e restrições necessárias à sua satisfação são:

DEFINE LAYER 1 Distrito

[17] A lógica de 1ª ordem (…rst-order logic) é uma linguagem de representação na qual as entidades a descrever são consideradas objectos, os quais possuem propriedades que os distinguem. Os objectos podem ser associados recorrendo a relações, algumas das quais representam funções que determinam univocamente determinado valor de saída [Russell e Norvig, 1995].

```
ATTR (ID_distrito, INT, UNIQUE), (nome_distrito, CHAR(20))
GEOMETRIC TYPE REGION
POSITIONING UTM      /* UTM é um sistema de coordenadas*/

DEFINE LAYER 2 Escola
ATTR (ID_escola, INT, UNIQUE), (nome_escola, CHAR(20))
GEOMETRIC TYPE POINT
POSITIONING UTM

DEFINE OBJECT CLASS Distritos_com_diversas_escolas
GEOMETRIC TYPE REGION,POINT
ON LAYERS Distrito, Escola
CONSTRAINT Diversas_Escolas

DEFINE CONSTRAINT Diversas_Escolas
ON Distrito
AMONG Distrito, Escola
AS (d 2 Distrito ^ 9x;y;z (x 2 Escola ^ y 2 Escola ^ z 2 Escola ^
    inside(x;d) ^ inside(y;d) ^ inside(z;d) ^ x 6= y ^ x 6= z ^ y 6= z))
```

2.3.3 A especi…cação formal no domínio geográ…co

Roman [Roman, 1990] propõe uma abordagem formal na modelação lógica de aplicações geográ…cas, na qual a utilização de predicados em lógica de 1 ª ordem permite a especi…cação dos requisitos de informação. Estes são compatíveis com as especi…cidades do PROLOG, permitindo a inferência lógica de informação. Predicados em lógica de 2 ª ordem[18] são utilizados para de…nir as regras que possibilitam a incorporação no sistema, de mecanismos de raciocínio espacial e temporal.

Esta abordagem assume que o mundo real pode ser modelado por uma colecção de objectos, que permitem retratar as diferentes perspectivas que os utilizadores têm das entidades geográ…cas. Estes objectos são formalmente de…nidos recorrendo a predicados de 1 ªordem, os quais constituem conhecimento básico (basic facts), representando os factos assumidos como verdadeiros, ou conhecimento virtual (virtual facts), que é construído recorrendo a outros factos (básicos ou virtuais).

Ao nível semântico, predicados em lógica de 2 ª ordem são utilizados na de…nição do conhecimento espacial e temporal a incorporar no sistema de raciocínio. O conhecimento semântico permite a de…nição do instante temporal e do espaço geográ…co em que determinado facto é verdadeiro.

[18] A lógica de 1 ª ordem apenas permite a quanti…cação de objectos. A lógica de 2 ªordem permite, além da quanti…cação de objectos, a quanti…cação de relações e funções (Ex: "dois objectos são iguais sse todas as propriedades aplicadas aos mesmos são equivalentes": $8_{x;y}$ (x = y) , (8_p p(x) , p(y))) [Russell e Norvig, 1995].

A consistência semântica entre os factos é conseguida recorrendo à especi…cação de restrições, que dependem do contexto em que a informação é utilizada. Esta aproximação permite a construção de diversos modelos sobre os dados, os quais retratam diferentes perspectivas dos mesmos.

Meta-factos e meta-restrições podem ser de…nidos em lógica de 2ª ordem, permitindo a especi…cação de conhecimento que não está limitado a determinado contexto (utilizador ou aplicação), mas que constitui os mecanismos básicos de raciocínio. Estes podem ainda ser agrupados em meta-modelos, que possibilitam o encapsulamento do conjunto de regras aplicáveis a determinado domínio de aplicação.

A especi…cação de um modelo de dados, utilizando a abordagem formal proposta por Roman [Roman, 1990], passa então pela de…nição de:

Conhecimento básico. Um facto pode assumir a forma de uma propriedade considerada verdadeira para um dado objecto, ou de uma propriedade veri…cada entre objectos, de…nida recorrendo a uma relação:

```
estrada(x1).
estrada(x2).
intersecta(x1,x2).
```

Restrições. Uma restrição tem como objectivo validar o conhecimento básico existente no sistema, detectando eventuais inconsistências e alertando o utilizador no caso destas ocorrerem:

```
(8 X;Y;Z):(concelho_de(X,Y)^ concelho_de(X,Z)^(Y6=Z)
          =) ERRO(pertence_a_dois_distritos,X)).
```

Conhecimento semântico. O conhecimento semântico permite a utilização de valores (quanti…cação) na quali…cação das propriedades dos objectos. Estes valores não podem ser tratados como objectos. A sua especi…cação passa pela extensão dos predicados, por forma a estes integrarem o valor do atributo e o respectivo objecto. Por exemplo, o conhecimento semântico associado a: "A temperatura média de Braga é 20 graus", pode ser especi…cado através de:

```
temperatura_média(20)(Braga).
```

Modelos. O conceito de modelo tem como objectivo permitir retratar diferentes perspectivas dos dados, ou alterações no domínio de aplicação, que requerem a re-interpretação dos mesmos. Os modelos permitem indicar o contexto em que os factos são válidos. Esta especi…cação é conseguida acrescentando o nome do modelo, seguido de um quali…cador (’), antes da de…nição do facto: celsius’temperatura_congelação(0)(x). Este facto indica que na escala Celsius, a temperatura de congelação do objecto x é de 0º.

Meta-factos (regras de inferência). Os meta-factos permitem especi…car axiomas[19] e regras de inferência que permitem o raciocínio num domínio semântico especí…co. Integram conhecimento, sob a forma de regras, que é independente de predicados particulares. A independência é conseguida recorrendo a predicados em lógica de 2ª ordem, que permitem a utilização de variáveis quanti…cadas universalmente, válidas para determinado conjunto de predicados. O meta-facto "um facto não explícito como sendo verdadeiro é assumido como sendo falso", pode ser especi…cado através de:

$$(8_{M;P;X}):(M'P(X)=) \ M'P(true)(X)),$$
$$(8_{M;P;X}):(MODELO(M)\wedge PREDICADO(P)\wedge OBJECTO(X)\wedge not(M'P(true)(X)) =) \ M'P(false)(X)).$$

Meta-restrições. As meta-restrições têm o mesmo objectivo das restrições, mas possuem a vantagem de poder utilizar na sua de…nição meta-factos já especi…cados. A meta-restrição "nenhum facto pode ser simultaneamente verdadeiro e falso", é conseguida através do predicado:

$$(8_{M;P;X}):(M'P(true)(X)\wedge M'P(false)(X) =) \ M'ERROR(contradição,P,X)).$$

Meta-modelos. Os meta-factos e as meta-restrições podem ser agrupadas num conjunto denominado de meta-regras, que quando integrado com o conhecimento semântico especí…co de um dado domínio de aplicação, constitui um meta-modelo. Os meta-modelos permitem que diferentes regras de inferência sejam utilizadas sobre os mesmos dados, aumentando a produtividade na avaliação de vários modelos.

2.4 Normalização em Informação Geográ…ca

A evolução ocorrida na área das TI e o consequente aumento da quantidade de dados armazenados, da diversidade de arquitecturas e dos requisitos de cada sistema, conduz inevitavelmente à necessidade de de…nição de normas que permitam aos utilizadores integrar dados provenientes de diferentes sistemas.

No caso especí…co da informação geográ…ca [CEN/TC-287, 1996c], a normalização tem como …nalidade permitir que a informação geográ…ca possa ser acedida por diferentes utilizadores, aplicações e sistemas, em diferentes localizações. Para tal, são necessárias regras que permitam descrever a informação geográ…ca, métodos para a sua estruturação e codi…cação, e ainda, mecanismos para aceder, transferir e actualizar a referida informação. Entre os benefícios associadas à de…nição de normas, para a área da informação geográ…ca, encontram-se:

- Clari…car os conceitos associados à informação geográ…ca;

[19] Os axiomas são habitualmente utilizados pelos matemáticos para de…nir os factos básicos acerca de um domínio e de…nir outros conceitos relacionados com estes factos, utilizando posteriormente os axiomas e de…nições para provar teoremas. Na área da inteligência arti…cial, são denominados de axiomas as sentenças que inicialmente se encontram na base de conhecimento utilizada [Russell e Norvig, 1995].

- Harmonizar o uso de informação geográ...ca e os diferentes métodos para a aceder;
- Permitir a integração de informação geográ...ca;
- Aumentar a disponibilidade de informação geográ...ca, incluindo a sua meta-informação;
- Utilizar a informação geográ...ca em diferentes aplicações;
- Permitir a transferência de informação geográ...ca e consequentemente a sua reutilização em diferentes contextos.

A normalização conduz à estruturação de conceitos e componentes, que permitem a de-...nição, descrição, estruturação, pesquisa, alteração e transferência de informação geográ...ca e sua meta-informação [Tom, 1994]. A próxima subsecção sistematiza as principais iniciativas de normalização, em curso na área da informação geográ...ca.

2.4.1 Principais grupos de trabalho

Os dois principais grupos de trabalho, na área da normalização da informação geográ...ca, incluem os grupos de trabalho CEN TC 287 e ISO TC 211. O primeiro, europeu, foi constituído em Outubro de 1991 e integra 22 países, entre os quais Portugal, representado pelo Instituto Português da Qualidade[20] (IPQ). O trabalho desta comissão, dividido em quatro grupos (WG - Working Group), deu origem às pré-normas europeias [CEN/TC-287, 1998d], que englobam os seguintes aspectos:

WG1 - Fundamentos	287001 Modelo de referência
	287002 Descrição geral
	287003 Vocabulário
WG2 - Descrição dos dados	287006 Regras para esquemas de aplicação
	287007 Geometria
	287008 Qualidade
	287009 Metadados
	287010 Transferência
WG3 - Referenciação	287011 Posicionamento directo
	287012 Tempo[21]
	287014 Sistemas de posicionamento indirecto
WG4 - Processamento	287013 Pesquisa e alteração

[20] A nível nacional, o IPQ constituiu a comissão técnica CT 134 (Geomática, Informação Geográ...ca e Cartogra...a), responsável pela tradução/adaptação das pré-normas para o contexto nacional.

[21] No que diz respeito a este item de trabalho, o CEN TC 287 suspendeu os trabalhos associados ao mesmo, decidindo que para este tópico será adoptado como EN (European Standard) o resultado do trabalho do ISO TC 211 ([CEN/TC-287, 1998d] p.13).

A comissão técnica ISO TC 211, com responsabilidades na produção de normas para Informação Geográ...ca e Geomática[22], foi constituída em Abril de 1994, tendo reunido pela primeira vez em Novembro do mesmo ano, e estando o termo dos seus trabalhos previsto para o ano 2001. Esta comissão é presidida pela Noruega, englobando 25 países colaboradores e 13 países observadores, sendo este último o estatuto de Portugal (mais uma vez representado pelo IPQ). Os documentos aprovados darão origem a norma ISO 15046 [ISO/TC-211, 1999a], cujos trabalhos estão divididos em 5 grupos, que englobam os seguintes aspectos:

WG1 - Enquadramento	15046-1 Modelo de referência 15046-2 Descrição geral 15046-3 Linguagem para o esquema conceptual 15046-4 Vocabulário 15046-5 Veri...cação e teste
WG2 - Per...s e normas funcionais	15046-6 Per...s
WG3 - Operadores e modelos de dados	15046-7 Esquema espacial 15046-8 Esquema temporal 15046-9 Regras para o esquema de aplicação
WG4 - Administração de dados	15046-10 Metodologia de catalogação 15046-11 Referenciação espacial por coordenadas 15046-12 Referenciação por identi...cadores geográ...cos 15046-13 Princípios de qualidade 15046-14 Procedimentos de avaliação da qualidade 15046-15 Metadados
WG5 - Serviços de informação geográ...ca	15046-16 Serviços de posicionamento 15046-17 Descrição de informação geográ...ca 15046-18 Codi...cação 15046-19 Serviços

As comissões CEN TC 287 e ISO TC 211 partilham membros, e estão relacionadas através de um acordo de cooperação, denominado Acordo de Viena, que visa evitar a duplicação de trabalho e garantir a harmonização das normas produzidas [Matos, 1997].

Ao nível da interoperabilidade entre sistemas, destaca-se o trabalho do OpenGIS, cujo objectivo é a de...nição de normas que garantam que a informação apresenta sempre o mesmo conteúdo, independentemente do sistema/aplicação utilizado. O OpenGIS tem-se debruçado sobre o desenvolvimento do Open Geodata Interoperability Speci...cation (OGIS) que se divide em dois grupos: Abstracção e Implementação [OGC, 1999a]. A abstracção tem como objectivo criar e documentar um modelo conceptual que permita a de...nição de especi...cações para a implementação.

Este consórcio integra vendedores de tecnologia SIG, investigadores e organizações priva-

[22] A Geomática surge de...nida como sendo a disciplina associada ao armazenamento, processamento, visualização e distribuição de dados geográ...cos [Santos, 1998].

das e governamentais, associadas à tecnologia SIG ou às ciências da computação [Tom, 1994]. Apesar de representar uma iniciativa de normalização privada (não vinculada a qualquer instituição de normalização), os seus trabalhos apresentam um elevado potencial, já que as suas resoluções são integradas nas aplicações comercializadas pelos produtores que integram o grupo de trabalho.

Entre as vantagens associadas ao trabalho desenvolvido pelo OpenGIS, destaca-se que pelo mesmo não estar condicionado a aprovação de nenhuma organização de normalização, as suas especi...cações são rapidamente elaboradas e testadas. Dentre os seus desenvolvimentos, destaca-se a de...nição de librarias com tipos de dados espaciais (pontos, linhas, polígonos) e operações sobre estes tipos (intersecção, sobreposição, ...) por forma a facilitar a troca de dados entre diferentes aplicações [Shekhar et al., 1997].

2.4.2 As pré-normas CEN TC 287 para Informação Geográ...ca

Estas pré-normas têm como principal objectivo permitir que a informação geográ...ca possa ser acedida por diferentes utilizadores, sistemas, aplicações e principalmente, de diferentes localizações. Para tal é necessário de...nir e descrever a informação geográ...ca de uma forma padronizada, de...nir os métodos e estruturas para o seu armazenamento e ainda, de...nir como esta informação pode ser acedida, alterada, pesquisada e transferida [CEN/TC-287, 1998e].

O trabalho de revisão dos documentos produzidos por esta comissão foram iniciados em meados de 1998. À data, os documentos disponíveis eram agrupados em dois grandes grupos: ENV (European Prestandards), constituindo as pré-normas até à votação formal das mesmas estar concluída; e CR (Comission Report), representando relatórios de suporte aos documentos ENV produzidos.

O trabalho de normalização foi dividido em quatro grupos: fundamentos (modelo de referência), descrição dos dados (esquemas de aplicação, espacial, qualidade, metadados e transferência), referenciação (posicionamento directo e sistemas de posicionamento indirecto) e processamento (pesquisa e alteração). A Tabela 2.3 sistematiza os documentos produzidos em cada um destes grupos de trabalho.

O modelo de referência

O modelo de referência tem como objectivo enquadrar os conceitos associados a informação geográ...ca, descrevendo, através da especi...cação de um esquema conceptual, os dados geográ...cos, os seus metadados, e ainda os serviços para a sua manipulação. O esquema conceptual representa o modelo conceptual para o domínio da informação geográ...ca, e compreende a integração de oito esquemas: semântico, espacial, qualidade, posicionamento directo, posicionamento indirecto, metadados, pesquisa e alteração, e por último, transferência [CEN/TC-287, 1996c]. A integração destes oito esquemas, para um domínio de aplicação especí...co, dá origem a um esquema de aplicação (Figura 2.11).

No esquema conceptual, os dados geográ...cos constituem representações de informação geográ...ca, que são descritas recorrendo a aspectos semânticos, espaciais e de qualidade:

- Aspectos semânticos. Caracterizam-se por permitir a identi...cação das entidades geo-

Grupos de Trabalho		Documentos
Fundamentos	287001: Modelo de referência	ENV 12009 – Modelo de referência
	287002: Descrição geral	CR 287002 – Descrição geral
	287003: Vocabulário	CR 287003 – Vocabulário
Descrição dos dados	287006: Regras para esquemas de aplicação	ENV 287006 – Esquemas de aplicação CR 287005 – Linguagem para esquemas conceptuais
	287007: Geometria	ENV 12160 – Esquema espacial
	287008: Qualidade	ENV 12656 – Qualidade
	287009: Metadados	ENV 12657 – Metadados
	287010: Transferência	ENV 12658 – Transferência
Referenciação	287011: Posicionamento directo	ENV 12762 – Posicionamento directo
	287014: Sistemas de posicionamento indirecto	ENV 12661 – Identificadores geográficos
Processamento	287013: Pesquisa e alteração	CR 12660 – Pesquisa e alteração: aspectos espaciais

Tabela 2.3: CEN TC 287: Grupos de trabalho e documentos produzidos

grá…cas e seus respectivos atributos, assim como os diferentes relacionamentos existentes entre as mesmas. A esquematização de toda esta informação, para um dado domínio de aplicação, dá origem ao esquema semântico [CEN/TC-287, 1996c]. Este esquema permite que diversos utilizadores possam adquirir um entendimento comum acerca das entidades identi…cadas.

- Aspectos espaciais. Os aspectos espaciais permitem a utilização de estruturas de representação da informação especí…cas, a identi…cação do posicionamento dos objectos, e a de…nição das características geométricas e topológicas associadas aos mesmos. O sistema de posicionamento pode ser directo, recorrendo ao uso de coordenadas, ou indirecto, permitindo a identi…cação de uma localização através da utilização de uma morada ou outro identi…cador geográ…co. Estes aspectos são de…nidos recorrendo ao esquema espacial [CEN/TC-287, 1996b], esquema de posicionamento directo [CEN/TC-287, 1998g] e ao esquema de identi…cadores geográ…cos [CEN/TC-287, 1998h], respectivamente. As características espaciais permitem que dados geográ…cos:

  – De diferentes origens possam ser integrados através da geo-referenciação;
  – Sejam analisados utilizando operadores e funções espaciais;
  – Sejam representados gra…camente em mapas.

- Aspectos de qualidade. Os parâmetros de qualidade permitem determinar a usabilidade de um conjunto de dados, através da caracterização dos mesmos segundo a sua: precisão (accuracy), exactidão (completeness) e actualidade (up-to-dateness). O esquema de qualidade [CEN/TC-287, 1998b] permite ao utilizador avaliar a qualidade da informação que lhe é disponibilizada, veri…cando se a mesma satisfaz os requisitos desejados.

A descrição dos metadados (através do esquema de metadados [CEN/TC-287, 1998a])

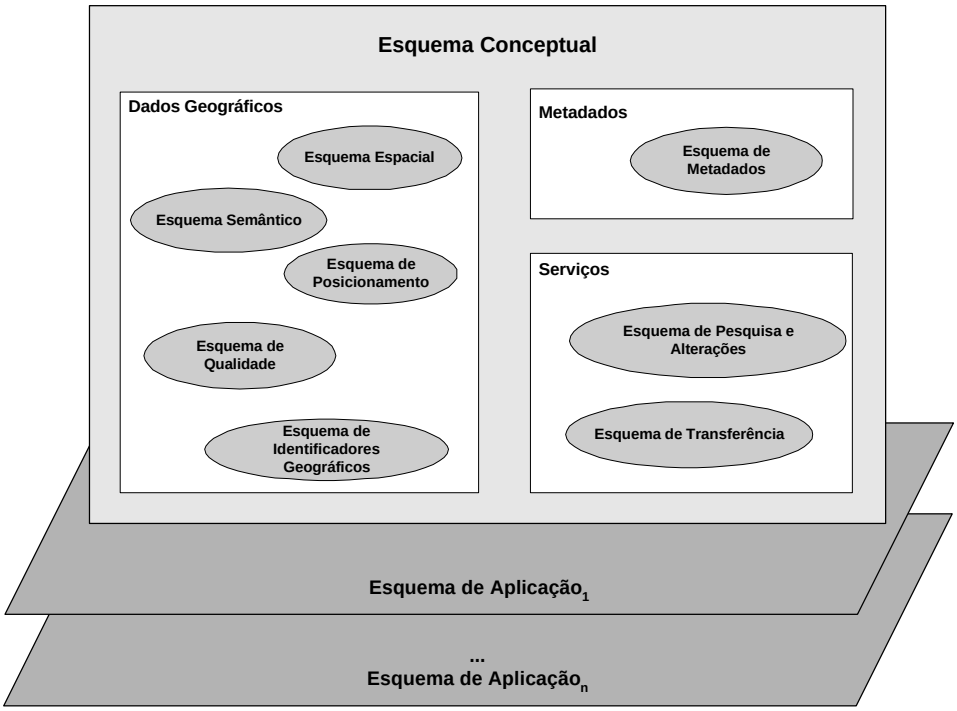


Figura 2.11: Esquema conceptual

permite caracterizar os dados através da identi…cação do seu proprietário, conteúdo e estrutura, e ainda, de…nir a sua disponibilidade e formas de distribuição.

Entre os serviços que podem ser disponibilizados, num serviço de informação geográ…ca, a comissão especi…ca os requisitos de um serviço de transferência e de um serviço de pesquisa e alteração:

- Serviço de pesquisa e alteração. Permite aos utilizadores colocar questões, ou receber respostas, relacionadas com os dados ou metadados geográ…cos armazenados num dado serviço de informação. A sua especi…cação dá origem ao esquema de pesquisa e alteração [CEN/TC-287, 1998f], e é suportado pelos serviços de transferência.
- Serviço de transferência. O esquema de transferência [CEN/TC-287, 1998c] de…ne as regras para a troca de dados geográ…cos entre sistemas, assim como dos seus respectivos metadados.

O esquema espacial e o esquema de identi…cadores geográ…cos

A descrição dos dados tem como principal objectivo permitir a comunicação de vários intervenientes, dentro de um sistema de informação ou entre sistemas de informação. Esta descrição especi…ca a estrutura dos dados e clari…ca a sua semântica através da modelação. Este processo inclui a construção de esquemas que de…nem e descrevem os dados, assim como as regras aplicáveis aos mesmos.

A linguagem escolhida[23] pelo CEN TC 287 para a de…nição dos esquemas é o EXPRESS[24] (ISO 10303-11) [CEN/TC-287, 1996a]. Esta linguagem permite modelar o esquema de aplicação, independentemente de qualquer implementação física.

O esquema de aplicação permite a diversos utilizadores obter um entendimento correcto dos dados modelados. Apesar das normas requererem a integração de oito esquemas, neste trabalho utilizam-se apenas dois: o espacial e o de identi…cadores geográ…cos. O esquema espacial porque permite especi…car os aspectos espaciais (como a topologia) relacionados aos identi…-cadores geográ…cos utilizados na referenciação da informação. O esquema de identi…cadores geográ…cos permite a de…nição do sistema de identi…cadores geográ…cos e do catálogo de localizações, que implementam o sistema de posicionamento indirecto adoptado. A integração destes esquemas permite a inclusão da componente geográ…ca associada aos identi…cadores utilizados, no processo de descoberta de conhecimento.

De seguida são apresentadas as especi…cações associadas ao esquema espacial e ao esquema de identi…cadores geográ…cos.

O esquema espacial

A caracterização dos aspectos espaciais da informação geográ…ca permite que dados de diferentes origens possam ser integrados através da geo-referenciação, ou ainda, que os mesmos possam ser analisados utilizando operadores e funções espaciais.

O esquema espacial agrega duas componentes, a geométrica e a topológica, para as quais disponibiliza estruturas especializadas necessárias à sua de…nição. Entre as primitivas geométricas encontra-se o ponto, a linha, a área ou o volume, no espaço bi ou tri-dimensional [CEN/TC-287, 1998e]. Estas primitivas passam a possuir topologia quando se relacionam umas com as outras.

A geometria permite descrever quantitativamente, através de coordenadas e funções matemáticas, os aspectos espaciais da informação geográ…ca[25]. A topologia permite descrever qualitativamente estes aspectos espaciais, permanecendo os mesmos invariantes a qualquer transformação do espaço. Entre as primitivas topológicas encontra-se o nodo, a aresta, a face e os anéis.

No que diz respeito aos nodos, estes podem ser de ligação ou isolados. Os nodos de ligação interligam duas ou mais arestas, podendo os mesmos ser intermédios, iniciais ou terminais, consoante a sua posição. Os nodos isolados podem pertencer a uma ou mais faces (Figura 2.12).

[23] A escolha foi baseada na análise de doze linguagens, das quais uma selecção prévia conduziu a comparação de seis: o EXPRESS, o IDEF1X, o INTERLIS, o NIAM, o ODMG-93 ODL e o SQL3. O método e critérios de avaliação utilizados conduziram a escolha do EXPRESS, podendo os mesmos ser consultados em [CEN/TC-287, 1996a]. Refere-se que o UML constitui a linguagem adoptada pelo ISO TC 211 e pelo OpenGIS, para a especi…cação das normas/especi…cações produzidas.

[24] O EXPRESS é uma linguagem de modelação baseada em objectos, que utiliza como componentes: esquemas, entidades, atributos, relacionamentos e restrições [Schenck e Wilson, 1994]. Um esquema de…ne um contexto, no qual um conjunto de entidades partilham de…nições semânticas. Uma entidade representa um objecto do mundo real ou um conceito de interesse. Possui atributos que a caracterizam e relacionamentos com outras entidades. As restrições são utilizadas para delimitar valores de atributos ou para restringir os relacionamentos existentes entre entidades. A modelação pode ser efectuada recorrendo a uma notação grá…ca e uma linguagem léxica, utilizada no processamento computacional do respectivo esquema.

[25] Uma vez que este trabalho utiliza identi…cadores qualitativos na geo-referenciação da informação, os aspectos geométricos, que podem ser especi…cados recorrendo ao esquema espacial, não são futuramente considerados.

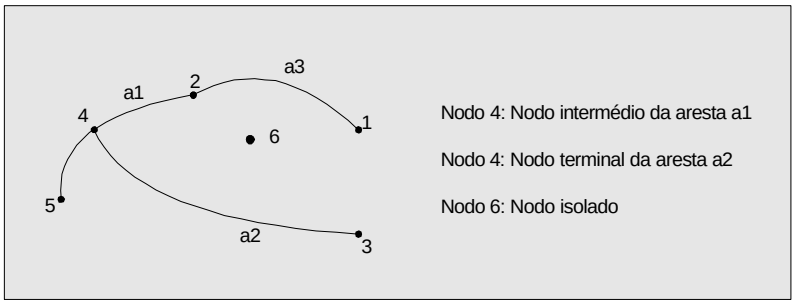


Figura 2.12: Diferentes tipos de nodos

As arestas são primitivas topológicas uni-dimensionais, que representam uma ligação orientada entre dois nodos (Figura 2.13), que podem por sua vez ser coincidentes. Em termos topológicos, uma aresta:

- tem associado um nodo inicial e um nodo …nal;
- pode conter 0 ou mais nodos intermédios;
- tem 0 ou 1 aresta à direita e 0 ou 1 uma aresta à esquerda;
- tem 0 ou 1 aresta anterior à esquerda e 0 ou 1 aresta anterior à direita;
- tem 0 ou mais faces à esquerda e 0 ou mais faces à direita;
- pode ser elemento de um ou mais anéis.

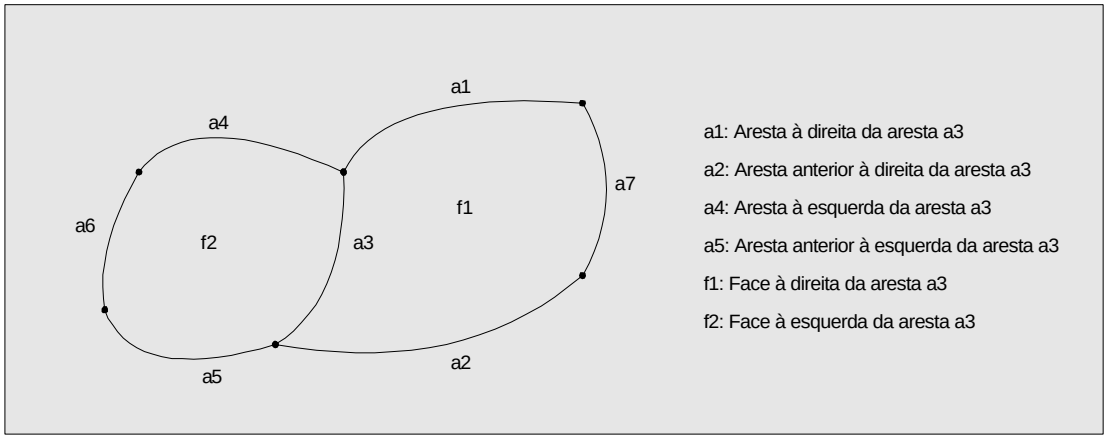


Figura 2.13: Caracterização da localização das arestas

Uma face é uma primitiva topológica bi-dimensional, descrita recorrendo a um anel exterior e 0 ou mais anéis interiores. Um anel integra um conjunto de uma ou mais arestas interligadas, que não se intersectam. Uma face pode ser de…nida recorrendo a relações entre

arestas e faces, recorrendo a relações entre arestas, ou recorrendo a relações entre faces, anéis e arestas. Em termos topológicos:

- um anel é composto por uma ou mais arestas;
- um anel é anel exterior de 0 ou 1 face;
- um anel é anel interior de 0 ou 1 face;
- uma face tem 1 anel exterior;
- uma face tem 0 ou mais anéis interiores;
- uma face pode conter 0 ou mais nodos isolados.

O esquema espacial pode ser construído recorrendo a um dos 8 esquemas espaciais prede-…nidos (G1..G8) pela comissão técnica [CEN/TC-287, 1996c]. Estes esquemas foram derivados de um esquema espacial genérico G0, que deve ser utilizado sempre que nenhum dos esquemas, G1 a G8, se adapte às especi…cidades do domínio geográ…co em causa. Os esquemas prede…nidos têm contextos de aplicação especí…cos, sendo eles:

Mapa de topologia planar (G1). Este esquema é utilizado sempre que o domínio geográ…co em causa for constituído por mapas planos[26] com faces que cobrem toda a superfície em análise, não existindo qualquer sobreposição entre as mesmas. Neste esquema são permitidos nodos isolados ou terminais, mas não permitem a existência de nodos intermédios. As faces têm de ser de…nidas sem qualquer orifício.

Rede linear de mapas planos (G2). Neste esquema apenas são permitidas arestas que não se interceptem, excepto na concordância de nodos iniciais e terminais. Permite a existência de nodos isolados, mas não de nodos intermédios.

Rede linear de mapas não planos (G3). Este esquema constitui uma excepção ao esquema G2, permitindo a existência de intersecções entre arestas, sem a existência de um nodo no local da intersecção. Não permite a de…nição de faces.

Rede linear de mapas não planos com superfícies (G4). Constitui uma extensão ao esquema G3, permitindo a existência de superfícies (faces) com sobreposições ou orifícios.

Esparguete (G5). Permite a de…nição de pontos e curvas, sem de…nição topológica e sem a imposição de restrições.

Rede triangular irregular (G6). Neste esquema, uma face é de…nida por 3 arestas. As faces são então triangulares, sem qualquer orifício ou sobreposição. Permite apenas a utilização de nodos terminais, sendo estes os únicos que são descritos recorrendo à componente geométrica.

[26] A fórmula de Euler de…ne que num mapa plano, número_faces + número_nodos - número_arestas = 1 [Shekhar et al., 1999].

Imagens raster (G7). Permite a de…nição espacial de imagens raster, não impondo restrições na utilização das primitivas geométricas (as quais não possuem qualquer caracterização topológica).

Grelha (G8). Este esquema é equivalente ao G7, mas não é especi…cada qualquer restrição em relação ao mesmo.

Da análise das características associadas a cada um dos esquemas anteriores, veri…ca-se que o esquema que satisfaz os requisitos do domínio geográ…co a caracterizar neste estudo, mapa plano com divisões administrativas de Portugal, é o mapa de topologia planar G1. As faces que constituem o mapa não possuem qualquer sobreposição, ou orifício ao nível de cobertura. Também não se veri…ca a existência de qualquer nodo intermédio, uma vez que as arestas que de…nem os limites administrativos não se intersectam.

O esquema espacial G1 possui duas rami…cações: uma geométrica e uma topológica. Uma vez que neste trabalho a geo-referenciação da informação é obtida através da utilização de identi…cadores geográ…cos, só se utiliza a rami…cação respeitante à topologia, apesar da Figura 2.14 apresentar as duas vertentes (já representadas recorrendo à linguagem UML).

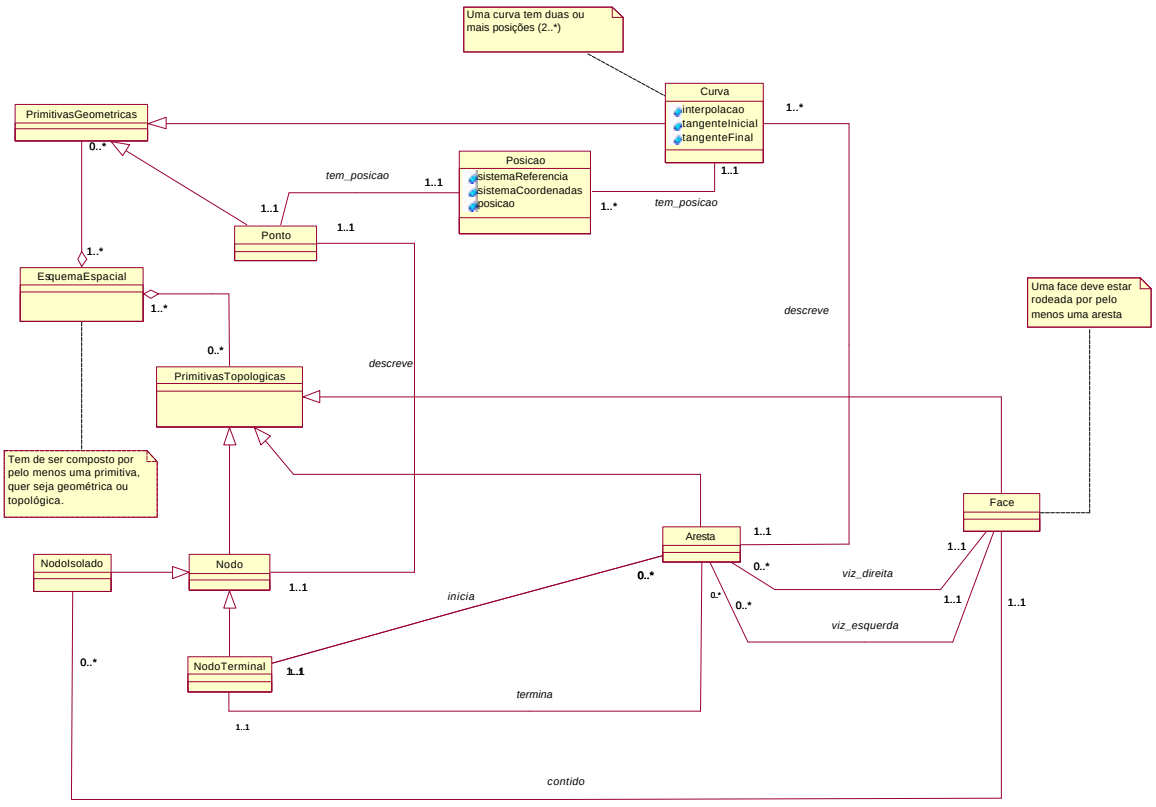


Figura 2.14: Esquema espacial: primitivas geométricas e topológicas

Da análise da Figura 2.14 constata-se que, a determinação da vizinhança existente entre as diversas faces é conseguida veri…cando a posição das arestas que as delimitam. Uma vez que a identi…cação das vizinhanças será sempre necessária (podendo ser efectuada automaticamente através de um SIG), o diagrama apresentado na Figura 2.14 foi alterado com vista a integrar os atributos necessários à de…nição das relações espaciais, necessárias ao sistema de raciocínio espacial qualitativo a implementar neste trabalho. A Figura 2.15 retrata o diagrama resultante (apenas a componente topológica), no qual se explicita a direcção[27] e distância existente entre faces adjacentes.

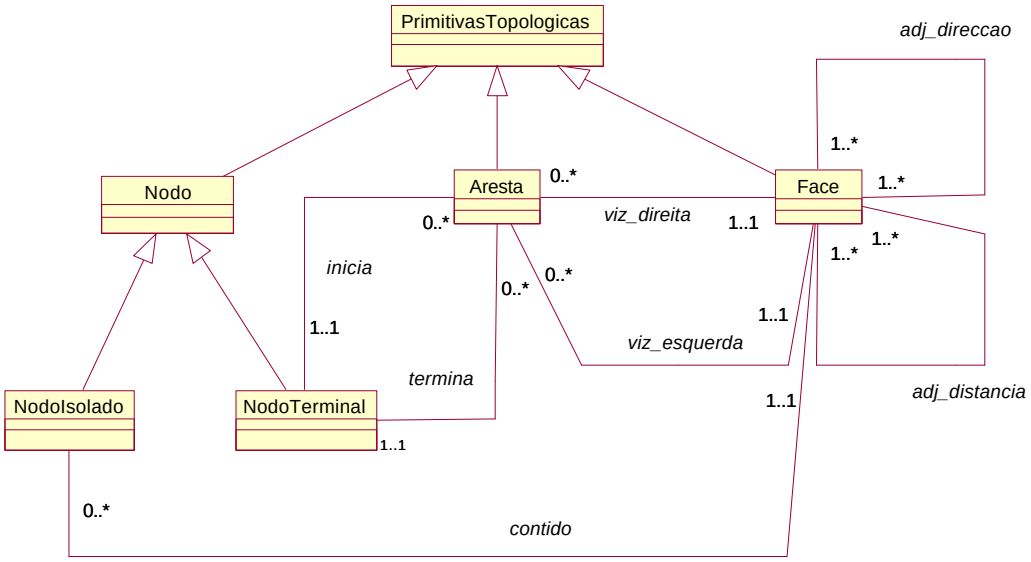


Figura 2.15: Alterações introduzidas à componente topológica do esquema espacial

Destaca-se que o novo diagrama apresentado na Figura 2.15 (evidenciando apenas a componente topológica) complementa as especi…cações proferidas pelas pré-normas, acrescentando informação que é extremamente importante na percepção dos relacionamentos espaciais existentes entre regiões, neste caso divisões administrativas.

A informação qualitativa adicionada à componente topológica pode ser obtida, no caso das pré-normas ISO TC 211, recorrendo a funções geométricas. Neste caso, características espaciais são descritas por um ou mais atributos espaciais, obtidos através de objectos geométricos (`GM_Object`) ou de objectos topológicos (`TP_Object`). Ao nível geométrico, as características espaciais incluem a de…nição da dimensão, posição, forma e orientação dos objectos. A título

[27] Chama-se a atenção para o facto de, como já referido anteriormente, a orientação de um dado objecto espacial ser uma característica geométrica. Contudo, a orientação como relacionamento espacial entre dois objectos, no contexto em que é aqui utilizada, não constitui uma característica passível de representação na componente geométrica do esquema espacial. Assim sendo, e dada a abordagem qualitativa aqui adoptada, a direcção e a distância podem ser automaticamente calculadas recorrendo a um SIG, num processo transparente ao utilizador. Neste caso, estes atributos são posteriormente armazenados na componente topológica do esquema espacial, os quais serão utilizados na sua forma qualitativa. O processo de determinação automática destes atributos é apresentada no Capítulo 5, no qual é descrita a implementação do sistema `Padrão`.

de exemplo salienta-se que, no vasto conjunto de funções geométricas de…nidas nos documentos de trabalho do ISO TC 211 [ISO/TC-211, 1999c], as mais relevantes para o trabalho aqui apresentado são:

`GM_Object::mbRegion() : GM_Object`. A função `mbRegion` permite conhecer a região que contém determinado objecto. Permite a de…nição de hierarquias conceptuais, no caso de desconhecimento das mesmas.

`GM_Object::distance(geometry:GM_Object) : Distance`. A função `distance` permite conhecer a distância existente entre dois objectos geográ…cos.

`GM_Object::centroid() : DirectPosition`. A função `centroid` determina o centróide da região identi…cada pelo `GM_Object`.

No caso especí…co do Padrão, a informação respeitante à direcção e à distância existente regiões, foi obtida automaticamente manipulando os objectos geográ…cos disponibilizados pelo SIG utilizado. As rotinas implementadas permitiram determinar os atributos em causa, veri…cando a orientação e distância existente entre os centróides das regiões analisadas (opção que é justi…cada no Capítulo 3).

O esquema de identi…cadores geográ…cos

Num sistema de referenciação espacial através de identi…cadores qualitativos, uma posição é indexada a uma localização recorrendo a um objecto real, sendo o seu identi…cador denominado identi…cador geográ…co. Neste contexto inserem-se nomes de ruas, de cidades, de monumentos, etc. Este sistema de referência é apelidado de indirecto, uma vez que não recorre à utilização de coordenadas.

O esquema de identi…cadores geográ…cos permite de…nir, através da construção de um sistema de identi…cadores geográ…cos, o conjunto de identi…cadores geográ…cos utilizados na georeferenciação da informação. Associado ao sistema de identi…cadores geográ…cos está um catálogo de localizações, que explicita os diversos identi…cadores utilizados e ainda os relacionamentos existentes entre os mesmos [CEN/TC-287, 1998h]. A Figura 2.16 apresenta a estrutura de um sistema de identi…cadores geográ…cos, realçando o conteúdo do catálogo de localizações, para um sistema de referenciação indirecto que considera divisões administrativas ao nível do País, Distrito e Concelho.

Um sistema de identi…cadores geográ…cos agrega uma colecção de classes de localização e seus respectivos identi…cadores geográ…cos. As localizações devem referenciar entidades reais, com qualquer dimensão geométrica (0 ou mais dimensões), devendo as mesmas ser independentes (existirem só por si) e contíguas (fornecendo uma divisão completa do domínio geográ…co caracterizado). Os identi…cadores geográ…cos devem permitir identi…car univocamente uma localização. O conjunto de instâncias de localização consideradas, e seus respectivos identi…cadores geográ…cos, devem ser compilados num catálogo de localizações.

A descrição do sistema de identi…cadores geográ…cos inclui a descrição do sistema propriamente dito e a descrição de cada classe de localização. A interpretação dos diagramas EXPRESS especi…cados no documento prENV 12661 ([CEN/TC-287, 1998h] Anexo A e Anexo C) permitiu

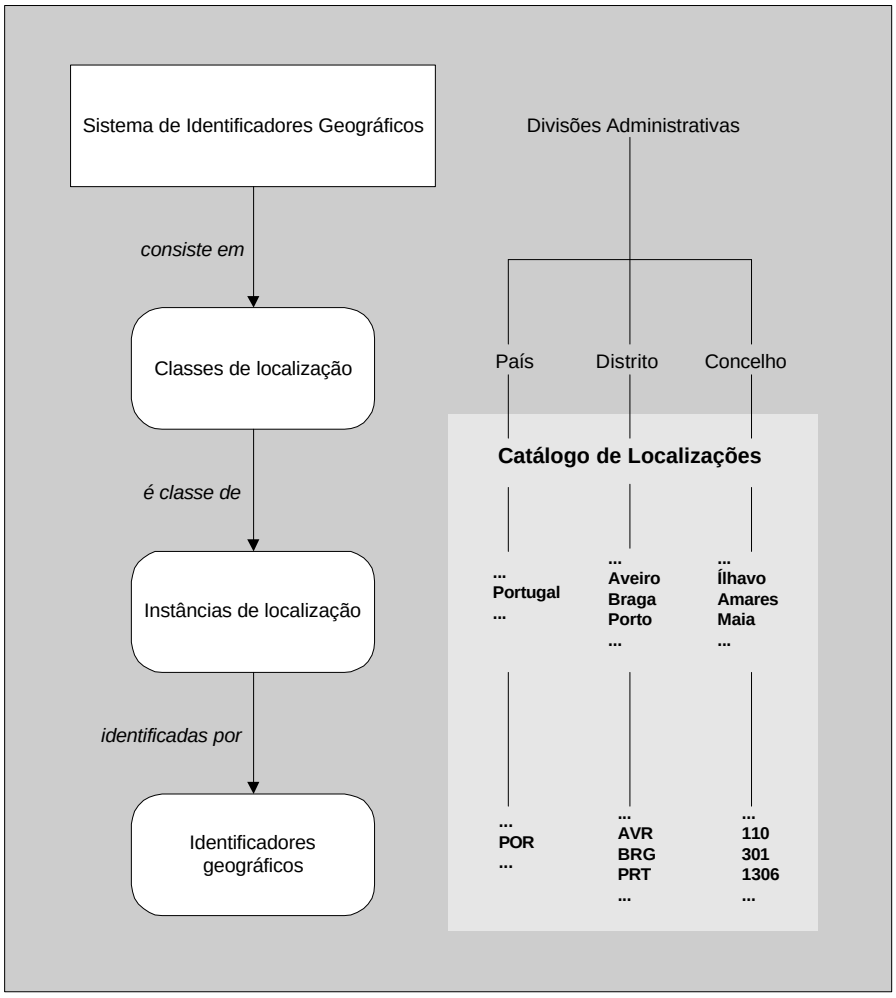


Figura 2.16: Sistema de Identi…cadores Geográ…cos e Catálogo de Localizações

a construção de um esquema de identi…cadores geográ…cos, que integra a especi…cação do sistema de identi…cadores geográ…cos utilizado e ainda o correspondente catálogo de localizações. A implementação deste esquema permite a qualquer utilizador conhecer todos os detalhes associados ao sistema de referenciação adoptado. A Figura 2.17 apresenta o esquema resultante, representado recorrendo, na linguagem UML, a um diagrama de classes.

Ao nível do esquema de identi…cadores geográ…cos, destaca-se que o trabalho desenvolvido pelo ISO TC 211 vem de encontro ao desenvolvido pelo CEN TC 287, requerendo a de…nição de tipos de localização, instâncias de localização e catálogo geográ…co [ISO/TC-211, 1999b], na construção de um sistema de referenciação espacial baseado em identi…cadores geográ…cos.

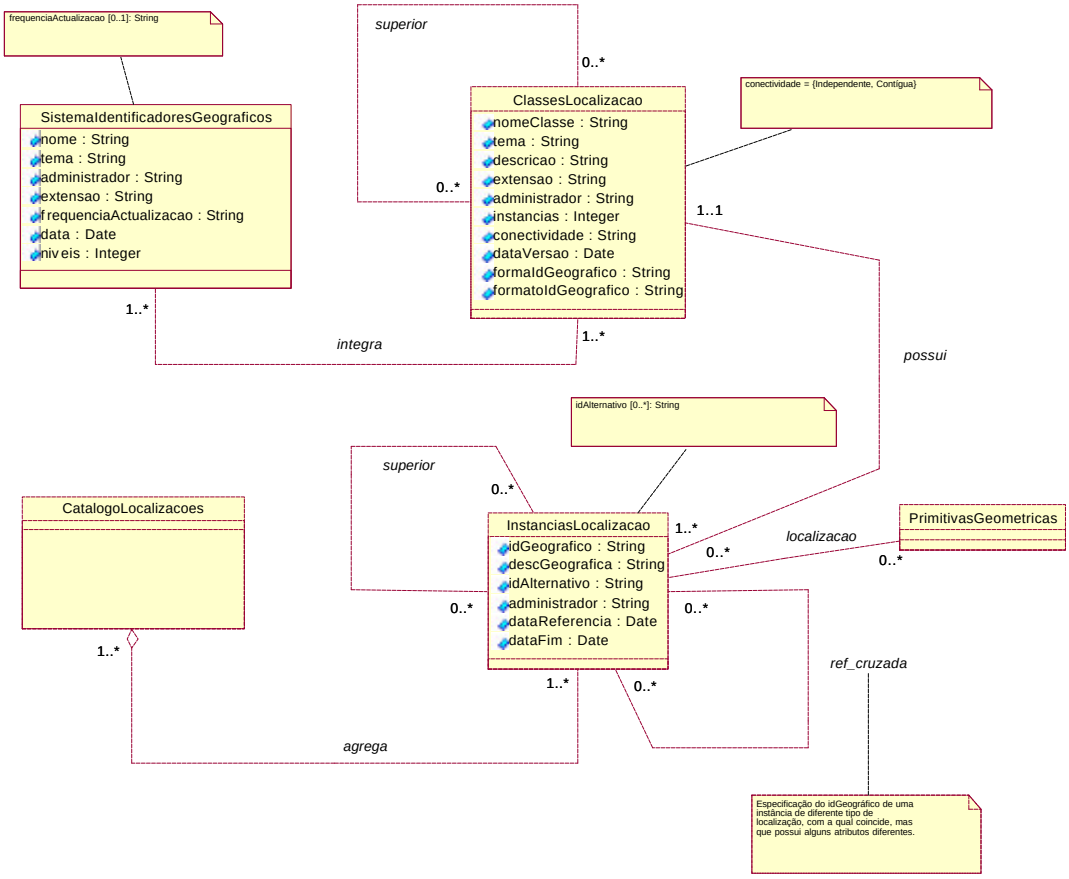


Figura 2.17: Esquema de Identi...cadores Geográ...cos

Capítulo 3

O raciocínio espacial qualitativo

Neste capítulo abordam-se os princípios nos quais se baseia o raciocínio espacial qualitativo, salientando os tipos de relações espaciais existentes e as abordagens ao raciocínio, homogénea, heterogénea e integrada, mais utilizadas. No caso da abordagem integrada, e dada a sua utilização neste trabalho, são analisados com detalhe dois sistemas de raciocínio espacial integrado, um que integra relações espaciais do tipo direcção e distância [Hong, 1994] e outro que integra relações espaciais do tipo direcção e topologia [Sharma, 1996]. O objectivo é veri…car os princípios que ditaram a construção destes dois sistemas, avaliando a viabilidade de integração dos mesmos, por forma a obter um sistema de raciocínio que integre a direcção, a distância e a topologia.

Este capítulo é organizado da seguinte forma. A secção 3.1 sistematiza alguns dos conceitos associados ao raciocínio espacial qualitativo. Na secção seguinte são apresentados mecanismos de representação de conhecimento espacial qualitativo, descrevendo o adoptado ao longo deste documento. A secção 3.3 apresenta os princípios que regem o raciocínio temporal qualitativo, uma vez que os mesmos foram amplamente utilizados no domínio espacial. Na secção 3.4 descrevem-se as relações espaciais do tipo direcção, distância e topologia, e ainda os princípios que ditam a construção das respectivas tabelas de composição. A secção 3.5 apresenta dois sistemas de raciocínio integrado e evidencia como os mesmos são utilizados na construção de um sistema de inferências, que utiliza a direcção, a distância e a topologia dos objectos no processo de raciocínio. A secção 3.6 descreve como características dos objectos, neste caso a dimensão, podem ser manipuladas por mecanismos qualitativos de raciocínio.

3.1 Princípios

O raciocínio espacial é o processo pelo qual a informação acerca dos objectos no espaço é utilizada para conhecer relacionamentos implícitos existentes entre os mesmos. Estes relacionamentos podem ser representados e manipulados quantitativamente, requerendo uma descrição completa da geometria dos objectos e seus respectivos relacionamentos, ou qualitativamente, através da utilização de identi…cadores qualitativos, os quais suprimem a necessidade de especi…cação geométrica das entidades envolvidas. Estas duas formas de representação de conhecimento espacial, e seus respectivos processos de raciocínio, retratam duas formas complementares de conceptualização e manipulação do espaço, devendo a escolha de uma ou outra ser ditada pelos objectivos que as mesmas visam servir [Sharma, 1996].

O raciocínio espacial qualitativo tem assumido um papel de destaque na área dos SIG, visão por computador, robótica, etc., uma vez que oferece mecanismos complementares de inferência, proporcionando ao utilizador o acesso a informação espacial desconhecida, e que de outra forma não estaria imediatamente disponível [Egenhofer, 1994b].

As aproximações qualitativas não manipulam coordenadas, podendo desta forma manusear dados imprecisos [Freksa, 1992]. Tornam-se particularmente úteis quando a precisão quantitativa não é necessária ou desejada, ou quando informação precisa, dados quantitativos, não estão disponíveis [Cohn, 1995] [Frank, 1996] [Papadias e Sellis, 1994].

O raciocínio espacial qualitativo tem sido proposto como um mecanismo complementar de inferência[1] de relações espaciais desconhecidas [Abdelmoty e El-Geresy, 1995]. É baseado na manipulação de um conjunto restrito de símbolos, como Norte, Sul, próximo, etc., para os quais tabelas de composição facilitam o raciocínio, permitindo a inferência de novas relações espaciais (abordagem particularmente útil em domínios de aplicação caracterizados por manipular grandes volumes de informação incompleta ou imprecisa).

A utilização, neste trabalho, destes princípios foi motivada não só pela constatação de que a referenciação geográ...ca da informação através do uso de identi...cadores qualitativos, como moradas ou códigos postais, ocorre na maioria das BD organizacionais, como também pelo facto de ser desejável que o resultado do processo de descoberta de conhecimento seja expresso em termos qualitativos. Refere-se ainda que, apesar de existir alguma perda de precisão no raciocínio qualitativo, este simpli...ca o processo de descoberta de conhecimento, evitando o desenvolvimento de novos algoritmos de DM e permitindo o raciocínio com informação geográ...ca incompleta ou imprecisa.

As relações espaciais [Freksa, 1992] [Papadias e Sellis, 1994] [Sharma, 1996] têm sido classi...cadas em vários tipos, entre os quais:

- relações de direcção, que descrevem orientações no espaço;
- relações de distância, que descrevem proximidade no espaço;
- relações topológicas, que descrevem vizinhança e sobreposição; e ainda,
- relações ordinais, que descrevem inclusão (sendo portanto um subconjunto das relações topológicas).

Uma operação essencial na manipulação destes relacionamentos é a composição[2] de factos, que permite inferir a relação espacial existente entre as entidades A e C, dados os relacionamentos existentes entre A e B, e B e C. Sempre que estas tabelas utilizem no processo de raciocínio apenas um tipo de relação espacial, direcção, distância ou topologia, representam raciocínio espacial homogéneo [Sharma, 1996]. Esta abordagem é caracterizada por gerar inferências pouco especí...cas, nas quais o resultado é normalmente constituído por um conjunto de inferências possíveis [Hernández, 1994] (Exemplo: $A\ Norte\ B; B\ Este\ C \Rightarrow A\ Norte\_Nordeste\_Este\ C$).

[1] Termo utilizado para representar o processo de raciocínio a partir do qual se extraiem conclusões a partir de premissas conhecidas (sendo uma premissa um facto ou uma regra) ou ainda para referir o resultado do referido processo [CT113, 1999].

[2] A composição de factos é neste trabalho representada pelo símbolo ";".

A abordagem heterogénea é caracterizada pela conjugação de dois tipos de relações espaciais. Esta abordagem permite construir tabelas de composição nas quais a relação espacial inferida pertence ao conjunto das relações iniciais (Exemplo: A contido_em B;B Noroeste C) A Noroeste C). A abordagem mista é semelhante à heterogénea, mas o tipo de relação espacial dos dois factos conhecidos tem de ser o mesmo, permitindo inferir outro tipo de relação espacial (Exemplo: A Norte B;B Noroeste C) A deslocado C) [Sharma, 1996].

A abordagem integrada constitui a aproximação mais precisa, uma vez que conjuga a cada instante dois tipos de relação espacial, tirando partido do relacionamento intrínseco existente entre os mesmos. As características de cada uma das relações são integradas e analisadas como um todo, gerando inferências mais exactas do que as abordagens anteriores (Exemplo: A Norte,próximo B;B Este,distante C) A Noroeste,distante C) [Sharma, 1996].

A abordagem homogénea é descrita na secção 3.4, onde são apresentados os três tipos de relações mais utilizadas no raciocínio espacial qualitativo. Na secção 3.5 são apresentados os princípios que regem a construção de sistemas de raciocínio espacial qualitativo integrado, dando particular ênfase à construção do sistema utilizado neste trabalho e que permite integrar a direcção, a distância e a topologia no processo de raciocínio.

3.2 Representação qualitativa de conhecimento espacial

A representação de conhecimento consiste no processo de codi…cação de conhecimento[3], com vista ao seu armazenamento numa base de conhecimento[4] [CT113, 1999]. Tal signi…ca que a representação de conhecimento acerca de um domínio, ao invés da representação do domínio propriamente dito, está associada ao facto do mundo ser acessível em termos formais através de conhecimento.

Linguagens formais como lógica de predicados permitem a abstracção das propriedades inerentes ao espaço. A especi…cação do domínio espacial através desta abstracção permite restringir a linguagem utilizada, por forma a esta ser semanticamente consistente com o domínio espacial. As leis do espaço estão implícitas nas representações adoptadas, permitindo focar apenas os aspectos considerados mais relevantes.

As representações qualitativas são caracterizadas por considerarem, apenas, o número de símbolos necessários num dado contexto e permitir, no caso do domínio espacial, retratar a con…guração existente entre entidades espaciais, preservando a sua localização no espaço, sem incorporar informação como a forma, o tamanho, a textura ou a cor dos objectos [Papadias e Sellis, 1994].

A representação qualitativa adoptada deverá permitir a imposição de restrições especí…cas do domínio espacial, as quais estarão implícitas no sistema de inferências criado para lidar com este conhecimento do espaço [Freksa, 1991]. A adopção de representações especí…cas para o raciocínio espacial tem como vantagem suprimir a necessidade de modelação de determinadas restrições espaciais. A abstracção considerada deverá permitir superar as restrições impostas

[3] Sendo este de…nido como uma colecção de factos, acontecimentos ou regras, organizados por forma a facilitar o seu uso sistemático [CT113, 1999].

[4] Uma base de conhecimento é uma base de dados que contém regras de inferência e informação referente à experiência e perícia humana num domínio especí…co. Nos sistemas evolutivos, a base de conhecimento contém também a informação resultante da resolução de problemas anteriores [CT113, 1999].

pelo espaço, sem a necessidade de as veri…car a cada instante. Por exemplo, a localização e disposição dos objectos, dadas por relações de vizinhança, constituem características importantes, e como tal devem ser consideradas na resolução de tarefas espaciais, uma vez que a movimentação no espaço apenas é possível para localizações vizinhas. Estas restrições não têm de ser veri…cadas explicitamente em cada tarefa, se o sistema de inferências for construído baseado nas mesmas. O sistema de raciocínio deve então ser construído fazendo uso do espaço e das suas propriedades intrínsecas.

A representação de conhecimento espacial deve preservar propriedades conceptuais do domínio, tais como:

- propriedades universais que não dependem de determinada situação, e que são denominadas de restrições de identidade:
  - cada objecto existe apenas uma vez;
  - uma localização coincide no máximo com um objecto;
- propriedades inerentes ao espaço físico e que estão associadas a restrições topológicas:
  - movimentações no espaço apenas são possíveis entre localizações vizinhas;
- propriedades inerentes à vizinhança das relações espaciais, e que permitem de…nir restrições conceptuais:
  - as quais impõem restrições à localização dos objectos após alterações do espaço. Um objecto apenas se pode movimentar para uma relação vizinha. Duas relações espaciais são vizinhas se elas podem ser derivadas uma da outra, sem necessidade de recorrer a uma terceira relação da mesma dimensão (No caso das relações de direcção, Norte e Nordeste são consideradas relações vizinhas).

Uma dada representação, para um determinado domínio de aplicação, é adequada se permite explicitar as restrições inerentes ao problema em causa [Hernández, 1991] [Hernández, 1994] e se facilita a sua resolução.

Diversos sistemas de representação podem ser comparados veri…cando o sistema de referência utilizado (Cartesiano, polar, ...), as primitivas geográ…cas que representam (pontos, objectos com extensão, ...), os aspectos representados (direcção, topologia, ...), a sua granularidade, a capacidade de representar informação incompleta ou imprecisa, a ‡exibilidade de integração de novo conhecimento, etc. [Freksa e Rohrig, 1993].

Os formalismos utilizados são, assim, independentes daquilo que é representado, sendo a selecção e estruturação do conhecimento do domínio independente da sua representação. Como tal, refere-se que:

- o formalismo utilizado na representação do conhecimento espacial utilizado neste projecto, passa pela utilização de predicados do tipo `A Norte B`, em que `A` e `B` representam entidades geográ…cas, e `Norte` representa a relação espacial existente entre as mesmas. Formalmente: `Objecto_primário [relação espacial] Objecto_de_referência`, podendo a

relação espacial ser do tipo direcção, distância ou topologia. Os conjuntos de indicadores qualitativos adoptados, para a caracterização de cada um destes tipos de relação espacial, são apresentados nas subsecções 3.4.1 a 3.4.3;

- a inferência de conhecimento é efectuada recorrendo a tabelas de composição especí…cas. Estas tabelas são construídas atendendo a determinados princípios, estando a sua utilização condicionada pela dependência existente entre as mesmas e o seu contexto de aplicação. Por exemplo, a adopção de determinado número de identi…cadores qualitativos, assim como os intervalos de validade associados aos mesmos, condiciona os resultados que podem ser obtidos.

3.3 O raciocínio temporal qualitativo

O raciocínio temporal qualitativo é aqui apresentado por ter motivado a extrapolação dos seus princípios ao domínio espacial. A contribuição dada por Allen [Allen, 1983], na de…nição de uma tabela de composição para a inferência de relações temporais qualitativas, permitiu a construção de tabelas de composição para o domínio espacial [Sharma, 1996], baseadas nas propriedades e primitivas de…nidas para o domínio temporal. Outras abordagens ao raciocínio temporal qualitativo podem ser encontradas em [Beek, 1992] [Frank, 1994] [Wijsen, 1998].

Através da observação de que o conhecimento temporal é frequentemente obtido através de comparações relativas, Allen [Allen, 1983] desenvolveu uma lógica temporal baseada em intervalos. Estes retratam acontecimentos com determinada duração, a qual é representada por um intervalo unidimensional. A veri…cação do conjunto das relações que podem existir entre os pontos iniciais e …nais de dois intervalos, permitiu identi…car um conjunto de treze relações (primitivas), que caracterizam todos os relacionamentos que podem existir entre dois intervalos temporais.

Esta abordagem assenta no princípio do conhecimento temporal relativo, suprimindo a utilização de datas explícitas e adoptando representações temporais sob a forma de intervalos. As primitivas temporais, utilizadas para descrever os acontecimentos, podem posteriormente ser manipuladas (compostas), permitindo inferir os intervalos temporais em que determinados factos ocorreram. Esta abordagem é ainda complementada com a possibilidade de de…nição de hierarquias temporais, as quais permitem a imposição de restrições à sequência cronológica em que determinados factos podem ocorrer. Esta catalogação de eventos permite modelar temporalmente uma sequência de acontecimentos, que utilizada conjuntamente com a tabela de composição para as relações temporais, facilita a inferência de datas desconhecidas sob a forma de intervalos.

Estes princípios revelam-se de particular importância na determinação de datas desconhecidas, necessárias, por exemplo, no processo de descoberta de conhecimento. Apesar de inferir datas aproximadas, uma vez que estão associadas a um intervalo de valores, a sua determinação é relevante em domínios de aplicação como a demogra…a, onde dada a "antiguidade" de algumas informações, muitas das datas associadas aos acontecimentos são desconhecidas. Esta aproximação temporal qualitativa é particularmente útil em domínios de aplicação com informação temporal desconhecida, onde é possível, a partir de outros factos conhecidos e da hierarquia temporal de…nida para o domínio, inferir o intervalo temporal em que determinado acontecimento

deverá ter ocorrido.

As primitivas temporais[5] de…nidas por Allen [Allen, 1983] são: anterior (before, b), igual (equal, e), durante (during, d), adjacente (meets, m), sobrepõe (overlaps, o), inicia (starts, s) e finaliza (…nishes, f), existindo as inversas posterior (after, a), contém (contains, di), adjacente_a (met-by, mb), sobreposto (overlapped-by, ob), iniciado (started-by, sb) e finalizado (…nished-by, fb) (Figura 3.1).

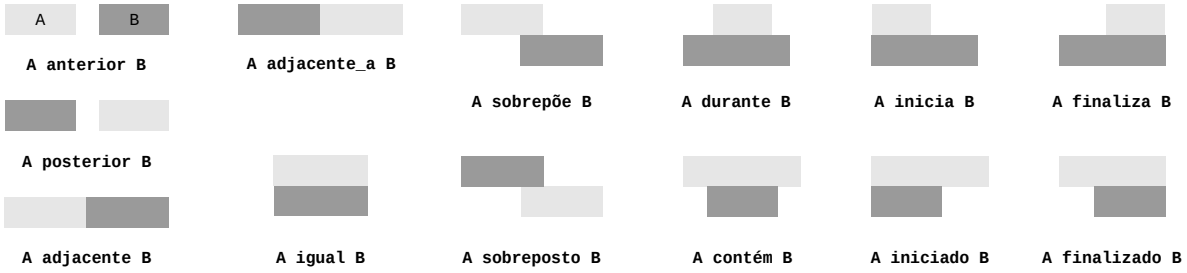


Figura 3.1: Primitivas temporais baseadas em intervalos

A tabela de composição (Tabela 3.1) para as inferências temporais foi obtida aplicando propriedades transitivas aos intervalos, e ainda, veri…cando detalhadamente a consistência do conjunto de relações temporais possíveis [Allen, 1983]. A composição de intervalos permite determinar a relação, ou conjunto de relações, que existe(m) entre os intervalos A e C, baseado no conhecimento das relações temporais existentes entre os intervalos A e B, e B e C. Como poderá ser constatado, ainda neste capítulo, esta tabela é utilizada por Sharma [Sharma, 1996] no domínio espacial, na integração de relações do tipo direcção e topologia (subsecção 3.5.2).

3.4 Tipos de relações espaciais

3.4.1 Direcção

Relações espaciais do tipo direcção de…nem orientações no espaço e são determinadas conjugando três entidades: dois objectos e um sistema de referência [Frank, 1992] [Frank, 1996] [Freksa, 1992] [Papadias e Sellis, 1994] [Sharma, 1996]. A união dos dois objectos determina uma linha recta, que permite que a direcção entre estes seja calculada veri…cando a orientação existente entre a recta obtida e o sistema de referência utilizado. No facto A Norte B, A constitui o objecto primário e B o objecto de referência. Norte integra as primitivas utilizadas pelo sistema de referência para as direcções cardinais. Apesar de neste trabalho ser utilizado um sistema de referência cardinal, refere-se que os sistemas de referência podem ser estabelecidos através de [Hernández, 1994]:

[5] Estas primitivas são identi…cadas, na respectiva tabela de composição, por abreviaturas derivadas das suas denominações originais em inglês.

	b	a	d	di	o	ob	m	mb	s	sb	f	fb
b	b	?	b,o,m,d,s	b	b	b,o,m,d,s	b	b,o,m,ds	b	b	b,o,m,d,s	b
a	?	a	a,ob,mb,d,f	a	a,ob,mb,d,f	a	a,ob,mb,d,f	a	a,ob,mb,d,f	a	a	a
d	b	a	d	?	b,o,m,d,s	a,ob,mb,d,f	b	a	d	a,ob,mbd,f	d	b,o,m,d,s
di	b,o,m,di,fb	a,ob,mb,di,sb	o,ob,d,s,f,di,sb,fb,e	di	o,di,fb	ob,di,sb	o,di,fb	ob,di,sb	di,fb,o	di	di,sb,ob	di
o	b	a,ob,mb,di,sb	o,d,s	b,o,m,di,fb	b,o,m	o,ob,d,s,f,di,sb,fb,e	b	ob,di,sb	o	di,fb,o	d,s,o	b,o,m
ob	b,o,m,di,fb	a	ob,d,f	a,ob,mb,disb	o,ob,d,s,f,di,sb,fb,e	a,ob,mb	o,di,fb	a	ob,d,f	ob,a,mb	ob	ob,disb
m	b	a,ob,mb,di,sb	o,d,s	b	b	o,d,s	b	f,fb,e	m	m	d,s,o	b
mb	b,o,m,di,fb	a	ob,d,f	a	ob,d,f	a	s,sb,e	a	d,f,ob	a	mb	mb
s	b	a	d	b,o,m,di,fb	b,o,m	ob,d,f	b	mb	s	s,sb,e	d	b,m,o
sb	b,o,m,di,fb	a	ob,d,f	di	o,di,fb	ob	o,di,fb	mb	s,sb,e	sb	ob	di
f	b	a	d	a,ob,mb,disb	o,d,s	a,ob,mb	m	a	d	a,ob,mb	f	f,fb,e
fb	b	a,ob,mb,di,sb	o,d,s	di	o	ob,di,sb	m	sb,ob,di	o	di	f,fb,e	fb

Tabela 3.1: Tabela de Composição para as Relações Temporais
Adaptado de: [Allen, 1983] p. 836

- uma orientação intrínseca, na qual o sistema de referência é estabelecido em relação ao objecto de referência;
- uma orientação extrínseca, na qual factores externos impõem a orientação do objecto de referência;
- um ponto de vista, caracterizado normalmente por constituir a localização de um observador. O sistema de referência é construído através do traçado de uma linha recta entre o objecto de referência e o ponto de vista.

No sistema de referência cardinal uma direcção é de…nida recorrendo a valores numéricos (0, 45, ...), que especi…cam graus no intervalo [0, 360º), ou pela utilização de identi…cadores qualitativos como Norte, Sul, etc., aos quais se associa uma dada região de aceitação (intervalo de validade quantitativo). Estas regiões de aceitação podem ser de…nidas recorrendo a um sistema de projecções ou ao sistema triangular (Figura 3.2).

Em ambos os sistemas existe uma divisão do plano num número limitado de áreas, às quais se associa explicitamente uma região de aceitação. Em termos quantitativos, a direcção entre dois pontos é de…nida recorrendo à geometria euclidiana, através da manipulação das coordenadas cartesianas dos referidos pontos. A direcção resulta da veri…cação da orientação do vector resultante da união dos dois pontos [Hong, 1994].

Sempre que os objectos em análise apresentarem extensão, vários factores podem in‡uenciar a orientação existente entre os mesmos. Entre eles, características como o tamanho ou

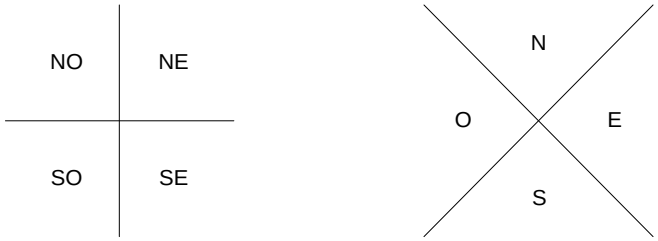


Figura 3.2: Direcções segundo o sistema de projecções e modelo triangular

forma das regiões. Nestes casos, a de…nição da direcção existente passa pela de…nição de áreas de aceitação, que representam o conjunto de valores para os quais, uma dada orientação é tida como uma descrição válida da direcção relativa existente entre os dois objectos [Sharma, 1996].

A de…nição das áreas de aceitação, para a determinação da orientação existente entre duas regiões, passa pela utilização do sistema de projecções ou do sistema triangular, apresentados anteriormente. Na utilização de projecções [Theodoridis et al., 1996], o sistema é constituído por nove áreas de aceitação, oito para as direcções e uma zona neutra utilizada para indicar a posição do objecto de referência. As projecções são construídas estendendo as linhas que delimitam o MBR (Minimum Bounding Rectangle) do referido objecto (Figura 3.3).

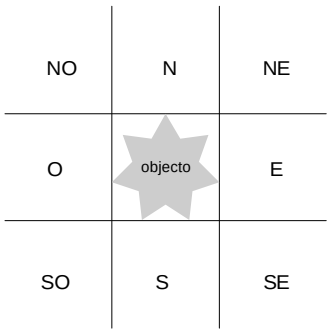


Figura 3.3: Sistema de projecções para a determinação de direcções entre objectos com extensão

A de…nição de direcções para objectos com extensão, utilizando o sistema triangular, passa pela veri…cação da orientação existente entre os centróides das referidas regiões. Esta aproximação permite transferir para o caso particular dos objectos com extensão, todos os princípios desenvolvidos para o manuseamento de objectos do tipo ponto ([Hernández, 1994] p.47). Uma outra vantagem, na utilização deste modelo, é o facto do mesmo permitir a de…nição de mais do que 8 áreas de aceitação (Figura 3.4), por exemplo 16, se a análise pretendida assim o determinar. Dado ser este o sistema utilizado neste trabalho, no qual as direcções e distâncias existentes entre regiões são calculadas veri…cando as posições dos respectivos centróides, as descrições e apresentações subsequentes estão associadas ao modelo triangular.

Na construção das tabelas de composição, para a inferência de relações espaciais desconhecidas, Hernández [Hernández, 1994] e Frank [Frank, 1992][Frank, 1996] adoptam duas estratégias

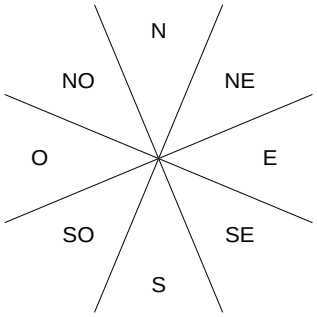


Figura 3.4: Modelo triangular com 8 regiões de aceitação

distintas. No primeiro caso, a regra que dita o resultado da inferência é obtida seguindo um princípio muito simples: conhecidos os factos A dir_x B e B dir_y C, o resultado da composição (A dir_z C) é conseguido veri…cando todas as áreas de aceitação que se encontram entre as direcções dir_x e dir_y (inclusive), seguindo o caminho mais curto existente entre as mesmas. Num sistema de referenciação que utilize 8 identi…cadores para a direcção, o número máximo de relações obtidas numa inferência é de quatro, excepto para o caso de direcções opostas, no qual as oito orientações são possíveis [Hernández, 1994]. A tabela de composição assim obtida é pouco especí…ca, uma vez que existe uma elevada probabilidade da inferência obtida ser constituída por um conjunto de relações alternativas.

Frank [Frank, 1992][Frank, 1996] utilizou outra abordagem, mais precisa, envolvendo o desenvolvimento de um método algébrico para a análise das várias direcções e suas propriedades. É baseado na manipulação caminhos, os quais são posteriormente transformados em direcções. Um caminho é representado por uma aresta que interliga um ponto origem P_1 a um ponto destino P_2. Entre as várias propriedades algébricas, que podem ser aplicadas a estes caminhos, encontra-se a associatividade, e ainda operações como a inversa ou a adição. A inversa do caminho de P_1 a P_2 é o caminho de P_2 para P_1. A composição combina dois caminhos, de P_1 para P_2 e de P_2 para P_3, dando como resultado o caminho de P_1 para P_3. A identidade é utilizada para caracterizar caminhos de um ponto para si próprio. Esta abordagem permite gerar como resultado, um único caminho bem de…nido (respostas unívocas).

A utilização de conceitos algébricos evita trabalhar com a direcção quantitativa entre os pontos, permitindo a construção de regras através da manipulação dos próprios símbolos que representam as direcções (N, S, NE, ...). O conjunto de operações e axiomas utilizados para determinar o resultado das composições pretendidas [Frank, 1996], são:

- Direcções Cardinais. Uma direcção é uma função binária que relaciona dois pontos no espaço (P_1,P_2) recorrendo a uma direcção simbólica D. O número de símbolos utilizado, para o sistema triangular, 4, 8, 16 ..., depende do grau de detalhe pretendido na análise. A um sistema com 8 símbolos D_8={N,NE,E,SE,S,SO,O,NO}, Frank [Frank, 1992][Frank, 1996] adiciona o símbolo I para caracterizar o caso particular do ponto origem e do ponto destino ser o mesmo. Este símbolo, que representa a identidade, simpli…ca as regras de inferência e permite retratar os casos em que a inferência não é possível (por exemplo, no caso de dois pontos estarem tão próximos, que é impossível determinar a orientação existente entre os mesmos).

- Inversa. Se a direcção é dada pela orientação de um segmento que une dois pontos P_1 e P_2, então a direcção entre P_2 e P_1 pode ser inferida através da operação inversa, de…nida como

  $$inv(dir(P_1,P_2)) = dir(P_2,P_1), \text{ e}$$
  $$inv(inv(dir(P_1,P_2))) = dir(P_1,P_2).$$

- Composição. Esta operação combina a direcção de dois segmentos de linha contíguos, de tal forma que o ponto …nal do primeiro segmento coincide com o ponto inicial do segundo segmento, $dir_1;dir_2 = dir_3 \; ! \; dir(P_1,P_2);dir(P_2,P_3) = dir(P_1,P_3)$. A composição apresenta as seguintes propriedades:

  - Associativa. A composição de mais do que duas direcções deve ser independente da ordem pela qual as mesmas são combinadas: $dir_1;(dir_2;dir_3) = (dir_1;dir_2);dir_3 = dir_1;dir_2;dir_3$.
  - Identidade. Representa a direcção de um ponto para ele próprio: $dir(P_1,P_1) = I$.
  - Inversa. A inversa de uma operação binária tem de ser de…nida de forma a que um valor, combinado com a sua inversa, dê como resultado a identidade: $inv(dir);dir = I$. A inversa é também utilizada para veri…car o resultado da composição de duas direcções, através da composição da inversa de cada uma das mesmas: $inv(dir_1;dir_2) = inv(dir_2);inv(dir_1)$.
  - Igualdade (Idempotent). A composição de dois segmentos com a mesma orientação deve dar como resultado a direcção composta: $dir;dir = dir$.

- Raciocínio Euclidianamente exacto. A veri…cação das inferências obtidas, comparando-as com os valores reais obtidos por métodos quantitativos, nomeadamente recorrendo à soma de vectores, é efectuada com o objectivo de averiguar se as regras qualitativas geradas são euclidianamente exactas ou aproximadas. Uma regra qualitativa é apelidada de exacta, se o resultado da sua aplicação é igual ao obtido pela conversão dos factos para valores quantitativos, e sua posterior manipulação por funções apropriadas. Se tal não acontecer, os resultados são apelidados de aproximados. No caso de exactos:

  $$dir(P_1,P_2);dir(P_2,P_3) = dir((P_1,P_2) + (P_2,P_3)).$$

Seguindo estes princípios, a Tabela 3.2 apresenta o conjunto de regras que permitem a inferência de direcções cardinais, para o modelo triangular. Nesta tabela, as relações representadas por letras minúsculas caracterizam inferências aproximadas, enquanto que I denota, como já referido, a identidade.

3.4.2 Distância

Distâncias representam valores quantitativos determinados por medição ou calculados a partir das coordenadas dos respectivos pontos. A de…nição quantitativa mais familiar de distância é

	N	NE	E	SE	S	SO	O	NO	I
N	N	n	ne	I	I	I	no	NO	N
NE	n	ne	ne	e	I	I	I	N	NE
E	ne	ne	E	e	se	I	I	I	E
SE	I	e	e	SE	se	s	I	I	SE
S	I	I	se	se	S	s	so	o	S
SO	I	I	I	s	s	SO	so	o	SO
O	no	I	I	I	so	so	O	o	O
NO	n	n	I	I	I	o	o	NO	NO
I	N	NE	E	SE	S	SO	O	NO	I

Tabela 3.2: Tabela de composição para a direcção
Adaptado de: [Frank, 1996] p. 278

dada pela distância Euclidiana, e consiste na determinação do caminho mais curto (em linha recta) entre dois pontos. A distância Euclidiana, no espaço bi-dimensional, pode ser de…nida através de $distância(A;B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$, onde $(x_A;y_A)$ e $(x_B;y_B)$ representam as coordenadas cartesianas dos pontos A e B, respectivamente ([Hong, 1994] p. 29).

Algumas di…culdades são encontradas quando se pretende determinar a distância existente entre objectos com extensão, uma vez que o tamanho dos mesmos in‡uencia o resultado. Para ultrapassar estas restrições, Hong [Hong, 1994] e Sharma [Sharma, 1996] sugerem que a distância entre duas regiões seja de…nida calculando:

- a distância existente entre os centros geográ…cos dos objectos (centróides);
- a distância média existente os dois objectos;
- a distância mais curta existente entre os objectos;
- a distância mais longa existente entre os objectos.

Cada uma destas alternativas pode criar con‡itos, principalmente se o domínio de aplicação em causa permitir a existência de regiões dentro de regiões, conduzindo à herança, pelas regiões contidas, da distância existente entre as regiões que as contêm [Sharma, 1996]. Não sendo este o cenário veri…cado neste trabalho, e atendendo ao facto de se pretender transformar os valores quantitativos obtidos em valores qualitativos, optou-se pela primeira alternativa, uma vez que a utilização de um SIG[6], na atribuição automática dos centróides de cada região (para o caso dos centróides não terem ainda sido de…nidos), permitirá obter estes valores automaticamente.

Os valores quantitativos obtidos são posteriormente convertidos em identi…cadores qualitativos, como próximo ou distante, permitindo a utilização de princípios qualitativos na construção das respectivas tabelas de composição. A cada um dos símbolos qualitativos corresponde um

[6] Os SIG são frequentemente utilizados na atribuição do centro geográ…co ou centróide das regiões, o que se reveste de particular importância nos casos em que é necessário calcular a distância entre duas regiões. Em situações muito particulares, normalmente associadas a polígonos de forma muito enrolada (convoluted), a atribuição automática pode localizar o centróide fora dos limites da região, solicitando o seu reposicionamento [Gatrell, 1991].

determinado intervalo de validade [Gahegan, 1995], o qual caracteriza o conjunto das distâncias quantitativas representado pelo mesmo. A de…nição destes intervalos deverá permitir a comparação qualitativa dos respectivos identi…cadores. Outra particularidade é que o tamanho dos intervalos deverá manter-se, ou mesmo aumentar, a medida que representam identi…cadores qualitativos sucessivos. A Figura 3.5 apresenta um sistema de distâncias que utiliza quatro símbolos qualitativos, mp (muito próximo), p (próximo), d (distante) e md (muito distante), na representação da distância existente entre dois objectos.

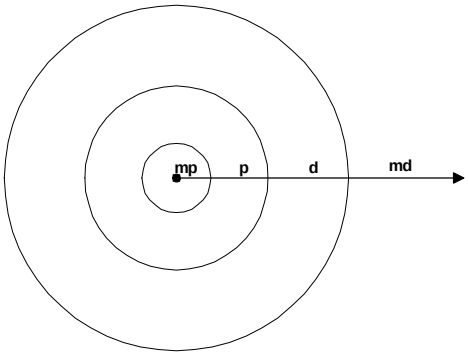


Figura 3.5: Distâncias qualitativas

A de…nição qualitativa de distância entre dois objectos é dependente do contexto em que a mesma é utilizada. Não depende exclusivamente da posição absoluta dos dois objectos (e da distância métrica existente entre os mesmos), mas também dos seus tamanhos, forma e do sistema de referência utilizado [Hernández et al., 1995]. As aproximações quantitativas suprimem o contexto, reduzindo toda a informação a uma escala métrica.

Representando P um conjunto de pontos e $D = \{dist_0, \ldots, dist_n\}$ um conjunto de n + 1 intervalos para a representação qualitativa das distâncias, a distância entre um ponto P_1 e um ponto P_2 é dada pela função $dist: P \times P \rightarrow D$, que identi…ca a distância existente entre P_1 e P_2.

Os símbolos utilizados, e como já referido anteriormente, devem ser ordenados tal que $dist_0 < dist_1 < \ldots < dist_n$. Nesta sequência, o anterior (ant) e posterior (pos) de uma dada distância pode ser calculado por $pos(dist_i) = dist_{i+1}$, para $i < n$ e $pos(dist_n) = dist_n$. Da mesma forma, o anterior é de…nido por $ant(dist_i) = dist_{i-1}$, para $i > 0$ e $ant(dist_0) = dist_0$.

A de…nição dos intervalos de validade quantitativos, associados a cada um dos identi…cadores qualitativos utilizados, passa pela de…nição de uma região de aceitação que circunda o ponto de referência P_1, de tal forma que todos os pontos de P_2 inseridos nessa região, sejam identi…cados pela mesma distância $dist_i$.

A construção das regras de inferência passa pela veri…cação de três axiomas [Sharma, 1996], os quais conceptualizam as métricas espaciais para a manipulação da distância, através das seguintes propriedades:

- Re‡exiva, a distância de um ponto para ele próprio é igual a zero:

	mp	**p**	**d**	**md**
mp	mp, p	mp, p, d	p, d, md	d, md
p	mp, p, d	mp, p, d	mp, p, d, md	d, md
d	p, d, md	mp, p, d, md	mp, p, d, md	mp, p, d, md
md	d, md	d, md	mp, p, d, md	mp, p, d, md

Tabela 3.3: Tabela de composição para a distância
Adaptado de: [Sharma, 1996] p. 61

$dist(P_1,P_1) = 0.$

- Simétrica, a distância do ponto P_1 ao ponto P_2 é igual à distância de P_2 para P_1:

  $dist(P_1,P_2) = dist(P_2,P_1).$

- Igualdade triangular, a distância do ponto P_1 ao ponto P_3 é inferior ou igual à soma das distâncias de P_1 a P_2 e de P_2 a P_3:

  $dist(P_1,P_2) + dist(P_2,P_3) \geq dist(P_1,P_3).$

Estas propriedades apenas são verificadas num espaço isotrópico. Nestas superfícies, o esforço de movimentação é o mesmo em todas as direcções [Hernández et al., 1995]. A utilização destes axiomas como base para a definição de distâncias qualitativas, e construção da respectiva tabela de composição, conduz à constatação de que a propriedade da igualdade triangular requer uma definição cautelosa da adição de identificadores qualitativos. O objectivo é o de prevenir inconsistências. Por exemplo, os factos $d(A,C)$, $p(A,B)$ e $p(B,C)$ só podem ser simultaneamente verdadeiros se $p + p = d$ [Sharma, 1996].

Atendendo a estes pressupostos, foi possível definir uma tabela de composição (Tabela 3.3) para a inferência de relações espaciais do tipo distância, na qual a direcção existente entre os objectos referenciados não é considerada. Por este motivo, o resultado obtido para uma dada inferência é na maior parte dos casos constituído por mais do que um identificador (resultados pouco específicos).

3.4.3 Topologia

Relações espaciais do tipo topológico permitem definir como dois objectos se relacionam no espaço, verificando a existência de sobreposições entre os mesmos. Estas relações permitem definir vizinhança, adjacência e sobreposição entre objectos, sendo estas relações caracterizadas por permanecerem invariantes a transformações do espaço, como rotação ou mudança de escala [Papadias e Theodoridis, 1997].

Ignorando a orientação que pode existir entre dois objectos, as suas projecções podem ser associadas através de algumas relações, como [Hernández, 1994] [Sharma, 1996]:

- estão distantes um do outro;
- estão próximos, mas não se tocam;
- tocam-se;
- sobrepõem-se;
- um deles está contido no outro.

Existem 8 relações topológicas (Figura 3.6) na caracterização de duas regiões planares[7]: deslocado (disjoint), contém (contains), dentro (inside), igual (equal), adjacente (meet), cobre (covers), coberto (covered by) e sobrepõe (overlap) [Egenhofer e Sharma, 1993].

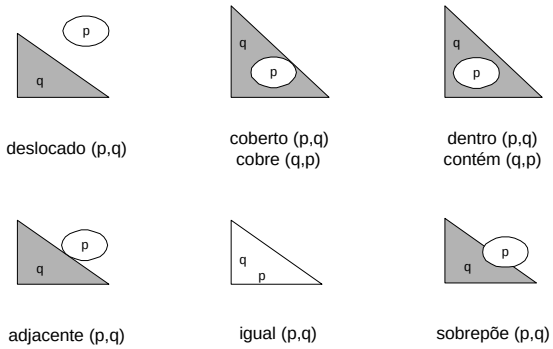


Figura 3.6: Relações topológicas

Relações topológicas constituem um conceito qualitativo, independente de qualquer medida quantitativa. A procura de um modelo formal capaz de representar o conhecimento espacial topológico, e ainda a necessidade do mesmo permitir raciocinar qualitativamente, conduziu Egenhofer [Egenhofer, 1994b] a basear todo o sistema de raciocínio na teoria de conjuntos, no qual as relações existentes entre dois objectos são derivadas a partir das propriedades existentes entre os seus subconjuntos.

A representação adoptada permite a de…nição da relação topológica existente entre dois objectos, através da veri…cação da sobreposição existente entre o interior, o limite e o exterior dos mesmos [Egenhofer e Sharma, 1993] [Egenhofer, 1994b][8]. Um caso particular é veri…cado na caracterização topológica [Egenhofer e Sharma, 1993] [Hernández, 1994] das relações existentes

[7] Refere-se que podem ser de…nidas 18 relações topológicas na caracterização de regiões com buracos, 33 relações topológicas na caracterização de objectos do tipo linha e 19 relações topológicas na caracterização de linhas e regiões com buracos [Egenhofer, 1994b].

[8] Refere-se que para esta representação, o ISO TC 211, no seu documento de trabalho N 818 [ISO/TC-211, 1999c], de…ne diversas funções para a veri…cação da relação topológica existente entre dois objectos, através da veri…cação da matriz de intersecções que caracteriza a relação entre os mesmos. A intersecção existente entre o interior, o limite e o exterior dos objectos é confrontada com as relações possíveis identi…cadas na `intersectionPatternMatrix`. O resultado é uma variável booleana, obtida através da seguinte função: `Boolean eRelate(TP_Object, TP_Object, intersectionPatternMatrix)`.

deslocado	$\begin{vmatrix} ? & ? \\ ? & ? \end{vmatrix}$	adjacente	$\begin{vmatrix} :? & ? \\ ? & ? \end{vmatrix}$
contém	$\begin{vmatrix} ? & ? \\ :? & :? \end{vmatrix}$	cobre	$\begin{vmatrix} :? & ? \\ :? & :? \end{vmatrix}$
dentro	$\begin{vmatrix} ? & :? \\ ? & :? \end{vmatrix}$	coberto	$\begin{vmatrix} :? & :? \\ ? & :? \end{vmatrix}$
igual	$\begin{vmatrix} :? & ? \\ ? & :? \end{vmatrix}$	sobrepõe	$\begin{vmatrix} :? & :? \\ :? & :? \end{vmatrix}$

Tabela 3.4: Intersecções existentes entre o interior e o limite de dois objectos sem buracos
Adaptado de: [Egenhofer e Sharma, 1993]

entre regiões sem buracos, na qual apenas é considerado o limite e interior dos objectos em causa.

Uma relação topológica r entre dois objectos A e B (isto é, entre dois conjuntos de pontos) é de...nida por um conjunto de 4 intersecções, I, que conjuga o interior e os limites de A, A^o e @A, com o interior e os limites de B, B^o e @B [Egenhofer e Sharma, 1993]:

$$I = \begin{vmatrix} @A \setminus @B & @A \setminus B^o \\ A^o \setminus @B & A^o \setminus B^o \end{vmatrix}$$

As oito relações topológicas apresentadas anteriormente, são então caracterizadas recorrendo as intersecções existentes entre duas regiões. A Tabela 3.4 apresenta cada uma das relações e ainda a interpretação geométrica correspondente, a qual é obtida veri...cando se o resultado de cada uma das intersecções é ou não nulo, ? e :? respectivamente.

A construção das regras de inferência, para as relações topológicas, é conseguida aplicando as propriedades transitivas inerentes à teoria dos conjuntos, a cada um dos subconjuntos que compõem uma intersecção. A composição é aqui expressa em termos de intersecção de conjuntos, onde a intersecção $I_z[P^A_\alpha;P^C_\alpha]$ é derivada das intersecções $I_x[P^A_\alpha;P^B_\alpha]$ e $I_y[P^B_\alpha;P^C_\alpha]$ que de...nem as relações topológicas A r_x B e B r_y C (onde $P^A_i;P^A_j \in \{@A;A^o\}$ e $P^A_i \neq P^A_j$, $P^B_l;P^B_m \in \{@B;B^o\}$ e $P^B_l \neq P^B_m$, $P^C_o;P^C_p \in \{@C;C^o\}$ e $P^C_o \neq P^C_p$) [Hernández, 1994].

O resultado da composição das intersecções é comparado com o conjunto das oito intersecções possíveis (Tabela 3.4), permitindo identi...car a(s) relação(ões) topológica(s) resultante(s). A Tabela 3.5 apresenta a tabela de composição que permite a inferência de relações topológicas, cujas regras foram obtidas atendendo aos princípios acima descritos.

	Deslocado	Adjacente	Igual	Dentro	Coberto	Contém	Cobre	Sobrepõe
	d	m	e	i	cb	ct	cv	o
Deslocado	d, m, e, i, cb, ct, cv, o	d, m, i, cb, o	d	d, m, i, cb, o	d, m, i, cb, o	d	d	d, m, i, cb, o
Adjacente	d, m, ct, cv, o	d, m, e, cb, cv, o	m	i, cb, o	m, i, cb, o	d	d, m	d, m, i, cb, o
Igual	d	m	e	i	cb	ct	cv	o
Dentro	d	d	i	i	i	d, m, e, i, cb, ct, cv, o	d, m, i, cb, o	d, m, i, cb, o
Coberto	d	d, m	cb	i	i, cb	d, m, ct, cv, o	d, m, e, cb, cv, o	d, m, i, cb, o
Contém	d, m, ct, cv, o	ct, cv, o	ct	e, i, cb, ct, cv, o	ct, cv, o	ct	ct	ct, cv, o
Cobre	d, m, ct, cv, o	m, ct, cv, o	cv	i, cb, o	e, cb, cv, o	ct	ct, cb	ct, cv, o
Sobrepõe	d, m, ct, cv, o	d, m, ct, cv, o	o	i, cb, o	i, cb, o	d, m, ct, cv, o	d, m, ct, cv, o	d, m, e, i, cb, ct, cv, o

Tabela 3.5: Inferências Topológicas
Adaptado de: [Egenhofer e Sharma, 1993]

Existem domínios de aplicação nos quais não é necessária a utilização das oito relações topológicas, na caracterização de um dado contexto espacial [Grigni et al., 1995]. Por exemplo, em aplicações de cadastro, a diferença topológica existente entre *dentro* e *coberto* pode não ser relevante, quando o que se pretende é identificar as parcelas de terreno inseridas em determinada região. Outro caso particular ocorre em domínios de aplicação, nos quais as regiões geográficas consideradas representam subdivisões administrativas. Nestes casos, as entidades consideradas apenas se podem relacionar através das primitivas *deslocado*, *contém* e *adjacente*, para as quais a relação *contém* é utilizada apenas quando se analisam simultaneamente diversos níveis hierárquicos.

Dado ser este o contexto verificado neste trabalho, onde as regiões analisadas correspondem a subdivisões administrativas, a Tabela 3.6 apresenta a tabela de composição para este subconjunto de relações topológicas. A primitiva *contém* é aqui incluída, apesar desta relação não ser posteriormente utilizada no processo de raciocínio[9].

3.5 Abordagem integrada ao raciocínio

O raciocínio espacial integrado é caracterizado pela utilização simultânea de mais do que um tipo de relação espacial, na caracterização dos objectos e nas inferências produzidas. Um exemplo deste tipo de raciocínio resulta da inferência do par de relações espaciais (*direcção*,*distância*) existente *A* e *C*, a partir dos pares (*direcção*,*distância*) existentes entre *A* e *B* e entre *B* e *C*. Esta abordagem combina diferentes tipos de relações espaciais, tirando partido das dependências existentes entre as mesmas [Sharma, 1996].

[9] Uma vez que no processo de descoberta de conhecimento, a agregação da informação é realizada até determinado nível hierárquico, não se analisando dois níveis simultaneamente.

	Deslocado	Adjacente	Contém
	d	m	ct
Deslocado	d, m, ct	d, m	d
Adjacente	d, m, ct	d, m,ct	d,m
Contém	d, m, ct	ct,m	ct

Tabela 3.6: Subconjunto das relações topológicas, para o caso particular das regiões representarem subdivisões administrativas

Adaptado de: [Grigni et al., 1995]

Duas abordagens integradas são analisadas nas próximas subsecções. A primeira integra a direcção e a distância, enquanto que a segunda integra a direcção e a topologia. O objectivo é veri…car os princípios que ditaram a construção destes sistemas integrados, e averiguar a possibilidade de desenvolvimento de um sistema que integre a direcção, a distância e a topologia no processo de raciocínio.

3.5.1 Integração da direcção e distância

Raciocinar qualitativamente em termos de distâncias envolve necessariamente a veri…cação da direcção existente entre os objectos referenciados. Por exemplo, os factos A *muito distante* B e B *muito distante* C não permitem inferir com precisão o relacionamento existente entre A e C. A pode estar *próximo* ou *muito próximo* de C ou A pode estar *distante* ou *muito distante* de C, consoante a orientação existente entre A e B, e entre B e C (Figura 3.7).

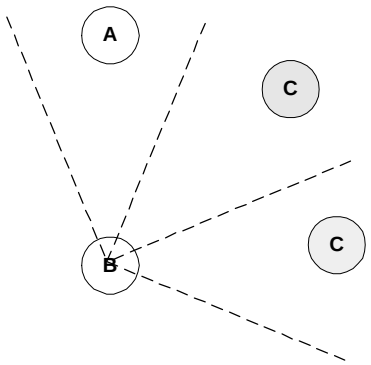


Figura 3.7: In‡uência da direcção na determinação da distância entre objectos

O desenvolvimento de mecanismos de raciocínio, que permitam a integração destes dois tipos de relações espaciais, conduziu Hong [Hong, 1994] a adopção de dois conjuntos de identi…cadores qualitativos para a caracterização das relações a integrar. Cada identi…cador é posteriormente associado a um intervalo de validade quantitativo. Em relação às direcções, os identi…cadores são obtidos utilizando um sistema de referência cardinal, no qual os símbolos,

Norte, Sul, ..., representam um conjunto de graus no referido sistema. É utilizado um conjunto de oito símbolos, no qual os identi…cadores válidos são: N, NE, E, SE, S, SO, O e NO.

No que diz respeito às distâncias, e em conformidade com o descrito para o caso do raciocínio homogéneo (subsecção 3.4.2), a de…nição dos intervalos de validade deve seguir diversos critérios, os quais in‡uenciam a robustez das inferências obtidas. Os quatro símbolos qualitativos utilizados são: mp, p, d e md.

A integração destas relações espaciais conduz à construção de um sistema de localização com 32 áreas de aceitação (8x4 símbolos), representadas na Figura 3.8 por r_{ij}, em que $0 \cdot i \cdot 3$ e $0 \cdot j \cdot 7$.

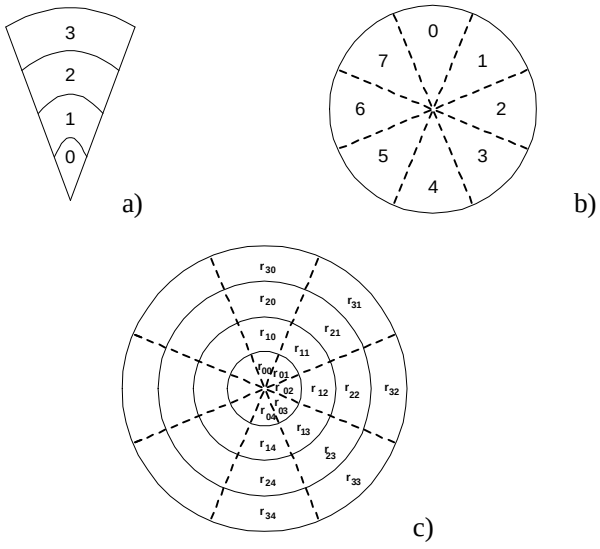


Figura 3.8: Integração da direcção e distância: a) distâncias; b) direcções; c) sistema de localização com 32 áreas de aceitação

A análise das características associadas aos intervalos de validade que podem ser de…nidos para as distâncias, permitiu a Hong [Hong, 1994] simular vários conjuntos de intervalos, os quais podem ser utilizados e/ou manipulados pelos utilizadores, por forma a se adaptarem ao contexto no qual serão utilizados. Os diversos ratios simulados permitem a construção de intervalos de dimensão regular (Tabela 3.7), nos quais existe um ratio constante entre o comprimento de dois intervalos vizinhos ($ratio = comp(dist_i)/comp(dist_{i-1})$).

Os intervalos simulados permitem a construção de novos intervalos, através da ampliação dos intervalos originais por determinado factor. Por exemplo, o conjunto de valores obtidos para o ratio 4 pode ser multiplicado pelo factor 10, originando um novo conjunto de intervalos: $dist_0$ (0, 10], $dist_1$ (10,50], $dist_2$ (50, 210] e $dist_2$ (210, 850]. Uma vez que o mesmo factor amplia os limites de todos os intervalos (relações quantitativas), as inferências qualitativas permanecem inalteradas seja qual for o factor utilizado.

ratio	$dist_0$	$dist_1$	$dist_2$	$dist_3$
1	(0,1]	(1,2]	(2,3]	(3,4]
2	(0,1]	(1,3]	(3,7]	(7,15]
3	(0,1]	(1,4]	(4,13]	(13,40]
4	(0,1]	(1,5]	(5,21]	(21,85]
5	(0,1]	(1,6]	(6,31]	(31,156]
6	(0,1]	(1,7]	(7,43]	(43,259]
7	(0,1]	(1,8]	(8,57]	(57,400]
8	(0,1]	(1,9]	(9,73]	(73,585]
9	(0,1]	(1,10]	(10,91]	(91,820]
10	(0,1]	(1,11]	(11,111]	(111,1111]
20	(0,1]	(1,21]	(21,421]	(421,8421]
50	(0,1]	(1,51]	(51,2551]	(2551,127551]
100	(0,1]	(1,101]	(101,10101]	(10101,1010101]

Tabela 3.7: Intervalos de validade quantitativos para distâncias qualitativas
Adaptado de: [Hong, 1994]

A construção da tabela de inferências, que integra a direcção e a distância no processo de raciocínio, passa pela adopção de princípios quantitativos [Hong, 1994] [Hong et al., 1995]. A transformação dos indicadores qualitativos utilizados em valores quantitativos, associa características geométricas às direcções e distâncias qualitativas. Estas características podem posteriormente ser manipuladas por métodos quantitativos, soma de vectores, sendo os valores obtidos posteriormente transformados em identi…cadores qualitativos. As regras qualitativas obtidas são então utilizadas no processo de raciocínio, não voltando a ser utilizado qualquer método quantitativo.

As regras construídas são caracterizadas por originarem um resultado com uma única inferência, par (direcção,distância), uma vez que quantitativamente é utilizado o ponto médio dos intervalos no processo de construção das regras, e não todos os valores contidos nos mesmos. Por exemplo, ao símbolo qualitativo Este é associado o valor quantitativo 90^oe não todos os valores do intervalo 67.5^oe 112.5^o. Desta forma, existe apenas um par de valores quantitativos, para uma dada relação qualitativa, sendo o resultado da composição dado por apenas um valor qualitativo.

Uma vez que a tabela de composição que dita as inferências possíveis depende dos intervalos de validade considerados para a distância, os diversos ratios de…nidos por Hong [Hong, 1994] permitiram-lhe a de…nição de várias tabelas de inferência. A Tabela 3.8 apresenta o conjunto de inferências possíveis para o ratio 4^{10} entre intervalos. O símbolo I é utilizado para representar a identidade, inferência da mesma localização (direcção e distância).

Pela análise da referida tabela veri…ca-se que os casos apresentados, ¢dir_0, ¢dir_1, ¢dir_2, ¢dir_3 e ¢dir_4, representam localizações com a mesma direcção até localizações com direcções

[10] Refere-se que este foi o ratio inicialmente considerado, por o mesmo permitir cobrir quantitativamente as distâncias que podem existir no espaço geográ…co analisado neste trabalho.

$4dir_0$	$4dir_1$	$4dir_2$	$4dir_3$	$4dir_4$
$r_{00}; r_{00} ! r_{00}$	$r_{00}; r_{01} ! r_{00}$	$r_{00}; r_{02} ! r_{01}$	$r_{00}; r_{03} ! r_{02}$	$r_{00}; r_{04} ! I$
$r_{10}; r_{00} ! r_{10}$	$r_{10}; r_{01} ! r_{10}$	$r_{10}; r_{02} ! r_{10}$	$r_{10}; r_{03} ! r_{10}$	$r_{10}; r_{04} ! r_{10}$
$r_{20}; r_{00} ! r_{20}$	$r_{20}; r_{01} ! r_{20}$	$r_{20}; r_{02} ! r_{20}$	$r_{20}; r_{03} ! r_{20}$	$r_{20}; r_{04} ! r_{20}$
$r_{30}; r_{00} ! r_{30}$	$r_{30}; r_{01} ! r_{30}$	$r_{30}; r_{02} ! r_{30}$	$r_{30}; r_{03} ! r_{30}$	$r_{30}; r_{04} ! r_{30}$
$r_{10}; r_{00} ! r_{20}$	$r_{00}; r_{11} ! r_{11}$	$r_{00}; r_{12} ! r_{12}$	$r_{00}; r_{13} ! r_{13}$	$r_{00}; r_{14} ! r_{14}$
$r_{20}; r_{00} ! r_{20}$	$r_{10}; r_{11} ! r_{21}$	$r_{10}; r_{12} ! r_{11}$	$r_{10}; r_{13} ! r_{12}$	$r_{10}; r_{14} ! I$
$r_{30}; r_{00} ! r_{30}$	$r_{20}; r_{11} ! r_{20}$	$r_{20}; r_{12} ! r_{20}$	$r_{20}; r_{13} ! r_{20}$	$r_{20}; r_{14} ! r_{20}$
$r_{20}; r_{00} ! r_{30}$	$r_{30}; r_{11} ! r_{30}$	$r_{30}; r_{12} ! r_{30}$	$r_{30}; r_{13} ! r_{30}$	$r_{30}; r_{14} ! r_{30}$
$r_{30}; r_{00} ! r_{30}$	$r_{00}; r_{21} ! r_{21}$	$r_{00}; r_{22} ! r_{22}$	$r_{00}; r_{23} ! r_{23}$	$r_{00}; r_{24} ! r_{24}$
$r_{30}; r_{00} ! r_{30}$	$r_{10}; r_{21} ! r_{21}$	$r_{10}; r_{22} ! r_{22}$	$r_{10}; r_{23} ! r_{23}$	$r_{10}; r_{24} ! r_{24}$
	$r_{20}; r_{21} ! r_{31}$	$r_{20}; r_{22} ! r_{21}$	$r_{20}; r_{23} ! r_{22}$	$r_{20}; r_{24} ! I$
	$r_{30}; r_{21} ! r_{30}$	$r_{30}; r_{22} ! r_{30}$	$r_{30}; r_{23} ! r_{30}$	$r_{30}; r_{24} ! r_{30}$
	$r_{00}; r_{31} ! r_{31}$	$r_{00}; r_{32} ! r_{32}$	$r_{00}; r_{33} ! r_{33}$	$r_{00}; r_{34} ! r_{34}$
	$r_{10}; r_{31} ! r_{31}$	$r_{10}; r_{32} ! r_{32}$	$r_{10}; r_{33} ! r_{33}$	$r_{10}; r_{34} ! r_{34}$
	$r_{20}; r_{31} ! r_{31}$	$r_{20}; r_{32} ! r_{32}$	$r_{20}; r_{33} ! r_{33}$	$r_{20}; r_{34} ! r_{34}$
	$r_{30}; r_{31} ! r_{31}$	$r_{30}; r_{32} ! r_{31}$	$r_{30}; r_{33} ! r_{32}$	$r_{30}; r_{34} ! I$

Tabela 3.8: Conjunto de Inferências para o *ratio* 4
Adaptado de: [Hong, 1994] p. 139-144

opostas, respectivamente. As restantes inferências são geradas [Hong, 1994] partindo do pressuposto que $¢dir_1 = ¢dir_7$, $¢dir_2 = ¢dir_6$ e $¢dir_3 = ¢dir_5$ (Figura 3.9).

Analisando as composições apresentadas na Tabela 3.8 e alterando a notação utilizada para os símbolos grá…cos (Figura 3.10), utilizados neste trabalho, a Tabela 3.9 apresenta o conjunto de inferências para o sistema de localização de…nido anteriormente. Na representação apresentada, a identidade I utilizada por Hong [Hong, 1994] foi substituída pela inferência do conjunto de todas as direcções possíveis, uma vez que pelas restrições de identidade referidas na secção 3.2, uma localização coincide no máximo com um objecto.

Segundo Hong [Hong, 1994], num sistema com oito direcções, e desde que sejam respeitadas as regras que ditam a de…nição dos intervalos de validade para as distâncias, aplicando

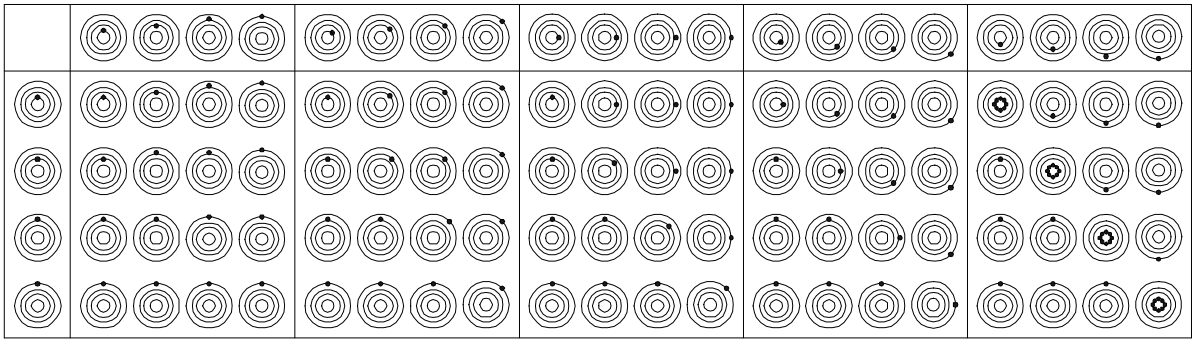


Tabela 3.9: Tabela de Composição para a integração da direcção e distância

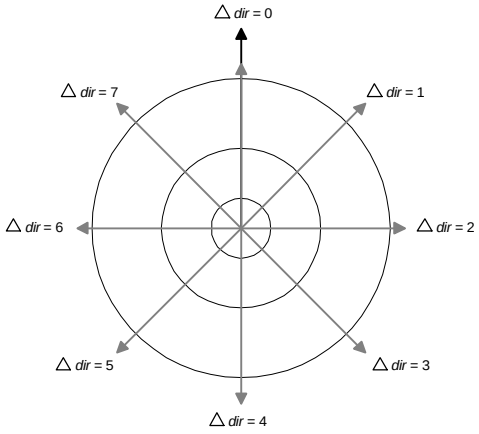


Figura 3.9: Diferenças entre direcções qualitativas

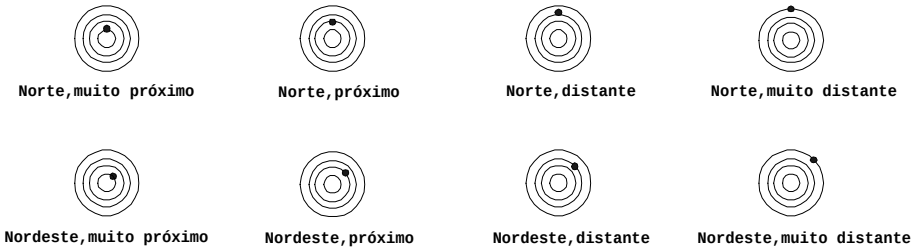


Figura 3.10: Símbolos grá…cos utilizados na representação da integração da direcção e distância

uma rotação de 1/8 ao resultado da composição de N;N, obtemos a inferência correspondente à composição NE;NE. Desta forma, é formulada a regra de que sempre que a diferença entre direcções se mantiver, as restantes inferências, não especi…cadas pelo autor, são obtidas por rotação das respectivas direcções. A título de exemplo, a Tabela 3.10 apresenta as regras de inferência obtidas por rotação das respectivas direcções (zona sombreada).

Apesar de ter sido identi…cada a regra que permite completar a tabela de composição que integra a direcção e distância, julga-se pertinente descrever os mecanismos quantitativos que ditaram a construção do sistema de inferências. Como já referido anteriormente, todo o processo foi baseado na utilização de técnicas quantitativas, as quais são de imediato apresentadas e utilizadas, na veri…cação das regras de inferência para as composições associadas aos grupos de direcções ¢dir$_3$ e ¢dir$_5$.

A localização de dois objectos no espaço pode ser representada, no espaço euclidiano, através de um vector que une os referidos objectos. Adoptando determinada escala quantitativa, a relação que caracteriza a localização destes dois objectos, pode ser calculada veri…cando o comprimento e a orientação do referido vector. O comprimento representa a distância que separa os dois objectos, enquanto que a orientação do vector retrata a direcção existente entre os

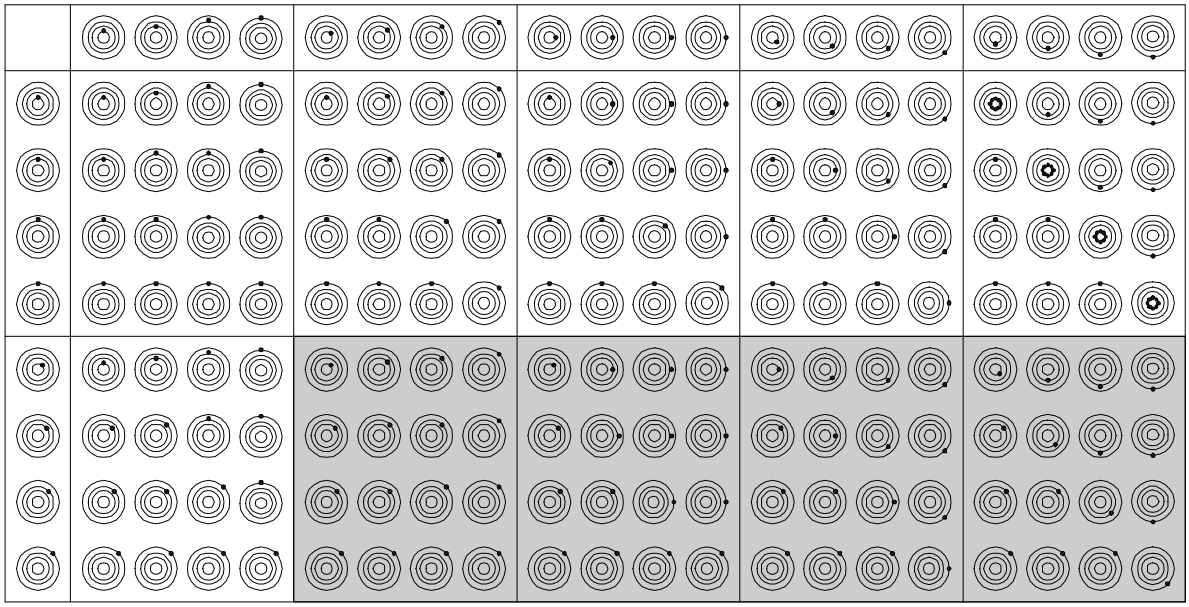


Tabela 3.10: Regras de Inferência obtidas por rotação das direcções

mesmos, num dado sistema de referência. Dados três objectos no espaço, A, B e C, a construção da regra de inferência que dita a relação de localização existente entre A e C, é obtida quantitativamente, a partir das relações existentes entre A e B e entre B e C, pela soma dos vectores V_{AB} e V_{BC}, sendo $V_{AC} = V_{AB} + V_{BC}$ [Maling, 1991] (Figura 3.11).

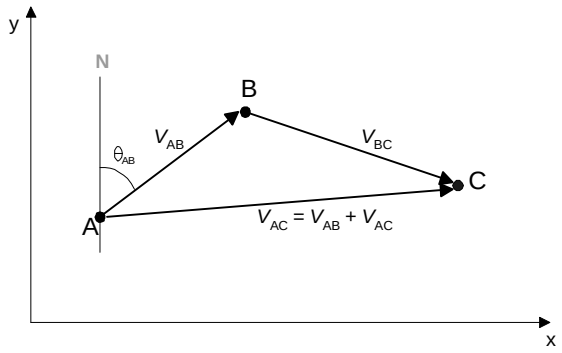


Figura 3.11: Soma de vectores

Adoptando como sistema de referência um sistema de coordenadas cartesianas, as coordenadas $(x;y)$ do vector V_{AB}, com comprimento $jV_{AB}j$ e direcção μ_{AB}, podem ser calculadas (em relação ao eixo dos yy, que de…ne a direcção Norte, e seguindo o sentido dos ponteiros do relógio), por:

$$V_{AB} = (jV_{AB}j \pounds sen(\mu_{AB}); jV_{AB}j \pounds cos(\mu_{AB})) = (V_{ABx}; V_{ABy})$$

Da mesma forma, as coordenadas $(x; y)$ do vector V_{BC} podem ser calculadas por:

$$V_{BC} = (|V_{BC}| \pounds \text{sen}(\mu_{BC}); |V_{BC}| \pounds \cos(\mu_{BC})) = (V_{BCx}; V_{BCy})$$

A direcção e a distância existente entre A e C é assim dada pelo vector V_{AC}, cujo comprimento e ângulo, em relação ao eixo dos yy, são dados por:

$$|V_{AC}| = \sqrt{(V_{ABx} + V_{BCx})^2 + (V_{ABy} + V_{BCy})^2}$$

$$\mu_{AC} = \tan^{-1}\left(\frac{V_{ABx} + V_{BCx}}{V_{ABy} + V_{BCy}}\right)$$

Utilizando estas fórmulas, é possível veri…car as regras que ditam as inferências qualitativas para as direcções pertencentes aos grupos $¢dir_3$ e $¢dir_5$, cujas composições são NE;S, E;SO, SE;O, S;NO, N;SO, NE;O e E;NO (permitindo con…rmar que as mesmas podem ser obtidas por rotação). Antes, é necessário estabelecer a escala utilizada, ou seja, de…nir os intervalos de validade quantitativos para cada um dos identi…cadores qualitativos utilizados.

Em relação à distância, e uma vez que as inferências produzidas dependem do *ratio* entre intervalos considerado, adopta-se o *ratio* 4 por ser um dos utilizados ao longo deste trabalho. Os intervalos de validade para este *ratio* são: mp (0,1], p (1,5], d (5,21] e md (21,85]. A título de exemplo, apenas se apresenta o processo de inferência para o caso da distância qualitativa entre os dois pares de objectos ser a mesma. É utilizado o identi…cador qualitativo mp e valor quantitativo 0:5 como ponto médio do referido intervalo.

Em relação à direcção, os intervalos de validade quantitativos para cada indicador qualitativo, de N a NO seguindo o sentido dos ponteiros do relógio, são: $[337.5^o, 22.5^o)$, $[22.5^o, 67.5^o)$, $[67.5^o, 112.5^o)$, $[112.5^o, 157.5^o)$, $[157.5^o, 202.5^o)$, $[202.5^o, 247.5^o)$, $[247.5^o, 292.5^o)$ e $[292.5^o, 337.5^o)$, sendo os pontos médios considerados de 0^o, 45^o, 90^o, 135^o, 180^o, 225^o, 270^oe 315^o, respectivamente [Hong, 1994]. A Tabela 3.11 evidencia os cálculos quantitativos efectuados e as respectivas direcções inferidas.

A impossibilidade de determinar com precisão as regras de inferência que caracterizam o conjunto das composições representadas por $¢dir_4$, para os casos em que a distância qualitativa entre os dois pares de objectos é a mesma, é ocasionada pelo facto dos objectos estarem tão próximos, que impossibilitam a determinação da direcção existente entre os mesmos [Frank, 1996].

No caso dos objectos considerados possuírem extensão, e representarem regiões administrativas, as quais não se sobrepõem, continua a não ser possível determinar as regras de inferência que ditam as composições para $¢dir_4$. A causa está no facto de se utilizar os pontos médios dos intervalos quantitativos, como valor de referência na construção das regras qualitativas. Os factos A N,mp B e B S,mp C só por si não permitem inferir a direcção existente entre A e C. As oito direcções qualitativas consideradas no sistema de localização, constituem opções válidas para a inferência. A Figura 3.12 apresenta quatro destes casos, sendo os restantes obtidos por troca da posição de A com a posição de C.

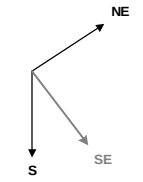
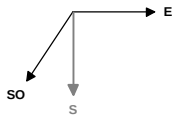
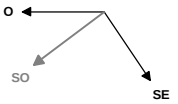
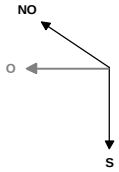
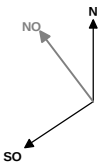
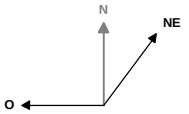
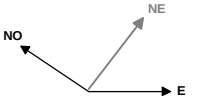
Composição	Cálculos	Resultado
NE;S	V_{AB} V_{BC} distância 0,5 0,5 direcção 225 0 Ang_{AC}= 112,5 $\|V_{AC}\|$= 0,382683	
E;SO	V_{AB} V_{BC} distância 0,5 0,5 direcção 270 45 Ang_{AC}= 157,5 $\|V_{AC}\|$= 0,382683	
SE;O	V_{AB} V_{BC} distância 0,5 0,5 direcção 315 90 Ang_{AC}= 202,5 $\|V_{AC}\|$= 0,382683	
S;NO	V_{AB} V_{BC} distância 0,5 0,5 direcção 0 135 Ang_{AC}= 247,5 $\|V_{AC}\|$= 0,382683	
N;SO	V_{AB} V_{BC} distância 0,5 0,5 direcção 180 45 Ang_{AC}= 292,5 $\|V_{AC}\|$= 0,382683	
NE;O	V_{AB} V_{BC} distância 0,5 0,5 direcção 225 90 Ang_{AC}= 337,5 $\|V_{AC}\|$= 0,382683	
E;NO	V_{AB} V_{BC} distância 0,5 0,5 direcção 270 135 Ang_{AC}= 22,5 $\|V_{AC}\|$= 0,382683	

Tabela 3.11: Veri...cação das regras de inferência

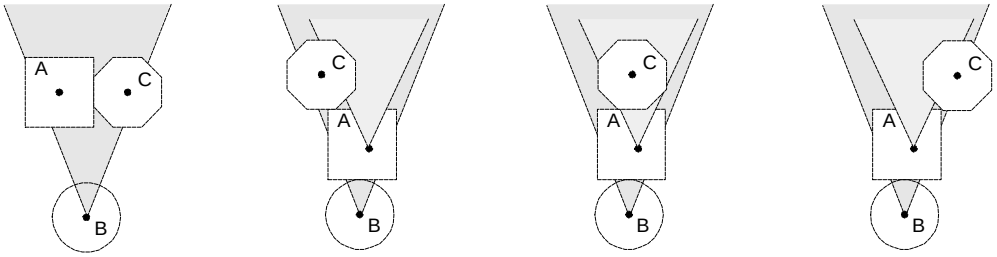


Figura 3.12: Direcções possíveis para as inferências em ¢dir4

Veri…ca-se então que para o caso em que A e C estão precisamente à mesma distância qualitativa de B, a integração da direcção e distância não permite inferir a localização existente entre estas duas regiões. A integração da direcção, distância e topologia permitirá melhorar alguns destes casos, uma vez que a topologia existente entre os objectos limitará o conjunto de direcções possíveis (esta abordagem é apresentada na subsecção 3.5.3).

3.5.2 Integração da direcção e topologia

A posição relativa de dois objectos, no espaço bi-dimensional, pode ser determinada veri…cando a dimensão dos objectos e a orientação existente entre os mesmos. Considerar estes dois factores isoladamente, signi…ca utilizar duas classes de relações espaciais: relações topológicas (que ignoram direcções) e relações de direcção (que ignoram a extensão dos objectos, sendo os mesmos tratados como pontos) [Hernández, 1994].

A integração destas duas classes de relações espaciais conduz à construção de um sistema de raciocínio qualitativo, que permite descrever a posição relativa existente entre os objectos em análise, e ainda especi…car como os limites (fronteiras) destes objectos se relacionam [Hernández, 1994] [Sharma, 1996]. A posição dos objectos é assim descrita por um par de primitivas (direcção, topologia).

Para a integração destes dois tipos de relação espacial, Sharma [Sharma, 1996] recorreu aos princípios estabelecidos por Allen [Allen, 1983] para o domínio temporal, adaptando-os ao domínio espacial. Esta adaptação passa pela análise das primitivas temporais (unidimensionais) ao longo de duas dimensões (eixo dos xx e eixo dos yy).

Raciocinar de forma integrada, em termos de direcção e topologia, requer a de…nição de um esquema de representação comum, que facilite a construção das tabelas de composição. Enquanto que as relações topológicas são independentes da ordem existente entre os objectos, ao longo de um dado eixo, as relações de direcção são dependentes e de…nidas neste sistema veri…cando a ordem dos objectos a longo de determinado eixo. Desta forma, está última permite conhecer a direcção existente entre dois objectos, enquanto que a topologia de…ne a conectividade existente entre os mesmos.

A representação dos pares (direcção, topologia) é efectuada recorrendo às primitivas temporais estabelecidas por Allen [Allen, 1983]. Esta abordagem permite a construção das regras de inferência, para a integração da direcção e topologia, recorrendo a tabela de composição de…nida para o domínio temporal. A transformação das características unidimensionais das primitivas temporais para o espaço bi-dimensional, passa pela veri…cação do par de primitivas

temporais que consegue caracterizar o comportamento do par (direcção, topologia) ao longo do eixo dos xx e do eixo dos yy (Figura 3.13).

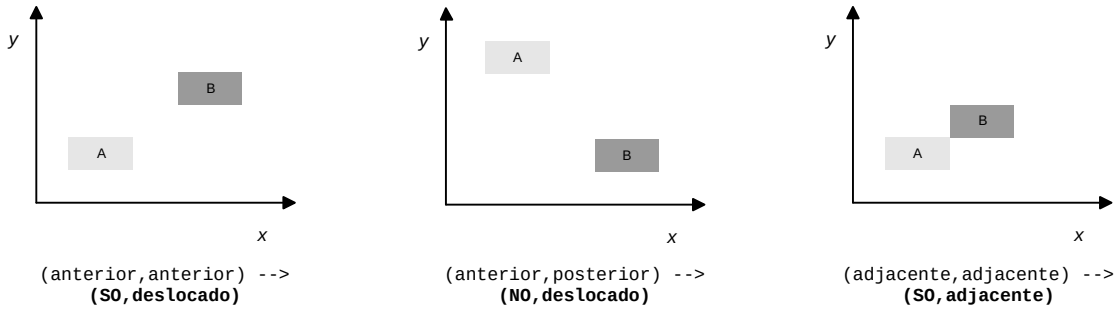


Figura 3.13: Representação da integração da direcção e topologia recorrendo a intervalos temporais

Restringindo o domínio da integração ao caso de entidades geográficas que representem subdivisões administrativas, verifica-se, como já referido anteriormente, que as relações topológicas possíveis são o deslocado e o adjacente, as quais podem ser representadas pelos intervalos temporais anterior e adjacente, e pelas respectivas primitivas inversas. Na caracterização da direcção, todas as primitivas temporais definidas por Allen [Allen, 1983] podem ser utilizadas, originando as várias combinações apresentadas na Figura 3.14.

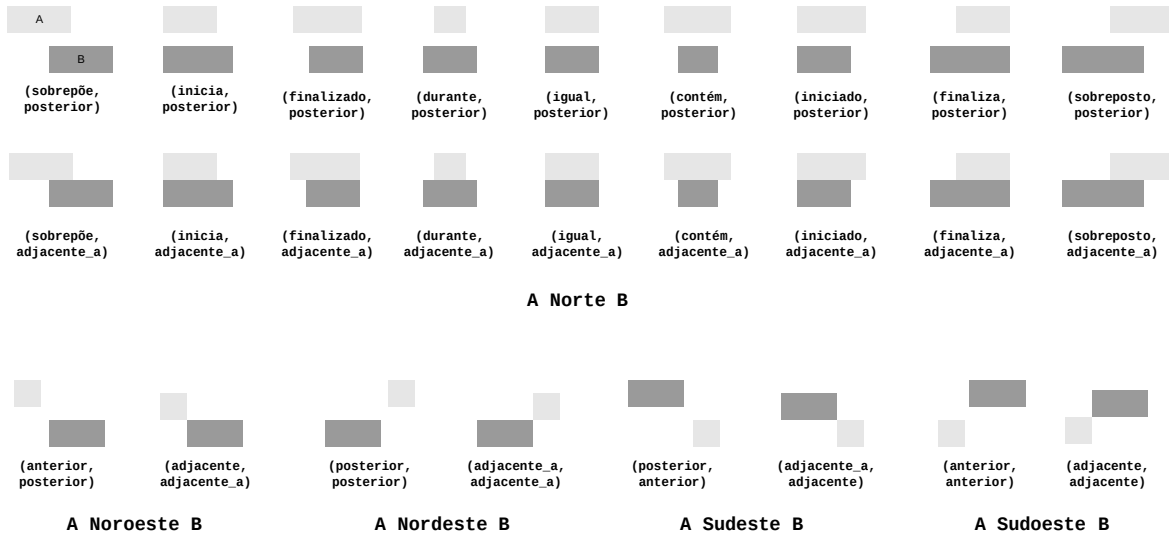


Figura 3.14: Primitivas temporais na caracterização da direcção e topologia (Adaptado de: [Sharma, 1996] p. 83)

A construção das regras de inferência, que ditam o resultado da composição de dois pares (direcção, topologia), passa, em primeiro lugar, pela caracterização dos mesmos recorrendo

a primitivas temporais. Posteriormente, estas primitivas são compostas, consultando-se a Tabela 3.1 (tabela de composição para o domínio temporal) para a determinação do resultado. Para os pares (Nordeste, deslocado) e (Sudeste, adjacente), a regra[11] que dita a inferência resultante da composição dos mesmos é obtida por:

$$
\begin{aligned}
(NE,d);(SE,m) &= (a,a);(mb,m) \\
&= (a;mb)\pounds(a;m) \\
&= (a)\pounds\{a,ob,mb,d,f\} \\
&= (a,a)\_(a,ob)\_(a,mb)\_(a,d)\_(a,f) \\
&= (NE,d)\_(E,d)
\end{aligned}
$$

Através deste processo, Sharma [Sharma, 1996] construiu as várias tabelas de composição, que integram a direcção com os pares topológicos deslocado;deslocado, deslocado;adjacente, adjacente;deslocado e adjacente;adjacente. A Figura 3.15 apresenta o conjunto de símbolos utilizados neste trabalho para a representação desta integração, enquanto que a Tabela 3.12 apresenta o conjunto das inferências possíveis, resultantes da integração dos oito identi…cadores qualitativos considerados para a direcção, com o par topológico deslocado;deslocado. As restantes tabelas, pares deslocado;adjacente, adjacente;deslocado e adjacente;adjacente, podem ser consultadas no Apêndice A.

Norte, deslocado | **Nordeste, adjacente** | **Sudeste, deslocado** ou **Sudeste, adjacente** | Direcção indeterminada, **deslocado** ou **adjacente**

Figura 3.15: Símbolos grá…cos utilizados na representação da integração da direcção e topologia

3.5.3 Integração da direcção, distância e topologia num sistema de raciocínio qualitativo

A integração da direcção e distância permitiu a construção de regras de inferência que originam como resultado um único par (direcção, distância). A única excepção é veri…cada na composição de pares de relações cujas direcções são opostas e cujas distâncias qualitativas são semelhantes.

No caso da integração da direcção e topologia constata-se que, apesar do resultado da composição não ser unívoco, para o caso das direcções opostas, a topologia existente entre os objectos em análise permite limitar o conjunto de direcções resultantes da composição.

A análise destas duas integrações permite constatar que a integração das mesmas, ou seja, a construção de tabelas de inferência para a composição integrada da direcção, distância

[11] Na construção das regras utilizam-se as abreviaturas originais das primitivas (provenientes do termo em inglês), apresentadas na descrição do raciocínio temporal qualitativo (secção 3.3).

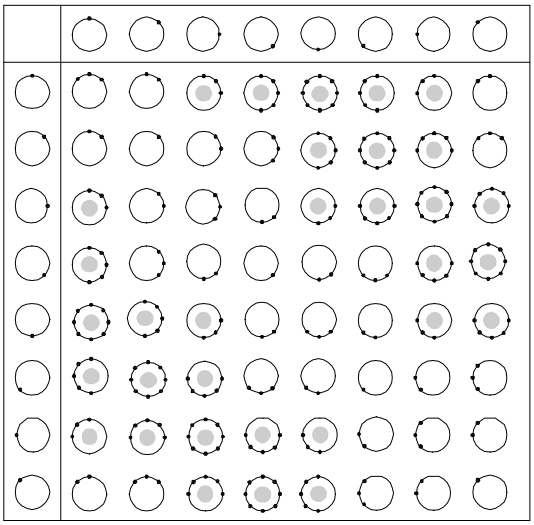


Tabela 3.12: Tabela de composição para a integração da direcção com o par topológico deslocado; deslocado

Adaptado de: [Sharma, 1996] p. 117

e topologia, conduzirá a tabelas de composição mais precisas. A integração não pode contudo ser efectuada, sem proceder à alterações a um dos conjuntos envolvidos. Tal veri…ca-se por Hong [Hong, 1994] ter baseado todo o seu sistema de raciocínio no modelo triangular, enquanto que Sharma [Sharma, 1996] utiliza um sistema de projecções na de…nição das relações entre os objectos.

Dado pretender-se utilizar o modelo triangular, os princípios adoptados por Sharma [Sharma, 1996] são neste projecto rede…nidos, por forma a caracterizarem a de…nição de pares (direcção, topologia), nos quais a direcção entre os objectos é determinada pela orientação existente entre os seus respectivos centróides.

A veri…cação dos intervalos temporais que permitem caracterizar estes pares (direcção, topologia), conduziu à de…nição de dois novos conjuntos de pares de primitivas temporais. A Figura 3.16 apresenta a representação por intervalos temporais adoptada para o caso particular da integração da direcção com a relação topológica deslocado, enquanto que a Figura 3.17 apresenta a integração da direcção com a relação topológica adjacente.

A atribuição de intervalos temporais apresentada visou caracterizar a localização dos centróides dos objectos envolvidos, tendo o cuidado de salientar que os mesmos representam neste caso particular regiões administrativas. Estas não apresentam contornos uniformes, podendo o centróide indicar determinada orientação, apesar da região em causa poder ter parcelas de terreno noutras áreas de aceitação. Deste modo, a escolha da primitiva durante para caracterizar as direcções Norte, Este, Sul e Oeste, foi motivada pelo facto desta permitir indicar que o centróide do objecto primário está incluído no intervalo de validade triangular, de…nido pelo objecto de referência para essas direcções.

Na de…nição dos intervalos temporais que caracterizam a integração da direcção com a

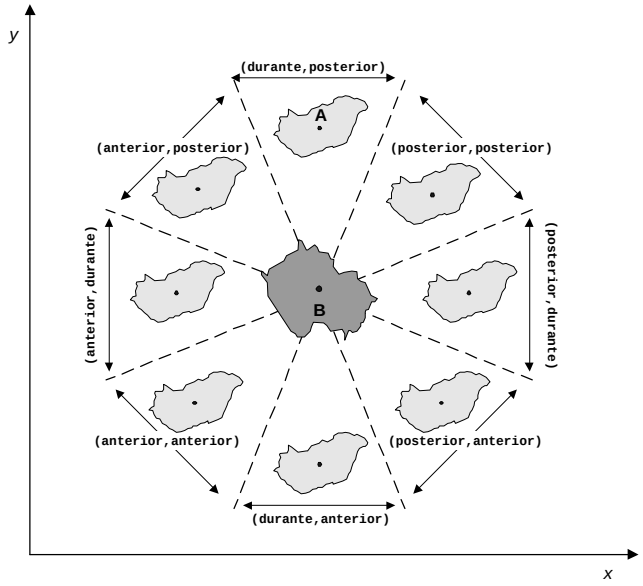


Figura 3.16: Caracterização por intervalos temporais da integração da direcção com a relação topológica deslocado

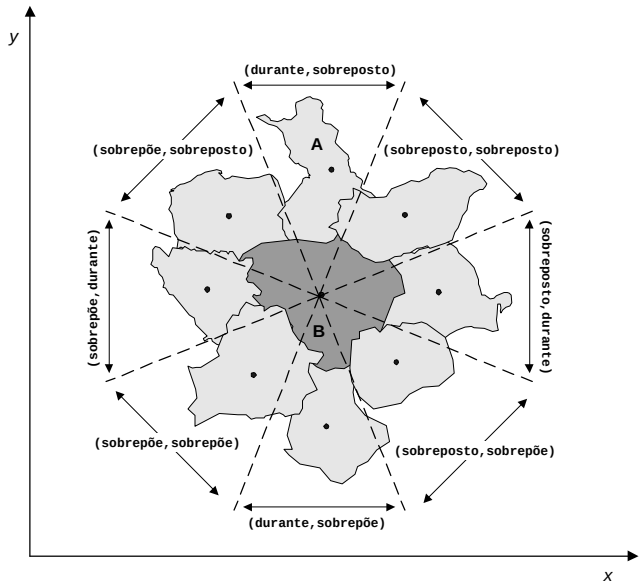


Figura 3.17: Caracterização por intervalos temporais da integração da direcção com a relação topológica adjacente

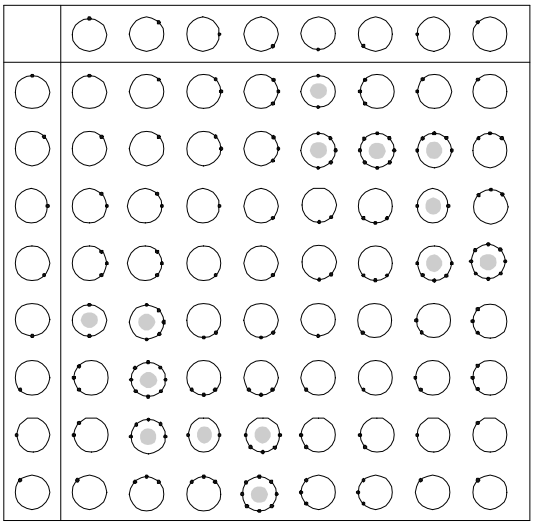


Tabela 3.13: Tabela de composição para a integração da direcção com o par topológico deslocado; deslocado, seguindo a abordagem proposta neste trabalho

relação topológica adjacente, veri…ca-se mais uma vez, que pelo facto dos objectos geográ…cos representarem regiões administrativas, poderá existir alguma sobreposição das duas entidades ao longo dos eixos, quando analisadas pela perspectiva temporal. Este facto motivou a escolha das primitivas temporais sobrepõe e sobreposto, ao invés de adjacente e adjacente_a como utilizado por Sharma [Sharma, 1996]. A adopção destes intervalos, e como poderá ser constatado pela análise das tabelas de inferência obtidas, impõe uma grande ‡exibilidade ao sistema de inferências (analisado posteriormente nesta subsecção).

A construção das tabelas de inferência, seguindo estes intervalos, é apresentada em detalhe no Apêndice B. A Tabela 3.13 apresenta as regras de composição para o caso particular da integração da direcção, com o par topológico deslocado;deslocado (obtidas utilizando as primitivas temporais acima descritas, nas …guras 3.16 e 3.17).

Após a construção das tabelas de composição para a integração da direcção e topologia, segundo os princípios do modelo triangular, impõe-se a integração destas com a tabela que integra a direcção e distância (Tabela 3.9). Os símbolos utilizados, para representar a integração destes três tipos de relações espaciais, são apresentados na Figura 3.18.

O processo de integração foi precedido de uma avaliação às características do domínio de aplicação, no qual o sistema integrado será utilizado. Assim, refere-se que a distância qualitativa existente entre regiões adjacentes está limitada a muito próximo ou próximo, já que o ratio escolhido para a de…nição dos intervalos de validade quantitativos assim o determinou[12]. Veri…ca-se então que:

[12] Julga-se contudo que esta deverá ser sempre a situação veri…cada, uma vez que a escolha dos ratios deverá permitir antecipar a relação topológica existente entre duas regiões. Por exemplo, o facto da distância entre duas regiões ser de distante, deve implicar que estas não se tocam. Para o caso da distância próximo, as regiões poderão

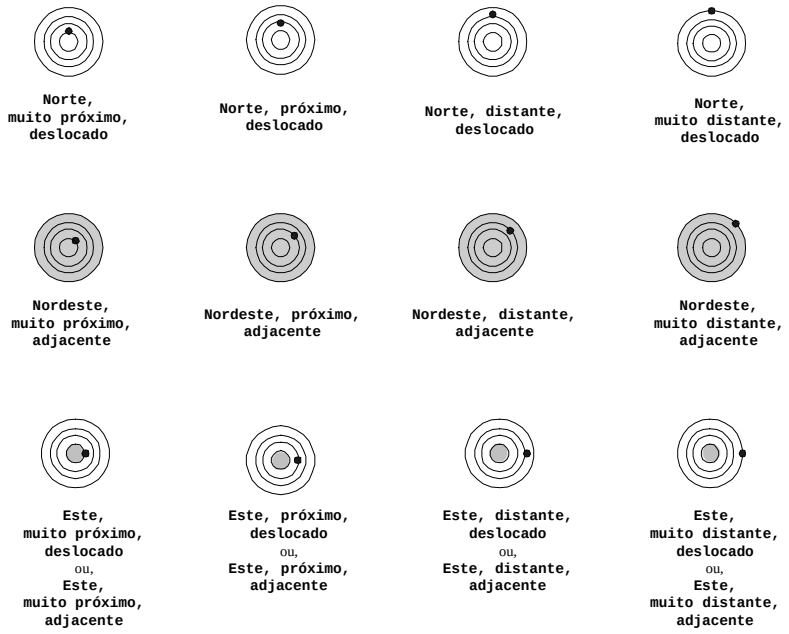


Figura 3.18: Símbolos utilizados na integração das relações espaciais direcção, distância e topologia.

- sempre que o resultado da integração da direcção e distância originar uma relação com distância qualitativa de distante ou muito distante, a relação topológica associada é deslocado;
- o identi…cador qualitativo muito próximo está associado a regiões adjacentes;
- o identi…cador qualitativo próximo poderá estar associado à primitiva topológica deslocado ou à primitiva topológica adjacente.

Pelo que a integração:

- do par topológico adjacente;adjacente, com a direcção e distância, apenas faz sentido para os casos das distâncias mp;mp, mp;p, p;mp e p;p;
- do par topológico deslocado;deslocado, com a direcção e distância, passa pela veri…cação dos casos p;p, uma vez que todas as restantes combinações de p com d e md, têm implícito o resultado topológico deslocado;
- dos pares topológicos deslocado;adjacente e adjacente;deslocado, com a direcção e distância, é efectuada para os casos mp;p, p;mp e p;p.

ou não ser adjacentes, dependendo da extensão das mesmas. A primitiva muito próximo está limitada à relação topológica adjacente, mais uma vez condicionada pelos intervalos de validade que podem ser de…nidos.

O critério de integração destas três relações espaciais baseia-se no facto de que a direcção resultante da integração da direcção e distância é a mesma que a inferida pela integração da direcção e topologia, ou então encontra-se no conjunto de direcções possíveis. Tal permite veri…car que, para os casos em que a distância in‡uencia a direcção, a integração da direcção e topologia apresenta, como resultado da composição, um conjunto de direcções possíveis, enquanto que para os restantes casos, apenas é inferida uma direcção qualitativa.

A junção destes dois pares de relacionamentos permite assim acrescentar a topologia, à direcção e distância já conhecidas. Sempre que a integração da direcção e topologia originar como resultado mais do que uma direcção, a direcção a considerar é a indicada pela integração da direcção e distância (é esta a direcção que comanda o processo de integração, uma vez que é mais precisa). A Figura 3.19 evidencia o processo de integração. A tabela de composição resultante, e que integra relações espaciais do tipo direcção, distância e topologia, segundo os princípios do raciocínio espacial qualitativo, pode ser consultada no Apêndice B (no qual pode ainda ser analisado o processo de construção de todo o sistema de raciocínio).

A integração efectuada permitiu limitar, em algumas composições, o leque de possíveis respostas para o caso das direcções opostas. Contudo, e pretendendo-se inferir com clareza o relacionamento existente entre duas regiões, refere-se que a implementação do sistema de inferências construído será efectuada com vista a atingir este objectivo. A implementação é descrita em detalhe no Capitulo 5, subsecção 5.3.2, mas adianta-se que, sempre que a inferência da relação espacial existente entre A e C, não puder ser obtida a partir das relações existentes entre A e B, e entre B e C, esta será determinada veri…cando outras relações vizinhas. No Capítulo 6 é apresentada uma avaliação detalhada ao sistema de inferências, permitindo veri…car o desempenho do mesmo, isto é, a validade das inferências produzidas.

Uma avaliação preliminar da tabela de composição construída, permite constatar que a integração aqui proposta caracteriza relações de localização, existentes entre regiões adjacentes de tamanhos não uniformes. O facto das regiões apresentarem grandes diferenças na extensão, permite que regiões de pequena dimensão sejam contornadas por regiões de grande dimensão, o que adiciona di…culdades ao processo de raciocínio. Para estes casos, o sistema de inferências construído demonstra ser capaz de os retratar, ao contrário da abordagem proposta por Sharma [Sharma, 1996]. Veri…que-se, por exemplo, o caso evidenciado na Figura 3.20. O concelho 109 está a Norte do concelho 116, que por sua vez está a Norte do concelho 113. A discrepância existente entre a dimensão destas regiões, pode conduzir a inferências erradas. Através da Figura B.3 (Apêndice B) é possível inferir que o concelho 109 está a Norte do concelho 113, sendo os mesmos adjacentes. Tal não acontece no sistema de inferências de…nido por Sharma, cujo resultado topológico seria erradamente deslocado.

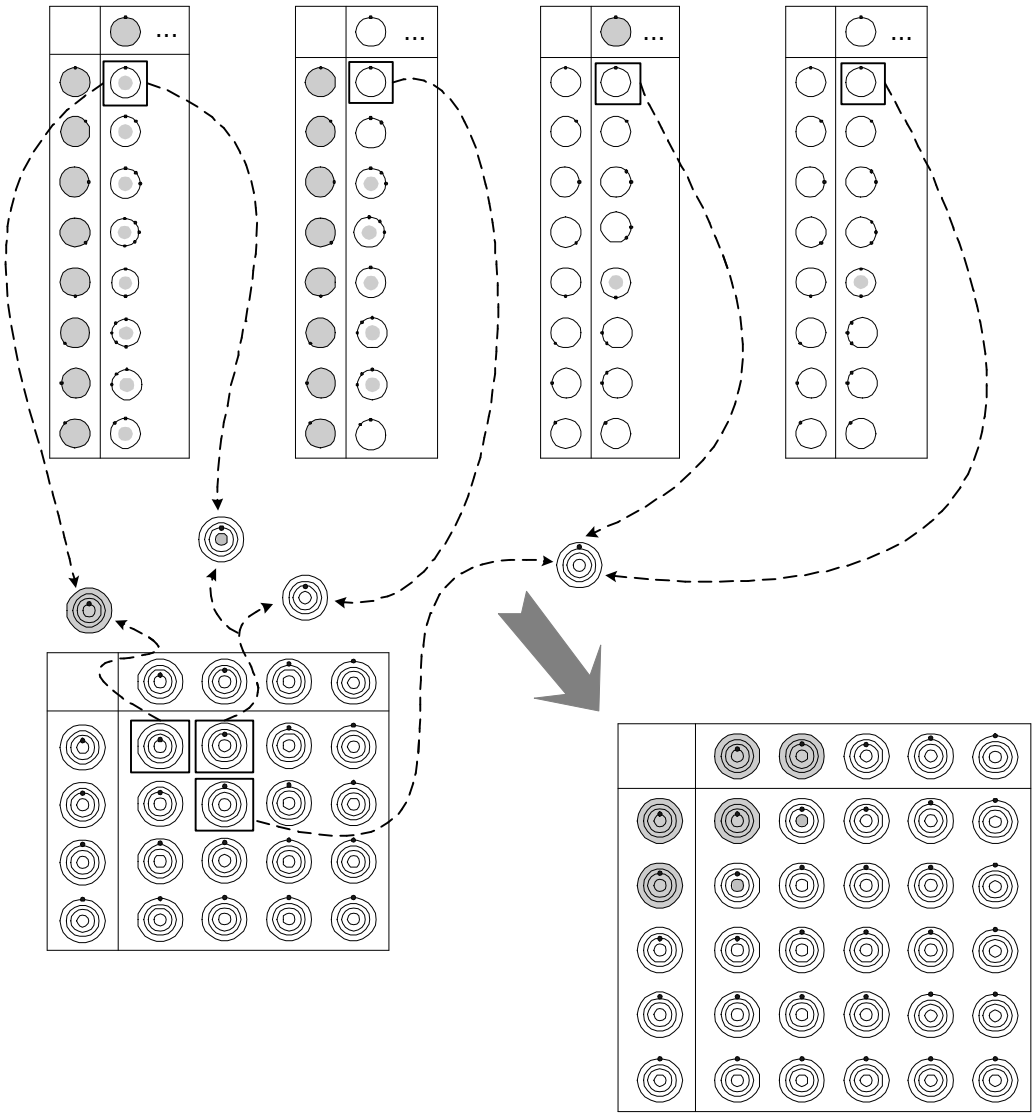


Figura 3.19: Processo de integração das tabelas de composição da direcção e topologia, com a tabela de composição da direcção e distância

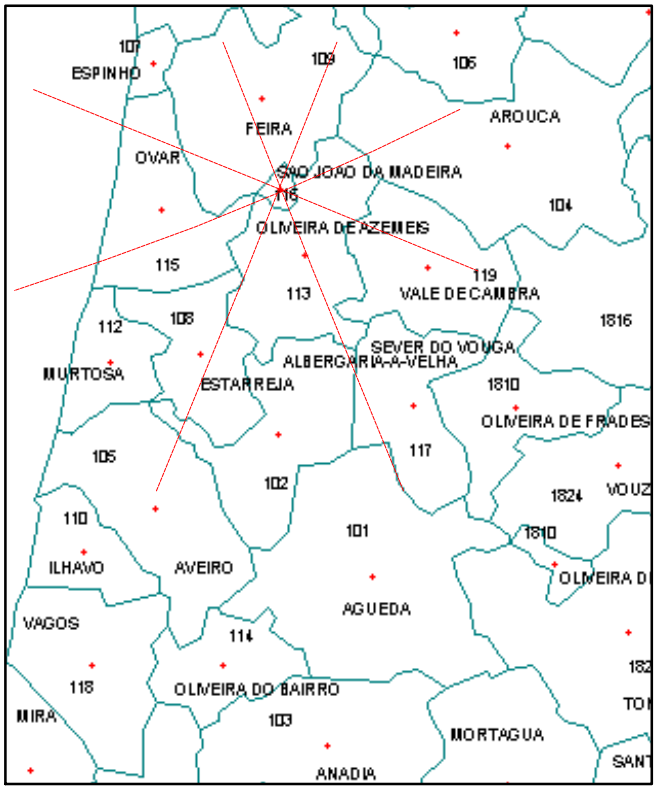


Figura 3.20: Avaliação preliminar da tabela de composição

3.6 A dimensão dos objectos

A dimensão dos objectos não é considerada um tipo de relação espacial. É contudo uma característica que in‡uencia, de forma determinante, o tipo de relação espacial (principalmente para a direcção) que pode existir entre objectos. Apesar de não ter sido considerada no sistema integrado construído na secção anterior, julga-se pertinente apresentar alguns dos conceitos que poderiam ser utilizados num sistema de raciocínio qualitativo que pretendesse tratar esta informação.

Algumas relações topológicas, como contém, inclui,..., apenas podem ser veri…cadas se os objectos envolvidos possuírem extensões apropriadas [Hernández, 1994]. A dimensão pode ainda in‡uenciar a identi…cação da direcção qualitativa existente entre dois objectos.

Uma abordagem qualitativa à utilização da dimensão dos objectos, consiste em adoptar intervalos de magnitude para as dimensões, os quais podem ser posteriormente manipulados e comparados qualitativamente através de operadores como $<$, $>$, $=$. A de…nição dos intervalos estará sempre dependente do contexto no qual os mesmos serão utilizados, limitando o resultado das operações a esse mesmo contexto.

Uma representação qualitativa do tamanho dos objectos é apresentada por Zimmermann [Zimmermann, 1993] [Zimmermann, 1995] [Zimmermann e Freksa, 1996], a qual estabelece diferenças entre dimensões, através da de…nição de parcelas de dimensão. A relação $A > ,tam_1 B$ representa o facto de A ser maior que B no tamanho tam_1, já que $jAj = jBj + jtam_1j$. Esta apro-

ximação, além de utilizar diferentes indicadores para referenciar os diversos tamanhos, permite explicitar a diferença qualitativa na dimensão existente entre diferentes regiões.

Esta representação permite a construção de regras que facilitam a inferência de dimensões desconhecidas. A de…nição dos intervalos de magnitude conduz ao estabelecimento de regras como:

- $tam_3 = tam_2 + tam_1$
- $tam_2 = tam_1 + tam_1$

Estas regras permitem inferir a relação existente entre as magnitudes de A e C a partir, por exemplo, dos factos $A >, tam_2 B$ e $C >, tam_1 B$:

$$A = B + tam_2 = B + tam_1 + tam_1 = C + tam_1 =) A >, tam_1 C$$

Estes mecanismos revelam-se de particular importância, dado que em determinados domínios de aplicação pode ser relevante analisar a área das regiões geográ…cas referenciadas, e estas podem não ser todas conhecidas. As áreas disponíveis podem ser facilmente obtidas via SIG, as quais depois de convertidas à identi…cadores qualitativos, podem ser integradas, por exemplo, no processo de descoberta de conhecimento.

A de…nição da magnitude dos intervalos, e consequente de…nição das regras que de…nem as associações entre os vários intervalos, deverá ser precedida da veri…cação do conjunto de valores quantitativos a representar, e ainda da de…nição do número de indicadores a utilizar.

Capítulo 4

A descoberta de conhecimento em bases de dados

O acentuado desenvolvimento das capacidades informáticas ao nível do armazenamento de dados tem sido acompanhado, e em parte motivado, pela capacidade e necessidade organizacional de gerar e manipular grandes quantidades de dados. A necessidade de recolher e armazenar dados de diversos tipos e proveniências superou a capacidade humana de analisar, sintetizar e extrair conhecimento a partir desses dados. Enquanto as BD fornecem as ferramentas necessárias ao armazenamento e utilização de grandes quantidades de dados, a compreensão e análise dos mesmos requer a utilização de ferramentas apropriadas, que automatizem o processo de análise dos dados e descoberta de conhecimento [Fayyad e Uthurusamy, 1996].

Os princípios associados à DCBD conjugam fundamentos provenientes de diversas áreas, tais como a inteligência arti…cial, a aprendizagem automática, o reconhecimento de padrões, a estatística, as BD, os sistemas de informação, entre outras. As aplicações de DCBD integram teorias, métodos e algoritmos provenientes destas diferentes áreas, tendo como objectivo a extracção de conhecimento a partir de grandes BD [Fayyad et al., 1996a].

Os algoritmos utilizados para procurar padrões nos dados são denominados de algoritmos de DM. O processo global de DCBD, que se desenvolve em várias fases, inclui a gestão dos algoritmos de DM e a interpretação dos padrões encontrados pelos mesmos, os quais são posteriormente utilizados no suporte à tomada de decisão.

Neste capítulo abordam-se os conceitos associados ao processo de DCBD, destacando as principais tarefas associadas ao DM, e as técnicas tradicionalmente utilizadas para a sua execução. Descrevem-se, ainda, as principais abordagens em curso, na área da exploração de dados referenciados espacialmente.

4.1 O processo de descoberta de conhecimento

4.1.1 Os princípios

Fayyad et al. ([Fayyad et al., 1996b] pág.6) de…nem a DCBD como "o processo não trivial de identi…cação de padrões válidos e potencialmente úteis, perceptíveis a partir dos dados". Nesta

de…nição, um padrão pode ser caracterizado por modelos, relações ou estruturas nos dados, que devem ser perceptíveis, se não imediatamente, após determinado período de processamento. Os dados representam um conjunto de factos armazenados na BD, na qual subconjuntos do mesmo são responsáveis pela caracterização de diversos padrões. O termo processo está associado à execução de diversos passos iterativos, que vão desde a selecção dos dados a analisar até a interpretação de resultados. O processo é assumido como não trivial uma vez que pode envolver a procura de estruturas, modelos, padrões ou parâmetros. Os padrões descobertos deverão ser:

- válidos quando aplicados a novos dados (isto é, dados não considerados na construção do modelo ou determinação do padrão);
- desconhecidos do sistema utilizado na sua detecção e preferencialmente do utilizador; e ainda,
- úteis para o utilizador, auxiliando o processo de tomada de decisão.

Além de iterativo (uma vez que pode existir retrocesso à etapas anteriores), este processo é também interactivo, já que requer a participação do utilizador sempre que é necessária a tomada de decisão. As diversas fases do processo de DCBD (descritas na subsecção 4.1.2) encontram-se representadas na Figura 4.1.

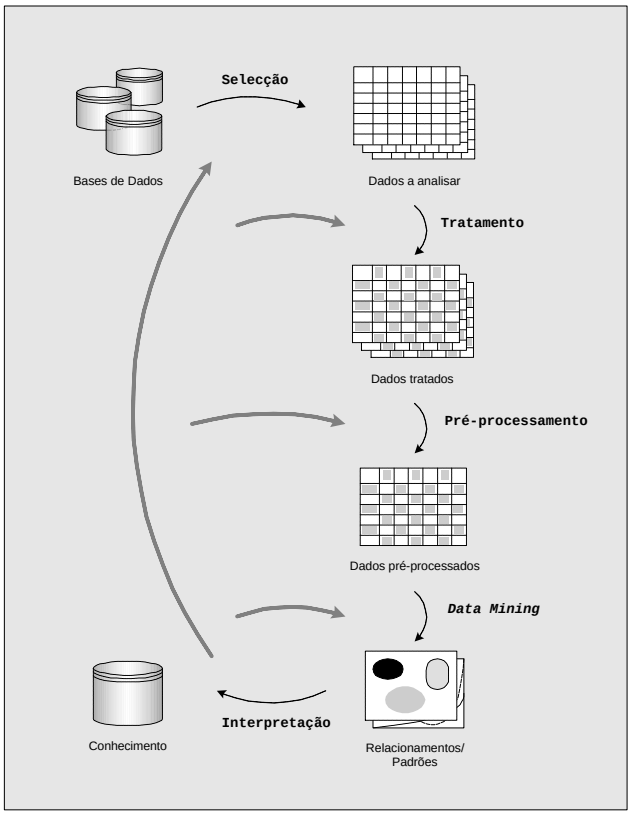


Figura 4.1: Fases do processo de DCBD (Adaptado de: [Fayyad et al., 1996b])

Para Matheus et al. [Matheus et al., 1993], um Sistema de Descoberta de Conhecimento (SDC) é um sistema que encontra conhecimento que lhe é desconhecido, isto é, conhecimento não implícito nos seus algoritmos ou explícito no conhecimento do domínio existente. Estes autores de…niram uma arquitectura genérica para SDC, que agrega seis componentes principais: o controlador, a interface, a base de conhecimento, o foco, a extracção de padrões e a avaliação (Figura 4.2). Nesta arquitectura, o utilizador assume um papel de particular importância na condução do processo de DCBD, já que a maioria dos algoritmos disponíveis, para extrair padrões dos dados, não são completamente autónomos.

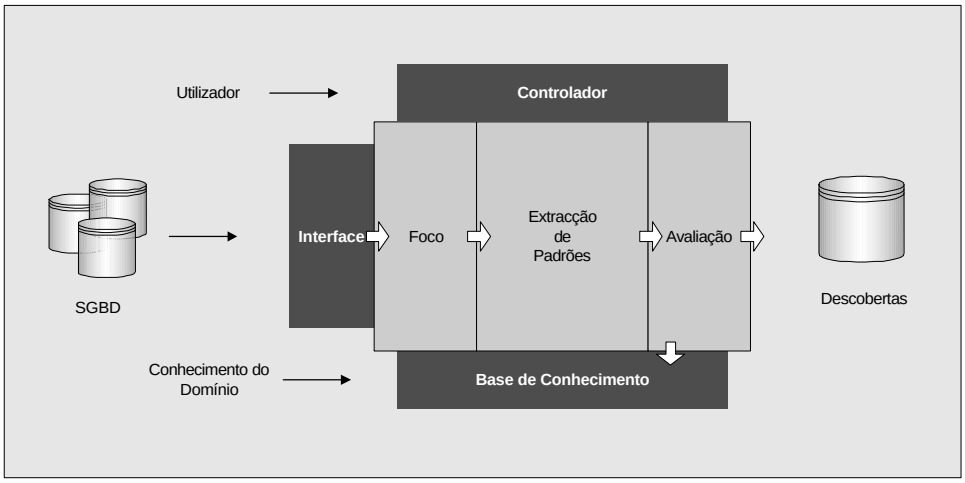


Figura 4.2: Arquitectura genérica para um SDC (Adaptado de: [Matheus et al., 1993])

Nesta arquitectura, a informação que circula no SDC tem três proveniências: os comandos do utilizador fornecidos ao controlador, os dados provenientes do SGBD e o conhecimento do domínio armazenado na base de conhecimento do sistema. Os dados de interesse provenientes do SGBD são explorados e os padrões encontrados avaliados, com o objectivo de veri…car a sua relevância para o utilizador. Sempre que estes sejam catalogados como relevantes, serão armazenados na base de conhecimento do sistema, permitindo a sua reutilização em próximas iterações.

O arquitectura apresentada representa uma abstracção dos componentes necessários num sistema de DCBD. Apesar de na realidade nem sempre ser possível separar claramente os diversos componentes, o modelo de…nido permite comparar sistemas existentes, realçando características como a autonomia (grau de dependência do utilizador) e a versatilidade (algoritmos e técnicas disponíveis). Segundo os autores [Matheus et al., 1993], o aumento da autonomia dos sistemas conduz à diminuição da versatilidade dos mesmos, uma vez que passam a estar vocacionados para a execução de tarefas muito especí…cas. Os sistemas mais versáteis dependem muito do utilizador, mas disponibilizam um conjunto diversi…cado de técnicas, permitindo a sua utilização em diversos domínios de aplicação.

Os vários componentes, já referidos anteriormente, da arquitectura genérica descrita por Matheus et al. [Matheus et al., 1993] são caracterizados por:

- Base de Conhecimento. A base de conhecimento deve incluir um dicionário de dados, com informação relativa aos dados armazenados nas diferentes BD analisadas. Este dicionário deverá incluir o nome e o tipo dos atributos armazenados, e ainda as diversas restrições aplicáveis aos mesmos. Nesta base de conhecimento devem, ainda, ser armazenados os objectivos do utilizador. Esta colecção de informações constitui o conhecimento do domínio. O principal objectivo deste conhecimento é auxiliar o processo de descoberta de relações nos dados, direccionando a pesquisa para determinados atributos. Este direccionamento poderá condicionar pela negativa todo o processo, já que padrões potencialmente interessantes poderão deixar de ser encontrados. Idealmente, os SDC deveriam estar preparados para analisar os dados sem qualquer auxilio do utilizador ou do conhecimento do domínio, armazenando no sistema o conhecimento encontrado e permitindo, ainda, a sua reutilização em posteriores análises.

- Controlador. Este componente condiciona a autonomia do sistema através das especi…cações provenientes do utilizador. Estas especi…cações são interpretadas pelo controlador e utilizadas para direccionar o foco, a extracção de padrões e a avaliação de resultados. Em sistemas cujas tarefas estejam bem de…nidas e permaneçam estáticas, o controlador limita-se a executar um conjunto de operações prede…nidas. Nos sistemas mais versáteis, o controlador pode assumir grandes responsabilidades, requerendo a intervenção do utilizador sempre que se justi…que a tomada de decisão.

- Interface. A interface com o SGBD é utilizada para permitir a selecção dos dados necessários ao componente de extracção de padrões. Esta selecção é efectuada através da especi…cação de uma questão (query), na qual atributos provenientes de diversas tabelas são seleccionados, integrados e disponibilizados para exploração. Esta selecção poderá ser efectuada pelo próprio SDC ou então recorrendo ao SGBD (alternativa mais apropriada sempre que se utilizam dados de diversas proveniências, ou então, quando o próprio tamanho da BD limita o desempenho do SDC).

- Foco. Este componente sugere os dados que devem ser seleccionados para análise. Especi…ca as tabelas a aceder, os atributos a seleccionar de cada uma das mesmas, e ainda, quais ou quantos registos são necessários. Na realização desta tarefa, o foco consulta a informação armazenada na base de conhecimento, nomeadamente a estrutura das tabelas e o tipo de informação requerida pelos diversos algoritmos, por forma a que os dados sejam disponibilizados no formato adequado.

- Extracção de padrões. O termo padrão refere-se a qualquer relacionamento que possa existir entre elementos da BD. Diferentes técnicas e algoritmos para análise dos dados são disponibilizados em SDC, condicionando a versatilidade de cada um dos mesmos.

- Avaliação. As BD estão repletas de padrões, mas poucos deles têm realmente interesse. Um padrão é interessante se for preciso, desconhecido (até a data) e útil aos objectivos do utilizador. O componente de avaliação deve determinar o grau de interesse dos padrões encontrados e decidir quais serão apresentados ao utilizador. Na maioria dos sistemas, a avaliação do interesse de um dado padrão é determinada pelo próprio algoritmo que o encontra, o qual possui restrições de qualidade ou signi…câncias a cumprir. A signi…cância estatística permite medir o interesse de um padrão, na medida em que valida o grau de certeza do relacionamento encontrado.

4.1.2 As fases do processo de descoberta de conhecimento

As diferentes fases do processo de descoberta de conhecimento, já apresentadas na Figura 4.1, incluem a selecção dos dados, tratamento dos dados, pré-processamento dos dados, data mining, e ainda a interpretação de resultados [Fayyad et al., 1996c]. Cada uma destas fases é de seguida brevemente descrita.

Selecção dos dados

Nesta fase são seleccionados os dados operacionais armazenados nos diversos repositórios de dados, desde sistemas transaccionais, armazéns de dados (data warehouses, data marts), etc., necessários aos algoritmos de DM. A selecção dos dados tem como principal objectivo limitar o espaço de pesquisa, eliminado atributos que não têm qualquer interesse no processo de descoberta de conhecimento. Entre os atributos que podem à partida ser eliminados, encontram-se aqueles que têm carácter meramente informativo, como nomes, números de contribuinte, etc.

Tratamento dos dados

No tratamento procede-se à limpeza dos dados. Entre os procedimentos habituais nesta fase [Adriaans e Zantinge, 1996], destaca-se a des-duplicação de registos, normalmente originada por negligência na introdução dos dados, pelo incorrecto fornecimento dos mesmos ou por um erro de digitação. Esta des-duplicação é incluída no processo de limpeza dos dados [Simoudis et al., 1995], no qual se identi…ca e corrige informação incorrecta e incompleta na BD. A veri…cação de inconsistências permite a identi…cação de registos que possuem atributos com valores que não fazem sentido no contexto em que estão a ser utilizados, como, por exemplo, datas erradas.

Pré-processamento dos dados

O pré-processamento dos dados passa essencialmente pela redução do espaço de pesquisa, isto é, pela diminuição do número de linhas/colunas a analisar. Esta redução é conseguida transformando atributos com valores contínuos em atributos com valores discretos, nomeadamente através da substituição de idades por faixas etárias, ou através da generalização de atributos, como regiões, o que permite que os dados sejam agrupados e analisados ao nível de freguesias, concelhos, ou mesmo distritos.

A forma como os dados são codi…cados/agregados in‡uencia fortemente os resultados obtidos. A tarefa de codi…cação é tida como uma tarefa de grande criatividade e que tem de ser repetida sucessivamente por forma a melhorar os resultados [Famili et al., 1997].

Assim, na fase de pré-processamento, os dados provenientes da fase anterior são agrupados ou transformados em valores discretos, por forma a tornar a pesquisa mais e…ciente. Dados muito detalhados são transformados por forma a constituírem agrupamentos, e permitirem a sua análise pelos algoritmos de DM. Nesta fase é requerida a de…nição de hierarquias de conceitos e de classes de valores [Adriaans e Zantinge, 1996].

Além de preocupações com a redução do tamanho da amostra, esta fase é ainda caracterizada por averiguar se existe ou não a necessidade de integração de dados externos à organização

(ou BD em análise). O objectivo é enriquecer eventuais resultados, incluindo dados demográ…cos, socio-económicos, etc., nos dados a analisar.

Data Mining

Esta é a fase de procura, na qual os dados provenientes da fase de pré-processamento são analisados. A veri…cação do tipo de resultados pretendido, tarefa a executar, permite a identi…cação da técnica a utilizar. Para atingir os objectivos propostos pode ser necessário utilizar mais do que uma técnica, já que a qualidade e o tipo dos dados disponíveis in‡uencia de forma decisiva os resultados que podem ser encontrados.

Interpretação de resultados

Nesta fase procede-se à análise dos resultados obtidos pelos algoritmos utilizados na fase anterior. Os modelos encontrados são aplicados a novos conjuntos de dados, permitindo veri…car o desempenho dos mesmos com dados desconhecidos para o sistema.

A ocorrência de falhas ao longo do processo de descoberta de conhecimento, originadas por decisões que se revelam inapropriadas, são normalmente traduzidas na obtenção de modelos que não satisfazem o interesse do utilizador (subjectivo, já que em termos objectivos os algoritmos utilizados veri…cam quantitativamente o interesse das regras), ou que apenas retratam o comportamento dos dados analisados, não podendo ser aplicados a dados desconhecidos. Nestes casos, existe a possibilidade de retrocesso a fases anteriores para recti…car as decisões tomadas ou para incluir novos dados na análise. O processo é então retomado, permitindo identi…car novos modelos que resultam das alterações efectuadas.

Apesar dos algoritmos disponíveis possuírem critérios objectivos de avaliação da qualidade das regras, a introdução de medidas de interesse subjectivas [Silberschatz e Tuzhilin, 1995] [Silberschatz e Tuzhilin, 1996b] tem como objectivo limitar o conjunto de resultados a apresentar ao utilizador. A de…nição de medidas de interesse subjectivas, e que dependem de utilizador para utilizador, tendem a aumentar o grau de envolvimento do utilizador no processo de descoberta de conhecimento [Silberschatz e Tuzhilin, 1996a], tendo como contrapartida o aumento do interesse das diversas regras encontradas. Duas medidas de interesse subjectivas são o grau de surpresa, salientando que um padrão é interessante se ele é inesperado pelo utilizador, e a utilidade do padrão, sendo este interessante se o utilizador o pode utilizar em sua vantagem. Apesar deste tipo de medida de interesse não se encontrar ainda implementado, ele é baseado no pressuposto de que todos os padrões inesperados são úteis e que os padrões mais úteis são os inesperados, captando o grau de utilidade através do grau de surpresa do mesmo.

Padmanabhan e Tuzhilin [Padmanabhan e Tuzhilin, 1999] apresentam uma outra abordagem, que permite a adopção de medidas de interesse subjectivas, sendo esta implementada através da procura de padrões inesperados. O sistema desenvolvido parte da especi…cação das crenças do utilizador, daquilo que acredita ser verdade para o domínio de aplicação em causa, e procura regras que sejam baseadas na negação das referidas crenças. Todos os fundamentos desta aproximação assentam na de…nição de padrões inesperados como sendo aqueles que contradizem o conhecimento inicialmente adquirido. O conhecimento inicialmente adquirido pode ser

especi…cado pelo utilizador ou assimilado através de algoritmos de aprendizagem automática[1], constituindo em ambos dos casos, o conhecimento do domínio que é negado e utilizado para conduzir a procura.

4.1.3 A importância do conhecimento do domínio

A autonomia dos SDC pode ser determinada veri…cando a independência, em relação ao utilizador, de uma dada ferramenta na procura de padrões, e na avaliação do interesse dos resultados obtidos com os mesmos. Deixar a procura ao critério da ferramenta pode tornar estas aplicações ine…cientes, uma vez que a grande quantidade de dados a explorar requer a participação do utilizador, na de…nição do conhecimento do domínio necessário ao direccionamento do processo de descoberta de conhecimento [Anand et al., 1995].

O conhecimento do domínio [Han et al., 1992] [Rainsford e Roddick, 1996] constitui um recurso essencial em qualquer SDC. Pode ser utilizado para conduzir o processo, podendo o conhecimento existente ser complementado com o conhecimento obtido na descoberta de conhecimento, através da introdução das regras encontradas no sistema. As hierarquias conceptuais constituem a principal fonte de conhecimento do domínio, as quais são normalmente complementadas com regras do domínio.

Hierarquias conceptuais

As hierarquias conceptuais representam conhecimento do domínio necessário ao processo de generalização. Os diferentes níveis que as constituem estão organizados segundo determinada taxinomia, normalmente ordenada dos conceitos mais gerais para os mais particulares. O conceito mais geral é representado por um descritor nulo, normalmente uma palavra reservada, ANY, enquanto que os conceitos mais especí…cos correspondem a valores de atributos na BD.

Uma representação usual das hierarquias de conceitos passa pela especi…cação de diversos níveis de $A \frac{1}{2} B$, indicando que B é uma generalização de A [Han et al., 1992].

As hierarquias conceptuais permitem a descrição de atributos recorrendo a diversos níveis de abstracção, surgindo naturalmente em vários domínios de aplicação. Veja-se, por exemplo, o caso das divisões administrativas de Portugal. Ao nível geográ…co podem ser caracterizadas por uma hierarquia que considere subdivisões do território em distritos, concelhos e freguesias. Esta hierarquia pode ser apresentada em forma de árvore (Figura 4.3), ou ainda, utilizando a notação introduzida anteriormente, por:

freguesia ½ concelho ½ distrito ½ ANY(local)
{Agadão, Espinhel, ...} ½ Águeda
{Besteiros, Caldelas, ...} ½ Amares
{Benfeita, Cepos, ...} ½ Arganil
...
{Águeda, Arouca, ...} ½ Aveiro
{Amares, Barcelos, ...} ½ Braga

[1] As regras encontradas devem ser validadas pelo utilizador, con…rmando-as ou não como regras do negócio.

{Arganil, Mira, ...} ½ Coimbra

...

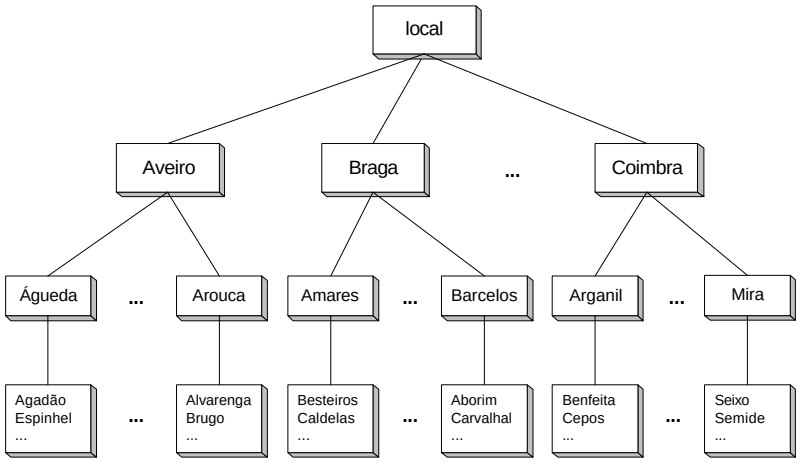


Figura 4.3: Hierarquia conceptual para subdivisões administrativas de Portugal

Estas hierarquias permitem de…nir agrupamentos lógicos de atributos, os quais facilitam o processo de selecção e análise dos dados, para além de facilitarem uma das técnicas utilizadas pelos algoritmos de DM, a indução de regras, através da generalização (roll up) ou especialização (drill down) dos conceitos ao longo da hierarquia.

Regras do domínio

As regras do domínio são de grande utilidade no processo de descoberta de conhecimento, uma vez que permitem descrever relações conhecidas entre atributos, assim como de…nir as técnicas de DM mais apropriadas para a realização de determinada tarefa. Tradicionalmente utilizadas na tomada de decisão, a sua explicitação facilita ainda o processo de descoberta de conhecimento [Anand et al., 1995].

4.1.4 Di…culdades encontradas no processo de DCBD

A exploração de dados armazenados em BD reais adiciona algumas di…culdades aos algoritmos de DM, uma vez que os mesmos têm de lidar com problemas existentes nos dados ou com a falta destes (dados). Entre as di…culdades mais usuais encontra-se a informação insu…ciente (originada pelo tamanho da amostra ou pelos atributos disponíveis), e ainda os dados corrompidos, caracterizados por possuírem ruído ou estarem incompletos [Holsheimer e Kersten, 1994]. Nas próximas subsecções são analisados os principais problemas associados a cada um destes grupos de di…culdades.

Informação insu…ciente

Os SDC procuram regras nos dados, as quais são construídas baseadas nos dados armazenados na BD analisada. A qualidade e disponibilidade destes dados, in‡uencia os resultados que podem ser encontrados. As principais limitações encontradas ao nível da informação insu…ciente são:

- Informação incompleta. A construção de regras para classi…cação dos dados está condicionada pelo valores disponíveis na BD, e a partir dos quais as várias classes são de…nidas. Contudo, e principalmente no contexto de grandes BD, existem valores de atributos que são desconhecidos e que passam a não ser considerados na determinação dos limites das classes. Esta situação faz com que nem sempre seja possível construir regras que classi…quem correctamente as amostras. Neste casos, as regras indicam a probabilidade de uma dada entidade pertencer a determinada classe.

- Dados dispersos. A determinação dos limites das classes, na construção de regras de classi…cação, está condicionada pelos valores veri…cados na BD para um dado atributo. O facto de poderem existir muitos valores para o referido atributo, dados dispersos, faz com que a determinação das regras seja grandemente di…cultada, não só pela determinação dos limites, com também do próprio número de classes a considerar. Como consequência, os limites das classes não podem ser determinados com exactidão, podendo nalguns casos estar mesmo incorrectos.

- Tamanho da amostra. A amostra/BD utilizada no processo de DCBD é normalmente dividida em dois conjuntos de dados: dados de treino, utilizados na construção das descrições (modelos, padrões, ...), e dados de teste, utilizados para veri…car a validade das descrições encontradas[2]. Esta divisão permite veri…car o comportamento das descrições, quando utilizadas para classi…car dados desconhecidos. As BD normalmente exploradas pelos algoritmos de DM são bastante extensas, facilitando a divisão da amostra nos dois conjuntos de dados necessários. Nem sempre este requisito é veri…cado, implicando a construção de dois conjuntos de dados de tamanho reduzido, que limitam as capacidades dos algoritmos na identi…cação dos modelos, assim como a avaliação dos mesmos. O facto das BD estarem constantemente a ser alteradas adiciona novas di…culdades, já que as descrições encontradas anteriormente podem tornar-se inconsistentes. O ajustamento das regras deverá ser periodicamente realizado, no sentido de estas permanecerem válidas. Regras inconsistentes podem ser eliminadas e geradas novas regras, ou então pode ser utilizada uma abordagem mais e…ciente, que passa pela aprendizagem incremental, na qual as regras já conhecidas são utilizadas no processo de aprendizagem que conduz à sua reformulação/actualização, por forma a eliminar as inconsistências.

[2] Os modelos construídos em exercícios de DM podem veri…car dois tipos de problemas, tradicionalmente denominados de sobre-ajustamento (over…tting) e sub-ajustamento (under…tting) [Elder e Pregibon, 1996]. O sobre-ajustamento do modelo de dados ocorre quando o modelo gerado utiliza particularidades dos dados, na previsão de resultados. Ocorre normalmente quando o conjunto de dados de treino apresenta uma dimensão reduzida, produzindo bons resultados com este conjunto, mas não conseguindo modelar dados desconhecidos (capacidade reduzida de previsão). O sub-ajustamento ocorre quando o modelo gerado é demasiado genérico, não realçando particularidades interessantes nos dados.

Dados corrompidos

É usual surgirem erros nos dados armazenados nas BD organizacionais, os quais introduzem ruído no processo de descoberta de conhecimento. Além dos erros, é com alguma frequência que são encontrados atributos com grande parte dos seus valores omissos, isto é, dados que não foram preenchidos pelo utilizador. Estes dois casos originam problemas diferentes, que são normalmente tratados da forma de seguida descrita.

- Ruído. O ruído ocasiona problemas de dois tipos, que se re‡ectem na:
  - construção de modelos a partir de amostras com ruído;
  - utilização destes modelos na classi…cação de dados com ruído.

No primeiro caso, o sistema deverá ser alertado para o facto da amostra utilizada possuir ruído, o que permitirá aos algoritmos utilizados aplicarem as estratégias adequadas a esta situação[3], gerando descrições que tentam ultrapassar estas falhas. O ruído nos dados exerce uma considerável in‡uência negativa na construção dos modelos, já que a determinação dos limites das classes, por exemplo, deixa de ser a mais apropriada, diminuindo o desempenho das regras quando utilizadas na classi…cação de novos dados. A identi…cação e correcção destas falhas deverá, sempre que possível, ser efectuada no conjunto de dados de treino, permitindo a construção de descrições mais correctas.

No segundo caso, as descrições obtidas podem ser utilizadas na classi…cação de dados com ruído, já que as mesmas apresentam desempenhos superiores quando comparadas com descrições geradas a partir de amostras sem ruído, e que são utilizadas para classi…car dados com ruído.

- Valores omissos. Os valores desconhecidos para um dado atributo podem ser:
  - eliminados da amostra, retirando os registos em causa da BD;
  - substituídos, através da construção de descrições que permitam prever o valor do atributo em falta, partindo dos valores dos outros atributos da amostra [Quinlan, 1986]. Esta aproximação permite preencher os dados desconhecidos, sendo o conjunto de dados resultante utilizado para construir as descrições;
  - etiquetados com uma marca, por exemplo "desconhecido", originando um novo valor para o atributo, que é desta forma considerado na construção das descrições.

4.2 Data Mining

DM consiste na procura de relacionamentos e padrões que existem em grandes BD, mas que estão escondidos no elevado volume de dados armazenado. Estes relacionamentos representam conhecimento acerca da BD explorada e das entidades nela contidas. Decidir se os padrões

[3] Quinlan [Quinlan, 1986] utiliza técnicas estatísticas para decidir se um dado valor de um atributo deve ser ou não considerado, isto é, se deve ser incluído numa árvore de decisão. Os fundamentos utilizados pelo autor assentam na ideia de que pequenos conjuntos de dados, com valores que constituem excepções, são considerados como tendo sido originados por ruído, e como tal devem ser ignorados.

encontrados re‡ectem ou não conhecimento útil, é uma das fases do processo de DCBD, na qual a participação do utilizador é normalmente requerida [Fayyad et al., 1996a]. Neste trabalho, o DM engloba o processo exaustivo de análise de grandes quantidades de dados, com o objectivo de inferir automaticamente modelos e regras que representam conhecimento implícito acerca dos dados analisados.

A metodologia utilizada pelos algoritmos de DM permite encontrar descrições lógicas ou matemáticas, eventualmente de natureza complexa, de padrões num conjunto de dados [Decker e Focardi, 1995]. Existem duas formas básicas de funcionamento destes algoritmos. A primeira abordagem baseia-se na aprendizagem supervisionada, na qual a participação do utilizador é veri…cada na especi…cação dos atributos de interesse e nas classes a utilizar no processo de construção das regras. Entre as técnicas mais utilizadas neste tipo de aprendizagem encontram-se as redes neuronais e os algoritmos de indução de regras (as diversas técnicas encontram-se descritas na subsecção 4.2.3). A segunda abordagem diz respeito a aprendizagem não supervisionada, na qual os padrões são encontrados partindo apenas de uma caracterização lógica do tipo de resultado pretendido. Nesta abordagem não é conhecida a estrutura dos padrões, nem especi…cados quaisquer atributos ou classes. A aproximação de vizinhanças, que permite a segmentação (clustering) dos dados, representa uma das técnicas mais utilizadas neste tipo de abordagem.

Um dos principais problemas com que se deparam os algoritmos de DM é que o número de relacionamentos que podem ser encontrados é extremamente elevado, ocultando por vezes os mais importantes. Como tal, as estratégias de pesquisa têm de ser inteligentes, pelo que são normalmente baseadas em técnicas de aprendizagem automática (Machine Learning) [Holsheimer e Siebes, 1994].

4.2.1 Dedução, Indução e Data Mining

As BD constituem repositórios de dados que devem permitir, entre outras operações, a pesquisa e…ciente da informação armazenada. Em alguns casos, a informação seleccionada não representa uma cópia exacta dos dados armazenados, constituindo muitas vezes informação que é inferida da BD. A inferência de informação pode ser efectuada através de dedução e de indução [Holsheimer e Siebes, 1994]. A:

- Dedução permite inferir informação que é uma consequência lógica da informação armazenada na BD. A maioria dos SGBD disponibilizam operadores que permitem a dedução de informação, permitindo a construção de novo conhecimento a partir do conhecimento inicial (por exemplo, a junção de duas tabelas, permite inferir a relação existente entre duas entidades, a partir da relação existente entre estas e uma terceira entidade)[4]. As inferências dedutivas permitem derivar uma conclusão lógica, a partir de determinado conjunto de factos ou regras [CT113, 1999].
- Indução permite inferir informação que constitui uma generalização da informação exis-

[4] As BD dedutivas permitem a dedução de novos factos a partir de um dado conjunto de factos armazenado na BD [Goh et al., 1996]. Estas BD possuem mecanismos de dedução que se revelam de particular importância sempre que factos necessários não são conhecidos. Contudo, os mecanismos de dedução utilizados podem gerar grandes quantidades de factos. Goh et al. [Goh et al., 1996] apresentam uma abordagem ao processo de descoberta de conhecimento em BD dedutivas, na qual é suprimida a necessidade de armazenamento dos novos factos deduzidos pela BD dedutiva.

> tente na BD, conduzindo à construção de regras que descrevem propriedades dos objectos analisados. Basicamente, a indução permite construir a descrição de um conceito, a partir do qual todas as instâncias positivas possam ser derivadas, mas nenhuma das instâncias negativas o seja. As regras assim construídas permitem prever o valor de um atributo, baseadas nos valores de outros atributos conhecidos[5]. A inferência indutiva deriva conceitos gerais a partir de um dado conjunto de factos ou regras [CT113, 1999].

A inferência de informação em BD está fortemente condicionada pelo tamanho das mesmas, requerendo a utilização de SGBD para a sua execução. Contudo, estes sistemas permitem a dedução de informação, não estando capacitados para as tarefas de indução. As ferramentas de DM permitem a execução da indução, facilitando a construção de regras que re‡ectem conhecimento implícito existente na BD explorada.

Sistemas cognitivos percebem o ambiente que os rodeia, construindo uma simpli…cação do mesmo, normalmente denominada modelo. A construção deste modelo é baseada em técnicas de aprendizagem indutiva (Inductive Learning)[6]. Durante a fase de aprendizagem, o sistema cognitivo observa o seu ambiente e reconhece semelhanças entre os objectos e os eventos que o caracterizam. Posteriormente, agrupa objectos similares em classes, construindo regras que prevêem o comportamento dos mesmos [Holsheimer e Siebes, 1994].

A aprendizagem indutiva pode ser implementada através de mecanismos supervisionados ou não (aprendizagem supervisionada e aprendizagem não supervisionada), constituindo uma área largamente explorada pela aprendizagem automática.

4.2.2 Tarefas de Data Mining

As tarefas associadas ao DM [Fayyad et al., 1996a] agrupam-se em dois grandes grupos: descrição e previsão. A descrição permite encontrar regras que descrevem ou caracterizam o comportamento dos dados analisados. A previsão utiliza determinadas variáveis da BD para prever o valor de outros atributos de interesse[7]. Esta distinção está relacionada com o objectivo do exercício de DM, que pode permitir aumentar o conhecimento do utilizador acerca dos dados, descrição, ou automatizar o processo de tomada de decisão, previsão, através da construção de modelos capazes de estimar um valor [Berry e Lino¤, 2000]. Apesar da complexidade que pode estar associada ao processo de DCBD, as tarefas tradicionalmente executadas pelos algoritmos de DM incluem a:

- Classi…cação. A classi…cação consiste no enquadramento dos dados, armazenados na BD explorada, dentro de classes prede…nidas. O modelo de classi…cação construído permite

[5] A indução de informação a partir de BD tem motivado grandemente a investigação na área das BD indutivas [Bergadano, 1993]. Nestas BD, parte dos registos que pertencem a uma dada classe são armazenados como fazendo parte dos casos positivos, enquanto que registos que não pertencem à referida classe são armazenados como constituindo parte dos casos negativos. A partir dos casos positivos e negativos é possivel induzir as regras que descrevem os dados armazenados.

[6] Dzeroski e Lavrac [Dzeroski e Lavrac, 1993] apresentam um sistema de aprendizagem indutiva aplicado a BD dedutivas.

[7] No caso da previsão, os dados são classi…cados atendendo ao que se espera seja o seu comportamento futuro. A precisão dos modelos obtidos para a previsão apenas é testada decorrido algum tempo sobre a classi…cação, o que permite veri…car o desvio ocorrido entre o valor real e o valor previsto [Berry e Lino¤, 2000].

determinar em que classe determinado elemento se enquadra. Este modelo pode posteriormente ser utilizado na previsão da classe a que pertencem registos ainda não classi…cados [Chen et al., 1996]. A construção dos agrupamentos, processo que permite encontrar propriedades comuns em conjuntos de objectos, é efectuada veri…cando a variável, ou variáveis, de…nidas pelo utilizador [Agrawal et al., 1993a] e as classes a considerar. A classi…cação lida com valores discretos. Sempre que a variável a classi…car possuir valores contínuos, a tarefa de DM apresenta a designação de estimação[8], permitindo enquadrar determinado atributo num valor real previsível.

- Segmentação. A segmentação (clustering) constitui uma tarefa não supervisionada, já que o utilizador não exerce qualquer in‡uência na mesma, que permite encontrar um conjunto …nito de classes ou segmentos que classi…cam os dados analisados. Basicamente, permite enquadrar os dados em diversas classes, determinadas a partir dos dados existentes, ao contrário da classi…cação onde as mesmas são prede…nidas pelo utilizador. Os segmentos são de…nidos veri…cando agrupamentos naturais que são detectados nos dados e que obedecem normalmente a métricas de similaridade [Chen et al., 1996].

- Sumariação. A sumariação permite descrever determinado subconjunto de dados, fornecendo descrições resumidas do mesmo. Estas descrições são obtidas através da generalização dos dados, permitindo uma descrição sumária dos mesmos [Chen et al., 1996]. Entre os exemplos mais simples de sumariação, encontra-se a determinação da média ou desvio padrão de uma amostra [Fayyad et al., 1996a].

- Associação. A associação permite determinar relacionamentos entre atributos da BD, veri…cando a correlação que existe entre os mesmos. As correlações encontradas devem satisfazer o nível de suporte e con…ança exigido pelo utilizador [Agrawal et al., 1993a] [Chen et al., 1996].

- Sequenciação. A sequenciação é utilizada para determinar relações temporais em conjuntos de dados que apresentam várias transacções separadas no tempo. É, assim, utilizada na análise de séries temporais e tem como objectivo modelar os desvios e evolução dos seus registos ao longo do tempo [Agrawal et al., 1996]. Regras de sequência podem ser vistas como um caso particular de regras de associação, nas quais o(s) antecedente(s) e o consequente de uma dada regra estão relacionados por componentes temporais.

4.2.3 Técnicas de Data Mining

Existe uma grande variedade de técnicas de Data Mining [Fayyad et al., 1996b]. Neste trabalho, as técnicas (descritas nas próximas subsecções) são agrupadas em quatro grandes grupos: indução de regras (que inclui as árvores de decisão e as regras de associação), redes neuronais, algoritmos genéticos e aproximação de vizinhanças. A escolha da técnica a utilizar

[8] A regressão constitui uma das principais técnicas de previsão [Berry e Lino¤, 2000]. A sua forma mais comum, regressão linear, combina numa equação todas as variáveis de entrada (variáveis independentes) por forma a determinar o valor da variável de saída (variável dependente). Na regressão, os dados são analisados de uma forma geométrica. Os valores dos atributos considerados são utilizados para de…nir pontos no espaço. A equação de regressão descreve a linha que melhor se adapta a estes dados, isto é, a que minimiza a distância média dos pontos à linha. A regressão, disponibilizada como técnica em diversas ferramentas de DM, apenas pode ser utilizada com dados numéricos e é muito sensível à distribuição apresentada pelos dados.

in‡uencia a ‡exibilidade de representação do modelo encontrado, condicionando a interpretação do mesmo por parte do utilizador …nal. Entre os modelos de representação mais populares, devido à clareza de exposição dos resultados, encontram-se as árvores de decisão e as regras de associação.

Indução de regras

Na indução de regras, que permite gerar árvores de decisão ou regras de associação, os registos da BD são tratados como regras, que são sucessivamente generalizadas por forma a resumirem o conteúdo da BD [Rainsford e Roddick, 1996]. Entre as vantagens da utilização desta técnica encontra-se: i) a facilidade de interpretação dos resultados, apresentados em regras de fácil compreensão; ii) a facilidade de incorporação de conhecimento do domínio, explícito em regras, no processo de descoberta de conhecimento; e iii) a facilidade de armazenamento das regras encontradas numa base de conhecimento.

As regras obtidas podem ser derivadas utilizando indução top-down ou bottom-up. Na aproximação bottom-up o sistema inicia o processo de aprendizagem considerando todos os registos da BD como regras, as quais são posteriormente generalizadas. Na estratégia top-down o processo é iniciado partindo de conceitos gerais, que descrevem os dados, os quais são posteriormente re…nados por um processo de especialização.

Árvores de Decisão. São constituídas por estruturas em árvore que representam um conjunto de decisões. Possuem uma representação simples, sendo facilmente interpretadas pelos utilizadores. Os algoritmos de indução de árvores permitem gerar regras de classi…cação dos dados, baseados na informação armazenada na BD. Para além de poderem lidar com grandes quantidades de dados, permitem utilizar directamente o resultado/conhecimento nelas explícito [Russell e Norvig, 1995] [Adriaans e Zantinge, 1996].

Numa árvore de decisão, nos nodos das árvores encontram-se os atributos a classi…car, enquanto que os ramos descrevem os valores possíveis para esses atributos. As folhas da árvore agrupam as diversas classes em que cada registo pode ser classi…cado [Rainsford e Roddick, 1996]. A Figura 4.4 apresenta uma árvore de decisão, que caracteriza o comportamento de uma instituição que fornece crédito para a aquisição de bens de consumo. Pela análise da referida árvore é possível identi…car os casos em que a organização concede o crédito, atendendo ao per…l dos clientes.

As árvores de decisão podem ainda ser representadas por conjuntos de regras. Cada folha da árvore dá origem a uma regra, sendo o seu conteúdo apresentado na parte que diz respeito à consequência (resultado). A parte antecedente da regra constitui uma conjunção de valores, respeitantes aos atributos existentes no ramo que liga esta folha à raiz da árvore. No exemplo apresentado anteriormente na Figura 4.4, as regras associadas à concessão de crédito são:

Se Bem Financiado = ”Electrodoméstico” e Estado Civil = ”Solteiro”
Então Conceder

Se Bem Financiado = ”Móveis”
Então Conceder

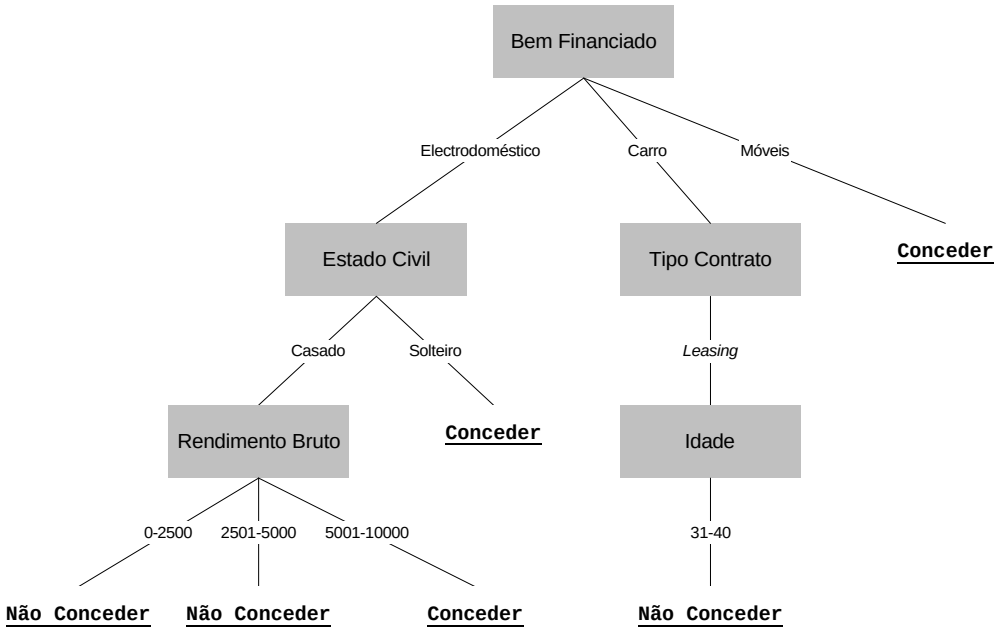


Figura 4.4: Árvore de decisão

Se Bem Financiado = "Electrodoméstico" e Estado Civil = "Casado" e
Rendimento Bruto = "5001-10000"
Então Conceder

Esta técnica, além de fornecer resultados compreensíveis pelo utilizador, apresenta ainda a vantagem de incluir na árvore apenas aqueles atributos que são realmente importantes no processo de tomada de decisão. Tal permite seleccionar de todos os atributos considerados, aqueles que efectivamente contribuíram para a determinação das classes. Esta identi…cação revela-se de particular importância quando se pretende treinar uma rede neuronal (com custos computacionais elevados), já que permite, através da selecção dos atributos relevantes, reduzir o espaço de pesquisa.

Regras de Associação. A descoberta de regras de associação em grandes BD foi inicialmente analisada por Agrawal et al. [Agrawal et al., 1993b], com o objectivo de determinar regras que relacionem uma conclusão (por exemplo, a compra de um produto) com um conjunto de condições (por exemplo, a compra de outros produtos).

As regras de associação [Adriaans e Zantinge, 1996] permitem encontrar relacionamentos entre os atributos existentes numa BD, representado-os na forma duma regra Se X Então Y ou $X \Rightarrow Y$. Dada uma BD T, que armazena um conjunto de atributos A, tal que $A = A_1, A_2, \ldots, A_n$, uma regra de associação existente em T é representada por $X \Rightarrow Y$, sendo X um subconjunto de atributos de A. Y representa um único atributo de A, não presente em X [Agrawal et al., 1993b]. A regra $X \Rightarrow Y$ é veri…cada no conjunto das transacções de T com um factor de con…ança c, $0 \cdot c \cdot 1$, se c% das transacções de T que satisfazem X também satisfazem Y

[Mohan, 2000]. O suporte de uma regra indica o número de transacções de T que suportam a regra, isto é, o subconjunto de registos que satisfazem a união dos atributos que integram a parte antecedente (X) e consequente (Y) da regra. Estas duas medidas de interesse, das regras encontradas, permitem conhecer a força de uma regra, con…ança, e a sua signi…cância estatística, suporte. Devem ser analisadas simultaneamente, uma vez que o suporte pode ser elevado (percentagem de registos que satisfazem a regra), e a regra possuir uma associação fraca, con…ança pequena, sendo diminuto o número de registos em que é possível prever Y, conhecendo X.

A regra de associação "pão & manteiga =) leite (173:17%, 0.84)" indica que 173 registos veri…cam pão, manteiga e leite (ou seja, estes produtos são comprados simultaneamente), representando 17% da população analisada. A con…ança da regra, representando o grau de certeza da mesma, indica que 84% dos clientes que compram pão e manteiga também compram leite.

No contexto das actuais ferramentas de DM, as regras de associação têm interesse se o utilizador tiver uma ideia, ainda que vaga, daquilo que procura. Isto porque não existe nenhum algoritmo que pesquise automaticamente (sem qualquer indicação do utilizador) relacionamentos de interesse na BD [Adriaans e Zantinge, 1996]. A procura de regras de associação, com restrições sintáticas, permite direccionar a procura para regras que veri…quem um dado A_i em X, um dado A_j em Y, ou ainda, uma conjunção das duas situações anteriores [Agrawal et al., 1993b].

Nesta técnica, os algoritmos começam por gerar um conjunto inicial de regras, que são posteriormente validadas, através da sua apresentação ao conjunto de dados. Este procedimento permite a especialização do conjunto inicial, através da introdução de novas condições nas regras. As regras obtidas são iterativamente validadas, até que veri…quem um nível de con…ança e de suporte mínimo (c_{min} e s_{min}) especi…cado pelo utilizador ou prede…nido no algoritmo utilizado [Han e Kamber, 2001].

Redes Neuronais

Uma rede neuronal consiste numa "rede de unidades elementares de processamento ligadas por conexões ponderadas com pesos ajustáveis, na qual cada unidade produz um valor pela aplicação de uma função não linear aos seus valores de entrada, e o transmite a outras unidades ou o apresenta como uma saída" ([CT113, 1999] p. 6).

Redes neuronais são sistemas de classi…cação modelados segundo os princípios do sistema nervoso humano. São compostas por um conjunto de unidades, organizadas em níveis. As diversas unidades (nodos) encontram-se conectadas através de ligações, as quais têm associado um determinado peso. Os diversos nodos que constituem uma rede, encontram-se agrupados em três grandes grupos: Nodos de entrada encarregues de receber os dados a analisar, nodos de saída que transmitem os sinais à saída da rede e um número ilimitado de níveis intermédios que contêm os nodos intermédios (Figura 4.5) [Adriaans e Zantinge, 1996] [Lu et al., 1996].

Existem dois estágios distintos na utilização destas redes. O primeiro diz respeito à aprendizagem, no qual a rede é treinada para a execução de determinada tarefa. A segunda fase prende-se com a previsão, na qual a rede é utilizada para classi…car registos desconhecidos [Adriaans e Zantinge, 1996].

Existem várias arquitecturas para as redes neuronais, as quais diferem essencialmente

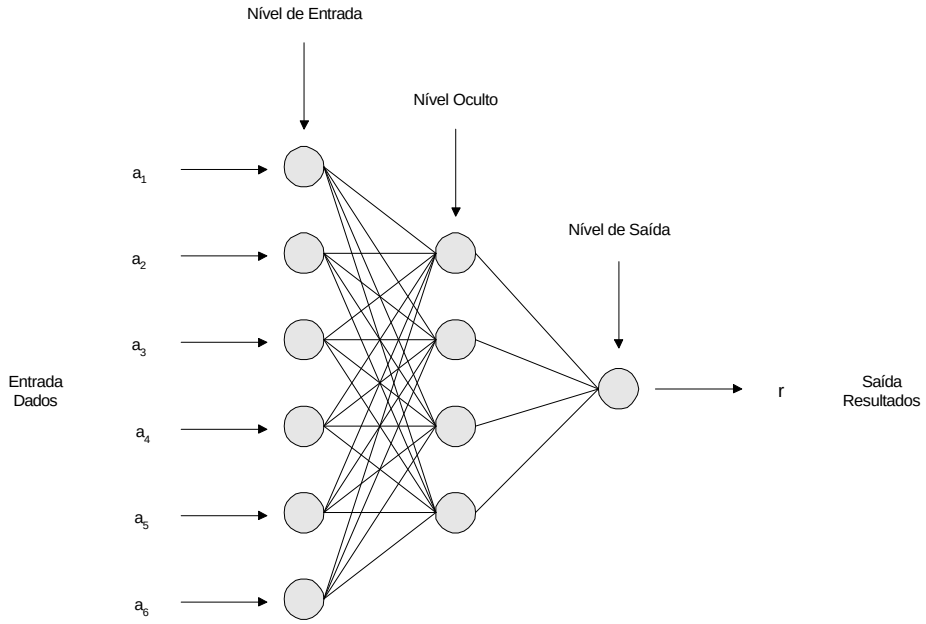


Figura 4.5: Con…guração de uma rede neuronal

no número de níveis intermédios[9] permitidos. Nas redes do tipo perceptrão (perœptron) não existe qualquer nível intermédio (apenas o de entrada e o de saída), o que torna o processo de aprendizagem mais simples, mas ao mesmo tempo condiciona o tipo de tarefa em que podem ser utilizadas[10]. As redes que apresentam um ou mais níveis intermédios são denominadas de perceptrão multi-nível [Russell e Norvig, 1995], permitindo aproximar qualquer função não-linear.

A aprendizagem de uma rede é iniciada pela atribuição de pesos semelhantes a todas as ligações da rede. A rede é posteriormente treinada com um conjunto de dados, denominado de dados de treino. Em cada iteração do processo de aprendizagem, a saída da rede é comparada com a saída desejada (explícita nos casos conhecidos armazenados no conjunto de dados de treino). O resultado desta comparação é propagado na rede, sendo os pesos das ligações gradualmente alterados. À medida que a aprendizagem progride, a rede …ca cada vez mais precisa na replicação dos resultados conhecidos. Este processo é repetido até que a rede apresente níveis de desempenho apropriados (normalmente já prede…nidos nos algoritmos, mas que podem ser especi…cados pelo utilizador). Uma vez estável a estrutura da rede, dá-se por terminado o processo de aprendizagem, encontrando-se a rede treinada e pronta para classi…car casos desconhecidos.

Estas redes necessitam de um conjunto mínimo[11] de dados de treino durante o processo de

[9] Também conhecidos por níveis ocultos, dado não poder ser observado o seu estado.

[10] Estando a sua utilização restrita a problemas resolvidos por funções lineares.

[11] Numa rede neuronal com um nível oculto e um nodo de saída, o número de pesos da rede é dado pela fórmula h * (n + 2) + 1, onde h representa o número de unidades ocultas e n o número de unidades de entrada. Cada ligação entre as unidades tem associado um peso. Se existirem mais pesos do que instâncias disponíveis para o treino da rede, existe uma grande problabilidade da rede memorizar o conjunto de dados de treino (sobre-ajustamento). Como regra geral, deverão existir no mínimo 10 instâncias de treino por cada peso na rede, sendo desejável ter

aprendizagem. Possuem a desvantagem de não transmitirem a sua aprendizagem (modelo) num formato perceptível ao utilizador. Podem ser tratadas como "caixas negras" que dão respostas, mas que não transmitem qualquer conhecimento acerca do processo que conduziu à obtenção das mesmas.

Na década de 80 surge uma nova versão das redes neuronais, conhecida como redes auto-organizáveis ou redes Kohonen, Kohonen self-organizing maps [Kohonen, 1989] (Figura 4.6). Estas redes são constituídas por um conjunto de nodos que se encontram directamente ligados a todos os nodos vizinhos. São normalmente representadas por estruturas bi-dimensionais, nas quais cada nodo tem associada uma determinada posição física na estrutura. Inicialmente, cada nodo possui uma posição aleatória, a qual é sucessivamente ajustada durante a fase de aprendizagem. Esta rede pode ser vista como uma estrutura auto-ajustável, que se organiza por forma a considerar todos os casos da amostra. São extremamente úteis em tarefas de segmentação, nas quais não são conhecidas as classes (clusters).

As redes Kohonen não possuem níveis intermédios. O número de nodos no nível de entrada é calculado em função dos atributos de entrada, sendo o número de nodos no nível de saída igual ao número de clusters obtidos na aprendizagem. Nesta fase, cada nodo de saída compete com os outros nodos para ganhar a classi…cação de um dado registo, sendo os pesos de ligação ajustados em função do sucesso ou insucesso de cada nodo. O processo de modelação da rede conduz ao agrupamento de nodos em vectores, que após a aprendizagem representam as classes encontradas [Lobo e Pires, 1998]. Os pesos obtidos para as ligações permitem veri…car a in‡uência que cada atributo teve na identi…cação das classes.

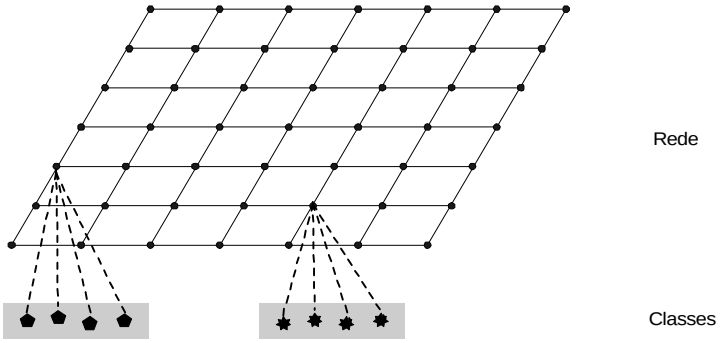


Figura 4.6: Rede neuronal do tipo Auto-organizáveis

Algoritmos Genéticos

A técnica dos algoritmos genéticos conjuga princípios da biologia e das ciências da computação. Na Origem das Espécies, Darwin descreve a teoria da evolução recorrendo à selecção natural. Cada espécie possui um número elevado de indivíduos, e num instinto de sobrevivência, aqueles que sobrevivem são os que melhor se adaptam ao ambiente. As alterações do ambiente

este número de instâncias por cada valor que a variável de saída possa tomar. Numa rede com 50 variáveis de entrada e 10 nodos ocultos, existem 521 pesos, pelo que deverão existir pelo menos 5210 instâncias de treino. No caso de uma saída binária, é apropriado um conjunto de treino com 10420 instâncias [Berry e Lino¤, 2000].

provocam mutações genéticas nas espécies, por forma a estas se adaptarem às alterações do seu meio ambiente [Santos, 2000].

É a partir dos princípios da mutação genética, veri…cando os mecanismos que permitem a transmissão hereditária de informação, que são de…nidos os princípios dos algoritmos genéticos [Adriaans e Zantinge, 1996].

Tradicionalmente, os algoritmos genéticos iniciam o processo de evolução com um conjunto de regras, as quais são submetidas a operadores de selecção e reprodução, por forma a desenvolverem regras mais apuradas. Estas regras são etiquetadas com determinado valor de utilidade (…tness), que facilita a selecção das mesmas. Este valor é calculado a partir de uma função de avaliação, de…nida para o domínio de aplicação em causa, que recebe uma regra como entrada e determina a saída, que representa o valor numérico que retrata a sua utilidade [Russell e Norvig, 1995].

Uma regra é representada por um conjunto …nito de caracteres (genes), normalmente restritos ao alfabeto binário (0, 1)[12]. A estratégia de selecção é na maioria das vezes aleatória, sendo a probabilidade de selecção proporcional à utilidade. A reprodução é efectuada através de cruzamento e mutação. O cruzamento consiste em construir pares aleatórios de regras, no conjunto das que foram seleccionadas para reprodução. Para cada par da ascendência, é escolhido aleatoriamente um ponto N de cruzamento. Este ponto N de…ne que na reprodução, a primeira regra que surge como descendente deste par é obtida seleccionando os genes 1 a N da primeira regra, e os restantes genes da segunda regra. A segunda regra, que surge ainda como descendente deste par, é obtida seleccionando os genes 1 a N da segunda regra da ascendência, e os restantes genes da primeira. A última fase do processo diz respeito à mutação, na qual um dado gene pode ser casualmente alterado para outro caracter do alfabeto. Este procedimento é resumidamente apresentado na Figura 4.7.

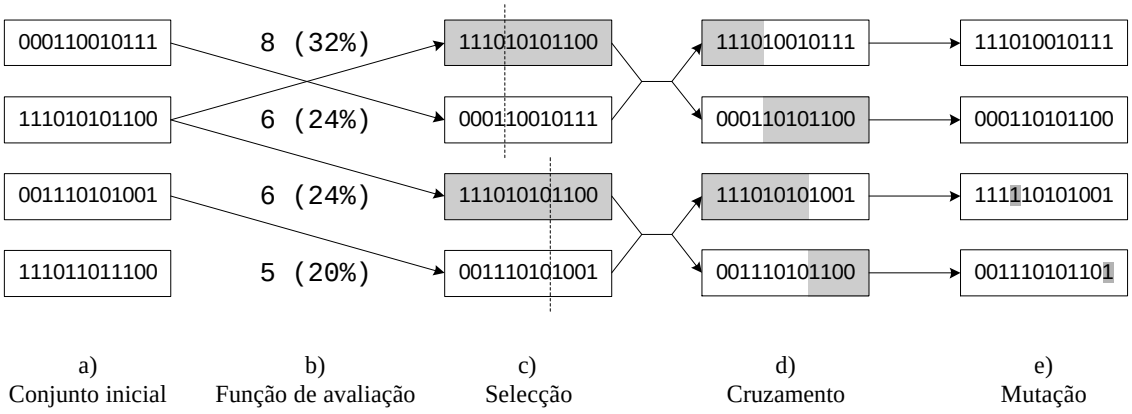


Figura 4.7: Modo de operação dos algoritmos genéticos. a) conjunto inicial com 4 regras; b) classi…cação das regras segundo a função de avaliação: a primeira regra foi classi…cada com 8, o que signi…ca que apresenta a probabilidade de 32% de ser seleccionada; c) construção dos pares e de…nição do ponto de cruzamento; d) regras produzidas por cruzamento, e e) mutação de dois caracteres (Adaptado de: [Russell e Norvig, 1995] p. 621).

[12] Podendo, contudo, existir alfabetos com mais valores.

Aproximação de vizinhanças

A técnica de aproximação de vizinhanças (nearest neighbour) é baseada no princípio de que registos semelhantes estão próximos uns dos outros, quando analisados numa perspectiva espacial. A veri…cação da localização dos registos, interpretados como pontos no espaço, permite a identi…cação de regiões, denominadas classes (ou segmentos), que de…nem características comuns para os registos que representam [Adriaans e Zantinge, 1996]. A complexidade desta técnica aumenta à medida que cresce o número de registos a analisar, já que cada registo é comparado com os restantes registos da amostra.

A implementação desta técnica passa pela construção de partições, dos objectos armazenados na BD, num conjunto de k classes, sendo k um parâmetro de entrada. Cada classe pode ser representada pelo seu centro de gravidade (estratégia k-means), isto é, pela localização média de todos os membros do segmento, ou por um dos objectos da classe próximo do seu centro (estratégia k-medoid) [Ester et al., 1998b].

Para determinar as classes, cada registo é transformado num ponto no espaço, apresentando este tantas dimensões quantos os atributos em análise. O valor de cada campo é interpretado como a distância da origem até a sua localização num dado eixo [Berry e Lino¤, 2000].

O processo de obtenção das classes é iniciado com centróides em posições aleatórias, as quais são optimizadas iterativamente através da movimentação dos centros. No caso do algoritmo k-medoid, as classes são inicialmente representadas por objectos aleatórios, que vão sendo substituídos por outros objectos, se a qualidade do segmento for melhorada. As estratégias utilizadas para a construção das classes veri…cam a distância média dos objectos ao centro ou objecto que representa a classe (Figura 4.8).

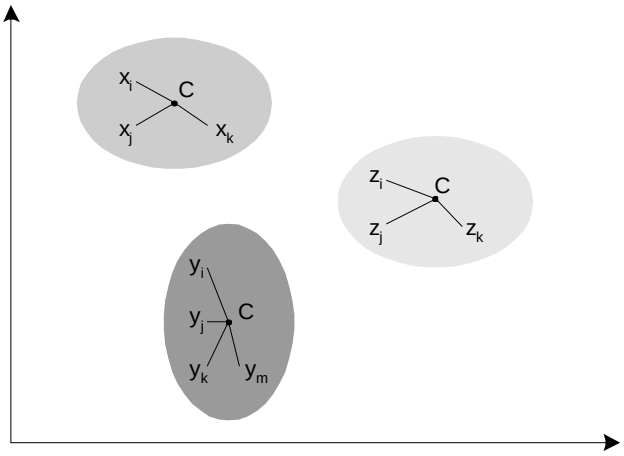


Figura 4.8: Partição dos objectos em classes

4.2.4 Síntese

Depois de apresentadas as principais tarefas associadas ao processo de DM, e ainda as principais técnicas utilizadas na sua implementação, julga-se conveniente sistematizar as descrições efectua-

Técnicas / Tarefas	Classificação	Segmentação	Sumariação	Associação	Sequenciação
Árvores de decisão	X		X		
Regras de associação				X	X
Redes neuronais	X	X			
Algoritmos genéticos	X		X		
Aprox. vizinhanças		X	X		

Tabela 4.1: Tarefas e técnicas de DM

das, referindo que técnicas podem ser utilizadas na execução de determinada tarefa. A Tabela 4.1 apresenta um quadro resumo com as tarefas e técnicas anteriormente descritas, destacando a possibilidade, ou não, de utilização de uma dada técnica numa dada tarefa.

4.3 A descoberta de conhecimento em bases de dados espaciais

Uma caso particular da DCBD diz respeito à exploração de dados geo-referenciados, 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 (posições relativas, adjacências, distâncias, direcções, etc.) e da influência que esta componente exerce nos restantes dados explorados. A principal diferença existente entre a análise da dados espaciais e dados não espaciais está associada ao facto das entidades geográficas endereçadas poderem ser afectadas por características de entidades vizinhas. A influência mútua que duas entidades exercem entre si depende de factores como a topologia, a distância e a direcção existente entre as mesmas [Ester et al., 1999b].

A semântica associada à localização dos factos, a análise dessas localizações com o objectivo de perceber o seu porquê, faz com a utilidade do conhecimento obtido, no processo de descoberta de conhecimento, seja largamente melhorada através da integração de dados espaciais e dados não espaciais [Dey e Roberts, 1996].

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 BDE [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.

4.3.1 Principais tarefas e abordagens

As tarefas tradicionalmente [Koperski et al., 1996] [Ester et al., 1998a] [Han e Kamber, 2001] associadas ao processo de DCBDE 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 veri…cada nos restantes registos da BD.

- a veri…cação 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 geográ…cas (uma regra discriminante compara, por exemplo, a variação de preços nas habitações em diversas regiões geográ…cas).

- o estabelecimento de relações entre dados espaciais, e entre dados espaciais e dados não espaciais ! associação espacial. A associação espacial permite identi…car a relação que existe entre um conjunto de objectos espaciais e um conjunto de dados não espaciais, ou entre dois conjuntos de dados espaciais, de…nindo 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.

- a veri…cação 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, veri…cada 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 veri…cados pelos atributos permite determinar tendências espaciais nos dados[13].

As próximas subsecções apresentam as principais abordagens à DCBDE. Como poderá ser constatado pela análise das mesmas, as diversas estratégias descritas baseiam o seu processo de descoberta de conhecimento na execução de generalizações. Encontrar distribuições espaciais nos dados não espaciais, envolve a procura de padrões gerais nos dados e a sua posterior segmentação em grupos. Uma forma de o conseguir é generalizando os dados não espaciais (por exemplo, classi…cando a população em termos a salários baixos, médios e altos) e de…nindo os segmentos espaciais (regiões) que caracterizam as generalizações encontradas. O inverso é também possível, iniciando o processo de generalização nos dados espaciais, ao qual é seguida a de…nição dos segmentos não espaciais que os caracterizam.

A DCBDE baseada em generalizações

Lu et al. [Lu et al., 1993] investigaram o processo de DCBDE baseado em generalizações. A estratégia implementada integra a indução orientada aos atributos não espaciais e a sua posterior caracterização espacial, com a generalização dos dados espaciais.

[13] No caso de todos os atributos a analisar serem numéricos, a regressão linear poderia ser uma das técnicas a utilizar para determinar tendências espaciais nos dados, já que a diferença entre as distâncias pode ser considerada a variável independente, constituindo a diferença entre os atributos a variável dependente [Ester et al., 1998a].

A técnica de indução orientada aos atributos, inicialmente desenvolvida para BD relacionais[14] [Han et al., 1992], é neste trabalho estendida a BDE. O processo de descoberta de conhecimento é iniciado a partir de um pedido do utilizador, o qual redirecciona a procura, e da especi…cação do conhecimento do domínio necessário à sua implementação. Entre o conhecimento requerido, encontra-se a de…nição de hierarquias de conceitos, espaciais e não espaciais. A indução é efectuada por ascensão nas diversas hierarquias, a qual permite, posteriormente, a sumariação dos relacionamentos encontrados entre dados espaciais e dados não espaciais.

O sistema implementado [Lu et al., 1993] permite encontrar relacionamentos nos dados, os quais podem ser utilizados na veri…cação da correlação existente entre diversas características espaciais.

Os dados não espaciais são armazenados em BD relacionais, enquanto que os dados espaciais se encontram armazenados em estruturas próprias (BDE), estando ligados aos dados não espaciais através de uma arquitectura do tipo SAND.

A movimentação pelas diferentes hierarquias é conseguida recorrendo a funções do tipo pai(atributo) e filho(atributo), as quais devolvem o conceito que está imediatamente acima do atributo na hierarquia e o conjunto de conceitos no nível imediatamente inferior, respectivamente.

Dois algoritmos foram implementados: indução orientada aos atributos não espaciais e indução orientada aos atributos espaciais. Em ambos os casos um dos conjuntos de dados é generalizado, sendo o outro conjunto ajustado por forma a caracterizar a generalização obtida.

Indução orientada aos atributos não espaciais

Neste caso, o processo de descoberta de conhecimento é iniciado pela selecção dos atributos não espaciais relevantes à análise (especi…cados pelo utilizador numa questão, que apresenta uma sintaxe similar ao SQL). Estes atributos são posteriormente generalizados até ao nível desejado, atendendo às hierarquias não espaciais inicialmente de…nidas. A agregação de registos similares permite a selecção dos objectos espaciais referenciados pelos mesmos. Estes objectos são posteriormente generalizados, atendendo às hierarquias espaciais, por forma a se obter um número reduzido de conceitos espaciais diferentes, ou até se atingir o nível de generalização pretendido pelo utilizador.

Indução orientada aos atributos espaciais

Esta abordagem permite que a generalização seja iniciada a partir de atributos espaciais. Para tal, seleccionam-se os dados espaciais especi…cados pelo utilizador, os quais são posteriormente generalizados atendendo às hierarquias espaciais de…nidas. Uma vez concluído o processo de generalização de dados espaciais, seleccionam-se os dados não espaciais associados aos dados espaciais generalizados. Estes dados não espaciais são também generalizados, por forma a reduzir o número de conceitos não espaciais diferentes, na regra assim encontrada.

As técnicas desenvolvidas por Lu et al. [Lu et al., 1993], indução orientada a atributos não espaciais e indução orientada a atributos espaciais, evidencia que o processo de DCBDE pode

[14] A versão original, desenvolvida para BD relacionais, solicitava dados relevantes, conhecimento do domínio e tipo de representação para os resultados, na especi…cação da tarefa a realizar [Han et al., 1992].

ser implementado adaptando técnicas disponíveis para BD relacionais. Em ambos os casos, o pedido do utilizador é especi…cado numa questão com a seguinte sintaxe:

```
extract região
from mapa_precipitação
where distrito=''Braga'' and período=''primavera'' and ano=1995
in relevance to precipitação and região
```

Esta questão permite caracterizar a precipitação veri…cada no distrito de Braga, durante a primavera do ano de 1995. Para a sua execução, o sistema deverá conhecer as hierarquias de conceitos referentes à precipitação, região, período, etc. que serão utilizadas nas generalizações a efectuar.

Integração de BD espaciais e BD não espaciais

Uma outra abordagem à DCBDE, proposta por Dey e Roberts [Dey e Roberts, 1996], passa pela integração de BD relacionais e BDE com o propósito de descoberta de conhecimento. Esta integração lógica é conseguida através de duas interfaces (Figura 4.9), uma para aceder a dados espaciais e outra para aceder aos dados não espaciais. Esta arquitectura é justi…cada pelo tamanho das BD exploradas em processos de descoberta de conhecimento[15], e pela necessidade de estruturas especí…cas para o armazenamento da informação espacial e não espacial.

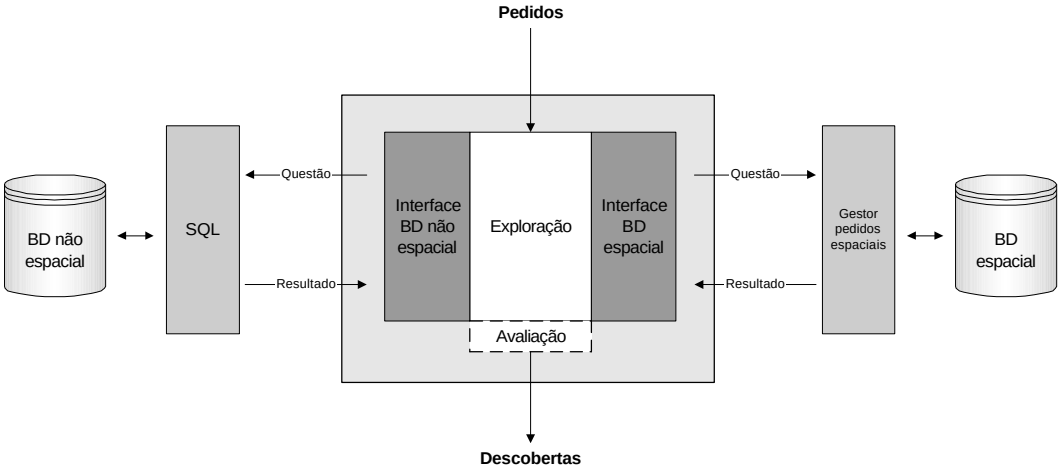


Figura 4.9: Integração de BD espaciais e não espaciais através de interfaces apropriadas (Adaptado de: [Dey e Roberts, 1996])

Na arquitectura apresentada na Figura 4.9 é assumido que a BD não espacial contém, entre os seus atributos, identi…cadores geográ…cos, que especi…cam localizações na BDE. O processo

[15] A possibilidade de construção de um repositório único de dados, para onde são transferidos os dados a analisar pela aplicação de descoberta de conhecimento, torna todo o processo mais e…ciente em termos de velocidade de pesquisa, mas é incorportável quando se analisam grandes conjuntos de dados [Dey e Roberts, 1996].

de descoberta de conhecimento pode ser iniciado a partir de dados não espaciais ou de dados espaciais.

Partindo dos dados não espaciais, identi…cam-se os atributos não espaciais a explorar (especi…cados pelo utilizador ou gerados aleatoriamente), os quais são posteriormente agrupados por forma a construírem classes com características semelhantes. Estas classes são posteriormente analisadas sobre a perspectiva espacial, no sentido de se identi…carem padrões espaciais nas mesmas.

Quando o processo de pesquisa é iniciado a partir de dados espaciais, procede-se à identi…cação de uma característica espacial a investigar (de…nida pelo utilizador ou gerada aleatoriamente), a qual permite a selecção dos identi…cadores geográ…cos associados à região ou característica espacial investigada. Através desta, é ainda possível seleccionar os dados não espaciais associados aos identi…cadores geográ…cos identi…cados.

A abordagem proposta [Dey e Roberts, 1996] evita o armazenamento dos dados não espaciais na BD espacial, ou a necessidade de copiar os relacionamentos espaciais, sob a forma de atributos, para a BD não espacial. Ambas as BD permanecem …sicamente independentes, permitindo a manutenção e alteração das mesmas separadamente.

Não é referido pelos autores o estado de desenvolvimento da arquitectura, ao nível da implementação. Apenas é reforçada a ideia de que algumas questões necessitam ser revistas, como seja a optimização do processo de identi…cação das características interessantes a explorar.

GeoMiner: um protótipo para DCBDE

O GeoMiner, um protótipo para a DCBDE, tem sido desenvolvido como uma extensão ao DBMiner[16] [Han et al., 1994], uma ferramenta de descoberta de conhecimento para BD relacionais [Han et al., 1997].

O protótipo inclui algoritmos que lhe permitem encontrar três tipos de regras: características, discriminantes e de associação, estando prevista na arquitectura do sistema, a integração de duas novas tarefas: classi…cação e segmentação. Utiliza a arquitectura SAND na modelação da BDE e o GMQL (Geo-Mining Query Language) como linguagem proprietária da aplicação, para a especi…cação das questões que conduzem o processo de descoberta de conhecimento. Estas questões devem ser fornecidas pelo utilizador, e devem indicar o tipo de regras a procurar, os atributos a seleccionar e ainda os predicados espaciais a considerar. Uma interface do sistema com o utilizador permite a interacção dos dois intervenientes, e ainda, a visualização dos resultados [Han et al., 1997].

O GeoMiner está a ser construído sobre o DBMiner, sendo este último o responsável pelo tratamento dos dados não espaciais. As tarefas de DM nos dados espaciais são realizadas pelo GeoMiner, o qual gere ainda os relacionamentos existentes entre os dados espaciais e os dados não espaciais (Figura 4.10).

A Figura 4.10 evidencia a arquitectura do GeoMiner, que integra:

- Uma interface grá…ca com o utilizador, que permite a análise dos dados e a visualização dos resultados na forma de mapas, grá…cos, tabelas, etc.

[16] Inicialmente denominado de DBLearn.

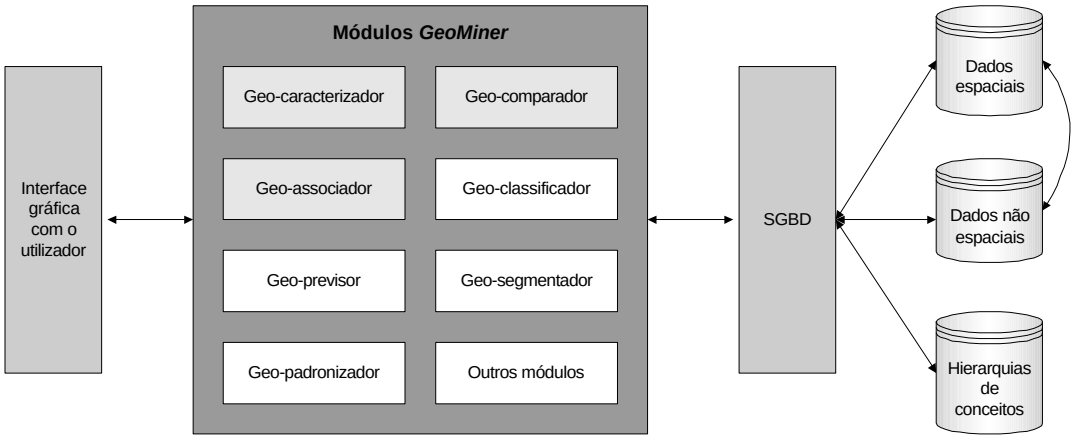


Figura 4.10: Arquitectura do GeoMiner (Adaptado de: [Han et al., 1997])

- Um conjunto de módulos para a exploração dos dados, estando já implementados o geo-caracterizador, o geo-comparador e o geo-associador. Os outros quatro módulos já planeados são: o geo-classificador, o geo-previsor, o geo-segmentador e geo-padronizador.
- Um SGBD para dados espaciais e dados não espaciais.
- Os dados e conhecimento do domínio necessários ao processo de descoberta de conhecimento, nomeadamente, os dados espaciais, dados não espaciais e as hierarquias de conceitos.

Os três módulos já implementados são, de seguida, sumariamente descritos:

Geo-caracterizador. Este módulo permite procurar regras que descrevem características de conjuntos de dados. Permite, ainda, a visualização dos dados em diferentes níveis de abstracção. O processo de descoberta de conhecimento pode ser conduzido pela generalização dominada pelos atributos espaciais ou pela generalização dominada pelos atributos não espaciais, representando duas orientações diferentes no processo de generalização dos dados.

Geo-comparador. Este módulo permite determinar um conjunto de regras que discriminam as diferenças existentes entre diversas classes de dados na BD. A abordagem utilizada compara uma das classes (classe origem), com as restantes classes existentes (classes de contraste).

Geo-associador. Permite encontrar um conjunto de regras de associação espacial, implícitas na BDE analisada. As regras de associação espacial de…nem a dependência probabilística que pode existir entre dados espaciais, ou entre dados espaciais e dados não espaciais. O algoritmo para a extracção de regras de associação em dados relacionais [Agrawal et al., 1993b] tem sido sucessivamente adaptado, apresentando Koperski e Han [Koperski e Han, 1995] [Koperski e Han, 1996] a versão para dados espaciais utilizada pelo GeoMiner. Nesta

versão, as regras de associação espacial são construídas recorrendo a predicados espaciais e a predicados não espaciais.

Os vários módulos implementados permitem a generalização da informação para os diversos níveis de abstracção especi…cados nas hierarquias de conceitos. A expressão GMQL, de seguida apresentada, permite ao utilizar especi…car a tarefa de aprendizagem requerida, assim como os atributos necessários à sua realização:

```
mine spatial characteristics
as ''Distribuição de temperaturas e precipitação''
analyze geo
in relevance to temperatura, precipitação
from meteorologia
where estação=''verão'' and ano=1997 and local=''Braga''
```

Esta questão deverá ser processada pelo módulo geo-caracterizador, o qual deverá seleccionar, da BDE, todos os registos meteorológicos associados ao local em causa. Baseado nos dados seleccionados, o sistema procede à generalização não espacial dos atributos, seguindo as hierarquias temperatura e precipitação. Esta generalização conduz à construção de classes de regiões, que apresentam os mesmos valores (generalizados) de temperatura e precipitação, conforme requerido pelo utilizador.

Classi…cação de dados espaciais através do método two-step

Nesta abordagem, Koperski et al. [Koperski et al., 1998] sugerem um algoritmo para a classi…cação de objectos armazenados em BDE, que recorre à utilização de árvores de decisão. A classi…cação permite encontrar regras que dividem o conjunto de objectos em grupos, de forma a que um dado objecto seja maioritariamente caracterizado pelo comportamento do grupo onde está inserido. Os objectos podem ser classi…cados recorrendo a diferentes tipos de informação, como seja atributos não espaciais associados aos mesmos, ou ainda, recorrendo a predicados e/ou funções espaciais.

A técnica de classi…cação mais utilizada está associada à construção de árvores de decisão. Uma vez que dados geográ…cos integram objectos espaciais e descrições não espaciais desses mesmos objectos, a classi…cação espacial permite encontrar regras para as classes, que não sejam unicamente de…nidas pelas propriedades não espaciais dos objectos, mas que considerem as relações espaciais existentes entre os objectos classi…cados.

O processamento das relações espaciais existentes entre os objectos pode ser extremamente dispendioso, quando analisado numa perspectiva computacional [Koperski et al., 1998]. Para acelerar este processo, é proposta uma estratégia em dois passos (two-step), que passa por determinar de uma forma genérica as regras existentes, as quais são avaliadas por forma a seleccionar as mais interessantes. As regras seleccionadas são sucessivamente melhoradas por forma a aumentar a sua qualidade, produzindo árvores de decisão mais apuradas.

Nesta abordagem, a detecção das relações espaciais entre objectos passa pela utilização de técnicas de aproximação de vizinhanças aos MBR dos objectos em análise. As relações espaciais

encontradas são transformadas em predicados do tipo perto_de (X,Y), sendo estes posteriormente considerados na construção das árvores de decisão. A classi…cação propriamente dita, é realizada veri…cando os valores agregados (atendendo às hierarquias de…nidas) dos atributos não espaciais inseridos em regiões vizinhas.

Os próximos desenvolvimentos deste método de classi…cação passam por permitir de…nir conjuntos de valores para os predicados espaciais, por exemplo, um intervalo de distâncias, e ainda por integrar o algoritmo desenvolvido no protótipo do GeoMiner [Han et al., 1997].

Segmentação de dados espaciais

A técnica de segmentação assenta em princípios estatísticos, e tem como grande vantagem o facto de permitir encontrar classes nos dados, sem recorrer a qualquer conhecimento do domínio, como por exemplo, hierarquias de conceitos [Koperski et al., 1996].

A segmentação tem sido amplamente utilizada em áreas como o reconhecimento de padrões e processamento de imagens, sendo reconhecida como uma das principais e mais utilizadas técnicas na DCBDE [Murray e Estivill-Castro, 1998] [Wang et al., 1997] [Son et al., 1998] [Xu et al., 1998]. Os algoritmos desenvolvidos para BD relacionais não são e…cientes quando utilizados em BDE, primeiro por processarem todos os objectos em memória central (na determinação dos segmentos), e em segundo lugar por se apresentarem ine…cientes mesmo quando aplicados a grandes BD, ainda que não espaciais [Ester et al., 1998b].

No caso especí…co das BDE, a maioria das BD analisadas por estes algoritmos armazenam imagens provenientes de telescópios, satélites, etc. A facilidade de recolha deste tipo de informação conduz a um crescimento exponencial da quantidade de dados armazenados nas mesmas. Tradicionalmente, estes algoritmos têm sido utilizados, na área dos dados espaciais, para criar classes de objectos que partilham características comuns, sendo necessária a manipulação de pontos, linhas e polígonos, armazenados em estruturas de dados especí…cas.

Entre as estruturas mais utilizadas encontram-se as árvores R*, que permitem gerir rectângulos com n dimensões, em vez de chaves numéricas de uma única dimensão, como acontece nas árvores B. As árvores R* permitem manipular objectos com extensão, como por exemplo polígonos (utilizando aproximações aos MBR), ou então pontos, tratados como um caso especial de rectângulos.

Ng e Han [Ng e Han, 1994] utilizaram o CLARANS[17] (Clustering Large Applications based on RANdomized Search) em BDE, para o qual desenvolveram dois algoritmos: o SD(CLARANS) e o NSD(CLARANS). No primeiro, a realização da segmentação é conduzida pelos dados espaciais (SD - Spatial Dominant), enquanto que no segundo, o processo de segmentação é dominado pelos dados não espaciais (NSD - Non-Spatial Dominant).

No SD(CLARANS), e após a selecção dos dados espaciais relevantes para a análise, procede-se à segmentação dos mesmos recorrendo ao CLARANS. Posteriormente, é efectuada a indução orientada aos atributos não espaciais[18] dos objectos em cada classe. Como resultado, determina-se uma descrição não espacial de cada um dos segmentos espaciais encontrados.

[17] Algoritmo de segmentação que suprime a necessidade de veri…cação de todos os registos no processo de classi…cação, implementando um sistema aleatório de procura [Ng e Han, 1994].

[18] Todas as operações sobre os dados não espaciais são efectuadas no DBMiner [Han et al., 1994].

O NSD(CLARANS) efectua a generalização dos atributos não espaciais, produzindo um conjunto de registos generalizados. Para cada um dos registos são seleccionados os objectos espaciais associados, os quais são posteriormente segmentados. As classes assim obtidas são complementadas com as generalizações dos dados não espaciais, permitindo a de…nição das características não espaciais de cada uma das classes espaciais encontradas.

Em ambos os algoritmos é requerida a especi…cação dos atributos de interesse, os quais conduzem o processo de selecção dos dados espaciais e não espaciais. Como se pode veri…car, estes algoritmos utilizam ainda conhecimento do domínio, nomeadamente as hierarquias de conceitos utilizadas no processo de generalização, tornando o processo de segmentação em BDE dependente do utilizador.

4.3.2 Síntese

As abordagens apresentadas anteriormente são representativas dos esforços desenvolvidos na área da DCBDE. Apesar de não terem sido descritas exaustivamente todas as aproximações existentes, resumiram-se as principais estratégias adoptadas na análise de dados espaciais, com o objectivo de descoberta de conhecimento. Estas estratégias passam essencialmente pela:

- construção de novos algoritmos de DM;
- adaptação de algoritmos e/ou ferramentas de descoberta de conhecimento existentes; ou pela
- integração de SGBDE no processo de descoberta de conhecimento, os quais permitem a manipulação de dados espaciais, fornecendo as estruturas de dados necessárias ao seu armazenamento e as funções espaciais que possibilitam a sua análise.

Uma síntese de abordagens à DCBDE é de seguida apresentada, as quais são sistematizadas em duas tabelas. A Tabela 4.2 apresenta as estratégias que derivam da implementação de algoritmos de DM espacial, ou da adaptação de algoritmos e/ou ferramentas já existentes, por forma a permitem a manipulação de dados espaciais. A Tabela 4.3 resume um conjunto de referências, cuja estratégia de desenvolvimento passa pela estruturação de ambientes integrados, para os quais ainda não se conhece o tipo de algoritmos utilizado ou os resultados que é possível obter com a implementação dos mesmos.

Um número considerável de abordagens propostas para o DME utilizam técnicas de generalização, aplicadas a dados espaciais e dados não espaciais, permitindo a de…nição de relações entre estes dois conjuntos de dados. Lu et al. [Lu et al., 1993] utilizam hierarquias de conceitos na indução de regras, através de dois algoritmos: indução orientada aos dados espaciais e indução orientada aos dados não espaciais, sendo o resultado do processo de descoberta de conhecimento in‡uenciado pela ordem em que a indução é efectuada. As regras obtidas descrevem, no primeiro caso, propriedades dos objectos espaciais e a sua relação com os atributos não espaciais analisados. No segundo caso, as regras descrevem propriedades das classes não espaciais encontradas, e a sua relação com os objectos espaciais referenciados. Em ambos os casos, o processo de aprendizagem é iniciado a partir de um pedido do utilizador, que especi…ca numa questão os atributos a explorar e o tipo de resultados pretendido.

Autores	Abordagem	Técnicas utilizadas
Lu et al. [Lu et al., 1993]	Indução orientada aos atributos: - espaciais - não espaciais	Generalização
Ng e Han [Ng e Han, 1994]	Algoritmos de segmentação: - SD(CLARANS) - NSD(CLARANS)	Generalização Segmentação
Koperski et al. [Koperski e Han, 1995] [Koperski et al., 1996] [Koperski et al., 1998]	Ferramenta GeoMiner	Generalização Segmentação Regras de Associação

Tabela 4.2: Algoritmos e ferramentas para o DME

Autores	Abordagem
Abraham e Roddick [Abraham e Roddick, 1997] [Abraham e Roddick, 1998]	Construção de meta-regras para a previsão de tendências
Andrienko e Andrienko [Andrienko e Andrienko, 1998]	Integração de ferramentas de DM com SIG
Dey e Roberts [Dey e Roberts, 1996]	Integração lógica de BD espaciais e não espaciais
Ester et al. [Ester et al., 1997] [Ester et al., 1999a] [Ester et al., 1999b]	Integração do conceito de vizinhança no processo de DCBD

Tabela 4.3: Ambientes integrados no DME

Koperski et al. [Koperski e Han, 1995] [Koperski et al., 1996] [Koperski et al., 1998] investigam a adaptação dos princípios utilizados na descoberta de conhecimento em BD relacionais, por forma aos mesmos permitirem a análise de dados espaciais. O GeoMiner [Han et al., 1997] surge assim como uma extensão ao DBMiner [Han et al., 1994] para a análise de dados espaciais, no qual estão já disponíveis três módulos de descoberta de conhecimento: o geo-caracterizador, o geo-associador e o geo-comparador. Na arquitectura deste sistema, dados espaciais e dados não espaciais são armazenados em BD diferentes, sendo a sua integração conseguida através de estruturas do tipo SAND. Além da especi…cação do conhecimento do domínio necessário ao processo de generalização, o utilizador especi…ca numa questão GMQL, o tipo de regra requerida e os atributos de interesse para a análise. Entre o tipo de regras que podem ser encontradas, estão as regras de associação espacial [Koperski e Han, 1995]. Estas regras incluem pelo menos um atributo espacial, e predicados espaciais como próximo ou dentro_de, na descrição de associações. O algoritmo é baseado em estratégias de indução, requerendo como conhecimento do domínio as hierarquias de conceitos a utilizar no processo de generalização/especialização das regras. Entre os módulos a acrescentar ao GeoMiner destaca-se o geo-classificador. O algoritmo já implementado [Koperski et al., 1998] permite a classi…cação de objectos espaciais através da construção de árvores de decisão.

A inclusão de técnicas de segmentação no processo de descoberta de conhecimento conduziu Ng e Han [Ng e Han, 1994] a apresentação de duas versões do algoritmo CLARANS, uma dominada pelos dados espaciais SD(CLARANS), e outra dominada pelos dados não espaciais NSD(CLARANS). A primeira executa a segmentação dos dados espaciais, aplicando posteriormente o DBMiner [Han et al., 1994] aos dados não espaciais associados a cada um dos segmentos encontrados. Esta aproximação permite a descrição não espacial dos segmentos espaciais obtidos. A aproximação dominada pelos atributos não espaciais começa por analisar estes dados com o DBMiner, aplicando posteriormente o CLARANS aos atributos espaciais associados às generalizações não espaciais obtidas. Como resultado identi…cam-se os segmentos espaciais que existem dentro de cada generalização não espacial obtida pelo DBMiner.

Ester et al. [Ester et al., 1997] [Ester et al., 1999a] [Ester et al., 1999b] apresentam uma abordagem alternativa, que passa pela introdução de um conjunto de operações num SGBDE, por forma a estes permitirem a algoritmos de DM manipular, e como tal incluir no processo de descoberta de conhecimento, dados espaciais. Para tal descrevem conceitos como índices, grafos e caminhos de vizinhança, e ainda, conjuntos de operações a executar sobre os mesmos. Segundo os autores, a extensão dos SGBDE de forma a integrar este tipo de conceitos e operações, permitirá a integração da noção de vizinhança no processo de DCBD.

Dey e Roberts [Dey e Roberts, 1996] sugerem uma abordagem que estende a arquitectura genérica, para SDC, proposta por Matheus et al. [Matheus et al., 1993], adicionando duas interfaces, uma para aceder a dados espaciais e outra para aceder a dados não espaciais (armazenados em diferentes BD). A integração das diferentes BD é lógica, e é conseguida pelo próprio processo de DCBD. A BD não espacial contém identi…cadores geográ…cos, como por exemplo moradas, que referenciam localizações na BD espacial. O processo de descoberta de conhecimento pode ser iniciado partindo dos dados espaciais ou dos dados não espaciais.

A construção de um sistema de DCBDE genérico, capaz de lidar com diferentes tipos de algoritmos e estruturas de armazenamento de informação espacial, é analisada por Abraham e Roddick [Abraham e Roddick, 1997] [Abraham e Roddick, 1998]. Neste sistema, os autores dão particular ênfase a construção de uma BD de padrões, na qual são armazenados os resul-

tados do processo de descoberta de conhecimento, que permita veri…car tendências nos dados. Estas tendências são estabelecidas através da construção de meta-regras espaço-temporais, que comparam conjuntos de regras obtidos anteriormente, em diferentes períodos temporais. As meta-regras assim obtidas caracterizam a evolução ou persistência de determinados padrões, permitindo a previsão de tendências para o domínio de aplicação modelado.

Uma outra abordagem, dos autores Andrienko e Andrienko [Andrienko e Andrienko, 1998], passa pela integração de ferramentas de DM tradicionais com SIG. O objectivo é disponibilizar ao utilizador mecanismos de visualização da informação referenciada espacialmente (dados e resultados), por forma permitir a detecção de descrições espaciais, nos padrões não espaciais encontrados pela ferramenta de DM utilizada. A solução proposta pelos autores apenas permite a descoberta de conhecimento a partir de dados não espaciais. Qualquer padrão espacial deverá ser detectado posteriormente pelo utilizador, na visualização grá…ca (em mapas) dos padrões não espaciais encontrados. A abordagem proposta fornece, assim, uma integração parcial dos dados referenciados espacialmente no processo de DCBD.

Capítulo 5

PADRÃO: Um sistema de descoberta de conhecimento

Neste capítulo é apresentado o enquadramento estrutural que justi…ca a arquitectura do sistema Padrão proposta neste trabalho. Posteriormente, é descrita a arquitectura do sistema Padrão, salientando os seus principais componentes, e ainda os diagramas de classes e os diagramas de caso de uso, que de…nem a estrutura das BD utilizadas e caracterizam o modo de funcionamento do sistema, respectivamente.

Os diagramas de caso de uso permitem de…nir a sequência de acções que um actor[1] realiza num sistema para conseguir determinado resultado. Possibilitam, assim, a descrição do que um sistema faz (ou parte deste), mas não como é que tal é internamente realizado. Estes diagramas são particularmente importantes na organização e modelação do comportamento de um sistema. Os diagramas de classes são utilizados para representar a estrutura comum de um conjunto de objectos, permitindo neste trabalho, construir o desenho lógico[2] das BD que integram o Padrão. Estes dois tipos de diagramas permitem documentar[3] o sistema, através da conceptualização daquilo que faz e das estruturas de dados utilizadas para suportar o seu funcionamento.

Após a apresentação e caracterização da arquitectura do sistema, este capítulo prossegue com a implementação do Padrão, descrevendo como a mesma foi conseguida no ambiente de trabalho do sistema Clementine.

[1] Actores são utilizadores ou qualquer outro sistema que possa interagir com o sistema em causa.

[2] Os diagramas de classes são utilizados para modelar uma visão estática do sistema [Booch et al., 1999], a qual pode ser efectuada de três formas distintas: i) pela modelação do vocabulário do sistema, permitindo de…nir as abstracções que fazem parte do mesmo; ii) pela modelação de colaborações, que são executadas por um conjunto de classes e suas relações; ou, iii) pela modelação do esquema lógico da BD.

[3] O UML é uma linguagem de modelação que permite visualizar, especi…car, construir e documentar os componentes de uma aplicação. No âmbito da documentação, o UML permite documentar a arquitectura de um sistema, assim como os detalhes associados à mesma [Booch et al., 1999].

5.1 Enquadramento do sistema PADRÃO

A revisão teórica realizada nos capítulos anteriores (nomeadamente no Capítulo 2, Capítulo 3 e Capítulo 4) permite identificar o conjunto de funcionalidades desejadas para o sistema Padrão, e ainda, justificar as diversas opções estruturais em que o mesmo se baseia.

A análise de dados espaciais com o objectivo de descoberta de conhecimento requer a utilização de técnicas específicas, que possibilitem a inclusão da semântica espacial, implícita na posição e dimensão dos objectos geográficos referenciados, no referido processo. Estas técnicas, e conforme descrição já efectuada no Capítulo 4, baseiam-se essencialmente: no desenvolvimento de novos algoritmos de DM capazes de incluir a semântica espacial no processo de descoberta de conhecimento; ou, na integração de SGBDE com ferramentas de descoberta de conhecimento, os quais permitem a manipulação dos dados espaciais e consequente junção dos resultados com os restantes dados não espaciais analisados.

Estas abordagens requerem a descrição geométrica dos diversos objectos geográficos referenciados, uma vez que se baseiam em estratégias de raciocínio espacial quantitativo, nas quais o posicionamento dos objectos é conseguido manipulando as respectivas coordenadas de pontos.

Constatando-se que grande parte das organizações possuem BD nas quais a dimensão espacial é referenciada recorrendo a identificadores geográficos qualitativos, como moradas, para os quais as mesmas não necessitam (e como tal não possuem), no seu funcionamento diário, da descrição geométrica dos objectos, uma abordagem quantitativa ao raciocínio espacial adiciona novas dificuldades às organizações que pretendem analisar os seus dados, com o objectivo de descoberta de conhecimento.

A existência destes identificadores qualitativos, nas BD, permite a utilização de sistemas indirectos de posicionamento geográfico, que suprimem a necessidade de caracterização geométrica das diversas entidades geográficas referenciadas, e ainda, evitam o desenvolvimento de novos algoritmos de DM para a análise dos dados.

Para que o raciocínio espacial seja possível, isto é, para que a semântica espacial associada aos identificadores geográficos utilizados seja efectivamente integrada no processo de descoberta de conhecimento, é necessário utilizar estratégias de raciocínio espacial qualitativo, que permitam raciocinar com informação geográfica incompleta ou imprecisa. Uma vez que o raciocínio espacial qualitativo utiliza tabelas de composição na inferência de relações espaciais desconhecidas, esta abordagem apenas requer a definição de um conjunto mínimo de relações espaciais existentes entre objectos, sendo as restantes inferidas à medida que vão sendo necessárias.

Neste trabalho, e uma vez que foram adoptados os princípios estabelecidos nas pré-normas europeias CEN TC 287 para informação geográfica, a implementação do esquema de identificadores geográficos e do esquema espacial permite a definição das relações espaciais existentes entre entidades geográficas adjacentes.

Ao nível das relações espaciais, as tabelas de composição utilizadas pelo sistema Padrão integram relações espaciais do tipo direcção, distância e topologia, uma vez que consideradas em conjunto permitem aumentar a precisão das inferências.

A Figura 5.1 sistematiza as diversas opções estruturais em que se baseia o sistema Padrão, antecipando, ainda, parte da sua arquitectura, na qual é possível verificar as diversas BD que o sistema integra e as fases do processo de descoberta de conhecimento consideradas pelo mesmo.

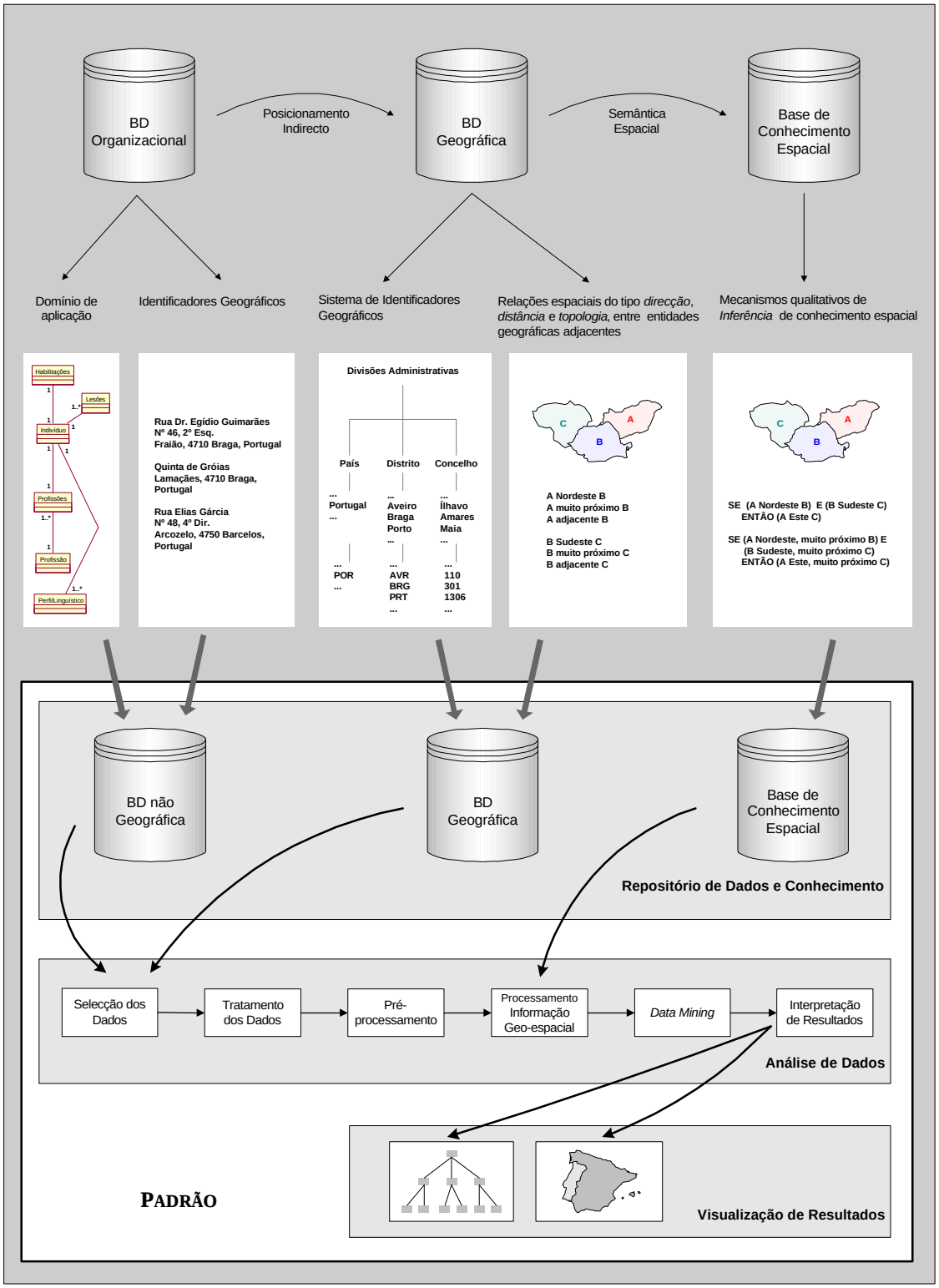


Figura 5.1: Enquadramento do sistema Padrão

O enquadramento realizado, e que evidencia as opções estruturais identi…cadas para o sistema Padrão, permite diferenciar a abordagem proposta neste trabalho, e ainda, salientar as suas principais vantagens. Destaca-se que a utilização de estratégias de raciocínio espacial qualitativo na análise de dados espaciais, com o objectivo de descoberta de conhecimento, constitui uma inovação. Esta estratégia permite que dados organizacionais sejam analisados com o objectivo de descoberta de conhecimento, independentemente da disponibilidade de algoritmos de DM especí…cos e da geometria dos objectos geográ…cos referenciados. Contudo, a principal vantagem desta aproximação está associada ao facto da mesma permitir que dados não geográ…cos e dados geo-espaciais sejam analisados simultaneamente pelos algoritmos de DM, não condicionando ou limitando os resultados que podem ser obtidos. Refere-se, ainda, que esta abordagem permite a utilização de uma diversidade de algoritmos de DM já disponíveis em ferramentas de descoberta de conhecimento tradicionais, possibilitando o uso de um vasto conjunto de técnicas na análise dos dados.

5.2 Arquitectura do sistema PADRÃO

A arquitectura do Padrão agrega três componentes principais: i) Repositório de Dados e Conhecimento; ii) Análise de Dados, e iii) Visualização de Resultados. A Figura 5.2 apresenta uma visão global do sistema, cujos componentes são descritos nas próximas subsecções.

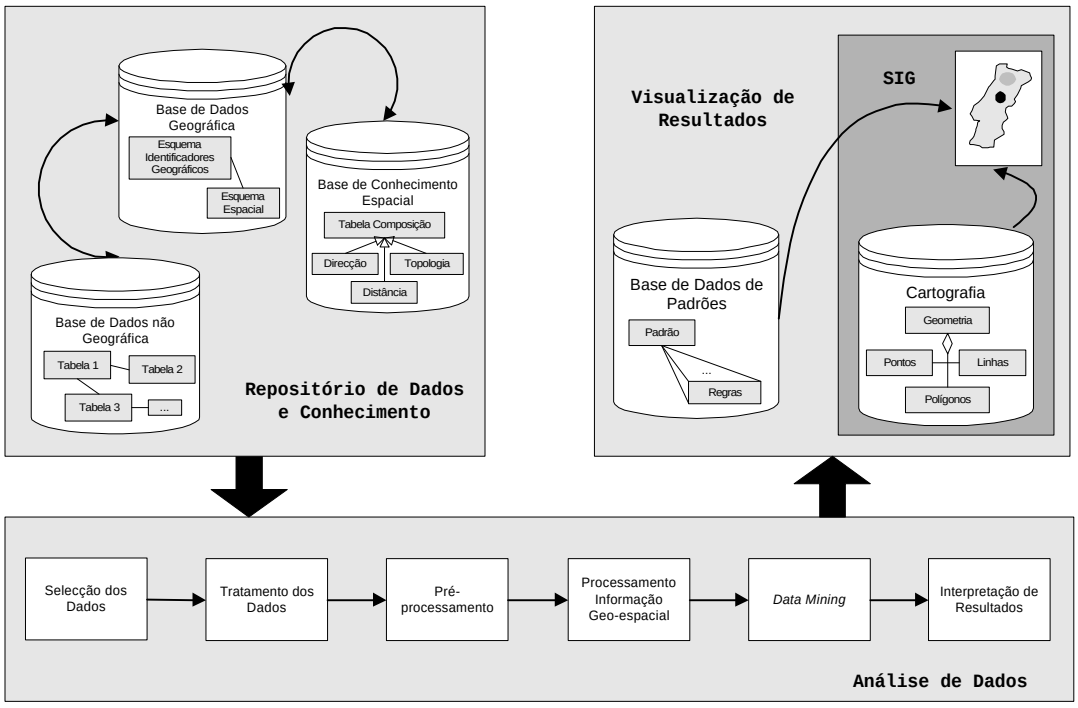


Figura 5.2: Arquitectura do sistema Padrão

Genericamente, o sistema recorre a três BD, as quais armazenam os dados e o conhecimento necessários ao processo de descoberta de conhecimento. As várias etapas que integram

este processo permitem a identi…cação de padrões ou outros relacionamentos implícitos, existentes nas BD analisadas. As regras que caracterizam determinado padrão podem por sua vez ser armazenadas, procedimento que permite a visualização das mesmas em mapas das regiões referenciadas, e ainda, a sua posterior utilização em exercícios de DM. A reutilização do conhecimento previamente obtido pode conduzir à construção de meta-regras, que descrevem a evolução dos padrões ao longo do tempo.

Como poderá ser constatado ao longo da descrição[4] de cada um dos componentes que integram o Padrão, o sucesso ou insucesso de um projecto de DM está condicionado pela pessoa[5] ou equipa que o leva a cabo. A diversidade de conhecimentos e aptidões que são requeridas nestes exercícios, sugere a necessidade de criação de normas para o processo de DM, que auxiliem o utilizador na condução do mesmo. Entre outras tarefas, estas normas devem auxiliar o processo de transformação dos objectivos do negócio em tarefas de DM, sugerir mecanismos apropriados para o tratamento dos dados e algoritmos de DM a utilizar, e ainda, permitir avaliar os resultados obtidos e documentar convenientemente o exercício.

O projecto CRoss Industry Standard Process for Data Mining[6] (CRISP-DM) tem como objectivo auxiliar a resolução destes problemas [Wirth e Hipp, 2000], de…nindo um modelo para o processo de descoberta de conhecimento que é independente do domínio de aplicação ou da tecnologia utilizada.

O modelo de referência do CRISP-DM fornece uma perspectiva do ciclo de vida dos projectos de DM, descrevendo cada uma das fases, as suas tarefas, e os resultados esperados de cada uma das mesmas. As fases consideradas neste modelo são: compreensão do negócio (Business Understanding), compreensão dos dados (Data Understanding), preparação dos dados (Data Preparation), modelação (Modelling), avaliação (Evaluation) e desenvolvimento (Development).

5.2.1 O componente Repositório de Dados e Conhecimento

O Repositório de Dados e Conhecimento é o responsável pelo armazenamento dos dados geo-espaciais e dados não geográ…cos utilizados no sistema. Integra, ainda, o conhecimento espacial necessário à implementação dos mecanismos qualitativos de inferência utilizados no sistema. Este componente integra três BD:

Uma Base de Dados Geográ…ca (BDG), construída segundo os princípios do CEN TC 287. Atendendo às directivas do CEN, foi possível implementar uma BDG na qual o posicionamento espacial dos dados geográ…cos é conseguido recorrendo a um sistema de identi…cadores geográ…cos. Neste sistema, caracterizam-se as subdivisões administrativas de Portugal, ao nível dos Concelhos e Distritos. A componente espacial associada aos identi…cadores geográ…cos utilizados permitiu a de…nição das relações espaciais do tipo

[4] Na qual o actor investigador participa em todas as tarefas associadas ao processo de descoberta de conhecimento.

[5] Como referido por Brachman e Anand [Brachman e Anand, 1996], apesar de a longo prazo ser desejável que os sistemas de descoberta de conhecimento sejam autónomos, a realidade é que actualmente, o utilizador desempenha um papel extremamente importante em todo este processo. Os autores sugerem que as descrições do processo de DCBD devem ser mais orientadas ao utilizador, indicando como este processo deve ser conduzido e enumerando as tarefas que devem ser executadas.

[6] Este projecto está a ser parcialmente suportado pela Comissão Europeia através do programa ESPRIT. O conjunto de parceiros é constituído pelo consórcio DaimlerChrysler AG, SPSS, NCR e OHRA.

direcção, distância e topologia, existentes entre Concelhos[7] adjacentes. A estrutura da BDG integra os dados provenientes do esquema espacial e do esquema de identi…cadores geográ…cos, já apresentados no Capítulo 2 nos seus respectivos diagramas de classes.

A construção da BDG requer a participação de dois actores[8]. O investigador[9] é o responsável pela identi…cação das características do domínio geográ…co em causa, o qual condiciona os diversos esquemas a implementar. O comité CEN TC 287 é considerado um actor, pelo facto das suas directivas interagirem com o sistema, fornecendo as especi…cações necessárias à implementação dos diversos esquemas identi…cados pelo investigador. No diagrama de caso de uso[10] apresentado na Figura 5.3, a BDG é também considerada um actor, pelo facto de esta interagir com o sistema. No caso especí…co de BD, optou-se por representar estes actores através de estereótipos[11] (stereotypes), um mecanismo de extensão permitido pelo UML [Booch et al., 1999], representados por um ícone que os diferencia dos restantes actores. A utilização de estereótipos permite diferenciar o papel, neste caso dos diversos actores, no sistema.

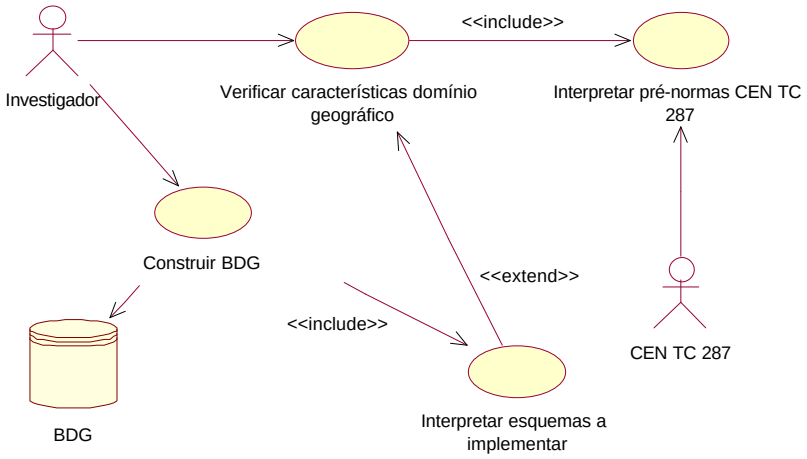


Figura 5.3: Diagrama de caso de uso: construção da BDG

[7] As relações espaciais são de…nidas entre Concelhos, por este conjunto de regiões representar o nível inferior da hierarquia geográ…ca utilizada neste trabalho. As relações espaciais dos restantes níveis podem ser inferidas a partir destas.

[8] Um actor representa um conjunto coerente de papéis, que um utilizador de um caso de uso leva a cabo quando interage com esse caso de uso. Tipicamente, um actor representa o papel que um indivíduo, um equipamento ou mesmo outro sistema, desempenha no sistema [Booch et al., 1999].

[9] A denominação investigador é aqui adoptada por referenciar a pessoa responsável pela implementação do sistema Padrão. Independentemente da designação adoptada, refere-se que este actor deve possuir valências que lhe permitam interagir com o sistema. Estas valências variam, sendo especí…cas do domínio de intervenção em que as mesmas são requisitadas.

[10] Os diagramas de caso de uso e de classes apresentados, foram construídos recorrendo ao Rational Rose 98. O Rational Rose é um produto da Rational Software Corporation. Mais detalhes podem ser encontrados em http://www.rational.com/rose.

[11] Os estereótipos estendem o vocabulário do UML, permitindo a de…nição de novos elementos, especí…cos do problema em causa [Booch et al., 1999].

Como já referido anteriormente, a BDG integra os dados provenientes do esquema de identi…cadores geográ…cos e do esquema espacial. Do conjunto de entidades que agregam estes esquemas, existem algumas que se revelam de extrema importância, uma vez que serão utilizadas no processo de descoberta de conhecimento. A Figura 5.4 apresenta este conjunto de entidades. Pela análise da referida …gura é possível veri…car que a Primitiva Topológica associada a determinada Instância de Localização é do tipo Nodo (em particular um Nodo Isolado, que representa o centróide da região referenciada pelo identi…cador geográ…co). Este relacionamento identi…ca a Face associada ao Nodo Isolado, permitindo então conhecer as relações espaciais existentes no espaço geográ…co analisado.

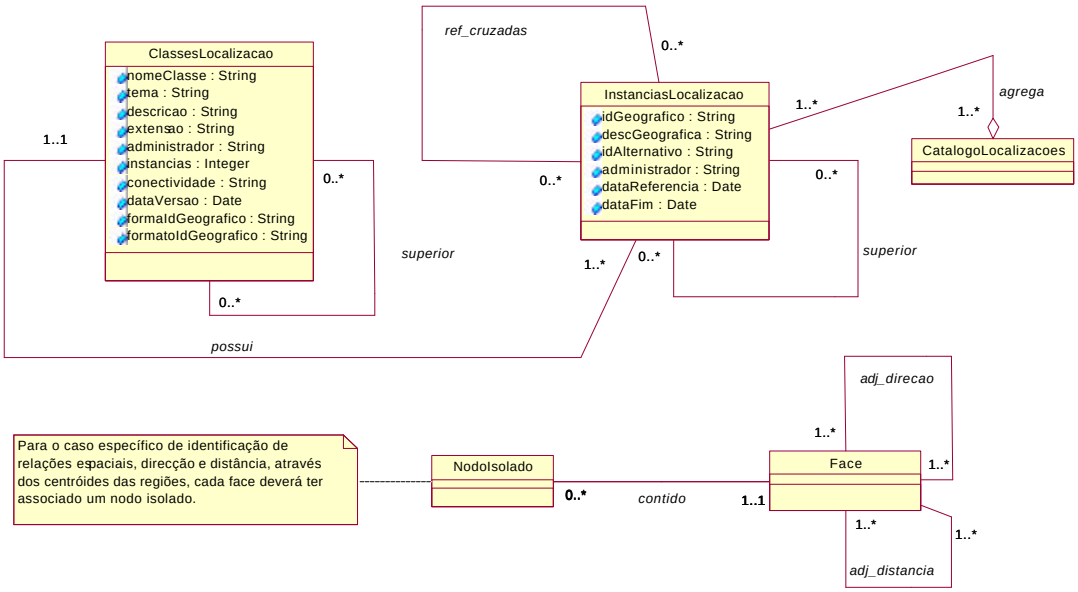


Figura 5.4: Entidades da BDG relevantes no processo de descoberta de conhecimento

Uma Base de Conhecimento Espacial (BCE) que armazena os mecanismos de raciocínio qualitativo que permitem a inferência de relações espaciais desconhecidas. Dentre o conhecimento disponível nesta base, encontram-se as tabelas de composição, que integram a direcção, a distância e a topologia segundo os princípios do raciocínio espacial qualitativo (conforme integração efectuada no Capítulo 3, subsecção 3.5.3), os identi…cadores qualitativos utilizados, e ainda, o intervalo de validade quantitativo associado a cada um dos mesmos. A construção da BCE requer a participação activa do investigador, o qual deve identi…car as relações espaciais de interesse, de…nir as características do sistema qualitativo a implementar, e construir as tabelas de composição que permitem a inferência de relações espaciais desconhecidas (Figura 5.5).

A Figura 5.6 apresenta o diagrama de classes que retrata a estrutura da BCE. Na referida …gura é possível veri…car que o sistema de raciocínio agrega o conhecimento espacial associado

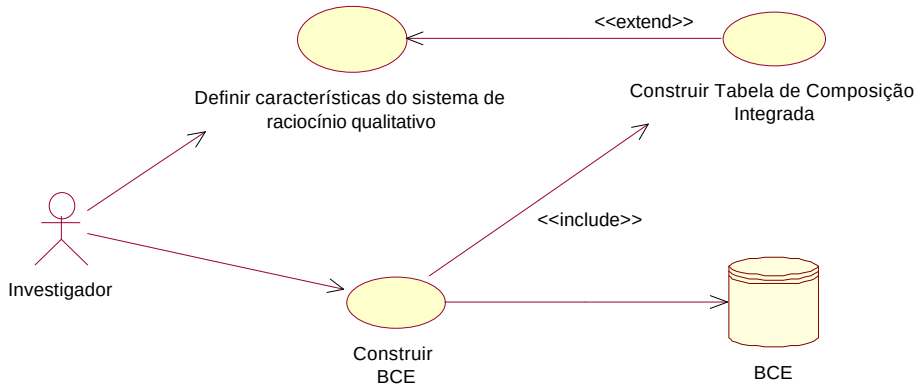


Figura 5.5: Diagrama de caso de uso: construção da BCE

a cada uma das relações espaciais consideradas. A de…nição dos intervalos de validade quantitativos, associados a cada identi…cador qualitativo, permite que o contexto quantitativo se altere em função dos objectivos da análise a efectuar.

Uma Base de Dados não Geográ…ca (BDnG) cujo conteúdo depende do domínio de aplicação em causa. Ao longo deste documento são referidas duas BDnG distintas. Ainda neste capítulo, na descrição da implementação do sistema Padrão, é referida uma BD demográ…ca, que armazena os registos paroquiais do distrito de Aveiro, datados entre 1690 e 1990. Posteriormente, no Capítulo 7, é apresentado um estudo de caso, no qual a BDnG está associada ao Sistema de Administração do Pessoal do Exército.

5.2.2 O componente Análise de Dados

O componente de Análise de Dados é implementado no Clementine, conforme descrição efectuada na secção 5.3 deste mesmo capítulo. O módulo de descoberta de conhecimento resultante, caracteriza-se por passar por 6 grandes etapas: selecção dos dados, tratamento dos dados, pré-processamento dos dados, processamento da informação geo-espacial, DM e interpretação de resultados. Estas fases são descritas nas próximas subsecções.

Funcionalmente, o processo de descoberta de conhecimento, nomeadamente as suas interacções com os actores, é caracterizado por diversos diagramas de caso de uso, os quais são apresentados à medida que as fases a que dizem respeito são descritas. Para que o processo de descoberta de conhecimento seja iniciado, é necessário que o utilizador[12] estabeleça o(s) objectivo(s) da análise. Esta tarefa é efectuada com a colaboração do investigador, que tem a seu cargo a responsabilidade de conduzir o processo e transformar os objectivos da análise em objectivos do DM. A identi…cação do subconjunto de dados necessário é então desencadeada, os quais são posteriormente tratados, pré-processados e integrados com a informação geo-espacial necessária ao cumprimento da tarefa. Posteriormente, e atendendo às características dos dados

[12] Perito do domínio de aplicação.

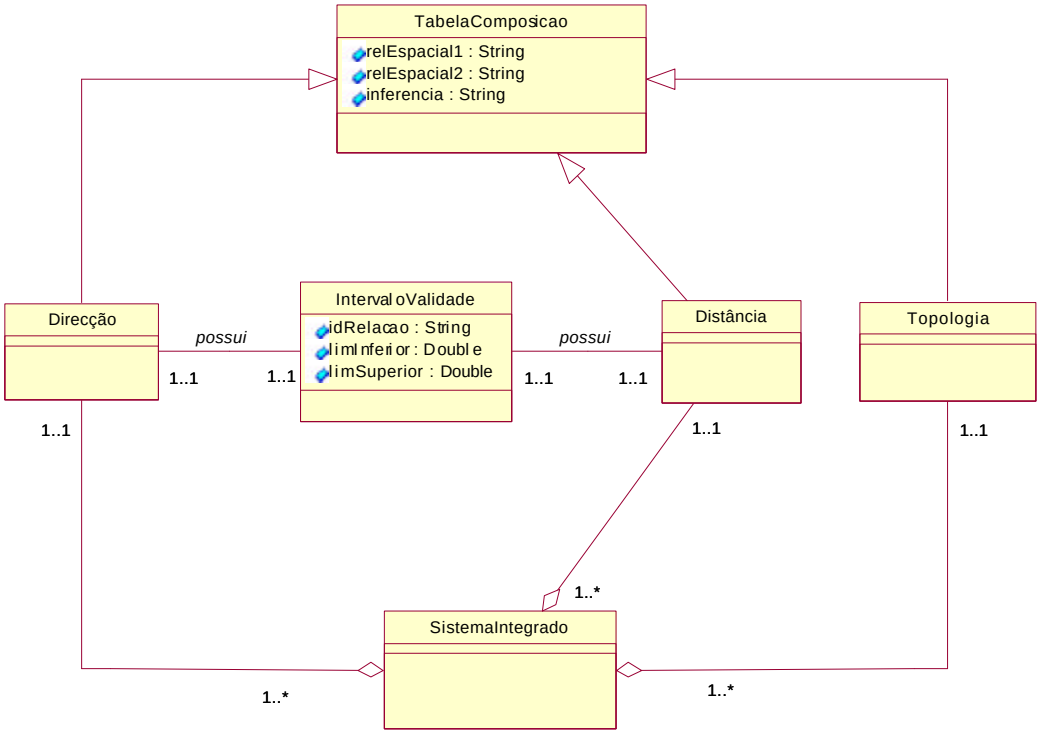


Figura 5.6: Diagrama de classes: estrutura da BCE

a analisar e o objectivo a atingir, o investigador selecciona o algoritmo de DM mais apropriado a tarefa em causa. A execução deste algoritmo poderá conduzir à identi…cação de padrões ou outros relacionamentos nos dados, os quais são posteriormente analisados por forma a avaliar a sua utilidade. Esta avaliação permite ao investigador medir o desempenho das decisões tomadas ao longo do processo, as quais podem ser revistas e melhoradas visando a optimização do mesmo.

Selecção dos dados

Nesta fase são identi…cados os dados não geográ…cos[13] e os dados geo-espaciais considerados relevantes para a tarefa de DM a efectuar (Figura 5.7). Nesta fase deverá ser equacionado:

Que dados devem ser seleccionados, atendendo aos objectivos da análise. A identi…cação do subconjunto de atributos relevante à análise depende dos objectivos que a mesma visa servir. Existe contudo, um conjunto de atributos com valor meramente informativo, como nomes, descrições, etc., que não têm qualquer interesse no processo de descoberta de conhecimento, e como tal podem à partida ser omitidos.

[13] Assume-se que os dados não geográ…cos a analisar possuem entre os seus atributos, um identi…cador geográ…co que os relaciona com os dados geo-espaciais.

O horizonte temporal a analisar ou a periodicidade de recolha das amostras. Diversos horizontes podem ser de…nidos. Para cada um deles, um conjunto de regras modela os dados analisados, o que permite a construção de meta-regras que avaliam a evolução ocorrida ao longo dos diversos horizontes temporais analisados. Ao nível da periodicidade de recolha das amostras, tem de ser veri…cada a dinâmica do sistema, com o objectivo de determinar as várias amostras a recolher.

O tamanho da amostra a analisar. O tamanho da amostra depende essencialmente das características dos dados armazenados na BD e da complexidade do problema a tratar. Quantos mais atributos e quantos mais valores para os mesmos, mais dados serão necessários à construção dos modelos.

A disponibilidade da componente geográ…ca. De…nido o objectivo da análise, é necessário garantir a geo-referenciação dos dados, por forma a permitir a inclusão da componente geo-espacial no processo de descoberta de conhecimento.

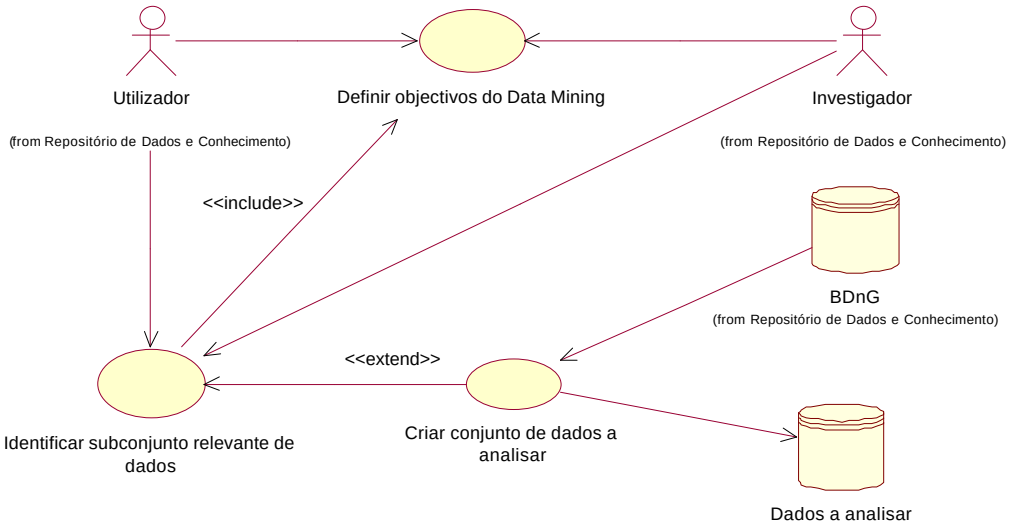


Figura 5.7: Diagrama de caso de uso para a fase de selecção dos dados

Tratamento dos dados

A fase de tratamento dos dados consiste essencialmente na limpeza dos mesmos. Os mecanismos de limpeza que podem ser utilizados dependem dos dados em causa, sendo inevitável a participação do utilizador (perito do domínio) nesta tarefa. Após a selecção do subconjunto de dados relevante à análise, estes devem ser inspeccionados por forma a detectar valores errados ou omissos. Os valores errados caem normalmente fora do intervalo de validade admissível para o domínio de aplicação em causa. Os valores omissos, e sempre que possível, são substituídos pelos

valores reais, ou então etiquetados com determinada descrição (por exemplo, 'desconhecido' ou '?'), que lhes permita ser considerados pelos algoritmos de DM (Figura 5.8).

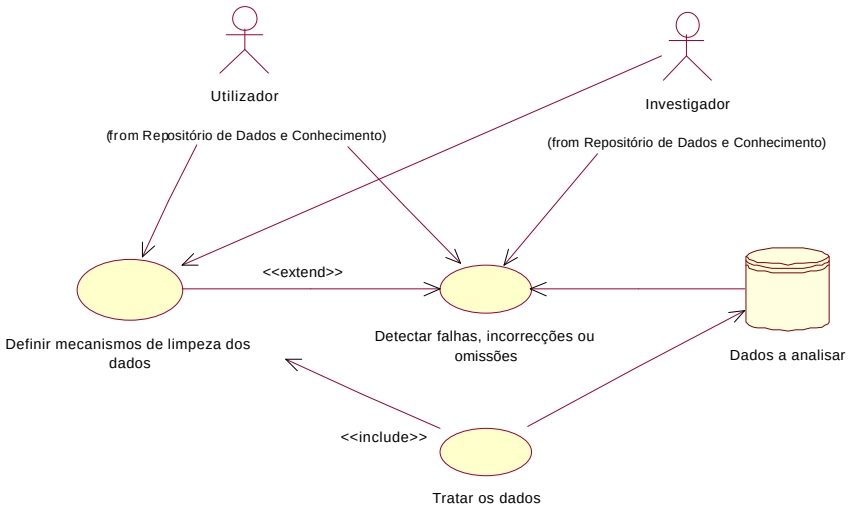


Figura 5.8: Diagrama de caso de uso para a fase de tratamento dos dados

Pré-processamento dos dados

É nesta fase que os dados são transformados no seu formato …nal, por forma a poderem ser explorados pelos algoritmos de DM. Esta etapa visa essencialmente a redução do tamanho da amostra, o que pode ser efectuado reduzindo o número de colunas e o número de linhas:

- a redução do número de colunas é conseguida através da remoção de atributos que não se revelam importantes para a satisfação do objectivo de…nido. A detecção de tais casos pode ser efectuada recorrendo à própria ferramenta de DM, a qual é utilizada para gerar uma árvore de decisão que se encarrega de apenas considerar na classi…cação de determinado atributo de saída, o subconjunto de atributos (de todos os seleccionados anteriormente) que realmente contribui para a previsão do mesmo. Outra possibilidade consiste em utilizar a teoria dos conjuntos irregulares[14] (Rough sets) [Rodrigues et al., 1999] para a determinação de tal subconjunto de atributos.

- a redução do número de linhas é conseguida agrupando registos com valor idêntico, e pode ser efectuada através de duas formas:

[14] A Teoria dos Conjuntos Irregulares proposta por Pawlak [Pawlak, 1991] é baseada na teoria de conjuntos, e permite classi…car objectos em conjuntos, baseado nos atributos dos mesmos. Constitui uma aproximação matemática para lidar com conceitos vagos e imprecisos. A sua utilização na análise de BD é adequada quando não existe conhecimento prévio acerca dos dados a analisar, uma vez que a abordagem proposta não necessita de qualquer conhecimento prévio dos dados ou das dependências existentes entre os mesmos. Esta teoria tem vindo a ser estudada e utilizada por Rodrigues [Rodrigues, 2000] na extracção de conhecimento em sistemas de informação imprecisos.

– generalização dos dados, recorrendo às hierarquias conceptuais de…nidas para o domínio de aplicação em causa;
– discretização dos dados, atendendo às classes de…nidas para atributos com valores contínuos (por exemplo, idade, número de …lhos, ...).

A participação do utilizador na fase de pré-processamento dos dados é evidenciada na Figura 5.9, na qual é possível veri…car que tanto as hierarquias conceptuais como as classes a utilizar na discretização dos atributos, são de…nidas com a participação deste. O seu conhecimento do domínio de aplicação é assim incluído no processo de descoberta de conhecimento.

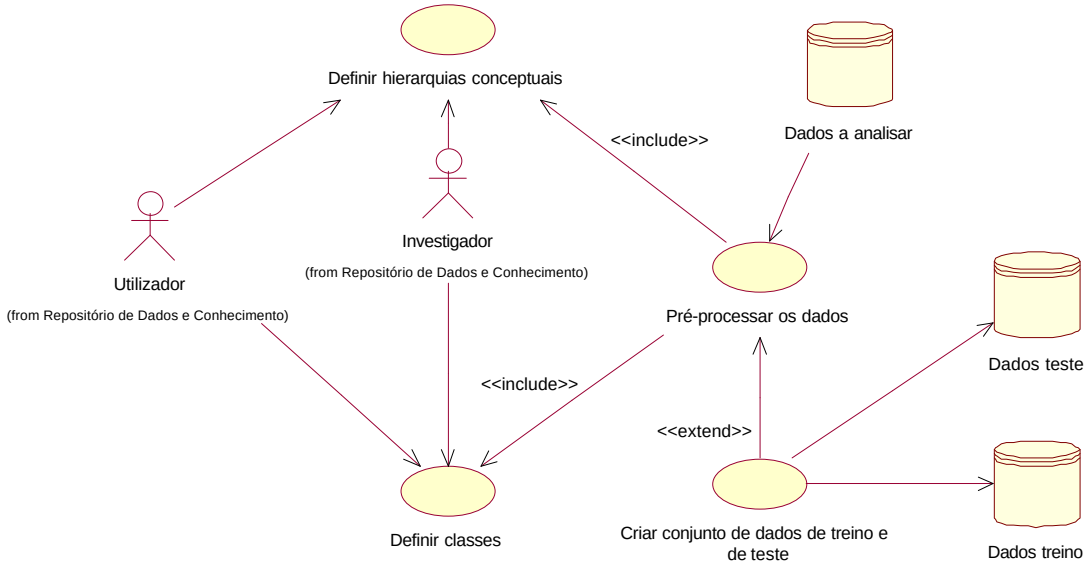


Figura 5.9: Diagrama de caso de uso para a fase de pré-processamento dos dados

Processamento da informação geo-espacial

Nesta fase veri…ca-se a informação geo-espacial disponível, no conjunto da informação necessária à etapa de DM. Uma vez que a BDG apenas armazena as relações espaciais existentes entre entidades geográ…cas adjacentes, todas as restantes, e consoante o requerido para a realização da tarefa em causa, são inferidas recorrendo às regras e conhecimento espacial armazenado na BCE.

A Figura 5.10 apresenta o diagrama de caso de uso que retrata o processo de veri…cação da informação geo-espacial disponível, e ainda, a inferência da informação necessária aos algoritmos de DM.

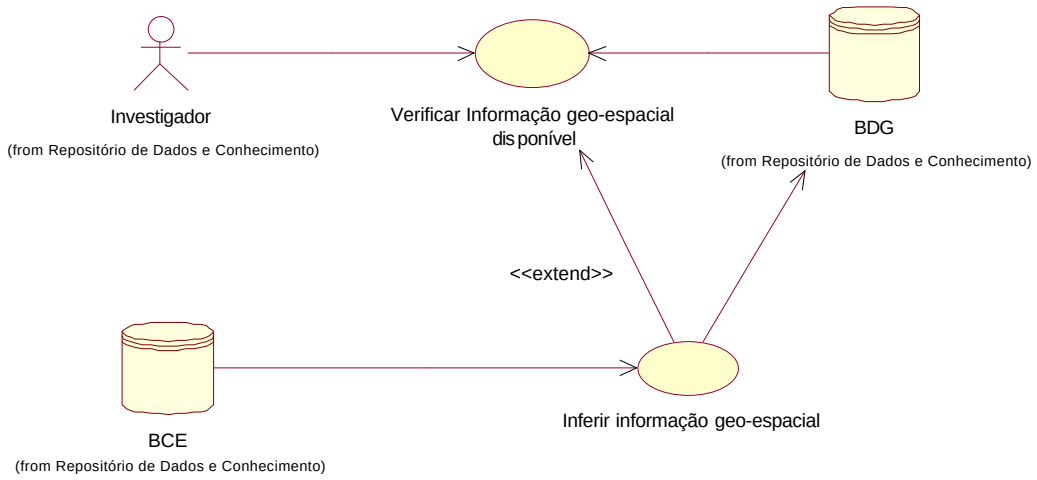


Figura 5.10: Diagrama de caso de uso para a fase de processamento da informação geo-espacial

Data Mining

Esta etapa tem como objectivo encontrar um modelo que se ajuste aos dados, ou que descreva padrões implícitos nos mesmos. A adopção de determinado algoritmo é efectuada atendendo às tarefas inicialmente de…nidas. Diferentes algoritmos podem ser utilizados, dependendo se se pretende classi…car, prever ou associar, ou mesmo para cada uma destas tarefas, a ferramenta de DM pode disponibilizar várias técnicas: árvores de decisão, redes neuronais, etc. O investigador tem um papel decisivo nesta fase, identi…cando o algoritmo que melhor se adapta à tarefa e aos dados. Depois de identi…cado o algoritmo a utilizar (Figura 5.11), dados geográ-…cos e dados não geográ…cos são integrados por forma a identi…car relacionamentos implícitos nos mesmos.

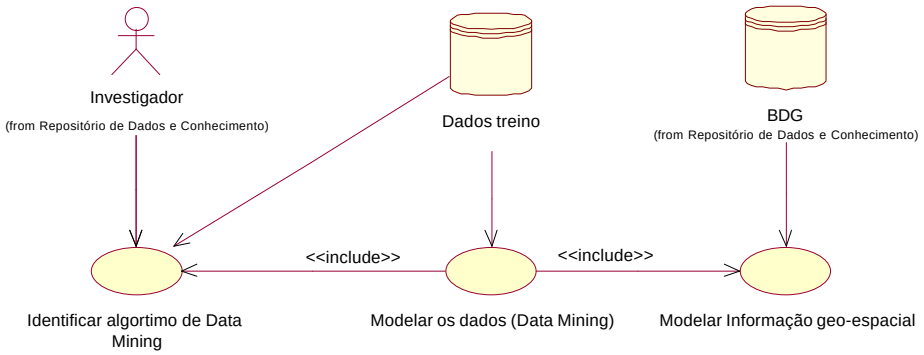


Figura 5.11: Diagrama de caso de uso para a fase de data mining

Interpretação de resultados

Esta é, provavelmente, a fase mais importante das seis que constituem o processo de descoberta de conhecimento, por permitir avaliar o desempenho de todas as opções/decisões tomadas nas fases anteriores. A avaliação da utilidade dos padrões encontrados é realizada pelo utilizador (que como perito do domínio verifica o interesse e relevância das descobertas) e pelo investigador (que como mediador de todo o processo pode alterar decisões, que serão consideradas em próximas iterações do processo). Os modelos encontrados são aplicados à amostra de teste, permitindo verificar a validade dos mesmos quando utilizados em dados desconhecidos do sistema. A avaliação efectuada conduz à identificação dos padrões considerados relevantes (Figura 5.12), os quais podem posteriormente ser armazenados na Base de Dados de Padrões (BDP), com vista à sua reutilização ou visualização em mapas.

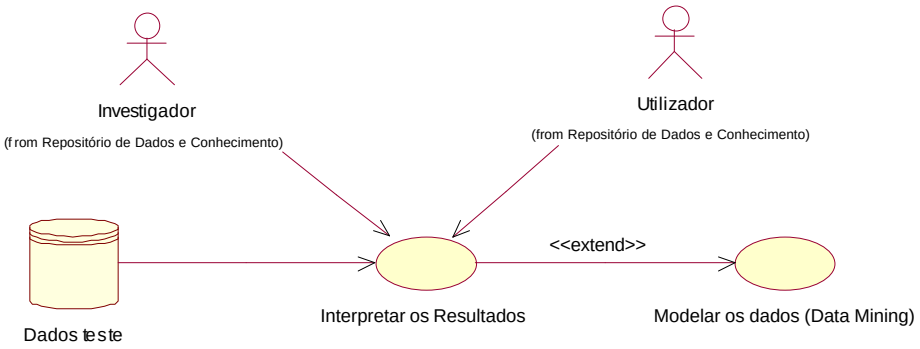


Figura 5.12: Diagrama de caso de uso para a fase de interpretação de resultados

5.2.3 O componente Visualização de Resultados

O componente de Visualização de Resultados permite o armazenamento dos padrões considerados relevantes na BDP e a sua posterior visualização em mapas das regiões analisadas. A primeira etapa, armazenamento dos padrões (Figura 5.13), cataloga todos os padrões a armazenar, aos quais relaciona posteriormente as regras que lhes estão associadas. A gestão do histórico dos padrões encontrados permite reutilizá-los em posteriores exercícios de DM.

A segunda etapa (Figura 5.14), visualização de padrões em SIG, adiciona uma nova facilidade ao sistema. O facto do Padrão analisar BD geo-referenciadas e permitir detectar padrões geo-referenciados, sugere que estes possam ser visualizados num formato adequado. Os mapas constituem este formato, permitindo neste caso uma visualização mais amigável dos resultados obtidos no processo de descoberta de conhecimento.

A BDP possui uma estrutura extremamente simples, na qual é possível catalogar temporalmente os padrões encontrados, assim como armazenar as regras que descrevem determinado padrão. Esta estrutura permite seleccionar a informação a visualizar em determinada ocasião, assim como utilizar conjuntos de regras em posteriores exercícios de DM. A Figura 5.15 apresenta o diagrama de classes que define a estrutura lógica desta BD.

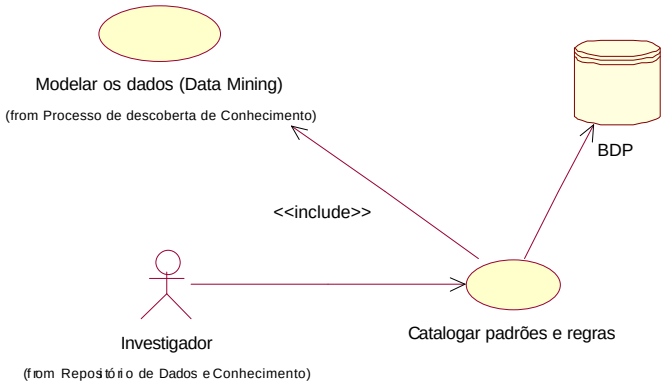


Figura 5.13: Diagrama de caso de uso para a etapa de armazenamento de padrões

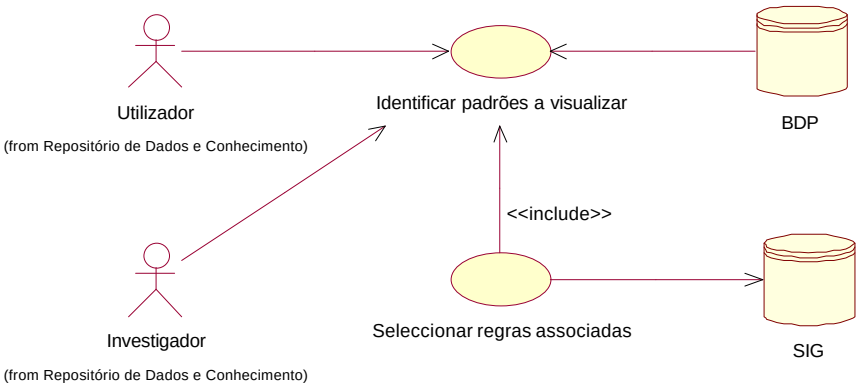


Figura 5.14: Diagrama de caso de uso para a etapa de visualização de padrões

5.3 Implementação do sistema PADRÃO

Em termos tecnológicos, o sistema Padrão foi implementado recorrendo a SGBD relacionais, nomeadamente ao Microsoft Access. As BD construídas são acedidas via ligações ODBC. A ferramenta de DCBD utilizada é o Clementine [SPSS, 1999b][SPSS, 1999a], na qual foi possível implementar todas as fases do componente de Análise de Dados do sistema Padrão. O SIG utilizado é o Geomedia Professional v3 [Intergraph, 1999b], no qual a BDP é integrada com a cartogra...a das regiões analisadas, permitindo a visualização dos padrões encontrados nos mapas das respectivas regiões.

As próximas subsecções descrevem a utilização dada às ferramentas seleccionadas para a implementação de cada um dos componentes do sistema Padrão.

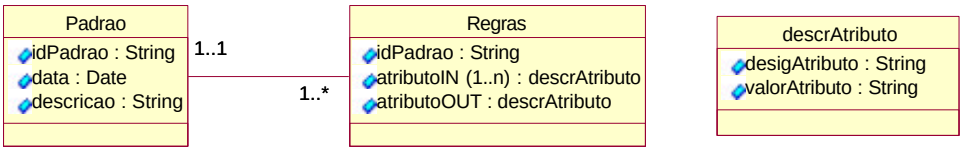


Figura 5.15: Diagrama de classes: estrutura da BDP

5.3.1 O componente Repositório de Dados e Conhecimento

O Repositório de Dados e Conhecimento integra, como já referido anteriormente, três BD centrais: a BDG, a BCE e a BDnG. A BDG e a BCE foram construídas em Microsoft Access. A BDnG, e uma vez que é fornecida por terceiros (proveniente de um domínio de aplicação específico), é no contexto do Padrão convertida para uma BD Access. A escolha do Microsoft Access foi motivada pelo facto desta ferramenta permitir o acesso aos dados em máquinas com menores capacidades computacionais (nomeadamente ao nível da memória central e velocidade de processamento). Contudo, os dados podem ser armazenados noutro SGBD, tal com o SQL Server ou o Oracle, desde que o mesmo suporte ligações ODBC para acesso aos dados.

O conteúdo da BDG e da BCE é de seguida descrito, salientando os módulos automáticos construídos para o carregamento de algumas das suas tabelas.

Base de Dados Geográfica

A BDG armazena a informação proveniente do esquema de identificadores geográficos e do esquema espacial apresentados no Capítulo 3. A Figura 5.4, apresentada anteriormente, destacou as entidades destes dois esquemas, relevantes para o processo de descoberta de conhecimento. Para o carregamento na BDG, da informação associada a estas entidades, verificaram-se duas situações:

- a informação geográfica estava disponível, uma vez que foi fornecida com os mapas das regiões que constituem o domínio geográfico analisado neste trabalho;
- a informação espacial estava implícita nos mapas utilizados.

No primeiro caso, apenas foi necessário transferir os dados existentes em folhas EXCEL[15], para as respectivas tabelas da BD. No segundo caso, e apenas ao que diz respeito à informação a integrar na componente topológica do esquema espacial, era necessário adoptar um mecanismo automático que permitisse extrair tais informações dos mapas utilizados.

Cada um destes casos é de seguida brevemente descrito. Antes de passar à descrição, convém recordar que o domínio geográfico considerado caracteriza as subdivisões administrativas

[15] Os mapas utilizados neste trabalho foram obtidos via SNIG (http://snig.cnig.pt). O proprietário dos mesmos, a Direcção Geral do Ambiente, disponibiliza um pacote que pode ser descarregado através da Internet, que contém além dos ficheiros com a cartografia, ficheiros com listagens dos dados associados, como sejam listas de freguesias, concelhos, etc.

de Portugal, ao nível dos Concelhos e Distritos[16].

O carregamento da componente da BDG respeitante ao esquema de identi...cadores geográ-...cos é extremamente simples, bastando apenas transferir a informação dos ...cheiros EXCEL para as respectivas tabelas da BD em Access. Após a transferência, o catálogo de localizações ...ca construído, armazenando não só os identi...cadores geográ...cos utilizados na geo-referenciação da informação, como também as hierarquias conceptuais existentes entre os diversos identi...cadores utilizados, isto é, entre as diversas subdivisões administrativas consideradas (Figura 5.16).

Instancias

idGeografico	classe
101	Concelho
102	Concelho
103	Concelho
104	Concelho
105	Concelho
106	Concelho
107	Concelho
108	Concelho
109	Concelho

Hierarquias

idGeografico	idSuperior
101	AVR
102	AVR
103	AVR
104	AVR
105	AVR
106	AVR
107	AVR
108	AVR
109	AVR

ID_Alternativos

idGeografico	idAlternativo
AVR	Aveiro
BGC	Braganca
BJA	Beja
BRG	Braga
CBR	Coimbra
CTB	Castelo Branco
EVR	Evora
FAR	Faro
GRD	Guarda

Figura 5.16: Excerto das tabelas Instâncias, Hierarquias e Identificadores Alternativos que integram o Catálogo de Localizações

Na componente da BDG referente ao esquema espacial, é necessário explicitar as relações espaciais existentes entre faces (regiões) adjacentes. Estas são obtidas recorrendo aos centróides das respectivas regiões. Para que estes dados possam ser carregados automaticamente nas respectivas tabelas, é necessário transferir[17] os mapas disponibilizados pela Direcção Geral do Ambiente (DGA) para o Geomedia. Após a transferência, e uma vez que nesta fase o mapa não é mais do que um agregado de linhas e pontos[18], estas linhas têm de ser agrupadas por forma a constituírem áreas. Cada área é posteriormente classi...cada como uma região, à qual é atribuída a sua identi...cação. Este procedimento é repetido para todo o mapa, isto é, até estarem de...nidas todas as faces que o constituem. Refere-se que a execução deste procedimento foi necessária pelo facto dos mapas utilizados não terem ainda de...nida a topologia dos objectos. A Figura 5.17 evidencia o processo de identi...cação de uma dada região, neste caso uma freguesia de uma das ilhas dos Açores, cujos limites foram devidamente seleccionados por forma a integrarem uma face.

[16] Esta opção deriva do facto de se ter adoptado o Concelho, como ponto de referência para a generalização das hierarquias geográ...cas. No processo de descoberta de conhecimento, todas as análises serão efectuadas ao nível do Concelho ou Distrito, permitindo uma maior redução do tamanho da amostra.

[17] A transferência só é possível depois dos mapas terem sido convertidos para um formato adequado. Neste caso, a conversão foi efectuada recorrendo ao MicroStation [Intergraph, 1995], o qual permitiu criar um ...cheiro dgn, já manipulável pelo Geomedia.

[18] Para o caso dos centróides se encontrarem disponíveis na cartogra...a inicial.

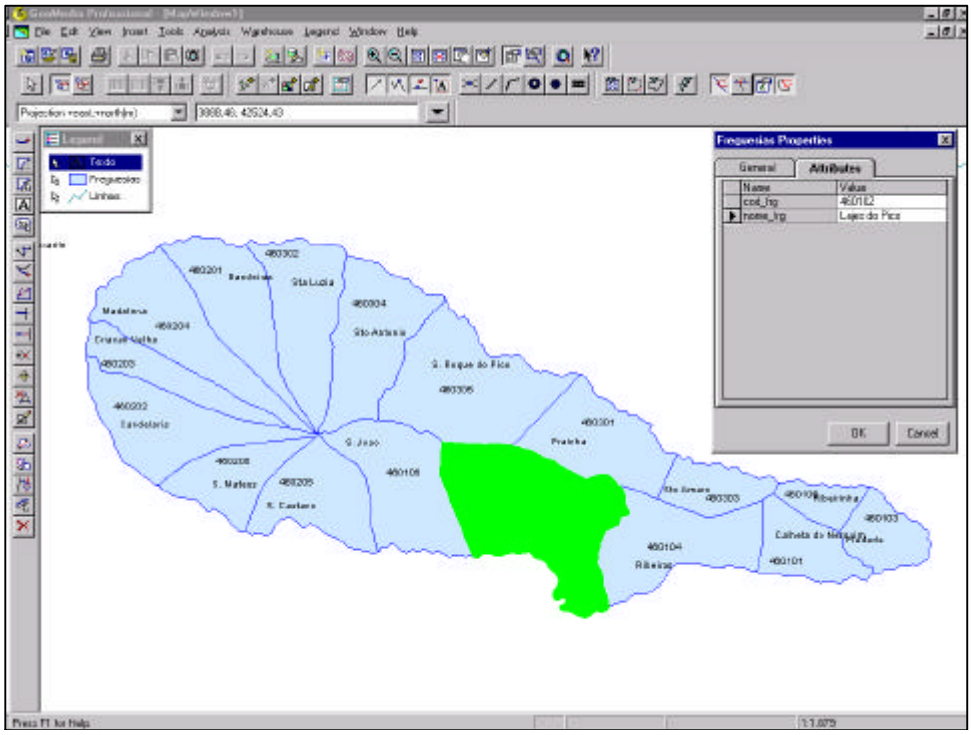


Figura 5.17: Identi…cação de uma face no Geomedia

No caso dos pontos, e representando estes os centróides das várias regiões, é necessário relacioná-los logicamente com as suas faces, isto é, caracterizá-los também em termos topológicos. Duas abordagens são aqui possíveis:

1. seleccionar ponto a ponto, designando para cada um deles a face a que correspondem (processo manual);
2. construir uma aplicação em Visual Basic[19] (VB), que automaticamente determine o ponto que está geometricamente contido numa dada região. Como resultado, a tabela Nodo Isolado é preenchida automaticamente, relacionando os centróides com as respectivas regiões, ou seja, os nodos isolados com as faces que os contêm.

Esta última foi a abordagem seguida, uma vez que utilizando os mesmos princípios foi possível preencher a tabela Faces, a qual armazena as relações espaciais, do tipo direcção e distância, existentes entre regiões adjacentes (estando desta forma implícita a relação topológica existente entre as mesmas).

A construção de rotinas em VB é motivada pelo facto deste conseguir manipular os objectos geográ…cos disponibilizados pelo Geomedia, permitindo, externamente ao SIG, manipu-

[19] A opção de utilização do Visual Basic advém do facto do Geomedia [Intergraph, 1999a] disponibilizar um conjunto de bibliotecas de objectos (baseadas na arquitectura COM, Component Object Model), que permite a construção de aplicações personalisadas, implementadas para a execução de uma dada tarefa. A versão do Visual Basic utilizada foi o Microsoft Visual Basic 6.0, for 32-bit Windows Development.

lar toda a informação armazenada por este. Neste projecto, foram construídos três módulos: o primeiro (módulo AssociaCentróide), associa cada um dos centróides existentes no mapa, à face a que corresponde (carregamento da tabela NodoIsolado). O segundo (módulo DetAdjacentes), percorre todas as faces existentes no mapa, e para cada uma delas determina as que lhe são adjacentes, e ainda, a direcção e a distância existente entre elas. O último módulo (CalculoCentróides), e apesar de não ter sido utilizado[20] no caso geográ…co em estudo, permite criar, para o caso destes não existirem no mapa inicial, os centróides de cada uma das regiões. Cada um destes módulos é de seguida brevemente descrito.

Módulo AssociaCentróide

O módulo AssociaCentróide permite associar um ponto, que geometricamente representa o centróide de determinada região, à face a que pertence. Basicamente, percorre todas as regiões existentes na BDG (já de…nidas como faces) e identi…ca o ponto que está no seu interior. Esta rotina recorre ao operador espacial gmsqContains, como pode ser veri…cado na Figura 5.18, uma vez que este permite identi…car a área geométrica na qual determinado ponto está inserido. A execução deste módulo preenche integralmente a tabela NodoIsolado. O código VB correspondente ao módulo AssociaCentróide pode ser consultado no Apêndice C.

```
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "SpatialPoint"
    .Table = "Centroides"
    .SpatialFilter = geomBlob
    .SpatialOperator = gmsqContains
End With

Set geomBlob = Nothing
Set objPoint = objOPipe1.OutputRecordset
Set objOPipe1 = Nothing
If Not objPoint.EOF Then
    objPoint.MoveFirst
    objResult.AddNew            'Upgrade tabela NodosIsolados
    With objResult
        .GFields(0).Value = objPoint(1).Value       'ID ponto
        .GFields(1).Value = objRecord1(4).Value     'ID face
    End With
    objResult.Update

End If
```

Figura 5.18: Fragmento do módulo AssociaCentróide

[20] Este módulo foi construído com o objectivo de estar disponível, caso os mapas a utilizar não apresentem a geometria dos centróides já de…nida.

Módulo DetAdjacentes

Este módulo percorre todas as faces (regiões) existentes no mapa e determina para cada uma delas, as faces que lhe são adjacentes e ainda, a direcção e a distância existente entre elas. A direcção e a distância são quantitativamente obtidas recorrendo à geometria Euclidiana, na qual são consideradas as coordenadas cartesianas dos pontos que representam cada um dos centróides. O procedimento DetCoordenadasCentróides (Apêndice C) determina o ângulo que descreve a posição do centróide do objecto de referência, em relação ao centróide do objecto primário, e ainda, a distância existente entre estes centróides.

A Figura 5.19 evidencia um pequeno fragmento do código que integra este módulo. Pela análise da referida …gura, é possível veri…car que o operador espacial gdbTouches é utilizado na identi…cação das regiões que são adjacentes a uma dada região. A listagem com o código VB que integra este módulo pode ser consultada no Apêndice C.

```
' determinação do conjunto de registos que satisfazem a condição
      objConnet1.CreateOriginatingPipe objOPipe2
      With objOPipe2
          .GeometryFieldName = "SpatialArea"
          .Table = "Limites"
          .SpatialFilter = geomBlob
          .SpatialOperator = gdbTouches
      End With

' carregamento dos registos para objRecord2
      Set objRecord2 = objOPipe2.OutputRecordset
      Set objOPipe2 = Nothing
      objRecord2.MoveLast
      objRecord2.MoveFirst

' armazenar todos os concelhos adjacentes ao concelho analisado
      If Not (objRecord2.EOF And objRecord2.BOF) Then
          Do Until objRecord2.EOF
              If Not (objRecord1.GFields(4).Value = objRecord2.GFields(4).Value)
                objResult.AddNew
                Call DetCoordenadasCentroides
                With objResult
                   .GFields(0).Value = objRecord1(4).Value
                   .GFields(1).Value = objRecord2(4).Value
                   .GFields(2).Value = CInt(angulo)
                   .GFields(3).Value = CInt(distancia)
                End With
                objResult.Update
              End If
              objRecord2.MoveNext
      Loop
```

Figura 5.19: Fragmento do módulo DetAdjacentes

Módulo CalculoCentróides

Este módulo permite determinar o centróide de uma dada região. Para os casos em que este é necessário, e não está presente na geometria do mapa utilizado, o objecto geográ…co CenterPointPipe permite criar tal geometria (Figura 5.20). Mais uma vez, o Apêndice C integra a listagem com o código VB deste módulo.

```
Set objRecord = objOPipe.OutputRecordset

Set objDB = CreateObject("Access.GDatabase")
objDB.OpenDatabase "d:\maribel\doutoramento\BasesDados\BD_Geografica.mdb"
Set objResultTransf = objDB.OpenRecordset("PontosCentroides", gdbOpenDynaset)
If Not objResultTransf.EOF Then
   objResultTransf.MoveLast
End If

Set objCentroid = CreateObject("GeoMedia.CenterPointPipe")

With objCentroid
   Set .InputRecordset = objOPipe.OutputRecordset
   .InputGeometryFieldName = objOPipe.GeometryFieldName
   .OutputGeometryFieldName = "Centroid"
End With

Set objResult = objCentroid.OutputRecordset
```

Figura 5.20: Fragmento do módulo CalculoCentróides

Os módulos apresentados anteriormente permitem carregar automaticamente as tabelas da BDG, nomeadamente as relacionadas ao esquema espacial. No sentido de exempli…car o conteúdo das mesmas, a Figura 5.21 apresenta pequenos fragmentos das tabelas Face e NodoIsolado, cujos registos foram automaticamente armazenados pelos módulos acima descritos.

Face

idFace	faceAdjacente	direccao	distancia
1602	1610	40	10
1610	1608	43	12
1608	1604	77	15
1604	1603	82	18
1604	1605	43	18
1603	1601	38	21
1609	1607	73	15
1605	1607	4	17
1601	1607	49	25

NodoIsolado

idNodo	idFace
46	1602
58	1610
78	1608
102	1604
140	1603
47	1609
80	1605
106	1601
76	1607

Figura 5.21: Excerto das tabelas Face e NodoIsolado do Esquema Espacial

Base de Conhecimento Espacial

A BCE armazena a tabela de composição que integra relações espaciais do tipo direcção, distância e topologia, segundo os princípios do raciocínio espacial qualitativo. Armazena, ainda, os

intervalos de validade quantitativos associados aos identi…cadores qualitativos utilizados, para as relações espaciais do tipo direcção e distância (já que a topologia constitui um conceito qualitativo). A tabela de composição SistemaIntegrado (apresentada anteriormente no diagrama de classes da Figura 5.6) armazena as regras de inferência obtidas no Capítulo 3, secção 3.5.3, na construção do sistema de raciocínio espacial integrado.

No que diz respeito à tabela IntervaloValidade, esta armazena os intervalos de validade quantitativos associados a cada um dos identi…cadores qualitativos utilizados. Para o caso da direcção, e como já referido anteriormente, os identi…cadores adoptados são N, NE, E, SE, S, SO, O e NO, sendo os intervalos de validade associados de [337.5º, 22.5º), [22.5º, 67.5º), [67.5º, 112.5º), [112.5º, 157.5º), [157.5º, 202.5º), [202.5º, 247.5º), [247.5º, 292.5º) e [292.5º, 337.5º), respectivamente.

A de…nição dos intervalos de validade para os indicadores de distância está intimamente associada ao contexto no qual os dados vão ser analisados, devendo sempre ser precedida da identi…cação da extensão da região geográ…ca objecto de estudo. No caso da análise visar, por exemplo, todo Portugal continental, a de…nição dos intervalos deverá considerar a distância mínima e máxima que pode existir entre regiões.

Para o caso de Portugal continental, a veri…cação da distância mínima e máxima que separa dois concelhos, permite adoptar o ratio 4 (Tabela 3.7, Capítulo 3), ampliado por um factor 10. Neste caso, e para os identi…cadores mp, p, d e md, os intervalos de validade considerados são (0, 10], (10,50], (50, 210] e (210, 850], respectivamente.

Se por outro lado, o contexto geográ…co em análise se restringir a um único distrito, então a ampliação, a ser necessária, deverá ser efectuada por um factor inferior. Poderá, ainda, ser equacionada a escolha de um outro ratio, que melhor caracterize o espaço em análise. A Figura 5.22 evidencia o conteúdo das tabelas SistemaIntegrado e IntervaloValidade para os intervalos acima de…nidos.

SistemaIntegrado

relDir1	relDis1	relTop1	relDir2	relDis2	relTop2	infDir	infDis	infTop
N	mp	adj	N	mp	adj	N	mp	adj
N	mp	adj	N	p	desl	N	p	desl
N	mp	adj	N	d	desl	N	d	desl
N	mp	adj	N	md	desl	N	md	desl
N	p	adj	N	p	adj	N	d	desl
N	p	adj	N	p	desl	N	d	desl
N	p	adj	N	d	desl	N	d	desl
N	p	adj	N	md	desl	N	md	desl

IntervaloValidade

idRelacao	limInferior	limSuperior
N	337,5	22,5
NE	22,5	67,5
E	67,5	112,5
SE	112,5	157,5
S	157,5	202,5
SO	202,5	247,5
O	247,5	292,5
NO	292,5	337,5

Figura 5.22: Excerto do conteúdo das tabelas SistemaIntegrado e IntervaloValidade

5.3.2 O componente Análise de Dados

A implementação do componente de Análise de Dados foi integralmente realizada no Clementine, apenas recorrendo à construção de um módulo externo em VB, para auxiliar o Clementine no processo de inferência da informação geo-espacial. Este módulo foi necessário pelo facto do Clementine, à data de utilização neste trabalho, não possuir mecanismos de manipulação de arrays. Como poderá ser constatado posteriormente, este módulo apenas combina regiões cujas relações espaciais são desconhecidas, permitindo que as regras de inferência qualitativas possam ser utilizadas para a sua obtenção.

Este componente do Padrão foi implementado recorrendo a diversas streams, que realizam as diferentes fases do processo de descoberta de conhecimento. Antes de iniciar a descrição das mesmas, é apresentada uma breve introdução ao Clementine e ao seu modo de funcionamento.

O Clementine é um sistema de descoberta de conhecimento baseado em programação visual, que inclui diversas técnicas de aprendizagem automática, como seja a indução de regras ou as redes neuronais. Disponibiliza ferramentas para manipular, explorar, visualizar e construir modelos sobre os dados.

Toda a …loso…a de trabalho do Clementine assenta na construção de streams, nas quais cada operação sobre os dados é representada por um nodo. Nodos com funções similares encontram-se agrupados em palettes, permitindo ao utilizador escolher o nodo mais apropriado para a execução de determinada tarefa. As palettes disponíveis são:

- Acesso aos dados (Sources), na qual são disponibilizados diversos mecanismos de acesso aos dados, desde ligações ODBC, …cheiros de texto com atributos de tamanho …xo ou variável, etc.
- Operações sobre registos (Records Ops), cujos nodos permitem efectuar diversas operações sobre registos, como sejam, seleccionar os registos que veri…cam determinada condição, balancear os dados, etc.
- Operações sobre atributos (Fields Ops), na qual podem ser encontrados nodos para seleccionar atributos, derivar novos atributos a partir de atributos existentes, preencher campos atendendo a determinada condição, etc.
- Grá…cos (Graphs), cujos nodos permitem explorar os dados utilizando, por exemplo, histogramas;
- Modelação (Modelling), a qual disponibiliza diversos algoritmos de DM, que utilizam técnicas como redes neuronais, árvores de decisão, redes Kohonen, regras de associação, etc.
- Resultados (Output), cujas tabelas ou nodos com funções estatísticas permitem visualizar ou analisar os resultados associados à realização de determinada tarefa de DM.

A Figura 5.23 apresenta uma janela com a interface do Clementine, na qual é possível visualizar a área destinada à construção das streams, as diversas palettes disponíveis, e ainda, a palette que armazena os modelos gerados pelo Clementine. Estes modelos podem ser gravados

ou simplesmente seleccionados como outro nodo qualquer, permitindo a sua utilização em outras streams, isto é, noutras tarefas de DM.

Figura 5.23: Interface do Clementine

Selecção, tratamento e pré-processamento dos dados

As três primeiras fases do componente de Análise de Dados, selecção, tratamento e pré- processamento dos dados, são precedidas da de…nição do objectivo da descoberta. Esta de…nição é normalmente efectuada pelo utilizador ou perito do domínio de aplicação, e é posteriormente transformada no objectivo do DM. Uma vez de…nido, e dependendo do grau de di…culdade associado às tarefas a efectuar, os passos de selecção, tratamento e pré-processamento dos dados, podem ou não ser agrupados numa única stream.

As descrições de seguida apresentadas recorrem a exemplos que derivam da exploração de uma BD demográ…ca, [Santos e Amaral, 2000a] [Santos e Amaral, 2000d] [Santos e Amaral, 2000c] [Santos e Amaral, 2000b], que armazena os registos paroquiais datados entre 1690 e 1990 no distrito de Aveiro. Não se considera relevante, nesta fase, fornecer mais detalhes sobre a referida BD, uma vez que os exemplos apenas são utilizados para descrever como o Clementine é utilizado pelo Padrão. Contudo, a Figura 5.24 apresenta um pequeno fragmento da tabela Indivíduo, armazenada na BD demográ…ca explorada, a qual contextualiza os dados utilizados nas descrições de seguida apresentadas.

A Figura 5.25 apresenta a stream construída para seleccionar, tratar e pré-processar os dados considerados relevantes para a análise (nesta fase da descrição, sem preocupações quanto

Num	Name	S	Birth date	Birth place	Died	Died place	Occupation:	M	Ch
6224	JOAO ANTON	M	18-03-1790	Arada	01-10-1847	Arada	Oleiro	1	12
6232	TERESA LOF	F	13-05-1790	Coimbrão	08-06-1830	Quinta do Picai	Oleira	1	8
6233	ANTONIO DA	M	24-05-1790	Quinta do Picai	16-09-1864	Quinta do Picai	Oleiro	2	10
6235	JOSE FRANC	M	28-05-1790	Quinta do Picai	05-10-1849	Verdemilho	Lavrador	1	10
6239	MANUEL FR/	M	03-08-1790	Quinta do Picai	01-08-1830	Quinta do Picai	Lavrador	1	7
6241	ROSA DOS S	F	25-08-1790	Bom Sucesso	27-08-1830	Quinta do Picai	Lavradora	1	7
6249	MANUEL JOA	M	25-09-1790	Verdemilho	20-09-1841	Verdemilho	Lavrador	1	10
6250	ANTONIO SIN	M	21-09-1790	Arada	07-03-1874	Arada	Lavrador	1	9
6253	JOANA MARI	F	31-10-1790	Verdemilho	06-03-1863	Verdemilho	Lavradora	1	10
6257	FRANCISCO	M	10-11-1790	Bom Sucesso	26-05-1831	Bom Sucesso	Lavrador	1	4
6259	JOAQUINA M	F	17-12-1790	Verdemilho	24-03-1864	Verdemilho	Lavradora	1	9
6260	BERNARDO	M		Quinta da Gran	21-08-1843	Verdemilho	Lavrador	1	8
6261	JOAQUINA F	F	26-11-1767	Verdemilho	31-12-1823	Verdemilho	Lavradora	1	8
6267	PERPETUA M	F	08-03-1791	Verdemilho	10-12-1855	Verdemilho	Lavradora	1	10
6288	MARIA DE JE	F	06-06-1791	Arada	24-03-1877	Arada	Jornaleira	0	0
6299	JOSEFA DE .	F		Quinta do Picai	30-03-1836	Quinta do Picai	Lavradora	1	9
6314	JOANA TERE	F	05-11-1791	Arada	17-04-1870	Arada	Mendicante	1	4
6331	MAURICIO FI	M	09-06-1792	Quinta do Picai	27-02-1858	Quinta do Picai	Lavrador	1	7
6335	MARIA ROSA	F	07-08-1792	Quinta do Picai	20-05-1870	Quinta do Picai	Lavradora	1	9
6337	MIGUEL FER	M	25-08-1792	Arada	24-12-1876	Arada	Lavrador	1	11
6338	MANUEL DA	M	27-08-1792	Bom Sucesso	28-09-1855	Bom Sucesso	Jornaleiro	1	6
6340	ANTONIA DO	F	28-10-1792	Bom Sucesso	25-05-1868	Arada	Lavradora	1	11

Figura 5.24: Fragmento da tabela Indivíduo

aos objectivos da mesma). Refere-se que a stream apresentada começa por extrair uma amostra de treino, do conjunto de dados armazenado na tabela Indivíduo. Estes dados são acedidos via uma ligação ODBC, representada na stream pelo nodo DB_AVR:INDIVíDUO. Esta tarefa de construção da amostra é efectuada recorrendo ao nodo sample[21], que neste caso selecciona aleatoriamente, do conjunto de dados inicial, 50% dos registos.

A amostra obtida é posteriormente …ltrada com o nodo filter, permitindo seleccionar o conjunto de atributos relevante para a tarefa de DM que foi de…nida. O nodo filler é utilizado para substituir todos os valores omissos por uma marca ('?'), que os identi…ca como desconhecidos, e que permite aos algoritmos de DM que estes valores sejam tratados como tal. Nesta fase, é possível proceder à redução do tamanho da amostra, através da generalização dos dados e/ou transformação dos atributos com valores contínuos em atributos com valores discretos. No exemplo apresentado, recorre-se ao nodo Age para determinar a idade dos indivíduos, a partir da data de nascimento e da data de óbito dos mesmos. Esta idade é posteriormente utilizada pelo nodo Age_class, que designa a classe a que corresponde cada idade. Na mesma …gura, Figura 5.25, é ainda possível visualizar parte da tabela com os dados já seleccionados, tratados e pré-processados, atendendo às operações executadas sobre os mesmos.

O cálculo da idade dos indivíduos é efectuada, no nodo Age, recorrendo a duas funções disponibilizadas pelo Clementine. O Clementine inclui o CLEM (Clementine Language for Expression Manipulation), uma linguagem com estruturas e funções próprias, que permitem a manipulação dos dados. A Figura 5.26 apresenta o nodo Age e ainda o código CLEM utilizado na determinação da idade dos indivíduos. Pela análise da referida …gura veri…ca-se que as funções utilizadas são a intof e a date-years-difference. Esta última efectua a subtracção das duas datas, dando como resultado o número de anos. Este valor é posteriormente transformado, pela

[21] Este nodo é aqui incluído com o objectivo de exempli…car o processo através do qual é possível criar a amostra de treino e a amostra de teste.

função intof, num número inteiro.

Processamento da informação geo-espacial

Após o pré-processamento dos dados impõe-se a junção da componente geo-espacial, no processo de descoberta de conhecimento. As relações armazenadas na BDG podem ser utilizadas para limitar o contexto geográ…co em análise, como por exemplo, todos os concelhos próximos de ..., a Norte de ..., do distrito de ..., etc.

A selecção da componente geográ…ca pode ser efectuada recorrendo aos nodos de manipulação de dados disponibilizados pelo Clementine. Os dados seleccionados são posteriormente analisados e processados, com vista à inferência de relações espaciais desconhecidas, necessárias no processo de descoberta de conhecimento. Antes de poder utilizar a tabela de composição[22] construída neste trabalho para inferir relações espaciais do tipo direcção, distância e topologia, é necessário gerar os modelos (um para cada tipo de relação espacial) que permitam a utilização das regras contidas na referida tabela. Assim, e depois da BCE armazenar o conhecimento explícito nesta tabela, é possível utilizá-lo na construção de árvores de decisão, que integram este conhecimento.

A Figura 5.27 evidencia a stream que permite a construção de três árvores de decisão[23], cada uma delas tendo como função inferir um tipo de relação espacial. Analisando a stream construída, veri…ca-se que os factos armazenados na tabela SistemaIntegrado[24] são acedidos através de uma ligação ODBC. O nodo type é utilizado para de…nir o conjunto de atributos à entrada do processo de aprendizagem e aquele que representa a saída do mesmo. São assim necessários três ciclos de aprendizagem, um para cada uma das relações espaciais consideradas. O algoritmo utilizado foi o C5.0[25], criando três modelos, infDir, infDis e infTop, que se complementam no processo de raciocínio. Na mesma …gura é ainda possível visualizar um fragmento do conjunto de regras, que integram cada um dos modelos obtidos.

Após a aprendizagem das regras de inferência, passa-se a descrever a fase de processamento da informação geo-espacial, através da apresentação de três exercícios distintos. O primeiro visa evidenciar a stream que ciclicamente infere todas as relações espaciais desconhecidas, inseridas em determinado contexto geográ…co. No segundo exercício, utilizam-se as relações espaciais inferidas no primeiro exercício, na construção de um modelo que geogra…camente descreve uma dada região. O terceiro exercício, e com o objectivo de tirar partido das hierarquias geográ…cas existentes na BDG, permite inferir a localização dos distritos, expressa através de relações espaciais, a partir do conhecimento disponível para concelhos adjacentes. Estes três exercícios são de seguida apresentados.

---

[22] As diversas tabelas de composição utilizadas neste trabalho, nomeadamente para o ratio 2, ratio 4 e ratio 5 entre distâncias, são apresentadas no Apêndice D, secção D.4. Estas tabelas resultam do processo de avaliação efectuado, no Capítulo 6, ao sistema de inferências. Esta avaliação ao sistema de inferências permitiu aumentar o desempenho das regras, e foi efectuada antes de etapa de DM, apresentada de seguida neste capítulo, para não comprometer a validade das regras encontradas nesta fase.

[23] Este processo é semelhante ao apresentado em [Santos et al., 1999] para a inferência de relações espaciais do tipo direcção, e em [Santos e Amaral, 1999] para a integração de relações espaciais do tipo direcção e distância.

[24] Como poderá ser constatado mais tarde, existem várias tabelas, uma vez que estas dependem do ratio entre distâncias utilizado: ratio 2, ratio 4 ou ratio 5.

[25] O algoritmo C5.0 permite a construção de árvores de decisão, nas quais é possível prever o valor de um atributo de saída (output) baseado em diversos atributos de entrada (input).

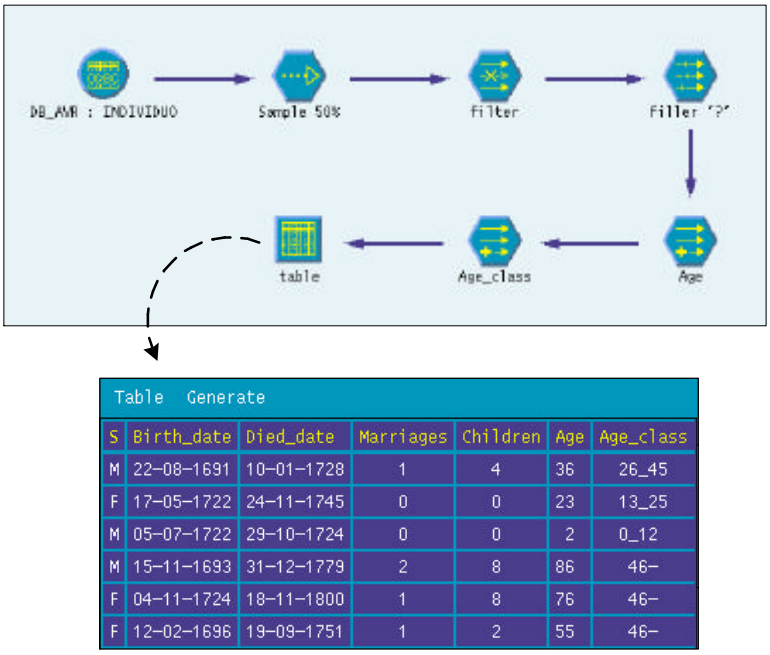


Figura 5.25: Stream para as fases de selecção, tratamento e pré-processamento dos dados

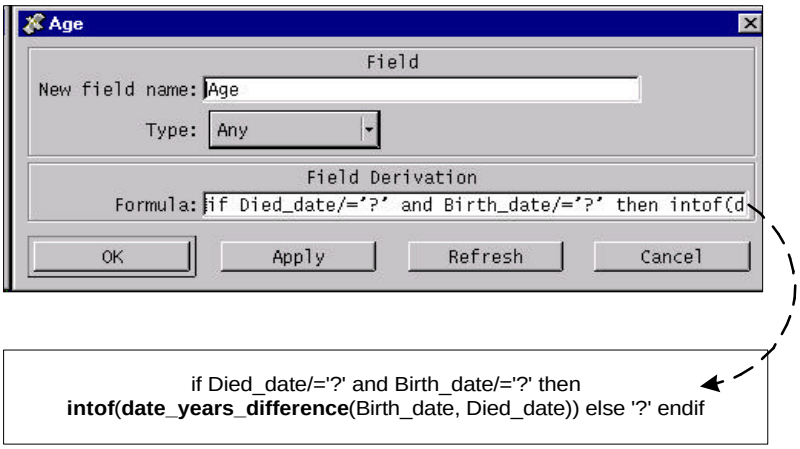


Figura 5.26: Funções CLEM utilizadas pelo nodo Age

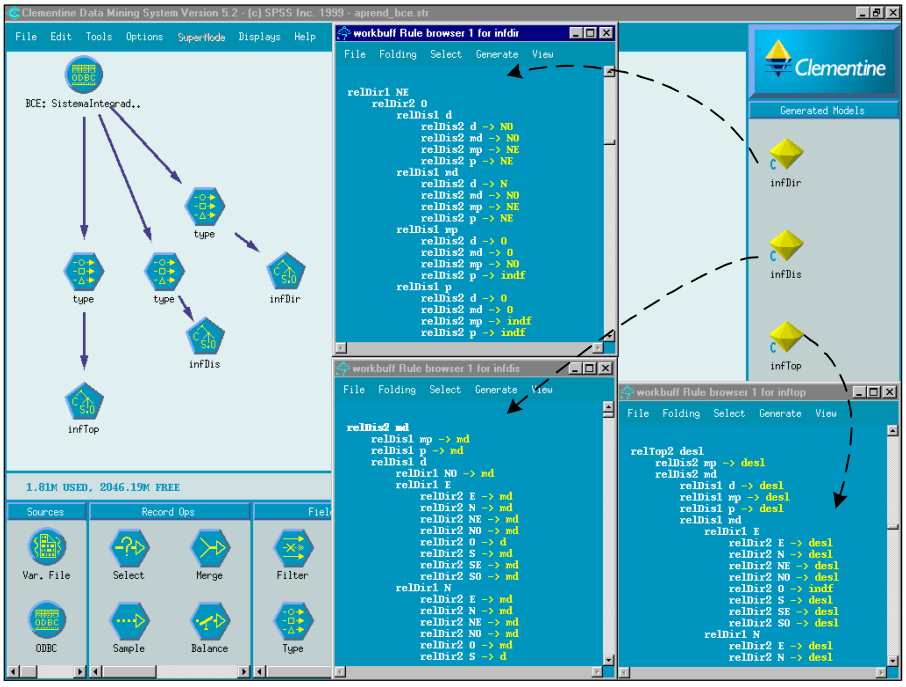


Figura 5.27: Processo de aprendizagem das regras de inferência armazenadas na tabela `Sistema Integrado`

1: Selecção do domínio geográ…co e inferência de todas as relações espaciais desconhecidas.

Neste exemplo é dado a conhecer o processo cíclico que permite combinar os factos armazenados na BDG, por forma a criar pares de regiões para as quais as relações espaciais não são conhecidas. Para estes pares, os modelos `infDir`, `infDis` e `infTop`, são utilizados na inferência das relações espaciais existentes entre os mesmos. A Figura 5.28 apresenta a stream construída para a selecção da componente geográ…ca. Na mesma, é possível veri…car que o nodo `merge` é utilizado para integrar a tabela Faces com a tabela `Hierarquias`, permitindo ao nodo `select` seleccionar o domínio geográ…co em causa. Neste exemplo, são seleccionadas as relações espaciais existentes entre os concelhos do distrito de Aveiro. Para este distrito, 78 relações espaciais se encontram explícitas na BDG. Após a selecção, de…niram-se os intervalos de validade quantitativos, que permitem a substituição das direcções e distâncias[26] quantitativas por identi…cadores qualitativos.

Esta transformação é efectuada pelos nodos `classeDir` e `classeDis`. Posteriormente, foi criado um novo atributo, `topologia`, que explicita a relação espacial do tipo `adjacente` existente entre os registos seleccionados da BDG. Finalmente, o nodo `BDG:geoAveiro` permite a transferência, via ligação ODBC, dos registos processados para uma nova tabela intitulada

[26] Para este distrito, e analisando a distância mínima e máxima que pode existir entre dois concelhos adjacentes, refere-se que os intervalos de validade adoptados para a distância foram: `mp` (0,7], `p` (7,22], `d` (22,51] e `md` (51,110], obtidos por ampliação, `factor` 7.3, dos intervalos correspondentes ao `ratio` 2. Em relação à direcção, os intervalos de validade quantitativos adoptados foram: (337.5,22.5], (22.5,67.5], (67.5,112.5], (112.5,157.5], (157.5,202.5], (202.5,247.5], (247.5,292.5] e (292.5,337.5], de `N` a `NO` respectivamente. Estas opções são justi…cadas no Capítulo 6, na avaliação do desempenho do sistema de inferências.

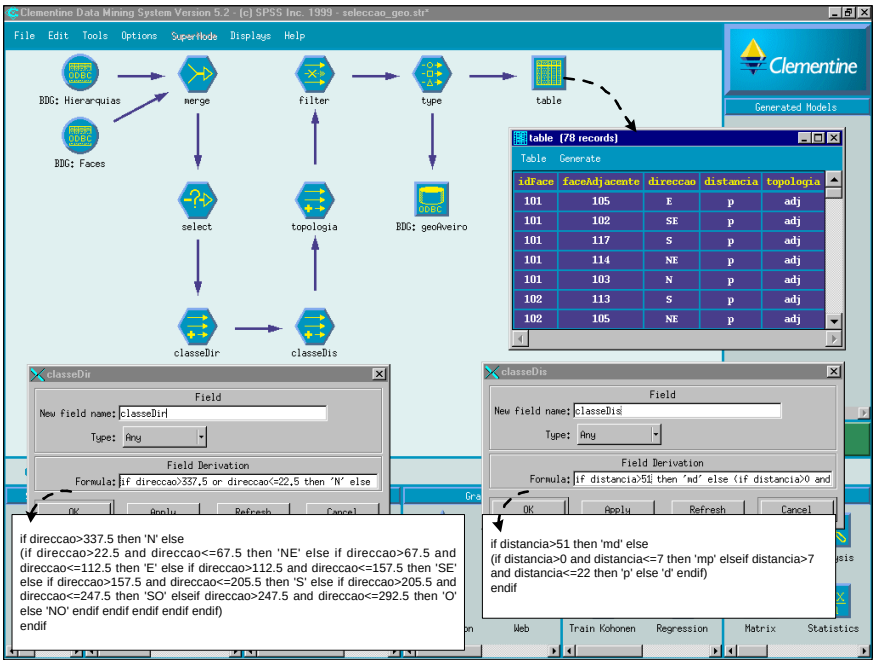


Figura 5.28: Selecção e tratamento da componente geográ…ca

geoAveiro, que será utilizada no processo de inferência. Esta primeira stream (Figura 5.28) permitiu a selecção da informação geográ…ca disponível na BDG para o distrito em análise, a transformação dos valores quantitativos associados à direcção e distância em valores qualitativos, e ainda, a explicitação da relação topológica existente entre as entidades geográ…cas seleccionadas. Na …gura apresentada, é possível visualizar uma tabela com um pequeno fragmento dos registos transformados pela stream e ainda, o código CLEM associado aos nodos `classeDir` e `classeDis`.

Os registos armazenados na tabela geoAveiro podem então ser manipulados e utilizados na inferência de relações espaciais desconhecidas. Para inferir as relações espaciais que podem existir entre todos os concelhos do distrito em análise, foi construída uma stream que é executada ciclicamente, até que não existam mais relações a inferir, isto é, relações entre concelhos desconhecidas. A execução cíclica é conseguida recorrendo a uma script em linguagem CLEM. A stream recorre a um programa externo[27], que permite a combinação de concelhos por forma a que a inferência das relações espaciais seja possível. O programa foi construído em Basic[28] e pode ser utilizado em qualquer stream, já que este é incorporado no Clementine recorrendo a um …cheiro de especi…cação (.spc, Speci…cation File). Este procedimento permite que o módulo Combina seja disponibilizado no Clementine, na palette Record Ops, possibilitando a sua utilização na manipulação de registos.

A Figura 5.29 apresenta a stream construída para a inferência de relações espaciais desconhecidas. Nesta …gura, é possível veri…car a utilização do nodo Combina e a sua integração na

[27] Uma vez que a linguagem CLEM não disponibiliza mecanismos de manipulação de arrays.

[28] O módulo em Basic construído foi compilado no Visual Basic 6, permitindo criar um …cheiro executável, .exe, que é posteriormente integrado no Clementine.

palette Record Ops. Os registos resultantes do processo de inferência são armazenados na tabela geoAveiro da BDG, permitindo a utilização dos mesmos nas próximas iterações do processo.

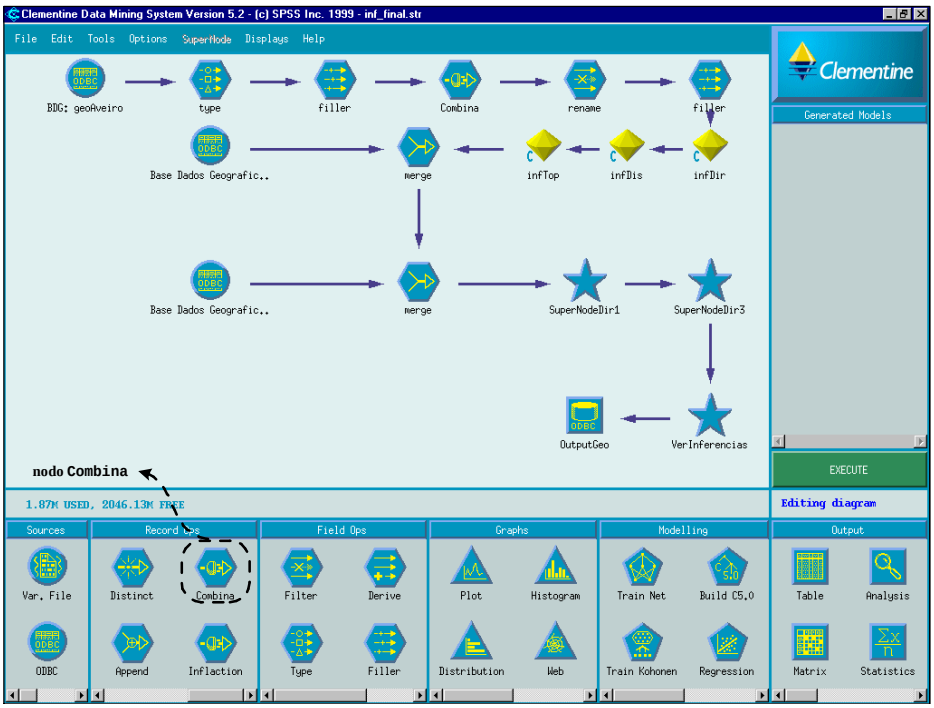


Figura 5.29: Processo de inferência de relações espaciais desconhecidas

Uma das características deste sistema de inferências é que uma mesma relação espacial entre entidades pode ser obtida através de diferentes caminhos, isto é, através da combinação de diferentes relações. Por exemplo, conhecendo-se os factos A Norte B, A Este D, B Este C e D Norte C, a direcção existente entre A e C pode ser inferida combinando A Norte B com B Este C ou combinando A Este D com D Norte C. Esta particularidade permite validar as inferências obtidas, uma vez que se o resultado é diferente, dependendo do percurso seguido, o mesmo não é considerado válido e como tal não é integrado no conjunto de resultados de uma dada iteração. A veri…cação destes casos é efectuada pelo super nodo[29] VerInferencias (Figura 5.29), o qual …ltra os casos em que não existe concordância de resultados.

Na avaliação do desempenho do sistema de inferências, descrita no Capítulo 6, constatou-se que a dimensão das regiões envolvidas na composição, desempenha um papel extremamente importante na determinação da direcção existente entre as mesmas. Tal facto permitiu a identi…cação de diversas regras (apresentadas no Apêndice D, subsecção D.5.3), que possibilitam a inclusão desta característica no processo de raciocínio. Uma vez que estas regras dizem respeito a casos muito especí…cos, nomeadamente para composições de direcções pertencentes ao grupo ¢dir1 e ¢dir3, optou-se por não proceder a qualquer alteração das árvores de decisão construídas anteriormente, infDir, infDis e InfTop. As regras obtidas são integradas em dois super

[29] Os super nodos disponibilizados pelo Clementine permitem a integração de diversos nodos, funcionando como procedimentos ou subrotinas que executam determinado conjunto de tarefas. Possuem a grande vantagem de simpli…car a leitura das streams construídas.

nodos, SuperNodeDir1 e SuperNodeDir3 (Figura 5.29), os quais alteram as inferências obtidas através das árvores, sempre que o tamanho das regiões envolvidas in‡uenciar o resultado.

As novas relações espaciais existentes entre concelhos não adjacentes, e inferidas através deste processo cíclico, constituem conhecimento já disponível na BDG, que pode ser utilizado no processo de descoberta de conhecimento. Apesar da avaliação ao desempenho do sistema de inferências ser apresentada mais tarde, no Capítulo 6, é possível constatar através da Figura 5.30, com um fragmento do conjunto de registos obtido e um mapa da região processada, que as inferências obtidas representam conhecimento válido, que pode ser utilizado nas próximas fases do processo de descoberta de conhecimento. O exemplo apresentado evidencia algumas das relações existentes entre o município 101 e outros concelhos do distrito, relações que foram inferidas pelo processo aqui apresentado.

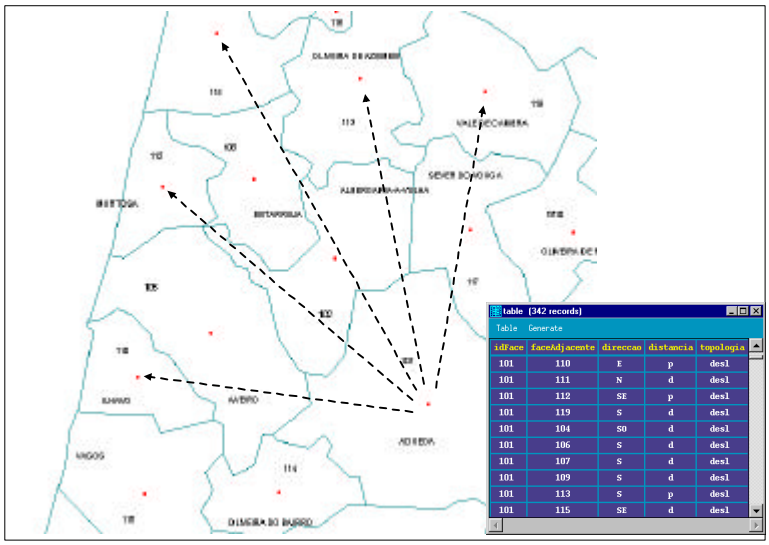


Figura 5.30: Veri…cação das inferências obtidas

A construção e inclusão do módulo Combina no Clementine permitiu adquirir experiência acerca do processo de integração de programas externos nesta ferramenta. Estes programas permitem adicionar novas funcionalidades, e podem ser escritos em qualquer linguagem de programação que permita a construção de um …cheiro executável (.exe). A Figura 5.31 apresenta o …cheiro de especi…cação construído para a incorporação do nodo Combina na palette Records Ops. Pela análise da referida especi…cação constata-se que uma vez que o módulo incorporado não constitui um nodo terminal, o mesmo recebe do Clementine um conjunto de registos (INPUT_DATA), o qual processa e devolve devidamente transformado (OUTPUT_DATA).

2: Construção de um modelo geográ…co.

Depois da identi…cação das relações espaciais que existem entre os concelhos de um determinado distrito, é possível utilizar este conhecimento na construção de um modelo que geogra…camente descreve a região em estudo, isto é, que localiza a posição de cada um dos concelhos, em relação ao distrito. Os modelos assim construídos podem posteriormente ser utilizados na fase de DM, na determinação de padrões ou outros relacionamentos implícitos existentes entre os dados geo-

```
SPECFILE
        NODE
              NAME    Combina
              TITLE   'Combina'
              TYPE PROCESS
              PALETTE RECORD
        ENDNODE

CORE
        PARAMETERS
               outstem pathname 'Regions'
        ENDPARAMETERS

        EXECUTE
               COMMAND 'd:/maribel/doutoramento/clementine/combreg.exe'
        ENDEXECUTE

        OPTIONS
               outstem [outstem]
        ENDOPTIONS

        CONTROLS
             outstem LABEL 'File'
        ENDCONTROLS

        INPUT_FIELDS
            INCLUDE ALL
        ENDINPUT_FIELDS

        OUTPUT_FIELDS
            EXTEND
                  FOREACH FIELD INCLUDE DIRECTION [IN]
                        CREATE NAME ['new' >< FIELD.NAME] TYPE [FIELD.TYPE]
                  ENDFOREACH
        ENDOUTPUT_FIELDS

        INPUT_DATA
             FILE_NAME [outstem]
             INC_FIELDS false
        ENDINPUT_DATA

        OUTPUT_DATA
            FILE_NAME [outstem >< '.res']
            SEPARATOR ','
            INC_FIELDS false
        ENDOUTPUT_DATA

ENDCORE

ENDSPECFILE
```

Figura 5.31: Ficheiro de especi…cação para o nodo Combina

espaciais e dados não geográ…cos.

Partindo do exercício anterior, no qual foi possível armazenar na tabela geoAveiro as relações espaciais que existem entre todos os concelhos do distrito, é possível, mais uma vez utilizando uma ligação ODBC, aceder aos dados armazenados na referida tabela e utilizá-los na construção do modelo geográ…co do distrito. A Figura 5.32 apresenta a stream construída para o efeito, na qual os dados da tabela geoAveiro são con…gurados no nodo type, permitindo de…nir o(s) atributo(s) de entrada e o atributo de saída, isto é, aquele que constitui o alvo do processo de aprendizagem. Os dados são então analisados pelo algoritmo C5.0, determinando o modelo geográ…co da região (geoAVR). As regras que integram este modelo, evidenciadas na mesma …gura, indicam a localização, no distrito, de cada um dos concelhos (em termos de direcção).

3: As hierarquias conceptuais na inferência de informação espacial.

Este exemplo visa evidenciar como é que as hierarquias geográ…cas podem ser utilizadas na inferência de relações espaciais desconhecidas, sem recorrer a quaisquer regras construídas segundo

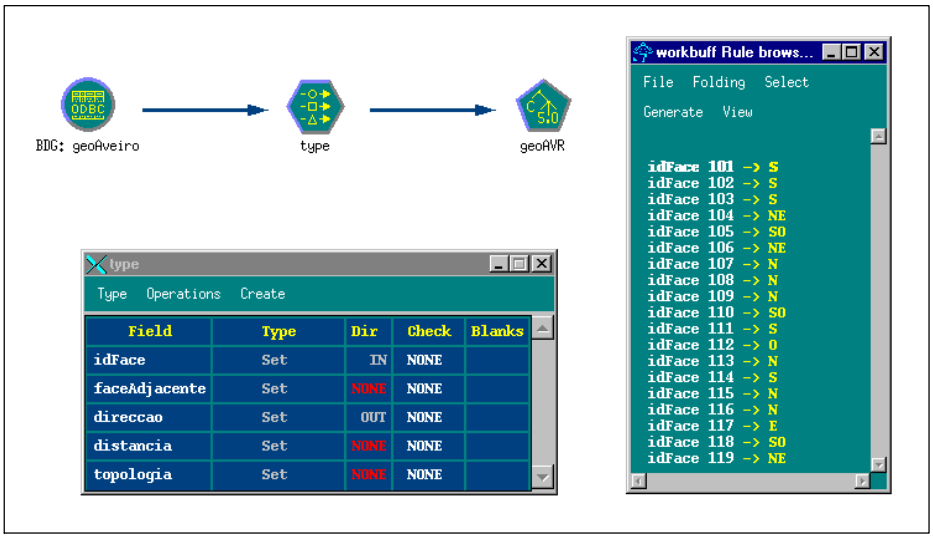


Figura 5.32: Modelo geográ…co construído para o distrito de Aveiro

os princípios do raciocínio espacial qualitativo.

A partir da informação existente entre concelhos adjacentes, armazenada na tabela Faces da BDG, é possível determinar, utilizando os mecanismos de aprendizagem disponibilizados pelo Clementine, as relações espaciais existentes entre os distritos que agregam tais concelhos. Para tal, apenas é necessário utilizar as de…nições constantes na tabela Hierarquias, a qual permite a generalização da informação explícita para o caso dos concelhos. Os registos resultantes são analisados pelo algoritmo C5.0, permitindo a identi…cação das relações espaciais existentes entre este nível hierárquico, o dos distritos.

A Figura 5.33 apresenta a stream construída para a execução deste exercício. Salienta-se que os nodos merge são utilizados para integrar a informação existente na tabela Faces com a tabela Hierarquias (para ambas as regiões, idFace e faceAdjacente), permitindo a generalização da informação explícita para o nível dos concelhos. Após a generalização, os valores quantitativos explícitos para a direcção e distância são transformados em identi…cadores qualitativos. A amostra assim construída é analisada pelo algoritmo de aprendizagem, permitindo a obtenção de um modelo, direcção, que caracteriza a relação espacial do tipo direcção, existente entre distritos adjacentes.

Uma tabela com um extracto dos resultados obtidos, com a execução da stream apresentada na Figura 5.33, é apresentada na Figura 5.34, na qual um mapa de parte da região analisada permite a avaliação de algumas das inferências obtidas. Na tabela apresentada nesta …gura, o atributo $C-classeDir indica a direcção inferida, enquanto que $CC-classeDir evidencia a con…ança do resultado. Apesar de em alguns casos este valor ser inferior a 0.5, originado pela utilização de um número reduzido de casos provenientes do nível hierárquico inferior, os resultados obtidos consideram-se satisfatórios por evidenciarem uma excelente aproximação à realidade.

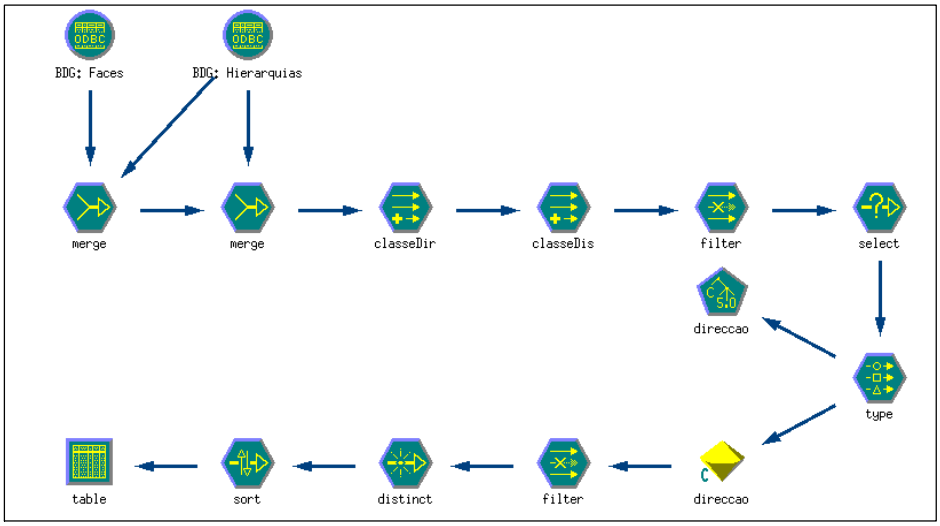


Figura 5.33: Identi…cação da direcção existente entre distritos, a partir das relações espaciais explícitas para concelhos adjacentes

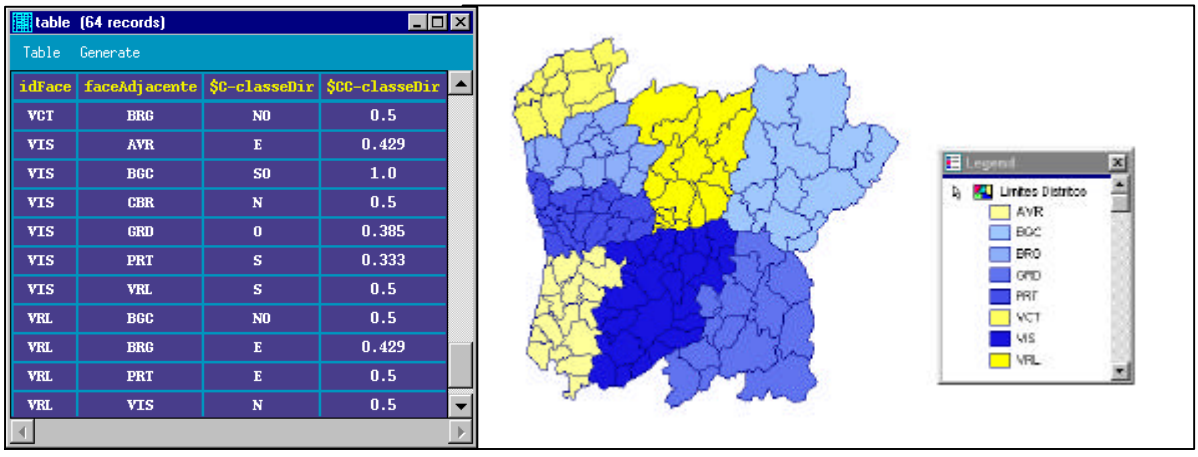


Figura 5.34: Direcção inferida para distritos adjacentes

Data Mining

Nesta fase, dados não geográficos e dados geo-espaciais são integrados por forma a permitirem a identificação de padrões ou outros relacionamentos existentes entre os mesmos. Para tal, diversas técnicas de DM podem ser utilizadas, dependendo da tarefa a executar. Para exemplificar o tipo de exercícios de DM realizados com a BD demográfica atrás mencionada, seleccionou-se um dos apresentados em Santos e Amaral [Santos e Amaral, 2000a], no qual se caracterizava a idade ao óbito dos indivíduos, atendendo ao século e região (município) em que viveram.

A Figura 5.35 apresenta a stream construída para satisfazer a tarefa de DM definida. Nesta figura, os dados provenientes da BD demográfica (DB_AVR) são integrados com o modelo geográfico da região (geoAVR), permitindo ao algoritmo C5.0 a definição de um conjunto de regras, que caracterizam o atributo idade ao óbito ao longo dos séculos analisados. Na mesma figura, é ainda possível visualizar o conjunto de regras obtido nesta etapa de DM.

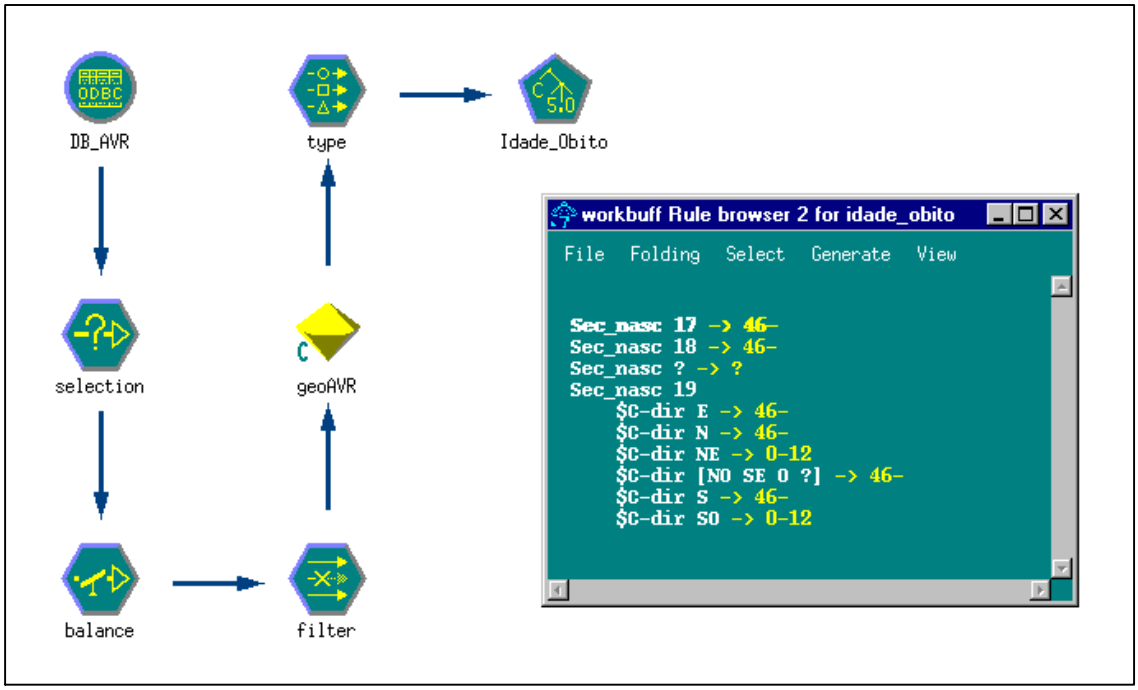


Figura 5.35: Caracterização da idade ao óbito ao longo dos séculos

As regras identificadas evidenciam uma distribuição geográfica bem demarcada, da idade ao óbito, no século 19. Nesta distribuição, os municípios localizados a NE e SO do distrito caracterizam-se por apresentar uma idade ao óbito inferior aos restantes concelhos.

Interpretação de resultados

A fase de interpretação de resultados consiste em avaliar a utilidade, para o domínio de aplicação em causa, das regras encontradas. Os modelos obtidos, e dependendo do tipo de tarefa de DM, devem ser validados com a amostra de teste, com o objectivo de verificar o seu desempenho

quando utilizados na classi…cação de dados desconhecidos. Neste caso particular, e uma vez que apenas se pretendia sistematizar o conhecimento implícito nos dados e evidenciar como a fase de DM é realizada no Clementine, não se avalia o desempenho das regras na classi…cação de dados desconhecidos[30].

Na tarefa de descrição realizada, os resultados obtidos retratam o comportamento dos dados armazenados na BD analisada. A utilidade destas regras, para o domínio de aplicação, deverá ser avaliada pelos historiadores demógrafos, os quais poderão ainda adicionar novos requisitos ao objectivo da tarefa de DM, implicando o retrocesso a uma das fases anteriores.

5.3.3 O componente Visualização de Resultados

O componente de Visualização de Resultados permite a transferência para a BDP, dos padrões relevantes encontrados na fase de DM. Simultaneamente, é possível visualizar os padrões detectados em mapas das regiões analisadas. Esta tarefa recorre à utilização de um SIG como ferramenta de integração, da BDP com a cartogra…a da região, e de suporte à visualização.

Para simpli…car a tarefa de visualização, foi construído, mais uma vez recorrendo ao VB, um módulo que permite ao utilizador a visualização das regras a partir do Clementine. Este programa externo, VisualPadrão, foi construído manipulando as bibliotecas de objectos geográ…cos disponibilizadas pelo Geomedia Professional, e foi integrado no Clementine recorrendo a um …cheiro de especi…cação, que permite que o mesmo esteja disponível na palette de Output. Esta aplicação foi construída com o objectivo de facilitar o processo de integração da BDP com a BD cartográ…ca, e disponibilizar ao utilizador um ambiente personalizado do SIG, o qual evidencia apenas a informação que transita do processo de descoberta de conhecimento (isto é, determinado conjunto de regras).

A Figura 5.36 apresenta a stream que permite a transferência das regras encontradas anteriormente (Figura 5.35) para a BDP. Nesta …gura, é ainda possível constatar que é utilizado um nodo user input, que permite ao utilizador descrever (Figura 5.37) o conjunto de regras que são armazenadas, indicando a data do exercício de DM, uma breve descrição sobre o mesmo e ainda, o nome da tabela que irá armazenar as referidas regras na BDP[31]. É a partir deste nodo de user input que deve ser chamado o VisualPadrão, uma vez que é neste que é especi…cado o nome da tabela que contém as regras a visualizar.

O processo de transferência das regras encontradas para a BDP apenas requer que o modelo gerado na fase de DM, Idade_Óbito, seja ligado à stream que lhe deu origem, permitindo que o mesmo seja relacionado com os registos da BD. Após este procedimento, é necessário …ltrar

[30] Este tipo de avaliação às regras não é neste capítulo equacionada, uma vez que é abordada no Capítulo 7, no qual se descreve o Padrão na análise de uma componente do Sistema de Administração do Pessoal do Exército.

[31] Apesar do diagrama de classes, para a BDP, apresentado na Figura 5.15 retratar a estrutura que se julga mais apropriada para o armazenamento dos resultados do processo de DM, na realidade, limitações tecnológicas provenientes do modo de funcionamento dos nodos ODBC (para saída de dados) do Clementine, motivaram a restruturação da BDP. Na versão do Clementine utilizada, estes nodos requerem que o número de colunas a transferir seja exactamente igual ao número de colunas da tabela. Nesta fase dos trabalhos, e uma vez que estes nodos deverão evoluir por forma a permitirem ao utilizador indicar que atributos serão transferidos e para que colunas da tabela destino, a concepção da BDP foi alterada, passando a integrar uma tabela de Padrões, na qual é descrita a data do exercício de DM, a descrição do mesmo, e ainda a tabela que armazena as regras resultantes do processo de descoberta de conhecimento. Para cada exercício de DM, é criada uma nova tabela na BDP, a qual é posteriormente integrada, recorrendo ao VisualPadrão, com a cartogra…a da região analisada.

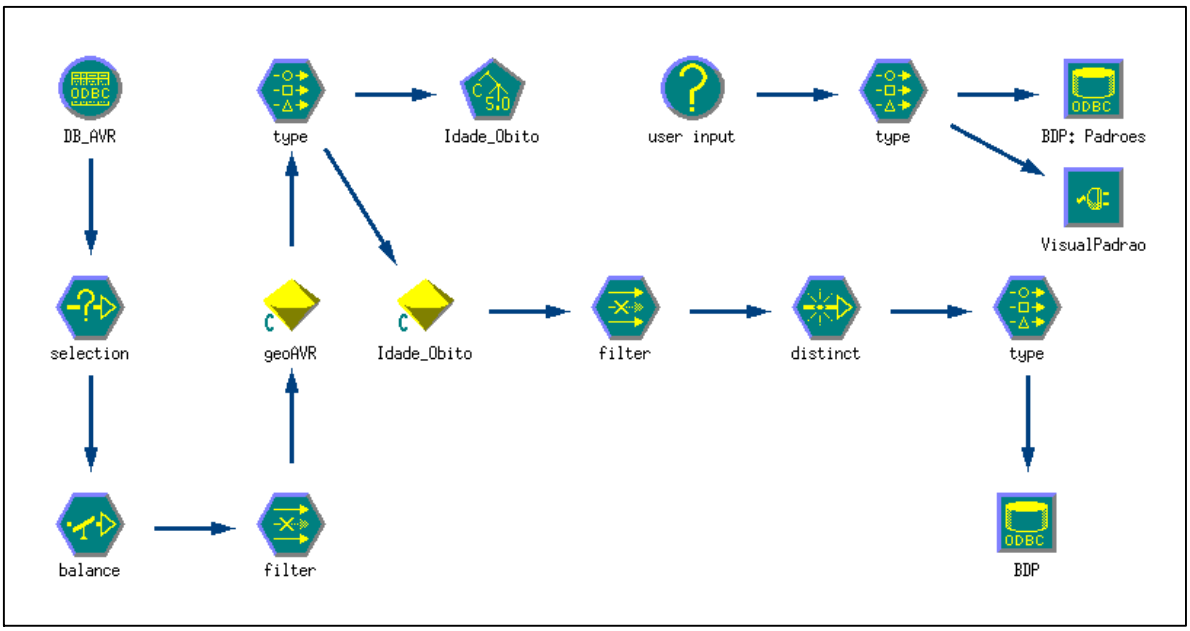


Figura 5.36: Stream para a visualização de resultados

os atributos de interesse, ou seja, seleccionar aqueles que efectivamente serão transferidos para a BDP. Esta transferência cria uma nova tabela na BD (Figura 5.37), cuja estrutura re‡ecte as opções do utilizador[32].

Na Figura 5.37, o nodo BDP: Padrões é utilizado para armazenar os parâmetros que caracterizam o exercício de DM, actualizando a tabela de Padrões. O nodo BDP é utilizado para criar uma nova tabela na BDP, para armazenar as regras, com a designação atribuída pelo utilizador (neste caso a tabela possui a designação IdadeÓbito).

Após a transferência das descrições e regras, o VisualPadrão pode ser executado. Este procedimento permite utilizar o Geomedia Professional a partir do Clementine, o qual é con…gurado por forma a permitir a visualização das regras armazenadas na nova tabela criada na BDP. A Figura 5.38 retrata o resultado deste processo, para o exercício de DM descrito até ao momento. Neste ambiente disponibilizado pelo VisualPadrão, o utilizador pode bene…ciar da utilização de um conjunto de facilidades, como seja armazenar ou imprimir o mapa em questão. É ainda possível estabelecer ligações a outras tabelas com informação geo-referenciada, permitindo a integração, e consequente análise espacial, de diversos padrões.

[32] A integração do modelo obtido na stream, adiciona duas novas colunas aos dados analisados ($C... e $CC...), permitindo ao VisualPadrão conhecer o atributo a contextualizar no mapa. No VisualPadrão, a integração da BDG com a BDP é realizada através do atributo com a designação idFace.

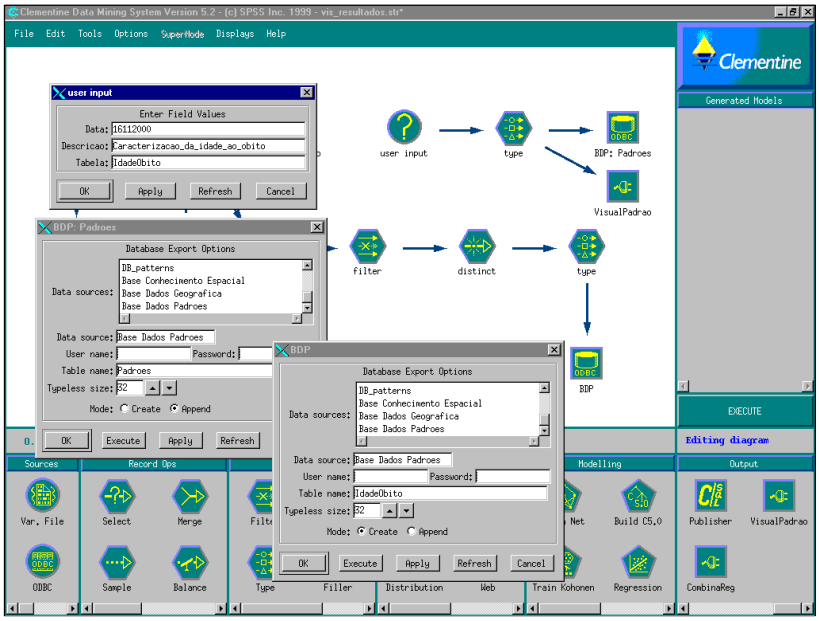


Figura 5.37: Processo de transferência das regras para a BDP

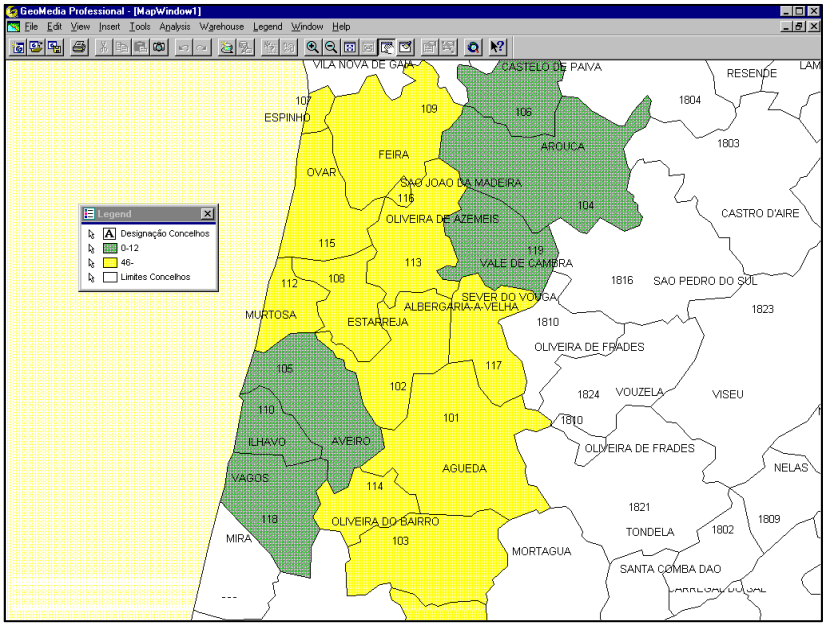


Figura 5.38: Visualização de resultados recorrendo ao módulo VisualPadrão

O código VB que integra o módulo VisualPadrão pode ser consultado no Apêndice C. O …cheiro de especi…cação, que permitiu a sua integração no Clementine, é apresentado na Figura 5.39.

```
SPECFILE
    NODE
        NAME    VisualPadrao
        TITLE   'VisualPadrao'
        TYPE TERMINAL
        PALETTE OUTPUT
    ENDNODE

    CORE
        PARAMETERS
            outstem pathname 'Padrao'
        ENDPARAMETERS

        EXECUTE
            COMMAND 'D:/Maribel/Doutoramento/Clementine/VisualPadrao.exe'
        ENDEXECUTE

        OPTIONS
            outstem [outstem]
        ENDOPTIONS

        CONTROLS
            outstem LABEL 'File'
        ENDCONTROLS

        INPUT_FIELDS
            INCLUDE ALL
        ENDINPUT_FIELDS

        INPUT_DATA
            FILE_NAME [outstem]
            INC_FIELDS true
        ENDINPUT_DATA

        RESULT
            RESULT_TYPE EXTERN
        ENDRESULT

    ENDCORE
ENDSPECFILE
```

Figura 5.39: Ficheiro de especi…cação do nodo VisualPadrão

Capítulo 6

Avaliação do desempenho do sistema PADRÃO

Após a concepção e implementação do sistema Padrão, apresentada no capítulo anterior, é necessário proceder à validação técnica do sistema, a qual será realizada em duas vertentes distintas. Em primeiro lugar, é necessário verificar o desempenho do sistema qualitativo de inferências utilizado pelo Padrão. Nesta avaliação pretende-se analisar a qualidade das inferências obtidas, e verificar a possibilidade da sua inclusão no processo de descoberta de conhecimento.

Posteriormente, é averiguada a capacidade, do sistema Padrão, de identificação de relacionamentos implícitos nos dados analisados. Esta validação tem como objectivo comprovar que a ferramenta de descoberta de conhecimento adoptada para a implementação do Padrão, o Clementine, pode ser utilizada na exploração de BD, com o objectivo de descoberta de conhecimento.

A avaliação do sistema qualitativo de inferências e a verificação da capacidade de identificação de padrões nos dados, apresentadas nas próximas secções, permitem validar tecnicamente o sistema Padrão, comprovando a viabilidade de implementação das diversas características identificadas na sua concepção. No próximo capítulo, Capítulo 7, o sistema será alvo de uma nova avaliação, na qual um estudo de caso permitirá verificar a utilidade do sistema na análise de BD organizacionais.

6.1 Avaliação do sistema qualitativo de inferências

Num sistema baseado em raciocínio espacial qualitativo é necessário encontrar representações que permitam o raciocínio com informação incompleta e imprecisa. Neste trabalho, a representação é baseada em relações espaciais explícitas, permitindo ao utilizador manipular informação espacial, independentemente da geometria dos objectos considerados.

A validade do sistema qualitativo de inferências, que permite a integração da componente geo-espacial no processo de descoberta de conhecimento, foi verificada comparando as relações espaciais reais, existentes entre as entidades geográficas, com as relações espaciais obtidas por inferência.

Para dois dos distritos que integram Portugal continental, foram geradas automaticamente, recorrendo a um módulo em VB, todas as relações espaciais que podem existir entre os seus concelhos. Este procedimento permitiu determinar com exactidão, a direcção, a distância[1] e a topologia existente entre os diversos concelhos. As relações espaciais assim determinadas foram comparadas com as obtidas pelo sistema de inferências do Padrão, permitindo verificar as diferenças existentes entre os dois conjuntos de relações.

Este processo permitiu, após a aprendizagem da tabela de composição que integra relações espaciais do tipo direcção, distância e topologia, avaliar a qualidade das inferências obtidas com a mesma. Os resultados produzidos são apresentados em detalhe nas próximas subsecções, permitindo verificar o desempenho das regras que integram o sistema de inferências. Os dois distritos analisados foram Aveiro e Braga, sendo a sua escolha justificada pelo facto do primeiro integrar regiões com dimensões bastante heterogéneas, enquanto que o segundo agrega regiões de dimensão mais similar.

6.1.1 Análise das inferências obtidas com o ratio 4

Dada a disprepância existente entre o tamanho das regiões que integram alguns dos distritos, verificou–se o comportamento do sistema de inferências para dois distritos, Aveiro e Braga, com características diferentes. Para estes dois distritos, utilizaram-se os intervalos de validade quantitativos para a distância de mp (0,10], p (10,40], d (40,130] e md (130,400], obtidos por ampliação, através do factor 10, dos intervalos que caracterizam o ratio 3[2]. As relações espaciais existentes entre concelhos adjacentes foram seleccionadas da BDG, sendo as mesmas submetidas ao processo de inferência.

As várias iterações necessárias para completar o processo de inferência foram analisadas em detalhe, permitindo identificar as alterações e/ou correcções que seria necessário efectuar. Verificando sempre o que acontece nos dois distritos seleccionados, diversas tabelas permitiram a quantificação dos resultados obtidos nas diversas iterações do processo de inferência. Esta quantificação é apresentada através de matrizes, nas quais as linhas indicam a relação espacial inferida, e as colunas a relação espacial real.

Em relação ao distrito de Aveiro, 78 relações espaciais se encontravam explícitas na BDG, sendo este número de 50 no que diz respeito ao distrito de Braga. Uma vez que o processo de inferência é cíclico, foi possível verificar os desvios ocorridos logo após a primeira iteração. Esta verificação é conseguida comparando os valores quantitativos reais, obtidos recorrendo a um módulo em VB[3], com os valores qualitativos obtidos pelo sistema de inferências. Os valores quantitativos são transformados em qualitativos atendendo aos intervalos definidos, permitindo a sua comparação com os obtidos por inferência (esta comparação é efectuada recorrendo ao nodo Matrix, um nodo terminal disponibilizado pelo Clementine). Três nodos Matrix foram necessários, um para cada tipo de relação espacial. A Figura 6.1 apresenta a stream Clementine construída para a verificação das inferência obtidas. Na referida figura é ainda possível visualizar as três matrizes que sintetizam os resultados alcançados, para o distrito de Aveiro, na primeira iteração deste processo.

[1] A direcção e a distância foram calculadas recorrendo à posição dos centróides das diversas regiões.

[2] Destaca-se que as regras de inferência para o ratio 3 e ratio 4 são semelhantes [Hong, 1994].

[3] Este módulo é muito semelhante aos apresentados no Capítulo 5, utilizados no carregamento da BDG. Contudo, o código que o integra pode ser consultado no Apêndice C (módulo VerRelações).

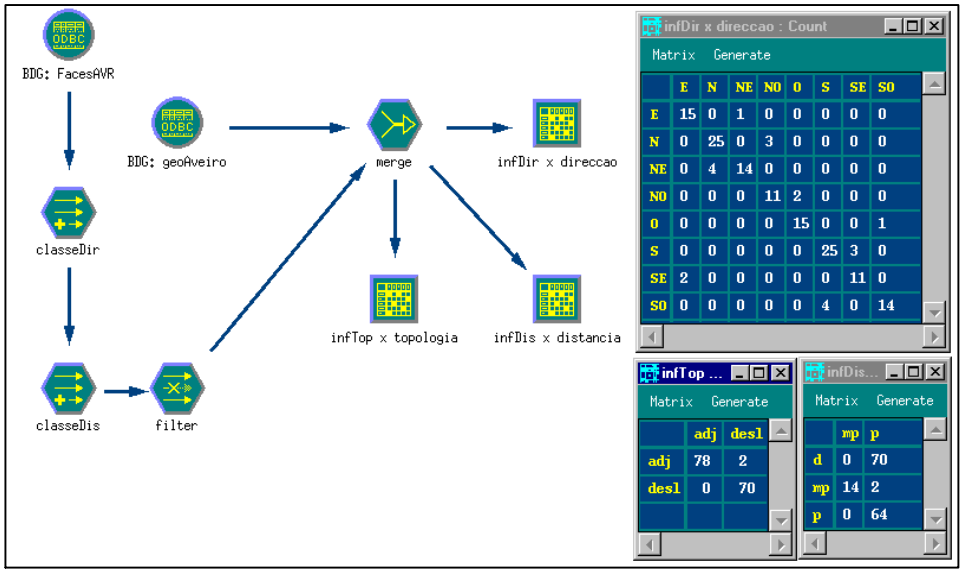


Figura 6.1: Stream que permite veri…car os desvios ocorridos no sistema qualitativo de inferências

Braga

	N	NE	E	SE	S	SO	O	NO
N	5	0	0	0	0	0	0	1
NE	1	8	0	0	0	0	0	0
E	0	4	14	0	0	0	0	0
SE	0	0	6	11	0	0	0	0
S	0	0	0	1	5	0	0	0
SO	0	0	0	0	1	8	0	0
O	0	0	0	0	0	3	14	0
NO	0	0	0	0	0	0	6	11

	mp	p	d	md
mp	4	2	0	0
p	0	46	0	0
d	0	46	1	0
md	0	0	0	0

	adj	desl
adj	50	2
desl	0	47

Tabela 6.1: Desempenho do processo de inferência, na primeira iteração, para o distrito de Braga

Para este distrito é possível constatar que, e no que diz respeito à direcção, em 20 casos (em 72 inferências) a relação espacial inferida é vizinha da que realmente deveria ter sido obtida. Em relação à distância, todas as inferências apresentam desvios, principalmente no que diz respeito à catalogação de relações como d, quando na realidade são p. Este facto é provavelmente motivado pelos limites de…nidos para os intervalos de validade, podendo estes ser inadequados para a região geográ…ca em análise. Em relação à topologia, detectaram-se apenas dois valores errados, mais uma vez podendo ser motivados pela escolha dos intervalos de validade para a distância.

Em relação ao distrito de Braga, e partindo dos mesmos intervalos de validade, a primeira iteração do processo de inferência originou também alguns desvios. A Tabela 6.1 apresenta as matrizes que quanti…cam os resultados obtidos. Pela análise da tabela é possível veri…car que continuam a detectar-se diversos desvios no que diz respeito à classi…cação das relações espaciais.

Pela análise dos resultados obtidos na primeira iteração para os dois distritos, veri…ca-se que em ambos existem variações no que diz respeito à relação inferida/relação real. Analisando

detalhadamente cada um dos casos, detectou-se que para a direcção, na maioria das composições do tipo ¢dir1, com a mesma distância qualitativa, a relação a inferir deveria ter sido a anterior à aquela que foi inferida. Por exemplo, todos os casos que envolviam pares NE;E, o resultado encontrado era E quando na realidade o correcto seria NE. Esta constatação chamou a atenção para o facto de, na veri…cação quantitativa realizada às regras de inferência propostas por Hong [Hong, 1994], apresentada no Apêndice D (secção D.1), nas composições para o grupo ¢dir1 com distâncias qualitativas iguais, os valores quantitativos obtidos para a direcção coincidirem precisamente com o limite de…nido para os intervalos. A alteração dos limites dos intervalos, no que diz respeito apenas ao "abrir" e "fechar" dos mesmos, resolveria esta situação, mas acarretaria também alterações nas composições do grupo ¢dir3, pelo que serão avaliadas novas iterações antes de estabelecer conclusões de…nitivas.

Ao nível da topologia, o distrito de Aveiro apresenta dois registos em que a relação inferida foi adjacente quando deveria ter sido deslocado. Analisando as regiões envolvidas, veri…cou-se que existe uma notável discrepância[4] entre a dimensão das mesmas. Duas das regiões representam os dois municípios mais pequenos do distrito, que estão a ser combinados na composição com uma região de grande dimensão, inferindo erradamente a relação topológica adjacente. Esta incorrecção pode ser resolvida optimizando os intervalos de validade para a distância, já que esta é a relação espacial que apresenta mais incorrecções no processo de inferência, e como tal impõe-se o seu ajuste.

No que diz respeito ao distrito de Braga, ao nível topológico veri…caram-se também duas anomalias, apesar das regiões apresentarem dimensões mais homogéneas. Em relação à direcção, veri…caram-se os mesmos problemas detectados no distrito de Aveiro, e que estão essencialmente associados ao grupo de composições ¢dir1, em que as distâncias envolvidas sejam p;p.

As diversas variações detectadas chamaram a atenção para o facto de poderem existir erros na tabela de composição, ou então, a escolha dos intervalos de validade para as distâncias não ter sido a mais apropriada. Cada uma destas hipóteses é averiguada nas próximas subsecções.

A continuação do processo de inferência introduz novos desvios no resultado, alguns originados pelos mesmos factores que originaram os ocorridos na primeira iteração, outros obtidos por composição de relações que não são as mais correctas. Apesar disto, partiu-se para uma nova iteração, no sentido de veri…car a degradação que vai sendo introduzida no sistema. Os resultados obtidos com esta segunda iteração são apresentados na Tabela 6.2, na qual é possível constatar que a classi…cação de direcções vizinhas continua a ocorrer, e que a distância d continua em muitos casos a ser erradamente inferida.

Uma análise mais cuidada a alguns dos registos obtidos, e que apresentaram desvios na inferência, permitiu veri…car que o grupo ¢dir3 apresenta os mesmos problemas do ¢dir1, para o caso de distâncias qualitativas iguais. Tal facto sugere que os limites quantitativos adoptados para a direcção devem ser reajustados.

Em relação à distância, e dado o elevado número de valores classi…cados como d em vez de p, os intervalos quantitativos utilizados devem também ser adaptados, ou então deve ser equacionada a utilização de outro *ratio*, por forma a que a composição p;p, para ¢dir1, in…ra a relação correcta.

[4] Para se ter uma ideia da discrepância que existe entre a dimensão das regiões envolvidas, refere-se que as duas mais pequenas apresentam áreas de 8.6 km^2 e 18.7 km^2, enquanto que a terceira região envolvida possui uma área de 218.3 km^2.

Aveiro

	N	NE	E	SE	S	SO	O	NO
N	**46**	1	0	0	0	0	0	10
NE	12	**32**	0	0	0	0	0	0
E	0	5	**20**	2	0	0	0	0
SE	0	0	3	**17**	0	0	0	0
S	0	0	0	9	**45**	6	0	0
SO	0	0	0	0	14	**24**	0	0
O	0	0	0	0	0	8	**20**	3
NO	1	0	0	0	0	0	3	**15**

	mp	p	d	md
mp	**14**	3	0	0
p	0	**64**	0	0
d	0	167	**19**	0
md	0	0	29	**0**

	adj	desl
adj	**78**	3
desl	0	**215**

Braga

	N	NE	E	SE	S	SO	O	NO
N	**8**	0	0	0	0	0	0	2
NE	2	**10**	0	0	0	0	0	0
E	0	6	**23**	0	0	0	0	0
SE	0	0	10	**16**	0	0	0	0
S	0	0	0	1	**6**	0	0	0
SO	0	0	0	0	4	**9**	0	0
O	0	0	0	0	0	7	**19**	0
NO	0	0	0	0	0	0	14	**15**

	mp	p	d	md
mp	**4**	2	0	0
p	0	**46**	0	0
d	0	74	**17**	0
md	0	0	9	**0**

	adj	desl
adj	**50**	2
desl	0	**100**

Tabela 6.2: Valores obtidos na segunda iteração

6.1.2 Análise às regras de inferência

Uma vez que existiram 20 e 23 registos com inferências deslocadas, para Aveiro e Braga respectivamente, ao nível da direcção, e 72 e 48 registos, para Aveiro e Braga respectivamente, com inferências deslocadas ao nível da distância, a primeira tarefa efectuada foi a veri…cação das regras explícitas na tabela de composição utilizada. Este procedimento permitiu veri…car que uma das regras proposta por Hong[5] [Hong, 1994], nomeadamente para a inferência das direcções pertencentes ao grupo ¢dir1, está errada. Para o caso particular da composição do par (N,mp);(NE,mp) a relação a inferir é (NE,mp) e não (N,mp). Esta alteração propaga-se a todas as restantes composições deste grupo, que são: (NE,mp);(E,mp), (E,mp);(SE,mp), (SE,mp);(S,mp), (S,mp);(SO,mp), (SO,mp);(O,mp), (O,mp);(NO,mp) e (NO,mp);(N,mp). A Tabela 6.3 apresenta os cálculos quantitativos que permitiram a detecção desta incorrecção. Todas as restantes regras de inferência foram também veri…cadas, para os grupos ¢dir0, ¢dir2, ¢dir3 e ¢dir4, apresentando o Apêndice D, secção D.1, os cálculos quantitativos que certi…cam a veracidade das mesmas. Recorda-se que estas regras de inferência foram obtidas com os intervalos de validade para a direcção de [337.5,22.5), [22.5,67.5), [67.5,112.5), [112.5,157.5), [157.5,202.5), [202.5,247.5), [247.5,292.5) e [292.5,337.5), de N a NO respectivamente. No que diz respeito à distância, utilizou-se o `ratio` 4.

[5] Recorda-se que a tabela de composição proposta por Hong [Hong, 1994], e que integra relações espaciais do tipo direcção e distância, é neste trabalho utilizada para construir um sistema que integra relações espaciais do tipo direcção, distância e topologia.

(N,mp);(NE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	NE,mp	

(N,mp);(NE,p)	V_{AB}	V_{BC}
distância	0,5	3
direcção	180	225
Ang_{AC}=	38,9817	
$\|V_{AC}\|$=	3,37214	
	NE,p	

(N,mp);(NE,d)	V_{AB}	V_{BC}
distância	0,5	13
direcção	180	225
Ang_{AC}=	43,4834	
$\|V_{AC}\|$=	13,3582	
	NE,d	

(N,mp);(NE,md)	V_{AB}	V_{BC}
distância	0,5	53
direcção	180	225
Ang_{AC}=	44,6203	
$\|V_{AC}\|$=	53,3547	
	NE,md	

(N,p);(NE,mp)	V_{AB}	V_{BC}
distância	3	0,5
direcção	180	225
Ang_{AC}=	6,01826	
$\|V_{AC}\|$=	3,37214	
	N,p	

(N,p);(NE,p)	V_{AB}	V_{BC}
distância	3	3
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	5,54328	
	NE,d	

(N,p);(NE,d)	V_{AB}	V_{BC}
distância	3	13
direcção	180	225
Ang_{AC}=	37,0143	
$\|V_{AC}\|$=	15,2694	
	NE,d	

(N,p);(NE,md)	V_{AB}	V_{BC}
distância	3	53
direcção	180	225
Ang_{AC}=	42,7961	
$\|V_{AC}\|$=	55,1621	
	NE,md	

(N,d);(NE,mp)	V_{AB}	V_{BC}
distância	13	0,5
direcção	180	225
Ang_{AC}=	1,51663	
$\|V_{AC}\|$=	13,3582	
	N,d	

(N,d);(NE,p)	V_{AB}	V_{BC}
distância	13	3
direcção	180	225
Ang_{AC}=	7,98572	
$\|V_{AC}\|$=	15,2694	
	N,d	

(N,d);(NE,d)	V_{AB}	V_{BC}
distância	13	13
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	24,0209	
	NE,md	

(N,d);(NE,md)	V_{AB}	V_{BC}
distância	13	53
direcção	180	225
Ang_{AC}=	36,5922	
$\|V_{AC}\|$=	62,8681	
	NE,md	

(N,md);(NE,mp)	V_{AB}	V_{BC}
distância	53	0,5
direcção	180	225
Ang_{AC}=	0,37967	
$\|V_{AC}\|$=	53,3547	
	N,md	

(N,md);(NE,p)	V_{AB}	V_{BC}
distância	53	3
direcção	180	225
Ang_{AC}=	2,20392	
$\|V_{AC}\|$=	55,1621	
	N,md	

(N,md);(NE,d)	V_{AB}	V_{BC}
distância	53	13
direcção	180	225
Ang_{AC}=	8,40777	
$\|V_{AC}\|$=	62,8681	
	N,md	

(N,md);(NE,md)	V_{AB}	V_{BC}
distância	53	53
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	97,9312	
	NE,md	

Tabela 6.3: Cálculos quantitativos para o grupo de direcções ¢dir1, ratio 4

6.1.3 Análise aos limites quantitativos dos intervalos de validade

A alteração dos limites dos intervalos de validade para a distância não conduz a qualquer aumento do desempenho do sistema de inferências, uma vez que a tabela de composição permanece inalterada para todos os intervalos construídos sobre o ratio 3 ou o ratio 4. Partiu-se, então, à procura de um outro ratio, que permitisse caracterizar e…cientemente o espaço geográ…co analisado. Veri…caram-se as regras de inferência para o ratio 2 e para o ratio 5, uma vez que estes dois ratios permitem construir intervalos de validade, para a distância, com características diferentes.

O Apêndice D, secções D.2 e D.3, apresenta os cálculos quantitativos realizados para a determinação das regras de inferência associadas a cada um destes ratios. Estas regras foram obtidas utilizando os intervalos de validade para a direcção de: (337.5, 22.5], (22.5, 67.5], (67.5, 112.5], (112.5, 157.5], (157.5, 202.5], (202.5, 247.5], (247.5, 292.5] e (292.5, 337.5], de N a NO respectivamente. No caso da distância, o ratio 2 é caracterizado pelos intervalos de validade (0, 1], (1, 3], (3, 7] e (7, 15], de mp a md respectivamente, enquanto que para o ratio 5, os intervalos de validade utilizados são: (0, 1], (1, 6], (6, 31] e (31, 156], de mp a md respectivamente.

O Apêndice D apresenta as tabelas de composição que permitem a integração da direcção, distância e topologia, para os diferentes ratios considerados. Estas tabelas armazenam as regras de inferência utilizadas neste trabalho para a inclusão da componente geo-espacial dos dados no processo de descoberta de conhecimento. A subsecção D.4.1, do referido apêndice, apresenta as

regras de inferência para o conjunto de distâncias representadas pelo `ratio 2`; a subsecção D.4.2 apresenta as regras de inferência para o `ratio 4`, enquanto que a subsecção D.4.3 apresenta as regras de inferência para o `ratio 5`. Recorda-se que estes três conjuntos de regras utilizam os novos limites quantitativos, de…nidos nesta subsecção, para a direcção.

6.1.4 Análise das inferências obtidas com o `ratio 2`

Veri…cando a distância mínima e máxima que existe entre concelhos adjacentes, para os dois distritos analisados, veri…cou-se que em Aveiro a distância mínima entre centróides é de 6 km e a máxima de 22 km. No distrito de Braga, a distância mínima entre centróides é de 9 km, enquanto que a máxima é também de 22 km.

Para estes dois distritos, e atendendo a um dos pressupostos em que se baseia o sistema de raciocínio qualitativo implementado, nomeadamente ao facto das entidades adjacentes estarem sempre associadas aos indicadores `mp` ou `p`, optou-se por estabelecer os intervalos de validade para a distância de: `mp` (0,7], `p` (7,22], `d` (22,51] e `md` (51,110], que equivalem a ampliação pelo `factor 7.3` dos intervalos correspondentes ao `ratio 2`.

A Tabela 6.4 sintetiza os resultados obtidos na primeira iteração do processo de inferência, utilizando o `ratio 2`, com os novos intervalos de validade para a direcção de…nidos na subsecção anterior. Pela análise da referida tabela constata-se que nos dois distritos, as relações topológicas comportam-se correctamente. No que diz respeito à distância, as melhorias foram bastante signi…cativas, existindo um número muito reduzido de casos em que a classi…cação veri…ca desvios. Chama-se, contudo, a atenção para o facto de, como pode ser veri…cado através da análise dos valores quantitativos reais, a maioria destes casos representarem situações em que a distância real coincide com o limite quantitativo 22, que divide o identi…cador qualitativo p do d. Esta situação pode ser constatada na Figura 6.2, na qual é possível visualizar um dos registos em questão e ainda, o valor quantitativo real. A tabela localizada à esquerda da referida …gura confronta os valores obtidos por inferência, `infDir`, `infDis` e `infTop`, com os obtidos substituindo os valores reais (na tabela à direita), `direcção`, `distância` e `topologia`, pelos respectivos identi…cadores qualitativos (para o caso da direcção e distância).

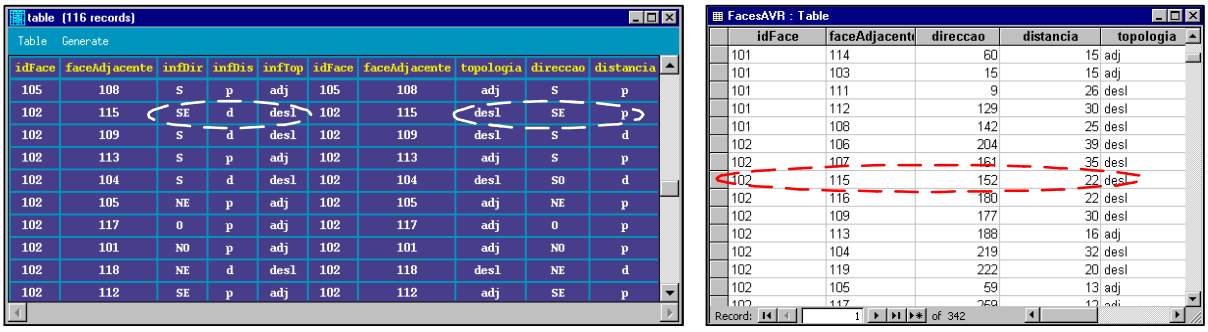


table (116 records)

Table Generate

idFace	faceAdjacente	infDir	infDis	infTop	idFace	faceAdjacente	topologia	direccao	distancia
105	108	S	p	adj	105	108	adj	S	p
102	115	SE	d	desl	102	115	desl	SE	p
102	109	S	d	desl	102	109	desl	S	d
102	113	S	p	adj	102	113	adj	S	p
102	104	S	d	desl	102	104	desl	S0	d
102	105	NE	p	adj	102	105	adj	NE	p
102	117	0	p	adj	102	117	adj	0	p
102	101	N0	p	adj	102	101	adj	N0	p
102	118	NE	d	desl	102	118	desl	NE	d
102	112	SE	p	adj	102	112	adj	SE	p

FacesAVR : Table

idFace	faceAdjacent	direccao	distancia	topologia
101	114	60	15	adj
101	103	15	15	adj
101	111	9	26	desl
101	112	129	30	desl
101	108	142	25	desl
102	106	204	39	desl
102	107	161	35	desl
102	115	152	22	desl
102	116	180	22	desl
102	109	177	30	desl
102	113	188	16	adj
102	104	219	32	desl
102	119	222	20	desl
102	105	59	13	adj

Record: 1 of 342

Figura 6.2: Erros provocados por incidência nos limites dos intervalos

Em relação à direcção, veri…ca-se que para o distrito de Aveiro, e antes da alteração aos limites dos intervalos quantitativos para esta relação, em 72 inferências, 20 estavam desfasadas. Após a alteração aos limites, 26 em 79 inferências estão desfasadas. Veri…ca-se aqui um

Aveiro

	N	NE	E	SE	S	SO	O	NO
N	**24**	4	0	0	0	0	0	0
NE	0	**12**	4	0	0	0	0	0
E	0	0	**16**	1	0	0	0	0
SE	0	0	0	**13**	4	0	0	0
S	0	0	0	0	**25**	4	0	0
SO	0	0	0	0	0	**12**	4	0
O	0	0	0	0	0	0	**16**	1
NO	4	0	0	0	0	0	0	**13**

	mp	p	d	md
mp	**4**	0	0	0
p	0	**74**	0	0
d	0	12	**67**	0
md	0	0	0	**0**

	adj	desl
adj	**78**	0
desl	0	**79**

Braga

	N	NE	E	SE	S	SO	O	NO
N	**4**	0	0	0	0	0	0	0
NE	0	**12**	2	0	0	0	0	0
E	0	0	**18**	2	0	0	0	0
SE	0	0	0	**11**	3	0	0	0
S	0	0	0	0	**4**	0	0	0
SO	0	0	0	0	0	**11**	2	0
O	0	0	0	0	0	0	**18**	2
NO	3	0	0	0	0	0	0	**11**

	mp	p	d	md
mp	**0**	0	0	0
p	0	**50**	0	0
d	0	4	**49**	0
md	0	0	0	**0**

	adj	desl
adj	**50**	0
desl	0	**53**

Tabela 6.4: Valores obtidos, na primeira iteração, para o `ratio 2`

pequeno aumento dos casos, que não é de forma alguma acompanhado pelo outro distrito analisado. Muito pelo contrário, no distrito de Braga, veri…ca-se que após a alteração dos limites, em 53 inferências, apenas 14 apresentaram desfasamentos, diminuindo em quase 50% os casos anteriormente veri…cados (23 em 49 inferências).

A diferença de comportamento, veri…cada entre estes dois distritos, chama a atenção para o facto da alteração dos limites dos intervalos, só por si, não resolver o problema da inferência na direcção vizinha. Este problema existirá sempre, uma vez que a abordagem utilizada para construir as regras de inferência, que integram a direcção e a distância, se baseia em pontos médios.

Não se pode, contudo, ignorar que as regiões em causa representam subdivisões administrativas, e como tal possuem limites com contornos irregulares. Esta situação faz com que se veri…quem casos em que uma dada região está posicionada em mais do que uma área de aceitação, mas a sua direcção é decidida pelo seu centróide. A Figura 6.3 apresenta alguns dos casos em que ocorreram desvios na classi…cação da direcção. O mapa apresentado permite constatar a diferença de dimensão existente entre as regiões e ainda, a di…culdade de determinação da direcção existente entre regiões, sem atender aos seus centróides, dados os contornos apresentados pelas mesmas. Além de todas estas di…culdades, neste distrito existem ainda regiões cujo centróide se encontra posicionado no limite das áreas de aceitação de…nidas pelo objecto de referência (Figura 6.4), o que obviamente di…culta ainda mais o processo de raciocínio qualitativo.

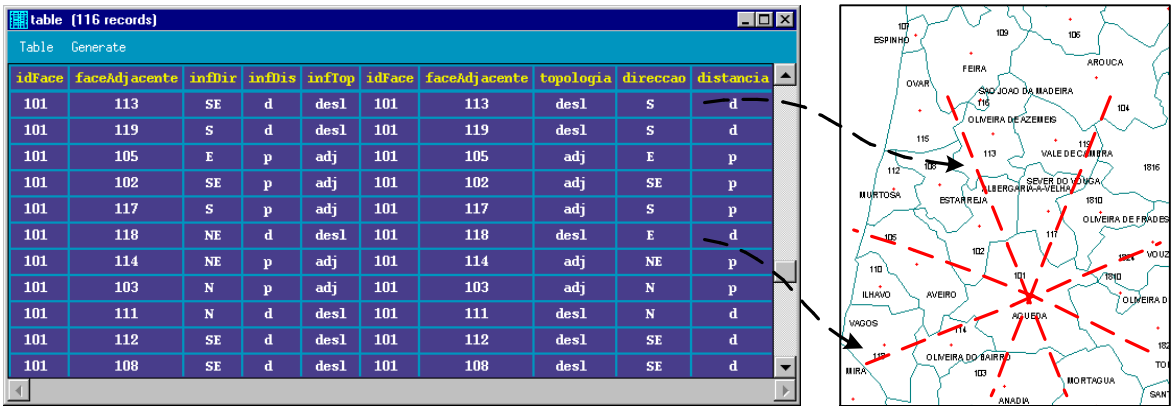


Figura 6.3: Desvios ocorridos no sistema de inferências

6.1.5 Incompatibilidades veri…cadas na integração da direcção e distância com a integração da direcção e topologia

Nesta fase importa veri…car porque é que em três casos, a direcção inferida pela integração da direcção e distância não coincide, ou não está no grupo das direcções inferidas pela integração da direcção e topologia.

A topologia in‡uencia as integrações respeitantes às distâncias `mp` e `p`. No caso da distância `p`, as regiões envolvidas podem estar adjacentes ou não. Para estas situações, veri…ca-se que o tamanho das regiões assume um papel de destaque na identi…cação da direcção existente entre as regiões envolvidas. Esta questão é avaliada nas próximas subsecções, nas quais são apresentados os casos em que foram veri…cadas incompatibilidades.

Composição `(N,p,desl);(NE,p,desl)`

Para o caso da composição `(N,p,desl);(NE,p,desl)`, a tabela que integra a direcção e distância (Apêndice D, secção D.2) indica que a direcção a inferir é `N`, enquanto que a tabela que integra a direcção e topologia (Apêndice B, subsecção B.1.5) aponta `NE` como a direcção a inferir. Estando o tamanho das regiões implícito na relação topológica existente entre entidades, e estando as regiões próximas, sem se tocarem, então, a dimensão das mesmas deve permitir que possam existir outras regiões entre elas. A Figura 6.5 apresenta vários cenários de localizações para estas regiões. Não é óbvia a identi…cação da direcção a inferir. Basta uma pequena alteração na distância que separa as regiões, para a direcção a inferir se alterar. Se a distância for semelhante, então o centróide de `A` está muito próximo do limite que separa a área de aceitação de `N` para `NE`. Se `C` se aproximar de `B`, então a direcção correcta a inferir é `N`. Se pelo contrário `A` é que está ligeiramente mais próximo de `B`, então a direcção correcta passa a ser `NE`.

Torna-se evidente que não é possível, com os factos conhecidos, proceder a qualquer alteração na regra de inferência que já estava de…nida para esta composição, uma vez que existirão sempre casos em que a direcção real é vizinha da direcção inferida pelo sistema.

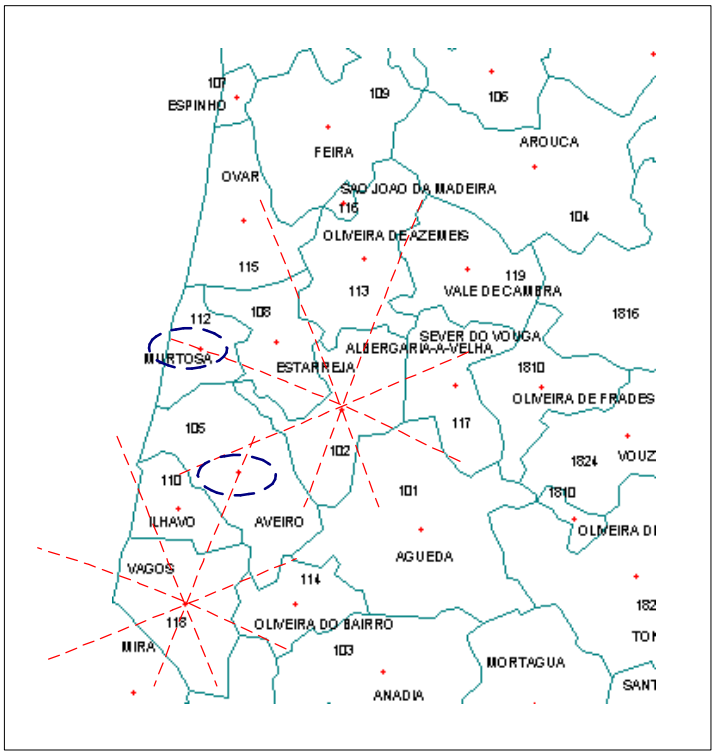


Figura 6.4: Centróides localizados nos limites das áreas de aceitação

Composição (N,p,adj);(NE,p,desl)

Para o caso da composição (N,p,adj);(NE,p,desl), a tabela que integra a direcção e distância (Apêndice D, secção D.2) indica que a direcção a inferir é N, enquanto que a tabela que integra a direcção e topologia (Apêndice B, subsecção B.1.5) aponta NE como a direcção a inferir. Neste caso, existem duas regiões que são adjacentes (A e B), sendo a distância qualitativa entre elas igual à existente entre B e C, apesar destas duas últimas não se tocarem. Pressupõe-se então que a dimensão de A e/ou B é superior à dimensão de C, ou que a distância quantitativa existente entre os centróides das regiões varia. A Figura 6.6 evidencia diferentes cenários de localizações para estas regiões. Mais uma vez, basta uma pequena alteração na distância existente entre as regiões, para variar a direcção a inferir.

Composição (N,p,adj);(SE,p,desl)

Para o caso da composição (N,p,adj);(SE,p,desl), a tabela que integra a direcção e distância (Apêndice D, secção D.2) indica que a direcção a inferir é NE, enquanto que a tabela que integra a direcção e topologia (Apêndice B, subsecção B.1.5) aponta E ou SE como direcções possíveis. Também neste caso, pressupõe-se que as regiões envolvidas na primeira relação apresentam maior dimensão. A Figura 6.7 apresenta vários cenários de localizações para estas regiões. Este é o único caso de incompatibilidade em que se poderia partir do princípio que, para a maior parte dos casos, a direcção a inferir seria E ao invés de NE.

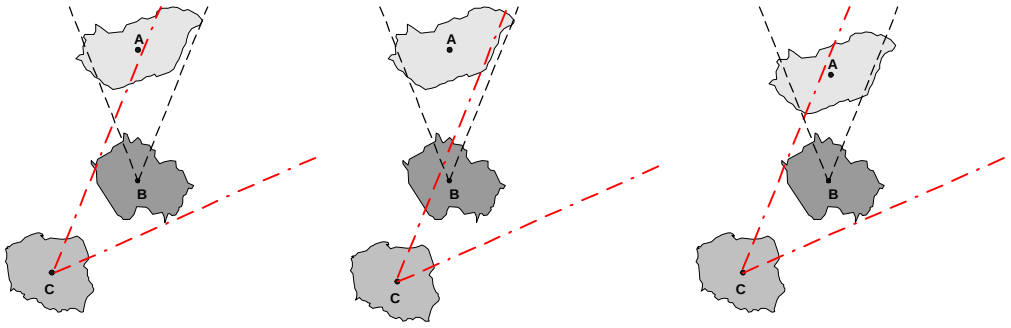


Figura 6.5: Composição (N,p,desl); (NE,p,desl): Di…culdades no estabelecimento da direcção

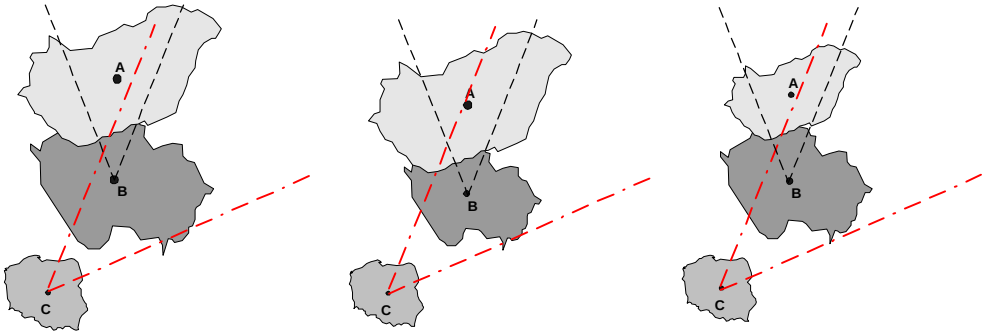


Figura 6.6: Composição (N,p,adj); (NE,p,desl): Di…culdades no estabelecimento da direcção

Resolução de incompatibilidades

A identi…cação das incompatibilidades existentes entre a integração da direcção e distância com a integração da direcção e topologia, permitiu constatar que para um dos três casos detectados, a indicação dada pela integração da direcção e topologia é a mais adequada. Contudo, e como já veri…cado anteriormente na rede…nição dos limites quantitativos dos intervalos para a direcção, um mesmo conjunto de regras não se comporta da mesma forma em diferentes contextos geográ…cos.

O que distingue o distrito de Braga do distrito de Aveiro?

À primeira vista, e além da diferença existente entre o tamanho das regiões que integram cada um dos distritos, constata-se que os contornos dos concelhos que integram o distrito de Aveiro são mais disformes, existindo diversos casos, em que uma dada região contorna outra em mais do que uma direcção. No distrito de Braga tal não acontece, uma vez que as regiões apresentam contornos mais regulares.

Veri…cando as características dos restantes distritos de Portugal, veri…cou-se que Aveiro é o distrito em que esta situação está provavelmente mais acentuada, não constituindo contudo um caso isolado.

Tal facto não permite tirar conclusões de…nitivas, em relação aos factores que diferenciam

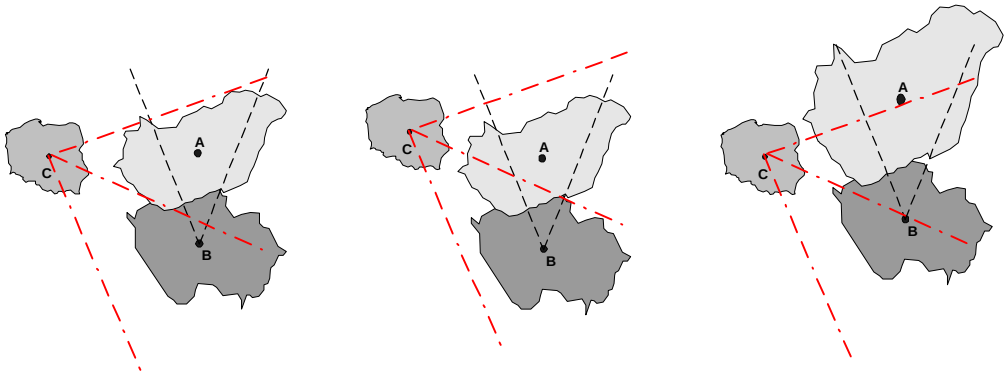


Figura 6.7: Composição (N,p,adj); (SE,p,desl): Di…culdades no estabelecimento da direcção

os dois distritos analisados. Relembra-se que a inferência é realizada com base nos centróides das regiões, apesar da adjacência poder ocorrer em mais do que uma área de aceitação.

Qual a solução encontrada?

Uma vez que estamos a lidar com uma abordagem qualitativa, da qual não se podem esperar resultados exactos, decidiu-se optimizar o processo de inferência através da inclusão no raciocínio, do tamanho das regiões. Nesta fase, esta integração é efectuada apenas para o caso das direcções ¢dir1 e ¢dir3, uma vez que estas representam os grupos de composições com problemas. Esta abordagem é apresentada na próxima subsecção.

6.1.6 Optimização do processo de inferência através da integração da dimensão das regiões

Como pode ser constatado na Figura 6.8, o tamanho das regiões assume um papel decisivo na determinação da direcção existente entre as mesmas. Se atendermos à dimensão das regiões, e para um determinado intervalo qualitativo, veri…ca-se que têm de existir variações em termos quantitativos das distâncias existentes entre as entidades. Em vez de adoptar o ponto médio de um dado intervalo na determinação das regras de inferência, optou-se por atribuir diferentes valores na identi…cação de uma dada regra, consoante o tamanho das regiões envolvidas. Por exemplo, se A>B e B>C e se A mp B e B mp C então, em termos quantitativos, a distância existente entre A e B é superior à distância existente entre B e C. Esta abordagem permite variar a distância quantitativa existente entre as regiões, na determinação das regras de inferência. Para o caso acima referido, em vez de se utilizar o ponto médio do intervalo, 0.5, utiliza-se 0.6 para a distância existente entre A e B, e 0.4 para a distância existente entre B e C. Nesta fase, não é relevante se 0.6 e 0.4 são as distâncias mais apropriadas, basta apenas que exista uma ligeira diferença entre as mesmas, para que o tamanho das regiões seja considerado na construção das regras. Posteriormente, e se esta abordagem se mostrar e…ciente, o sistema pode ser mais re…nado analisando se, por exemplo, uma região é muito maior do que a outra. Tal permitiria adoptar valores que destacassem esta diferença, como por exemplo 0.75 e 0.25, apesar de tal só servir para evidenciar a discrepância de tamanhos, uma vez que as regras de inferência permanecem inalteradas (quando comparadas com as obtidas com 0.6 e 0.4).

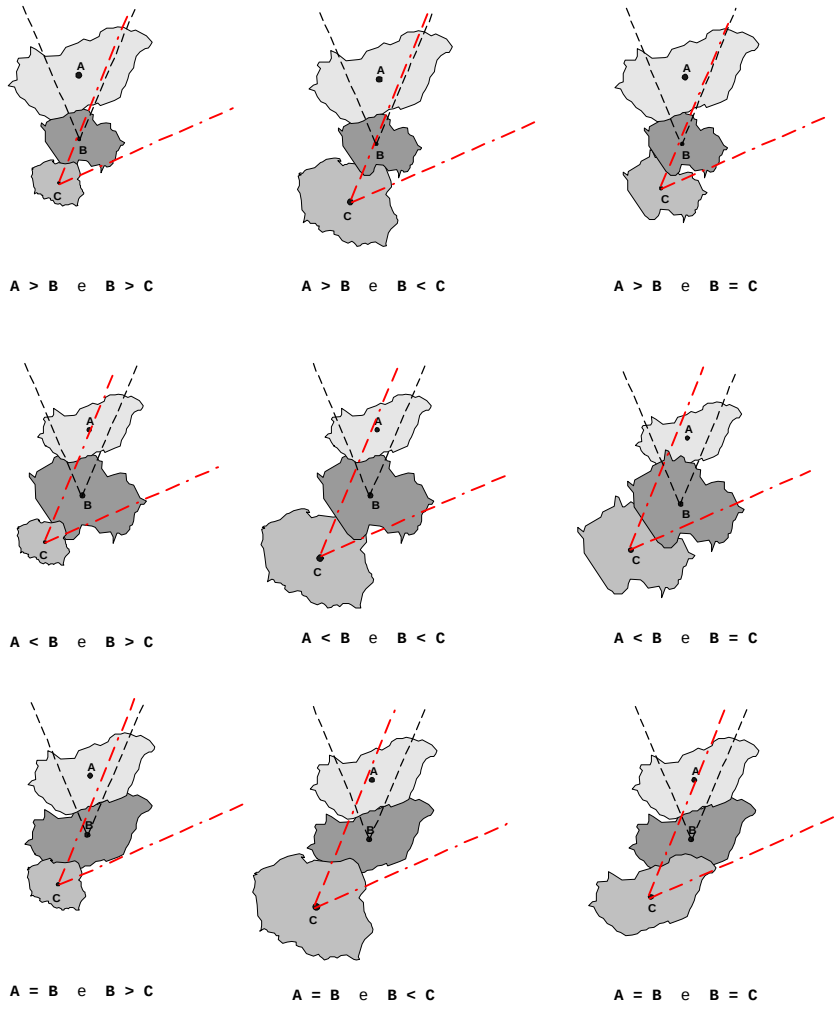


Figura 6.8: In‡uência do tamanho das regiões na determinação da direcção existente entre as mesmas

A determinação das regras de inferência, atendendo ao tamanho dos objectos, para os grupos ¢dir1 e ¢dir3 é apresentada no Apêndice D, secção D.5. Contudo, a Tabela 6.5 resume o processo de identi…cação das regras de inferência, que integram a direcção e distância, para o caso particular da composição (N,mp);(NE,mp). Como pode ser constatado pela análise da referida tabela, a integração do tamanho das regiões através da variação da distância quantitativa, faz com que as regras deixem de ser simétricas, isto é, o resultado da composição (N,mp);(NE,mp) não é igual ao resultado da composição (NE,mp);(N,mp). A região de maior dimensão, A ou C, in‡uencia determinantemente a direcção a inferir.

A aproximação apresentada permite que em apenas uma situação, o tamanho de A = C, a direcção inferida coincida com o limite quantitativo dos intervalos. Contudo, e como estamos a lidar com subdivisões administrativas, nas quais as dimensões das regiões são muito diversas, serão muito raros, ou mesmo inexistentes, os casos de inferência nesta situação.

(N,mp);(NE,mp)

A>B e B>C		
	V_{AB}	V_{BC}
distância	0,6	0,4
direcção	180	225
Ang_{AC}=	17,76428	
$\|V_{AC}\|$=	0,927044	
	N,mp	

A>B e B<C e A=C		
	V_{AB}	V_{BC}
distância	0,6	0,6
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	1,108655	
	N,mp	

A>B e B=C		
	V_{AB}	V_{BC}
distância	0,6	0,5
direcção	180	225
Ang_{AC}=	20,3435	
$\|V_{AC}\|$=	1,016988	
	N,mp	

A<B e B>C e A=C		
	V_{AB}	V_{BC}
distância	0,4	0,4
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,739104	
	N,mp	

A<B e B<C		
	V_{AB}	V_{BC}
distância	0,4	0,6
direcção	180	225
Ang_{AC}=	27,23572	
$\|V_{AC}\|$=	0,927044	
	NE,mp	

A<B e B=C		
	V_{AB}	V_{BC}
distância	0,4	0,5
direcção	180	225
Ang_{AC}=	25,13511	
$\|V_{AC}\|$=	0,832372	
	NE,mp	

A=B e B>C		
	V_{AB}	V_{BC}
distância	0,5	0,4
direcção	180	225
Ang_{AC}=	19,86489	
$\|V_{AC}\|$=	0,832372	
	N,mp	

A=B e B<C		
	V_{AB}	V_{BC}
distância	0,5	0,6
direcção	180	225
Ang_{AC}=	24,6565	
$\|V_{AC}\|$=	1,016988	
	NE,p	

A=B e B=C		
	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	N,mp	

(NE,mp);(N,mp)

A>B e B>C		
	V_{AB}	V_{BC}
distância	0,6	0,4
direcção	225	180
Ang_{AC}=	27,23572	
$\|V_{AC}\|$=	0,927044	
	NE,mp	

A>B e B<C e A=C		
	V_{AB}	V_{BC}
distância	0,6	0,6
direcção	225	180
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	1,108655	
	N,mp	

A>B e B=C		
	V_{AB}	V_{BC}
distância	0,6	0,5
direcção	228	180
Ang_{AC}=	26,3178	
$\|V_{AC}\|$=	1,005723	
	NE,mp	

A<B e B>C e A=C		
	V_{AB}	V_{BC}
distância	0,4	0,4
direcção	225	180
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,739104	
	N,mp	

A<B e B<C		
	V_{AB}	V_{BC}
distância	0,4	0,6
direcção	225	180
Ang_{AC}=	17,76428	
$\|V_{AC}\|$=	0,927044	
	N,mp	

A<B e B=C		
	V_{AB}	V_{BC}
distância	0,4	0,5
direcção	225	180
Ang_{AC}=	19,86489	
$\|V_{AC}\|$=	0,832372	
	N,mp	

A=B e B>C		
	V_{AB}	V_{BC}
distância	0,5	0,4
direcção	225	180
Ang_{AC}=	25,13511	
$\|V_{AC}\|$=	0,832372	
	NE,mp	

A=B e B<C		
	V_{AB}	V_{BC}
distância	0,5	0,6
direcção	225	180
Ang_{AC}=	20,3435	
$\|V_{AC}\|$=	1,016988	
	N,p	

A=B e B=C		
	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	225	180
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	N,mp	

Tabela 6.5: Cálculos quantitativos para a composição (N,mp);(NE,mp) e (NE,mp);(N,mp)

Aveiro

	N	NE	E	SE	S	SO	O	NO
N	**28**	2	0	0	0	0	0	2
NE	2	**19**	2	0	0	0	0	0
E	0	0	**17**	2	0	0	0	0
SE	0	0	1	**16**	1	0	0	0
S	0	0	0	2	**27**	2	0	0
SO	0	0	0	0	2	**19**	2	0
O	0	0	0	0	0	0	**17**	2
NO	1	0	0	0	0	0	1	**16**

	mp	p	d	md
mp	**4**	0	0	0
p	0	**114**	9	0
d	0	6	**50**	0
md	0	0	0	**0**

	adj	desl
adj	**78**	0
desl	0	**105**

Braga

	N	NE	E	SE	S	SO	O	NO
N	**7**	0	0	0	0	0	0	0
NE	0	**12**	0	0	0	0	0	0
E	0	0	**16**	0	0	0	0	0
SE	0	0	1	**12**	1	0	0	0
S	0	0	0	0	**7**	0	0	0
SO	0	0	0	0	0	**11**	0	0
O	0	0	0	0	0	0	**16**	0
NO	1	0	0	0	0	0	1	**12**

	mp	p	d	md
mp	**0**	0	0	0
p	0	**54**	5	0
d	0	0	**38**	0
md	0	0	0	**0**

	adj	desl
adj	**50**	0
desl	0	**47**

Tabela 6.6: Valores obtidos na primeira iteração, para o `ratio` 2, após inclusão da dimensão das regiões

Após a inclusão no processo de raciocínio das regras (Apêndice D, subsecção D.5.3) que permitem considerar o tamanho das regiões, voltou-se a medir o desempenho do sistema de inferências, avaliando os desvios ocorridos durante a classi…cação. Veri…cando mais uma vez os distritos de Aveiro e Braga, a Tabela 6.6 resume os resultados obtidos após a primeira iteração do processo.

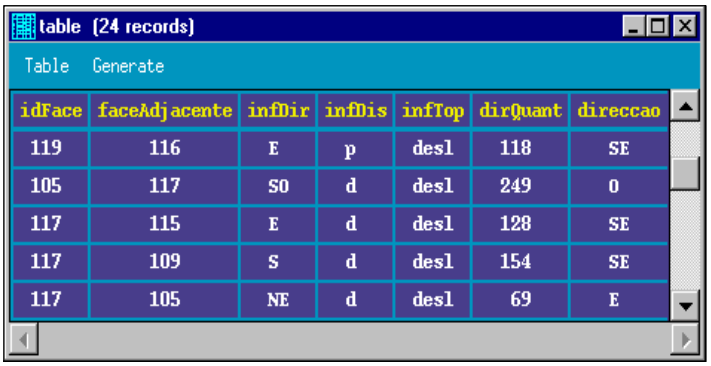
Esta primeira iteração permitiu inferir 105[6] relações espaciais para o distrito de Aveiro, das quais 24 apresentam desfasamentos. Para o distrito de Braga, em 47 inferências[7], 4 estão deslocadas. Para ambos os distritos, constata-se que existiu uma diminuição dos casos veri…cados anteriormente, continuando Aveiro com 23% de desfasamentos, para 9% no distrito de Braga.

A Figura 6.9, na qual é apresentado um subconjunto de registos dos 24 detectados para Aveiro, permite veri…car que a análise dos desvios através de comparação com os valores quantitativos reais não deve ser levada ao extremo, catalogando as inferências deslocadas como erradas, uma vez que os valores quantitativos reais (`dirQuant`) estão em alguns casos muito próximos dos limites quantitativos de…nidos para os intervalos. Esta situação certi…ca que as regiões em causa têm partes do seu território em ambas as direcções.

No …nal do processo de raciocínio (Tabela 6.7), foram veri…cadas melhorias nas inferências

[6] Após a integração das regras que incluem o tamanho dos objectos geográ…cos no processo de raciocínio, veri…ca-se um aumento da coincidência de resultados (isto é, relações inferidas por composição de factos diferentes), incrementando o número de relações inferidas em cada iteração.

[7] Aqui veri…ca-se uma diminuição do número de inferências, mas um aumento signi…cativo da percentagem de acerto.



idFace	faceAdjacente	infDir	infDis	infTop	dirQuant	direccao
119	116	E	p	desl	118	SE
105	117	SO	d	desl	249	O
117	115	E	d	desl	128	SE
117	109	S	d	desl	154	SE
117	105	NE	d	desl	69	E

Figura 6.9: Comparação da direcção inferida com a direcção real

de ambos os distritos, sendo estas mais evidentes no distrito de Braga.

Os resultados obtidos permitem a utilização das regras de inferência que integram a direcção, a distância e a topologia, considerando ainda a dimensão das regiões, no processo de descoberta de conhecimento, uma vez que as relações espaciais obtidas através de mecanismos de raciocínio qualitativo são, na pior das hipóteses, em 75% dos casos precisas. Os restantes 25% indicam a relação vizinha, permitindo uma aproximação válida à realidade, e ainda, su…ciente para os objectivos que este sistema de inferências visa servir, a inclusão da componente espacial associada aos dados geográ…cos no processo de descoberta de conhecimento.

Para comprovar esta a…rmação com um critério objectivo, isto é, con…rmar que não deverá existir outro distrito com piores resultados do que os apresentados para o distrito de Aveiro, veri…caram-se as discrepâncias existentes na dimensão das regiões que integram cada um dos distritos de Portugal continental. Após a análise das mesmas (apresentadas no Apêndice D, subsecção D.5.4), veri…cou-se que Santarém é o distrito que apresenta o menor ratio, existente entre a dimensão mínima e a dimensão máxima dos concelhos que o integram. O ratio entre estes dois valores é de 0,01, evidenciando a enorme diferença que existe entre o tamanho das regiões que representam estes extremos. Refere-se que o concelho com menor dimensão apresenta uma área de 13,52 km^2, contrastando com os 1120,47 km^2 do município de maior dimensão.

No distrito de Santarém, a distância quantitativa mínima e máxima existente entre regiões adjacentes é de 7 km e 37 km respectivamente, permitindo a de…nição dos intervalos de validade para a distância de `mp` (0,9], `p` (9,37], `d` (37,120] e `md` (120,370], os quais são obtidos por ampliação, através do `factor` 9.25, dos intervalos associados ao `ratio` 3. A utilização deste `ratio` é motivada pelo facto do mesmo permitir de…nir intervalos de validade que limitam os identi…cadores qualitativos `mp` e `p`, a relações topológicas do tipo `adjacente`.

O processo de inferência foi realizado para este distrito, permitindo identi…car as relações espaciais existentes entre as regiões que o integram. A partir das 90 relações espaciais explícitas na BDG, foram geradas 330 novas relações. Os resultados obtidos são apresentados na Tabela 6.8, na qual pode ser constatado que no …nal do processo, o desvio máximo veri…cado em relação à direcção é de 25%. Em relação à distância, veri…ca-se um incremento do número de casos deslocados, mas que obviamente deriva da grande discrepância existente entre a dimensão das regiões, di…cultando o processo de raciocínio e, ainda, a tarefa de de…nição dos intervalos de

Aveiro

	N	NE	E	SE	S	SO	O	NO
N	**67**	6	0	0	0	0	0	8
NE	7	**37**	3	0	0	0	0	0
E	0	0	**19**	4	0	0	0	0
SE	0	0	1	**18**	2	0	0	0
S	0	0	0	8	**64**	7	0	0
SO	0	0	0	0	9	**36**	4	0
O	0	0	0	0	0	0	**18**	3
NO	1	0	0	0	0	0	1	**19**

	mp	p	d	md
mp	**4**	0	0	0
p	0	**114**	18	0
d	0	6	**153**	15
md	0	0	5	**27**

	adj	desl
adj	**78**	0
desl	0	**264**

Braga

	N	NE	E	SE	S	SO	O	NO
N	**8**	0	0	0	0	0	0	0
NE	0	**15**	0	0	0	0	0	0
E	0	1	**32**	1	0	0	0	0
SE	0	0	3	**16**	2	0	0	0
S	0	0	0	0	**8**	0	0	0
SO	0	0	0	0	0	**16**	0	0
O	0	0	0	0	0	0	**32**	1
NO	2	0	0	0	0	0	3	**16**

	mp	p	d	md
mp	**0**	0	0	0
p	0	**56**	9	0
d	0	0	**73**	3
md	0	0	8	**7**

	adj	desl
adj	**50**	0
desl	0	**106**

Tabela 6.7: Valores obtidos no …nal do processo de inferência, para o `ratio` 2, após inclusão da dimensão das regiões

validade quantitativos para os indicadores de distância. Para o caso da distância, estes resultados podem ser optimizados se na restruturação da integração da relação topológica (sugerida na próxima subsecção), se permitir que o indicador qualitativo d possa estar associado a entidades adjacentes. Tal permitirá de…nir intervalos mais pequenos, e como se constatou anteriormente, mais apropriados para a região analisada. Contudo, e em termos de localização, quando as direcções inferidas não são as exactas, indicam a direcção vizinha, na qual as entidades não têm localizado o seu centróide, mas sim partes da sua região.

6.1.7 Avaliação do desempenho na inferência de relações topológicas

Apesar das relações topológicas existentes entre regiões adjacentes se encontrarem explícitas na BDG, poderão existir domínios de aplicação em que nem todas as relações sejam conhecidas. Nestes casos, o sistema de inferências deverá ser utilizado para conhecer tais relações. Esta situação não é aqui avaliada, por se considerar que tal só deverá acontecer depois da optimização do sistema de inferências, uma vez que são diversos os casos em que a relação topológica não é univocamente inferida.

Esta optimização, que pode ser efectuada recorrendo mais uma vez ao tamanho dos objectos, permitirá clari…car muitas das situações em dúvida. Esta reconstrução das regras poderá seguir o procedimento adoptado neste trabalho, e que transforma o conjunto integrado (`direcção, topologia`) em intervalos temporais. As primitivas temporais deverão ser rede…nidas, por forma a caracterizarem mais convenientemente a dimensão das entidades geográ…cas. A

Santarém

	N	NE	E	SE	S	SO	O	NO
N	**42**	5	0	0	0	0	0	1
NE	10	**73**	11	0	0	0	0	0
E	0	3	**33**	4	0	0	0	0
SE	0	0	3	**21**	3	0	0	0
S	0	0	0	1	**45**	7	0	0
SO	0	0	0	0	10	**69**	12	0
O	0	0	0	0	0	3	**32**	3
NO	4	0	0	0	0	0	3	**22**

	mp	p	d	md
mp	**8**	0	0	0
p	0	**151**	21	0
d	0	53	**172**	0
md	0	0	15	**0**

	adj	desl
adj	**90**	0
desl	0	**330**

Tabela 6.8: Valores obtidos no …nal do processo de inferência, para o distrito de Santarém, utilizando o `ratio 3`

Figura 6.10 apresenta algumas destas situações, sugerindo as primitivas temporais que poderiam ser utilizadas.

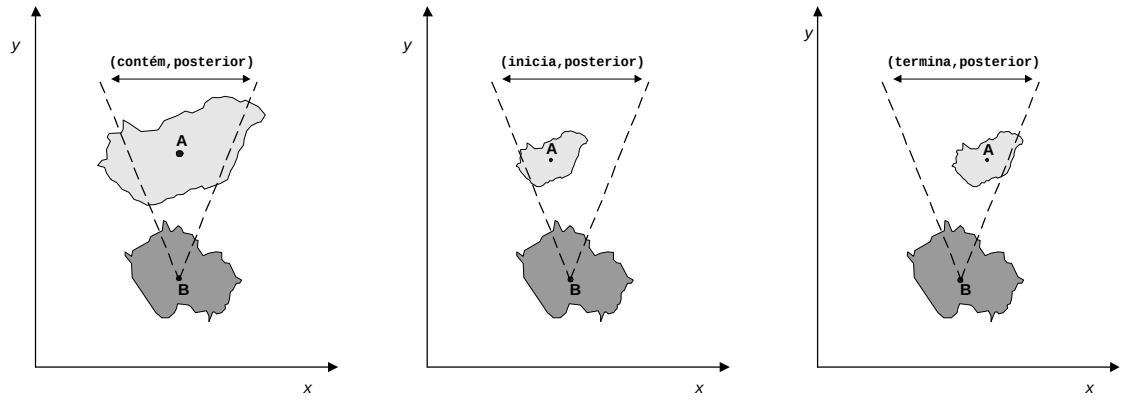


Figura 6.10: Intervalos temporais que caracterizam a integração da direcção e topologia, considerando a dimensão das regiões

A de…nição de novos intervalos temporais, nos quais seja integrada a dimensão dos objectos, permite ainda a adopção de um cone com 16 direcções, cuja utilização evitaria muitas das situações de inferência na relação vizinha. O uso de um cone com 16 áreas de aceitação permite a de…nição de intervalos de validade de menor dimensão, e como tal, a de…nição de relações de direcção mais especí…cas (Figura 6.11).

Após a rede…nição das regras de inferência, o processo de avaliação do seu desempenho poderia passar por remover aleatoriamente algumas das relações explícitas na BDG, para entidades adjacentes. Posteriormente, veri…car-se-ia a validade das relações inferidas, por comparação com as relações reais.

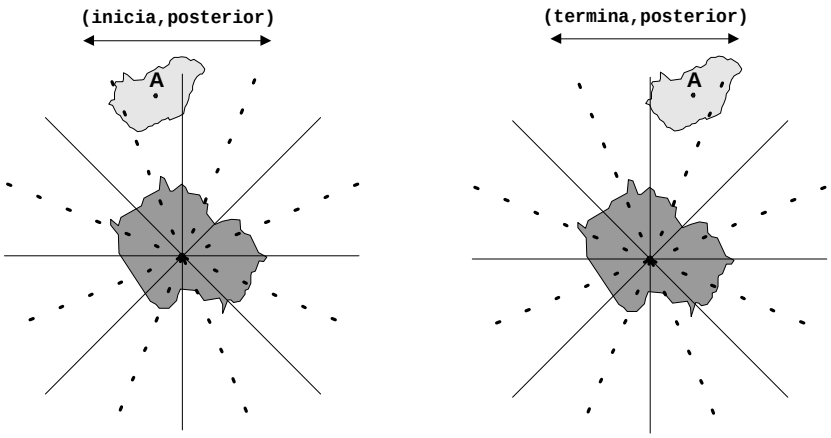


Figura 6.11: Cone de 16 direcções na de…nição das primitivas temporais

6.2 Avaliação do processo de descoberta de conhecimento

Nesta secção pretende-se avaliar, independentemente da componente geográ…ca dos dados, se o Padrão identi…ca, ou não, regras que realmente estão implícitas nas BD analisadas. Neste processo, é explorado um conjunto de dados de avaliação. Estes conjuntos de dados são construídos a partir de um grupo especí…co de regras, passando as mesmas a estar implícitas nos dados.

Dos inúmeros conjuntos de dados construídos para auxiliar o desenvolvimento de novos algoritmos ou testar novas técnicas de DM (http://www.kdnuggets.com/datasets), para esta avaliação seleccionou-se um dos utilizados nas aulas da Opção III - Tecnologias e Sistemas de Informação, leccionada no 5ºano da Licenciatura em Informática de Gestão da Universidade do Minho. Os dados foram preparados[8] com o objectivo de auxiliar o processo de assimilação dos conceitos associados à DCBD e ainda, das diversas técnicas e algoritmos disponíveis no Clementine.

O conjunto de dados seleccionado agrupa 3.031 registos que caracterizam os clientes de uma empresa de …nanciamento, que fornece crédito para a aquisição de bens. Para estes dados, foram de…nidos os seguintes objectivos:

Objectivo do negócio: minimizar o risco de incumprimento que advém do …nanciamento concedido aos clientes.

Objectivo do DM: conseguir determinar o per…l dos clientes, por forma a minimizar o risco de investimento da empresa.

6.2.1 Compreensão dos dados

Antes de prosseguir com as diversas fases do processo de descoberta de conhecimento, é necessário analisar os dados a explorar, por forma a compreender o signi…cado de cada um dos atributos, e de…nir estratégias de análise para os mesmos.

[8] Pela empresa NTech (http://www.ntech.pt), responsável pela leccionação da referida Opção III.

Descrição dos dados

Os atributos que integram os dados a analisar são: identi…cação, número …scal, número dealer, estatuto, nome, bem …nanciado, tipo de contrato, duração, rendimento bruto, valor do crédito, tipo de pagamento, credito à habitação, valor da prestação, estado civil, número de …lhos, idade e incumprimento.

Globalmente, refere-se que além da identi…cação dos clientes, à qual é associado o número de …lhos, é referido o bem …nanciado, o tipo de pagamento seleccionado pelo cliente, o valor da prestação e ainda, se o cliente possui um outro …nanciamento para a habitação. O atributo incumprimento é utilizado para assinalar os clientes que veri…caram anomalias no pagamento das respectivas prestações.

Exploração dos dados

Nesta etapa pretende-se detectar anomalias nos dados, veri…cando o conjunto de valores que cada atributo armazena e ainda, a sua distribuição. A exploração dos dados foi realizada no Clementine, recorrendo aos nodos Distribution e Histrogram da palette Graphs, através da stream apresentada na Figura 6.12. Nesta …gura é ainda possível veri…car a qualidade dos dados que, excluindo três atributos com valor informativo, Nome, N_Fiscal, Estatuto, se encontram completamente preenchidos.

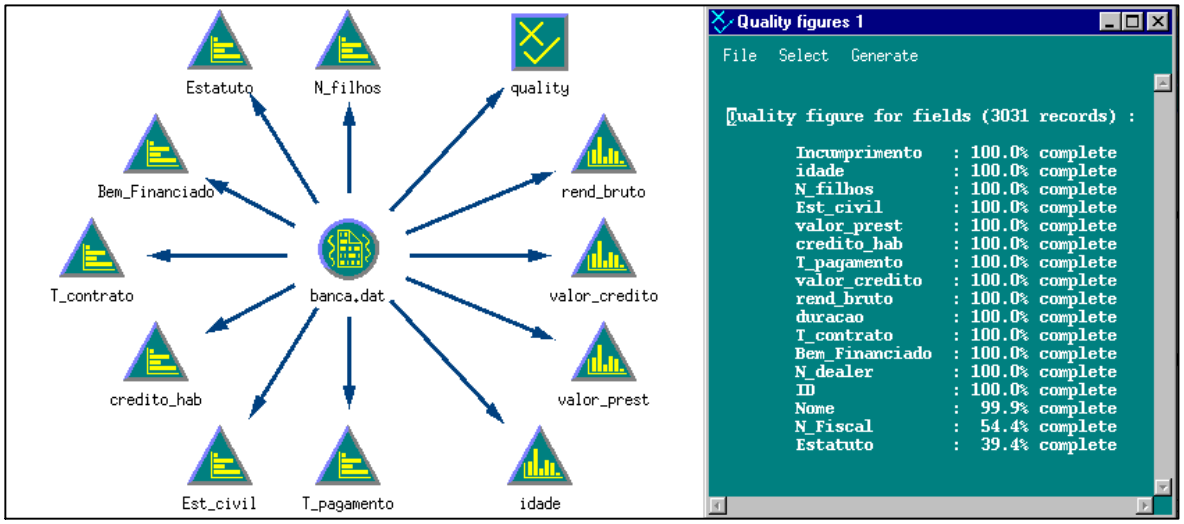


Figura 6.12: Exemplo Financiamento: Exploração dos dados

Os resultados obtidos em cada um dos nodos Distribution, utilizados para veri…car a distribuição de atributos com valores categóricos, são sintetizados na Figura 6.13. Pela análise da referida …gura constata-se que:

- No atributo que indica se o cliente possui ou não crédito à habitação (credito_hab), com os valores 1 ou 0 respectivamente, existe um registo com o valor 2, o qual deverá ser removido uma vez que representa um erro nos dados;

[2] No atributo Estatuto existem cinco casos de …nanciamento concedido a empresas, os quais não podem ser analisados em conjunto com os restantes casos de …nanciamento concedido a particulares. Para além do conjunto de regras que dita a concessão de …nanciamento a estes dois tipos de clientes ser diferente, o reduzido número de casos disponíveis para o cliente empresa também não permite que os mesmos sejam considerados na análise, e como tal têm de ser removidos da amostra.

Para os restantes atributos não foram detectadas quaisquer anomalias, apresentado os mesmos uma distribuição que resulta do normal funcionamento da empresa.

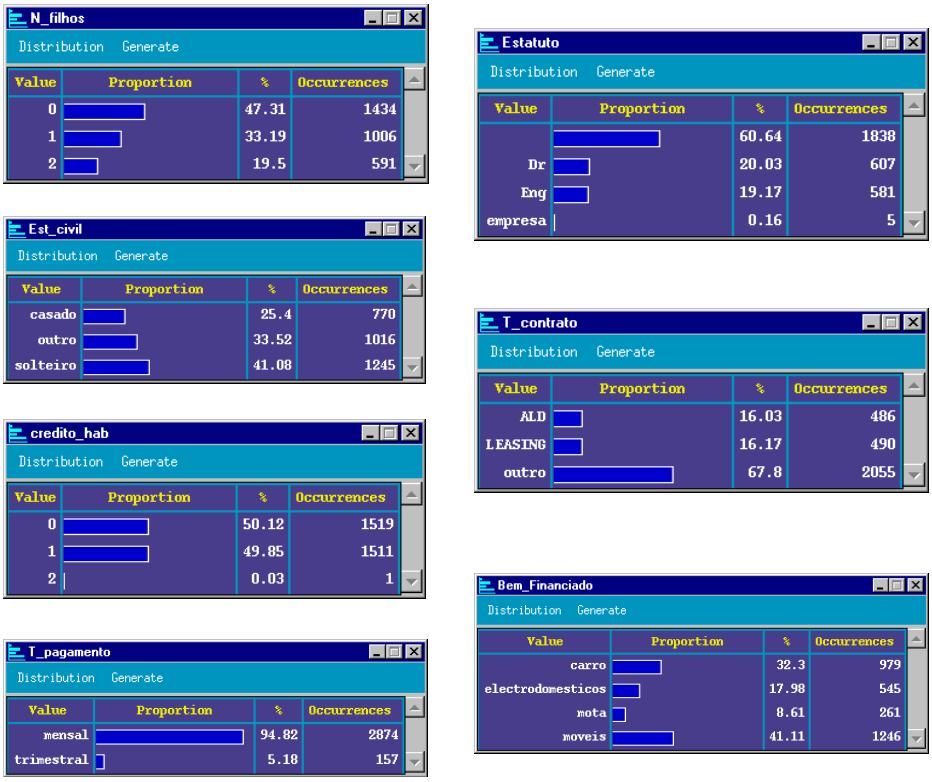


Figura 6.13: Exemplo Financiamento: Distribuição dos dados categóricos

No caso dos atributos com valores contínuos, a Figura 6.14 apresenta os histogramas que permitem analisar a distribuição dos mesmos, e de…nir as classes a utilizar na transformação dos atributos com valores contínuos, em atributos com valores discretos. A análise dos histogramas apresentados permitiu adoptar as classes apresentadas na Tabela 6.9.

O exercício de compreensão e exploração dos dados conduziu à identi…cação dos atributos a analisar e, à de…nição das classes a utilizar na etapa de pré-processamento dos dados. As próximas subsecções apresentam as diversas fases do processo de descoberta de conhecimento, que conduziram à detecção de padrões nos dados.

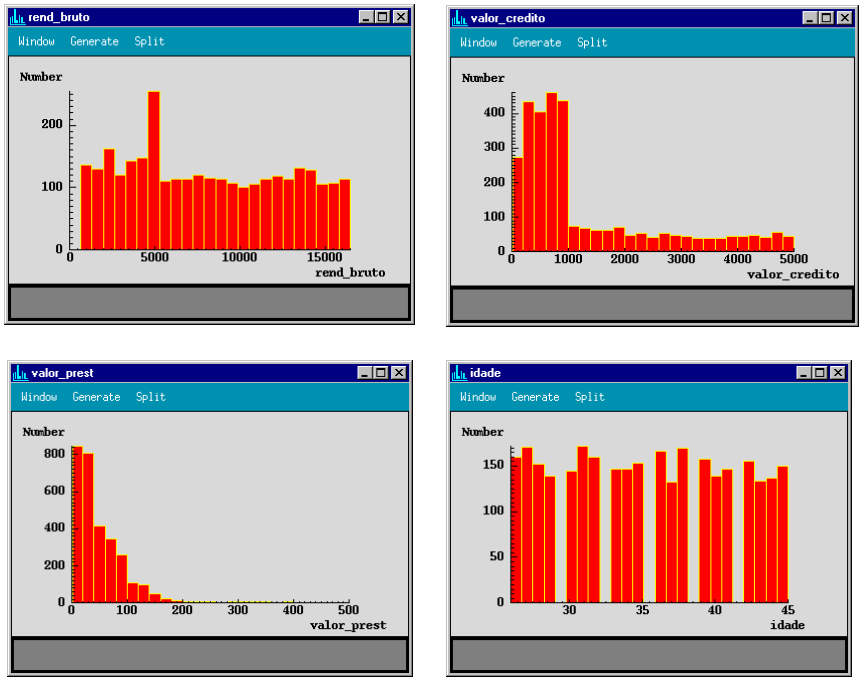


Figura 6.14: Exemplo Financiamento: Histogramas para os atributos com valores contínuos

6.2.2 Selecção e tratamento dos dados

A fase de selecção dos dados permite eliminar todos os atributos que não têm interesse no processo de descoberta de conhecimento. São estes a identi…cação, o número …scal, o número dealer, o estatuto e o nome. Os restantes atributos são seleccionados, com o objectivo de avaliar a sua contribuição na determinação do per…l de cliente.

A fase de tratamento dos dados consiste basicamente no tratamento de dados omissos e dados corrompidos. No exemplo em análise, apenas em dois casos foram detectadas anomalias, como já referido anteriormente, conduzindo à eliminação do registo com valor 2 no atributo crédito à habitação (credito_hab), e à remoção do valor empresa no atributo estatuto (Estatuto).

Atributos	Classes
Idade	(25..30] → **26-30**, (30..40] → **31-40**, (40..45] → **41-45**
Rendimento bruto	(0..2500] → **0-2500**, (2500..5000] → **2501-5000**, (5000..10000] → **5001-10000**, (10000..17000] → **10001-17000**
Valor crédito	(0..100] → **0-100**, (100..500] → **101-500**, (500..1000] → **501-1000**, (1000..2500] → **1001-2500**, (2500..5000] → **2501-5000**
Valor prestação	(0..10] → **0-10**, (10..20] → **11-20**, (20..50] → **21-50**, (50..100] → **51-100**, (100..500] → **101-500**

Tabela 6.9: Exemplo Financiamento: Classes para os atributos com valores contínuos

A Figura 6.15 apresenta a stream construída para as fases de selecção e tratamento dos dados, atendendo às tarefas acima especi…cadas. Como resultado, é criado o …cheiro DadosBanca, com os dados a utilizar nas próximas fases do processo de descoberta de conhecimento.

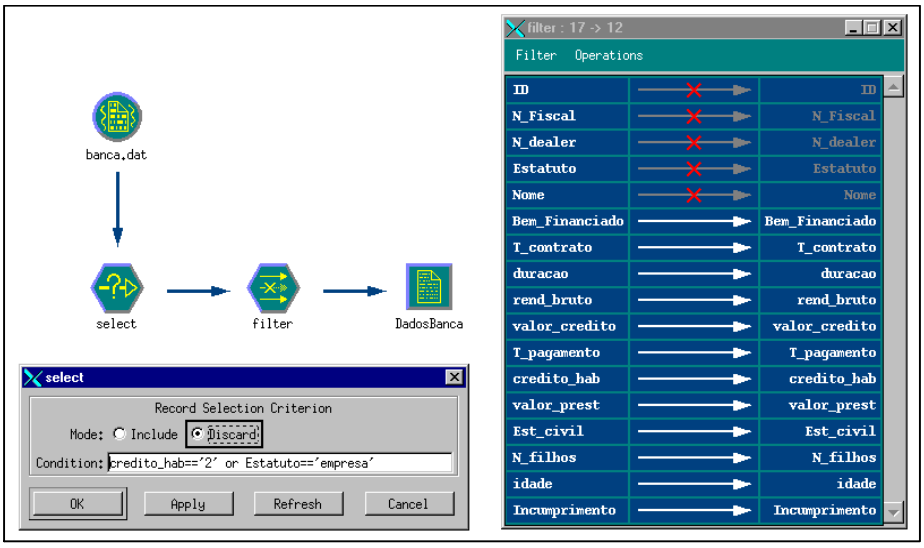


Figura 6.15: Exemplo Financiamento: Selecção e tratamento dos dados

6.2.3 Pré-processamento dos dados

Na fase de pré-processamento dos dados (Figura 6.16), os atributos com valores contínuos são transformados em atributos com valores discretos, atendendo às classes de…nidas na Tabela 6.9. Nesta fase são, ainda, utilizados nodos Web na exploração dos dados. Esta exploração permite identi…car associações entre os atributos, que indiciam a relevância dos mesmos na identi…cação do per…l dos clientes. A última tarefa, efectuada nesta etapa, consiste na divisão aleatória dos dados em dois …cheiros, Treino e Teste, que serão utilizados na construção dos modelos que caracterizam os dados e na sua validação, respectivamente.

Os nodos Web gerados (Figura 6.17) apenas permitem constatar que existe uma associação fraca entre o estado civil solteiro e o incumprimento (atributo incumprimento com valor 1) (Figura 6.17 a)), e ainda, entre o rendimento bruto caracterizado pela classe 10001-17000 e o incumprimento (Figura 6.17 c)). Estes dois casos permitem concluir que nesta empresa de …nanciamento, os solteiros e as pessoas com maiores rendimentos são os mais cumpridores. No que diz respeito às associações existentes entre os bens …nanciados e o incumprimento, constata-se que o bem móveis não tem qualquer associação com o incumprimento (Figura 6.17 b)), salientando que não existe qualquer anomalia no …nanciamento deste bem. Em relação ao valor da prestação (Figura 6.17 d)) não é possível tirar qualquer conclusão, uma vez que tanto o cumprimento como o incumprimento, apresentam associações fortes com as diversas classes que caracterizam os valores das prestações.

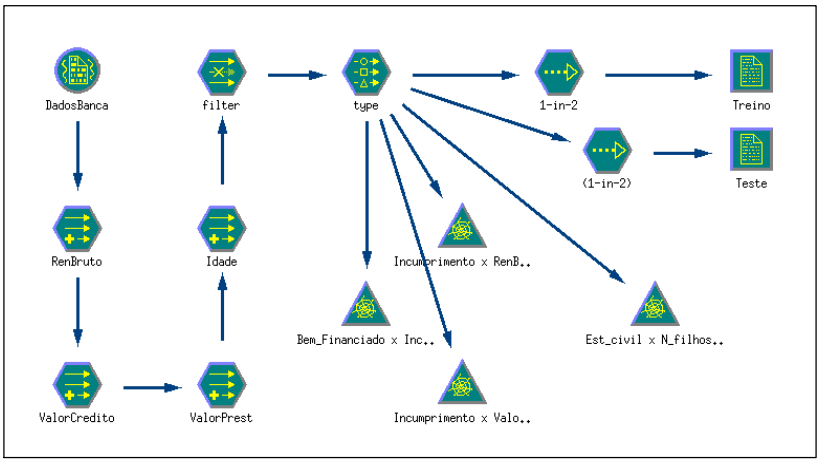
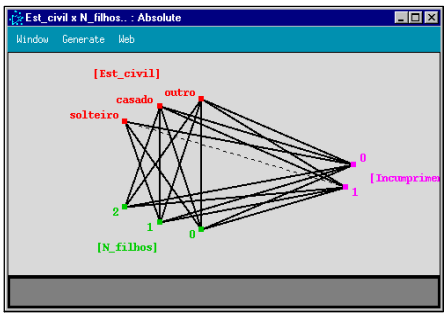
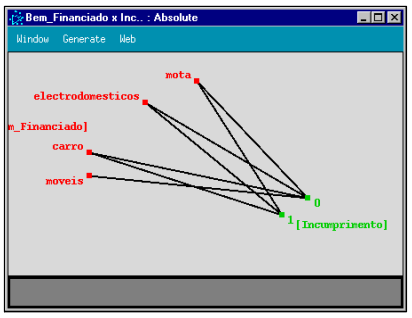


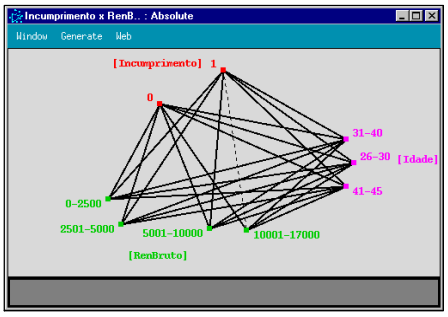
Figura 6.16: Exemplo Financiamento: Pré-processamento dos dados



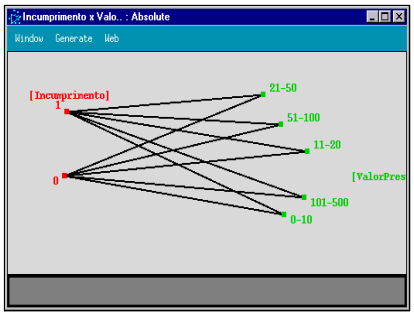
a) **Estado civil** x **Nº Filhos** x **Incumprimento**



b) **Bem Financiado** x **Incumprimento**



c) **Idade** x **Rendimento Bruto** x **Incumprimento**



d) **Valor Prestação** x **Incumprimento**

Figura 6.17: Exemplo Financiamento: Exploração dos dados com nodos Web

6.2.4 Data Mining

Na fase de DM (Figura 6.18), o …cheiro Treino é utilizado na construção de três modelos que caracterizam os dados. O primeiro, construído recorrendo ao algoritmo C5.0, tem como função determinar o conjunto de atributos relevante para a previsão do atributo incumprimento, e ainda, identi…car as regras que caracterizam o per…l dos clientes …nanciados, principalmente, dos incumpridores. O modelo obtido[9] (evidenciado na Figura 6.19 na forma de regras) identi…ca as oito regras que caracterizam o per…l dos clientes catalogados como incumpridores. Quatro destas regras apresentam um grau de con…ança superior a 97%, e restringem, nalguns casos, as regras mais genéricas que apresentam um grau de con…ança menor.

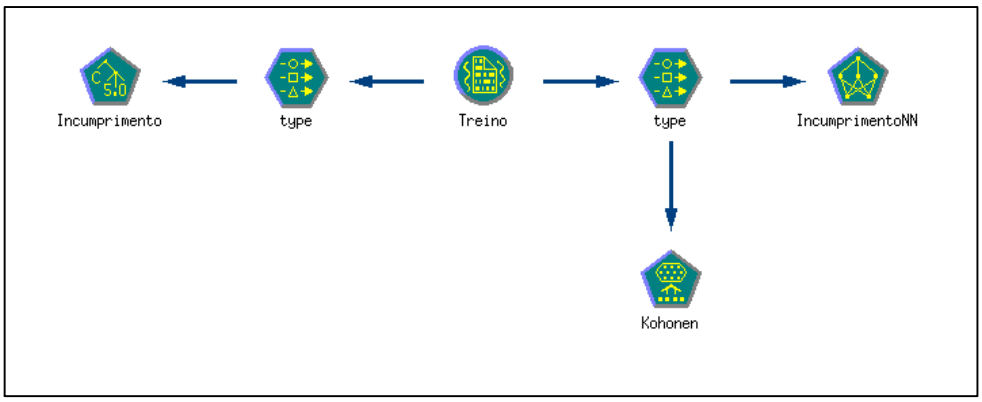


Figura 6.18: Exemplo Financiamento: Data Mining

A análise das regras obtidas com o algoritmo C5.0 permite identi…car o conjunto de atributos relevante para a caracterização do incumprimento. São eles Bem_financiado, T_contrato, RenBruto, Credito_hab, Est_civil, N_filhos e Idade. Estes atributos são utilizados no treino de uma rede neuronal (nodo IncumprimentoNN na stream da Figura 6.18), que complementa o modelo obtido com o C5.0, na previsão do atributo incumprimento.

O último modelo construído (nodo Kohonen na stream da Figura 6.18) utiliza as redes neuronais do tipo Kohonen, na identi…cação de segmentos de clientes. Os segmentos obtidos caracterizam grupos de clientes com per…l idêntico, para os quais é possível determinar, recorrendo ao algoritmo C5.0, as regras que classi…cam os clientes num ou noutro segmento. Este procedimento é necessário uma vez que os modelos construídos recorrendo ao algoritmo de Kohonen, apenas acrescentam aos registos analisados, as coordenadas (x, y) que os posicionam em determinado segmento. Os segmentos obtidos podem ser visualizados recorrendo a um nodo Plot (Figura 6.20), que permite a selecção dos mesmos e, ainda, a sua utilização na identi…cação das regras que os caracterizam. Na Figura 6.20 visualizam-se, claramente, três segmentos associados ao incumprimento. A título de exemplo, um dos segmentos foi seleccionado e utilizado pelo algoritmo C5.0, na determinação das regras que o caracterizam.

A Figura 6.21 apresenta a stream construída para a visualização, num nodo Plot, dos

[9] O modelo foi construído recorrendo a opção Boosting, que recorre a utilização de classi…cadores. Cada um destes classi…cadores é consultado na tarefa de previsão. Este procedimento permite obter modelos mais precisos, sendo necessário mais tempo para a construção dos mesmos.

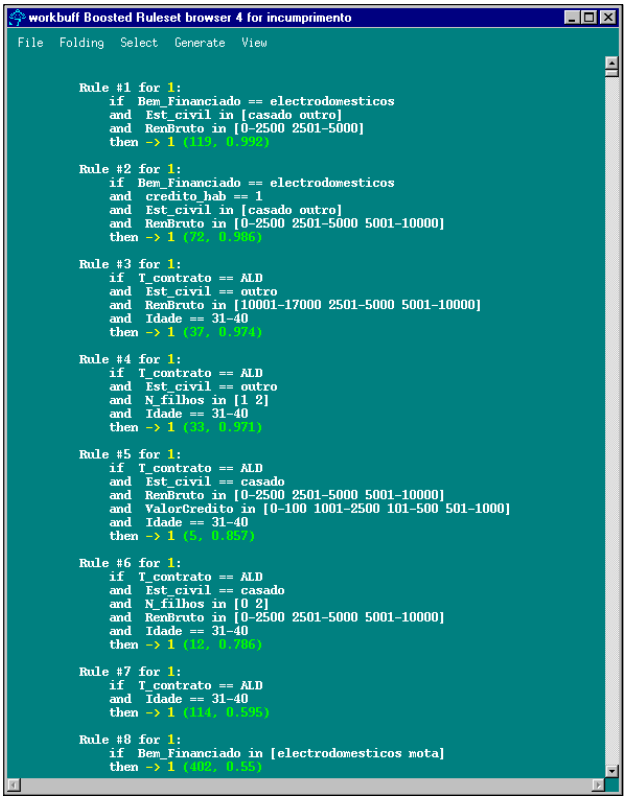


Figura 6.19: Exemplo Financiamento: Regras obtidas recorrendo ao algoritmo C5.0

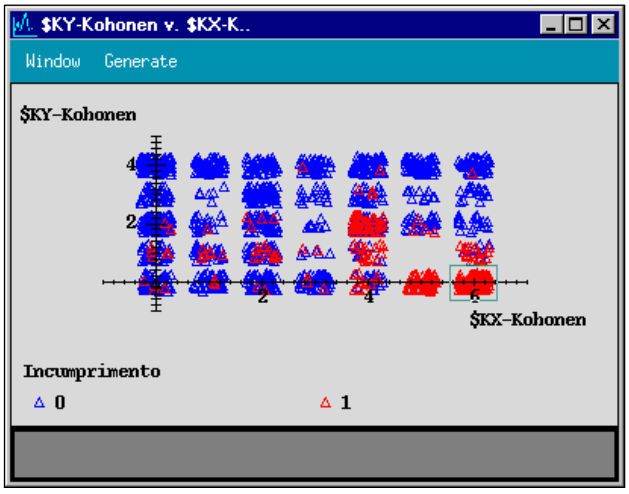


Figura 6.20: Exemplo Financiamento: Segmentos obtidos com a rede neuronal do tipo Kohonen

segmentos identi…cados pelo algoritmo Kohonen, e das regras que caracterizam o segmento seleccionado na Figura 6.20.

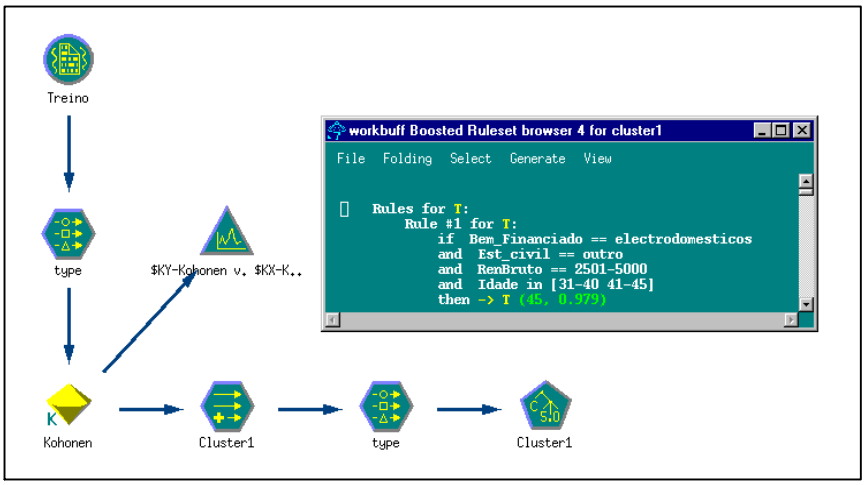


Figura 6.21: Exemplo Financiamento: Regras que caracterizam um dado segmento

6.2.5 Interpretação de resultados

Os modelos construídos na fase de DM, recorrendo ao algoritmo C5.0 e redes neuronais, são nesta fase utilizados na classi…cação de casos desconhecidos, com o objectivo de avaliar o seu desempenho na previsão do incumprimento dos clientes. O …cheiro Teste, gerado na fase de pré-processamento dos dados, é nesta etapa utilizado na veri…cação da con…ança das regras, encontradas na fase de DM.

A stream construída para a fase de interpretação de resultados é apresentada na Figura 6.22. Nesta …gura é possível visualizar o resumo do desempenho de cada um dos modelos e ainda, o desempenho dos dois modelos se utilizados integradamente. Neste último caso, o desempenho na previsão apresenta uma percentagem de acerto de 99.21%, contra os 98.51% evidenciado pelo modelo gerado pelo algoritmo C5.0 e os 98.64% obtidos com a rede neuronal.

A utilização conjunta dos dois modelos, na previsão do incumprimento dos clientes, permite obter resultados mais precisos. Tal deve-se ao facto de cada um destes modelos apresentar desempenhos diferentes, que dependem do bem …nanciado. Esta situação pode ser con…rmada na Figura 6.23, na qual é possível veri…car que, por exemplo, para o bem carro, a con…ança na utilização do modelo obtido com o algoritmo C5.0 é de 98.00%, enquanto que o modelo obtido com a rede neuronal apresenta uma con…ança de 97.60%. No caso do bem mota esta situação inverte-se, apresentando a árvore de decisão uma con…ança de 95.28% e a rede neuronal uma con…ança de 97.64%. O conhecimento do desempenho de cada um dos modelos, atendendo ao bem …nanciado, permite que na previsão do comportamento dos clientes estes sejam utilizados integradamente, aumentando a con…ança global dos mesmos.

Com a veri…cação da con…ança que os modelos obtidos atingem na classi…cação de dados desconhecidos, isto é, dados não utilizados no treino dos modelos, con…rma-se a utilidade dos

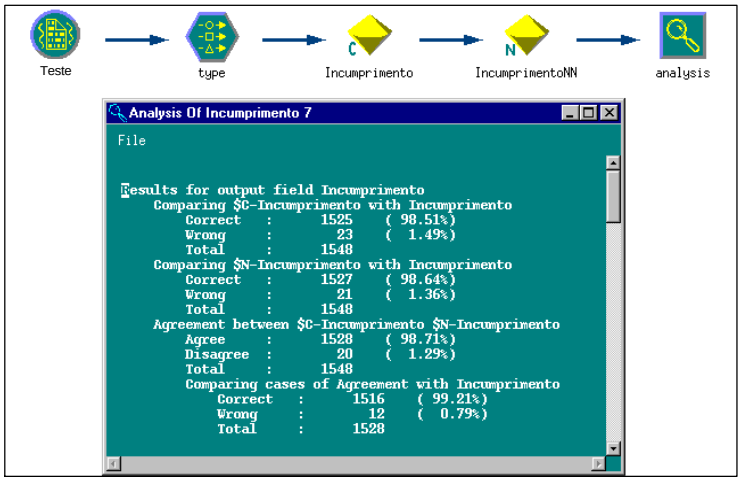


Figura 6.22: Exemplo Financiamento: Desempenho dos modelos construídos

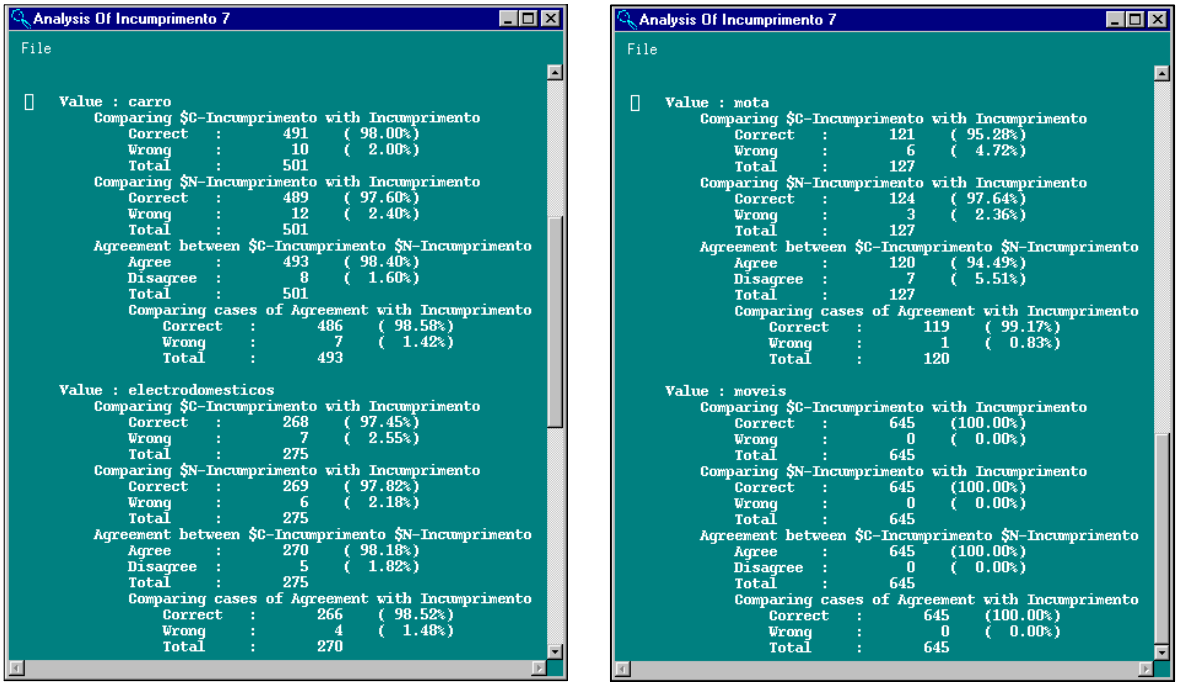


Figura 6.23: Exemplo Financiamento: Desempenho, por bem …nanciado, dos modelos construídos

mesmos na identi…cação do per…l dos clientes, determinando a sua predisposição para o incumprimento. Tal facto permite a utilização das regras apresentadas na Figura 6.19, na classi…cação dos clientes, nomeadamente no que diz respeito ao seu eventual cumprimento ou incumprimento do empréstimo solicitado.

Além deste exercício ter permitido evidenciar as diversas fases do processo de descoberta de conhecimento, recorrendo a um caso prático especí…co, o seu principal objectivo é o de veri…car se a ferramenta de descoberta de conhecimento adoptada para a implementação do sistema Padrão, o Clementine, efectivamente identi…ca relacionamentos implícitos nos dados.

Com a identi…cação do conjunto de regras a partir do qual estes dados foram gerados, con…rma-se a utilização do Clementine, como a ferramenta de descoberta de conhecimento utilizada pelo sistema Padrão.

6.2.6 A componente geo-espacial

Nas diversas fases do processo de descoberta de conhecimento, apresentadas nas subsecções anteriores, não foi incluída a fase de processamento da informação geo-espacial considerada pelo sistema Padrão. Tal deve-se, essencialmente, ao facto do …cheiro seleccionado para exploração não se encontrar geo-referenciado. Contudo, e como o objectivo deste exercício era averiguar a capacidade do Clementine de identi…car relacionamentos implícitos nos dados, a não realização desta fase não interfere com a validade dos resultados obtidos.

A capacidade de identi…cação de relacionamentos nos dados não depende do tipo dos dados, uma vez que a abordagem adoptada no Padrão, para a inclusão da componente espacial no processo de descoberta de conhecimento, passa por integrar novos atributos, neste caso os geo-espaciais, aos dados a analisar. Estes atributos explicitam a informação geo-espacial associada a cada registo, sendo o conjunto de dados resultante analisado pelos algoritmos de DM, sem existir qualquer distinção entre o tipo dos mesmos.

Através deste processo, os algoritmos de DM poderão ou não identi…car relacionamentos nos dados, estando esta identi…cação apenas dependente do conteúdo dos dados (atributos explorados) e não do seu tipo.

Capítulo 7

Validação da utilidade do sistema PADRÃO

Neste capítulo pretende-se validar a utilidade do sistema Padrão na análise de BD organizacionais, representando esta uma avaliação às potencialidades do sistema. Para verificar esta utilidade foi analisada uma BD organizacional de grande dimensão, nomeadamente um componente do Sistema de Informação de Administração do Pessoal do Exército (SIAPE). Este componente diz respeito ao Cadastro Geral, no qual o Comando de Pessoal do Exército armazena os dados pessoais dos indivíduos que ...zeram/fazem parte do quadro permanente do Exército e ainda, os dados pessoais dos mancebos, que até à data de utilização neste trabalho, foram registados nesta BD.

A análise desta BD constitui o estudo de caso que permite complementar a validação do sistema Padrão já apresentada no capítulo anterior. Recorda-se que a utilização do estudo de caso visa completar a validação preliminar efectuada ao sistema, através da asserção. Objectivamente, e uma vez que já foi avaliado o sistema de inferências utilizado pelo Padrão, assim como averiguada a sua capacidade de identificação de relacionamentos implícitos nos dados, pretende-se com este estudo de caso confirmar a utilidade do sistema Padrão na análise de BD organizacionais geo-referenciadas, nomeadamente, na identificação de relacionamentos existentes entre os dados geo-espaciais e os dados não geográficos.

Antes de proceder com a execução das diversas fases previstas no Padrão, e que integram o processo de descoberta de conhecimento, é realizado o exercício de compreensão dos dados, o qual é apresentado na primeira secção deste capítulo. Este exercício permite a familiarização com o conteúdo e o tipo dos dados. O capítulo prossegue, segunda secção, com a definição dos objectivos do DM, aos quais é seguida a implementação das várias etapas do processo de descoberta de conhecimento, que permitem a sua concretização. Este capítulo culmina com a sistematização das principais dificuldades encontradas na análise desta BD.

7.1 Compreensão dos dados

Nesta fase procura-se compreender o conteúdo das tabelas que integram a BD (Figura 7.1), assim como a sua relevância para o processo de descoberta de conhecimento. Para cada uma

das tabelas que integram a BD, é de seguida listado o seu conteúdo, referindo os seus atributos, assim como o tipo e o signi…cado de cada um dos mesmos. É ainda destacado o número total de registos que integram cada tabela.

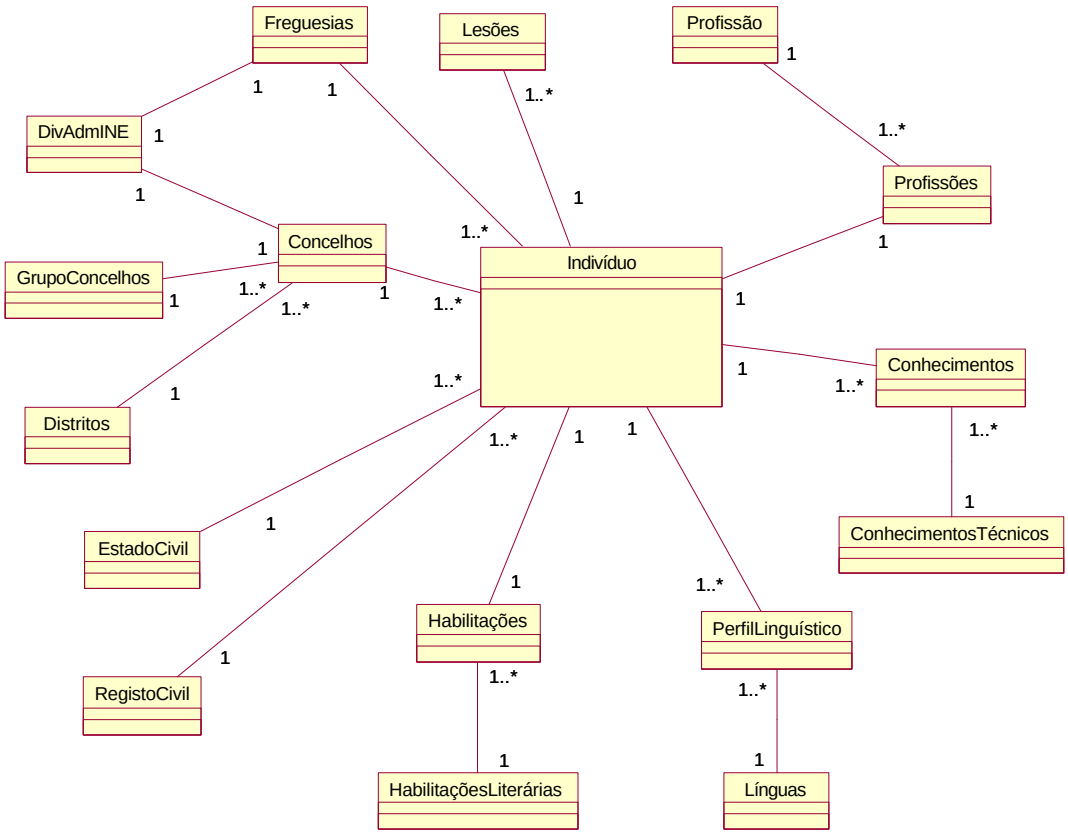


Figura 7.1: Estrutura lógica da BD a analisar

Tabela Freguesias: 4.269 registos

Concelho	Char(3)	Código do concelho
Freguesia	Char(2)	Código da freguesia
DesigFreguesia	Char(60)	Designação da freguesia
DivAdmINE	Long Integer	Divisão administrativa INE[1]

Tabela Concelhos: 313 registos

Concelho	Char(3)	Código do concelho
DesigConcelho	Char(60)	Designação do concelho
DistritoAdm	Integer	Identi…cação distrito administrativo
DivAdmINE	Long Integer	Divisão administrativa INE
GrupoConcelhos	Char(3)	Grupo de concelhos

[1] Instituto Nacional de Estatística.

Tabela GrupoConcelhos: 23 registos

GrupoConcelho	Char(3)	Código do grupo de concelhos
DesigGrupo	Char(60)	Designação do grupo de concelhos

Tabela Distritos: 21 registos

Distrito	Integer	Distrito administrativo
DesigDistrito	Char(26)	Designação do distrito
AbrevDistrito	Char(26)	Abreviatura do distrito administrativo

Tabela DivAdmINE: 4.581 registos

DivAdmINE	Long Integer	Divisão administrativa INE
DesigDivAdmINE	Char(39)	Designação da divisão administrativa INE

Tabela Línguas: 20 registos

Língua	Char(1)	Código da língua
DesigLíngua	Char(30)	Designação da língua

Tabela HabilitaçõesLiterárias: 1.425 registos

HabLiterária	Char(4)	Código da habilitação literária
DesigHabLiterária	Char(45)	Designação da habilitação literária
DescrHabLiterária	Char(130)	Descrição da habilitação literária
EscolaridadePrévia	Char(2)	Escolaridade prévia requerida
Conclusão	Char(2)	Conclusão
Bacharelato	Char(2)	Bacharelato
Doutoramento	Char(2)	Doutoramento
IdadeMáx	Char(2)	Idade Máxima
Admissão	Char(2)	Admissão

Tabela Profissão: 318 registos

Profissão	Char(4)	Código da pro...ssão
DesigProfissão	Char(60)	Designação da pro...ssão

Tabela RegistoCivil: 325 registos

ConservRC	Integer	Código da conservatória do registo civil
DesigConservRC	Char(26)	Designação da conservatória do registo civil
Abreviatura	Char(26)	Abreviatura da conservatória do registo civil

Tabela ConhecimentosTécnicos: 24 registos

ConhTécnico	Char(1)	Código do conhecimento técnico
DesigConhTécnico	Char(60)	Designação do conhecimento técnico

Tabela EstadoCivil: 7 registos

EstadoCivil	Char(1)	Código do estado civil
DesigEstadoCivil	Char(35)	Designação do estado civil

Tabela Indivíduos: 1.328.573 registos

Número	Long Integer	Número de identi...cação militar
DataNascimento	Date	Data de nascimento
Sexo	Char(1)	Sexo
Concelho	Char(3)	Código do concelho
Freguesia	Char(2)	Código da freguesia
ConservRC	Integer	Código da conservatória do registo civil
EstadoCivil	Char(1)	Código do estado civil
GrauAcadémico	Char(1)	Grau académico
GrupoSanguíneo	Char(3)	Grupo sanguíneo

Tabela PerfilLinguístico: 362.301 registos

Número	Long Integer	Número de identi...cação militar
Língua	Char(1)	Código da língua
GrauConhecimento	Char(1)	Grau conhecimento
GrauFala	Char(1)	Grau fala
GrauEscreve	Char(1)	Grau escrita
GrauTraduz	Char(1)	Grau traduz
GrauCompreensão	Char(1)	Grau compreensão
GrauLeitura	Char(1)	Grau leitura

Tabela Habilitações: 803.368 registos

Número	Long Integer	Número de identi...cação militar
HabLiterária	Char(4)	Código da habilitação literária
LocalHabLiterária	Char(10)	Local da habilitação literária obtida
FreqHabLiterária	Char(2)	Frequência da habilitação literária obtida
CompHabLiterária	Char(1)	Comprovação da habilitação literária obtida

Tabela Profissões: 372.525 registos

Número	Long Integer	Número de identi...cação militar
Profissão	Char(4)	Código da pro...ssão
VinProfissional	Char(1)	Vínculo pro...ssional
TempoServiço	Char(1)	Tempo de serviço
TempoAbandono	Char(1)	Tempo de abandono

Tabela Conhecimentos: 487.964 registos

Número	Long Integer	Número de identi...cação militar
ConhTécnico	Char(1)	Código do conhecimento técnico
GrauConhTécnico	Char(1)	Grau de conhecimento

Tabela Lesões: 492.204 registos

Número	Long Integer	Número de identi…cação militar
Lesão	Char(3)	Código da lesão[2]
SIVAGE	Char(1)	Factor SIVAGE
GrauLesão	Integer	Grau da lesão detectada

As próximas subsecções sintetizam as informações associadas aos indivíduos, e que permitem a sua caracterização. As tabelas sintetizadas são: Indivíduos, PerfilLinguístico, Habilitações, Profissões, Conhecimentos e Lesões.

7.1.1 Tabela Indivíduos

A tabela Indivíduos armazena o denominado Cadastro Geral do SIAPE, integrando 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. Chama-se a atenção, mais uma vez, para o facto desta tabela armazenar o cadastro geral do Exército, na qual se encontram dados referentes ao pessoal do quadro permanente e aos mancebos que, até à data em que esta BD foi disponibilizada para análise (Novembro de 2000), foram registados na mesma.

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 Indivíduos não se encontram preenchidos. A Figura 7.2 apresenta a distribuição por Concelho (uma vez que as freguesias são generalizadas até este nível ou até ao nível dos distritos), Estado Civil e Sexo, dos diversos registos que integram esta tabela.

Pela análise da Figura 7.2 veri…ca-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 em quaisquer tarefas de DM que venham a ser de…nidas. O mesmo acontece com o atributo Sexo, uma vez que dos 1.328.573 registos que integram esta tabela, 1.319.175 dizem respeito à 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 especi…camente as datas

[2] A tabela de descodi…cação destes códigos não foi fornecida pelo Comando de Pessoal do Exército, visto a mesma se encontrar em fase de restruturação. O resultado dos vários exames médicos efectuados aos indivíduos é transformado no atributo SIVAGE.

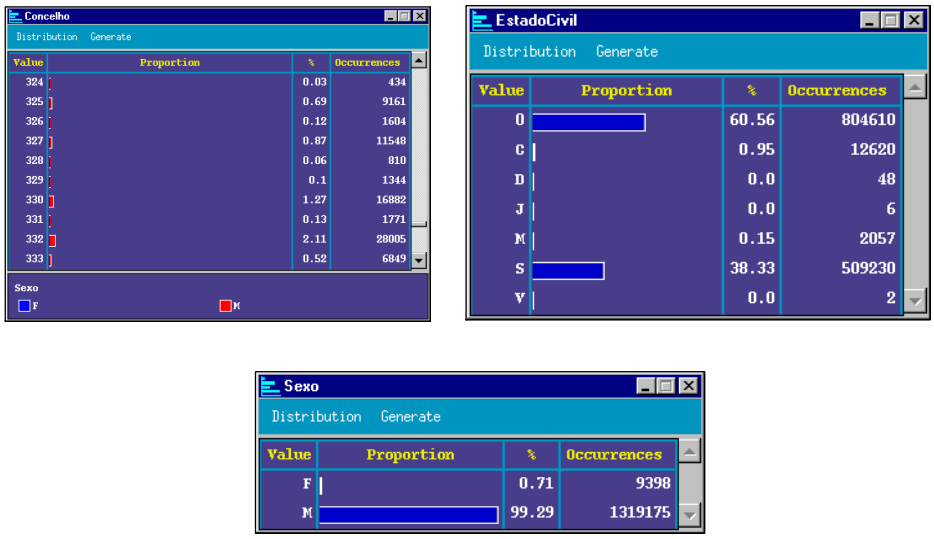


Figura 7.2: Distribuição dos indivíduos pelos atributos Concelho, EstadoCivil e Sexo

12/06/979, 08/12/2886, 03/01/1989 e 14/03/1994, podendo as mesmas ser recti…cadas para os valores correctos, 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'.

7.1.2 Tabela PerfilLinguístico

O per…l linguístico caracteriza o grau de conhecimento de uma determinada língua. Apesar de terem sido criados 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 classi…cado recorrendo aos indicadores: D (desconhecido), M (muito), R (regular) e P (pouco). A Figura 7.3 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.

Pela análise da Figura 7.3 veri…ca-se que o atributo GrauConhecimento apresenta um registo com código omisso, devendo o mesmo ser substituído pelo código D. Constata-se, ainda, que 679 registos apresentam o código N, o qual não consta do conjunto de valores possíveis de…nido 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 hipótese foram averiguadas junto do Comando de Pessoal, permitindo veri…car 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 7.3), veri…ca-se que existe uma predominância do conhecimento do Inglês, Francês, Espanhol e Alemão, existindo

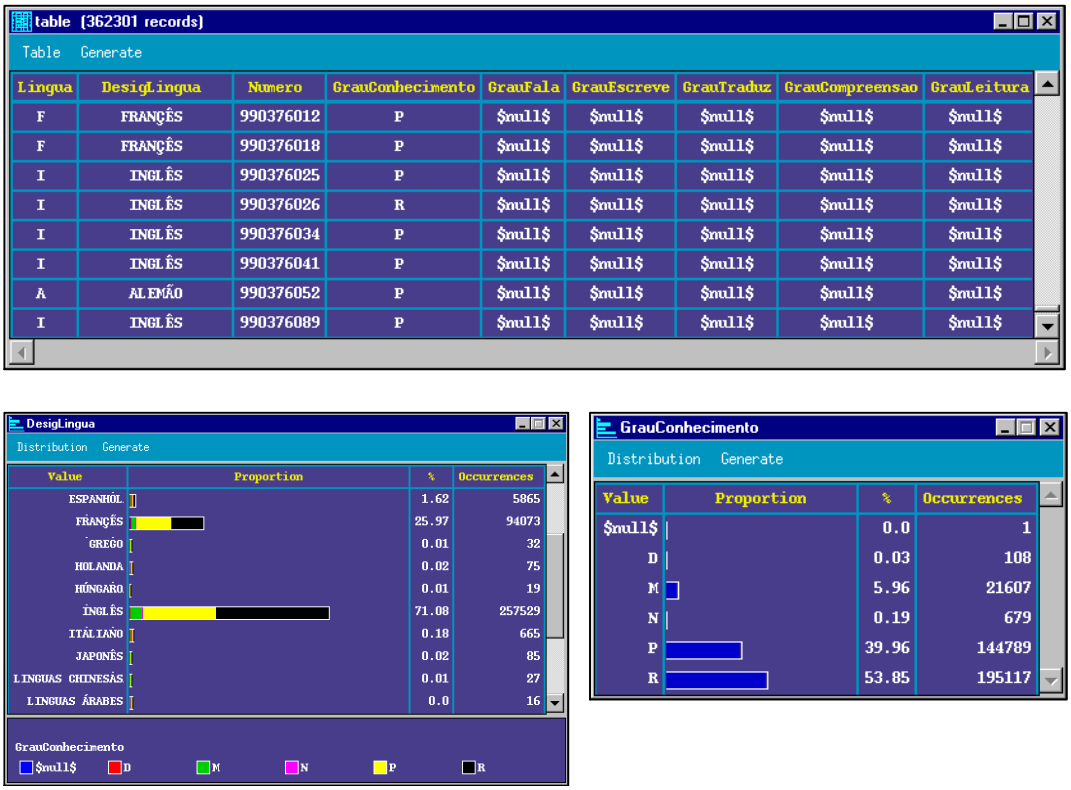


Figura 7.3: Conteúdo da tabela PerfilLinguístico

contudo, conhecimentos nas outras línguas consideradas.

7.1.3 Tabela Habilitações

A tabela Habilitações armazena as habilitações literárias que caracterizam os indivíduos. Além da indicação do código da habilitação e do local em que a mesma foi obtida (ou está a ser obtida), esta tabela armazena o nível de frequência atingido (ou o nível de frequência actual). Este atributo é caracterizado através de valores como: 1C, 1M, 4C ou 9C, representando o dígito o ano de estudo atingido ou o ano de estudo actual, e a letra se este está completo (C) ou se o indivíduo está matriculado (M) no mesmo.

Uma análise detalhada ao conteúdo desta tabela permitiu constatar que o atributo LocalHabLiterária não se encontra preenchido, não permitindo a sua inclusão no processo de descoberta de conhecimento. Os atributos FreqHabLiterária e CompHabLiterária não se encontram completamente preenchidos[3]. Ao nível das habilitações propriamente ditas, veri…ca-se uma distribuição dos indivíduos por 1.226 das 1.425 habilitações literárias de…nidas na tabela HabilitaçõesLiterárias, solicitando a de…nição de uma hierarquia para as mesmas, que

[3] No que diz respeito ao atributo CompHabLiterária, 45% dos registos apresentam este valor em falta. Em relação ao atributo FreqHabLiterária, apenas 1% dos registos não estão preenchidos. Contudo, considera-se que este atributo possui um carácter informativo, não se prevendo a sua inclusão no processo de descoberta de conhecimento.

permita a generalização de habilitações em áreas semelhantes. Este procedimento diminuirá o número de casos distintos, permitindo a exploração deste atributo pelos algoritmos de DM.

Apesar da generalização de atributos ser uma tarefa tradicionalmente executada na fase de pré-processamento dos dados, a mesma é neste momento apresentada por se considerar que a generalização destes dados, e a sua posterior distribuição por um número reduzido de classes, permitirá identi…car, ainda na fase de compreensão dos dados, as principais áreas de formação dos indivíduos.

A de…nição da hierarquia a utilizar permitiu caracterizar a população alvo do estudo através de duas vertentes distintas: os indivíduos que possuem habilitações ao nível da escolaridade obrigatória, básica ou ao nível do secundário, e os que têm formação especializada em determinado domínio, quer esta tenha sido obtida via ensino superior ou não. A Figura 7.4 evidencia a distribuição dos registos pelas diversas classes consideradas. Refere-se que a classe etiquetada pelo indicador Ensino superior, não agrupa os indivíduos da BD com formação superior, mas apenas aqueles cidadãos que não obtiveram o grau no país, e aos quais é atribuída equivalência, sem especi…car a área ou o curso em questão.

Value	%	Occurrences
Administracao	1,21	9745
Agricultura	0,83	6698
Alimentacao	0,11	860
Arquitectura	0,70	5601
Artes	1,21	9756
Biociencias	0,41	3254
Ciencias Pol/Edu/Sociais	0,42	3348
Ciencias do ambiente	0,95	7634
Comunicacao Social	1,35	10841
Construcao civil	1,50	12036
Desenho	0,42	3361
Desporto	0,47	3811
Economia e Gestao	6,36	51117
Electricidade	0,26	2068
Electronica	1,39	11164
Electrotecnia	1,17	9437
Ensino	0,92	7390
Ensino basico	0,26	2114
Ensino secundario	6,01	48308
Ensino superior	0,07	601
Escolaridade obrigatoria	30,63	246050
Estatistica	0,03	237
Filosofia	0,09	727
Fisica	0,39	3137
Geografia	0,12	967
Geologia/Geotecnia	0,05	390
Geral unificado 6-9	29,37	235910
Historia	0,20	1629
Informatica	3,16	25376
Linguas	0,35	2812
Maritimas	0,10	823
Marketing	0,20	1637
Matematica	0,48	3869
Mecanica	1,57	12651
Nao Fala Portugues	0,01	55
Nao sabe ler nem escrever	0,22	1735
Outras	3,55	28528
Quimica	0,63	5099
Religiao	0,08	620
Sabe ler e/ou escrever	0,04	299
Saude	2,62	21072
Textil	0,07	600

Figura 7.4: Distribuição dos indivíduos por grupo de habilitação literária

Salienta-se ainda que, dos 803.368 registos que integram esta tabela, 246.050 apresentam como habilitação literária a escolaridade obrigatória, enquanto que 235.910 possuem o nível geral uni…cado 6-9. As áreas de conhecimento em que os indivíduos apresentam maior incidência são a Economia e Gestão, a Informática e a Saúde.

7.1.4 Tabela Profissões

Ao nível das Profissões, esta tabela regista a pro…ssão de cada um dos indivíduos, o vínculo pro…ssional, o tempo de serviço e o tempo de abandono do mesmo, para cumprimento do serviço militar. Este último atributo encontra-se na maioria dos casos por preencher, não permitindo que o mesmo seja considerado nas análises que serão posteriormente efectuadas.

O vínculo pro…ssional é codi…cado segundo os valores: A, D e R[4]. Pela análise da Figura 7.5 constata-se que 92% dos casos dizem respeito a vínculos R, sendo os restantes 8% divididos entre as classes A e D. Veri…ca-se, ainda, que existe um registo classi…cado com o valor X, representando um erro de digitação. Mais uma vez, não existe uma distribuição homogénea dos registos pelas diversas classes.

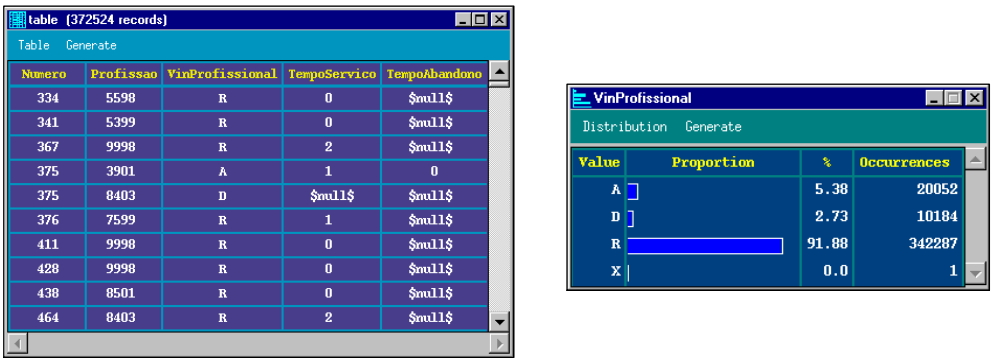


Figura 7.5: Distribuição dos indivíduos por vínculo pro…ssional

No que diz respeito às pro…ssões, foi possível constatar que os indivíduos se encontram distribuídos por 315 pro…ssões, das 318 possíveis de…nidas na tabela Profissão da BD. Tal sugere, mais uma vez, a criação de uma hierarquia que permita a generalização deste vasto conjunto de pro…ssões. Esta tarefa, realizada em duas etapas, permitiu criar dois níveis, CatProfissão e GrupoProfissão, que conduziram à de…nição de 53 grupos de pro…ssões. O nível intermédio, das categorias, auxiliou todo este processo, permitindo agrupar as pro…ssões com poucos indivíduos no grupo Diversos (este grupo integra as categorias com menos de 100 elementos).

A Figura 7.6 retrata a distribuição dos indivíduos pelos grupos de…nidos para as pro…ssões. A análise da mesma permite constatar que as pro…ssões que agregam mais elementos são as relacionadas com a Agricultura, área Comercial, Electricidade, Mecânica e Serralharia, evidenciando as saídas pro…ssionais dos indivíduos com habilitações literárias ao nível da escolaridade obrigatória ou nível geral uni…cado 6-9.

[4] Não foi possível obter o signi…cado destes códigos, uma vez que o Comando de Pessoal não os conseguiu identi…car. Tal não constitui um problema, já que este atributo não será utilizado no processo de descoberta de conhecimento, como poderá ser constatato mais adiante, na secção 7.2.

Value	%	Occurrences
Administracao/Contabilidade	3,78	14085
Agricultura	6,78	25257
Alfaiate	0,09	321
Arquitectura/Desenho	0,47	1749
Artes graficas	0,62	2301
Artesanato	0,17	625
Barbeiro/Cabeleireiro	0,22	822
Bombeiro	0,18	685
Borracha/Plasticos	0,61	2256
Carpinteiro	3,65	13583
Carteiro	0,43	1593
Cobrador/Revisor	0,03	113
Comercial	5,99	22315
Comunicacao social	0,11	421
Construcao civil	1,63	6059
Cozinheiro	0,83	3077
Desportista	0,37	1366
Direccao/Gestao	1,46	5454
Diversos	0,09	339
Electricista	5,20	19362
Engenheiro	0,12	454
Ferragens	0,03	115
Fotografia	0,15	574
Frezador	0,03	107
Industria calcado	1,04	3864
Informatica	1,03	3836
Jurista	0,03	120
Litografia	1,18	4378
Maquinista	0,67	2496
Marcenaria	2,10	7824
Mecanica	5,95	22158
Media	0,30	1102
Metais	1,63	6073
Motorista	3,02	11239
Musica	0,10	361
Nao especificado	15,36	57210
Padeiro/Pasteleiro	2,17	8084
Pedreiro	7,42	27643
Pintura/Escultura	2,56	9527
Professor	0,90	3357
Recepcionista	1,93	7183
Relojoaria/Ourivesaria	0,34	1258
Restauracao/Hotelaria	5,99	22303
Sapateiro	0,81	3036
Saude	0,45	1663
Seguranca	0,60	2244
Serralheiro	5,25	19553
Soldador	1,42	5286
Tecnicos diversos	0,10	390
Telefonista/Telegrafista	0,09	335
Textil	3,43	12778
Topografia	0,49	1816
Vidreiro	0,64	2374

Figura 7.6: Distribuição dos indivíduos por grupo de pro...ssão

7.1.5 Tabela Conhecimentos

A tabela Conhecimentos armazena os conhecimentos técnicos apresentados pelos indivíduos. A cada conhecimento técnico declarado é associado um grau de conhecimento, restringido aos valores: M (muito), P (pouco) e R (regular). A Figura 7.7 evidencia a distribuição dos registos armazenados nesta tabela, atendendo aos atributos ConhTécnico e GrauConhTécnico.

7.1.6 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

Value	%	Occurrences
ANDAR A CAVALO	2,62	12799
ANDAR DE MOTO	31,59	154132
BOMBEIRO	2,19	10662
CINEMA E/OU TV	0,88	4273
COZINHA	3,57	17414
DACTILOFRAFIA	5,46	26656
DESENHO DE CONSTRUÇÃO	2,06	10034
ELECTRICIDADE	8,02	39131
ELECTRÓNICA	5,59	27263
FARMÁCIA	0,34	1637
FOTOGRAFIA	2,01	9798
MECANICA DE PRECISÃO	0,44	2142
MECANICA GERAL	5,37	26198
MECANOGRAFIA	16,32	79651
MÚSICA-CORDAS (OBRIGATÓRIO SABER SOLFEJO)	1,49	7257
MÚSICA-PERCURSÃO (OBRIGATÓRIO SABER SOLFEJO)	0,86	4183
MÚSICA-SOPRO (OBRIGATÓRIO SABER SOLFEJO)	1,31	6393
PRIMEIROS SOCORROS	3,49	17007
RADIOAMADOR	0,45	2195
RÁDIO E SOM	2,32	11315
TOPOGRAFIA	0,81	3938
TRATAMENTO DE CAVALOS	1,11	5440
TRATAMENTO DE CÃES	1,73	8445

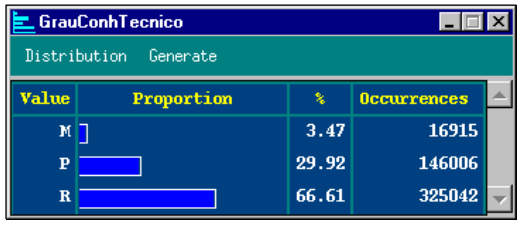


Figura 7.7: Distribuição dos indivíduos por conhecimento técnico e por grau de conhecimento

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 7.8 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 …gura 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 identi…cação do valor correcto do atributo SIVAGE nestes registos, os mesmos serão removidos do conjunto de dados a analisar.

Veri…cam-se, ainda, casos de duplicação no armazenamento de um determinado factor SIVAGE, para um dado indivíduo (como pode ser constatado na Figura 7.8, para o indivíduo com o número 73). Estas duplicações são ocasionadas pela veri…cação de diferentes lesões, que na sua codi…cação 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-

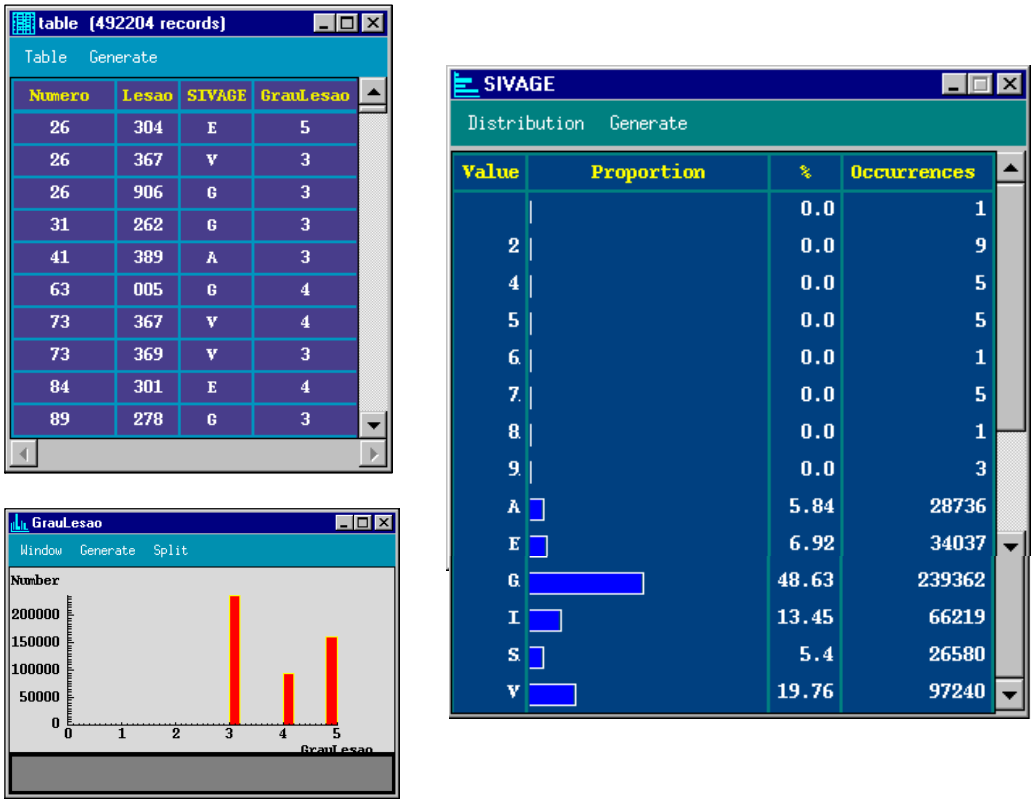


Figura 7.8: Distribuição dos indivíduos pelos atributos SIVAGE e grau de lesão

se, na análise, o registo que apresenta o maior grau, permitindo uma aproximação por excesso, e nunca por defeito, à realidade.

7.2 De…nição dos objectivos do DM

Como já referido anteriormente, no Capítulo 4, as tarefas de DM agrupam-se em dois grandes grupos: descrição e previsão. A descrição permite a identi…cação de regras que descrevem ou caracterizam os dados em análise, enquanto que a previsão permite utilizar algumas variáveis da BD, para prever o valor de um outro atributo da BD.

No sentido de explorar estes dois tipos de tarefas, e uma vez que a exploração de dados espaciais permite: i) determinar características espaciais; ii) efectuar análise espacial discriminante; iii) de…nir regras de associação espacial e, iv) detectar tendências espaciais nos dados, foram de…nidos como objectivos do DM a realização das seguintes tarefas:

Tarefa 1. Caracterização espacial do per…l linguístico dos indivíduos. Na realização desta tarefa pretende-se veri…car a distribuição espacial associada ao conhecimento de uma dada língua, nomeadamente o Alemão e o Francês. Este exercício permitirá identi…car as zonas geográ…cas que apresentam uma maior incidência de determinado grau de conhecimento. No que diz respeito à componente geográ…ca, a análise será efectuada ao nível dos distritos.

Tarefa 2. Análise espacial discriminante do per…l linguístico, para os diversos distritos que integram Portugal continental. A realização desta tarefa permitirá identi…car um conjunto de regras que identi…quem 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 pro…ssões exercidas pelos mesmos. Para a realização desta tarefa, e no sentido de efectuar a análise ao nível dos concelhos, a identi…cação das regras de associação espacial será elaborada para o distrito de Braga.

Tarefa 4. Detecção de tendências espaciais no factor SIVAGE, que permitam identi…car alterações regulares nos graus de lesão veri…cados pelos indivíduos. A realização desta tarefa permitirá identi…car alterações regulares dos graus de lesão, associados a regiões que sucessivamente se afastam de uma dada entidade geográ…ca. Ao nível geográ…co, esta tarefa será realizada analisando os concelhos que integram o distrito de Braga.

7.3 O processo de descoberta de conhecimento

As diversas fases do processo de descoberta de conhecimento, necessárias à satisfação dos objectivos do DM descritos anteriormente, e consideradas pelo sistema Padrão, são apresentadas nas próximas subsecções.

7.3.1 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, identi…caram-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 para cada uma das tarefas), permitindo optimizar 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 …les, …cheiros 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 Indivíduos, 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 recti…cadas 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".

- No pré-processamento dos dados foi criado um novo atributo, AnoNascimento, com o objectivo de converter as datas de nascimento em anos. Este novo atributo permitiu veri…car a distribuição dos indivíduos pelos diversos anos de nascimento, que vão desde 1858 a 1983, conduzindo à de…nição de 14 classes a utilizar na agregação dos mesmos. As classes identi…cadas foram: [1858, 1907], [1908, 1920], [1921, 1930], [1931, 1935], [1936, 1940], [1941, 1947], [1948, 1955], [1956, 1960], [1961, 1971], [1972, 1973], [1974, 1975], [1976, 1977], [1978, 1979] e [1980, 1983]. Estas 14 classes distribuem-se por dois grupos distintos. O primeiro grupo com 8 classes, que vai até ao ano de 1960, agrega indivíduos do quadro permanente, sendo a distribuição por cada uma classes de aproximadamente 3000 indivíduos. Em 1961 veri…ca-se um aumento signi…cativo do número de indivíduos nascidos neste ano, salientando o inicio do registo na BD dos dados dos mancebos. As 6 classes que integram este segundo grupo, apresentam uma distribuição média na ordem dos 200.000 indivíduos por classe.

Para a tabela Habilitações, e uma vez que já foi de…nida a hierarquia a considerar na generalização das habilitações literárias, apenas foi necessário proceder à eliminação dos atributos LocalHabLiterária, FreqHabLiterária e CompHabLiterária.

No caso da tabela Profissões, não foram considerados os atributos VinProfissional, TempoServiço e TempoAbando, por se considerar que os mesmos não são relevantes aos objectivos de…nidos para o DM. A hierarquia de conceitos a considerar para este atributo também já foi de…nida, na fase de compreensão dos dados apresentada na secção 7.1.

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.

Dada a elevada quantidade de dados armazenada nesta BD, e de por vezes ser necessário optimizar o processo de análise dos dados, refere-se que é possível seleccionar uma amostra representativa do universo[5] em questão, na construção dos conjuntos de dados de treino e de teste. A amostragem[6] por estrati…cação [Han e Kamber, 2001] permite a selecção desta amostra,

[5] Conjunto global de dados a analisar.

[6] A amostragem (sampling) [Han e Kamber, 2001] representa uma técnica de redução de dados, que permite que um conjunto de dados de grande dimensão seja representado por uma amostra aleatória, ou subconjunto

garantindo a representatividade do universo, já que a distribuição dos registos, em função da variável objectivo, permanece idêntica à distribuição do conjunto de dados global. A amostra representativa do conjunto inicial é construída limitando o número total de registos que a mesma pode conter. Apesar da amostragem não ter sido utilizada como técnica de redução dos dados, é aqui referida por poder ser utilizada em posteriores exercícios de análise destes dados e principalmente, porque os seus princípios foram utilizados na construção dos conjuntos de dados de treino e de teste, garantindo a distribuição real dos dados em ambos os conjuntos.

Para a satisfação da primeira tarefa, de…nida nos objectivos do DM, foi necessário proceder à integração de dados da tabela Indivíduos, 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 de…nição 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 geográ…cas), interferindo com o desempenho dos algoritmos de DM. Ao nível dos distritos, e considerando a análise geográ…ca restrita ao continente, é possível agregar a informação a analisar em 18 classes distintas, que representam os 18 distritos que integram Portugal continental. Apesar do elevado número de registos envolvidos, nesta tarefa especí…ca, não se optou por seleccionar uma amostra representativa dos dados, na construção do conjunto de dados de treino e de teste. Estes conjuntos foram gerados a partir de todos os dados disponíveis para esta tarefa. Tal permite ao utilizador especi…car, na fase de modelação, a língua alvo do estudo. Esta abordagem facilita 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 7.9. No …nal deste processo, o conjunto de dados de treino agrega 120.767 registos, enquanto que 241.534 registos podem ser utilizados para veri…car o desempenho dos modelos obtidos.

Para a realização da segunda tarefa, análise espacial discriminante do per…l linguístico, podem ser utilizados os conjuntos de dados de treino e de teste, construídos para a primeira tarefa, uma vez que estes não apresentam qualquer restrição geográ…ca ou de línguas a analisar.

Para a identi…cação de regras de associação espacial, que permitam encontrar associações espaciais entre as habilitações literárias e as pro…ssões dos indivíduos, construíram-se os conjuntos de dados de treino e de teste para esta tarefa, a partir dos 39.928 registos[7] disponíveis para o distrito de Braga. A distribuição destes registos, pelas diversas classes disponíveis para o ano de nascimento, permitiu veri…car que a classe '1956-1960' agrega um número bastante reduzido de registos (apenas 2), quando comparada com as restantes classes representadas (a partir do

dos dados, de dimensão mais reduzida. A amostragem por estrati…cação deve ser utilizada quando os dados não apresentam uma distribuição uniforme em relação à variável objectivo. Consiste, basicamente, em ordenar os registos a analisar pelos valores possíveis para a variável objectivo, o que permite construir os conjuntos de dados de treino e de teste seleccionando 1 em cada N registos para o conjunto de treino, e não seleccionando 1 em cada N registos para o conjunto de teste (1-N). No …nal deste processo, a amostra obtida está limitada a determinado número de registos especi…cado pelo utilizador. A distribuição dos registos pelos valores possíveis para a variável objectivo permanece semelhante à distribuição do conjunto inicial dos dados, garantindo a representatividade dos mesmos.

[7] Estes registos foram obtidos através da integração das tabelas Indivíduos, Habilitações e Profissões. Do conjunto global de registos obtidos, seleccionaram-se os respeitantes a indivíduos do distrito de Braga.

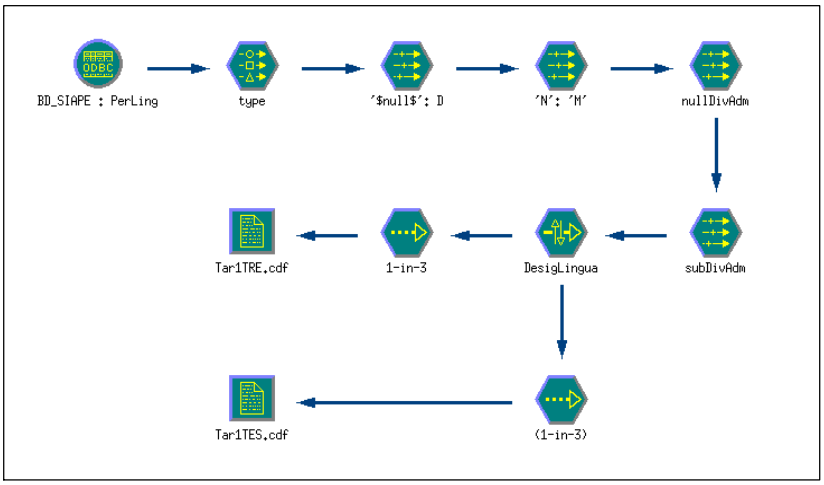


Figura 7.9: Conjunto de dados de treino e de teste para a caracterização do per…l linguístico

ano 1961). Dada a discrepância veri…cada, estes dois registos foram removidos do conjunto de dados a analisar. A Figura 7.10 apresenta a stream construída nesta etapa para a obtenção dos …cheiros de treino e de teste, necessários nas fases seguintes.

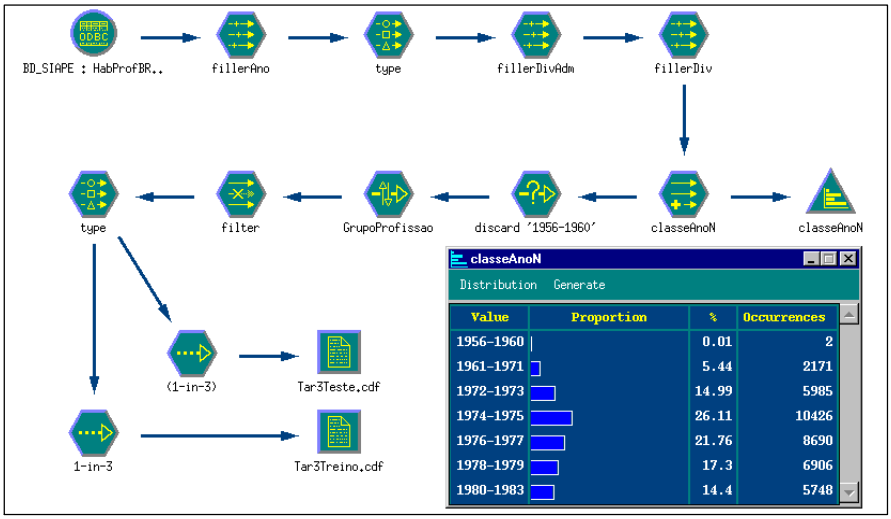


Figura 7.10: Conjunto de dados de treino e de teste para a identi…cação de regras de associação espacial

Para a satisfação da quarta tarefa de DM, foi necessário proceder à integração de dados armazenados na tabela Indivíduos, 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 7.11 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).

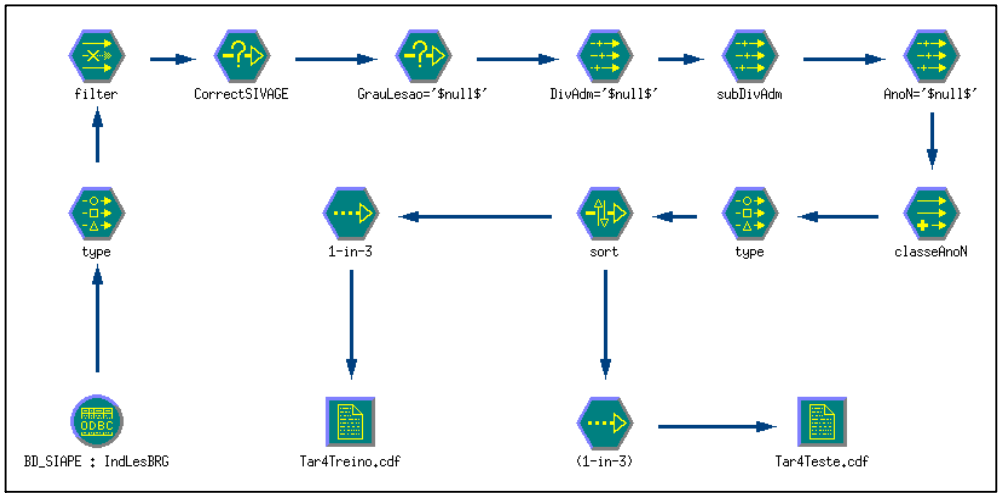


Figura 7.11: Conjunto de dados de treino e de teste para a identi…cação de tendências espaciais nos dados

7.3.2 Processamento da informação geo-espacial

A abordagem ao processamento da informação geo-espacial, já apresentada no Capítulo 5, inclui 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.

Esta aproximação apresenta como vantagem a possibilidade de utilização de modelos geográ…cos construídos em anteriores exercícios de DM, que se adequem à tarefa em causa.

No caso da realização da primeira e segunda tarefas, e dado que a componente geográ…ca foi generalizada até ao nível dos distritos, não é necessário proceder à inferência de informação geográ…ca. Nestas tarefas, a informação geográ…ca 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. Recorda-se que a tabela Faces, da BDG, 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 geográ…ca necessária, à satisfação destas tarefas, seja integrada na fase de DM (apresentada na próxima subsecção).

No que diz respeito à terceira tarefa, e uma vez que a mesma pretende identi…car regras de associação espacial, num determinado conjunto de registos do distrito de Braga, deverá

ser utilizado o conhecimento espacial já inferido[8] para o distrito de Braga, e que se encontra armazenado na tabela geoBraga da BDG. Esta situação é, também, a veri…cada na satisfação da quarta tarefa.

7.3.3 Data Mining

Na fase de DM apenas é necessário seleccionar o algoritmo (ou algoritmos) 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[9] resultante identi…cará a região, ou regiões, em que se veri…ca uma maior incidência no conhecimento de uma dada língua.

A Figura 7.12 apresenta a stream construída para a caracterização geográ…ca do per…l linguístico, nomeadamente no conhecimento do Alemão e do Francês. Na referida …gura é possível constatar que, ao …cheiro com os dados de treino (Tar1Tre.cdf) é integrada a tabela que armazena a hierarquia conceptual de…nida para o domínio geográ…co (BDG: Hierarquias). Posteriormente, a relação topológica existente entre distritos[10] (BDG: RelDistritos) é integrada com a generalização realizada, permitindo ao algoritmo de DM utilizado conhecer a relação topológica existente entre distritos.

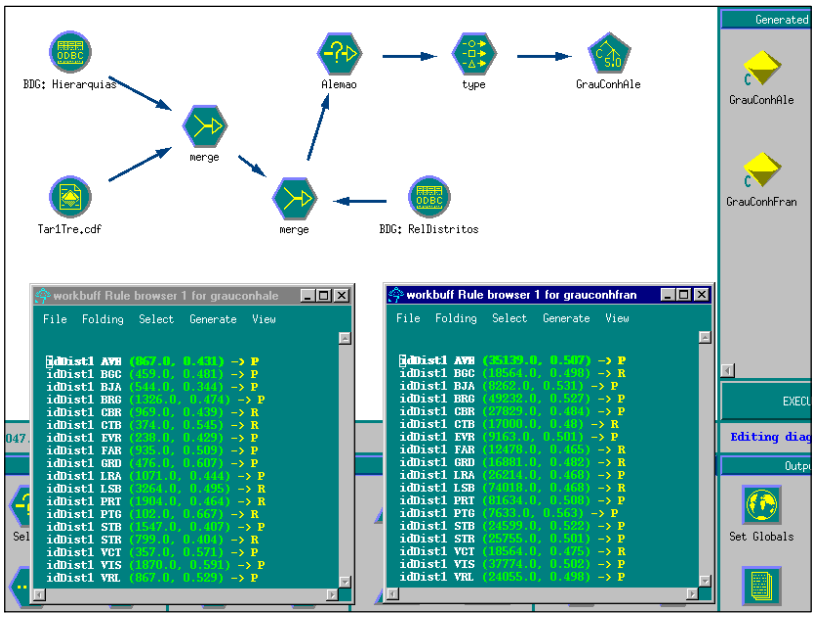


Figura 7.12: Utilização do algoritmo C5.0 na caracterização do conhecimento do Alemão e do Francês

[8] Este processo foi apresentado no Capítulo 6, na avaliação realizada ao sistema qualitativo de inferências.

[9] A frequente utilização de árvores de decisão, como técnica de DM, é justi…cada não só pelo facto desta técnica permitir analisar dados numéricos e alfanuméricos, mas principalmente, pela facilidade de interpretação dos resultados obtidos com a mesma.

[10] Como já referido anteriormente, a relação topológica existente entre distritos é determinada a partir da informação explícita na tabela Faces da BDG, que armazena as relações espaciais existentes entre concelhos adjacentes.

Na Figura 7.12 é ainda possível veri…car que foram construídos dois modelos, o primeiro, GrauConhAle, caracteriza a distribuição geográ…ca do conhecimento do Alemão, enquanto que o modelo GrauConhFran, caracteriza a distribuição geográ…ca do conhecimento do Francês. As regras explícitas em cada um destes modelos podem ser visualizadas na mesma …gura (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 con…ança associada às mesmas. A con…ança 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 classi…cação do conjunto de dados de teste.

A estratégia seguida para a construção do conjunto de dados de treino e de teste permitiu, na realização desta tarefa, que a mesma stream fosse utilizada na caracterização das duas línguas, Alemão e Francês, uma vez que o nodo select utilizado possibilita a especi…cação da língua a analisar.

Na segunda tarefa, e tendo como objectivo discriminar o grau de conhecimento das diversas línguas nos diferentes distritos analisados, utilizaram-se os princípios seguidos na execução da tarefa anterior, não sendo contudo necessário especi…car uma dada língua. Todas as línguas existentes na BD foram analisadas simultaneamente, permitindo ao algoritmo C5.0 discriminar o grau de conhecimento das diversas línguas, pelos diferentes distritos. A Figura 7.13 apresenta a stream construída para a fase de DM. Na mesma …gura é ainda evidenciada parte da árvore de decisão obtida, a qual discrimina o grau de conhecimento de uma dada língua, nas diferentes regiões analisadas.

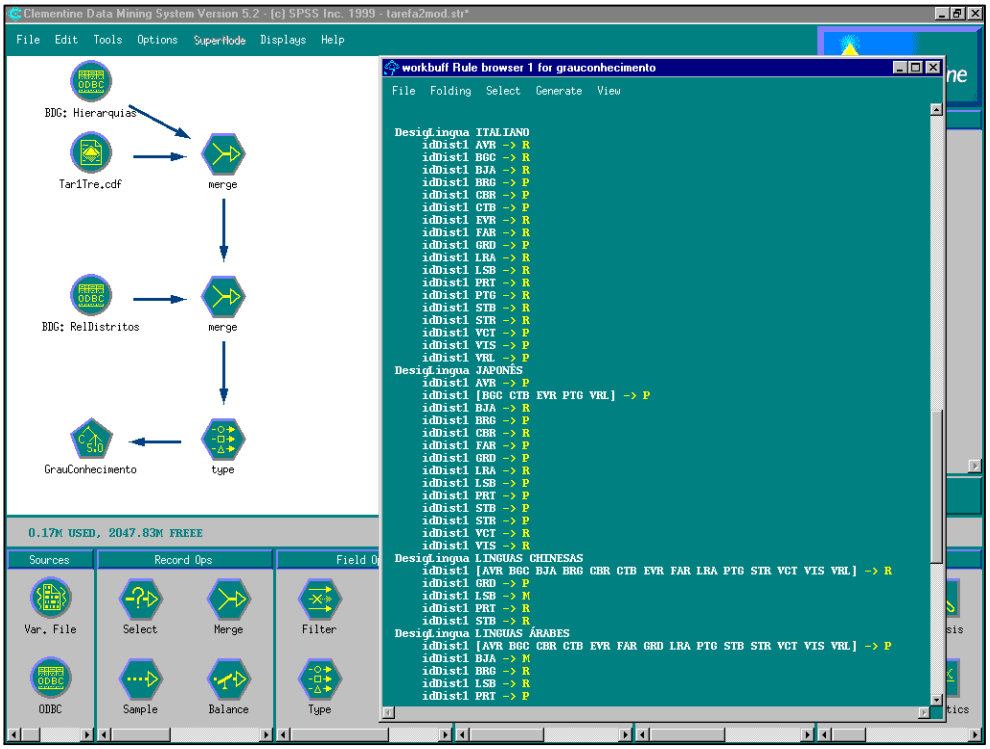


Figura 7.13: Utilização do algoritmo C5.0 na discriminação do per…l linguístico por região

A identi…cação de regras de associação espacial entre habilitações e pro…ssões, contextualizadas geogra…camente, foi conseguida recorrendo à stream apresentada na Figura 7.14. Nesta …gura, é possível veri…car que ao conjunto de dados de treino é integrada a componente geográ…ca (BDG: geoBraga) da região em estudo. Nesta tarefa, foi utilizado o algoritmo Generalised Rule Induction (GRI) disponibilizado pelo Clementine, para a indução de regras de associação[11]. Este algoritmo permite que os dados de entrada sejam numéricos ou simbólicos, sendo este último o tipo do atributo de saída. Apesar de não ter sido incluído nos objectivos desta tarefa, foi construída uma árvore de decisão, que permite veri…car a distribuição geográ…ca das habilitações no distrito, atendendo às pro…ssões dos indivíduos[12].

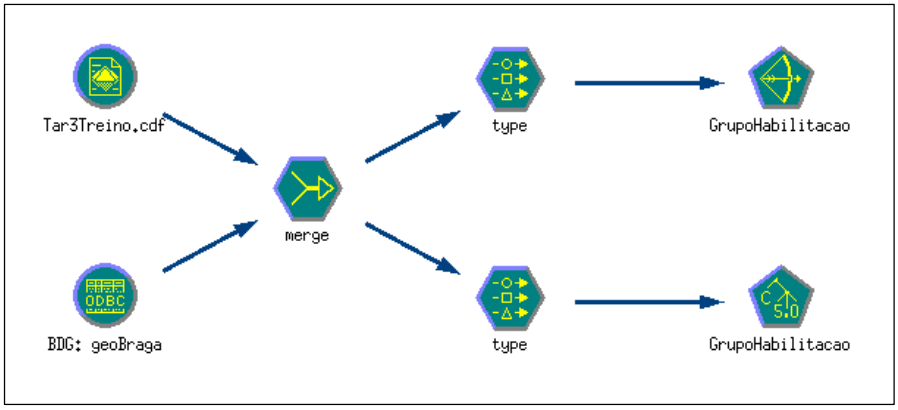


Figura 7.14: Identi…cação de regras de associação espacial com o algoritmo GRI

A Figura 7.15 evidencia as regras de associação espacial identi…cadas, e ainda, a árvore de decisão que permite identi…car a distribuição geográ…ca das habilitações, atendendo às pro…ssões dos indivíduos.

A detecção de tendências espaciais, no factor SIGAVE, tem como objectivo identi…car alterações regulares nos graus de lesão veri…cados 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 geográ…ca do distrito em análise, já disponível na BDG (geoBraga), com o conjunto de dados de treino. Este processo é evidenciado na Figura 7.16, na qual é ainda possível veri…car que o algoritmo C5.0 é utilizado na identi…cação de regras que descrevem tendências nos dados. No caso particular do código V (visão), veri…ca-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,

[11] O Clementine disponibiliza outro algoritmo de indução de regras de associação, o Apriori, que apenas permite manusear dados de entrada e de saída simbólicos.

[12] Destaca-se que o exercício inverso foi também efectuado, permitindo a identi…cação de um conjunto de regras nas quais eram previstas as pro…ssões, atendendo às habilitações dos indivíduos. No entanto, os modelos obtidos revelaram ser pouco capazes de prever a pro…ssão dada a habilitação, zona geográ…ca e ano de nascimento, já que a maioria dos indivíduos se encontram distribuídos por dois grupos de habilitação, escolaridade obrigatória e geral uni…cado 6-9. A con…ança dos modelos encontrados não foi superior aos 20%.

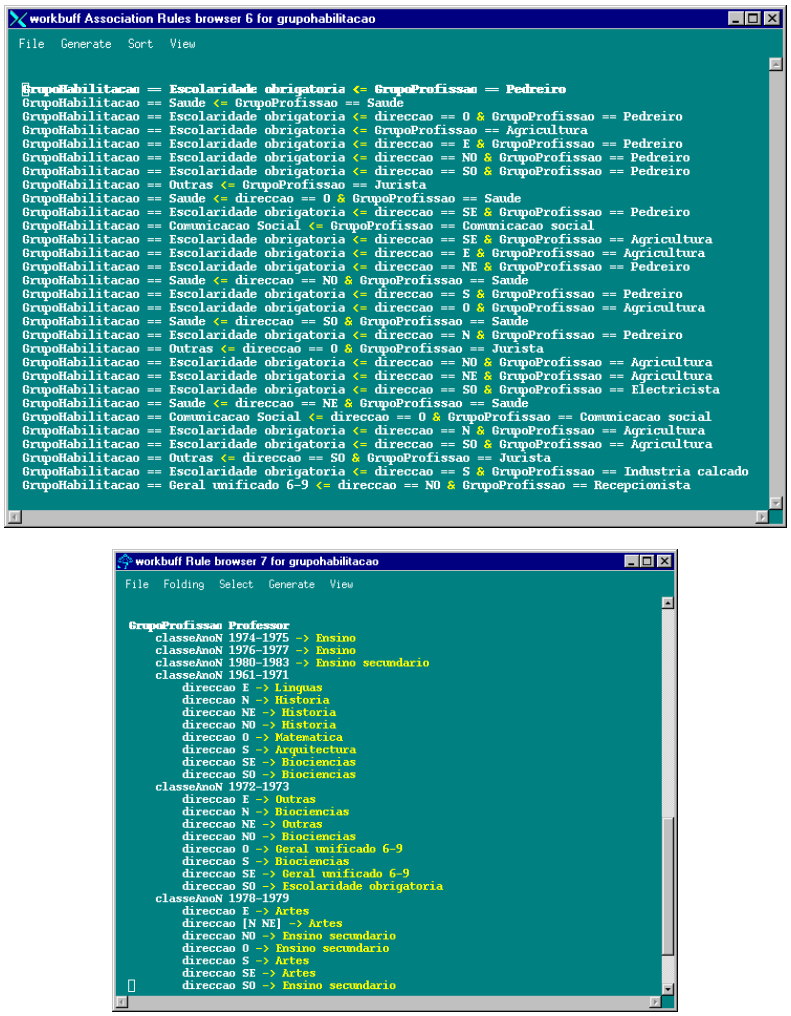


Figura 7.15: Regras de associação espacial e árvore de decisão, para a caracterização das habilitações literárias

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 identi…car 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 7.16. Posteriormente, no componente de Visualização de Resultados, serão visualizados gra…camente estes achados.

Estando já descrita a fase de DM, para cada uma das tarefas de…nidas, salienta-se que o tempo necessário à obtenção dos modelos depende da quantidade de dados a analisar, mas principalmente, da técnica seleccionada para a sua concretização. A Tabela 7.1 resume o tempo consumido na aprendizagem de quatro dos modelos apresentados anteriormente nesta subsecção. Pela análise da referida tabela constata-se que, o tempo necessário para o treino de árvores de decisão é consideravelmente inferior, ao tempo necessário à identi…cação de associações nos

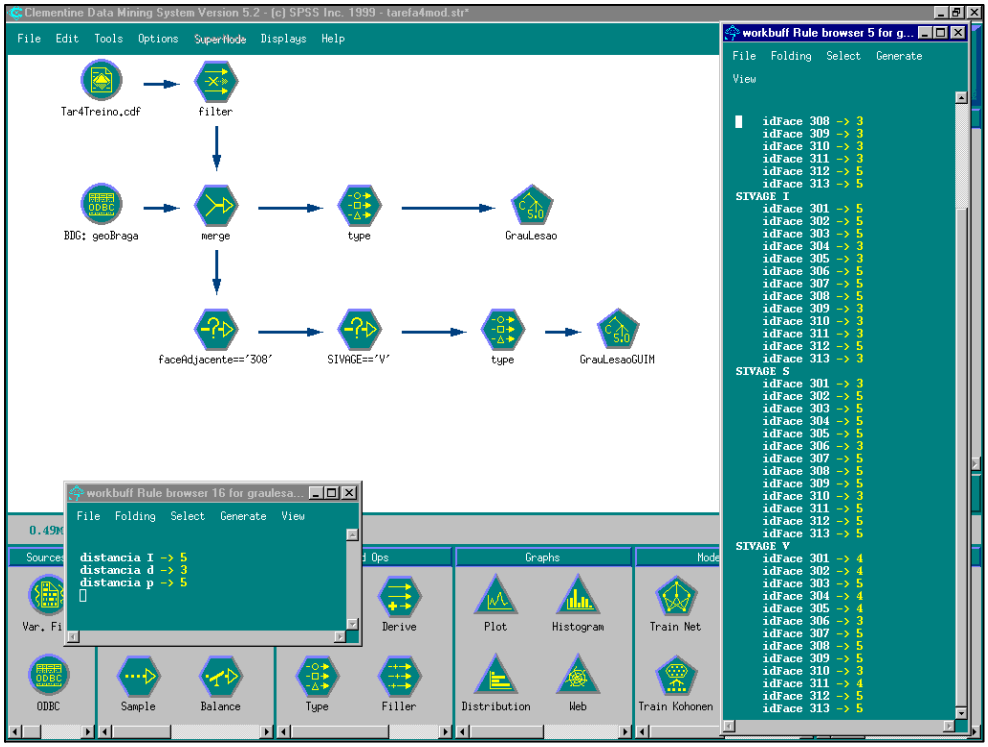


Figura 7.16: Detecção de tendências espaciais no factor SIVAGE

dados. Enquanto que as árvores de decisão se destinam essencialmente à classi…cação, sendo os seus modelos construídos seleccionando apenas os atributos (e valores) relevantes para tal, as regras de associação entre atributos são construídas a partir do processamento exaustivo dos dados, no qual todos os atributos seleccionados, e seus respectivos valores, são simultaneamente considerados na análise.

Modelo	Conjunto de treino	Tempo de aprendizagem
GrauConhecimento (C5.0)	120.767 registos	6 minutos, 28 segundos
GrupoHabilitação (GRI)	13.309 registos	20 minutos, 36 segundos
GrupoHabilitação (C5.0)		44 segundos
GrauLesão (C5.0)	12.253 registos	21 segundos

Tabela 7.1: Tempo de aprendizagem dos modelos

7.3.4 Interpretação de resultados

Os modelos GrauConhAle e GrauConhFran, construídos para a caracterização geográ…ca do conhecimento das línguas Alemão e Francês, respectivamente, são nesta fase utilizados para veri…car como é que os mesmos se comportam na classi…cação de um conjunto de dados desconhecido, nomeadamente o conjunto de dados de teste. A Figura 7.17 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 certi…cam 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%.

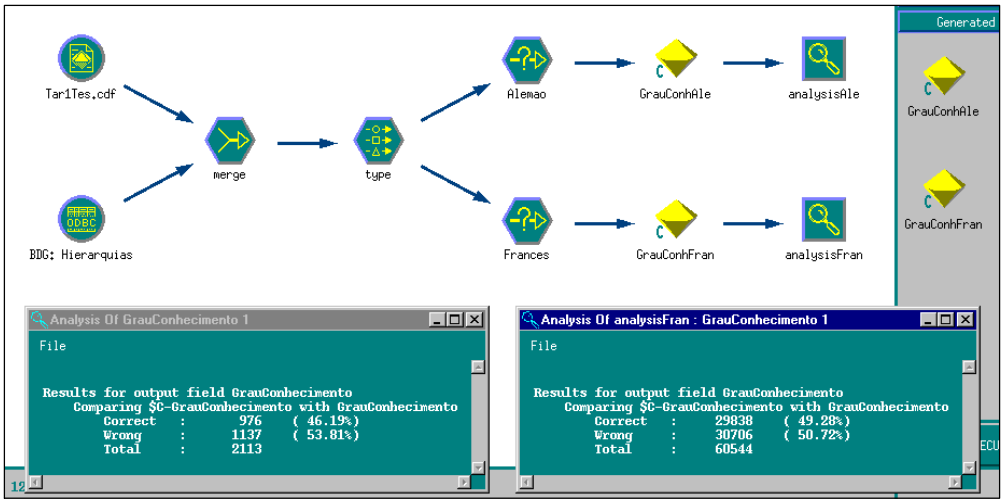


Figura 7.17: Análise dos modelos construídos para a caracterização do per…l linguístico

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 con…ança 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 avaliação do desempenho do modelo GrauConhecimento (Figura 7.18), que discrimina o grau de conhecimento de uma dada língua, em diferentes regiões geográ…cas, permitiu veri…car que a percentagem de acerto, na classi…cação do conjunto de dados de teste, é de 54.88%. Dada a quantidade de dados envolvida nesta tarefa, considera-se que a con…ança do modelo poderá ser melhorada, através do aumento do número de registos do conjunto de dados de treino.

A identi…cação do grupo de habilitação, a que cada indivíduo pertence, a partir do conhecimento da pro…ssão exercida, zona geográ…ca em que está inserido e do respectivo ano de nascimento, é efectuada considerando as regras explícitas na árvore de decisão, construída na tarefa de detecção de associações espaciais. A Figura 7.19 apresenta a stream construída para veri…car o grau de con…ança da referida árvore de decisão. O modelo em causa relevou, na classi…cação do conjunto de dados de teste, uma percentagem de acerto de 70.31% .

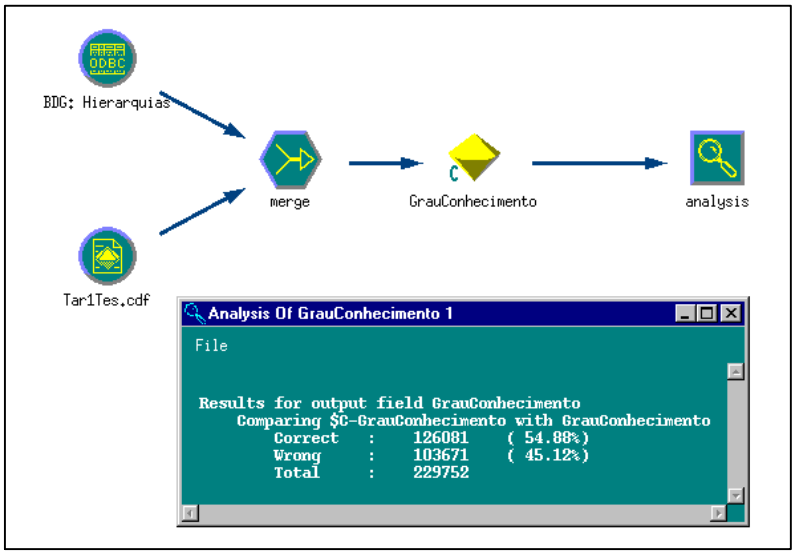


Figura 7.18: Avaliação do modelo GrauConhecimento

A árvore de decisão construída na satisfação da quarta tarefa, e que permitiu a identi…cação de tendências espaciais nos dados, foi utilizada para classi…car dados desconhecidos. Nesta avaliação (Figura 7.20), utilizando o conjunto de dados de teste, o modelo apresentou uma percentagem de acerto de 47.17%.

Os desempenhos apresentados nesta subsecção, e que sintetizam a con…ança dos modelos construídos na fase de DM, quando utilizados na classi…cação de dados desconhecidos, são in‡uenciados pela distribuição pouco homogénea dos dados analisados. Esta distribuição di…culta o processo de aprendizagem, e pode condicionar a utilização dos modelos obtidos em tarefas de previsão.

7.3.5 Visualização de resultados

No componente de Visualização de Resultados[13], que integra o sistema Padrão, é possível armazenar, na BDP, os modelos encontrados na fase de DM. Este procedimento permite a visualização das regras em mapas das regiões analisadas.

A Figura 7.21 apresenta o processo de armazenamento das regras, que descrevem a caracterização geográ…ca do conhecimento do Alemão. Na stream apresentada é possível veri…car 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 identi…ca a tabela da BDP que armazena as regras a visualizar.

[13] Este componente é descrito apenas para a primeira e quarta tarefa, por se considerar que a sua apresentação constitui uma repetição, já que o processo de armazenamento das regras, e a sua posterior visualização em mapas, é efectuado sempre da mesma forma.

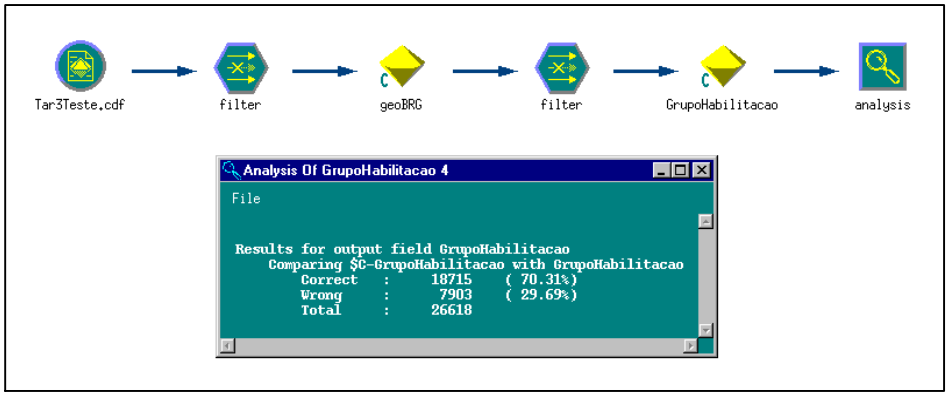


Figura 7.19: Desempenho da árvore de decisão, na identi…cação das habilitações dos indivíduos

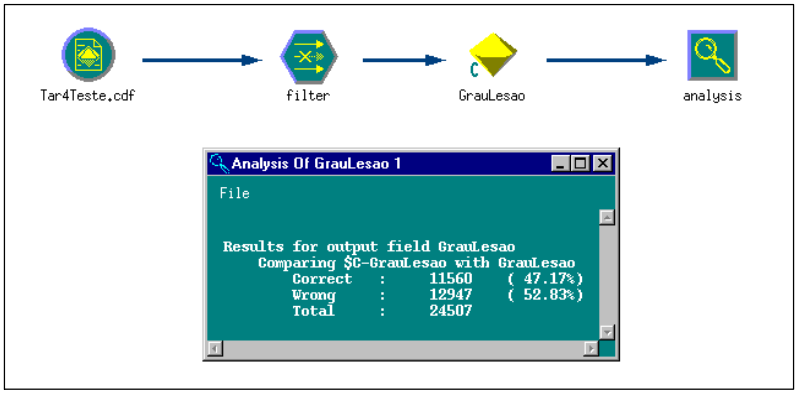


Figura 7.20: Desempenho da árvore de decisão que explicita tendências espaciais nos dados

A utilização do VisualPadrão na visualização grá…ca dos modelos GrauConhAle e GrauConhFran, conduziu à construção de dois mapas (Figura 7.22, 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 veri…car 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, veri…cam um grau de conhecimento menor na outra língua analisada. Os distritos de Lisboa e de Castelo Branco representam as duas excepções desta veri…cação.

A detecção de tendências espaciais no factor SIVAGE, explícita no modelo GrauLesão construído anteriormente, é nesta subsecção apresentada recorrendo à cartogra…a da região analisada. A Figura 7.23 apresenta a stream que permitiu a transferência das regras para a BDP, nomeadamente para a tabela TendEsp. A Figura 7.24 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.

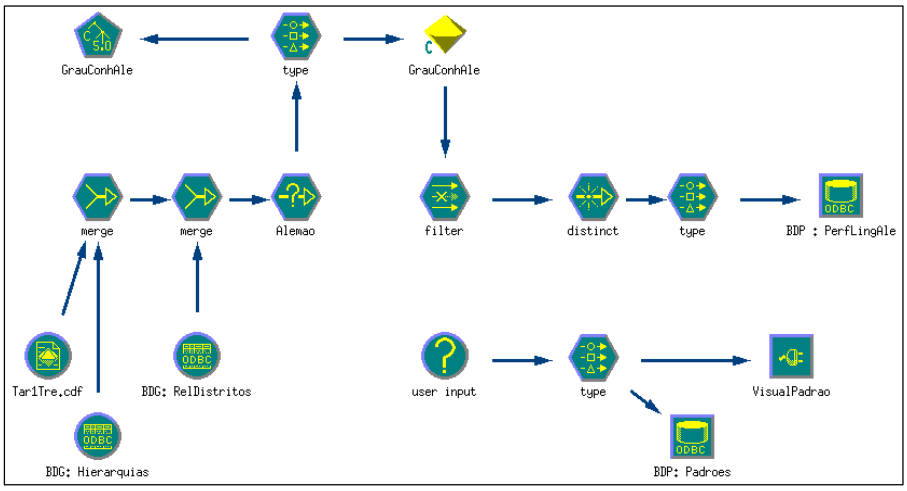


Figura 7.21: Transferência das regras que caracterizam o per…l linguístico para a BDP

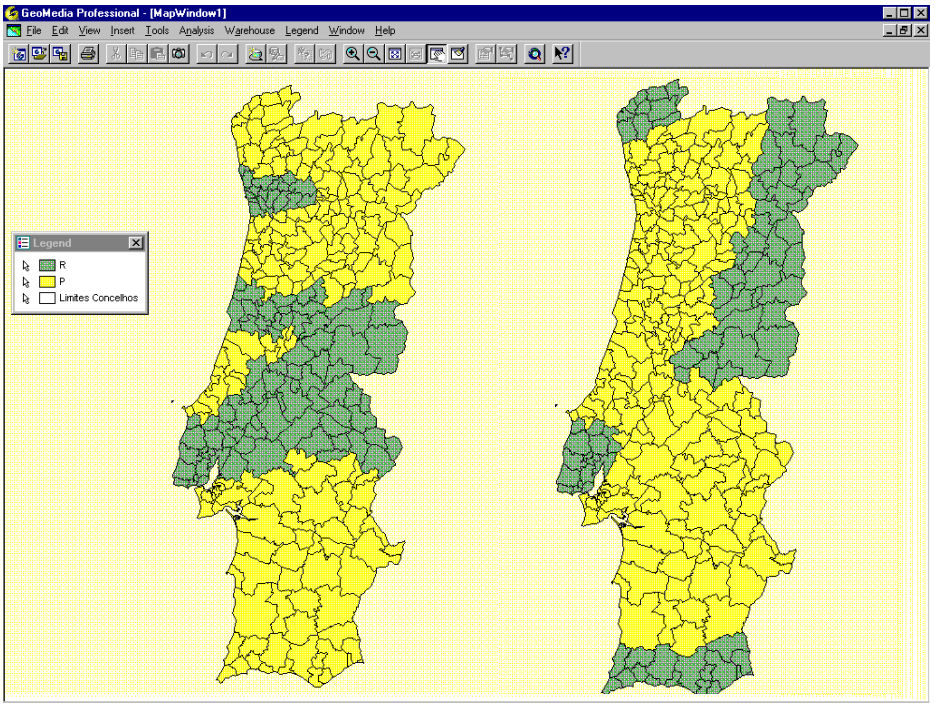


Figura 7.22: Mapas com a caracterização geográ…ca do conhecimento do Alemão e do Francês

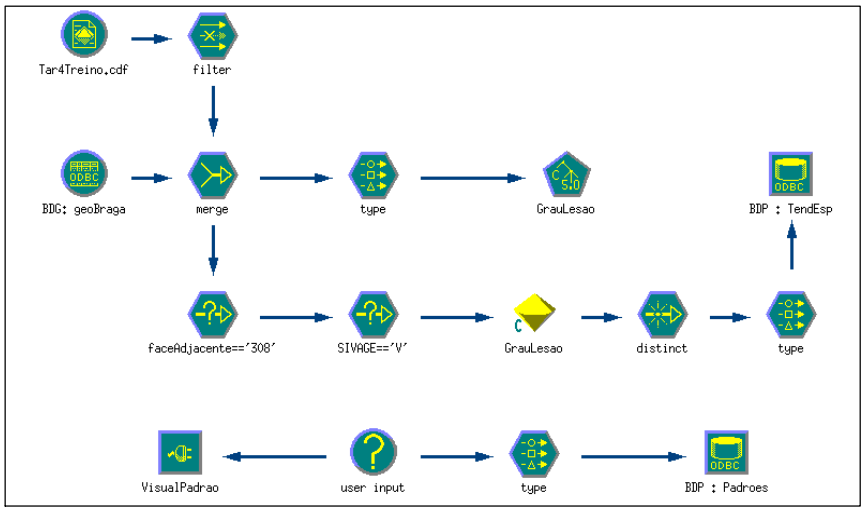


Figura 7.23: Transferência das regras que explicitam as tendências espaciais para a BDP

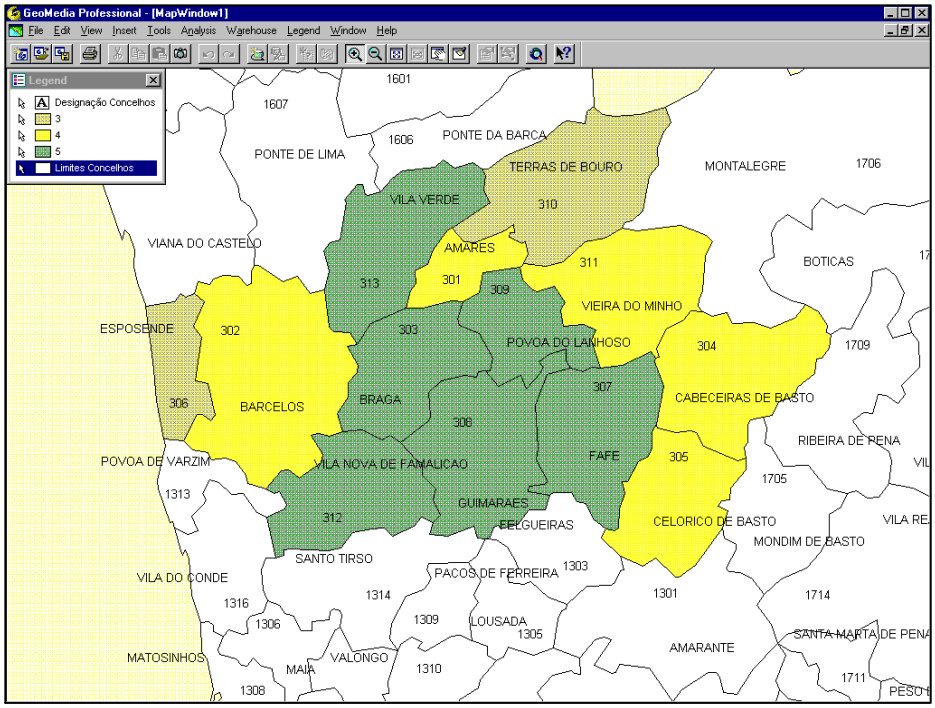


Figura 7.24: Tendências espaciais no código V do factor SIVAGE

7.4 Di…culdades encontradas

Após o processo de análise de uma BD real de grande dimensão, constata-se que as di…culdades encontradas, ao longo deste estudo de caso, estão relacionadas com o conteúdo dos dados analisados, e não com o processo de descoberta de conhecimento. Foi possível veri…car que apesar da quantidade de dados envolvida, não foram detectados problemas metodológicos, uma vez que as tarefas a executar em cada uma das fases estão, à partida, de…nidas.

Este estudo de caso comprovou que as maiores di…culdades que podem ser encontradas no processo de descoberta de conhecimento, derivam do facto dos dados terem sido armazenados com o objectivo de suportar processos operacionais, associados ao funcionamento da organização, e não com o objectivo de descoberta de conhecimento. O facto de não existirem normas quanto à informação que pode ser armazenada em determinado atributo, e sua respectiva validação, permite que valores com diferente interpretação semântica, possam ser atribuídos a um mesmo atributo.

Neste estudo de caso, na tabela Habilitações, o atributo HabLiterária foi utilizado em alguns casos para indicar o nível da habilitação obtida pelo indivíduo (nível básico, secundário,...), e noutros, para referir uma área especí…ca de conhecimento. Esta dualidade semântica obriga a uma cuidadosa de…nição e utilização dos atributos no processo de descoberta de conhecimento, cujos resultados devem ser interpretados atendendo sempre a esta dupla de…nição.

A documentação dos dados armazenados na BD, ou seja, a de…nição dos seus metadados, deverá receber mais atenção por parte das organizações, já que a falta dos mesmos conduz a situações caricatas, como as veri…cadas neste estudo de caso, no qual não era conhecida a descodi…cação dos valores que o atributo VinProfissional pode tomar.

Salienta-se ainda que, e no caso de atributos com grande diversidade de valores possíveis, como DesigHabLiterária e DesigProfissão, os valores para estes atributos deverão ser de…nidos de forma a minimizar a redundância, e facilitar o processo de construção de hierarquias conceptuais para os mesmos.

Capítulo 8

Conclusões

Este capítulo culmina a descrição do trabalho realizado ao longo deste projecto, apresentando uma síntese dos resultados e contribuições obtidas, atendendo à …nalidade e aos objectivos inicialmente impostos ao mesmo. Este capítulo inclui a identi…cação de projectos de trabalho futuro, cujo objectivo é dar continuidade ao trabalho aqui iniciado. Finalmente, são tecidas algumas considerações …nais sobre este projecto.

8.1 Síntese

Um caso particular da DCBD diz respeito à exploração de dados referenciados espacialmente, isto é, dados que incluem referências a objectos geográ…cos, localizações, ou partes de uma divisão territorial. A análise destes dados impõe a veri…cação da componente espacial associada aos mesmos (posições relativas, adjacências, direcções, distâncias, etc.), e da sua in‡uência nos restantes dados explorados, já que um objecto geográ…co pode ser afectado por acontecimentos veri…cados em objectos vizinhos.

A análise de dados espaciais com o objectivo de descoberta de conhecimento requer a utilização de técnicas especí…cas, capazes de incluir a semântica espacial, implícita na posição e dimensão dos objectos geográ…cos referenciados, no referido processo. Até ao momento, estas técnicas têm visado o desenvolvimento de novos algoritmos de DM (ou adaptação de algoritmos já existentes), e a integração de SGBDE, ou SIG, com ferramentas de descoberta de conhecimento. Estes últimos permitem a manipulação dos dados espaciais, e consequente transferência dos resultados para análise, na ferramenta de descoberta de conhecimento, com os restantes dados não espaciais explorados.

Estas abordagens requerem a descrição geométrica dos diversos objectos geográ…cos referenciados, uma vez que se baseiam em estratégias quantitativas de raciocínio espacial, que manipulam as coordenadas de pontos que descrevem as diversas entidades geográ…cas.

Constatando-se que em inúmeras BD organizacionais são utilizados identi…cadores geográ-…cos qualitativos, como moradas, na geo-referenciação da informação, a utilização de sistemas de posicionamento indirecto suprime a necessidade de desenvolvimento (ou adaptação) de algoritmos de DM, evitando, também, a utilização de SIG ou SGBDE para manipulação dos dados espaciais, e consequentemente, a obrigatoriedade de caracterização geométrica das entidades

geográficas.

Para que o raciocínio espacial fosse possível, isto é, para que a semântica espacial fosse efectivamente integrada no processo de descoberta de conhecimento, utilizaram-se os princípios associados ao raciocínio espacial qualitativo, os quais permitiram raciocinar com informação geográfica incompleta ou imprecisa.

Foi definida como finalidade deste trabalho, a concepção, implementação e validação de um sistema de descoberta de conhecimento em BD geo-referenciadas. Como características, o sistema deveria utilizar mecanismos indirectos de referenciação geográfica, e basear-se em estratégias de raciocínio espacial qualitativo, que permitam inferir informação geográfica desconhecida.

O Padrão, designação adoptada para o sistema proposto, não incluiu o desenvolvimento de novos algoritmos de DM adaptados à componente espacial dos dados, visando essencialmente o aproveitamento das capacidades de análise exploratória de dados conseguidas até ao momento pelas ferramentas de DCBD disponíveis no mercado, nomeadamente do Clementine, a ferramenta de descoberta de conhecimento adoptada para a sua implementação.

A utilização de ferramentas de descoberta de conhecimento já disponíveis, expande a quantidade de técnicas de DM que podem ser utilizadas para analisar os dados, neste caso, dados geo-espaciais e dados não geográficos. Os princípios defendidos neste trabalho, e nos quais é baseada a concepção do sistema Padrão, representam uma nova abordagem à análise de dados referenciados espacialmente, com o objectivo de descoberta de conhecimento.

Atendendo à finalidade deste trabalho, foi possível formular um conjunto de sete objectivos a atingir, cuja satisfação permitiu a obtenção dos resultados e contributos esperados para este projecto. As próximas subsecções sintetizam os diversos objectivos, realçando os resultados e contributos conseguidos com a realização de cada um dos mesmos.

8.1.1 O domínio geo-espacial

A proposta do sistema Padrão assenta na constatação de que a componente espacial associada aos dados geo-referenciados não é incorporada no processo de descoberta de conhecimento, já que os algoritmos tradicionalmente utilizados para explorar os dados, não incluem mecanismos que lhes permitam raciocinar em termos espaciais. Tal não permite, por exemplo, verificar a influência que as entidades geográficas exercem umas nas outras.

O domínio geográfico/espacial representa uma peça fundamental neste trabalho, disponibilizando conceitos e princípios utilizados ao longo do mesmo. Tal justifica que o primeiro objectivo a alcançar estivesse associado à revisão teórica/bibliográfica dos diversos conceitos associados a este domínio e utilizados neste projecto.

O enquadramento teórico realizado ao domínio geo-espacial visou essencialmente três áreas:

- o enquadramento conceptual do domínio. A este nível clarificaram-se alguns conceitos e definiu-se o significado do termo geo-espacial adoptado neste trabalho. No que diz respeito à tecnologia de BD espaciais, apresentaram-se:
  - os tipos de dados espaciais, nomeadamente o ponto, a linha e o polígono. Para

colecções de objectos geográficos, definiu-se a partição e a rede, representando um conjunto de regiões não sobrepostas e um grafo embebido num plano, respectivamente.

- os tipos de representação de dados espaciais, nomeadamente a representação baseada em células (modelo raster) e a representação baseada em objectos (modelo vectorial).
- linguagens de manipulação de dados espaciais, destacando extensões do SQL que permitem manusear dados espaciais. Estas extensões acrescentam à tradicional sintaxe `select ... from ... where`, operadores e funções espaciais.
- mecanismos de integração de dados espaciais e dados não espaciais, através da arquitectura SAND, na qual dois grupos de apontadores, entre dados espaciais e dados não espaciais e entre dados não espaciais e dados espaciais, facilitam a navegação e permitem que as pesquisas possam ser iniciadas a partir de qualquer um dos conjuntos de dados.
- os SIG, descrevendo os seus principais componentes e ainda, a sua capacidade de análise espacial.

- as técnicas de modelação de informação geográfica. A modelação de dados engloba um conjunto de actividades que conduz ao desenho da BD. Ao nível das técnicas de modelação para dados geográficos, descreveu-se

  - uma extensão do modelo E-R para dados geográficos, na qual objectos espaciais são representados por entidades, cujas propriedades estruturais e topológicas com outros objectos são representadas por relacionamentos.
  - o modelo geo-relacional, que estende o modelo relacional por forma a permitir modelar dados espaciais. Esta extensão incorpora no modelo componentes como níveis, relações, níveis virtuais, classes de objectos e restrições de integridade, que permitem a definição de modelos lógicos dos dados, assim como vistas sobre os mesmos.
  - uma abordagem formal para a modelação lógica de aplicações geográficas. Nesta abordagem, predicados em lógica de 1ª ordem permitem a especificação dos requisitos de informação, enquanto que predicados em lógica de 2ª ordem são utilizados para definir as regras que incorporam no sistema, mecanismos de raciocínio espacial e temporal.

- a normalização na área da informação geográfica. A normalização tem como objectivo permitir que a informação geográfica possa ser acedida por diferentes utilizadores, aplicações e sistemas, em diferentes localizações. Neste âmbito, um conjunto estruturado de regras permitem a definição, descrição, estruturação, pesquisa, alteração e transferência de informação geográfica, e ainda da sua meta-informação. Os principais grupos de trabalho incluem o CEN TC 287 e o ISO TC 211. Os trabalhos do CEN iniciaram-se em Outubro de 1991, enquanto que no caso do ISO, a comissão técnica apenas foi constituída em Abril de 1994. Estas duas comissões partilham membros, e através de um acordo de cooperação visam garantir a harmonização das normas produzidas e evitar a duplicação de trabalho nos dois grupos. Neste projecto adoptaram-se as pré-normas produzidas pelo CEN TC 287, das quais se utilizaram as directivas do esquema espacial e do esquema de identificadores geográficos, na construção da BDG que integra o componente Repositório de Dados e Conhecimento do sistema Padrão.

Em resumo, além do enquadramento teórico/bibliográ…co resultante da execução deste objectivo, apresenta-se como uma das contribuições deste trabalho, a utilização das pré-normas europeias para informação geográ…ca. A compreensão, interpretação e utilização das regras explícitas nos documentos produzidos pelo grupo de trabalho do CEN TC 287, constitui uma mais valia inquestionável, para além de contribuir para a construção de repositórios normalizados de informação geográ…ca.

8.1.2 O raciocínio espacial qualitativo

A utilização no Padrão de mecanismos de referenciação indirectos (recorrendo a identi…cadores geográ…cos) e relações espaciais qualitativas (que permitem a de…nição das relações espaciais existentes entre as entidades geo-referenciadas), conduz inevitavelmente à utilização de estratégias de raciocínio espacial qualitativo, que permitam raciocinar com informação geográ…ca incompleta ou imprecisa.

O segundo objectivo a atingir visou a revisão dos fundamentos teóricos subjacentes ao raciocínio espacial qualitativo. O enquadramento conceptual resultante permitiu:

- descrever mecanismos de representação de conhecimento espacial qualitativo. Neste trabalho adoptaram-se, como formalismo para a representação do conhecimento espacial, predicados da forma A Norte B, em que A e B representam entidades geográ…cas, e Norte descreve a relação espacial existente entre as mesmas (formalmente, objectoPrimário [relEspacial] objectoReferência).
- sintetizar os fundamentos associados ao raciocínio temporal qualitativo, uma vez que os mesmos foram amplamente adaptados ao domínio espacial.
- de…nir os tipos de relações espaciais abordadas neste trabalho, nomeadamente as relações espaciais do tipo direcção, distância e topologia. Para cada uma destas relações, apresentaram-se os identi…cadores qualitativos adoptados e ainda, os mecanismos que permitem a construção das tabelas de composição, utilizadas na inferência de informação geográ…ca desconhecida.
- apresentar duas abordagens de raciocínio espacial integrado, caracterizadas por permitirem raciocinar, simultaneamente, com mais do que um tipo de relação espacial. Descreveu-se a integração da direcção e distância proposta por Hong [Hong, 1994], e a integração da direcção e topologia proposta por Sharma [Sharma, 1996]. Para estes sistemas, analisaram-se detalhadamente os princípios que permitiram a construção das regras de inferência, explícitas nas respectivas tabelas de composição.
- salientar a importância das características das entidades geo-referenciadas, nomeadamente do seu tamanho, e dos mecanismos de raciocínio qualitativo que podem ser utilizados na construção das regras que permitem a inferência de dimensões desconhecidas.

A compreensão dos princípios que ditaram a construção dos sistemas integrados referidos anteriormente, permitiu de…nir novas tabelas de composição para a integração da direcção e topologia, utilizando o sistema triangular na de…nição da direcção existente entre os objectos (e

não o sistema de projecções como utilizado por Sharma [Sharma, 1996]). Novas primitivas temporais foram de…nidas para a caracterização da direcção e topologia, cuja composição permitiu a obtenção de novas regras de inferência.

Esta alteração visou a compatibilização dos dois sistemas integrados, nomeadamente no que diz respeito à utilização do sistema triangular. Tal permitiu a construção de um sistema de inferências, que conjuga relações espaciais do tipo direcção, distância e topologia, no processo de raciocínio. Este sistema, utilizado pelo Padrão na inferência de informação geográ…ca desconhecida, representa mais um contributo deste trabalho, uma vez que para além de integrar os três tipos de relações espaciais, permitiu que o mesmo pudesse ser utilizado na composição de entidades geográ…cas com extensão (já que a integração da direcção e distância apresentada por Hong [Hong, 1994], está associada a objectos geográ…cos sem extensão, isto é, entidades geométricas do tipo ponto). Ainda em relação ao sistema de inferências, foi possível veri…car, através da avaliação efectuada ao desempenho do mesmo, que a dimensão das entidades geográ…cas assume um papel primordial na identi…cação da relação do tipo direcção. Tal permitiu, para os casos particulares dos grupos de direcções ¢dir1 e ¢dir3, a identi…cação de um conjunto de regras que consideram o tamanho das regiões no processo de raciocínio, integrando à direcção, distância e topologia, a dimensão das regiões envolvidas.

8.1.3 A descoberta de conhecimento em bases de dados

A DCBD representa uma outra área de conhecimento fundamental neste trabalho, e cuja revisão teórica/bibliográ…ca consta do terceiro objectivo inicialmente de…nido.

O enquadramento conceptual elaborado permitiu sistematizar os princípios associados a esta área de conhecimento, dos quais foi dada particular importância:

- às diversas fases do processo de descoberta de conhecimento, desde a selecção dos dados até a interpretação de resultados, e ainda, às características desejáveis para um sistema de descoberta de conhecimento. Nestas, é dada particular ênfase ao utilizador, como responsável pela condução do processo e pelas várias decisões tomadas ao longo do mesmo.
- ao conhecimento do domínio de aplicação, utilizado na condução do processo de descoberta de conhecimento, constituindo portanto, um recurso essencial ao mesmo.
- às di…culdades mais frequentes encontradas no processo de descoberta de conhecimento, como sejam, a informação insu…ciente e os dados corrompidos. No primeiro caso, englobam-se os casos de informação incompleta, dados dispersos ou amostras de tamanho reduzido, enquanto que o segundo caso está associado a dados com ruído ou com valores omissos.
- à procura de relacionamentos e padrões nos dados, apresentando as tarefas de DM (classi…cação, segmentação, associação, ...) e as diversas técnicas que podem ser utilizadas para a sua execução. As técnicas descritas foram agrupadas em quatro grandes grupos: a indução de regras (que inclui as árvores de decisão e as regras de associação), as redes neuronais, os algoritmos genéticos e a aproximação de vizinhanças.
- a um caso particular da DCBD, e que diz respeito à exploração de dados referenciados espacialmente. O processo de DCBDE desempenha um papel fundamental na percepção

das características associadas aos dados espaciais, e principalmente, dos relacionamentos implícitos que existem entre dados geo-espaciais e dados não espaciais. As tarefas tradicionalmente associadas a este processo incluem:

- a descrição de distribuições espaciais nos dados não espaciais: características espaciais;
- a comparação de distribuições espaciais dos dados não espaciais: análise espacial discriminante;
- o estabelecimento de relações entre dados espaciais, e entre dados espaciais e dados não espaciais: associação espacial; e
- a verificação de alterações regulares de um ou mais atributos não espaciais, associados a um dado objecto espacial: detecção de tendências espaciais.

Para a execução destas tarefas, diversos algoritmos tiveram de ser construídos ou adaptados, por forma a permitirem a análise da componente espacial associada aos dados explorados. Noutros casos, SGBDE ou SIG tiveram de ser integrados nos sistemas de descoberta de conhecimento, por forma a permitirem a manipulação de dados espaciais (já que estes sistemas disponibilizam estruturas apropriadas para o seu armazenamento, e ainda, funções espaciais para a sua análise).

A descrição e sistematização das diversas iniciativas em curso na área da DCBDE, permitiu constatar que estas podem ser agrupadas em dois grandes grupos. O primeiro caracteriza-se pela implementação de novos algoritmos de DM ou adaptação de algoritmos existentes, por forma a permitirem a análise de dados espaciais. No segundo grupo, encontram-se as propostas de ambientes integrados, nos quais SIG ou SGBDE suportam a análise da componente espacial dos dados. Em ambos, o processo de descoberta de conhecimento é iniciado num dos conjuntos de dados, dados espaciais ou dados não espaciais, sendo sempre solicitada uma questão ao utilizador, que guie o processo de pesquisa. Salienta-se que, para diversas das aproximações integradas descritas neste documento, não se conhecem os algoritmos utilizados, ou os resultados que é possível obter com os mesmos, por estas propostas se encontrarem ainda na fase de arquitectura.

A abordagem utilizada no Padrão diferencia-se das aproximações atrás mencionadas, não só por não incluir a implementação de novos algoritmos de DM, ou adaptação de algoritmos já existentes, mas também por permitir que dados não espaciais e dados geo-espaciais possam ser analisados simultaneamente pelos algoritmos de DM (não condicionando a procura), e porque a utilização de um SIG ou um SGBDE não é requerida no processo de descoberta de conhecimento. Recorda-se que o Padrão apenas recorre a um SIG para a visualização de resultados, e apenas quando requerido pelo utilizador. Este facto revela-se de particular importância no caso de não se possuir a descrição geométrica das entidades geográficas referenciadas, como é necessário sempre que se adopta uma abordagem quantitativa ao raciocínio espacial.

8.1.4 Concepção, implementação e validação do sistema Padrão

A concepção, implementação e validação do sistema Padrão, representam respectivamente, o quarto objectivo, o quinto objectivo e o sexto objectivo definidos neste trabalho. Ao nível da concepção, destaca-se a definição da arquitectura do sistema, a qual é baseada em mecanismos

indirectos de referenciação da informação, os quais são complementados com estratégias qualitativas de raciocínio espacial, que permitem a inferência de informação geo-espacial desconhecida, e necessária aos algoritmos de DM.

A arquitectura do Padrão agrega três componentes: Repositório de Dados e Conhecimento, Análise de Dados e Visualização de Resultados. O Repositório de Dados e Conhecimento é o responsável pelo armazenamento dos dados geo-espaciais e dos dados não geográficos utilizados pelo sistema, e ainda, das regras que permitem a implementação dos mecanismos de inferência utilizados no processo de raciocínio qualitativo. Este componente agrega três BD:

- uma BDG construída segundo os princípios estabelecidos pelo CEN TC 287, e que permitiram a referenciação indirecta da informação através da implementação de um sistema de identificadores geográficos. Para este sistema foi ainda possível definir as relações espaciais do tipo direcção e distância, existentes entre entidades geográficas adjacentes. A estrutura desta BD integra o esquema de identificadores geográficos e o esquema espacial, definidos nas pré-normas europeias para informação geográfica.
- uma BCE que armazena os mecanismos qualitativos de raciocínio que permitem a inferência de relações espaciais desconhecidas. Esta base de conhecimento agrega as tabelas de composição que integram a direcção, a distância e a topologia, o conjunto de identificadores qualitativos utilizados, e ainda, os intervalos de validade quantitativos de cada um dos identificadores associados à direcção e à distância.
- uma BDnG que depende do domínio de aplicação em análise.

O componente de Análise de Dados caracteriza-se por passar por 6 grandes etapas: selecção dos dados, tratamento dos dados, pré-processamento dos dados, processamento da informação geo-espacial, DM e interpretação de resultados. Destaca-se que às cinco fases que tradicionalmente constituem o processo de descoberta de conhecimento, apenas foi acrescentada a que permite verificar a informação geo-espacial disponível e inferir a informação desconhecida, necessária na etapa de DM. Tal permite que dados geo-espaciais e dados não geográficos possam ser analisados simultaneamente, aumentando a probabilidade de identificar padrões desconhecidos do utilizador.

O componente de Visualização de Resultados permite o armazenamento na BDP dos padrões considerados relevantes, e a sua posterior visualização em mapas das regiões analisadas. Os padrões armazenados são devidamente catalogados, permitindo a sua posterior utilização em exercícios de DM. Esta catalogação pode conduzir à construção de meta-regras, que evidenciam a evolução dos padrões encontrados ao longo do tempo.

A documentação das estruturas de dados utilizadas pelo Padrão, nomeadamente das BD que o integram, e do modo de funcionamento do sistema, foi conseguida utilizando o UML, mais precisamente, os diagramas de classes e os diagramas de caso de uso. Estes últimos revelaram-se de particular importância, na definição das interacções existentes entre os diversos actores e o sistema.

Em termos tecnológicos, o sistema Padrão foi implementado recorrendo:

- a SGBD relacionais, nomeadamente ao Microsoft Access, no qual foram construídas e/ou mantidas as diversas BD utilizadas pelo Padrão. Destaca-se que poderia ter sido

utilizado outro SGBD, desde que o mesmo disponibiliza-se ligações ODBC para acesso aos dados, uma vez que foi este o meio utilizado para aceder às diversas BD utilizadas. Para o carregamento automático de algumas das tabelas que constituem a BDG, foram construídos diversos módulos em VB, os quais facilitaram e agilizaram este processo. Estes módulos veri…cam, na cartogra…a digital utilizada, as relações espaciais existentes entre as regiões, armazenando os dados resultantes nas respectivas tabelas.

- ao Clementine, a ferramenta de DCBD adoptada, que permitiu a implementação no seu ambiente de trabalho, de todas as fases do componente de Análise de Dados do Padrão. Refere-se que foi necessário recorrer à implementação de um módulo externo em VB, para auxiliar o Clementine no processo de inferência da informação geo-espacial. Esta necessidade foi motivada pelo facto de, até à data de utilização neste trabalho, o Clementine não possuir mecanismos de manipulação de arrays.

- ao Geomedia Professional, o SIG utilizado na visualização de resultados. Uma vez que o Geomedia disponibiliza uma biblioteca de objectos geográ…cos, que podem ser manipulados em linguagens de programação como o VB, foi possível construir o VisualPadrão, uma aplicação que foi integrada no Clementine, passando a constituir um dos nodos disponíveis no seu ambiente de trabalho. Esta abordagem permite, após o armazenamento na BDP dos relacionamentos encontrados, a visualização dos padrões em mapas das regiões analisadas, sendo todo o processo realizado no ambiente de trabalho do Clementine. Uma legenda contextualiza os achados, permitindo a compreensão grá…ca dos mesmos.

A incorporação de módulos externos no Clementine permitiu construir um ambiente único de trabalho, no qual o utilizador executa todas as fases do processo de descoberta de conhecimento consideradas no Padrão.

Chegada a fase de validação do sistema, foi necessário avaliar o seu desempenho em duas áreas distintas. A primeira, relacionada com a componente geo-espacial, tinha como objectivo veri…car o desempenho do sistema de inferências, avaliando a qualidade das inferências obtidas com o mesmo. Para tal, veri…cou-se o desvio ocorrido nas várias iterações do processo de inferência, para uma dada região geográ…ca. Refere-se que o resultado obtido por inferência foi comparado com o valor real, dado por um módulo[1] em VB elaborado para o efeito. A este nível foi possível veri…car que:

- a enorme discrepância existente entre o tamanho das entidades que integram uma dada região geográ…ca, di…culta o processo de raciocínio. Este foi melhorado integrando, no processo de raciocínio, a dimensão das entidades referenciadas. Este procedimento permitiu constatar que, para o caso da direcção e distância, sempre que é inferida uma relação que não é efectivamente a direcção ou distância real dada pelo centróide dos objectos, esta é em todos os casos uma relação que é vizinha da direcção ou distância real. A maioria destas ocorrências dizem respeito a regiões com partes do território em mais do que uma área de aceitação do sistema triangular de…nido, o que permite, tendo-se assumido uma abordagem qualitativa ao raciocínio, adoptar os resultados obtidos por inferência, e catalogá-los como válidos para os objectivos que os mesmos visam servir neste trabalho.

[1] Este módulo permitiu determinar a localização, em termos de direcção, distância e topologia, consultando directamente a cartogra…a da região, disponível no SIG utilizado.

- sempre que a região em análise integrar entidades geográficas de dimensão homogénea, isto é, sem diferenças muito significativas entre as mesmas, o sistema de inferências apresenta um comportamento bastante estável, sendo inferior a 10% a quantidade de inferências na relação vizinha.
- no caso da topologia, e sempre que os intervalos de validade quantitativos para a distância sejam adequadamente definidos, a inferência desta relação é em todos os casos correcta, não catalogando relações deslocadas como adjacentes.

Além da avaliação à componente geo-espacial, era necessário verificar se a ferramenta de DCBD adoptada para a implementação do Padrão, o Clementine, efectivamente detectava padrões implícitos nos dados analisados. Para tal, foi seleccionada e analisada uma amostra de validação, com o objectivo de verificar se as regras implícitas na mesma, eram efectivamente identificadas. Salienta-se que estas amostras, amplamente divulgadas e utilizadas na validação de novas técnicas ou novos algoritmos de DM, são construídas a partir das regras, pelo que as mesmas passam a estar implícitas nos dados produzidos. A amostra seleccionada permitiu avaliar o Clementine no processo de descoberta de conhecimento, e confirmar a sua utilização no âmbito do sistema Padrão.

A validação mais esperada ao sistema, proveniente da sua utilização em casos reais e consequentemente de interesse para as organizações, provém da análise de uma componente do SIAPE. Este estudo de caso confirmou a utilidade do sistema Padrão na exploração de BD reais de grande dimensão. A análise desta BD, e atendendo aos objectivos definidos para a fase de DM, permitiu realizar as tarefas tradicionalmente associadas ao DME: identificar características espaciais, efectuar análise espacial discriminante, estabelecer regras de associação espacial e detectar tendências espaciais nos dados. A utilização do Padrão neste estudo de caso culmina a validação efectuada ao sistema, confirmando a satisfação da finalidade deste trabalho, a concepção, implementação e validação do sistema Padrão, e a realização do principal contributo desta tese de doutoramento.

8.1.5 Projectos de trabalho futuro

O sétimo e último objectivo associado a este projecto está relacionado com a formulação e proposta de projectos de trabalho futuro, que visem a evolução do sistema Padrão, e que promovam a sua utilização.

O trabalho apresentado neste documento descreve as opções estruturais e tecnológicas que permitiram chegar ao sistema Padrão. A sua validação permitiu identificar um conjunto de situações que podem ser optimizadas, com vista ao aumento do desempenho global do sistema. As situações identificadas estão associadas com opções estruturais e com opções tecnológicas. Ao nível estrutural, estas visam essencialmente o aumento do desempenho do sistema qualitativo de inferências. Ao nível tecnológico, pretende-se avaliar as vantagens e desvantagens da adopção de um paradigma lógico de programação, na implementação do sistema Padrão.

A abordagem qualitativa utilizada pelo Padrão, na qual o conhecimento espacial necessário ao processo de raciocínio é incorporado através de regras, sugere que os princípios subjacentes à programação lógica indutiva sejam explorados, com o objectivo de avaliar as vantagens e as

desvantagens que podem decorrer da sua utilização. Refere-se que esta técnica será incluída em futuras versões do Clementine [Khabaza e Brewer, 2000].

Para o aumento do desempenho do sistema qualitativo de inferências, sugere-se:

- a incorporação da dimensão das regiões no processo de raciocínio. Para os grupos mais críticos, ¢dir1 e ¢dir3, esta característica foi incluída no processo de raciocínio[2], e as melhorias obtidas sugerem a sua integração em todas as composições, por forma a aumentar o desempenho global do sistema.
- a optimização da integração da direcção e topologia. Esta optimização pode ser obtida adoptando primitivas temporais, que evidenciem a dimensão das regiões envolvidas na composição. A caracterização da relação integrada (direcção, topologia) existente entre A e B, recorrendo a primitivas temporais, pode implicitamente referenciar o tamanho das entidades. Por exemplo, as primitivas temporais (contém, posterior) indicam que a dimensão de A é superior à dimensão de B, enquanto que as primitivas (contido, posterior) traduzem que a dimensão de B é superior à de A. Ambas permitem caracterizar o par integrado direcção, topologia (Norte, deslocado).
- a utilização de um cone com 16 áreas de aceitação para o sistema triangular, o qual permitirá a de…nição de intervalos de validade de dimensão inferior, e consequentemente, a relações de direcção mais especí…cas. Mais uma vez, esta sugestão conduz, ainda que implicitamente, à introdução da dimensão das regiões no processo de raciocínio.

A optimização do sistema qualitativo de inferências passa essencialmente, conforme sugestões acima referidas, pela introdução da dimensão das regiões no processo de raciocínio, já que esta característica provou in‡uenciar determinantemente a qualidade dos resultados. Esta realidade é ainda mais evidente no caso das entidades geográ…cas referenciarem subdivisões administrativas, nas quais a discrepância existente entre a dimensão das diversas regiões é uma constante.

O componente de Visualização de Resultados do Padrão constitui um peça importante do sistema, uma vez que permite armazenar e visualizar os padrões encontrados no processo de descoberta de conhecimento. Neste componente, o VisualPadrão possibilita a visualização de resultados em mapas das regiões analisadas. Considera-se que seria útil expandir este componente, por forma a permitir a gestão da BDP, nomeadamente a sua exploração, com vista à veri…cação de evoluções nas regras obtidas em anteriores exercícios de DM. Tal permitiria ao utilizador identi…car tendências nos dados, que podem posteriormente ser con…rmadas em exercícios de DM.

8.2 Considerações …nais

Actualmente, as instituições produzem e armazenam grandes quantidades de dados, resultantes da sua actividade diária. O facto destes dados re‡ectir o comportamento e evolução das organizações ao longo do tempo, chama a atenção para os benefícios que podem decorrer da

[2] Conforme validação apresentada no Capítulo 6.

compreensão do conhecimento implícito nos mesmos, e da sua consequente utilização na tomada de decisão.

A investigação na área da DCBD 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, cujos principais desenvolvimentos estão associados à implementação de novos algoritmos de DM, ou à adaptação de algoritmos já existentes, capazes de incluir a semântica associada à localização dos factos, no processo de descoberta de conhecimento.

A concepção, implementação e validação do Padrão, permitiu a construção de um sistema de descoberta de conhecimento para BD geo-referenciadas, baseado em mecanismos de posicionamento indirecto e estratégias de raciocínio espacial qualitativo. Os princípios estabelecidos para o 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, referenciação espacial através de identi…cadores geográ…cos, evita ainda, a necessidade de de…nição geométrica das entidades geográ…cas referenciadas. Esta de…nição geométrica é requerida em abordagens quantitativas, nas quais o posicionamento directo dos objectos é especi…cado através de coordenadas de pontos, que indicam a localização espacial dos mesmos.

A abordagem qualitativa à referenciação espacial, considerada pelo sistema Padrão, permite que as BD organizacionais sejam analisadas sobre uma perspectiva espacial, independentemente da disponibilidade da geometria das diversas entidades geográ…cas 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 que permitem a inferência de informação geográ…ca desconhecida. Esta aproximação possibilita a explicitação em termos qualitativos, dos resultados do processo de descoberta de conhecimento.

A concepção[3] do sistema Padrão foi iniciada com a de…nição da arquitectura do sistema, a qual integra três componentes principais: 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 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 e dados não geográ…cos sejam analisados simultaneamente, possibilitando a identi…cação de padrões implícitos nos mesmos.

Os resultados do processo de descoberta de conhecimento podem ser armazenados na BDP, que integra o componente de Visualização de Resultados. Este componente permite que os padrões catalogados na BDP sejam visualizados em mapas das regiões analisadas, facilitando o processo de interpretação das regras encontradas.

Em termos tecnológicos, as diversas BD que integram o sistema Padrão foram geridas recorrendo ao Microsoft Access, no qual o acesso aos dados foi conseguido através de ligações ODBC. A ferramenta de DCBD adoptada para a implementação do componente de Análise de

[3] O sistema Padrão foi desenvolvido através de uma abordagem incremental, que permitiu adicionar ao sistema novas facilidades e funcionalidades, cuja incorporação era realizada à medida que as mesmas eram identi…cadas.

Dados foi o Clementine, a qual permitiu a realização, no seu ambiente de trabalho, de todas as fases do processo de descoberta de conhecimento. O SIG Geomedia Professional foi utilizado na implementação do VisualPadrão, uma aplicação desenvolvida em VB, que após integração no Clementine, permite a visualização de resultados em mapas das regiões analisadas.

Além do VisualPadrão, foi implementado outro módulo externo em VB, o qual permitiu a construção de um ambiente único de trabalho, inserido no Clementine, no qual o utilizador executa todas as fases do processo de descoberta de conhecimento.

No que diz respeito à validação do sistema Padrão, foi avaliada a qualidade das inferências obtidas com o sistema de raciocínio qualitativo implementado, e que integra relações espaciais do tipo direcção, distância e topologia, e ainda, con…rmada a sua capacidade de identi…cação de relacionamentos implícitos nos dados.

A validação técnica efectuada possibilitou a utilização do sistema Padrão na análise de uma BD real de grande dimensão, permitindo constatar a sua utilidade na análise de BD geo-referenciadas. Este estudo de caso evidenciou a identi…cação de relacionamentos existentes entre dados geo-espaciais e dados não geográ…cos, nomeadamente a identi…cação de características espaciais nos dados, a elaboração de análise espacial discriminante, o estabelecimento de regras de associação espacial e a detecção de tendências espaciais nos dados.

Con…rmada a utilidade do sistema Padrão na análise de BD geo-referenciadas, resta referir que o Padrão constitui o principal contributo desta tese de doutoramento. Os seus princípios representam uma nova abordagem na análise de dados geo-referenciados, com o objectivo de descoberta de conhecimento. Além da inovação, esta nova abordagem apresenta como vantagens o facto de permitir utilizar uma diversidade de técnicas de DM, já disponíveis para dados não espaciais. Suprime a necessidade de caracterização geométrica das entidades geográ-…cas referenciadas, e principalmente, permite aos algoritmos de DM analisar simultaneamente dados geo-espaciais e dados não espaciais, não condicionando ou limitando os resultados que podem ser obtidos.

A compreensão, interpretação e utilização dos princípios estabelecidos nas pré-normas europeias para informação geográ…ca (CEN TC 287), facilitou a incorporação da componente espacial no processo de descoberta de conhecimento, apresentando-se como outro contributo desta tese, ao colaborar na construção de repositórios normalizados de informação geográ…ca.

A construção do sistema de inferências, que integra relações espaciais do tipo direcção, distância e topologia, representa outro importante contributo desta tese, uma vez que permite utilizar simultaneamente estes três tipos de relações no raciocínio, aumentando a precisão dos resultados. A dimensão dos objectos geográ…cos, nomeadamente a sua in‡uência no processo de raciocínio, permitiu, ainda, identi…car um conjunto de regras que contribuem para o aumento do desempenho do sistema de inferências construído.

Sistematizados os principais contributos desta tese, conclui-se com a satisfação de ver realizada a …nalidade que justi…cou este trabalho, e com a constatação do grati…cante que foi chegar até aqui e do estimulante que este projecto se mostrou.

Bibliografia

[Abbott et al., 1998] D. W. Abbott, I. P. Matkovsky, e J. F. Elder. An Evaluation of High-end Data Mining Tools for Fraud Detection. In IEEE International Conference on Systems, Man. and Cybernetics, San Diego, 12-14 October, 1998.

[Abdelmoty e El-Geresy, 1995] A. I. Abdelmoty e B. A. El-Geresy. A General Method for Spatial Reasoning in Spatial Databases. In Proceedings of the fourth International Conference on Information and Knowledge Management, 312–317, Baltimore, 1995.

[Abraham e Roddick, 1997] T. Abraham e J. F. Roddick. Discovering Meta-Rules in Mining Temporal and Spatio-Temporal Data. In Proceedings of the 8th. International Database Workshop, 29–31, Hong Kong, 1997.

[Abraham e Roddick, 1998] T. Abraham e J. F. Roddick. Opportunities for Knowledge Discovery in Spatio-Temporal Information Systems. Australian Journal of Information Systems, 5(2):3–12, 1998.

[Adam e Gangopadhyay, 1997] N. R. Adam e A. Gangopadhyay. Database Issues in Geographic Information Systems. Kluwer Academic Publishers, Massachusetts, 1997.

[Adriaans e Zantinge, 1996] P. Adriaans e D. Zantinge. Data Mining. Addison Wesley Longman, Edimburgo, 1996.

[Agrawal et al., 1993a] R. Agrawal, T. Imielinski, e A. Swami. Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6):914–925, 1993.

[Agrawal et al., 1993b] R. Agrawal, T. Imielinski, e A. Swami. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD Conference on Management of data, 207–216, Washington DC, 1993.

[Agrawal et al., 1996] R. Agrawal, M. Mehta, J. Shafer, e R. Srikant. The Quest Data Mining System. In Proceedings of the second International Conference on Knowledge Discovery and Data Mining, Portland, 1996.

[Allen, 1983] J. F. Allen. Maintaining Knowledge about Temporal Intervals. Communications of the ACM, 26(11):832–843, 1983.

[Alter, 1992] S. Alter. Information Systems: a management perspective. Addison-Wesley, 1992.

[Amorim e Correia, 1999] M. N. Amorim e A. Correia. Francisca Catarina (1846-1940): Vida e raízes em S. João do Pico - Biogra...a, Genealogia e Estudo de Comunidade. Instituto de Ciências Sociais, Universidade do Minho, 1999.

[Amorim et al., 2001] M. N. Amorim, A. Ferreira, F. Rodrigues, M. Santos, e P. Henriques. Reconstituição de Paróquias e Formação de uma Base de Dados Central. In ADEH'01, Castelo Branco, 2001.

[Amorim, 1992] M. N. Amorim. Evolução demográ...ca de três paróquias do Sul do Pico 1680-1980. Instituto de Ciências Sociais, Universidade do Minho, 1992.

[Anand et al., 1995] S. S. Anand, D. A. Bell, e J. G. Hughes. The Role of Domain Knowledge in Data Mining. In Proceedings of the fourth International Conference on Information and Knowledge Management, 37–43, Baltimore, 1995.

[Andrienko e Andrienko, 1998] G. L. Andrienko e N. V. Andrienko. Knowledge Extraction from Spatially Referenced Databases: a Project of an Integrated Environment. In Varenius Workshop on Status and Trends in Spatial Analysis, Sta. Barbara, CA, 1998.

[Arono¤, 1989] S. Arono¤. Geographic Information Systems: a management perspective. WDL Publications, Ottawa, 1989.

[Bailey, 1994] T. C. Bailey. A review of statistical spatial analysis in geographical information systems. In S. Fotheringham e P. Rogerson (Eds.), Spatial Analysis and GIS, 13–44. Taylor and Francis, London, 1994.

[Beek, 1992] P. V. Beek. Reasoning about Qualitative Temporal Information. Arti...cial Intelligence, 58:297–326, 1992.

[Bergadano, 1993] F. Bergadano. Inductive Databases Relations. IEEE Transactions on Knowledge and Data Engineering, 5(6):969–972, 1993.

[Berry e Lino¤, 2000] M. J. A. Berry e G. Lino¤. Mastering Data Mining: The Art and Science of Customer Relationship Management. Wiley Computer Publishing. John Wiley and Sons, Inc., 2000.

[Booch et al., 1999] G. Booch, J. Rumbaugh, e I. Jacobson. The Uni...ed Modeling Language User Guide. Addison Wesley Longman, Inc., 1999.

[Brachman e Anand, 1996] R. J. Brachman e T. Anand. The Process of Knowledge Discovery in Databases: a Human-Centered Approach. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining. The MIT Press, Massachusetts, 1996.

[Buckinghan et al., 1987] R. A. Buckinghan, R. A. Hirschheim, F. F. Land, e C. J. Tully. Information Systems Education: Recommendations and Implementation. Cambridge University Press, 1987.

[Burrough, 1986] P. A. Burrough. Principles of Geographical Information Systems for Land Resources Assessment. Oxford University Press, Oxford, 1986.

[CEN/TC-287, 1996a] CEN/TC-287. Geographic Information: Data description, Conceptual Schema Language. Technical Report CR 287005, European Committee for Standardisation, 1996.

[CEN/TC-287, 1996b] CEN/TC-287. Geographic Information: Data Description, Spatial Schema. Technical Report prENV 12160, European Committee for Standardisation, 1996.

[CEN/TC-287, 1996c] CEN/TC-287. Geographic Information: Reference Model. Technical Report prENV 12009, European Committee for Standardisation, 1996.

[CEN/TC-287, 1998a] CEN/TC-287. Geographic Information: Data Description, Metadata. Technical Report prENV 12657, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998b] CEN/TC-287. Geographic Information: Data Description, Quality. Technical Report prENV 12656, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998c] CEN/TC-287. Geographic Information: Data Description, Transfer. Technical Report prENV 12658, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998d] CEN/TC-287. Geographic Information: Fundamentals, Overview. Technical Report CR 287002, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998e] CEN/TC-287. Geographic Information: Fundamentals, Overview. Technical Report CR 287002, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998f] CEN/TC-287. Geographic Information: Processing, Query and Update: spatial aspects. Technical Report prENV 12660, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998g] CEN/TC-287. Geographic Information: Referencing, Direct Position. Technical Report prENV 12762, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998h] CEN/TC-287. Geographic Information: Referencing, Geographic Identi…ers. Technical Report prENV 12661, European Committee for Standardisation, 1998.

[CEN/TC-287, 1998i] CEN/TC-287. Geographic Information: Vocabulary. Technical Report CR 287003, European Committee for Standardisation, 1998.

[Chen et al., 1996] M.-S. Chen, J. Han, e P. S. Yu. Data Mining: an overview from a Database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6):866–883, 1996.

[Chen, 1976] P. P. Chen. The Entity-Relationship Model: Toward a Uni…ed View of Data. ACM Transactions on Database Systems, 1(1):9–35, 1976.

[Codd, 1970] E. F. Codd. A Relational Model for Large Shared Data Banks. Communications of the ACM, 13(6):377–387, 1970.

[Cohn, 1995] A. G. Cohn. The Challenge of Qualitative Spatial Reasoning. ACM Computing Surveys, 27(3):323–325, 1995.

[CT113, 1999] CT113. Tecnologias da Informação - Vocabulário (parte 28: Inteligência Arti…cial - Conceitos básicos e sistemas periciais). Norma Portuguesa 3003 prNP 3003-28, Instituto Português da Qualidade, 1999.

[Decker e Focardi, 1995] K. M. Decker e S. Focardi. Technology Overview: A Report on Data Mining. Technical report cscs tr-95-02, Swiss Scienti…c Computing Center, 1995.

[Dey e Roberts, 1996] S. Dey e S. A. Roberts. Combining Spatial and Relational Databases for Knowledge Discovery. In Proceedings of the …rst International Conference on GeoComputation, Leeds, 1996.

[Dzeroski e Lavrac, 1993] S. Dzeroski e N. Lavrac. Inductive Learning in Deductive Databases. IEEE Transactions on Knowledge and Data Engineering, 5(6):939–949, 1993.

[Egenhofer e Herring, 1991] M. J. Egenhofer e J. R. Herring. High-level spatial data structures for GIS. In D. J. Maguire, M. F. Goodchild, e D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications, Volume 1, 227–237. Longman Scienti…c and Technical, Harlow, 1991.

[Egenhofer e Sharma, 1993] M. J. Egenhofer e J. Sharma. Assessing the Consistency of Complete and Incomplete Topological Information. Geographical Systems, 1(1):47–68, 1993.

[Egenhofer, 1994a] M. Egenhofer. Spatial SQL: A Query and Presentation Language. IEEE Transactions on Knowledge and Data Engineering, 6(1):86–95, 1994.

[Egenhofer, 1994b] M. J. Egenhofer. Deriving the Composition of Binary Topological Relations. Journal of Visual Languages and Computing, 5(2):133–149, 1994.

[Elder e Pregibon, 1996] J. F. Elder e D. Pregibon. A Statistical Perspective on Knowledge Discovery in Databases. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, 83–113. The MIT Press, Massachusetts, 1996.

[Ester et al., 1997] M. Ester, H.-P. Kriegel, e J. Sander. Spatial Data Mining: A Database Approach. In Proceedings of the 5th International Symposium on Large Spatial Databases, 47–68, Berlin, Germany, 1997. Springer-Verlag.

[Ester et al., 1998a] M. Ester, A. Frommelt, H.-P. Kriegel, e J. Sander. Algorithms for Characterization and Trend Detection in Spatial Databases. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, 44–50. AAAI Press, 1998.

[Ester et al., 1998b] M. Ester, H.-P. Kriegel, J. Sander, e X. Xu. Clustering for Mining in Large Spatial Databases. KI-Journal, Special Issue on Data Mining, 1:18–24, 1998.

[Ester et al., 1999a] M. Ester, S. Gundlach, H.-P. Kriegel, e J. Sander. Database Primitives for Spatial Data Mining. In Proceedings of BTW'99 - International Conference on Databases in O¢ce, Engineerintg and Science, Germany, 1999.

[Ester et al., 1999b] M. Ester, H.-P. Kriegel, e J. Sander. Knowledge Discovery in Spatial Databases. In Proceedings of KI'99 - Conference on Arti…cial Intelligence, Bonn, Germany, 1999.

[Famili et al., 1997] A. Famili, W.-M. Shen, R. Weber, e E. Simoudis. Data preprocessing and Intelligent Data Analysis. Intelligent Data Analysis, 1(1), 1997.

[Fayyad e Uthurusamy, 1996] U. Fayyad e R. Uthurusamy. Data Mining and Knowledge Discovery in Databases. Communications of the ACM, 39(11):24–26, 1996.

[Fayyad et al., 1996a] U. Fayyad, G. Piatetsky-Shapiro, e P. Smyth. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11):27–34, 1996.

[Fayyad et al., 1996b] U. M. Fayyad, G. Piatetsky-Shapiro, e P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, 1–34. The MIT Press, Massachusetts, 1996.

[Fayyad et al., 1996c] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.). Advances in Knowledge Discovery and Data Mining. The MIT Press, Massachusetts, 1996.

[Fotheringham e Rogerson, 1994] S. Fotheringham e P. Rogerson (Eds.). Spatial analysis and GIS. Taylor and Francis, Bu¤alo, 1994.

[Frank, 1992] A. U. Frank. Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. Journal of Visual Languages and Computing, (3):343–371, 1992.

[Frank, 1994] A. U. Frank. Qualitative Temporal Reasoning in GIS: Ordered Time Scales. In Proceedings of the Sixth International Symposium on Spatial Data Handling (SDH'94), 410–430, Edinburgh, 1994.

[Frank, 1996] A. U. Frank. Qualitative Spatial Reasoning: cardinal directions as an example. International Journal of Geographical Information Systems, 10(3):269–290, 1996.

[Freksa e Rohrig, 1993] C. Freksa e R. Rohrig. Dimensions of Qualitative Spatial Reasoning. In Proceedings of the QUARDET - Qualitative Reasoning in Decision Technologies, Barcelona, 1993.

[Freksa, 1991] C. Freksa. Qualitative Spatial Reasoning. In D. M. Mark e A. U. Frank (Eds.), Cognitive and Linguistic Aspetcs of Geographic Space, 361–372. Kluwer Academic Publishers, 1991.

[Freksa, 1992] C. Freksa. Using Orientation Information for Qualitative Spatial Reasoning. In A. U. Frank, I. Campari, e U. Formentini (Eds.), Theories and Methods of Spatio-Temporal Reasoning in Geographic space, Lectures Notes in Computer Science 639. Springer-Verlag, Berlin, 1992.

[Gahegan, 1995] M. Gahegan. Proximity Operators for Qualitative Spatial Reasoning. In A. U. Frank e W. Kuhn (Eds.), Spatial Information Theory - A Theoretical Basis for GIS, Proceedings of the International Conference COSIT'95, Lectures Notes in Computer Science 988, 31–44. Springer-Verlag, Semmering, Austria, 1995.

[Gatrell, 1991] A. C. Gatrell. Concepts of space and geographical data. In D. J. Maguire, M. F. Goodchild, e D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications, Volume 1, 119–134. Longman Scienti…c and Technical, Harlow, 1991.

[Goebel e Gruenwald, 1999] M. Goebel e L. Gruenwald. A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations, ACM SIGKDD, 1(1):20–33, 1999.

[Goh et al., 1996] C.-L. Goh, M. Tsukamoto, e S. Nishio. Knowledge Discovery in Deductive Databases with Large Deduction Results: the …rst step. IEEE Transactions on Knowledge and Data Engineering, 8(6):952–961, 1996.

[Gouveia et al., 2001] C. Gouveia, P. Henriques, R. Nicolau, J. Rocha, e M. Santos. Moving from CEN TC 287 to ISO TC 211 - The approach of the Portuguese National Geographic Information Infrastructure. In Proceedings of the AGILE Conference - GI in Europe: Integrative, Interoperable, Interactive, Brno, República Checa, 2001.

[Grigni et al., 1995] M. Grigni, D. Papadias, e C. Papadimitriou. Topological Inference. In Proceedings of the International Joint Conference of Arti…cial Intelligence (IJCAI), Montreal, Canada, 1995.

[Güting, 1994] R. Güting. An Introduction to Spatial Database System. The VLDB Journal, 3(4):357–400, 1994.

[Hadzilacos e Tryfona, 1996] T. Hadzilacos e N. Tryfona. Logical Data Modelling for Geographical Applications. International Journal of Geographical Information Systems, 10(2):179–203, 1996.

[Haining, 1994] R. Haining. Designing spatial data analysis modules for geographical information systems. In S. Fotheringham e P. Rogerson (Eds.), Spatial Analysis and GIS, 45–63. Taylor and Francis, London, 1994.

[Han e Kamber, 2001] J. Han e M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[Han et al., 1992] J. Han, Y. Cai, e N. Cercone. Knowledge Discovery in Databases: An Attribute-Oriented Approach. In Proceedings of the 18th VLDB conference, 547–559, Vancouver, Canada, 1992.

[Han et al., 1994] J. Han, Y. Fu, Y. Huang, Y. Cai, e N. Cercone. DBLearn: a system prototype for Knowledge Discovery in Relational Databases. In Proceedings on the 1994 ACM SIGMOD International Conference on Management of Data, 516. ACM, 1994.

[Han et al., 1997] J. Han, K. Koperski, e N. Stefanovic. GeoMiner: A System Prototype for Spatial Data Mining. In Proceedings 1997 ACM-SIGMOD International Conference on Management of Data, Arizona, 1997.

[Hernández et al., 1995] D. Hernández, E. Clementini, e P. D. Felice. Qualitative Distances. In A. U. Frank e W. Kuhn (Eds.), Spatial Information Theory - A Theoretical Basis for GIS, Proceedings of the International Conference COSIT'95, Lectures Notes in Computer Science 988, 45–57, Semmering, Austria, 1995. Springer-Verlag.

[Hernández, 1991] D. Hernández. Relative representation of spatial knowledge: the 2-D case. In D. M. Mark e A. U. Frank (Eds.), Cognitive and Linguistic Aspetcs of Geographic Space, 373–385. Kluwer Academic Publishers, 1991.

[Hernández, 1994] D. Hernández. Qualitative Representations of Spatial Knowledge. Lecture Notes in Arti…cial Intelligence 804. Springer-Verlag, Berlin, 1994.

[Holsheimer e Kersten, 1994] M. Holsheimer e M. L. Kersten. Architectural support for data mining. Technical Report CS-R9429, Centrum voor Wiskunde en Informatica - Amsterdam, 1994.

[Holsheimer e Siebes, 1994] M. Holsheimer e A. Siebes. Data Mining: The search for Knowledge in Databases. Technical Report CS-R9406, Centrum voor Wiskunde en Informatica - Amsterdam, 1994.

[Hong et al., 1995] J.-H. Hong, M. J. Egenhofer, e A. U. Frank. On the Robustness of Qualitative and Direction Reasoning. In Proceedings of Auto-Carto 12, 301–310, Charlotte, North California, 1995.

[Hong, 1994] J.-H. Hong. Qualitative Distance and Direction Reasoning in Geographic Space. PhD thesis, University of Maine, 1994.

[Intergraph, 1995] Intergraph. MicroStation 95 version 05.05.02.23 Windows x86. Bentley Systems, Intergraph Corporation, 1995.

[Intergraph, 1999a] Intergraph. Developing Enterprise Solutions with GeoMedia Technology. Technical Report DJA075130, Intergraph Corporation, 1999.

[Intergraph, 1999b] Intergraph. Geomedia Professional v3, Reference Manual. Intergraph Corporation, 1999.

[ISO/TC-211, 1999a] ISO/TC-211. Geographic Information - Overview. Technical Report N723, International Standard Organisation, 1999.

[ISO/TC-211, 1999b] ISO/TC-211. Geographic Information - Spatial referencing by geographic identi…ers. Technical Report N822, International Standard Organisation, 1999.

[ISO/TC-211, 1999c] ISO/TC-211. Geographic Information - Spatial Schema. Technical Report N818, International Standard Organisation, 1999.

[ISO/TC-211, 1999d] ISO/TC-211. Geographic Information - Terminology (CD 19104.2). Technical Report N816, International Standard Organisation, 1999.

[Khabaza e Brewer, 2000] T. Khabaza e S. Brewer. ILP - Advanced Rule-based Modelling for Clementine. In Clementine User Group Conference, Woke…eld Park Conference Center, 2000.

[Kohonen, 1989] T. Kohonen. Self-Organization and Associative Memory. Springer-Verlag, Berlin, 3th. edition, 1989.

[Koperski e Han, 1995] K. Koperski e J. Han. Discovery of Spatial Association Rules in Geographic Information Systems. In Proc. 4th International Symposium on Large Spatial Databases (SSD95), 47–66, Maine, 1995.

[Koperski e Han, 1996] K. Koperski e J. Han. Data Mining Methods for the analysis of Large Geographic Databases. In Proceedings of the 10th Annual Conference on GIS, Vancouver, 1996.

[Koperski et al., 1996] K. Koperski, J. Adhikary, e J. Han. Spatial Data Mining: Progress and Challenges. In Proc. of the 1996 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, 1996.

[Koperski et al., 1998] K. Koperski, J. Han, e N. Stefanovic. An E¢cient Two-Step Method for Classi…cation of Spatial Data. In Proceedings of the International Symposium on Spatial Data Handling (SDH'98), Vancouver, Canada, 1998.

[Laurini e Thompson, 1992] R. Laurini e D. Thompson. Fundamentals of Spatial Information Systems. Academic Press, San Diego, 1992.

[Lobo e Pires, 1998] V. Lobo e F. M. Pires. SOM - Kohonen's Self-Organizing Maps. In Proceedings of the International Summer School on Knowledge Discovery in Databases and Data Mining: Methods and Applications, Volume II, Caminha, Portugal, 1998.

[Lu et al., 1993] W. Lu, J. Han, e B. C. Ooi. Discovery of General Knowledge in Large Spatial Databases. In Proc. of the 1993 Far East Workshop on Geographic Information Systems, 275–289, Singapura, 1993.

[Lu et al., 1996] H. Lu, R. Setiono, e H. Liu. E¤ective Data Mining using Neural Networks. IEEE Transactions on Knowledge and Data Engineering, 8(6):957–961, 1996.

[Maguire, 1991] D. J. Maguire. An overview and de…nition of GIS. In D. J. Maguire, M. F. Goodchild, e D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications, Volume 1, 9–20. Longman Scienti…c and Technical, Harlow, 1991.

[Maling, 1991] D. H. Maling. Coordinate Systems and Map Projections for GIS. In D. J. Maguire, M. F. Goodchild, e D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications, Volume 1, 135–146. Longman Scienti…c and Technical, Harlow, 1991.

[Matheus et al., 1993] C. J. Matheus, P. K. Chan, e G. Piatetsky-Shapiro. Systems for Knowledge Discovery in Databases. IEEE Transactions on Knowledge and Data Engineering, 5(6):903–913, 1993.

[Matos, 1997] J. Matos. Normalização em Cartogra…a. Ingenium, (Novembro), 1997.

[Mohan, 2000] C. K. Mohan. Frontiers of Expert Systems: Reasoning with Limited Knowledge. International Series in Engineering and Computer Science. Kluwer Academic Publishers, 2000.

[Murray e Estivill-Castro, 1998] A. T. Murray e V. Estivill-Castro. Cluster discovery techniques for exploratory data analysis. International Journal of Geographical Information Science, 12(5):431–443, 1998.

[Navathe, 1992] S. B. Navathe. Evolution of Data Modeling for Databases. Communications of the ACM, 35(9):112–123, 1992.

[Ng e Han, 1994] R. T. Ng e J. Han. E¢cient and E¤ective Clustering Methods for Spatial Data Mining. In Proceedings of the 20th Very Large Databases Conference, Santiago, Chile, 1994.

[OGC, 1999a] OGC. The OpenGIS Abstract Speci…cation. Topic 0: Abstract Speci…cation Overview. Version 4. Technical Report 99-100r1, Open GIS Consortium, 1999.

[OGC, 1999b] OGC. OpenGIS Simple Features Speci…cation for SQL. Revision 1.1. Technical Report 99-049, Open GIS Consortium, 1999.

[O'Kelly, 1994] M. E. O'Kelly. Spatial analysis and GIS. In S. Fotheringham e P. Rogerson (Eds.), Spatial Analysis and GIS, 66–79. Taylor and Francis, London, 1994.

[Openshaw, 1991] S. Openshaw. Developing appropriate spatial analysis methods for GIS. In D. J. Maguire, M. F. Goodchild, e D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications, Volume 1, 389–402. Longman Scienti…c and Technical, Harlow, 1991.

[Padmanabhan e Tuzhilin, 1999] B. Padmanabhan e A. Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, 27(3):303–318, 1999.

[Painho, 1996] M. Painho. Um novo olhar sobre a Terra. Ingenium, (Fevereiro):35–37, 1996.

[Painho, 1997] M. Painho. Texto de Apoio: Sistemas de Informação Geográ…ca. Bloco 3 - Natureza dos dados espaciais. URL: http://aurelia.si.fct.unl.pt/ www-unl/sid/recint/texta/sig1.htm, URL: http://aurelia.si.fct.unl.pt/ www-unl/sid/recint/texta/sig2.htm, URL: http://aurelia.si.fct.unl.pt/ www-unl/sid/recint/texta/sig3.htm, Instituto Superior de Estatística e Gestão da Informação, 1997.

[Papadias e Sellis, 1994] D. Papadias e T. Sellis. On the Qualitative Representation of Spatial Knowledge in 2D Space. Very Large Databases Journal, Special Issue on Spatial Databases, 3(4):479–516, 1994.

[Papadias e Theodoridis, 1997] D. Papadias e Y. Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. International Journal of Geographical Information Systems, 11(2):111–138, 1997.

[Pawlak, 1991] Z. Pawlak. Rough sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991.

[Pereira, 1997] J. L. Pereira. Tecnologia de Bases de Dados. FCA - Editora de Informática, Lisboa, 1997.

[Quinlan, 1986] J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81–106, 1986.

[Rainsford e Roddick, 1996] C. P. Rainsford e J. F. Roddick. A Survey of Issues in Data Mining. Technical Report CIS-96-006, School of Computer and Information Science, University of South Australia, 1996.

[Ravada e Sharma, 1999] S. Ravada e J. Sharma. Oracle8i Spatial: Experiences with Extensible Databases. In R. H. Güting, D. Papadias, e F. Lochovsky (Eds.), Advances in Spatial Databases, LNCS 1651, Proceedings of the 6th International Symposium on Spatial Databases (SSD'99), 355–359, Hong Kong, 1999.

[Rodrigues et al., 1999] M. F. Rodrigues, C. Ramos, e P. R. Henriques. An Intelligent System to Study Demographic Evolution. In Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology, 161–170, Orlando, Florida, 1999.

[Rodrigues, 2000] M. F. Rodrigues. Arquitectura Heterogénea para Extracção de Conhecimento a partir de Dados. Tese de Doutoramento, Universidade do Minho, 2000.

[Roman, 1990] G.-C. Roman. Formal speci…cation of geographic data processing requirements. IEEE Transactions on Knowledge and Data Engineering, 2(4):370–380, 1990.

[Russell e Norvig, 1995] S. Russell e P. Norvig. Arti…cial Intelligence - A Modern Approach. Prentice Hall International, Inc., New Jersey, 1995.

[Samet e Aref, 1995] H. Samet e W. G. Aref. Spatial Data Models and Query Processing. In W. Kim (Eds.), Modern Database Systems: The Object Model, Interoperability and Beyond, 338–360. Addison Wesley/ACM Press, 1995.

[Samet, 1995] H. Samet. Spatial Data Structures. In W. Kim (Eds.), Modern Database Systems: The Object Model, Interoperability and Beyond, 361–385. Addison Wesley/ACM Press, 1995.

[Santos e Amaral, 1999] M. Santos e L. Amaral. As Normas de Informação Geográ…ca e o Raciocínio Espacial Qualitativo na Inferência de Informação Geográ…ca Qualitativa. In Proceedings do V Encontro de Sistemas de Informação Geográ…ca, Lisboa, Portugal, 24-26 Novembro, 1999. Edição em CD-ROM.

[Santos e Amaral, 2000a] M. Santos e L. Amaral. Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning. In PADD'00 Proceedings of the 4th International Conference and Exhibition on Practical Applications of Knowledge Discovery and Data Mining, 73–88, Manchester, 2000. 11-13 April.

[Santos e Amaral, 2000b] M. Santos e L. Amaral. Knowledge Discovery in Spatial Databases: the Padrão's qualitative approach. Cities and Regions, GIS special issue(November):33–49, 2000.

[Santos e Amaral, 2000c] M. Santos e L. Amaral. O Padrão na Descoberta de Conhecimento em Bases de Dados Demográ…cas. In 1ra. Conferência da Associação Portuguesa de Sistemas de Informação, Guimarães, 25-27 Outubro, 2000. Edição em CD-ROM.

[Santos e Amaral, 2000d] M. Santos e L. Amaral. A Qualitative Spatial Reasoning Approach in Knowledge Discovery in Spatial Databases. In Proceedings of Data Mining 2000: Data Mining Methods and Databases for Engineering, Finance and Others Fields, 249–258, Cambridge University, 2000. WIT Press, 5-7 July.

[Santos et al., 1999] M. Santos, L. Amaral, e P. Pimenta. A Descoberta de Conhecimento em Bases de Dados Geográ…cas através da Explicitação Semântica. In GISBrasil'99 - V Congress and Exhibition of Latin America Geo-processing Users, Salvador, Brasil, 19-23 July, 1999. Edição em CD-ROM.

[Santos, 1998] M. C. Santos. A…nal, o que é a Geomática? URL: http://www.fatorgis.com.br, 1998.

[Santos, 2000] M. F. Santos. Sistemas de Classi…cação em Ambientes Distribuídos. Tese de Doutoramento, Universidade do Minho, 2000.

[Schenck e Wilson, 1994] D. Schenck e P. Wilson. Information Modeling: The EXPRESS Way. Oxford University Press, Oxford, 1994.

[Sharma, 1996] J. Sharma. Integrated Spatial Reasoning in Geographic Information Systems: Combining Topology and Direction. PhD Thesis, University of Maine, 1996.

[Shekhar et al., 1997] S. Shekhar, M. Coyle, B. Goyal, D.-R. Liu, e S. Sarkar. Data Models in Geographic Information Systems. Communications of the ACM, 10(4):103–111, 1997.

[Shekhar et al., 1999] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, e C. t. Lu. Spatial Databases - Accomplishments and Research Needs. IEEE Transactions on Knowledge and Data Engineering, 11(1):45–55, 1999.

[Shepherd, 1991] I. Shepherd. Information Integration and GIS. In D. J. Maguire, M. F. Goodchild, e D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications, Volume 1, 337–360. Longman Scienti…c and Technical, Harlow, 1991.

[Silberschatz e Tuzhilin, 1995] A. Silberschatz e A. Tuzhilin. On subjective Measures of Interestingness in Knowledge Discovery. In Proceedings of the …rst International Conference on Knowledge Discovery and Data Mining, 275–281, Montréal, 1995. AAAI Press.

[Silberschatz e Tuzhilin, 1996a] A. Silberschatz e A. Tuzhilin. User-Assisted Knowledge Discovery: How much should the user be involved. In ACM-SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996.

[Silberschatz e Tuzhilin, 1996b] A. Silberschatz e A. Tuzhilin. What makes patterns interesting in Knowledge Discovery Systems. IEEE Transactions on Knowledge and Data Engineering, 8(6):970–974, 1996.

[Simoudis et al., 1995] E. Simoudis, B. Livezey, e R. Kerber. Using Recon for Data Cleaning. In Proceedings of the …rst International Conference on Knowledge Discovery and Data Mining, 282–287, Montréal, 1995.

[Son et al., 1998] E.-J. Son, I.-S. Kang, T.-W. Kim, e K.-J. Li. A Spatial Data Mining Method by Clustering Analysis. In Proceedings of the 6th International Symposium on Advances in GIS, 157–158, Washington, 1998.

[Sousa et al., 2000] A. A. Sousa, J. L. Pereira, e J. Á. Carvalho. A Linguagem XML numa Perspectiva de Bases de Dados. In 1ra. Conferência da Associação Portuguesa de Sistemas de Informação, Guimarães, 25-27 Outubro, 2000. Edição em CD-ROM.

[SPSS, 1999a] SPSS. Clementine, Reference Manual, Version 5.2. SPSS Inc., 1999.

[SPSS, 1999b] SPSS. Clementine, User Guide, Version 5.2. SPSS Inc., 1999.

[Theodoridis et al., 1996] Y. Theodoridis, D. Papadias, e E. Stefanakis. Supporting Direction Relations in Spatial Database Systems. In Proceedings of the 7th. International Symposium on Spatial Data Handling (SDH'96), Netherlands, 1996.

[Tichy, 1998] W. F. Tichy. Should Computer Scientists Experiment More? IEEE Computer, (May):32–40, 1998.

[Tom, 1994] H. Tom. The Geographic Information Systems (GIS) Standards Infrastructure. StandardView, 2(3):133–139, 1994.

[Wang et al., 1997] W. Wang, J. Yang, e R. Muntz. STING: A Statistical Information Grig Approach to Spatial Data Mining. In Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997.

[Wijsen, 1998] J. Wijsen. Reasoning about qualitative trends in databases. Information Systems, 23(7):463–487, 1998.

[Wirth e Hipp, 2000] R. Wirth e J. Hipp. CRISP-DM: Towards a Standard Process Model for Data Mining. In PADD'00 Proceedings of the 4th International Conference and Exhibition on Practical Applications of Knowledge Discovery and Data Mining, 29–39, Manchester, 2000.

[Xu et al., 1998] X. Xu, M. Ester, H.-P. Kriegel, e J. Sander. A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. In Proceedings of the 14th International Conference on Data Engineering (ICDE), 324–331, Orlando, FL, 1998.

[Zelkowitz e Wallace, 1997] M. V. Zelkowitz e D. Wallace. Experimental Validation in Sofware Engineering. In Proceedings of the Empirical Assessment of Software Engineering Conference, Keele University, 1997.

[Zelkowitz e Wallace, 1998] M. V. Zelkowitz e D. R. Wallace. Experimental Models for Validating Technology. IEEE Computer, (May):23–31, 1998.

[Zimmermann e Freksa, 1996] K. Zimmermann e C. Freksa. Qualitative Spatial Reasoning Using Orientation, Distance and Path Knowledge. Applied Intelligence, (6):49–58, 1996.

[Zimmermann, 1993] K. Zimmermann. Enhancing Qualitative Spatial Reasoning: Combining Orientation and Distance. Technical Report 26, Hamburg University, 1993.

[Zimmermann, 1995] K. Zimmermann. Measuring without Measures: The D-Calculus. In A. U. Frank e W. Kuhn (Eds.), Spatial Information Theory - A Theoretical Basis for GIS, Proceedings of the International Conference COSIT'95, Lectures Notes in Computer Science 988, 59–67. Springer-Verlag, Semmering, Austria, 1995.

Índice de Autores

Apêndices

Apêndice A

Integração da Direcção e Topologia

Este apêndice sintetiza algumas das tabelas de composição construídas por Sharma [Sharma, 1996], para a integração de relações espaciais do tipo direcção e topologia. A descrição é limitada ao caso das entidades geográ…cas representarem subdivisões administrativas, para as quais as relações topológicas possíveis são deslocado e adjacente.

A.1 Relações topológicas deslocado;adjacente

A Tabela A.1 apresenta a tabela de composição que integra a direcção com o par topológico deslocado;adjacente.

A.2 Relações topológicas adjacente;deslocado

A Tabela A.2 apresenta a tabela de composição que integra a direcção com o par topológico adjacente;deslocado.

A.3 Relações topológicas adjacente;adjacente

A Tabela A.3 apresenta a tabela de composição que integra a direcção com o par topológico adjacente;adjacente.

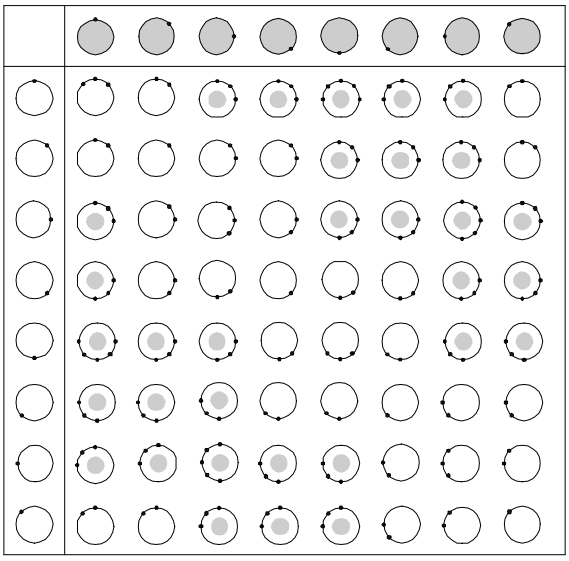


Tabela A.1: Integração da direcção com o par topológico deslocado;adjacente
Adaptado de: [Sharma, 1996] p. 118

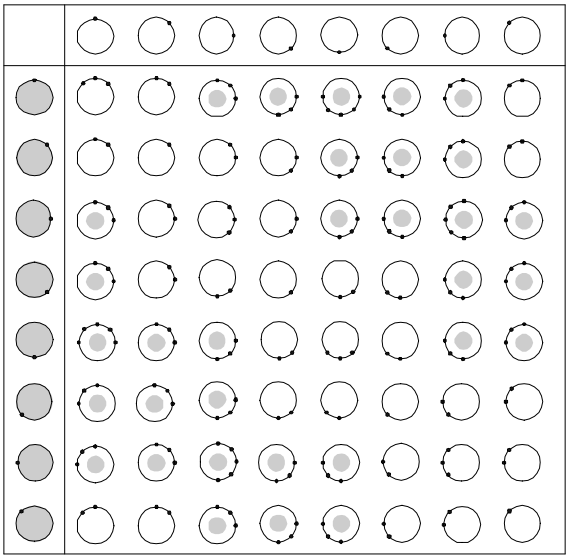


Tabela A.2: Integração da direcção com o par topológico adjacente;deslocado
Adaptado de: [Sharma, 1996] p. 119

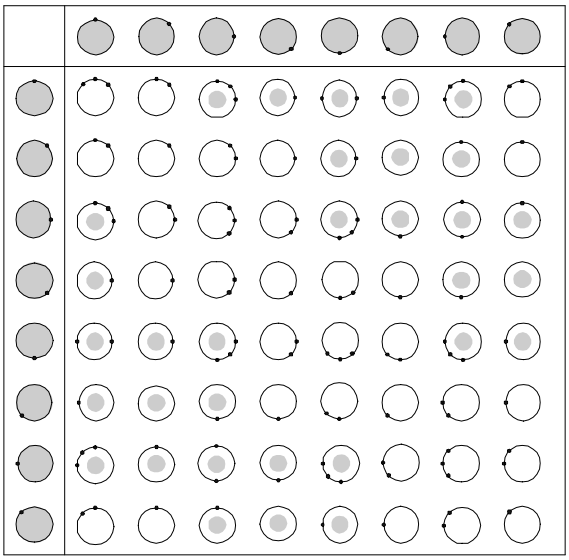


Tabela A.3: Integração da direcção com o par topológico adjacente;adjacente
Adaptado de: [Sharma, 1996] p. 120

Apêndice B

Integração da Direcção, Distância e Topologia

A integração dos três tipos de relações espaciais, direcção, distância e topologia, apenas é possível após a construção de um sistema de referenciação comum, no qual ambas as aproximações, integração de direcção e distância e integração de direcção e topologia, utilizem o modelo triangular.

A criação de novas tabelas de composição, que integrem a direcção e topologia segundo os princípios do raciocínio espacial qualitativo, passa pela utilização dos princípios temporais de…nidos por Allen [Allen, 1983]. A caracterização do domínio espacial, seguindo estes princípios, foi já efectuada no Capítulo 3, secção 3.5.3, pelo que nas próximas secções são apresentados os passos que conduziram à construção das regras de inferência, para a integração da direcção e topologia. Estas são apresentadas seguindo a ordem dos pares topológicos deslocado;deslocado, deslocado;adjacente, adjacente;deslocado e adjacente;adjacente.

B.1 Integração da direcção e topologia

O objectivo da construção de um sistema integrado de raciocínio espacial, que conjugue a direcção, distância e topologia, é o de permitir inferir relações de localização, o mais precisas possível, desconhecidas entre objectos. Neste caso particular, os objectos referenciam regiões administrativas. Tal facto adiciona algumas di…culdades à construção das regras, uma vez que as entidades geográ…cas consideradas possuem limites com contornos irregulares. Isto quer dizer que os pares de primitivas temporais considerados como oportunos na caracterização do domínio espacial, e a partir dos quais é iniciado o processo de construção das regras, podem não retratar todas as situações possíveis. As Figuras 3.16 e 3.17 (Capítulo 3) apresentam o conjunto de primitivas de…nidas para iniciar o processo de construção das regras, enquanto que a Figura B.1 evidencia algumas situações particulares, que podem existir entre regiões, e cujos pares de intervalos não estão considerados no conjunto inicial de…nido no Capítulo 3. Esta situação, ocasionada neste trabalho pelas diferentes dimensões das regiões envolvidas e pelo formato dos seus contornos, permite que um mesmo par de primitivas temporais, por exemplo (sobreposto, sobreposto), possa retratar duas situações distintas, (NE, deslocado) e (NE, adjacente). A

interpretação da relação correcta é efectuada no momento da integração da direcção e topologia com a direcção e distância, uma vez que a distância existente entre os objectos condiciona a relação topológica que pode existir entre os mesmos.

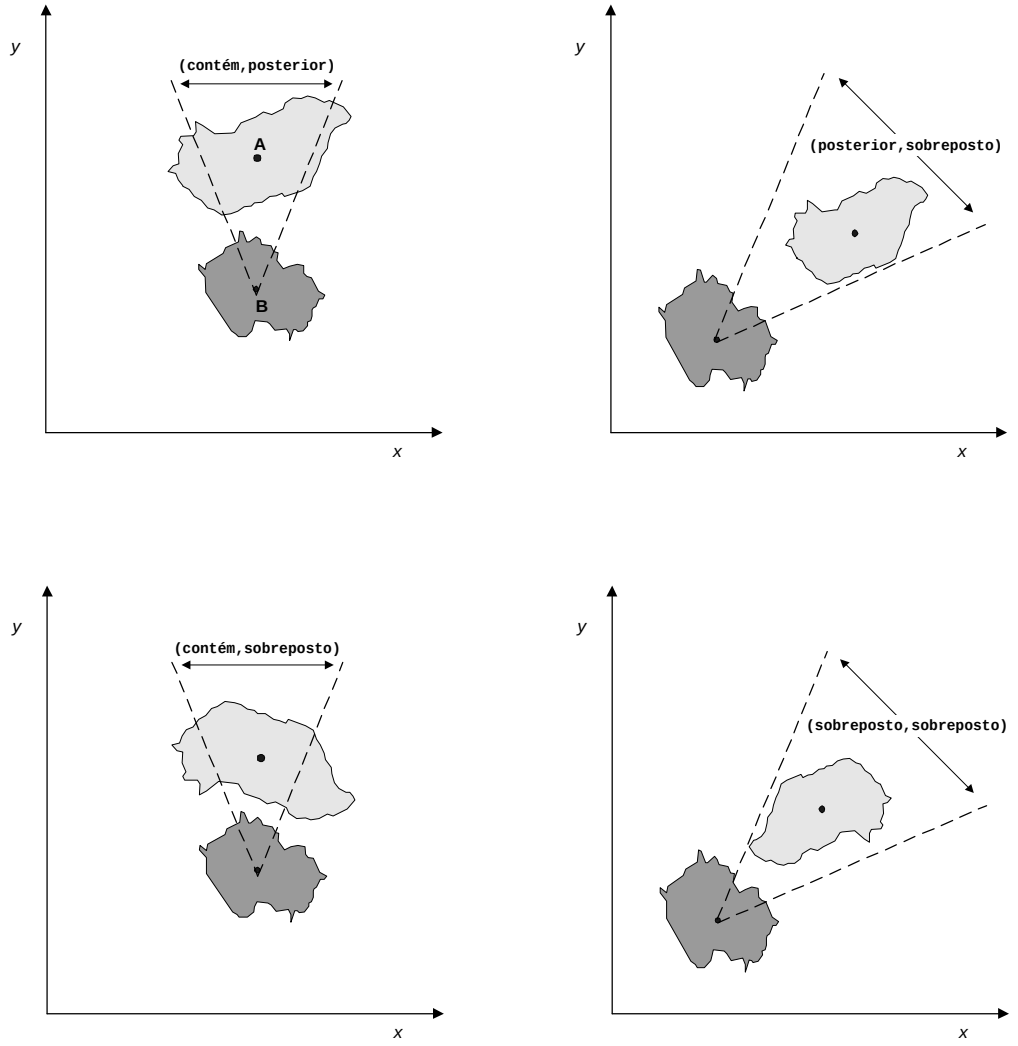


Figura B.1: Casos particulares de primitivas temporais na caracterização da direcção e topologia

A identi…cação destas situações, e a sua consequente inclusão na integração com a distância, impõe alguma ‡exibilidade ao sistema de inferências, permitindo que duas regiões possam estar próximas uma da outra, sem existir qualquer contacto entre os limites das mesmas. Tal permite ainda que o processo de construção das regras seja iniciado a partir das primitivas apresentadas nas Figuras 3.16 e 3.17 (Capítulo 3), e que o resultado da composição seja interpretado atendendo aos casos particulares apresentados anteriormente. A Tabela B.1 sistematiza o conjunto de primitivas temporais que permite a identi…cação de um dado par (direcção, topologia).

(N,deslocado)	(durante, posterior) (d,a) (contém, posterior) (di,a) (durante, sobreposto) (d,ob) (contém, sobreposto) (di,ob)	(NE,deslocado)	(posterior, posterior) (a,a) (posterior, sobreposto) (a,ob) (sobreposto, posterior) (ob,a) (sobreposto, sobreposto) (ob,ob)
(E,deslocado)	(posterior, durante) (a,d) (posterior, contém) (a,di) (sobreposto, durante) (ob,d) (sobreposto, contém) (ob,di)	(SE,deslocado)	(posterior, anterior) (a,b) (posterior, sobrepõe) (a,o) (sobreposto, anterior) (ob,b) (sobreposto, sobrepõe) (ob,o)
(S,deslocado)	(durante, anterior) (d,b) (contém, anterior) (di,b) (durante, sobrepõe) (d,o) (contém, sobrepõe) (di,o)	(SO,deslocado)	(anterior, anterior) (b,b) (anterior, sobrepõe) (b,o) (sobrepõe, anterior) (o,b) (sobrepõe, sobrepõe) (o,o)
(O,deslocado)	(anterior, durante) (b,d) (anterior, contém)(b,di) (sobrepõe, durante) (o,d) (sobrepõe, contém) (o,di)	(NO,deslocado)	(anterior, posterior) (b,a) (anterior, sobreposto) (b,ob) (sobrepõe, posterior) (o,a) (sobrepõe, sobreposto) (o,ob)

Tabela B.1: Primitivas temporais possíveis, na integração da direcção com a relação topológica deslocado

B.1.1 Par topológico deslocado;deslocado

Para a integração da direcção e topologia, a Tabela B.2 apresenta o conjunto das regras que permitem a composição da direcção Norte com as restantes direcções, para o caso do par topológico deslocado;deslocado. A Tabela B.3 apresenta a integração da direcção Nordeste com as restantes direcções, para o mesmo par topológico. Os restantes casos são obtidos por rotação das respectivas direcções, como pode ser veri…cado na tabela …nal apresentada na Tabela B.11.

Composição		Resultado
(N,deslocado);(N,deslocado)	= (d,a);(d,a) = (d;d) x (a;a) = (d,a)	
(N,deslocado);(NE,deslocado)	= (d,a);(a,a) = (d;a) x (a;a) = (a,a)	
(N,deslocado);(E,deslocado)	= (d,a);(a,d) = (d;a) x (a;d) = (a) x {a,ob,d,mb,f} = (a,a) _ (a,d) _ (a,ob)	
(N,deslocado);(SE,deslocado)	= (d,a);(a,b) = (d;a) x (a;b) = (a) x {?}	
(N,deslocado);(S,deslocado)	= (d,a);(d,b) = (d;d) x (a;b) = (d) x {?} = (d,a) _ (d,b) _ (d,ob)_ (d,o)	
(N,deslocado);(SO,deslocado)	= (d,a);(b,b) = (d;b) x (a;b) = (b) x {?} = (b,a) _ (b,d) _ (b,b)	
(N,deslocado);(O,deslocado)	= (d,a);(b,d) = (d;b) x (a;d) = (b) x {a,ob,mb,d,f} = (b,a) _ (b,ob) _ (b,d)	
(N,deslocado);(NO,deslocado)	= (d,a);(b,a) = (d;b) x (a;a) = (b,a)	

Tabela B.2: Integração da direcção Norte com o par topológico deslocado;deslocado

Composição		Resultado
(NE,deslocado);(E,deslocado)	= (a,a);(a,d) = (a;a) x (a,d) = (a) x {a,ob,d) = (a,a) _ (a,ob) _ (a,d)	
(NE,deslocado);(SE,deslocado)	= (a,a);(a,b) = (a;a) x (a;b) = (a) x {?} = (a,a) _ (a,d) _ (a,b)	
(NE,deslocado);(S,deslocado)	= (a,a);(d,b) = (a;d) x (a,b) = {a,ob,d) x {?} = (a,a) _ (a,d) _ (a,b) _ (ob,ob) _ (ob,o) _ (d,a) _ (d,b) _ (ob,d) _ (d,ob) _ (d,o)	
(NE,deslocado);(SO,deslocado)	= (a,a);(b,b) = (a;b) x (a,b) = {?} x {?}	
(NE,deslocado);(O,deslocado)	= (a,a);(b,d) = (a;b) x (a;d) = {?} x {a,ob,mb,d,f} = (a,a) _ (b,a) _ (ob,ob) _ (o,ob) _ (d,ob) _ (ob,d) _ (o,d) _ (a,d) _ (b,d) _ (d,a)	
(NE,deslocado);(NO,deslocado)	= (a,a);(b,a) = (a;b) x (a;a) = {?} x (a) = (a,a) _ (d,a) _ (b,a)	

Tabela B.3: Integração da direcção Nordeste com o par topológico deslocado;deslocado

B.1.2 Par topológico deslocado;adjacente

A Tabela B.4 apresenta o conjunto das regras que permitem a composição da direcção Norte com as restantes direcções, para o caso do par topológico deslocado;adjacente. A Tabela B.5 apresenta a integração da direcção Nordeste com as restantes direcções, para o mesmo par topológico. Os restantes casos são obtidos por rotação das respectivas direcções, como pode ser veri…cado na tabela …nal apresentada na Tabela B.11.

Composição		Resultado
(N,deslocado);(N,adjacente)	= (d,a);(d,ob) = (d;d) x (a;ob) = (d,a)	
(N,deslocado);(NE,adjacente)	= (d,a);(ob,ob) = (d;ob) x (a;ob) = {a,ob,d} x (a) = (a,a) _ (d,a)	
(N,deslocado);(E,adjacente)	= (d,a);(ob,d) = (d;ob) x (a;d) = {a,ob,d} x {a,ob,d} = (a,a) _ (a,d) _ (ob,ob) _ (ob,d) _ (d,a) _ (d,ob)	
(N,deslocado);(SE,adjacente)	= (d,a);(ob,o) = (d;ob) x (a;o) = {a,ob,d} x {a,ob,d} = (a,a) _ (a,d) _ (ob,ob) _ (ob,d) _ (d,a) _ (d, ob)	
(N,deslocado);(S,adjacente)	= (d,a);(d,o) = (d;d) x (a;o) = (d) x {a,ob,d} = (d,a) _ (d,ob)	
(N,deslocado);(SO,adjacente)	= (d,a);(o,o) = (d;o) x (a;o) = {b,o,d} x {a,ob,d} = (b,a) _ (b,d) _ (o,ob) _ (o,d) _ (d,a) _ (d,ob)	
(N,deslocado);(O,adjacente)	= (d,a);(o,d) = (d;o) x (a;d) = {b,o,d} x {a,ob,d} = (b,a) _ (b,d) _ (o,ob) _ (o,d) _ (d,a) _ (d,ob)	
(N,deslocado);(NO,adjacente)	= (d,a);(o,ob) = (d;o) x (a;ob) = {b,o,d} x (a) = (b,a) _ (d,a)	

Tabela B.4: Integração da direcção Norte com o par topológico deslocado;adjacente

Composição		Resultado
(NE,deslocado);(E,adjacente)	= (a,a);(ob,d) = (a;ob) x (a;d) = (a) x {a,ob,d} = (a,a) _ (a,d)	
(NE,deslocado);(SE,adjacente)	= (a,a);(ob,o) = (a;ob) x (a;o) = (a) x {a,ob,d} = (a,a) _ (a,d)	
(NE,deslocado);(S,adjacente)	= (a,a);(d,o) = (a;d) x (a;o) = {a,ob,d} x {a,ob,d} = (a,a) _ (a,d) _ (ob,ob) _ (ob,d) _ (d,a) _ (d,ob)	
(NE,deslocado);(SO,adjacente)	= (a,a);(o,o) = (a;o) x (a;o) = {a,ob,d} x {a,ob,d} = (a,a) _ (a,d) _ (ob,ob) _ (ob,d) _ (d,a) _ (d,ob)	
(NE,deslocado);(O,adjacente)	= (a,a);(o,d) = (a;o) x (a;d) = {a,ob,d} x {a,ob,d} = (a,a) _ (a,d) _ (ob,ob) _ (ob,d) _ (d,a) _ (d,ob)	
(NE,deslocado);(NO,adjacente)	= (a,a);(o,ob) = (a;o) x (a;ob) = {a,ob,d} x (a) = (a,a) _ (d,a)	

Tabela B.5: Integração da direcção Nordeste com o par topológico deslocado;adjacente

B.1.3 Par topológico adjacente;deslocado

A Tabela B.6 apresenta o conjunto das regras que permitem a composição da direcção Norte com as restantes direcções, para o caso do par topológico adjacente;deslocado. A Tabela B.7 apresenta a integração da direcção Nordeste com as restantes direcções, para o mesmo par topológico. Os restantes casos são obtidos por rotação das respectivas direcções, como pode ser veri…cado na tabela …nal apresentada na Tabela B.11.

Composição		Resultado
(N,adjacente);(N,deslocado)	= (d,ob);(d,a) = (d;d) x (ob;a) = (d,a)	
(N,adjacente);(NE,deslocado)	= (d,ob);(a,a) = (d;a) x (ob;a) = (a,a)	
(N,adjacente);(E,deslocado)	= (d,ob);(a,d) = (d;a) x (ob;d) = (a) x {ob,d,f} = (a,d) _ (a,ob)	
(N,adjacente);(SE,deslocado)	= (d,ob);(a,b) = (d;a) x (ob;b) = (a) x {b,o,m,di,fb} = (a,b) _ (a,o) _ (a,di)	
(N,adjacente);(S,deslocado)	= (d,ob);(d,b) = (d;d) x (ob;b) = (d) x {b,o,m,di,fb} = (d,b) _ (d,o)	
(N,adjacente);(SO,deslocado)	= (d,ob);(b,b) = (d;b) x (ob;b) = (b) x {b,o,m,di,fb} = (b,b) _ (b,o) _ (b,di)	
(N,adjacente);(O,deslocado)	= (d,ob);(b,d) = (d;b) x (ob;d) = (b) x {ob,d,f} = (b,d) _ (b,ob)	
(N,adjacente);(NO,deslocado)	= (d,ob);(b,a) = (d;b) x (ob;a) = (b,a)	

Tabela B.6: Integração da direcção Norte com o par topológico adjacente;deslocado

Composição		Resultado
(NE,adjacente);(E,deslocado)	= (ob,ob);(a,d) = (ob;a) x (ob;d) = (a) x {ob,d,f} = (a,d) _ (a,ob)	
(NE,adjacente);(SE,deslocado)	= (ob,ob);(a,b) = (ob;a) x (ob;b) = (a) x {b,o,m,di,fb} = (a,b) _ (a,o) _ (a,di)	
(NE,adjacente);(S,deslocado)	= (ob,ob);(d,b) = (ob;d) x (ob;b) = {ob,d,f} x {b,o,m,di,fb} = (ob,o) _ (ob,b) _ (ob,di) _ (d,b) _ (d,o)	
(NE,adjacente);(SO,deslocado)	= (ob,ob);(b,b) = (ob;b) x (ob;b) = {b,o,di,m,fb} x {b,o,m,di,fb} = (b,b) _ (b,o) _ (b,di) _ (o,b) _ _ (o,o) _ (o,di) _ (di,b) _ (di,o)	
(NE,adjacente);(O,deslocado)	= (ob,ob);(b,d) = (ob;b) x (ob;d) = {b,o,m,di,fb} x {ob,d,f} = (b,d) _ (b,ob) _ (o,ob) _ (o,d) _ (di,ob)	
(NE,adjacente);(NO,deslocado)	= (ob,ob);(b,a) = (ob;b) x (ob;a) = {b,o,di,m,fb} x (a) = (b,a) _ (di,a) _ (o,a)	

Tabela B.7: Integração da direcção Nordeste com o par topológico adjacente;deslocado

B.1.4 Par topológico adjacente;adjacente

No caso particular de regiões adjacentes, duas situações podem ocorrer:

- as regiões envolvidas estão muito próximas uma da outra (mp;mp), sendo adjacente a relação topológica resultante;
- a direcção é integrada com um dos pares mp;p ou p;mp, podendo a relação topológica resultante ser deslocado ou adjacente.

Estas duas situações sugerem que se veri…que o conjunto de resultados que faz sentido num dado contexto. Esta análise é efectuada nas próximas subsecções, nas quais se determina ainda, a tabela de inferências a utilizar num e noutro caso.

Integração da (direcção, topologia) com os pares (mp;p) ou (p;mp)

Neste caso especí…co não se veri…ca qualquer alteração na de…nição das primitivas temporais a utilizar na composição, ou ainda na interpretação dos resultados encontrados com as mesmas. Assim, a construção das regras de inferência, para este caso particular, será efectuada como até aqui tem vindo a ser realizado.

A Tabela B.8 apresenta o conjunto das regras que permitem a composição da direcção Norte com as restantes direcções, para o caso do par topológico adjacente;adjacente. A Tabela B.9 apresenta a integração da direcção Nordeste com as restantes direcções, para o mesmo par topológico. Os restantes casos são obtidos por rotação das respectivas direcções, como pode ser veri…cado na tabela …nal apresentada na Tabela B.11.

Composição		Resultado
(N,adjacente);(N,adjacente)	= (d,ob);(d,ob) = (d;d) x (ob,ob) = (d) x {a,ob,mb} = (d,a) _ (d,ob)	
(N,adjacente);(NE,adjacente)	= (d,ob);(ob,ob) = (d;ob) x (ob,ob) = {a,ob,mb,d,f} x {a,ob,mb} = (a,a) _ (ob,ob) _ (d,a) _ (d,ob)	
(N,adjacente);(E,adjacente)	= (d,ob);(ob,d) = (d;ob) x (ob;d) = {a,ob,mb,d,f} x {ob,d,f} = (a,ob) _ (a,d) _ (ob,ob) _ (ob,d) _ (d,ob)	
(N,adjacente);(SE,adjacente)	= (d,ob);(ob,o) = (d;ob) x (ob;o) = {a,ob,mb,d,f} x {o,ob,d} = (a,d) _ (ob,ob) _ (ob,o) _ (a,o) _ (ob,d) _ (d,o) _ (d,ob) _ (a,ob)	
(N,adjacente);(S,adjacente)	= (d,ob);(d,o) = (d;d) x (ob;o) = (d) x {o,ob,d,di}	
(N,adjacente);(SO,adjacente)	= (d,ob);(o,o) = (d;o) x (ob;o) = {b,o,d} x {o,ob,d} = (b,o) _ (b,ob) _ (b,d) _ (o,o) _ (o,ob) _ (o,d) _ (o,di) _ (d,o) _ (d,ob) _ (b,di)	
(N,adjacente);(O,adjacente)	= (d,ob);(o,d) = (d;o) x (ob;d) = (b,o,d) x {ob,d} = (b,ob) _ (b,d) _ (o,ob) _ (o,d) _ (d,ob)	
(N,adjacente);(NO,adjacente)	= (d,ob);(o,ob) = (d;o) x (ob;ob) = {b,o,d} x {a,ob} = (b,a) _ (o,ob) _ (d,a) _ (d,ob)	

Tabela B.8: Integração da direcção Norte com o par topológico adjacente;adjacente para o caso das distâncias mp;p ou p;mp

Composição		Resultado
(NE,adjacente);(E,adjacente)	= (ob,ob);(ob,d) = (ob;ob) x (ob;d) = {a,ob,mb} x {ob,d,f} = (a,ob) _ (a,d) _ (ob,ob) _ (ob,d)	
(NE,adjacente);(SE,adjacente)	= (ob,ob);(ob,o) = (ob;ob) x (ob;o) = {a,ob,mb} x {o,ob,d,s,f,di,sb,fb,e} = (a,d) _ (ob,o) _ (ob,ob) _ (ob,d) _ (a,ob) _ (a,o) _ (a,di) _ (ob,di)	
(NE,adjacente);(S,adjacente)	= (ob,ob);(d,o) = (ob;d) x (ob;o) = {ob,d,f} x {o,ob,d,s,f,di,sb,fb,e} = (ob,o) _ (ob,ob) _ (ob,d) _ (d,o) _ (d,ob)_ (ob,di)	
(NE,adjacente);(SO,adjacente)	= (ob,ob);(o,o) = (ob;o) x (ob;o) = {o,ob,d,s,f,di,sb,fb,e} x {o,ob,d,s,f,di,sb,fb,e}	
(NE,adjacente);(O,adjacente)	= (ob,ob);(o,d) = (ob;o) x (ob;d) = {o,ob,d,s,f,di,sb,fb,e} x {ob,d,f} = (o,ob) _ (o,d) _ (ob,ob) _ (ob,d) _ (d,ob) _ (di,ob)	
(NE,adjacente);(NO,adjacente)	= (ob,ob);(o,ob) = (ob;o) x (ob;ob) = {o,ob,d,s,f,di,sb,fb,e} x {a,ob,mb} = (o,ob) _ (ob,ob) _ (d,a) _ (d,ob) (o,a) _ (ob,a) _ (di,a) _ (di,ob)	

Tabela B.9: Integração da direcção Nordeste com o par topológico adjacente;adjacente para o caso das distâncias mp;p ou p;mp

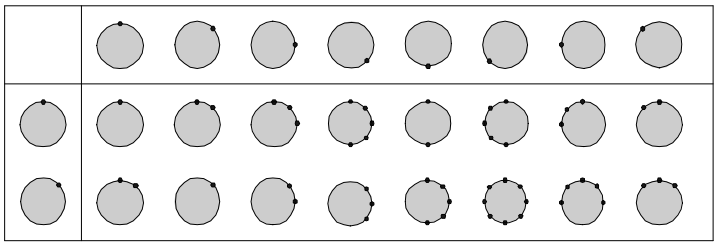


Tabela B.10: Integração da direcção e topologia para o caso particular mp;mp

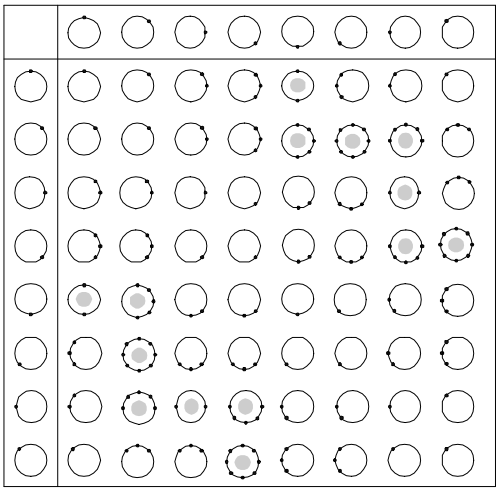
Integração da (direcção, topologia) para o caso particular mp;mp

Sempre que as regiões envolvidas no processo de inferência estão catalogadas como mp em termos de distância, e sendo um facto que a composição de mp;mp dá como resultado mp (seja esta relação considerada separadamente ou integrada com a direcção), o único resultado possível para a inferência da topologia é a primitiva adjacente, dada a distância que existe entre as regiões.

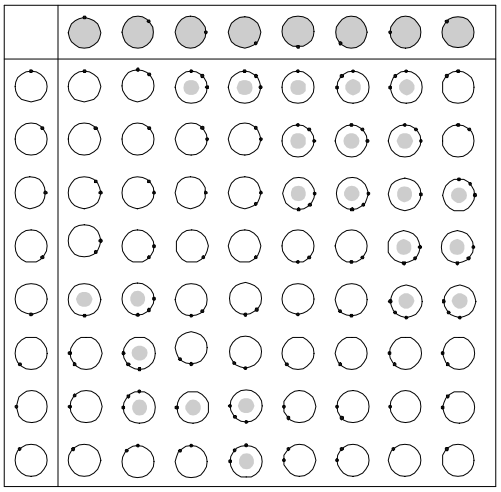
Este facto obriga a analisar as tabelas de composição apresentadas nas Tabelas B.8 e B.9, e a limitar o resultado da relação topológica a adjacente. A tabela de composição a adoptar neste caso particular é a apresentada na Tabela B.10 (evidenciando apenas as direcções N e NE. As restantes são obtidas por rotação das apresentadas).

B.1.5 Síntese

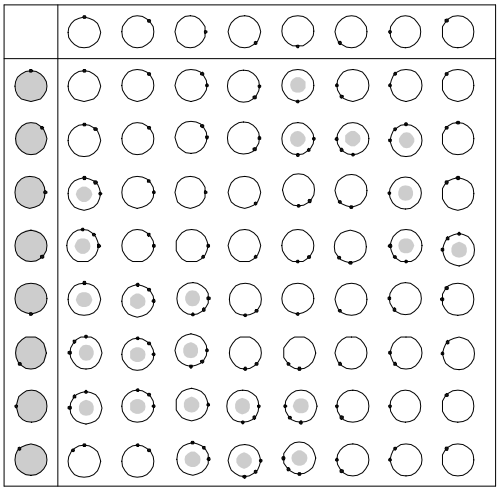
Nas secções anteriores apresentaram-se os princípios que ditaram a construção das regras de inferência para a integração de dois tipos de relações espaciais, direcção e topologia. A tabela B.11 sistematiza o conhecimento obtido, através da apresentação das tabelas de composição que permitem a inferência de relações espaciais desconhecidas.



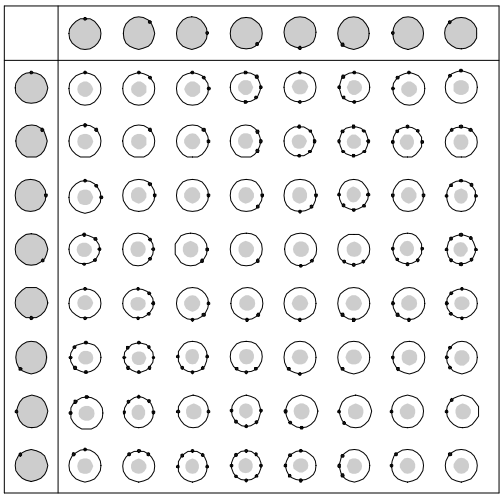
a) deslocado;deslocado



b) deslocado;adjacente



c) adjacente;deslocado



d) adjacente;adjacente

Tabela B.11: Tabelas de composição para a inferência integrada de relações espaciais do tipo direcção e topologia

B.2 Integração da direcção, distância e topologia

Após a construção das tabelas de composição que ditam a integração da direcção e topologia, é necessário proceder a integração das mesmas com a tabela que possui as regras de inferência para a integração da direcção e distância (recorda-se, como referido anteriormente no Capítulo 3, que é utilizado o `ratio` 4 entre distâncias). O processo de integração foi já descrito no Capítulo 3, secção 3.5.3, apresentando a Figura B.3, Figura B.4 e a Figura B.5 a tabela resultante de dita integração (dividida em três partes, para facilitar a sua visualização, conforme o esquema apresentado na Figura B.2).

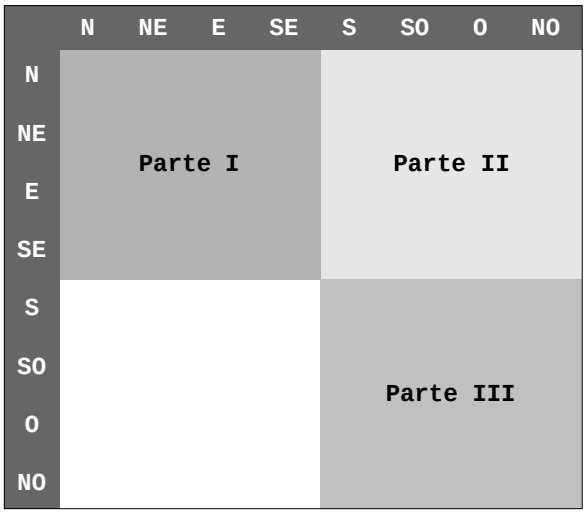


Figura B.2: Esquema de apresentação das três partes da tabela de composição

Destaca-se que dada a ‡exibilidade imposta ao sistema de inferências, materializada na selecção das primitivas temporais que caracterizam a integração da direcção e topologia, veri…ca-se que o sistema obtido é bastante ‡exível, uma vez que permite caracterizar situações "complicadas" num dado mapa. Estas situações são normalmente originadas pela não uniformidade existente entre a dimensão das regiões que caracterizam o domínio geográ…co considerado.

Novas restrições podem ser colocadas ao sistema de raciocínio, com o objectivo de o tornar cada vez mais exacto. Tal poderia ser conseguido incorporando, por exemplo, a dimensão dos objectos na análise.

Após a integração das três relações espaciais, veri…cou-se que ocorreram duas situações nas quais o conjunto de direcções inferidas como resposta na integração da direcção e topologia não coincidia com a direcção inferida através da integração da direcção e distância. Para estes casos, optou-se por considerar a direcção inferida por esta última, uma vez que à partida tem mais probabilidades de ser a correcta (por considerar a distância existente entre os objectos). Ambas as situações ocorreram na composição de dois factos com a distância qualitativa `próximo`. O primeiro veri…cou-se na integração do par topológico `deslocado;adjacente` com as direcções `NE;S`, enquanto que o segundo ocorre na integração do par topológico `adjacente;deslocado`

com as direcções N;SO. Identi…cadas estas situações, a estratégia é a de prosseguir com a tabela de composição, tal qual como resultou do processo de integração, implementá-la, e veri…car posteriormente o seu desempenho.

Após a avaliação veri…car-se-á se existe ou não a necessidade de reformular o sistema, e se sim, em que moldes é que tal reformulação deverá acontecer.

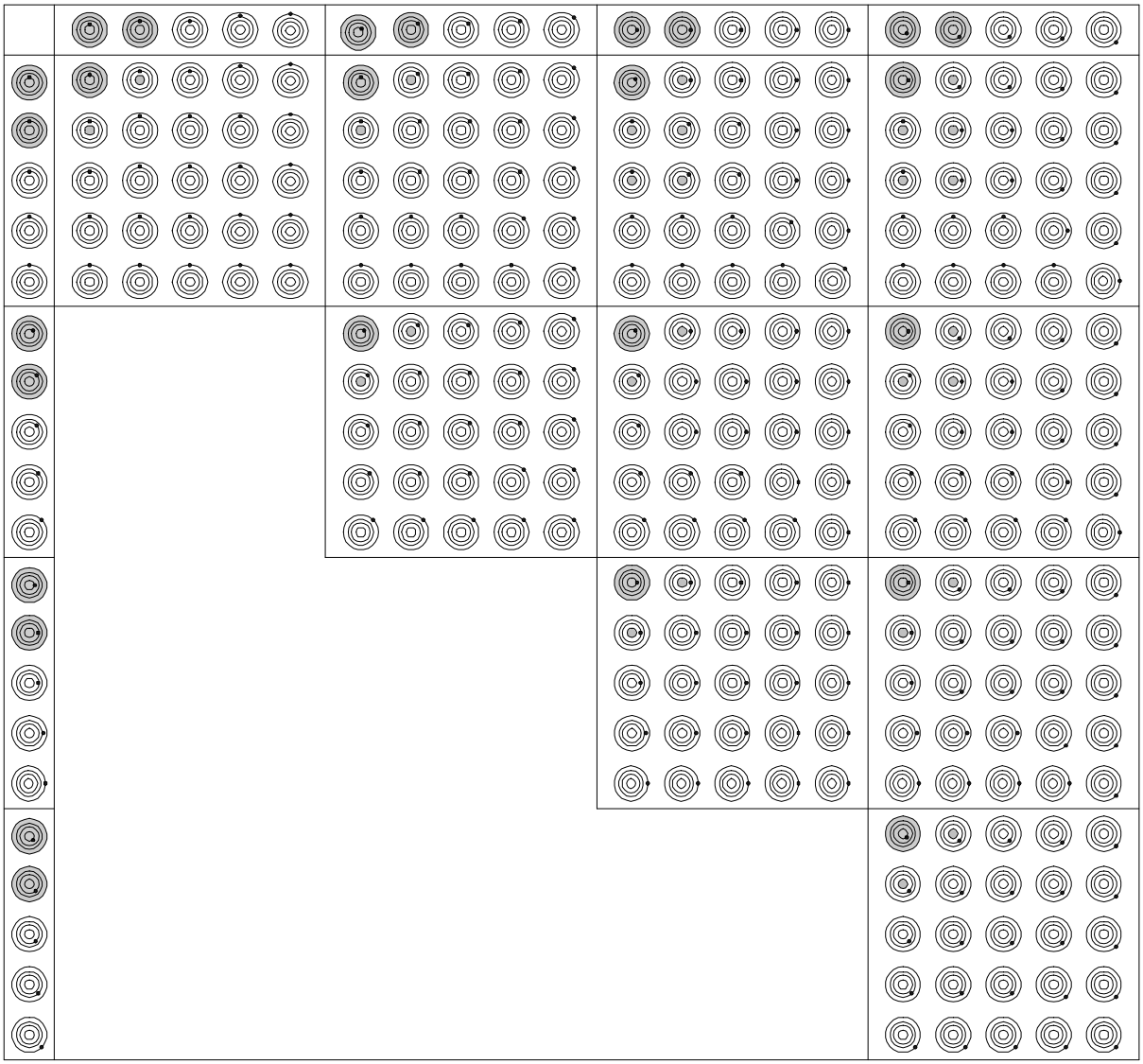


Figura B.3: Tabela de composição que integra a direcção, distância e topologia (Parte I)

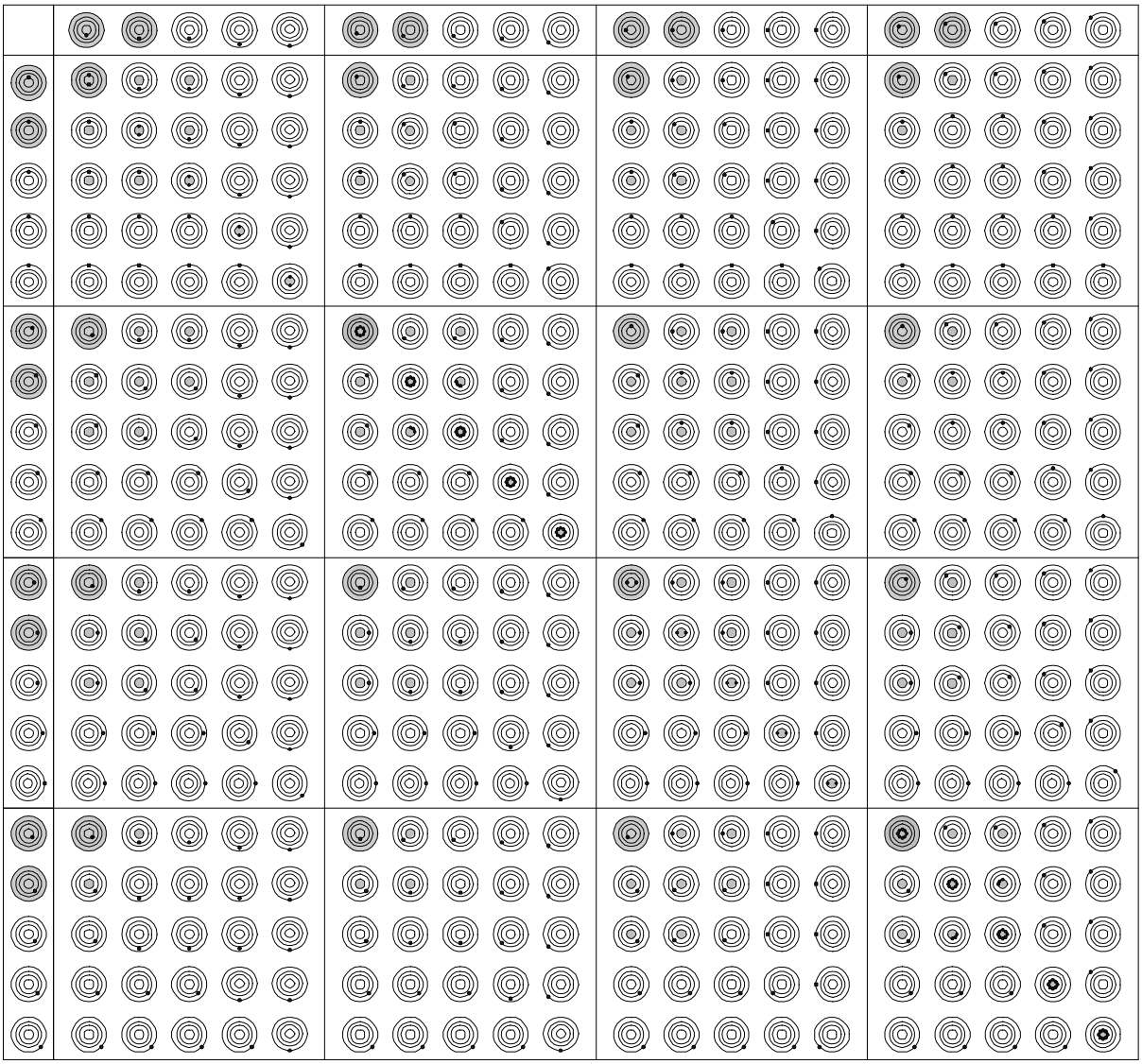


Figura B.4: Tabela de composição que integra a direcção, distância e topologia (Parte II)

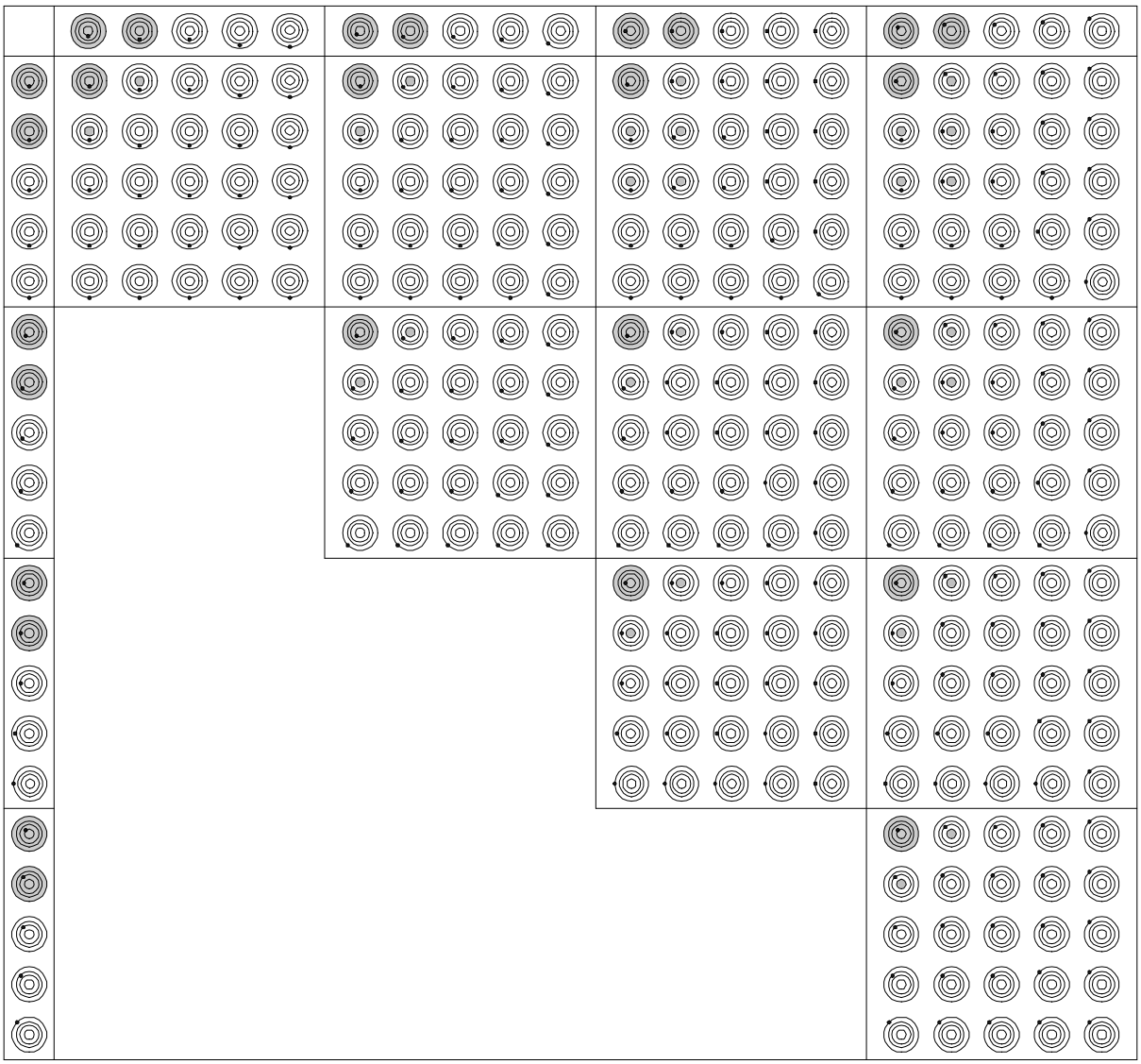


Figura B.5: Tabela de composição que integra a direcção, distância e topologia (Parte III)

Apêndice C

Módulos em Visual Basic

Neste apêndice encontram-se as listagens com o código VB que integra os vários módulos implementados para auxiliar o carregamento da BDG. É ainda incluído o código dos dois módulos externos integrados no Clementine, o Combina e o VisualPadrão.

C.1 Módulo AssociaCentróide

```
Attribute VB_Name = "Module3"
    Global distancia As Double, angulo As Double
    Global objRecord1 As GRecordset, objRecord2 As GRecordset
Sub main()
    AssociaCentroide
End Sub
Sub AssociaCentroide()
Dim objConnet1 As New Connection
Dim objGeometry1 As GeometryStorageService
Dim objOPipe1 As OriginatingPipe
Dim objGeom1 As Object
Dim objResult As GRecordset, objPoint As GRecordset
Dim objSpaFil As New SpatialFilter
Dim geomBlob As Variant
Dim objGSS As New GeometryStorageService
Dim objConResult As New Connection
Dim objDB As GDatabase
' Ligação à Base de Dados dos Concelhos, tabela de Limites e
' carregamento das faces que constituem os concelhos para objRecord1
With objConnet1
    .Location = "C:nWarehousesnBD_conc.mdb"
    .Mode = gmcModeReadOnly
    .Type = "Access.GDatabase"
    .ConnectionName = "Connet1"
```

```
    .Connect
End With
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "SpatialArea"
    .Table = "Limites"
End With
Set objRecord1 = objOPipe1.OutputRecordset
Set objOPipe1 = Nothing
objRecord1.MoveLast
objRecord1.MoveFirst
' Abertura da BD_Geogra...ca para armazenar a informação que vai sendo obtida
With objConResult
    .Location = "D:nMaribelnDoutoramentonBasesDadosnBD_Geogra...ca.mdb"
    .Mode = gmcModeReadWrite
    .Type = "Access.GDatabase"
    .ConnectionName = "ConResult"
    .Connect
End With
objConResult.CreateOriginatingPipe objOPipe1
With objOPipe1
    .Table = "NodosIsolados"
End With
Set objDB = CreateObject("Access.GDatabase")
objDB.OpenDatabase "D:nMaribelnDoutoramentonBasesDadosnBD_Geogra...ca.mdb"
Set objResult = objDB.OpenRecordset("NodosIsolados", gdbOpenDynaset)
If Not (objResult.EOF) Then
    objResult.MoveLast
End If
Do Until objRecord1.EOF
    ' Carregamento da geometria da face para objGeom1
    Set objGeometry1 = CreateObject("GeoMedia.GeometryStorageService")
    objGeometry1.GetGeometry objRecord1.GFields("SpatialArea"), objGeom1
    Set objSpaFil.Geometry = objGeom1
    With objConnet1
        Set SpatialFilter = objSpaFil
    End With
    Set objGeom1 = Nothing
    Set objGeometry1 = Nothing
    ' conversão da geometria do ...ltro da face para geomBlob
    objGSS.GeometryToStorage objSpaFil.Geometry, geomBlob
    Set objGSS = Nothing
    Set objSpaFil = Nothing
    objConnet1.CreateOriginatingPipe objOPipe1
    With objOPipe1
```

```
            .GeometryFieldName = "SpatialPoint"
            .Table = "Centroides"
            .SpatialFilter = geomBlob
            .SpatialOperator = gmsqContains
        End With
        Set geomBlob = Nothing
        Set objPoint = objOPipe1.OutputRecordset
        Set objOPipe1 = Nothing
        If Not objPoint.EOF Then
            objPoint.MoveFirst
            objResult.AddNew 'Upgrade tabela NodosIsolados
            With objResult
                .GFields(0).Value = objPoint(1).Value 'ID ponto
                .GFields(1).Value = objRecord1(4).Value 'ID face
            End With
            objResult.Update
        End If
        Set geomBlob = Nothing
        Set objConnet1.SpatialFilter = Nothing
        objRecord1.MoveNext
Loop
Set objRecord1 = Nothing
Set objConnet1 = Nothing
Set objResult = Nothing
End Sub
```

C.2 Módulo DetAdjacentes

```
Attribute VB_Name = "Module3"
    Global distancia As Double, angulo As Double
    Global objRecord1 As GRecordset, objRecord2 As GRecordset
    Global objConnet1 As New Connection
Sub main()
    DetAdjacentes
End Sub
Sub DetAdjacentes()
Dim objGeometry1 As GeometryStorageService
Dim objGeometry2 As GeometryStorageService
Dim objOPipe1 As OriginatingPipe, objOPipe2 As OriginatingPipe
Dim objGeom1 As Object
Dim objResult As GRecordset
Dim objSpaFil As New SpatialFilter
Dim geomBlob As Variant
Dim objGSS As New GeometryStorageService
```

```
Dim objConResult As New Connection
Dim objDB As GDatabase
' Ligação à Base de Dados dos Concelhos, tabela de Limites e
' carregamento das faces que constituem os concelhos para objRecord1
With objConnet1
    .Location = "C:nWarehousesnBD_conc.mdb"
    .Mode = gmcModeReadOnly
    .Type = "Access.GDatabase"
    .ConnectionName = "Connet1"
    .Connect
End With
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "SpatialArea"
    .Table = "Limites"
End With
Set objRecord1 = objOPipe1.OutputRecordset
Set objOPipe1 = Nothing
objRecord1.MoveLast
objRecord1.MoveFirst
' Abertura da BD_Geogra…ca para armazenar a informação que vai sendo obtida
With objConResult
    .Location = "D:nMaribelnDoutoramentonBasesDadosnBD_Geogra…ca.mdb"
    .Mode = gmcModeReadWrite
    .Type = "Access.GDatabase"
    .ConnectionName = "ConResult"
    .Connect
End With
objConResult.CreateOriginatingPipe objOPipe1
With objOPipe1
    .Table = "Faces"
End With
Set objDB = CreateObject("Access.GDatabase")
objDB.OpenDatabase "D:nMaribelnDoutoramentonBasesDadosnBD_Geogra…ca.mdb"
Set objResult = objDB.OpenRecordset("Faces", gdbOpenDynaset)
If Not (objResult.EOF) Then
    objResult.MoveLast
End If
Do Until objRecord1.EOF
    ' Carregamento da geometria para objGeom1
    Set objGeometry1 = CreateObject("GeoMedia.GeometryStorageService")
    objGeometry1.GetGeometry objRecord1.GFields("SpatialArea"), objGeom1
    Set objSpaFil.Geometry = objGeom1
    Set objGeom1 = Nothing
    Set objGeometry1 = Nothing
```

```
    ' conversão da geometria do ...ltro para geomBlob
    objGSS.GeometryToStorage objSpaFil.Geometry, geomBlob
    Set objGSS = Nothing
    Set objSpaFil = Nothing
    ' determinação do conjunto de registos que satisfazem a condição
    objConnet1.CreateOriginatingPipe objOPipe2
    With objOPipe2
        .GeometryFieldName = "SpatialArea"
        .Table = "Limites"
        .SpatialFilter = geomBlob
        .SpatialOperator = gdbTouches
    End With
    ' carregamento dos registos para objRecord2
    Set objRecord2 = objOPipe2.OutputRecordset
    Set objOPipe2 = Nothing
    objRecord2.MoveLast
    objRecord2.MoveFirst
    ' armazenar todos os concelhos adjacentes ao concelho analisado
    If Not (objRecord2.EOF And objRecord2.BOF) Then
        Do Until objRecord2.EOF
            If Not (objRecord1.GFields(4).Value = objRecord2.GFields(4).Value)
                And Not objRecord2.EOF And Not objRecord1.EOF Then
                objResult.AddNew
                Call DetCoordenadasCentroides
                With objResult
                    .GFields(0).Value = objRecord1(4).Value
                    .GFields(1).Value = objRecord2(4).Value
                    .GFields(2).Value = CInt(angulo)
                    .GFields(3).Value = CInt(distancia)
                End With
                objResult.Update
            End If
        objRecord2.MoveNext
        Loop
    Else
        MsgBox ("Sem adjacentes")
    End If
    Set objRecord2 = Nothing
    Set geomBlob = Nothing
    objRecord1.MoveNext
Loop
Set objRecord1 = Nothing
Set objConnet1 = Nothing
Set objConResult = Nothing
Set objResult = Nothing
```

```
Set objOPipe1 = Nothing
End Sub
```

Procedimento DetCoordenadasCentróides

```
Sub DetCoordenadasCentroides()
Dim objDB As GDatabase
Dim objGeometry1 As GeometryStorageService
Dim objPoint As New GeometryStorageService
Dim objGSS As New GeometryStorageService
Dim objRecord As GRecordset, objRec1 As GRecordset, objRec2 As GRecordset
Dim objResult As GRecordset
Dim objOPipe1 As OriginatingPipe, objOPipe2 As OriginatingPipe
Dim objGeom1 As Object, objGeom2 As Object, objGeomPnt As Object
Dim objPntLow1 As New Point, objPntHigh1 As New Point
Dim objPntLow2 As New Point, objPntHigh2 As New Point
Dim objMeas As MeasurementService
Dim num As Integer, i As Integer
Dim face As String, face_adj As String
Dim geomBlob As Variant
Dim objSpaFil As New SpatialFilter
Dim objCoord As GRecordset
face = objRecord1(4).Value
face_adj = objRecord2(4).Value
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "SpatialArea"
    .Table = "Limites"
    .Filter = "codconc = "' & face & ""'
End With
objConnet1.CreateOriginatingPipe objOPipe2
With objOPipe2
    .GeometryFieldName = "SpatialArea"
    .Table = "Limites"
    .Filter = "codconc = "' & face_adj & ""'
End With
Set objRec1 = objOPipe1.OutputRecordset
Set objRec2 = objOPipe2.OutputRecordset
Set objOPipe1 = Nothing
Set objOPipe2 = Nothing
objRec1.MoveFirst
objRec2.MoveFirst
' Geometria da face e identi…cacao respectivo centroide
Set objGeometry1 = CreateObject("GeoMedia.GeometryStorageService")
objGeometry1.GetGeometry objRec1.GFields("SpatialArea"), objGeom1
Set objSpaFil.Geometry = objGeom1
```

```
Set objGeometry1 = Nothing
Set objGeom1 = Nothing
Set objRec1 = Nothing
objGSS.GeometryToStorage objSpaFil.Geometry, geomBlob
Set objGSS = Nothing
Set objSpaFil = Nothing
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "Geometry1"
    .Table = "PontosCentroides"
    .SpatialFilter = geomBlob
    .SpatialOperator = gmsqContains
End With
Set geomBlob = Nothing
Set objResult = objOPipe1.OutputRecordset
Set objOPipe1 = Nothing
If Not objResult.EOF Then
    objResult.MoveFirst
    Set objPoint = CreateObject("GeoMedia.GeometryStorageService")
    objPoint.GetGeometry objResult.GFields("Geometry1"), objGeomPnt
    GetRange objGeomPnt, objPntLow1, objPntHigh1
    Set objPoint = Nothing
    Set objGeomPnt = Nothing
    ' Geometria da segunda face e respectivo centroide
    Set objGeometry1 = CreateObject("GeoMedia.GeometryStorageService")
    objGeometry1.GetGeometry objRec2.GFields("SpatialArea"), objGeom1
    Set objSpaFil.Geometry = objGeom1
    Set objGeometry1 = Nothing
    Set objGeom1 = Nothing
    Set objRec2 = Nothing
    objGSS.GeometryToStorage objSpaFil.Geometry, geomBlob
    Set objGSS = Nothing
    Set objSpaFil = Nothing
    objConnet1.CreateOriginatingPipe objOPipe1
    With objOPipe1
        .GeometryFieldName = "Geometry1"
        .Table = "PontosCentroides"
        .SpatialFilter = geomBlob
        .SpatialOperator = gmsqContains
    End With
    Set geomBlob = Nothing
    Set objResult = objOPipe1.OutputRecordset
    Set objOPipe1 = Nothing
    If Not objResult.EOF Then
        objResult.MoveFirst
```

```
Set objPoint = CreateObject("GeoMedia.GeometryStorageService")
objPoint.GetGeometry objResult.GFields("Geometry1"), objGeomPnt
GetRange objGeomPnt, objPntLow2, objPntHigh2
Set objPoint = Nothing
Set objGeomPnt = Nothing
' Calculo da distância e ângulo existente entre os centroides
Dim linha As New LineGeometry
Dim coordX As Double, coordY As Double
Dim PI As Double
angulo = 0
distancia = 0
PI = 4 * Atn(1)
linha.Start.X = objPntLow1.X
linha.Start.Y = objPntLow1.Y
linha.Start.Z = 0
linha.End.X = objPntHigh2.X
linha.End.Y = objPntHigh2.Y
linha.End.Z = 0
coordX = (linha.Start.X - linha.End.X) / 100
coordY = (linha.Start.Y - linha.End.Y) / 100
distancia = Sqr(coordX * coordX + coordY * coordY) / 1000
If coordX <> coordY Then
    angulo = Atn(coordX / coordY) * (180 / PI)
End If
' veri...car quadrante em que esta localizado este angulo
If angulo > 0 Then
    If coordX > 0 And coordY < 0 Then
        angulo = 90 + angulo
    Else
        If coordX < 0 And coordY < 0 Then
            angulo = 180 + angulo
        ElseIf coordX < 0 And coordY > 0 Then
            angulo = 270 + angulo
        End If
    End If
End If
If angulo < 0 Then
    If coordX > 0 And coordY < 0 Then
        angulo = 180 + angulo
    Else
        If coordX < 0 And coordY < 0 Then
            angulo = 270 + angulo
        Else
            If coordX < 0 And coordY > 0 Then
                angulo = 360 + angulo
```

```
                        ElseIf coordX > 0 And coordY > 0 Then
                            angulo = 90 + angulo
                        End If
                    End If
                End If
            End If
        End If
Else
    angulo = 0
    distancia = 0
End If
Set objPntLow1 = Nothing
Set objPntLow2 = Nothing
Set objPntHigh1 = Nothing
Set objPntHigh2 = Nothing
Set linha = Nothing
End Sub
```

C.3 Módulo CalculoCentróides

```
Sub CalculoCentroides()
Dim objGeom As Object
Dim objGeometry As GeometryStorageService
Dim objRecord As GRecordset, objResult As GRecordset
Dim objConnet As New Connection
Dim objOPipe As OriginatingPipe
Dim objCentroid As CenterPointPipe
Dim objPntLow As New Point, objPntHigh As New Point
Dim objConResult As New Connection
Dim objResultTransf As GRecordset
Dim objDB As GDatabase
' Abertura da BD e respectiva tabela com os limites dos concelhos
With objConnet
    .Location = "C:nWarehousesnBD_conc.mdb"
    .Type = "Access.GDatabase"
    .Connect
End With
objConnet.CreateOriginatingPipe objOPipe
With objOPipe
    .GeometryFieldName = "SpatialArea"
    .Table = "Limites"
End With
Set objRecord = objOPipe.OutputRecordset
Set objDB = CreateObject("Access.GDatabase")
```

```
objDB.OpenDatabase "d:nmaribeIndoutoramentonBasesDadosnBD_Geogra…ca.mdb"
Set objResultTransf = objDB.OpenRecordset("PontosCentroides", gdbOpenDynaset)
If Not objResultTransf.EOF Then
    objResultTransf.MoveLast
End If
Set objCentroid = CreateObject("GeoMedia.CenterPointPipe")
With objCentroid
    Set .InputRecordset = objOPipe.OutputRecordset
    .InputGeometryFieldName = objOPipe.GeometryFieldName
    .OutputGeometryFieldName = "Centroid"
End With
Set objResult = objCentroid.OutputRecordset
If Not objResult.EOF Then
    objResult.MoveFirst
End If
Do Until objResult.EOF
    ' Determinação da geometria do centroide e respectivas coordenadas
    Set objGeometry = CreateObject("GeoMedia.GeometryStorageService")
    objGeometry.GetGeometry objResult.GFields("Centroid"), objGeom
    GetRange objGeom, objPntLow, objPntHigh
    ' Armazenamento da informação na base de dados
    objResultTransf.AddNew
    With objResultTransf
        .GFields(0) = objResult.GFields("ID")
        .GFields(1) = objResult.GFields("codconc")
        .GFields(2) = (objPntLow.X) / 100
        .GFields(3) = (objPntHigh.Y) / 100
        .GFields(4) = objResult.GFields("Centroid")
    End With
    objResultTransf.Update
    objResult.MoveNext
Loop
End Sub
```

C.4 Módulo Combina

```
Sub Main()
Dim r1(1 To 4) As String
Dim r2(1 To 4) As String
Dim dir(1 To 4) As String
Dim dist(1 To 4) As String
Dim topo(1 To 4) As String
Dim enc As Integer
' carregamento da explicitacao semantica dos concelhos
```

```
Open "Regions" For Input As #1
Open "Regions.res" For Output As #2
Do While Not EOF(1)
    Input #1, r1(1), r2(1), dir(1), dist(1), topo(1)
    Open "Regions" For Input As #4
    Input #4, r1(2), r2(2), dir(2), dist(2), topo(2)
    Do
        If r2(1) = r1(2) Then 'regioes adjacentes
            enc = 0
            Open "Regions" For Input As #3
            Do While enc = 0 And Not EOF(3)
                Input #3, r1(4), r2(4), dir(4), dist(4), topo(4)
                If r1(4) = r1(1) And r2(4) = r2(2) Then
                    enc = 1
                End If
            Loop
            If enc <> 1 And r1(1) <> r2(2) Then 'para nao inferir a mesma regiao
                Write #2, r1(1), r2(1), dir(1), dist(1), topo(1), r1(2), r2(2), dir(2), dist(2), topo(2)
            End If
            Close #3
        End If
        If Not EOF(4) Then
            Input #4, r1(2), r2(2), dir(2), dist(2), topo(2)
        End If
    Loop While Not EOF(4)
    Close #4
Loop
Close #2
Close #1
End Sub
```

C.5 Módulo VisualPadrão

```
Sub Main()
    Call VisualThematic
End Sub

Sub VisualThematic()
Dim GeoApp As GeoMedia.Application
Dim objDB1 As GDatabase, objDB2 As GDatabase
Dim objConnet As New Connection
Dim objOPipe As OriginatingPipe, objOPipe2 As OriginatingPipe
Dim objRLE As RecordLegendEntry, objRLE2 As RecordLegendEntry
Dim objDoc As Document
Dim LeftRS As GRecordset, RightRS As GRecordset, JoinRS As GRecordset
```

```
Dim tabTexto As GRecordset, textoRLE As RecordLegendEntry
Dim JoinPipe As New EquijoinPipe, JoinFields(1, 0) As String
Dim data As Long, descr As String, tabela As String

On Error GoTo Erros
Set GeoApp = GetObject(, "Geomedia.Application")
GeoApp.Visible = True

' criar novo GeoWorkspace
Set objDoc = GeoApp.New
GeoApp.ActiveWindow.WindowState = gmwMaximize
With objDoc
    .BackgroundColor = RGB(255, 255, 204)
    .HighlightColor = RGB(255, 0, 0)
    .HandleColor = RGB(100, 100, 100)
    .SelectColor = RGB(0, 0, 50)
End With

'ligar a BD dos concelhos
Set objDB1 = CreateObject("Access.GDatabase")
objDB1.OpenDatabase "C:nWarehousesnBD_conc.mdb"
Set LeftRS = objDB1.OpenRecordset("Limites", 4)
Set tabTexto = objDB1.OpenRecordset("Nomes", 4)

'join com a tabela de padroes
Set JoinPipe = CreateObject("Geomedia.EquijoinPipe")
JoinFields(0, 0) = "codconc"

'ligacao ao ...cheiro com indicação do Padrão a visualizar
Open "D:nMaribelnDoutoramentonClementinenPadrao" For Input As #1
Do While Not EOF(1)
    Input #1, data, descr, tabela
Loop

Set objDB2 = CreateObject("Access.GDatabase")
objDB2.OpenDatabase "D:nMaribelnDoutoramentonBasesDadosnBD_Padroes"
Set RightRS = objDB2.OpenRecordset(tabela, 4)
JoinFields(1, 0) = "idFace"

With JoinPipe
    Set .LeftRecordset = LeftRS
    Set .RightRecordset = RightRS
    .JoinFieldNames = JoinFields
    .JoinType = gmejpInner
End With
Set JoinRS = JoinPipe.OutputRecordset
JoinRS.MoveLast
JoinRS.MoveFirst

'estilos para o display de informacao
Dim objStyle As AreaStyle
Set objStyle = GeoApp.CreateService("GeoMedia.AreaStyle")
```

```
With objStyle
    .FillType = gmsFPTransparent
    .BackColor = RGB(255, 255, 255)
    .Boundary.Color = RGB(0, 0, 0)
    .Boundary.Width = 1
    .Boundary.LineStyle = gmsLinearSolid
    .BoundaryOn = True
    .StyleUnits = gmsStyleUnitsView
End With
'criar e de...nir a legenda do mapa
Set objRLE = GeoApp.CreateService("Geomedia.RecordLegendEntry")
With objRLE
    Set .Style = objStyle
    Set .Recordset = LeftRS
    .Title = "Limites Concelhos"
    .GeometryFieldName = "SpatialArea"
End With
Set textoRLE = GeoApp.CreateService("Geomedia.RecordLegendEntry")
With textoRLE
    Set .Style = objStyle
    Set .Recordset = tabTexto
    .Title = "Designação Concelhos"
    .GeometryFieldName = "GraphicText"
End With
Set objStyle = Nothing
'Display da legenda na Map View
Dim objMV As Object
Set objMV = GeoApp.ActiveWindow.MapView
objMV.Legend.Visible = True
'designação das regiões geográ...cas
If textoRLE.ValidateSource Then
    If objMV.Legend.LegendEntries.Count = 0 Then
        objMV.Legend.LegendEntries.Append textoRLE
    Else
        objMV.Legend.LegendEntries.Append textoRLE, , 1
    End If
    textoRLE.LoadData
End If
'cartogra...a
If objRLE.ValidateSource Then
    If objMV.Legend.LegendEntries.Count = 0 Then
        objMV.Legend.LegendEntries.Append objRLE
    Else
        objMV.Legend.LegendEntries.Append objRLE, , 1
    End If
```

```
    objRLE.LoadData
End If
'entradas para os diversos valores do atributo saida
Dim campos As Integer, i As Integer
Dim regFinal As GRecordset, tabDef As GTableDef, objField As GField
Dim legenda(10) As String, j As Integer, enc As Integer, num As Integer
Dim colors(10) As Long
colors(3) = RGB(188, 188, 65)
colors(4) = RGB(213, 204, 187)
colors(1) = RGB(255, 255, 65)
colors(2) = RGB(128, 180, 128)
colors(5) = RGB(144, 127, 97)
colors(6) = RGB(155, 155, 200)
colors(7) = RGB(200, 188, 65)
colors(8) = RGB(213, 200, 187)
colors(9) = RGB(255, 180, 65)
colors(10) = RGB(100, 180, 128)
JoinRS.MoveLast
JoinRS.MoveFirst
campos = JoinRS.GFields.Count
num = 0
enc = 0
Do While Not JoinRS.EOF
    If num = 0 Then
        num = num + 1
        legenda(num) = JoinRS.GFields(campos - 2).Value
    Else
        enc = 0
        For j = 1 To num
            If legenda(j) = JoinRS.GFields(campos - 2).Value Then
                enc = 1
            End If
        Next
        If enc = 0 Then
            num = num + 1
            legenda(num) = JoinRS.GFields(campos - 2).Value
        End If
    End If
    JoinRS.MoveNext
Loop
Set tabDef = objDB2.CreateTableDef("tabFinal")
tabDef.Name = "tabFinal"
For i = 0 To campos - 1
    Set objField = tabDef.CreateField(JoinRS.GFields(i).Name)
    With objField
```

```
            .Type = JoinRS.GFields(i).Type
            If JoinRS.GFields(i).Type = 32 Then
                .SubType = JoinRS.GFields(i).SubType 'no caso de atributos espaciais é necessário de…nir
subtipos
            End If
        End With
        tabDef.GFields.Append objField
        Set objField = Nothing
Next
objDB2.GTableDefs.Append tabDef
Set regFinal = objDB2.OpenRecordset("tabFinal", gdbOpenDynaset)
For j = 1 To num        'criar legendas para cada um dos valores de saída armazenados no array legenda
        JoinRS.MoveLast
        JoinRS.MoveFirst
        Do While Not JoinRS.EOF
            If legenda(j) = JoinRS.GFields(campos - 2).Value Then
                regFinal.AddNew
                For i = 0 To campos - 1
                    regFinal.GFields(i).Value = JoinRS.GFields(i).Value
                Next
                regFinal.Update
            End If
            JoinRS.MoveNext
        Loop
        regFinal.MoveLast
        regFinal.MoveFirst
        Set objStyle = GeoApp.CreateService("GeoMedia.AreaStyle")
        With objStyle
            .FillType = gmsFPSolid
            .BackColor = colors(j)
            .Boundary.Color = RGB(0, 0, 0)
            .Boundary.Width = 1
            .Boundary.LineStyle = gmsLinearSolid
            .BoundaryOn = True
        End With
        'criar e de…nir a legenda do padrao
        Set objRLE2 = GeoApp.CreateService("Geomedia.RecordLegendEntry")
        With objRLE2
            Set .Style = objStyle
            Set .Recordset = regFinal
            .Title = legenda(j)
            .GeometryFieldName = "SpatialArea"
        End With
        If objRLE2.ValidateSource Then
            If objMV.Legend.LegendEntries.Count = 0 Then
```

```
                objMV.Legend.LegendEntries.Append objRLE2
            Else
                objMV.Legend.LegendEntries.Append objRLE2, , 1
            End If
            objRLE2.LoadData
        End If
        regFinal.MoveFirst
        Do While Not regFinal.EOF
            regFinal.Delete
            regFinal.MoveNext
        Loop
Next
regFinal.Close
Set regFinal = Nothing
Set tabDef = Nothing
Set tabFinal = Nothing
Set regFinal = Nothing
objDB2.GTableDefs.Delete "tabFinal"
objDB2.Close
objDB1.Close
Set objDB2 = Nothing
Set JoinRS = Nothing
Set LeftRS = Nothing
Set RightRS = Nothing
Set objStyle = Nothing
Set JoinPipe = Nothing
objMV.Legend.Fit
objMV.Legend.Visible = True
objMV.Fit
objDoc.RefreshViews
If Not objDoc.Saved Then
    objDoc.SaveAs "c:nGeoWorkspacesnPadrao.gws"
End If
Set objRLE = Nothing
Set objRLE2 = Nothing
Set objMV = Nothing
Set textoRLE = Nothing
GoTo Fim
Erros:
    If Err.Number = 429 Then
        Err.Clear
        Set GeoApp = CreateObject("GeoMedia.Application")
        Resume Next
    Else
        GoTo Fim
```

```
    End If
Fim:
    Set objDoc = Nothing
    Set GeoApp = Nothing
    Close #1
End Sub
```

C.6 Módulo VerRelações

```
Attribute VB_Name = "Module1"
    Global distancia As Double, angulo As Double
    Global objRecord1 As GRecordset, objRecord2 As GRecordset
    Global objConnet1 As New Connection
Sub main()
    VerRelacoes
End Sub
Sub VerRelacoes()
Dim objOPipe1 As OriginatingPipe, objOPipe2 As OriginatingPipe
Dim objResult As GRecordset
Dim objConResult As New Connection
Dim objDB As GDatabase
' Ligação à Base de Dados dos Concelhos, tabela de LimitesAVR e
' carregamento das faces que constituem os concelhos para objRecord1
With objConnet1
    .Location = "C:nWarehousesnBD_conc.mdb"
    .Mode = gmcModeReadOnly
    .Type = "Access.GDatabase"
    .ConnectionName = "Connet1"
    .Connect
End With
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "SpatialArea"
    .Table = "LimitesAVR"
End With
objConnet1.CreateOriginatingPipe objOPipe2
With objOPipe2
    .GeometryFieldName = "SpatialArea"
    .Table = "LimitesAVR"
End With
Set objRecord1 = objOPipe1.OutputRecordset
Set objRecord2 = objOPipe2.OutputRecordset
Set objOPipe1 = Nothing
Set objOPipe2 = Nothing
```

```
objRecord1.MoveLast
objRecord1.MoveFirst
objRecord2.MoveLast
objRecord2.MoveFirst
' Abertura da BD_Geogra…ca para armazenar a informação que vai sendo obtida
With objConResult
    .Location = "D:nMaribelnDoutoramentonBasesDadosnBD_Geogra…ca.mdb"
    .Mode = gmcModeReadWrite
    .Type = "Access.GDatabase"
    .ConnectionName = "ConResult"
    .Connect
End With
objConResult.CreateOriginatingPipe objOPipe1
With objOPipe1
    .Table = "FacesAVR"
End With
Set objDB = CreateObject("Access.GDatabase")
objDB.OpenDatabase "D:nMaribelnDoutoramentonBasesDadosnBD_Geogra…ca.mdb"
Set objResult = objDB.OpenRecordset("FacesAVR", gdbOpenDynaset)
If Not (objResult.EOF) Then
    objResult.MoveLast
End If
Do Until objRecord1.EOF
    Do Until objRecord2.EOF
        If objRecord1.GFields(4).Value <> objRecord2.GFields(4).Value Then
            ' se não for a mesma regiao
            objResult.AddNew
            Call DetRelEspaciais
            With objResult
                .GFields(0).Value = objRecord1(4).Value
                .GFields(1).Value = objRecord2(4).Value
                .GFields(2).Value = CInt(angulo)
                .GFields(3).Value = CInt(distancia)
            End With
            objResult.Update
        End If
        objRecord2.MoveNext
    Loop
    objRecord1.MoveNext
    objRecord2.MoveFirst
    objRecord2.MoveFirst
Loop
Set objRecord1 = Nothing
Set objRecord2 = Nothing
Set objConnet1 = Nothing
```

```
Set objConResult = Nothing
Set objResult = Nothing
End Sub
```

Procedimento DetRelEspaciais

```
Sub DetRelEspaciais()
Dim objDB As GDatabase
Dim objGeometry As GeometryStorageService
Dim objPoint As New GeometryStorageService
Dim objGSS As New GeometryStorageService
Dim objRecord As GRecordset, objRec1 As GRecordset, objRec2 As GRecordset
Dim objResult As GRecordset
Dim objOPipe1 As OriginatingPipe, objOPipe2 As OriginatingPipe
Dim objGeom1 As Object, objGeom2 As Object, objGeomPnt As Object
Dim objPntLow1 As New Point, objPntHigh1 As New Point
Dim objPntLow2 As New Point, objPntHigh2 As New Point
Dim num As Integer, i As Integer
Dim face1 As String, face2 As String
Dim geomBlob As Variant
Dim objSpaFil As New SpatialFilter
Dim objCoord As GRecordset
face1 = objRecord1(4).Value
face2 = objRecord2(4).Value
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "SpatialArea"
    .Table = "LimitesAVR"
    .Filter = "codconc = "' & face1 & ""'
End With
objConnet1.CreateOriginatingPipe objOPipe2
With objOPipe2
    .GeometryFieldName = "SpatialArea"
    .Table = "LimitesAVR"
    .Filter = "codconc = "' & face2 & ""'
End With
Set objRec1 = objOPipe1.OutputRecordset
Set objRec2 = objOPipe2.OutputRecordset
Set objOPipe1 = Nothing
Set objOPipe2 = Nothing
objRec1.MoveFirst
objRec2.MoveFirst
' Geometria da face e identi…cacao respectivo centroide
Set objGeometry1 = CreateObject("GeoMedia.GeometryStorageService")
objGeometry1.GetGeometry objRec1.GFields("SpatialArea"), objGeom1
Set objSpaFil.Geometry = objGeom1
```

```
Set objGeometry1 = Nothing
Set objGeom1 = Nothing
Set objRec1 = Nothing
objGSS.GeometryToStorage objSpaFil.Geometry, geomBlob
Set objGSS = Nothing
Set objSpaFil = Nothing
objConnet1.CreateOriginatingPipe objOPipe1
With objOPipe1
    .GeometryFieldName = "Geometry1"
    .Table = "PontosCentroides"
    .SpatialFilter = geomBlob
    .SpatialOperator = gmsqContains
End With
Set geomBlob = Nothing
Set objResult = objOPipe1.OutputRecordset
Set objOPipe1 = Nothing
If Not objResult.EOF Then
    objResult.MoveFirst
    angulo = 0
    distancia = 0
    Set objPoint = CreateObject("GeoMedia.GeometryStorageService")
    objPoint.GetGeometry objResult.GFields("Geometry1"), objGeomPnt
    GetRange objGeomPnt, objPntLow1, objPntHigh1
    Set objPoint = Nothing
    Set objGeomPnt = Nothing
    ' Geometria da segunda face e respectivo centroide
    Set objGeometry1 = CreateObject("GeoMedia.GeometryStorageService")
    objGeometry1.GetGeometry objRec2.GFields("SpatialArea"), objGeom1
    Set objSpaFil.Geometry = objGeom1
    Set objGeometry1 = Nothing
    Set objGeom1 = Nothing
    Set objRec2 = Nothing
    objGSS.GeometryToStorage objSpaFil.Geometry, geomBlob
    Set objGSS = Nothing
    Set objSpaFil = Nothing
    objConnet1.CreateOriginatingPipe objOPipe1
    With objOPipe1
        .GeometryFieldName = "Geometry1"
        .Table = "PontosCentroides"
        .SpatialFilter = geomBlob
        .SpatialOperator = gmsqContains
    End With
    Set geomBlob = Nothing
    Set objResult = objOPipe1.OutputRecordset
    Set objOPipe1 = Nothing
```

```
If Not objResult.EOF Then
    objResult.MoveFirst
    Set objPoint = CreateObject("GeoMedia.GeometryStorageService")
    objPoint.GetGeometry objResult.GFields("Geometry1"), objGeomPnt
    GetRange objGeomPnt, objPntLow2, objPntHigh2
    Set objPoint = Nothing
    Set objGeomPnt = Nothing
    ' Calculo da distância e ângulo existente entre os centroides
    Dim linha As New LineGeometry
    Dim coordX As Double, coordY As Double
    Dim PI As Double
    PI = 4 * Atn(1)
    linha.Start.X = objPntLow1.X
    linha.Start.Y = objPntLow1.Y
    linha.Start.Z = 0
    linha.End.X = objPntHigh2.X
    linha.End.Y = objPntHigh2.Y
    linha.End.Z = 0
    coordX = (linha.Start.X - linha.End.X) / 100
    coordY = (linha.Start.Y - linha.End.Y) / 100
    distancia = Sqr(coordX * coordX + coordY * coordY) / 1000
    If coordX <> coordY Then
        angulo = Atn(coordX / coordY) * (180 / PI)
    End If
    ' veri...car quadrante en que esta localizado este angulo
    If angulo > 0 Then
        If coordX > 0 And coordY < 0 Then
            angulo = 90 + angulo
        Else
            If coordX < 0 And coordY < 0 Then
                angulo = 180 + angulo
            ElseIf coordX < 0 And coordY > 0 Then
                angulo = 270 + angulo
            End If
        End If
    End If
    If angulo < 0 Then
        If coordX > 0 And coordY < 0 Then
            angulo = 180 + angulo
        Else
            If coordX < 0 And coordY < 0 Then
                angulo = 270 + angulo
            Else
                If coordX < 0 And coordY > 0 Then
                    angulo = 360 + angulo
```

```
                        ElseIf coordX > 0 And coordY > 0 Then
                            angulo = 90 + angulo
                        End If
                    End If
                End If
            End If
        End If
Else
    angulo = 0
    distancia = 0
End If
Set objPntLow1 = Nothing
Set objPntLow2 = Nothing
Set objPntHigh1 = Nothing
Set objPntHigh2 = Nothing
Set linha = Nothing
End Sub
```

Apêndice D

Veri...cação e Identi...cação das regras de inferência

Neste apêndice são apresentados os cálculos quantitativos que permitiram averiguar a veracidade das regras (ratio 4) de inferência propostas por Hong [Hong, 1994] (secção D.1) e ainda, os cálculos quantitativos que possibilitaram a determinação das regras de inferência para o ratio 2 entre distâncias (secção D.2) e para o ratio 5 entre distâncias (secção D.3). Este apêndice apresenta ainda, secção D.4, as tabelas de composição utilizadas pelo Padrão no processo de descoberta de conhecimento, nomeadamente na fase de processamento da informação geo-espacial. O apêndice culmina com a integração da dimensão das regiões, secção D.5, no sistema de raciocínio qualitativo implementado.

D.1 Cálculos quantitativos para a veri...cação das regras de inferência, ratio 4

Esta secção apresenta os cálculos quantitativos que permitiram determinar a validade das regras de inferência propostas por Hong [Hong, 1994]. As composições, para cada um dos grupos de direcções ¢dir0 a ¢dir4 são apresentadas nas próximas subsecções, e referem-se às distâncias associadas ao ratio 4.

D.1.1 Grupo de composições para ¢dir0

A Tabela D.1 apresenta os cálculos quantitativos necessários à identi...cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir0, aqui representadas pelo caso particular N;N.

D.1.2 Grupo de composições para ¢dir1

A Tabela D.2 apresenta os cálculos quantitativos necessários à identi...cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir1, aqui representadas pelo caso particular N;NE.

(N,mp);(N,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	1	
	N,mp	

(N,mp);(N,p)	V_{AB}	V_{BC}
distância	0,5	3
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	3,5	
	N,p	

(N,mp);(N,d)	V_{AB}	V_{BC}
distância	0,5	13
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	13,5	
	N,d	

(N,mp);(N,md)	V_{AB}	V_{BC}
distância	0,5	53
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	53,5	
	N,md	

(N,p);(N,mp)	V_{AB}	V_{BC}
distância	3	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	3,5	
	N,p	

(N,p);(N,p)	V_{AB}	V_{BC}
distância	3	3
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	6	
	N,d	

(N,p);(N,d)	V_{AB}	V_{BC}
distância	3	13
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	16	
	N,d	

(N,p);(N,md)	V_{AB}	V_{BC}
distância	3	53
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	56	
	N,md	

(N,d);(N,mp)	V_{AB}	V_{BC}
distância	13	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	13,5	
	N,d	

(N,d);(N,p)	V_{AB}	V_{BC}
distância	13	3
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	16	
	N,d	

(N,d);(N,d)	V_{AB}	V_{BC}
distância	13	13
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	26	
	N,md	

(N,d);(N,md)	V_{AB}	V_{BC}
distância	13	53
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	66	
	N,md	

(N,md);(N,mp)	V_{AB}	V_{BC}
distância	53	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	53,5	
	N,md	

(N,md);(N,p)	V_{AB}	V_{BC}
distância	53	3
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	56	
	N,md	

(N,md);(N,d)	V_{AB}	V_{BC}
distância	53	13
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	66	
	N,md	

(N,md);(N,md)	V_{AB}	V_{BC}
distância	53	53
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	106	
	N,md	

Tabela D.1: Cálculos quantitativos para o grupo de direcções ¢dir0, ratio 4

(N,mp);(NE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	NE,mp	

(N,mp);(NE,p)	V_{AB}	V_{BC}
distância	0,5	3
direcção	180	225
Ang_{AC}=	38,9817	
$\|V_{AC}\|$=	3,37214	
	NE,p	

(N,mp);(NE,d)	V_{AB}	V_{BC}
distância	0,5	13
direcção	180	225
Ang_{AC}=	43,4834	
$\|V_{AC}\|$=	13,3582	
	NE,d	

(N,mp);(NE,md)	V_{AB}	V_{BC}
distância	0,5	53
direcção	180	225
Ang_{AC}=	44,6203	
$\|V_{AC}\|$=	53,3547	
	NE,md	

(N,p);(NE,mp)	V_{AB}	V_{BC}
distância	3	0,5
direcção	180	225
Ang_{AC}=	6,01826	
$\|V_{AC}\|$=	3,37214	
	N,p	

(N,p);(NE,p)	V_{AB}	V_{BC}
distância	3	3
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	5,54328	
	NE,d	

(N,p);(NE,d)	V_{AB}	V_{BC}
distância	3	13
direcção	180	225
Ang_{AC}=	37,0143	
$\|V_{AC}\|$=	15,2694	
	NE,d	

(N,p);(NE,md)	V_{AB}	V_{BC}
distância	3	53
direcção	180	225
Ang_{AC}=	42,7961	
$\|V_{AC}\|$=	55,1621	
	NE,md	

(N,d);(NE,mp)	V_{AB}	V_{BC}
distância	13	0,5
direcção	180	225
Ang_{AC}=	1,51663	
$\|V_{AC}\|$=	13,3582	
	N,d	

(N,d);(NE,p)	V_{AB}	V_{BC}
distância	13	3
direcção	180	225
Ang_{AC}=	7,98572	
$\|V_{AC}\|$=	15,2694	
	N,d	

(N,d);(NE,d)	V_{AB}	V_{BC}
distância	13	13
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	24,0209	
	NE,md	

(N,d);(NE,md)	V_{AB}	V_{BC}
distância	13	53
direcção	180	225
Ang_{AC}=	36,5922	
$\|V_{AC}\|$=	62,8681	
	NE,md	

(N,md);(NE,mp)	V_{AB}	V_{BC}
distância	53	0,5
direcção	180	225
Ang_{AC}=	0,37967	
$\|V_{AC}\|$=	53,3547	
	N,md	

(N,md);(NE,p)	V_{AB}	V_{BC}
distância	53	3
direcção	180	225
Ang_{AC}=	2,20392	
$\|V_{AC}\|$=	55,1621	
	N,md	

(N,md);(NE,d)	V_{AB}	V_{BC}
distância	53	13
direcção	180	225
Ang_{AC}=	8,40777	
$\|V_{AC}\|$=	62,8681	
	N,md	

(N,md);(NE,md)	V_{AB}	V_{BC}
distância	53	53
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	97,9312	
	NE,md	

Tabela D.2: Cálculos quantitativos para o grupo de direcções ¢dir1, ratio 4

(N,mp);(E,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	270
Ang_{AC}=		45
$\|V_{AC}\|$=		0,70711
		NE,mp

(N,mp);(E,p)	V_{AB}	V_{BC}
distância	0,5	3
direcção	180	270
Ang_{AC}=		80,5377
$\|V_{AC}\|$=		3,04138
		E,p

(N,mp);(E,d)	V_{AB}	V_{BC}
distância	0,5	13
direcção	180	270
Ang_{AC}=		87,7974
$\|V_{AC}\|$=		13,0096
		E,d

(N,mp);(E,md)	V_{AB}	V_{BC}
distância	0,5	53
direcção	180	270
Ang_{AC}=		89,4595
$\|V_{AC}\|$=		53,0024
		E,md

(N,p);(E,mp)	V_{AB}	V_{BC}
distância	3	0,5
direcção	180	270
Ang_{AC}=		9,46232
$\|V_{AC}\|$=		3,04138
		N,p

(N,p);(E,p)	V_{AB}	V_{BC}
distância	3	3
direcção	180	270
Ang_{AC}=		45
$\|V_{AC}\|$=		4,24264
		NE,p

(N,p);(E,d)	V_{AB}	V_{BC}
distância	3	13
direcção	180	270
Ang_{AC}=		77,0054
$\|V_{AC}\|$=		13,3417
		E,d

(N,p);(E,md)	V_{AB}	V_{BC}
distância	3	53
direcção	180	270
Ang_{AC}=		86,7603
$\|V_{AC}\|$=		53,0848
		E,md

(N,d);(E,mp)	V_{AB}	V_{BC}
distância	13	0,5
direcção	180	270
Ang_{AC}=		2,2026
$\|V_{AC}\|$=		13,0096
		N,d

(N,d);(E,p)	V_{AB}	V_{BC}
distância	13	3
direcção	180	270
Ang_{AC}=		12,9946
$\|V_{AC}\|$=		13,3417
		N,d

(N,d);(E,d)	V_{AB}	V_{BC}
distância	13	13
direcção	180	270
Ang_{AC}=		45
$\|V_{AC}\|$=		18,3848
		NE,d

(N,d);(E,md)	V_{AB}	V_{BC}
distância	13	53
direcção	180	270
Ang_{AC}=		76,2184
$\|V_{AC}\|$=		54,5711
		E,md

(N,md);(E,mp)	V_{AB}	V_{BC}
distância	53	0,5
direcção	180	270
Ang_{AC}=		0,54051
$\|V_{AC}\|$=		53,0024
		N,md

(N,md);(E,p)	V_{AB}	V_{BC}
distância	53	3
direcção	180	270
Ang_{AC}=		3,2397
$\|V_{AC}\|$=		53,0848
		N,md

(N,md);(E,d)	V_{AB}	V_{BC}
distância	53	13
direcção	180	270
Ang_{AC}=		13,7816
$\|V_{AC}\|$=		54,5711
		N,md

(N,md);(E,md)	V_{AB}	V_{BC}
distância	53	53
direcção	180	270
Ang_{AC}=		45
$\|V_{AC}\|$=		74,9533
		NE,md

Tabela D.3: Cálculos quantitativos para o grupo de direcções ¢dir2, ratio 4

D.1.3 Grupo de composições para ¢dir2

A Tabela D.3 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir2, aqui representadas pelo caso particular N;E.

D.1.4 Grupo de composições para ¢dir3

A Tabela D.4 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir3, aqui representadas pelo caso particular N;SE.

D.1.5 Grupo de composições para ¢dir4

A Tabela D.5 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir4, aqui representadas pelo caso particular N;S.

D.2 Identi…cação das regras de inferência, que integram a direcção e distância, para o ratio 2

Esta secção apresenta os cálculos quantitativos necessários à obtenção das regras de inferência, para os intervalos de validade para a distância obtidos a partir do ratio 2. Consideram-

(N,mp);(SE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,38268	
	E,mp	

(N,mp);(SE,p)	V_{AB}	V_{BC}
distância	0,5	3
direcção	180	315
Ang_{AC}=	127,391	
$\|V_{AC}\|$=	2,66996	
	SE,p	

(N,mp);(SE,d)	V_{AB}	V_{BC}
distância	0,5	13
direcção	180	315
Ang_{AC}=	133,399	
$\|V_{AC}\|$=	12,6514	
	SE,d	

(N,mp);(SE,md)	V_{AB}	V_{BC}
distância	0,5	53
direcção	180	315
Ang_{AC}=	134,615	
$\|V_{AC}\|$=	52,6476	
	SE,md	

(N,p);(SE,mp)	V_{AB}	V_{BC}
distância	3	0,5
direcção	180	315
Ang_{AC}=	7,6094	
$\|V_{AC}\|$=	2,66996	
	N,p	

(N,p);(SE,p)	V_{AB}	V_{BC}
distância	3	3
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	2,2961	
	E,p	

(N,p);(SE,d)	V_{AB}	V_{BC}
distância	3	13
direcção	180	315
Ang_{AC}=	123,966	
$\|V_{AC}\|$=	11,0836	
	SE,d	

(N,p);(SE,md)	V_{AB}	V_{BC}
distância	3	53
direcção	180	315
Ang_{AC}=	132,613	
$\|V_{AC}\|$=	50,9229	
	SE,md	

(N,d);(SE,mp)	V_{AB}	V_{BC}
distância	13	0,5
direcção	180	315
Ang_{AC}=	1,60139	
$\|V_{AC}\|$=	12,6514	
	N,d	

(N,d);(SE,p)	V_{AB}	V_{BC}
distância	13	3
direcção	180	315
Ang_{AC}=	11,0341	
$\|V_{AC}\|$=	11,0836	
	N,d	

(N,d);(SE,d)	V_{AB}	V_{BC}
distância	13	13
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	9,94977	
	E,d	

(N,d);(SE,md)	V_{AB}	V_{BC}
distância	13	53
direcção	180	315
Ang_{AC}=	123,149	
$\|V_{AC}\|$=	44,7617	
	SE,md	

(N,md);(SE,mp)	V_{AB}	V_{BC}
distância	53	0,5
direcção	180	315
Ang_{AC}=	0,38477	
$\|V_{AC}\|$=	52,6476	
	N,md	

(N,md);(SE,p)	V_{AB}	V_{BC}
distância	53	3
direcção	180	315
Ang_{AC}=	2,38749	
$\|V_{AC}\|$=	50,9229	
	N,md	

(N,md);(SE,d)	V_{AB}	V_{BC}
distância	53	13
direcção	180	315
Ang_{AC}=	11,8507	
$\|V_{AC}\|$=	44,7617	
	N,md	

(N,md);(SE,md)	V_{AB}	V_{BC}
distância	53	53
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	40,5644	
	E,md	

Tabela D.4: Cálculos quantitativos para o grupo de direcções ¢dir3, ratio 4

(N,mp);(S,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	360
Ang_{AC}=	#DIV/0!	
$\|V_{AC}\|$=	6,1E-17	
	Indf	

(N,mp);(S,p)	V_{AB}	V_{BC}
distância	0,5	3
direcção	180	360
Ang_{AC}=	180	
$\|V_{AC}\|$=	2,5	
	S,p	

(N,mp);(S,d)	V_{AB}	V_{BC}
distância	0,5	13
direcção	180	360
Ang_{AC}=	180	
$\|V_{AC}\|$=	12,5	
	S,d	

(N,mp);(S,md)	V_{AB}	V_{BC}
distância	0,5	53
direcção	180	360
Ang_{AC}=	180	
$\|V_{AC}\|$=	52,5	
	S,md	

(N,p);(S,mp)	V_{AB}	V_{BC}
distância	3	0,5
direcção	180	360
Ang_{AC}=	360	
$\|V_{AC}\|$=	2,5	
	N,p	

(N,p);(S,p)	V_{AB}	V_{BC}
distância	3	3
direcção	180	360
Ang_{AC}=	#DIV/0!	
$\|V_{AC}\|$=	3,7E-16	
	Indf	

(N,p);(S,d)	V_{AB}	V_{BC}
distância	3	13
direcção	180	360
Ang_{AC}=	180	
$\|V_{AC}\|$=	10	
	S,d	

(N,p);(S,md)	V_{AB}	V_{BC}
distância	3	53
direcção	180	360
Ang_{AC}=	180	
$\|V_{AC}\|$=	50	
	S,md	

(N,d);(S,mp)	V_{AB}	V_{BC}
distância	13	0,5
direcção	180	360
Ang_{AC}=	360	
$\|V_{AC}\|$=	12,5	
	N,d	

(N,d);(S,p)	V_{AB}	V_{BC}
distância	13	3
direcção	180	360
Ang_{AC}=	360	
$\|V_{AC}\|$=	10	
	N,d	

(N,d);(S,d)	V_{AB}	V_{BC}
distância	13	13
direcção	180	360
Ang_{AC}=	#DIV/0!	
$\|V_{AC}\|$=	1,6E-15	
	Indf	

(N,d);(S,md)	V_{AB}	V_{BC}
distância	13	53
direcção	180	360
Ang_{AC}=	180	
$\|V_{AC}\|$=	40	
	S,md	

(N,md);(S,mp)	V_{AB}	V_{BC}
distância	53	0,5
direcção	180	360
Ang_{AC}=	360	
$\|V_{AC}\|$=	52,5	
	N,md	

(N,md);(S,p)	V_{AB}	V_{BC}
distância	53	3
direcção	180	360
Ang_{AC}=	360	
$\|V_{AC}\|$=	50	
	N,md	

(N,md);(S,d)	V_{AB}	V_{BC}
distância	53	13
direcção	180	360
Ang_{AC}=	360	
$\|V_{AC}\|$=	40	
	N,md	

(N,md);(S,md)	V_{AB}	V_{BC}
distância	53	53
direcção	180	360
Ang_{AC}=	#DIV/0!	
$\|V_{AC}\|$=	6,5E-15	
	Indf	

Tabela D.5: Cálculos quantitativos para o grupo de direcções ¢dir4, ratio 4

(N,mp);(N,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	1	
	N,mp	

(N,mp);(N,p)	V_{AB}	V_{BC}
distância	0,5	2
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	2,5	
	N,p	

(N,mp);(N,d)	V_{AB}	V_{BC}
distância	0,5	5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	5,5	
	N,d	

(N,mp);(N,md)	V_{AB}	V_{BC}
distância	0,5	11
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	11,5	
	N,md	

(N,p);(N,mp)	V_{AB}	V_{BC}
distância	2	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	2,5	
	N,p	

(N,p);(N,p)	V_{AB}	V_{BC}
distância	2	2
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	4	
	N,d	

(N,p);(N,d)	V_{AB}	V_{BC}
distância	2	5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	7	
	N,d	

(N,p);(N,md)	V_{AB}	V_{BC}
distância	2	11
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	13	
	N,md	

(N,d);(N,mp)	V_{AB}	V_{BC}
distância	5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	5,5	
	N,d	

(N,d);(N,p)	V_{AB}	V_{BC}
distância	5	2
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	7	
	N,d	

(N,d);(N,d)	V_{AB}	V_{BC}
distância	5	5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	10	
	N,md	

(N,d);(N,md)	V_{AB}	V_{BC}
distância	5	11
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	16	
	N,md	

(N,md);(N,mp)	V_{AB}	V_{BC}
distância	11	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	11,5	
	N,md	

(N,md);(N,p)	V_{AB}	V_{BC}
distância	11	2
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	13	
	N,md	

(N,md);(N,d)	V_{AB}	V_{BC}
distância	11	5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	16	
	N,md	

(N,md);(N,md)	V_{AB}	V_{BC}
distância	11	11
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	22	
	N,md	

Tabela D.6: Cálculos quantitativos para o grupo de direcções ¢dir0, ratio 2

se os intervalos de validade para a direcção de (337.5,22.5], (22.5,67.5], (67.5,112.5], (112.5,157.5], (157.5,202.5], (202.5,247.5], (247.5,292.5] e (292.5,337.5], de N a NO respectivamente. Os intervalos de validade adoptados para o ratio 2 são: mp (0,1], p (1,3], d (3,7] e md (7,15].

D.2.1 Grupo de composições para ¢dir0

A Tabela D.6 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir0, aqui representadas pelo caso particular N;N.

D.2.2 Grupo de composições para ¢dir1

A Tabela D.7 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir1, aqui representadas pelo caso particular N;NE.

D.2.3 Grupo de composições para ¢dir2

A Tabela D.8 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir2, aqui representadas pelo caso particular N;E.

(N,mp);(NE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	N,mp	

(N,mp);(NE,p)	V_{AB}	V_{BC}
distância	0,5	2
direcção	180	225
Ang_{AC}=	36,4568	
$\|V_{AC}\|$=	2,37996	
	NE,p	

(N,mp);(NE,d)	V_{AB}	V_{BC}
distância	0,5	5
direcção	180	225
Ang_{AC}=	41,2216	
$\|V_{AC}\|$=	5,36522	
	NE,d	

(N,mp);(NE,md)	V_{AB}	V_{BC}
distância	0,5	11
direcção	180	225
Ang_{AC}=	43,2164	
$\|V_{AC}\|$=	11,3591	
	NE,md	

(N,p);(NE,mp)	V_{AB}	V_{BC}
distância	2	0,5
direcção	180	225
Ang_{AC}=	8,54315	
$\|V_{AC}\|$=	2,37996	
	N,p	

(N,p);(NE,p)	V_{AB}	V_{BC}
distância	2	2
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	3,69552	
	N,d	

(N,p);(NE,d)	V_{AB}	V_{BC}
distância	2	5
direcção	180	225
Ang_{AC}=	32,5663	
$\|V_{AC}\|$=	6,56827	
	NE,d	

(N,p);(NE,md)	V_{AB}	V_{BC}
distância	2	11
direcção	180	225
Ang_{AC}=	38,5009	
$\|V_{AC}\|$=	12,4945	
	NE,md	

(N,d);(NE,mp)	V_{AB}	V_{BC}
distância	5	0,5
direcção	180	225
Ang_{AC}=	3,77838	
$\|V_{AC}\|$=	5,36522	
	N,d	

(N,d);(NE,p)	V_{AB}	V_{BC}
distância	5	2
direcção	180	225
Ang_{AC}=	12,4337	
$\|V_{AC}\|$=	6,56827	
	N,d	

(N,d);(NE,d)	V_{AB}	V_{BC}
distância	5	5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	9,2388	
	N,md	

(N,d);(NE,md)	V_{AB}	V_{BC}
distância	5	11
direcção	180	225
Ang_{AC}=	31,3292	
$\|V_{AC}\|$=	14,9593	
	NE,md	

(N,md);(NE,mp)	V_{AB}	V_{BC}
distância	11	0,5
direcção	180	225
Ang_{AC}=	1,78363	
$\|V_{AC}\|$=	11,3591	
	N,md	

(N,md);(NE,p)	V_{AB}	V_{BC}
distância	11	2
direcção	180	225
Ang_{AC}=	6,49905	
$\|V_{AC}\|$=	12,4945	
	N,md	

(N,md);(NE,d)	V_{AB}	V_{BC}
distância	11	5
direcção	180	225
Ang_{AC}=	13,6708	
$\|V_{AC}\|$=	14,9593	
	N,md	

(N,md);(NE,md)	V_{AB}	V_{BC}
distância	11	11
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	20,3253	
	N,md	

Tabela D.7: Cálculos quantitativos para o grupo de direcções ¢dir1, ratio 2

(N,mp);(E,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	0,70711	
	NE,mp	

(N,mp);(E,p)	V_{AB}	V_{BC}
distância	0,5	2
direcção	180	270
Ang_{AC}=	75,9638	
$\|V_{AC}\|$=	2,06155	
	E,p	

(N,mp);(E,d)	V_{AB}	V_{BC}
distância	0,5	5
direcção	180	270
Ang_{AC}=	84,2894	
$\|V_{AC}\|$=	5,02494	
	E,d	

(N,mp);(E,md)	V_{AB}	V_{BC}
distância	0,5	11
direcção	180	270
Ang_{AC}=	87,3974	
$\|V_{AC}\|$=	11,0114	
	E,md	

(N,p);(E,mp)	V_{AB}	V_{BC}
distância	2	0,5
direcção	180	270
Ang_{AC}=	14,0362	
$\|V_{AC}\|$=	2,06155	
	N,p	

(N,p);(E,p)	V_{AB}	V_{BC}
distância	2	2
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	2,82843	
	NE,p	

(N,p);(E,d)	V_{AB}	V_{BC}
distância	2	5
direcção	180	270
Ang_{AC}=	68,1986	
$\|V_{AC}\|$=	5,38516	
	E,d	

(N,p);(E,md)	V_{AB}	V_{BC}
distância	2	11
direcção	180	270
Ang_{AC}=	79,6952	
$\|V_{AC}\|$=	11,1803	
	E,md	

(N,d);(E,mp)	V_{AB}	V_{BC}
distância	5	0,5
direcção	180	270
Ang_{AC}=	5,71059	
$\|V_{AC}\|$=	5,02494	
	N,d	

(N,d);(E,p)	V_{AB}	V_{BC}
distância	5	2
direcção	180	270
Ang_{AC}=	21,8014	
$\|V_{AC}\|$=	5,38516	
	N,d	

(N,d);(E,d)	V_{AB}	V_{BC}
distância	5	5
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	7,07107	
	NE,d	

(N,d);(E,md)	V_{AB}	V_{BC}
distância	5	11
direcção	180	270
Ang_{AC}=	65,556	
$\|V_{AC}\|$=	12,083	
	NE,md	

(N,md);(E,mp)	V_{AB}	V_{BC}
distância	11	0,5
direcção	180	270
Ang_{AC}=	2,60256	
$\|V_{AC}\|$=	11,0114	
	N,md	

(N,md);(E,p)	V_{AB}	V_{BC}
distância	11	2
direcção	180	270
Ang_{AC}=	10,3048	
$\|V_{AC}\|$=	11,1803	
	N,md	

(N,md);(E,d)	V_{AB}	V_{BC}
distância	11	5
direcção	180	270
Ang_{AC}=	24,444	
$\|V_{AC}\|$=	12,083	
	NE,md	

(N,md);(E,md)	V_{AB}	V_{BC}
distância	11	11
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	15,5563	
	NE,md	

Tabela D.8: Cálculos quantitativos para o grupo de direcções ¢dir2, ratio 2

(N,mp);(SE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,38268	
	NE,mp	

(N,mp);(SE,p)	V_{AB}	V_{BC}
distância	0,5	2
direcção	180	315
Ang_{AC}=	122,881	
$\|V_{AC}\|$=	1,68398	
	SE,p	

(N,mp);(SE,d)	V_{AB}	V_{BC}
distância	0,5	5
direcção	180	315
Ang_{AC}=	130,649	
$\|V_{AC}\|$=	4,65988	
	SE,d	

(N,mp);(SE,md)	V_{AB}	V_{BC}
distância	0,5	11
direcção	180	315
Ang_{AC}=	133,098	
$\|V_{AC}\|$=	10,6523	
	SE,md	

(N,p);(SE,mp)	V_{AB}	V_{BC}
distância	2	0,5
direcção	180	315
Ang_{AC}=	12,1195	
$\|V_{AC}\|$=	1,68398	
	N,p	

(N,p);(SE,p)	V_{AB}	V_{BC}
distância	2	2
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	1,53073	
	NE,p	

(N,p);(SE,d)	V_{AB}	V_{BC}
distância	2	5
direcção	180	315
Ang_{AC}=	113,476	
$\|V_{AC}\|$=	3,85459	
	SE,d	

(N,p);(SE,md)	V_{AB}	V_{BC}
distância	2	11
direcção	180	315
Ang_{AC}=	126,608	
$\|V_{AC}\|$=	9,68955	
	SE,md	

(N,d);(SE,mp)	V_{AB}	V_{BC}
distância	5	0,5
direcção	180	315
Ang_{AC}=	4,35132	
$\|V_{AC}\|$=	4,65988	
	N,d	

(N,d);(SE,p)	V_{AB}	V_{BC}
distância	5	2
direcção	180	315
Ang_{AC}=	21,524	
$\|V_{AC}\|$=	3,85459	
	N,d	

(N,d);(SE,d)	V_{AB}	V_{BC}
distância	5	5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	3,82683	
	NE,d	

(N,d);(SE,md)	V_{AB}	V_{BC}
distância	5	11
direcção	180	315
Ang_{AC}=	109,655	
$\|V_{AC}\|$=	8,25943	
	E,md	

(N,md);(SE,mp)	V_{AB}	V_{BC}
distância	11	0,5
direcção	180	315
Ang_{AC}=	1,90201	
$\|V_{AC}\|$=	10,6523	
	N,md	

(N,md);(SE,p)	V_{AB}	V_{BC}
distância	11	2
direcção	180	315
Ang_{AC}=	8,39244	
$\|V_{AC}\|$=	9,68955	
	N,md	

(N,md);(SE,d)	V_{AB}	V_{BC}
distância	11	5
direcção	180	315
Ang_{AC}=	25,3445	
$\|V_{AC}\|$=	8,25943	
	NE,md	

(N,md);(SE,md)	V_{AB}	V_{BC}
distância	11	11
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	8,41904	
	NE,md	

Tabela D.9: Cálculos quantitativos para o grupo de direcções ¢dir3, ratio 2

D.2.4 Grupo de composições para ¢dir3

A Tabela D.9 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir3, aqui representadas pelo caso particular N;SE.

D.2.5 Grupo de composições para ¢dir4

A Tabela D.10 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir4, aqui representadas pelo caso particular N;S.

D.3 Identi…cação das regras de inferência, que integram a direcção e distância, para o ratio 5

Esta secção apresenta os cálculos quantitativos necessários à obtenção das regras de inferência, para os intervalos de validade para a distância, obtidos a partir do ratio 5. Consideram-se os intervalos de validade para a direcção de (337.5,22.5], (22.5,67.5], (67.5,112.5], (112.5,157.5], (157.5,202.5], (202.5,247.5], (247.5,292.5] e (292.5,337.5], de N a NO respectivamente. Os intervalos de validade adoptados para o ratio 5 são: mp (0,1], p (1,6], d (6,31] e md (31,156].

(N,mp);(S,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		6,1E-17
		Indf

(N,mp);(S,p)	V_{AB}	V_{BC}
distância	0,5	2
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		1,5
		S,p

(N,mp);(S,d)	V_{AB}	V_{BC}
distância	0,5	5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		4,5
		S,d

(N,mp);(S,md)	V_{AB}	V_{BC}
distância	0,5	11
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		10,5
		S,md

(N,p);(S,mp)	V_{AB}	V_{BC}
distância	2	0,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		1,5
		N,p

(N,p);(S,p)	V_{AB}	V_{BC}
distância	2	2
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		2,5E-16
		Indf

(N,p);(S,d)	V_{AB}	V_{BC}
distância	2	5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		3
		S,p

(N,p);(S,md)	V_{AB}	V_{BC}
distância	2	11
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		9
		S,md

(N,d);(S,mp)	V_{AB}	V_{BC}
distância	5	0,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		4,5
		N,d

(N,d);(S,p)	V_{AB}	V_{BC}
distância	5	2
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		3
		N,p

(N,d);(S,d)	V_{AB}	V_{BC}
distância	5	5
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		6,1E-16
		Indf

(N,d);(S,md)	V_{AB}	V_{BC}
distância	5	11
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		6
		S,d

(N,md);(S,mp)	V_{AB}	V_{BC}
distância	11	0,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		10,5
		N,md

(N,md);(S,p)	V_{AB}	V_{BC}
distância	11	2
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		9
		N,md

(N,md);(S,d)	V_{AB}	V_{BC}
distância	11	5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		6
		N,d

(N,md);(S,md)	V_{AB}	V_{BC}
distância	11	11
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		1,3E-15
		Indf

Tabela D.10: Cálculos quantitativos para o grupo de direcções ¢dir4, ratio 2

D.3.1 Grupo de composições para ¢dir0

A Tabela D.11 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir0, aqui representadas pelo caso particular N;N.

D.3.2 Grupo de composições para ¢dir1

A Tabela D.12 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir1, aqui representadas pelo caso particular N;NE.

D.3.3 Grupo de composições para ¢dir2

A Tabela D.13 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir2, aqui representadas pelo caso particular N;E.

D.3.4 Grupo de composições para ¢dir3

A Tabela D.14 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir3, aqui representadas pelo caso particular N;SE.

(N,mp);(N,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	1	
N,mp		

(N,mp);(N,p)	V_{AB}	V_{BC}
distância	0,5	3,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	4	
N,p		

(N,mp);(N,d)	V_{AB}	V_{BC}
distância	0,5	18,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	19	
N,d		

(N,mp);(N,md)	V_{AB}	V_{BC}
distância	0,5	93,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	94	
N,md		

(N,p);(N,mp)	V_{AB}	V_{BC}
distância	3,5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	4	
N,p		

(N,p);(N,p)	V_{AB}	V_{BC}
distância	3,5	3,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	7	
N,d		

(N,p);(N,d)	V_{AB}	V_{BC}
distância	3,5	18,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	22	
N,d		

(N,p);(N,md)	V_{AB}	V_{BC}
distância	3,5	93,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	97	
N,md		

(N,d);(N,mp)	V_{AB}	V_{BC}
distância	18,5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	19	
N,d		

(N,d);(N,p)	V_{AB}	V_{BC}
distância	18,5	3,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	22	
N,d		

(N,d);(N,d)	V_{AB}	V_{BC}
distância	18,5	18,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	37	
N,md		

(N,d);(N,md)	V_{AB}	V_{BC}
distância	18,5	93,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	112	
N,md		

(N,md);(N,mp)	V_{AB}	V_{BC}
distância	93,5	0,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	94	
N,md		

(N,md);(N,p)	V_{AB}	V_{BC}
distância	93,5	3,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	97	
N,md		

(N,md);(N,d)	V_{AB}	V_{BC}
distância	93,5	18,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	112	
N,md		

(N,md);(N,md)	V_{AB}	V_{BC}
distância	93,5	93,5
direcção	180	180
Ang_{AC}=	-7E-15	
$\|V_{AC}\|$=	187	
N,md		

Tabela D.11: Cálculos quantitativos para o grupo de direcções ¢dir0, ratio 5

(N,mp);(NE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
N,mp		

(N,mp);(NE,p)	V_{AB}	V_{BC}
distância	0,5	3,5
direcção	180	225
Ang_{AC}=	39,7579	
$\|V_{AC}\|$=	3,86974	
NE,p		

(N,mp);(NE,d)	V_{AB}	V_{BC}
distância	0,5	18,5
direcção	180	225
Ang_{AC}=	43,9257	
$\|V_{AC}\|$=	18,8569	
NE,d		

(N,mp);(NE,md)	V_{AB}	V_{BC}
distância	0,5	93,5
direcção	180	225
Ang_{AC}=	44,7842	
$\|V_{AC}\|$=	93,8542	
NE,md		

(N,p);(NE,mp)	V_{AB}	V_{BC}
distância	3,5	0,5
direcção	180	225
Ang_{AC}=	5,24206	
$\|V_{AC}\|$=	3,86974	
N,p		

(N,p);(NE,p)	V_{AB}	V_{BC}
distância	3,5	3,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	6,46716	
N,p		

(N,p);(NE,d)	V_{AB}	V_{BC}
distância	3,5	18,5
direcção	180	225
Ang_{AC}=	38,2707	
$\|V_{AC}\|$=	21,1204	
NE,d		

(N,p);(NE,md)	V_{AB}	V_{BC}
distância	3,5	93,5
direcção	180	225
Ang_{AC}=	43,5229	
$\|V_{AC}\|$=	96,0068	
NE,md		

(N,d);(NE,mp)	V_{AB}	V_{BC}
distância	18,5	0,5
direcção	180	225
Ang_{AC}=	1,07432	
$\|V_{AC}\|$=	18,8569	
N,d		

(N,d);(NE,p)	V_{AB}	V_{BC}
distância	18,5	3,5
direcção	180	225
Ang_{AC}=	6,72935	
$\|V_{AC}\|$=	21,1204	
N,d		

(N,d);(NE,d)	V_{AB}	V_{BC}
distância	18,5	18,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	34,1835	
N,md		

(N,d);(NE,md)	V_{AB}	V_{BC}
distância	18,5	93,5
direcção	180	225
Ang_{AC}=	38,0027	
$\|V_{AC}\|$=	107,381	
NE,md		

(N,md);(NE,mp)	V_{AB}	V_{BC}
distância	93,5	0,5
direcção	180	225
Ang_{AC}=	0,21584	
$\|V_{AC}\|$=	93,8542	
N,md		

(N,md);(NE,p)	V_{AB}	V_{BC}
distância	93,5	3,5
direcção	180	225
Ang_{AC}=	1,47714	
$\|V_{AC}\|$=	96,0068	
N,md		

(N,md);(NE,d)	V_{AB}	V_{BC}
distância	93,5	18,5
direcção	180	225
Ang_{AC}=	6,99731	
$\|V_{AC}\|$=	107,381	
N,md		

(N,md);(NE,md)	V_{AB}	V_{BC}
distância	93,5	93,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	172,765	
N,md		

Tabela D.12: Cálculos quantitativos para o grupo de direcções ¢dir1, ratio 5

(N,mp);(E,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	0,70711	
	NE,mp	

(N,mp);(E,p)	V_{AB}	V_{BC}
distância	0,5	3,5
direcção	180	270
Ang_{AC}=	81,8699	
$\|V_{AC}\|$=	3,53553	
	E,p	

(N,mp);(E,d)	V_{AB}	V_{BC}
distância	0,5	18,5
direcção	180	270
Ang_{AC}=	88,4518	
$\|V_{AC}\|$=	18,5068	
	E,d	

(N,mp);(E,md)	V_{AB}	V_{BC}
distância	0,5	93,5
direcção	180	270
Ang_{AC}=	89,6936	
$\|V_{AC}\|$=	93,5013	
	E,md	

(N,p);(E,mp)	V_{AB}	V_{BC}
distância	3,5	0,5
direcção	180	270
Ang_{AC}=	8,1301	
$\|V_{AC}\|$=	3,53553	
	N,p	

(N,p);(E,p)	V_{AB}	V_{BC}
distância	3,5	3,5
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	4,94975	
	NE,p	

(N,p);(E,d)	V_{AB}	V_{BC}
distância	3,5	18,5
direcção	180	270
Ang_{AC}=	79,2869	
$\|V_{AC}\|$=	18,8282	
	E,d	

(N,p);(E,md)	V_{AB}	V_{BC}
distância	3,5	93,5
direcção	180	270
Ang_{AC}=	87,8562	
$\|V_{AC}\|$=	93,5655	
	E,md	

(N,d);(E,mp)	V_{AB}	V_{BC}
distância	18,5	0,5
direcção	180	270
Ang_{AC}=	1,54816	
$\|V_{AC}\|$=	18,5068	
	N,d	

(N,d);(E,p)	V_{AB}	V_{BC}
distância	18,5	3,5
direcção	180	270
Ang_{AC}=	10,7131	
$\|V_{AC}\|$=	18,8282	
	N,d	

(N,d);(E,d)	V_{AB}	V_{BC}
distância	18,5	18,5
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	26,163	
	NE,d	

(N,d);(E,md)	V_{AB}	V_{BC}
distância	18,5	93,5
direcção	180	270
Ang_{AC}=	78,808	
$\|V_{AC}\|$=	95,3126	
	E,md	

(N,md);(E,mp)	V_{AB}	V_{BC}
distância	93,5	0,5
direcção	180	270
Ang_{AC}=	0,30639	
$\|V_{AC}\|$=	93,5013	
	N,md	

(N,md);(E,p)	V_{AB}	V_{BC}
distância	93,5	3,5
direcção	180	270
Ang_{AC}=	2,14376	
$\|V_{AC}\|$=	93,5655	
	N,md	

(N,md);(E,d)	V_{AB}	V_{BC}
distância	93,5	18,5
direcção	180	270
Ang_{AC}=	11,192	
$\|V_{AC}\|$=	95,3126	
	N,md	

(N,md);(E,md)	V_{AB}	V_{BC}
distância	93,5	93,5
direcção	180	270
Ang_{AC}=	45	
$\|V_{AC}\|$=	132,229	
	NE,md	

Tabela D.13: Cálculos quantitativos para o grupo de direcções ¢dir2, ratio 5

(N,mp);(SE,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,38268	
	NE,mp	

(N,mp);(SE,p)	V_{AB}	V_{BC}
distância	0,5	3,5
direcção	180	315
Ang_{AC}=	128,589	
$\|V_{AC}\|$=	3,16625	
	SE,p	

(N,mp);(SE,d)	V_{AB}	V_{BC}
distância	0,5	18,5
direcção	180	315
Ang_{AC}=	133,884	
$\|V_{AC}\|$=	18,1499	
	SE,d	

(N,mp);(SE,md)	V_{AB}	V_{BC}
distância	0,5	93,5
direcção	180	315
Ang_{AC}=	134,783	
$\|V_{AC}\|$=	93,1471	
	SE,md	

(N,p);(SE,mp)	V_{AB}	V_{BC}
distância	3,5	0,5
direcção	180	315
Ang_{AC}=	6,4112	
$\|V_{AC}\|$=	3,16625	
	N,p	

(N,p);(SE,p)	V_{AB}	V_{BC}
distância	3,5	3,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	2,67878	
	NE,p	

(N,p);(SE,d)	V_{AB}	V_{BC}
distância	3,5	18,5
direcção	180	315
Ang_{AC}=	126,221	
$\|V_{AC}\|$=	16,2151	
	SE,d	

(N,p);(SE,md)	V_{AB}	V_{BC}
distância	3,5	93,5
direcção	180	315
Ang_{AC}=	133,443	
$\|V_{AC}\|$=	91,0588	
	SE,md	

(N,d);(SE,mp)	V_{AB}	V_{BC}
distância	18,5	0,5
direcção	180	315
Ang_{AC}=	1,11617	
$\|V_{AC}\|$=	18,1499	
	N,d	

(N,d);(SE,p)	V_{AB}	V_{BC}
distância	18,5	3,5
direcção	180	315
Ang_{AC}=	8,77923	
$\|V_{AC}\|$=	16,2151	
	N,d	

(N,d);(SE,d)	V_{AB}	V_{BC}
distância	18,5	18,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	14,1593	
	NE,d	

(N,d);(SE,md)	V_{AB}	V_{BC}
distância	18,5	93,5
direcção	180	315
Ang_{AC}=	125,761	
$\|V_{AC}\|$=	81,4755	
	SE,md	

(N,md);(SE,mp)	V_{AB}	V_{BC}
distância	93,5	0,5
direcção	180	315
Ang_{AC}=	0,21747	
$\|V_{AC}\|$=	93,1471	
	N,md	

(N,md);(SE,p)	V_{AB}	V_{BC}
distância	93,5	3,5
direcção	180	315
Ang_{AC}=	1,55743	
$\|V_{AC}\|$=	91,0588	
	N,md	

(N,md);(SE,d)	V_{AB}	V_{BC}
distância	93,5	18,5
direcção	180	315
Ang_{AC}=	9,23923	
$\|V_{AC}\|$=	81,4755	
	N,md	

(N,md);(SE,md)	V_{AB}	V_{BC}
distância	93,5	93,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	71,5618	
	NE,md	

Tabela D.14: Cálculos quantitativos para o grupo de direcções ¢dir3, ratio 5

(N,mp);(S,mp)	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		6,1E-17
		Indf

(N,mp);(S,p)	V_{AB}	V_{BC}
distância	0,5	3,5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		3
		S,p

(N,mp);(S,d)	V_{AB}	V_{BC}
distância	0,5	18,5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		18
		S,d

(N,mp);(S,md)	V_{AB}	V_{BC}
distância	0,5	93,5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		93
		S,md

(N,p);(S,mp)	V_{AB}	V_{BC}
distância	3,5	0,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		3
		N,p

(N,p);(S,p)	V_{AB}	V_{BC}
distância	3,5	3,5
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		4,3E-16
		Indf

(N,p);(S,d)	V_{AB}	V_{BC}
distância	3,5	18,5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		15
		S,d

(N,p);(S,md)	V_{AB}	V_{BC}
distância	3,5	93,5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		90
		S,md

(N,d);(S,mp)	V_{AB}	V_{BC}
distância	18,5	0,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		18
		N,d

(N,d);(S,p)	V_{AB}	V_{BC}
distância	18,5	3,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		15
		N,d

(N,d);(S,d)	V_{AB}	V_{BC}
distância	18,5	18,5
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		2,3E-15
		Indf

(N,d);(S,md)	V_{AB}	V_{BC}
distância	18,5	93,5
direcção	180	360
Ang_{AC}=		180
$\|V_{AC}\|$=		75
		S,md

(N,md);(S,mp)	V_{AB}	V_{BC}
distância	93,5	0,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		93
		N,md

(N,md);(S,p)	V_{AB}	V_{BC}
distância	93,5	3,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		90
		N,md

(N,md);(S,d)	V_{AB}	V_{BC}
distância	93,5	18,5
direcção	180	360
Ang_{AC}=		360
$\|V_{AC}\|$=		75
		N,md

(N,md);(S,md)	V_{AB}	V_{BC}
distância	93,5	93,5
direcção	180	360
Ang_{AC}=		#DIV/0!
$\|V_{AC}\|$=		1,1E-14
		Indf

Tabela D.15: Cálculos quantitativos para o grupo de direcções ¢dir4, ratio 5

D.3.5 Grupo de composições para ¢dir4

A Tabela D.15 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, que integram a direcção e distância, para o grupo de direcções ¢dir4, aqui representadas pelo caso particular N;S.

D.4 Tabelas de composição que integram a direcção, distância e topologia

A identi…cação do conjunto de regras que permitem a integração da direcção e distância, para o ratio 2 e para o ratio 5, possibilita a construção de duas novas tabelas de composição. Estas tabelas, que integram a direcção, distância e topologia para estes ratios, permitem raciocinar qualitativamente em contextos geográ…cos com características diferentes ao nível da distância. A subsecção D.4.1 apresenta as tabelas …nais para o ratio 2 e a subsecção D.4.2 apresenta as tabelas …nais para o ratio 4, considerando os novos intervalos de validade adoptados para a direcção. A subsecção D.4.3 apresenta as tabelas de composição para o ratio 5.

D.4.1 Tabela de composição para o ratio 2

As regras que permitem a composição de relações integradas, direcção, distância e topologia, foram obtidas utilizando os intervalos de validade quantitativos para a direcção de (337.5, 22.5], (22.5, 67.5], (67.5, 112.5], (112.5, 157.5], (157.5, 202.5], (202.5, 247.5], (247.5, 292.5] e (292.5, 337.5], de N a NO respectivamente. Os intervalos de validade

iniciais, ratio 2, para a distância são: mp (0, 1], p (1, 3], d (3, 7] e md (7, 15]. As Figuras D.1, D.2 e D.3 sintetizam as regras de composição utilizadas para este ratio.

D.4.2 Tabela de composição para o ratio 4

As regras que permitem a composição de relações integradas, direcção, distância e topologia, foram obtidas utilizando os intervalos de validade quantitativos para a direcção de (337.5, 22.5], (22.5, 67.5], (67.5, 112.5], (112.5, 157.5], (157.5, 202.5], (202.5, 247.5], (247.5, 292.5] e (292.5, 337.5], de N a NO respectivamente. Os intervalos de validade iniciais, ratio 4, para a distância são: mp (0, 1], p (1, 5], d (5, 21] e md (21, 85]. As Figuras D.4, D.5 e D.6 sintetizam as regras de composição utilizadas para este ratio.

D.4.3 Tabela de composição para o ratio 5

As regras que permitem a composição de relações integradas, direcção, distância e topologia, foram obtidas utilizando os intervalos de validade quantitativos para a direcção de (337.5, 22.5], (22.5, 67.5], (67.5, 112.5], (112.5, 157.5], (157.5, 202.5], (202.5, 247.5], (247.5, 292.5] e (292.5, 337.5], de N a NO respectivamente. Os intervalos de validade iniciais, ratio 5, para a distância são: mp (0, 1], p (1, 6], d (6, 31] e md (31, 156]. As Figuras D.7, D.8 e D.9 sintetizam as regras de composição utilizadas para este ratio.

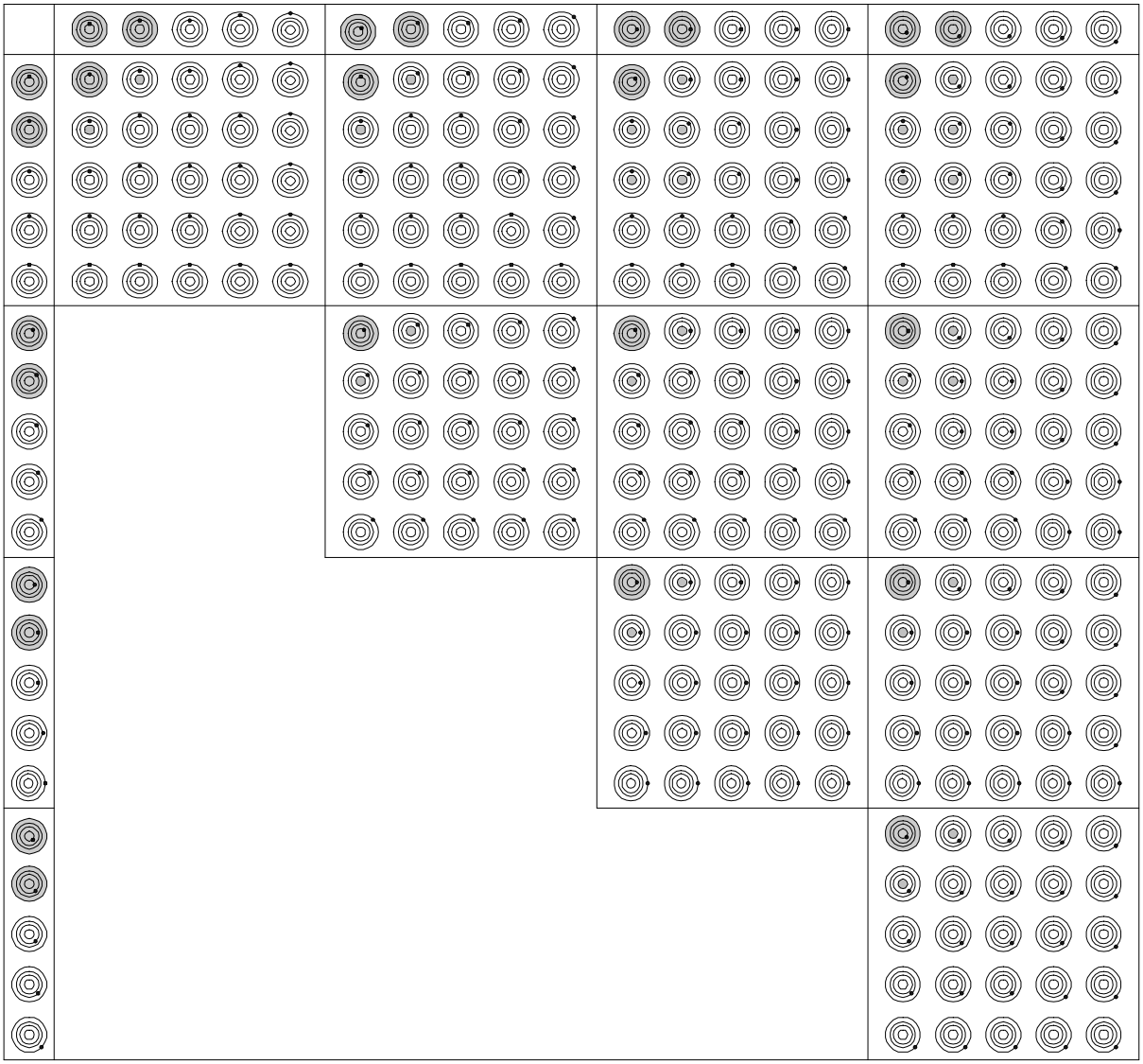


Figura D.1: Tabela de composição para o Ratio 2 (Parte I)

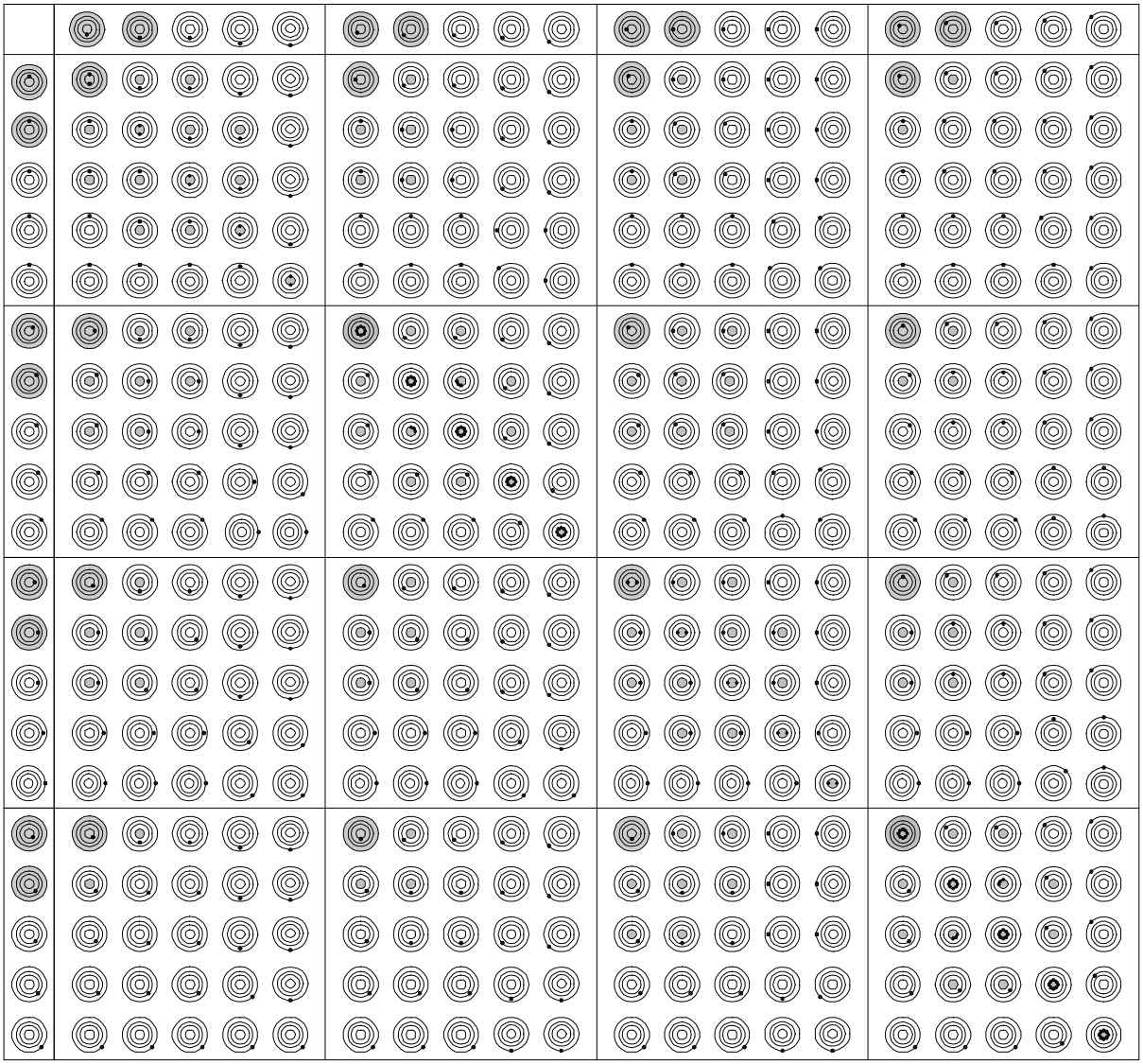


Figura D.2: Tabela de composição para o Ratio 2 (Parte II)

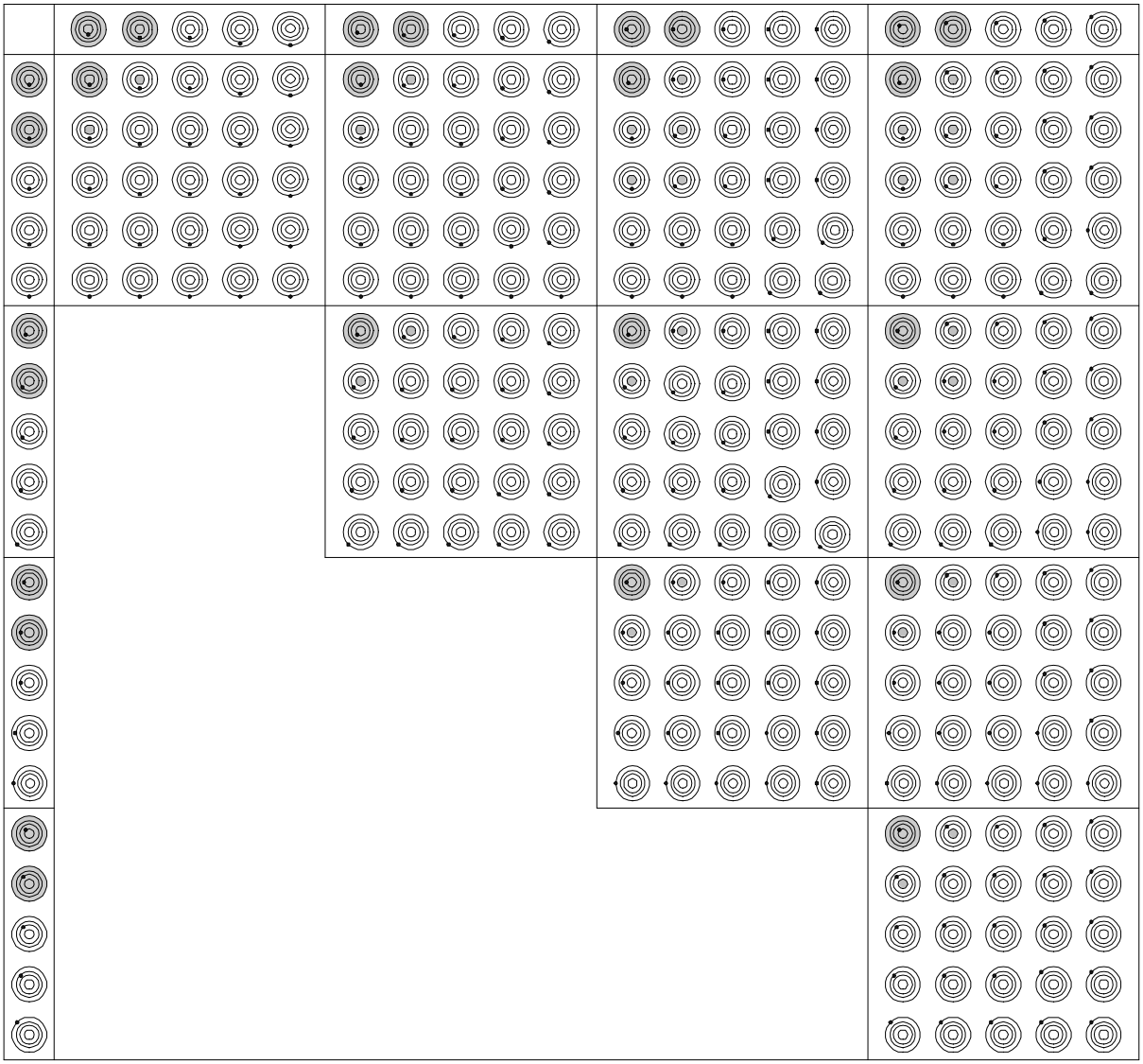


Figura D.3: Tabela de composição para o Ratio 2 (Parte III)

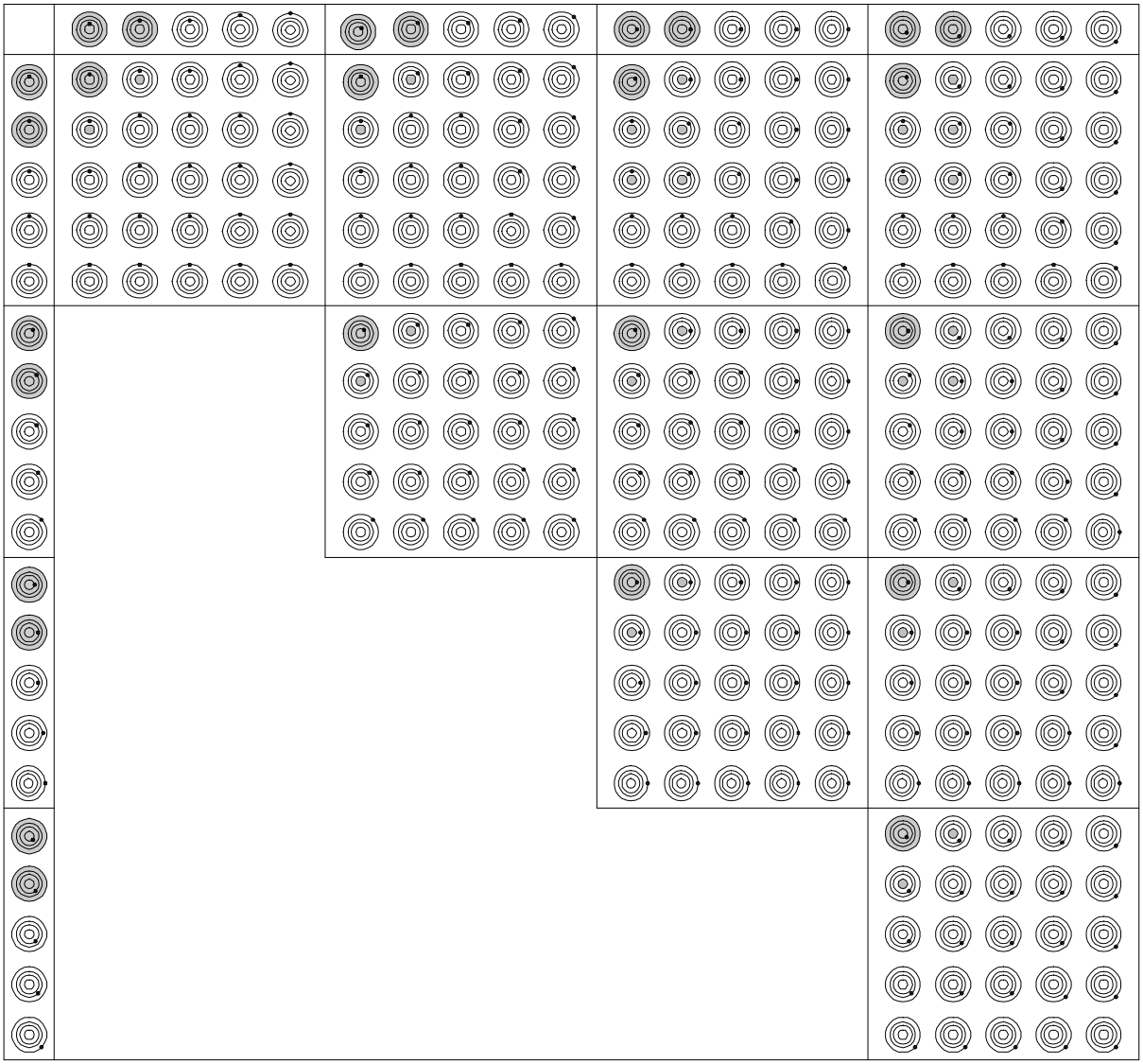


Figura D.4: Tabela de composição para o Ratio 4 (Parte I)

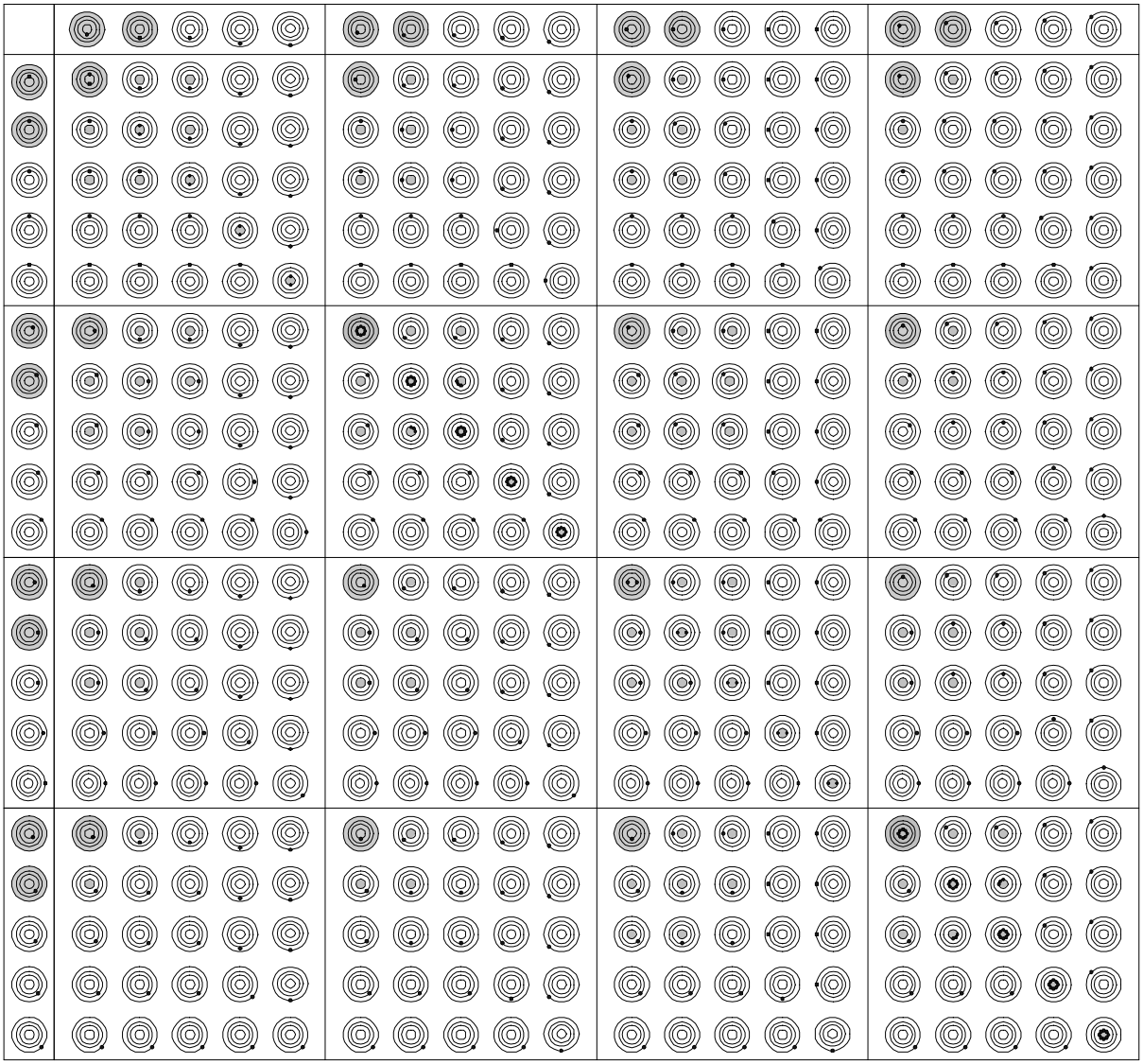


Figura D.5: Tabela de composição para o Ratio 4 (Parte II)

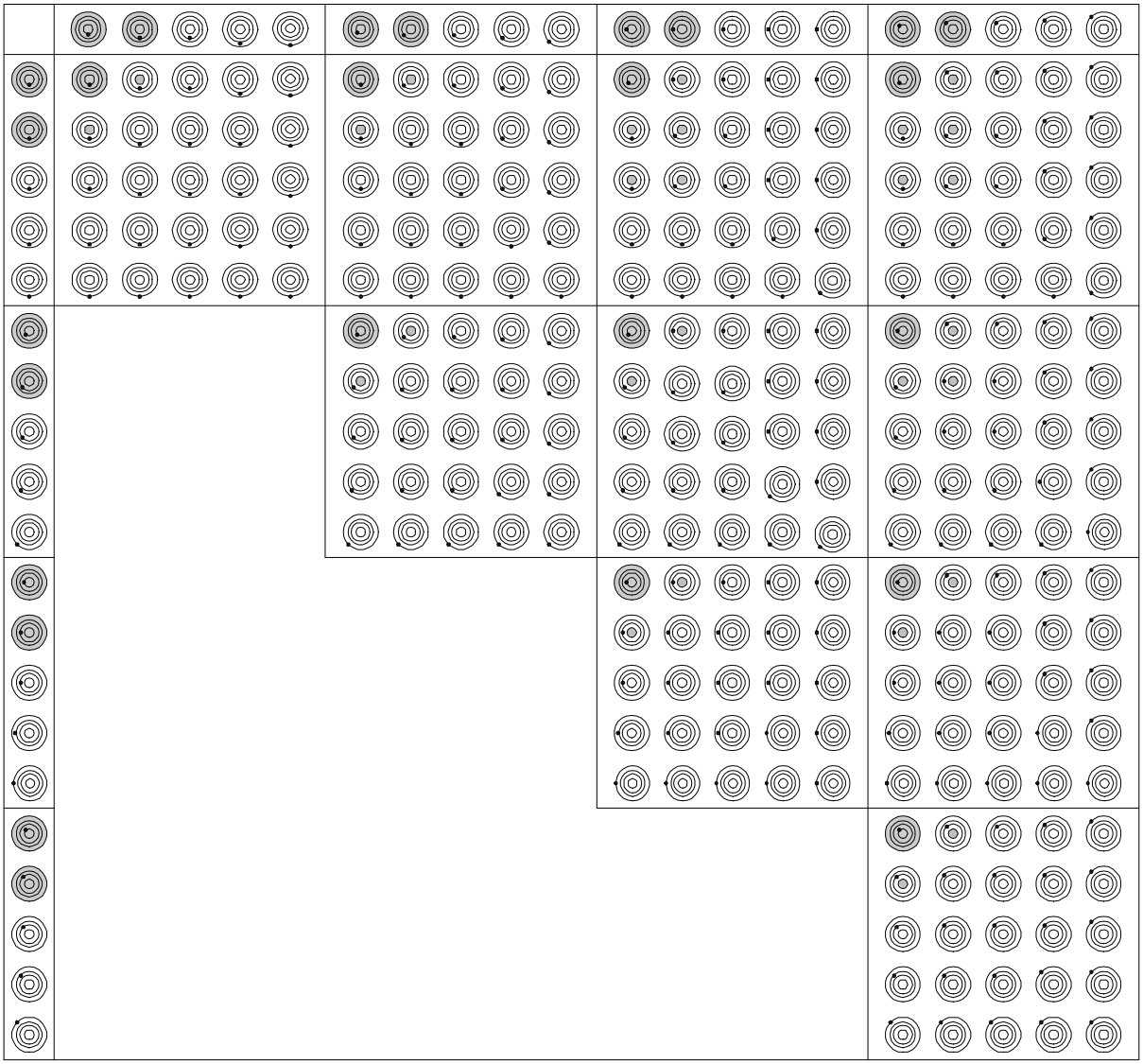


Figura D.6: Tabela de composição para o Ratio 4 (Parte III)

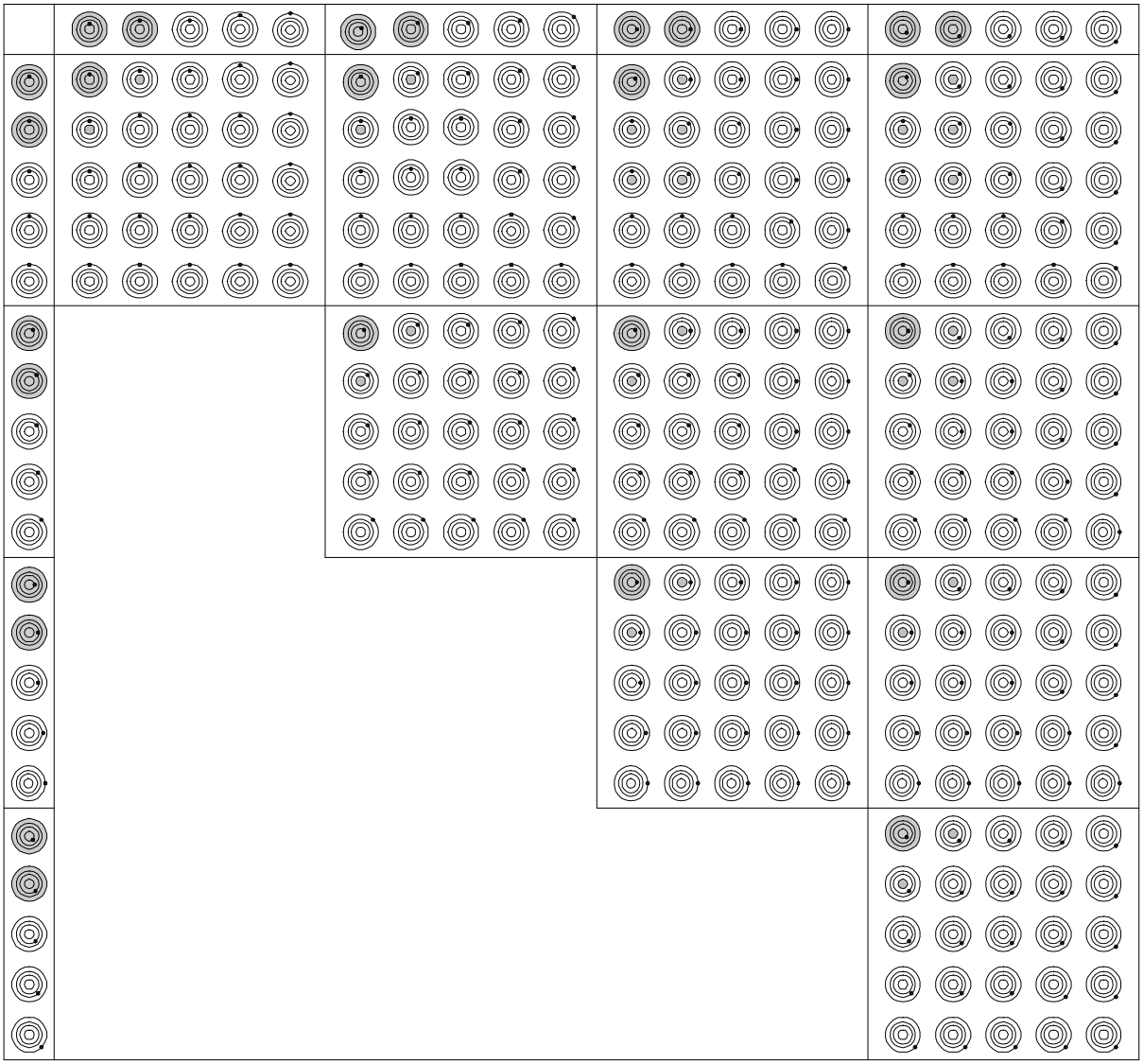


Figura D.7: Tabela de composição para o Ratio 5 (Parte I)

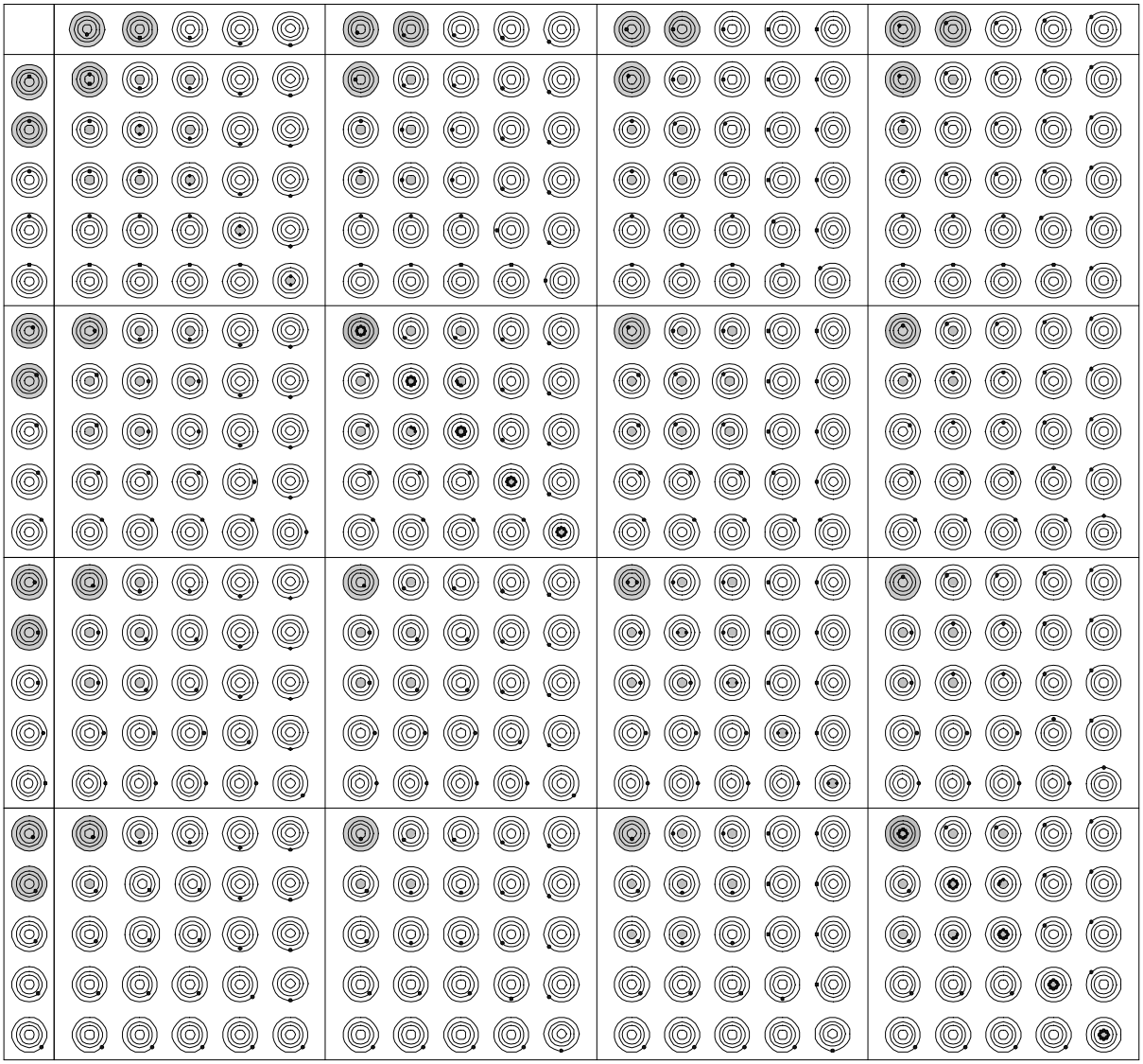


Figura D.8: Tabela de composição para o Ratio 5 (Parte II)

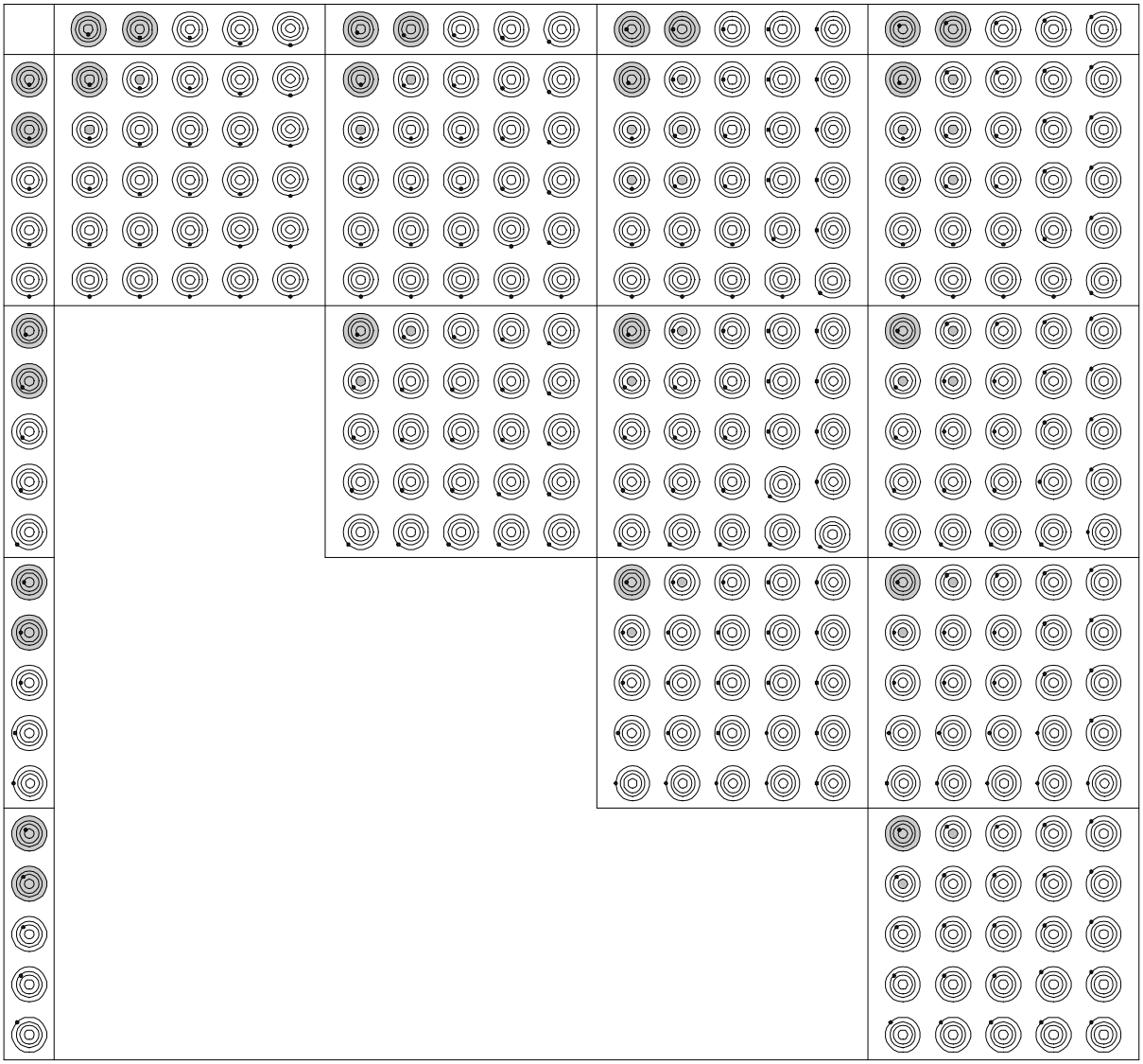


Figura D.9: Tabela de composição para o Ratio 5 (Parte III)

D.5 Integração da dimensão das regiões no processo de raciocínio

O tamanho das regiões assume um papel determinante na identi…cação da direcção existente entre regiões, principalmente nos grupos de direcções ¢dir1 e ¢dir3 com distâncias qualitativas iguais. As próximas secções apresentam as regras de inferência obtidas para estes conjuntos, considerando, no processo de raciocínio, a dimensão das regiões envolvidas.

D.5.1 Grupo de composições para ¢dir1

A Tabela D.16 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, para o grupo de composições associadas às direcções caracterizadas em ¢dir1, com o caso particular das distâncias mp;mp (ratio 2). Para as restantes composições, as inferências, em termos de direcção, são obtidas por rotação das apresentadas na Tabela D.16, enquanto que em termos de distância mantêm-se as apresentadas na secção D.2.

D.5.2 Grupo de composições para ¢dir3

A Tabela D.17 apresenta os cálculos quantitativos necessários à identi…cação das regras de inferência, para o grupo de composições associadas às direcções caracterizadas em ¢dir3, com o caso particular das distâncias mp;mp (ratio 2). Para as restantes composições, as inferências, em termos de direcção, são obtidas por rotação das apresentadas na Tabela D.17, enquanto que em termos de distância mantêm-se as apresentadas na secção D.2.

D.5.3 Regras de composição que integram a dimensão das regiões

Após a determinação das regras de inferência para os grupos ¢dir1 e ¢dir3, apresentadas nas subsecções anteriores, estas foram incluídas no Clementine através de diversos nodos Filler (agregados em dois super nodos, SuperNodeDir1 e SuperNodeDir3), que permitem recti…car as inferências obtidas, com a tabela de composição, veri…cando o tamanho das regiões e a sua in‡uência na direcção existente entre as mesmas. A Figura D.10 apresenta as diversas regras utilizadas, as quais estão listadas no Annotation Editor de cada um dos respectivos super nodos.

D.5.4 Análise à dimensão das regiões

Nesta subsecção são apresentados alguns cálculos matemáticos, que permitiram a análise por distrito, da dimensão das regiões que os integram. Pela análise das Tabelas D.18, D.19 e D.20 é possível veri…car que Santarém é o distrito que apresenta a maior diferença (valor máximo e valor mínimo), entre a dimensão dos seus concelhos.

(N,mp);(NE,mp)

A>B e B>C	V_{AB}	V_{BC}
distância	0,6	0,4
direcção	180	225
Ang_{AC}=	17,76428	
$\|V_{AC}\|$=	0,927044	
	N,mp	

A>B e B<C e A=C	V_{AB}	V_{BC}
distância	0,6	0,6
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	1,108655	
	N,mp	

A>B e B=C	V_{AB}	V_{BC}
distância	0,6	0,5
direcção	180	225
Ang_{AC}=	20,3435	
$\|V_{AC}\|$=	1,016988	
	N,mp	

A<B e B>C e A=C	V_{AB}	V_{BC}
distância	0,4	0,4
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,739104	
	N,mp	

A<B e B<C	V_{AB}	V_{BC}
distância	0,4	0,6
direcção	180	225
Ang_{AC}=	27,23572	
$\|V_{AC}\|$=	0,927044	
	NE,mp	

A<B e B=C	V_{AB}	V_{BC}
distância	0,4	0,5
direcção	180	225
Ang_{AC}=	25,13511	
$\|V_{AC}\|$=	0,832372	
	NE,mp	

A=B e B>C	V_{AB}	V_{BC}
distância	0,5	0,4
direcção	180	225
Ang_{AC}=	19,86489	
$\|V_{AC}\|$=	0,832372	
	N,mp	

A=B e B<C	V_{AB}	V_{BC}
distância	0,5	0,6
direcção	180	225
Ang_{AC}=	24,6565	
$\|V_{AC}\|$=	1,016988	
	NE,p	

A=B e B=C	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	225
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	N,mp	

(NE,mp);(N,mp)

A>B e B>C	V_{AB}	V_{BC}
distância	0,6	0,4
direcção	225	180
Ang_{AC}=	27,23572	
$\|V_{AC}\|$=	0,927044	
	NE,mp	

A>B e B<C e A=C	V_{AB}	V_{BC}
distância	0,6	0,6
direcção	225	180
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	1,108655	
	N,mp	

A>B e B=C	V_{AB}	V_{BC}
distância	0,6	0,5
direcção	228	180
Ang_{AC}=	26,3178	
$\|V_{AC}\|$=	1,005723	
	NE,mp	

A<B e B>C e A=C	V_{AB}	V_{BC}
distância	0,4	0,4
direcção	225	180
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,739104	
	N,mp	

A<B e B<C	V_{AB}	V_{BC}
distância	0,4	0,6
direcção	225	180
Ang_{AC}=	17,76428	
$\|V_{AC}\|$=	0,927044	
	N,mp	

A<B e B=C	V_{AB}	V_{BC}
distância	0,4	0,5
direcção	225	180
Ang_{AC}=	19,86489	
$\|V_{AC}\|$=	0,832372	
	N,mp	

A=B e B>C	V_{AB}	V_{BC}
distância	0,5	0,4
direcção	225	180
Ang_{AC}=	25,13511	
$\|V_{AC}\|$=	0,832372	
	NE,mp	

A=B e B<C	V_{AB}	V_{BC}
distância	0,5	0,6
direcção	225	180
Ang_{AC}=	20,3435	
$\|V_{AC}\|$=	1,016988	
	N,p	

A=B e B=C	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	225	180
Ang_{AC}=	22,5	
$\|V_{AC}\|$=	0,92388	
	N,mp	

Tabela D.16: Cálculos quantitativos para o grupo de direcções ¢dir1, ratio 2, considerando a dimensão das regiões

(N,mp);(SE,mp)

A>B e B>C		
	V_{AB}	V_{BC}
distância	0,6	0,4
direcção	180	315
Ang_{AC}=	41,72677	
$\|V_{AC}\|$=	0,424957	
	NE,mp	

A>B e B<C e A=C		
	V_{AB}	V_{BC}
distância	0,6	0,6
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,45922	
	NE,mp	

A>B e B=C		
	V_{AB}	V_{BC}
distância	0,6	0,5
direcção	180	315
Ang_{AC}=	55,12133	
$\|V_{AC}\|$=	0,430971	
	NE,mp	

A<B e B>C e A=C		
	V_{AB}	V_{BC}
distância	0,4	0,4
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,306147	
	NE,mp	

A<B e B<C		
	V_{AB}	V_{BC}
distância	0,4	0,6
direcção	180	315
Ang_{AC}=	93,27323	
$\|V_{AC}\|$=	0,424957	
	E,mp	

A<B e B=C		
	V_{AB}	V_{BC}
distância	0,4	0,5
direcção	180	315
Ang_{AC}=	82,51586	
$\|V_{AC}\|$=	0,356591	
	E,mp	

A=B e B>C		
	V_{AB}	V_{BC}
distância	0,5	0,4
direcção	180	315
Ang_{AC}=	52,48414	
$\|V_{AC}\|$=	0,356591	
	NE,mp	

A=B e B<C		
	V_{AB}	V_{BC}
distância	0,5	0,6
direcção	180	315
Ang_{AC}=	79,87867	
$\|V_{AC}\|$=	0,430971	
	E,mp	

A=B e B=C		
	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	180	315
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,382683	
	NE,mp	

(SE,mp);(N,mp)

A>B e B>C		
	V_{AB}	V_{BC}
distância	0,6	0,4
direcção	315	180
Ang_{AC}=	93,27323	
$\|V_{AC}\|$=	0,424957	
	E,mp	

A>B e B<C e A=C		
	V_{AB}	V_{BC}
distância	0,6	0,6
direcção	315	180
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,45922	
	NE,mp	

A>B e B=C		
	V_{AB}	V_{BC}
distância	0,6	0,5
direcção	315	180
Ang_{AC}=	79,87867	
$\|V_{AC}\|$=	0,430971	
	E,mp	

A<B e B>C e A=C		
	V_{AB}	V_{BC}
distância	0,4	0,4
direcção	315	180
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,306147	
	NE,mp	

A<B e B<C		
	V_{AB}	V_{BC}
distância	0,4	0,6
direcção	315	180
Ang_{AC}=	41,72677	
$\|V_{AC}\|$=	0,424957	
	NE,mp	

A<B e B=C		
	V_{AB}	V_{BC}
distância	0,4	0,5
direcção	315	180
Ang_{AC}=	52,48414	
$\|V_{AC}\|$=	0,356591	
	NE,mp	

A=B e B>C		
	V_{AB}	V_{BC}
distância	0,5	0,4
direcção	315	180
Ang_{AC}=	82,51586	
$\|V_{AC}\|$=	0,356591	
	E,mp	

A=B e B<C		
	V_{AB}	V_{BC}
distância	0,5	0,6
direcção	315	180
Ang_{AC}=	55,12133	
$\|V_{AC}\|$=	0,430971	
	NE,mp	

A=B e B=C		
	V_{AB}	V_{BC}
distância	0,5	0,5
direcção	315	180
Ang_{AC}=	67,5	
$\|V_{AC}\|$=	0,382683	
	NE,mp	

Tabela D.17: Cálculos quantitativos para o grupo de direcções ¢dir3, `ratio 2`, considerando a dimensão das regiões

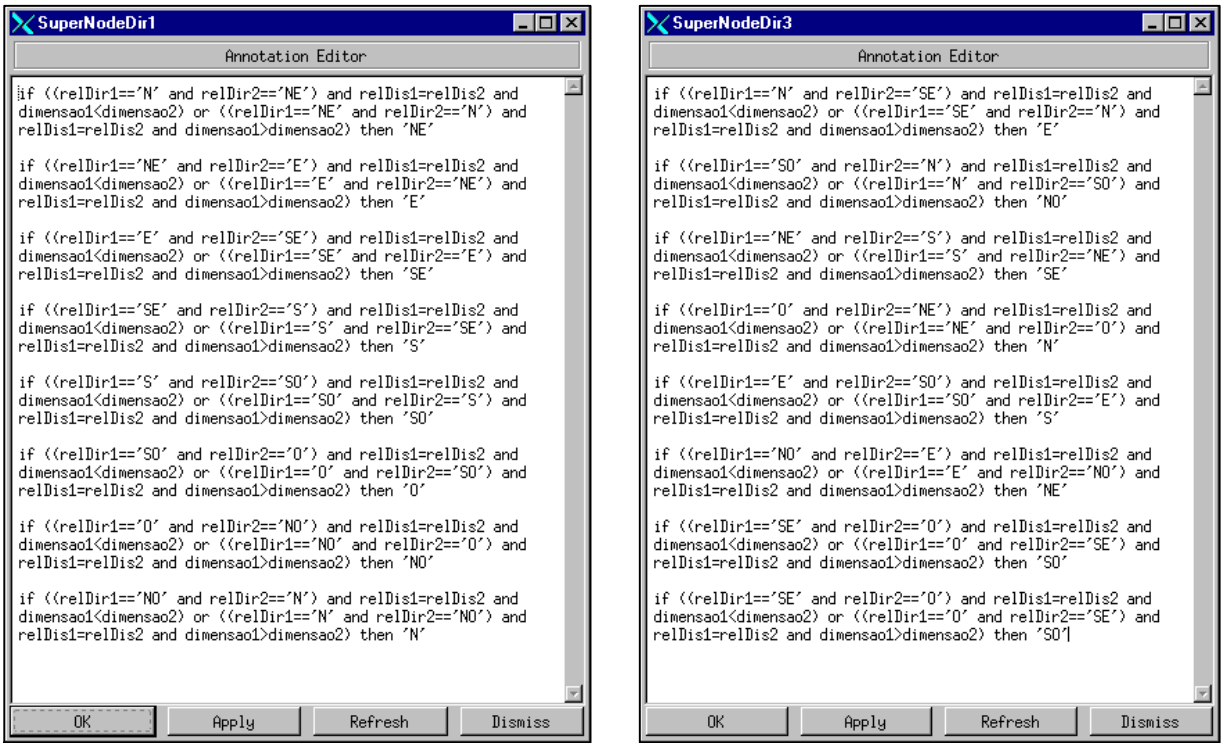


Figura D.10: Regras que permitem a integração da dimensão das regiões no processo de inferência

Concelho	Área (km2)		
101	326,02	**Aveiro**	
102	157,01		
103	215,58	Média	145,95
104	325,94	Mínimo	8,58
105	196,41	Máximo	326,02
106	119,87		
107	18,62	**Ratio Min/Max**	**0,03**
108	109,15		
109	217,23		
110	72,65		
111	109,74		
112	71,41		
113	161,89		
114	87,64		
115	143,59		
116	8,58		
117	125,62		
118	159,83		
119	146,20		
201	455,49	**Beja**	
202	764,40		
203	275,81	Média	730,56
204	179,18	Mínimo	167,99
205	1138,35	Máximo	1697,92
206	567,68		
207	167,99	**Ratio Min/Max**	**0,10**
208	656,17		
209	1281,11		
210	964,01		
211	1697,92		
212	672,13		
213	1086,82		
214	320,79		
301	76,80	**Braga**	
302	375,86		
303	185,24	Média	207,44
304	243,70	Mínimo	76,80
305	181,02	Máximo	375,86
306	93,31		
307	218,01	**Ratio Min/Max**	**0,20**
308	249,74		
309	133,99		
310	279,53		
311	228,95		
312	200,68		
313	229,94		
401	313,49	**Bragança**	
402	1205,80		
403	272,11	Média	552,12
404	225,47	Mínimo	225,47
405	700,26	Máximo	1205,80
406	490,82		
407	661,22	**Ratio Min/Max**	**0,19**
408	750,29		
409	525,66		
410	266,95		
411	501,56		
412	711,84		

Concelho	Área (km2)		
501	123,43	**Castelo Branco**	
502	1414,86		
503	508,43	Média	599,47
504	752,15	Mínimo	123,43
505	1389,84	Máximo	1414,86
506	479,19		
507	572,51	**Ratio Min/Max**	**0,09**
508	386,50		
509	456,37		
510	184,00		
511	326,86		
601	334,19	**Coimbra**	
602	395,31		
603	302,28	Média	231,40
604	141,87	Mínimo	116,87
605	358,01	Máximo	395,31
606	262,39		
607	139,74	**Ratio Min/Max**	**0,30**
608	116,87		
609	127,57		
610	230,19		
611	237,20		
612	377,77		
613	243,04		
614	127,49		
615	251,67		
616	204,68		
617	83,50		
701	546,16	**Évora**	
702	688,10		
703	148,14	Média	527,08
704	521,14	Mínimo	148,14
705	1296,73	Máximo	1296,73
706	1224,89		
707	442,34	**Ratio Min/Max**	**0,11**
708	281,55		
709	597,82		
710	366,69		
711	453,36		
712	233,39		
713	377,66		
714	201,17		
801	136,44	**Faro**	
802	587,11		
803	313,94	Média	312,31
804	297,94	Mínimo	60,62
805	198,21	Máximo	758,14
806	87,21		
807	215,44	**Ratio Min/Max**	**0,08**
808	758,14		
809	399,24		
810	133,29		
811	179,13		
812	148,93		
813	688,58		
814	617,83		
815	174,85		
816	60,62		

Tabela D.18: Dimensão das regiões que integram cada distrito

Concelho	Área (km2)		
902	520,38		
903	239,65	Média	411,42
904	517,78	Mínimo	106,70
905	135,05	Máximo	817,60
906	298,42		
907	727,05	**Ratio Min/Max**	**0,13**
908	106,70		
909	287,44		
910	493,68		
911	817,60		
912	434,63		
913	364,22		
914	405,83		
1001	485,62	**Leiria**	
1002	155,04		
1003	184,62	Média	221,93
1004	99,75	Mínimo	66,94
1005	86,51	Máximo	630,91
1006	253,70		
1007	66,94	**Ratio Min/Max**	**0,11**
1008	171,80		
1009	555,92		
1010	176,24		
1011	76,48		
1012	147,70		
1013	126,26		
1014	72,01		
1015	630,91		
1016	261,43		
1101	304,49	**Lisboa**	
1102	75,78		
1103	249,09	Média	180,66
1104	177,65	Mínimo	26,94
1105	92,50	Máximo	411,96
1106	82,24		
1107	196,00	**Ratio Min/Max**	**0,07**
1108	146,23		
1109	289,43		
1110	41,98		
1111	312,28		
1112	51,30		
1113	411,96		
1114	252,10		
1115	26,94		
1201	360,00	**Portalegre**	
1202	325,27		
1203	599,59	Média	405,70
1204	254,96	Mínimo	155,08
1205	269,73	Máximo	837,27
1206	406,19		
1207	629,19	**Ratio Min/Max**	**0,19**
1208	245,30		
1209	290,49		
1210	155,08		
1211	414,89		
1212	578,48		
1213	837,27		
1214	451,36		
1215	267,72		

Concelho	Área (km2)		
1302	166,19		
1303	122,94	Média	124,52
1304	130,11	Mínimo	33,43
1305	99,77	Máximo	214,41
1306	83,94		
1307	197,21	**Ratio Min/Max**	**0,16**
1308	53,57		
1309	73,92		
1310	154,46		
1311	209,91		
1312	33,43		
1313	76,78		
1314	214,41		
1315	74,37		
1316	139,74		
1317	161,65		
1401	721,69	**Santarém**	
1402	128,92		
1403	221,82	Média	321,39
1404	102,49	Mínimo	13,52
1405	541,73	Máximo	1120,47
1406	152,21		
1407	747,02	**Ratio Min/Max**	**0,01**
1408	80,22		
1409	1120,47		
1410	13,52		
1411	193,49		
1412	73,28		
1413	403,48		
1414	273,99		
1415	249,43		
1416	554,07		
1417	93,30		
1418	347,38		
1419	268,61		
1420	46,48		
1421	415,52		
1501	1422,10	**Setúbal**	
1502	113,10		
1503	70,71	Média	356,78
1504	33,75	Mínimo	33,75
1505	795,48	Máximo	1422,10
1506	56,62		
1507	51,97	**Ratio Min/Max**	**0,02**
1507	276,46		
1508	461,81		
1509	1071,14		
1510	99,48		
1511	193,80		
1512	169,78		
1513	178,77		
1602	120,43		
1603	252,23	Média	248,25
1604	214,50	Mínimo	100,48
1605	140,27	Máximo	613,05
1606	191,03		
1607	319,14	**Ratio Min/Max**	**0,16**
1608	112,90		
1609	302,11		
1610	100,48		

Tabela D.19: Dimensão das regiões que integram cada distrito, continuação

Concelho	Área (km2)		
1701	287,42	**Vila Real**	
1702	325,43		
1703	613,05	Média	309,29
1704	27,38	Mínimo	27,38
1705	168,49	Máximo	829,96
1706	829,96		
1707	190,81	**Ratio Min/Max**	**0,03**
1708	90,18		
1709	245,70		
1710	155,27		
1711	67,73		
1712	545,32		
1713	402,59		
1714	380,76		
1801	108,77	**Viseu**	
1802	111,34		
1803	379,17	Média	207,93
1804	240,46	Mínimo	102,53
1805	172,16	Máximo	507,08
1806	215,53		
1807	219,54	**Ratio Min/Max**	**0,20**
1808	248,17		
1809	119,96		
1810	150,65		
1811	140,86		
1812	131,27		
1813	126,59		
1814	102,53		
1815	264,80		
1816	342,00		
1817	200,21		
1818	227,74		
1819	142,36		
1820	103,62		
1821	367,22		
1822	171,65		
1823	507,08		
1824	196,66		

Tabela D.20: Dimensão das regiões que integram cada distrito, continuação