

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

**Área Departamental de Engenharia de Electrónica e Telecomunicações e de
Computadores**

**Extração de conhecimento em bases de dados geográficas
com recurso a geo-ontologia**

ANTÓNIO JOSÉ DA SILVA PAREDES

Licenciado

Dissertação para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

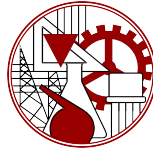
Orientador : Doutor Paulo Manuel Trigo Cândido da Silva

Júri:

Presidente: Doutor Manuel Martins Barata

Vogais: Doutor Gonçalo Caetano Marques
Doutor Paulo Manuel Trigo Cândido da Silva

outubro, 2017



INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

**Área Departamental de Engenharia de Electrónica e Telecomunicações e de
Computadores**

**Extração de conhecimento em bases de dados geográficas
com recurso a geo-ontologia**

ANTÓNIO JOSÉ DA SILVA PAREDES

Licenciado

Dissertação para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientador : Doutor Paulo Manuel Trigo Cândido da Silva

Júri:

Presidente: Doutor Manuel Martins Barata

Vogais: Doutor Gonçalo Caetano Marques
Doutor Paulo Manuel Trigo Cândido da Silva

outubro, 2017

Agradecimentos

Ao Doutor Paulo Trigo, pela disponibilidade e pela qualidade do apoio prestado, ao longo do desenvolvimento do presente trabalho..

À Adelaide, pelo incentivo e motivação, ao longo deste percurso.

À minha mãe, pela compreensão da minha ausência, tantas vezes sentida, durante esta etapa.

Resumo

A extração de conhecimento de dados geolocalizados (*geo-mining*) tem contribuído para a resolução de problemas existentes em diversos domínios tais como, na saúde, meteorologia ou segurança, *e.g.*, na identificação de locais de concentração de crimes como meio de prevenção, e também no planeamento urbano ou proteção civil, *e.g.*, na elaboração de modelos de riscos de incêndio como suporte ao planeamento das ações de busca e salvamento.

Neste trabalho desenvolve-se uma plataforma que a partir de dados geolocalizados extrai conhecimento sob a forma de regras de associação espacial. As regras extraídas são constituídas por relacionamentos espaciais, topológicos e de distância. Os relacionamentos são obtidos entre entidades espaciais existentes na superfície terrestre e elementos espaciais que caracterizam um fenómeno, *e.g.*, um foco de gripe aviária ou acidentes rodoviários, adicionados de um *buffer* espacial de dimensão variável. A informação geoespacial utilizada foi obtida do projeto colaborativo *OpenStreetMap*. Os dados *OpenStreetMap* são coletados a partir de colaboração voluntária segundo o conceito *Volunteered Geographic Information*.

Um processo de extração de regras de associação produz, normalmente, um elevado número de regras de associação o que dificulta a perceção de novo conhecimento. Para abordar este problema é constituída uma geo-ontologia, com relacionamentos espaciais detetados '*a priori*', que originarão padrões conhecidos. Estes relacionamentos são obtidos através da análise do *dataset* produzido. A plataforma desenvolvida permite também integrar conhecimento do utilizador. A metodologia utilizada para extração de regras de associação espacial permite obter do espaço de pesquisa, um número de regras que não excede um valor definido pelo utilizador.

Apresenta-se um estudo de caso sobre o problema da gripe aviária. As experiências realizadas mostram resultados que seguem um padrão que foi confrontado, com documentação especializada na área, que validam os resultados obtidos.

Palavras-chave: bases de dados geográficas, *data mining*, geo-ontologias.

Abstract

The extraction of geo-mining data has contributed towards the resolution of issues concerning a variety of domains, such as health, meteorology or security, for example, in identifying crime concentration sites, as a means of prevention and also in urban planning or civil protection, i.e., in preparing fire risk models as support towards the planning of search and rescue measures.

This study focuses on the development of a platform that extracts knowledge in the form of rules of spatial association, taken from geo-mining data. The extracted rules consist of spatial, topological and distance relationships. The relationships are obtained from spatial entities that exist on the terrestrial surface and spatial elements that characterise a phenomenon, *e.g.*, a spout of avian flu or road accidents, with the addition of a spatial buffer of variable size. The geospatial information used was obtained from a collaborative project - *OpenStreetMap*. The *OpenStreetMap* data is collected according to the concept of *Volunteered Geographic Information*, taken from voluntary cooperation.

A given process of extracting rules of association, usually produces a high number of rules of association, which makes it difficult to perceive new knowledge. In order to address this issue, a geo-ontology is created, with spatial relationships detected beforehand, which will originate known patterns. These relationships are obtained through the analysis of the dataset produced. The developed platform also enables the integration of user knowledge. The methodology used for extracting rules of association enables one to obtain a series of rules, from the search area, that do not exceed a user-defined value.

Shown here, is a case study on the problem of avian flu. The experiences conducted show results that follow a pattern, which were then compared against documentation from experts in this field, thus validating the results obtained.

Keywords: geographic databases, data mining, geo-ontologies.

Índice

Índice	xi
Lista de Figuras	xv
Lista de Tabelas	xvii
Lista de Algoritmos	xxi
Glossário	xxiii
Siglas	xxiii
1 Introdução	1
1.1 Organização do documento	3
2 Trabalho relacionado	5
2.1 Fases do processo de descoberta de conhecimento	5
2.2 O conceito de <i>Dataset</i>	6
2.3 <i>Market basket analysis</i>	8
2.3.1 Regras de associação	10
2.3.2 <i>Suporte, confiança e lift</i>	10
2.4 Sistema de suporte à informação espacial	11
2.5 Tipos de objetos espaciais	12
2.6 Modelo de dados espacial	12

2.7	Operadores topológicos, de distância e de direção	13
2.8	Seleção de operadores topológicos	14
2.9	Níveis de granularidade	15
2.10	<i>Market basket analysis</i> no ‘ <i>mundo geo</i> ’	16
2.11	Extração de conhecimento	17
2.12	Aprendizagem supervisionada e aprendizagem não supervisionada	19
3	Modelo proposto	21
3.1	Requisitos	21
3.2	<i>Dataset</i> - Modelo teórico	22
3.3	Arquitetura do modelo conceptual	23
3.4	Elementos do modelo conceptual	25
3.5	Geo-ontologia	27
3.5.1	Análise de dependências	27
3.5.2	Análise formal	29
3.6	Processo de pesquisa de regras de associação	31
4	Implementação	33
4.1	Arquitetura do modelo implementado	33
4.1.1	Domínio dependente (DD)	33
4.1.2	Domínio independente (DInD)	35
4.1.3	Definição de linhas e colunas da matriz	35
4.1.4	Operadores topológicos e de distância	36
4.1.5	Especificação de pontos de surto	37
4.1.6	Exemplo de <i>dataset</i>	38
4.1.7	Construção da geo-ontologia	39
4.1.8	Filtro do processo de extração	41
4.1.9	Algoritmo de pesquisa	41
4.1.10	Variação significativa do número de regras	43
4.1.11	Utilização da geo-ontologia	46

<i>ÍNDICE</i>	xiii
4.2 Informação geográfica	46
4.2.1 Modelo de dados do OpenStreetMap	46
4.2.2 Processamento de dados <i>OpenStreetMap</i>	47
4.2.3 Base de dados espacial	49
5 Estudo de caso	51
5.1 A infecção da gripe aviária	51
5.2 Dados do fenómeno	54
5.3 Experiências	54
5.3.1 Experiência 1	59
5.3.2 Experiência 2	60
5.3.3 Experiência 3	61
5.3.4 Experiência 4	62
5.4 Validação dos resultados	63
6 Conclusões e trabalho futuro	67
Referências	71
A Função do processo de filtragem	i
B Ficheiro de configuração da <i>Framework IMPOSM3</i>	iii
C Resultados da experiência 1	vii
D Resultados da experiência 2	xiii
E Resultados da experiência 3	xvii
F Resultados da experiência 4	xxiii

Lista de Figuras

2.1	Fases do modelo de referência CRISP-DM.	6
2.2	Tipos básicos de abstrações: ponto, linha e polígono.	12
2.3	Exemplo de modelo de dados espacial.	13
2.4	Algumas relações topológicas, de distância e de direção.	14
2.5	Hierarquias de conceito e de operadores.	15
2.6	Um exemplo de taxonomia.	16
3.1	Arquitetura do modelo conceptual.	24
4.1	Arquitetura do bloco <i>Domain Dependent</i>	34
4.2	Arquitetura do bloco <i>Domain Independent</i>	35
4.3	Camadas geográficas extraídas relativas a um ponto geográfico de surto.	48
5.1	Áreas não disjuntas concêntricas em pontos de surto.	62

Lista de Tabelas

2.1	Atividades desenvolvidas em cada fase do processo CRISP-DM . . .	7
2.2	Registo de transações.	9
2.3	Matriz esparsa com a mesma informação da tabela 2.2.	9
2.4	Representação compacta da informação da tabela 2.2.	9
2.5	Exemplo de <i>dataset</i>	17
2.6	Relações topológicas possíveis e obrigatórias.	19
2.7	Relações topológicas possíveis.	19
3.1	<i>Itemsets</i> frequentes com suporte 50%.	28
3.2	Algumas regras geradas com a dependência estudada.	28
3.3	Sequência de pontos analisados.	31
4.1	Operadores selecionados	39
4.2	Exemplo de <i>dataset</i>	39
4.3	Caracterização de situações de análise.	44
4.4	Relação de tabelas existentes na base de dados espacial.	49
5.1	Número de ocorrências por países.	55
5.2	Posicionamento relativo das áreas concêntricas em pontos de surto.	56
5.3	Configuração da função filtro do processo de extração.	58
5.4	Configuração de linhas e colunas da matriz.	59

5.5	Camadas relacionadas com padrões espaciais relevantes extraídos na primeira fase da experiência 1.	59
5.6	Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 1.	60
5.7	Valores de configuração de linhas e colunas da matriz.	60
5.8	Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 2.	61
5.9	Valores de configuração de linhas e colunas da matriz.	61
5.10	Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 3.	61
5.11	Número de ocorrências utilizado na experiência 4.	63
5.12	Valores de configuração de linhas e colunas da matriz.	63
5.13	Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 4.	63
C.1	Regras de associação espacial obtidas em zonas de surto do tipo <i>poultry</i> e do tipo <i>wild_birds</i>	vii
C.2	Regras de associação espacial obtidas em zonas de surto do tipo <i>poultry</i> e do tipo <i>wild_birds</i> (continuação da tabela C.1).	viii
C.3	Geo-ontologias utilizadas na extração de regras de associação espacial em zonas de surto do tipo <i>poultry</i> e do tipo <i>wild_birds</i>	ix
C.4	Geo-ontologia utilizada na extração de regras de associação espacial em zonas de surto do tipo <i>poultry</i> e do tipo <i>wild_birds</i> (continuação da tabela C.3).	x
C.5	Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela C.3.	xi
C.6	Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas na geo-ontologia mostrada na tabela C.4 (continuação da tabela C.5)	xii
D.1	Regras de associação espacial obtidas em zonas de surto do tipo <i>poultry</i>	xiii

D.2	Geo-ontologias utilizadas na extração de regras de associação espacial em zonas de surtos do tipo <i>poultry</i>	xiv
D.3	Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela D.2.	xv
D.4	Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela D.2.	xvi
E.1	Regras de associação espacial obtidas em zonas de surto do tipo <i>wild_birds</i>	xvii
E.2	Regras de associação espacial obtidas em zonas de surto do tipo <i>wild_birds</i>	xviii
E.3	Geo-ontologias utilizadas na extração de regras de associação espacial em zonas de surtos do tipo <i>wild_birds</i>	xix
E.4	Geo-ontologia utilizada na extração de regras de associação espacial em zonas de surtos do tipo <i>wild_birds</i> (continuação da tabela E.3).	xx
E.5	Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas nas tabelas E.3 e E.4.	xxi
F.1	Regras de associação espacial obtidas de zonas que intercetam pelo menos uma zona de outro tipo de surto.	xxiii
F.2	Geo-ontologias utilizadas na extração de regras de associação espacial em zonas que intercetam pelo menos uma zona de outro tipo de surto.	xxiv
F.3	Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela F.2.	xxv

Lista de Algoritmos

1	Algoritmo de pesquisa de predicados com 100% de <i>suporte</i>	40
2	Algoritmo de pesquisa de dependências geográficas.	40
3	Algoritmo de pesquisa de regras de associação.	42
4	Algoritmo de obtenção de regras de associação.	43
5	Algoritmo de ajuste do valor de suporte.	45
6	Algoritmo de remoção de predicados com 100% de <i>suporte</i>	46
7	Algoritmo de identificação de regras que contêm dependências. . .	47

Glossário

camada Em SIG, é uma coleção de elementos georreferenciados, todos do mesmo tipo e relacionados com um único tema, *e.g., road.* .

Siglas

CRISP-DM *Cross - Industry Standard Process for Data Mining.*

HPAI *Highly Pathogenic Avian Influenza.*

LPAI *Low Pathogenic Avian Influenza.*

MBA *Market Basket Analysis.*

OGC *Open Geospatial Consortium.*

OSM *OpenStreetMap.*

PBF *Protocolbuffer Binary Format.*

POLY *Polygon Filter File Format.*

SBDE *Sistema de Bases de Dados Espaciais.*

SIG *Sistemas de Informação Geográfica.*

VGI *Volunteered Geographic Information.*

XML *eXtensible Markup Language.*



Introdução

A extração de conhecimento em bases de dados que armazenam dados georreferenciados pode conseguir-se através da extração de regras de associação espacial. Esta técnica tem sido utilizada em domínios tão diversos como a saúde, a segurança, o planeamento urbano ou a proteção civil. No domínio da saúde procuraram-se relações envolvendo as regiões geográficas mais afetadas pela mortalidade provocada pelo cancro (Vinnakota & Lam, 2006). No domínio da segurança, analisou-se a ocorrência e localização do crime como forma de determinar rotas de patrulha policial (Asmai, Roslin, Abdullah, & Ahmad, 2014). No domínio do planeamento urbano utilizaram-se censos georreferenciados para auxiliar a definição de políticas públicas de planeamento urbano e acessibilidade de transportes (Appice, Ceci, Lanza, Lisi, & Malerba, 2003). No domínio da proteção civil procuraram-se construir modelos de risco de incêndios como forma de antecipação de cenários de risco e no planeamento das ações de busca e salvamento (Karasová, Krisp, & Virrantaus, 2005).

Nos últimos anos, assiste-se a uma crescente coleta de dados georreferenciados, para fins comerciais ou conhecimento partilhado. Esta informação para além de poder servir fins específicos, *e.g.*, determinação de rotas, contribui para um conhecimento mais detalhado da superfície terrestre. O projeto *OpenStreetMap* (OSM) constitui um exemplo relevante nesta área (*OpenStreetMap@ONLINE*, n.d.).

O crescimento destas bases de dados, em tamanho e camadas, trouxe a oportunidade de pesquisa de padrões espaciais, analogamente ao que é realizado em uma

base de dados transacional.

Neste trabalho desenvolve-se uma plataforma de extração de conhecimento na forma de regras de associação espacial. As regras de associação extraídas são compostas por relacionamentos espaciais. Os relacionamentos espaciais podem ser topológicos ou de distância. Os relacionamentos são procurados entre recursos espaciais que representam ocorrências de um fenómeno, *e.g.*, gripe aviária, e tipos de recursos geográficos existentes na superfície terrestre, *e.g.*, *roads*.

Os recursos espaciais representativos de ocorrências de fenómenos apresentam-se aumentados espacialmente com um *buffer* de dimensão variável. As regras de associação obtidas são extraídas dos padrões frequentes encontrados nas áreas definidas pelas geometrias que caracterizam o fenómeno aumentadas por um *buffer* espacial imposto.

A descoberta de conhecimento, na forma de regras de associação, tem como característica o fato de gerar um elevado número de regras, o que dificulta a análise dos resultados. Em bases de dados geográficas este problema acentua-se devido às relações naturais intrínsecas à natureza dos dados. No entanto, muitas destas regras são redundantes ou são constituídas por correlações conhecidas, *e.g.*, *contains_port* \Rightarrow *contains_waterbody*. Para diminuir este problema, nesta plataforma, é constituída uma geo-ontologia que agrega relacionamentos obtidos pela análise do *dataset* construído. Pela análise dos relacionamentos presentes na geo-ontologia cabe ao utilizador decidir quais os padrões de regras que devem ser excluídos do resultado.

Com o objetivo de pesquisar regras de associação em todo o espaço de *suporte* e *confiança* e facilitar a análise dos resultados pelo utilizador, foi desenvolvido um algoritmo de pesquisa. Este algoritmo percorre todo o espaço de soluções, analisando o número de regras extraídas em pontos desse espaço. Um ponto neste espaço é caracterizado por (*suporte*, *confiança*). A distância entre pontos consecutivos é parametrizada por um valor de salto em *suporte* ou em *confiança*. A extração de regras realizada pelo algoritmo, em cada ponto, cessa quando é ultrapassado um número de regras máximo, requerido pelo utilizador. Caso o número de regras extraídas não seja superior ao número definido pelo utilizador, as regras são-lhe devolvidas. Caso contrário, são descartadas.

A plataforma desenvolvida é aplicada no estudo do problema da gripe aviária. A gripe aviária é o nome de uma doença causada por um subtipo do vírus *Influenza A*. Este vírus tem a capacidade de contaminar aves e ocasionalmente humanos, podendo causar pandemias (epidemias que se propagam em largam escala). As

ocorrências deste fenómeno são expressas através de pontos geográficos que localizam surtos. As regras de associação extraídas são obtidas de padrões frequentes detetados em áreas concêntricas nos pontos de surto. A área concêntrica, em cada ponto, é determinada pelo ponto geográfico que localiza o surto e por um *buffer* espacial.

A recolha ou criação de informação geográfica, com a finalidade de ser disponibilizada *online*, realizada por cidadãos eventualmente voluntários e sem qualificações, constitui um movimento designado por *Volunteered Geographic Information* (VGI). Um dos projetos de VGI com mais notoriedade é o OSM. Pelo fato dos colaboradores do projeto, eventualmente não disporem das qualificações profissionais necessárias, é possível que estes dados apresentem diferentes níveis de detalhe e exatidão. Como por exemplo, duas zonas distintas previsivelmente semelhantes descritas com diferentes níveis de detalhe. Por outro lado, são passíveis de revelar inconsistências, tais como: designar uma floresta ser diferente do que se encontra numa zona florestal.

A informação geográfica utilizada neste trabalho foi obtida do projeto VGI designado por *OpenStreetMap*, através das coletas disponibilizadas pela organização *Geofabrik*. Esta informação foi segmentada em camadas, *e.g.*, *landusages*, *buildings*, *places*, *transport_areas*, *aeroways*, ..., que podem ser selecionadas de acordo com o objetivo da descoberta.

Na plataforma desenvolvida são procurados relacionamentos topológicos e de distância. Os operadores topológicos estão selecionados de acordo com o tipo de geometria que caracteriza cada camada. O operador de distância é utilizado em todas as camadas. No entanto, esta escolha pode ser alterada pelo utilizador, com o intuito de pesquisar regras de associação espacial em padrões que contenham relacionamentos específicos.

1.1 Organização do documento

Este relatório está organizado em seis capítulos. Nesta secção indicam-se os conteúdos desenvolvidos em cada capítulo.

No capítulo um apresenta-se o problema, descreve-se em linhas gerais as características da plataforma a desenvolver, caracteriza-se o movimento produtor dos dados espaciais utilizados e apresenta-se o estudo de caso.

No capítulo dois analisa-se trabalho relacionado relevante para o desenvolvimento do presente estudo.

No capítulo três apresenta-se e descreve-se a arquitetura do modelo conceptual. Neste capítulo é também apresentada a justificação formal do conteúdo da geontologia e o processo de pesquisa de regras de associação.

No capítulo quatro trata-se da apresentação e descrição do modelo implementado. Ainda neste capítulo, descreve-se o processo utilizado de conversão de informação geográfica para uma base de dados espacial.

O capítulo cinco é iniciado com um estudo sobre o fenómeno da gripe aviária. Seguem-se experiências de utilização da plataforma desenvolvida com informação de surtos em aves do tipo *poultry* e do tipo *wild bird*. Caracterizam-se as experiências realizadas e analisam-se os resultados obtidos.

O capítulo seis apresenta as conclusões do trabalho realizado e aponta trabalho futuro.

2

Trabalho relacionado

Neste capítulo analisa-se trabalho desenvolvido com relevância para este estudo.

2.1 Fases do processo de descoberta de conhecimento

O processo *Cross - Industry Standard Process for Data Mining* (CRISP-DM) é não-proprietário, documentado e oferece uma estrutura para o desenvolvimento de projetos que envolvam tarefas de *data mining* (Shearer, 2000; Shafique & Qaiser, 2014). O processo CRISP-DM envolve seis fases que são ilustradas na figura 2.1.

A seguir, apresentam-se as atividades a desenvolver em cada uma das fases do processo CRISP-DM (Jackson, 2002; CORRÊA & Sferra, 2003).

Business understanding - Visa o entendimento dos objetivos e requisitos do projeto na perspectiva do problema que se propõe ajudar a resolver.

Data understanding: Esta fase inclui uma coleta inicial de dados e atividades que os tornam familiares, assim como possibilita a identificação de problemas na sua qualidade. Ainda nesta fase, podem detetar-se os primeiros discernimentos e subconjuntos interessantes que formam hipóteses de informação escondida.

Data preparation - Inclui as atividades que levam à construção do *dataset* final.

Figura 2.1: Fases do modelo de referência CRISP-DM.



Fonte: Extraído de (Chapman et al., 2000)

Modeling - Nesta fase, selecionam-se os algoritmos a executar, ajustam-se parâmetros e implementa-se o modelo.

Evaluation - Fase de avaliação do resultado de cada passo e certificação do modelo.

Deployment - Depois do modelo construído e avaliado, podem ser propostas ações com base na visão do modelo e seus resultados ou propor a aplicação do modelo a outros conjuntos de dados.

A sequência das fases do processo não é rígida, como se pode observar na figura 2.1. O círculo externo simboliza a natureza cíclica do processo.

Neste trabalho foi adotado o processo de desenvolvimento CRISP-DM. Na tabela 2.1, são referidas as atividades desenvolvidas, em cada uma das fases do processo.

2.2 O conceito de *Dataset*

O conceito de *dataset* é comum a muitas áreas de investigação e de produção onde os dados são a base de atividades de pesquisa e análise. Este fato não significa que áreas diferentes tenham uma mesma definição para o conceito de *dataset*. Em

Tabela 2.1: Atividades desenvolvidas em cada fase do processo CRISP-DM

Fases	Atividades
<i>Business understanding</i>	<ul style="list-style-type: none"> • Objetivo <ul style="list-style-type: none"> - Extração de conhecimento em bases de dados espaciais segundo a técnica de regras de associação espacial • Requisitos <ul style="list-style-type: none"> - Desenvolvimento de plataforma de descoberta de conhecimento em bases de dados espaciais segundo a técnica de regras de associação espacial - Desenvolvimento de um processo de descoberta de conhecimento, iterativo e interativo • Critérios de avaliação do projeto <ul style="list-style-type: none"> - Confrontar os resultados obtidos com o conhecimento detido por peritos do domínio de aplicação
<i>Data understanding</i>	<ul style="list-style-type: none"> • Elaboração de uma simulação com um protótipo desenvolvido, utilizando os dados existentes no artigo que orientou o trabalho • Levantamento de necessidades ao nível dos dados para o projeto • Procurar <i>providers</i> de informação geográfica que forneçam dados por camadas espaciais • Avaliação da qualidade dos dados <ul style="list-style-type: none"> - Qualidade externa - Qualidade interna - Completude, consistência lógica, precisão de posicionamento, precisão temática
<i>Data preparation</i>	<ul style="list-style-type: none"> • Obtenção da informação espacial necessária e adequada ao conjunto de teste • Transformações dos dados de forma a serem integrados no modelo relacional estendido • Seleção de atributos • Análise de valores em falta • Conjunto de dados para teste da plataforma
<i>Modeling</i>	<ul style="list-style-type: none"> • Seleção dos algoritmos a executar e ajuste de parâmetros • Desenvolver projeto de teste
<i>Evaluation</i>	<ul style="list-style-type: none"> • Confrontar resultados obtidos com os emitidos por peritos
<i>Deployment</i>	<ul style="list-style-type: none"> • Produzir relatório final

engenharia informática é comum aceitar a seguinte definição para *dataset*: repositório para integração de dados heterogêneos em formatos variados e de diversas fontes. Numa atividade de *data mining*, um *dataset* pode assumir a forma de uma matriz. No entanto, segundo (Renear, Sacchi, & Wickett, 2010), nas diversas definições do conceito encontradas em documentação técnica e em literatura científica podem encontrar-se as seguintes quatro características comuns:

- **Agrupamento** - Conjunto de dados que são tratados como uma unidade.
- **Conteúdo** - Os conjuntos de dados referem-se a um contexto.
- **Relacionamento** - Os dados estão relacionados uns com os outros com um relacionamento que vai para além do agrupamento.
- **Objetivo** - Além de cada um dos dados servir para registar informações, o *dataset* tem uma aplicação maior, como fornecer um caminho para investigação científica.

O *dataset* na forma de uma matriz constitui o modelo de dados de um processo de *data mining*. As dimensões da matriz, ($N \times P$), indicam um número N de instâncias e um número P de atributos (Liu & Motoda, 2012). Os valores da matriz podem ser binários, contínuos ou discretos.

2.3 *Market basket analysis*

A área do negócio vive com a preocupação de melhorar a relação com o cliente tentando responder às suas expectativas. Este anseio por parte do gestor de negócio pode passar por conhecer os hábitos de compra do cliente. Com o massivo aumento de dados dos clientes, coletados e armazenados, abre-se uma janela de análise dos hábitos de compra dos clientes. As análises de vendas realizadas são habitualmente utilizadas para melhorar decisões do negócio, tais como, o que colocar à venda ou como colocar os produtos nas prateleiras para maximizar o lucro (Agrawal, Imieliński, & Swami, 1993).

A *Market Basket Analysis* (MBA) é uma técnica que permite analisar a relação comercial entre uma entidade e um cliente (Haughton et al., 2003). Esta análise, na sua forma mais simples, permite extrair padrões frequentes. Padrões frequentes são grupos de produtos comprados pelos clientes, que são encontrados na análise de vendas, e que ocorrem com a frequência imposta na análise. Dos padrões encontrados são extraídos grupos de produtos correlacionados, permitindo determinar que produtos são comprados juntos. Na tabela 2.2 estão representadas as vendas de um determinado negócio. As tabelas 2.3 e 2.4 utilizam outras formas de representação da informação contida na tabela 2.2.

Tabela 2.2: Registo de transações.

TID	Item
T1	B
T2	A
T2	C
T2	D
T3	B
T3	C

Tabela 2.3: Matriz esparsa com a mesma informação da tabela 2.2.

TID \ Item	A	B	C	D
T1	0	1	0	0
T2	1	0	1	1
T3	0	1	1	0

Tabela 2.4: Representação compacta da informação da tabela 2.2.

TID	Item
1	B
2	A, C, D
3	B, C

Considerem-se as seguintes definições:

- Chama-se *item* a um produto indicado numa transação.
- Um *itemset* é uma lista de *itens*.
- Um *itemset* com k elementos é chamado de *k-itemset*.
- Diz-se que uma transação T contém um *itemset* I se $I \subseteq T$. Por exemplo, a terceira transação da tabela acima suporta os *itemsets* $\{B\}$, $\{C\}$ e $\{B, C\}$.

A MBA consiste em encontrar grupos de itens correlacionados, regras de associação, que ocorrem juntos numa transação. Para tal, os dados transacionais constituirão um ficheiro que será o *input* de um algoritmo de extração de padrões frequentes. O formato de organização dos dados no ficheiro poderá ser o de uma matriz, tabela 2.3, ou em representação compacta, tabela 2.4, de acordo com os requisitos do algoritmo. As regras de associação serão obtidas por aplicação de um outro algoritmo sobre os padrões frequentes obtidos. As funções de cada um dos algoritmos, segundo (Agrawal et al., 1993), são sumariadas em seguida.

1. Geração de todas as combinações de *itens* com valor de suporte, (cf., secção 2.3.2), superior a um determinado *threshold*. As combinações obtidas passam a designar-se por *itemsets* frequentes.

2. Por cada *itemset* frequente encontrado no passo anterior, gerar todas as regras de associação (cf., secção 2.3.1). Considerando um determinado *k-itemset* frequente, $Y = \{I_1, I_2, \dots, I_k\}$, as regras de associação a extrair terão que ter um antecedente e um conseqüente que sejam *subsets* de Y . Se o antecedente X tiver $k-1$ *itens* então o conseqüente terá que ter $Y - X$ *itens*. Se a regra gerada tiver um factor de confiança, (cf., secção 2.3.2), superior a um determinado *threshold* a regra é considerada forte, caso contrário é descartada.

2.3.1 Regras de associação

Considerando o exposto na secção 2.3, descreve-se a definição de regra de associação.

Seja:

- $I = \{i_1, i_2, \dots, i_M\}$, conjunto dos M itens vendidos.
- $T = \{t_1, t_2, \dots, t_N\}$ conjunto das N transações onde $t_j \subseteq I$, e.g., t_j é um conjunto de *itens*.

Uma regra de associação consiste numa implicação da forma $X \Rightarrow Y$, onde $X \subseteq I$, $Y \subseteq I$ e $X \cap Y = \emptyset$. O termo X denomina-se *antecedente* e o termo Y *conseqüente*.

2.3.2 Suporte, confiança e lift

Sobre a regra de associação, $X \Rightarrow Y$, definem-se as seguintes medidas quantitativas (Han, Pei, & Kamber, 2011):

- **suporte** ($X \Rightarrow Y$) = $P(X \cup Y)$
- **confiança** ($X \Rightarrow Y$) = $P(Y | X)$
- **lift** ($X \Rightarrow Y$) = $P(Y | X) / P(Y)$

O *suporte* indica a probabilidade dos *itemsets* X e Y ocorrerem juntos numa transação. A *confiança* é uma probabilidade condicionada, i.e., mede a probabilidade de ocorrer o *itemset* Y quando ocorre o *itemset* X . O *lift* relaciona a probabilidade do *conseqüente* ocorrer sozinho com a probabilidade deste ocorrer no contexto da regra. Se o *conseqüente* tiver menor probabilidade de ocorrer sozinho do que no

contexto da regra, a regra é interessante, caso contrário a regra não traz valor acrescentado.

A regra $X \Rightarrow Y$ é considerada forte, no conjunto de transações T , se cumulativamente verificar as seguintes condições (Agrawal et al., 1993):

- O suporte s , $0 \leq s \leq 1$, do *itemset* $\{ X, Y \}$, é superior a um determinado *threshold* chamado *minsupport*, i.e., a percentagem de transações que satisfazem o *itemset* $\{ X, Y \}$;
- A confiança c , $0 \leq c \leq 1$, é superior a um determinado *threshold* chamado *minconfidence*, i.e., se pelo menos $c\%$ das transações que satisfazem X também satisfazem Y .

2.4 Sistema de suporte à informação espacial

A informação geográfica existente na superfície terrestre é constituída por camadas. A cada camada corresponde uma entidade, e.g., *roads, buildings, places, transport areas, aeroways, ...* Os dados geográficos, constituem recursos espaciais e podem ser armazenados em bases de dados espaciais, seguindo a abordagem do modelo relacional estendido, (Shekhar & Chawla, 2003), como se descreve na secção 2.6. A informação espacial está sempre relacionada com dados não espaciais. Um Sistema de Bases de Dados Espaciais (SBDE) verifica os requisitos descritos a seguir (Güting, 1994).

- Sistema de bases de dados com recursos adicionais.
- Fornece um conjunto de tipos de dados espaciais, e.g.. *POINT, LINE, REGION*, fundamental para modelar a estrutura de entidades geométricas no espaço e uma linguagem de consulta.
- Fornece indexação espacial e algoritmos eficientes para junções espaciais.
- Capaz de recuperar uma coleção de objetos que se encontram numa área particular sem necessidade de procurar em todo o conjunto de objetos.

2.5 Tipos de objetos espaciais

Para a representação espacial de entidades geográficas, são necessários três tipos de abstrações: ponto, linha e polígono (Güting, 1994). Um objeto ponto representa uma entidade, onde somente é relevante a sua localização geográfica, *e.g.*, localização de uma escola numa cidade. Uma linha é utilizada para representar entidades geográficas que permitem movimentação no espaço ou ligação entre dois pontos deste, *e.g.*, um rio, um cabo telefónico. Uma linha deve ser entendida com a possibilidade de significado de uma curva. Neste caso, a linha é constituída por uma sequência de segmentos, que também se designa por polilinha. Um polígono é uma abstração que representa uma entidade que tem uma extensão no espaço bidimensional, *e.g.*, um país. Um polígono pode ter buracos ou ser constituído por vários polígonos, constituindo um multipolígono. A figura 2.2 ilustra as abstrações apresentadas.

Figura 2.2: Tipos básicos de abstrações: ponto, linha e polígono.



Fonte: Extraído de (Güting, 1994).

2.6 Modelo de dados espacial

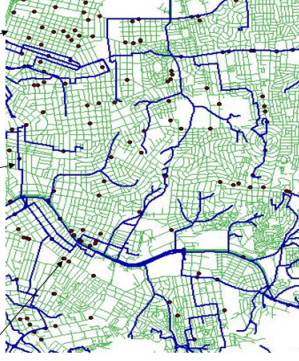
Numa base de dados espacial, os dados espaciais são guardados em tabelas, de acordo com o seu tipo. A cada entidade geográfica, *e.g.*, *roads*, *buildings*, corresponde uma tabela no modelo de dados espacial (Shekhar & Chawla, 2003). Assim, cada tabela representa um plano de informação, denominado por camada. A figura 2.3 ilustra um exemplo de modelo de dados espacial. Na figura 2.3, observa-se que cada tabela contém um atributo espacial e um conjunto de atributos alfanuméricos. O atributo espacial contém a definição de um objeto espacial, *e.g.*, *Point*, *Line*, *Polygon*, incluindo a sua referência geográfica. Os atributos alfanuméricos contêm informação que permitem contextualizar o objeto espacial.

Figura 2.3: Exemplo de modelo de dados espacial.

Road					
id	name	type	shape		
1	A	pedestrian	Multiline [(x1, y1),(x2,y2),...]		
2	B	primary	Multiline [(x1, y1),(x2,y2),...]		

WaterResource					
id	name	type	shape		
1	Jacui	water	Multiline [(x1,y1),(x2,y2),...]		
2	Guaiba	basin	Multiline [(x1,y1),(x2,y2),...]		
3	Uruguai	reservoir	Multiline [(x1,y1),(x2,y2),...]		

GasStation					
id	name	type	volDiesel	volGas	shape
1	Elf	small	20000	85000	Point{(x1,y1)}
2	Esso	complete	30000	95000	Point{(x1,y1)}



Fonte: Adaptado de (Bogorny et al., 2008).

2.7 Operadores topológicos, de distância e de direção

Numa base de dados geográfica, os atributos espaciais têm implicitamente codificadas relações espaciais (Bogorny et al., 2008). Assim, um determinado objeto espacial pode estar submetido à influência dos objetos espaciais vizinhos. Esta é a principal diferença na extração de conhecimento numa base de dados relacional e numa base de dados espacial (Ester, Frommelt, Kriegel, & Sander, 2000).

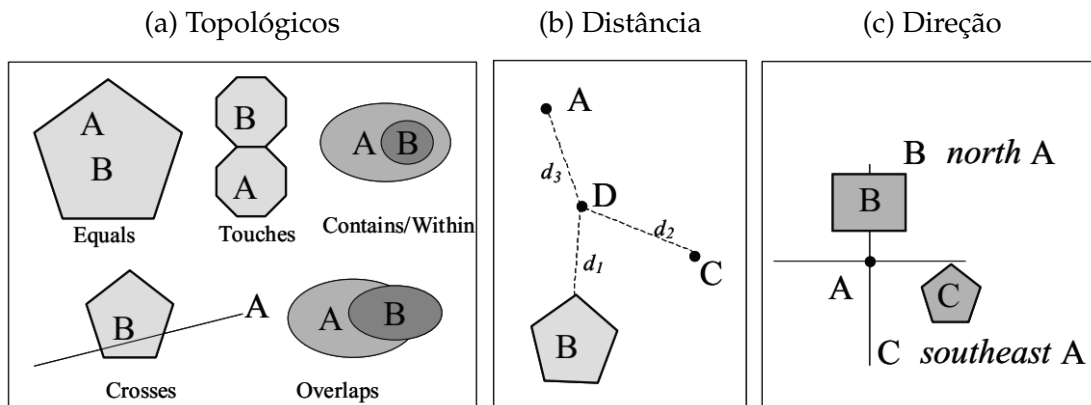
As relações espaciais podem ser classificadas em três tipos básicos: topológicas, de distância e de direção (Bogorny et al., 2008; Ester et al., 2000). São relações binárias porque determinam relações entre pares de objetos.

O *Open Geospatial Consortium* (OGC) é uma organização que desenvolve *standards* para o desenvolvimento de soluções interoperáveis entre sistemas de informação geográfica. O OGC, definiu em (Inc, 2011), o seguinte conjunto de operadores topológicos: *Equals, Disjoint, Intersects, Touches, Crosses, Within, Contains, Overlaps, Relate, LocateAlong* e *LocateBetween*. As relações topológicas têm a característica de se manter invariantes quando a ambos os objetos são aplicadas transformações de rotação, translação ou escala.

A distância espacial mede a posição relativa entre objetos espaciais (Duan, Zang, Pu, & Shi, 2010). A distância espacial pode ser medida no espaço Euclideano e significa a distância mais curta entre dois dados pontos.

Os operadores de direção lidam com a ordem no espaço, i.e., referenciam um objeto em relação a outro, utilizando os pontos cardeais (norte, sul, este, oeste), como se mostra na figura 2.4.

Figura 2.4: Algumas relações topológicas, de distância e de direção.



Fonte: Extraído de (Bogorny et al., 2008).

2.8 Seleção de operadores topológicos

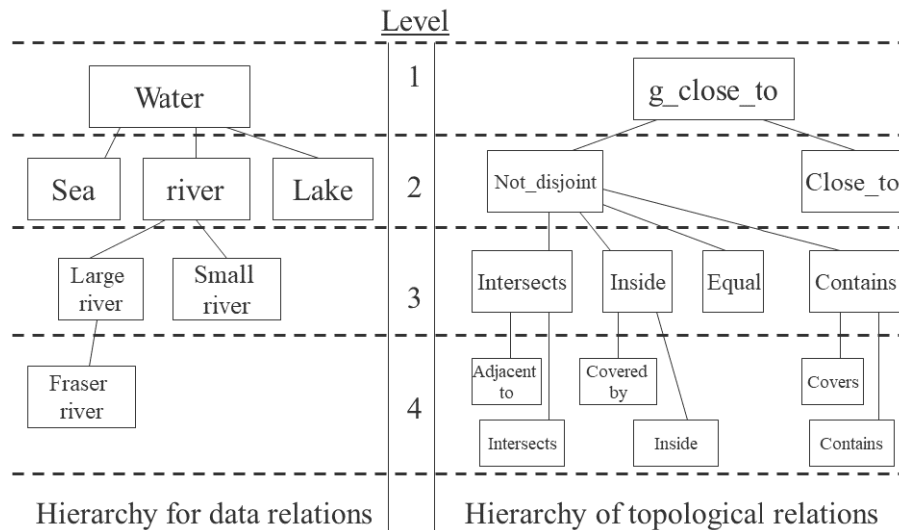
As regras de associação espacial obtidas num processo de descoberta de conhecimento representam correlações entre relações topológicas. Os operadores topológicos selecionados para obtenção dessas correlações influenciam os padrões encontrados.

À partida todos os operadores topológicos, entre recursos espaciais, devem ser testados para detetar quais se verificam (Bogorny et al., 2008). No entanto, a definição do conjunto de operadores, a incluir no processo, admite análises com vista à definição de uma estratégia de seleção de operadores. Uma possibilidade pode passar por identificar os operadores que representam restrições de integridade espacial obrigatórias existentes nos dados, e.g., *island within water*. Deste modo torna-se possível excluir estes operadores, por produzirem resultados conhecidos (Bogorny et al., 2008). Pode ser realizada uma análise semântica de identificação de relações proibidas entre recursos espaciais. Como exemplo, um *bus_stop* não está dentro de uma *road*, logo a relação, *bus_stop within road*, não faz sentido ser testada.

O problema da determinação dos operadores topológicos a considerar num processo de descoberta de conhecimento foi abordado por (Koperski & Han, 1995). Este autor propõe a extração de regras de associação espaciais, em diferentes níveis de granularidade, recorrendo a hierarquias de conceitos e de relações topológicas. A hierarquia de relações topológicas acompanha a hierarquia do conceito, de forma a obter resultados de relações espaciais mais específicas, à medida que o algoritmo utilizado, for diminuindo a granularidade. Na figura 2.5 mostra-se um

exemplo de hierarquia de operadores topológicos que acompanha a hierarquia do conceito *water*. Na figura 2.5, o operador *g_close_to* define um operador que integra os operadores *intersect*, *adjacent_to*, *contains* e *close_to*.

Figura 2.5: Hierarquias de conceito e de operadores.



2.9 Níveis de granularidade

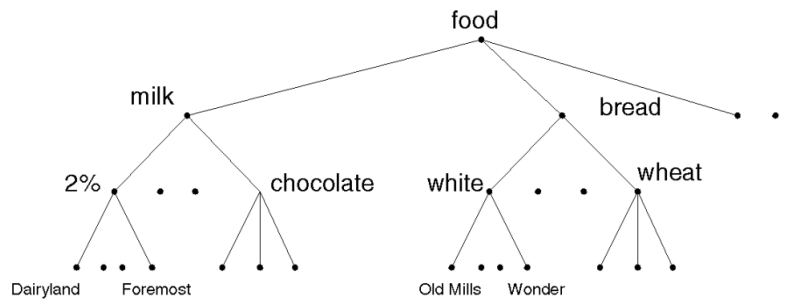
Um sistema de extração de regras de associação pode permitir a extração de associações em múltiplos níveis. Como exemplo, considerem-se as seguintes regras: “80% dos clientes que compram leite também compram pão” e “2% dos clientes que compram leite, compram pão de trigo”. A segunda regra está expressa, relativamente ao conceito “pão”, num menor nível de abstração e transporta informação mais concreta (adaptado de (Han & Fu, 1999)). Para que um sistema possa extrair regras de associação multinível, deverá ser estabelecida uma hierarquia de conceitos. Um exemplo de taxonomia é apresentado na figura 2.6.

Um sistema que explore a extração de regras de associação multinível, de acordo com (Han & Fu, 1999), deverá apresentar as seguintes características:

1. Dados em múltiplos níveis de abstração.
2. Métodos eficientes para extração de regras de associação multi-nível.

Cada tabela do modelo de dados espacial abstrai um conceito do mundo real (cf., secção 2.6). A coleção de atributos em cada tabela tem normalmente um atributo

Figura 2.6: Um exemplo de taxonomia.



Fonte: Extraído de (Han & Fu, 1999).

denominado *type* que permite categorizar os tipos de objetos espaciais existentes na relação (Koperski & Han, 1995). Como exemplo, a tabela *Road* contém instâncias espaciais do tipo *pedestrian* e do tipo *primary*. Esta forma de modelação da informação permite hierarquizar o conceito em níveis de abstracção e estabelecer uma taxonomia.

2.10 Market basket analysis no 'mundo geo'

O processo MBA com informação espacial foi apresentado em (Bogorny et al., 2008). Na extração de conhecimento em bases de dados transacionais, a organização dos dados que constituem o *input* do processo mostra que cada linha representa uma transação e as colunas representam produtos (cf., secção 2.3). Em processo idêntico, mas utilizando bases de dados espaciais, cada linha representa uma instância de um recurso geográfico e as colunas representam predicados. Os predicados podem ser espaciais ou não espaciais. Os predicados espaciais são obtidos a partir dos operadores topológicos, de distância ou de direcção (cf., secção 2.7). Um predicado espacial, construído com um operador topológico, indica a existência de uma relação topológica de um tipo de recurso espacial, *e.g.*, *river*, com a instância espacial representada em linha. Um predicado não espacial envolve uma condição que inclui um atributo não espacial, *e.g.*, *population*, relativo à instância representada em linha. A tabela 2.5 ilustra um exemplo de *dataset* para extração de conhecimento construído a partir de uma base de dados espacial.

O *dataset* construído, na forma de matriz, a partir de informação espacial será o *input* do algoritmo de extração de padrões frequentes. A partir dos padrões frequentes obtidos, outro algoritmo procederá à extração de regras de associação espacial. Uma regra de associação espacial é uma regra que contém pelo menos

Tabela 2.5: Exemplo de *dataset*.

Áreas geográficas	Predicados espaciais				
	Cidade 1	contains_school	contains_hospital	contains_roads	crosses_waterways
Cidade 2			contains_roads	crosses_waterways	contains_amenities
Cidade 3	contains_school	contains_hospital	contains_roads	crosses_waterways	
Cidade 4	contains_school	contains_hospital	contains_roads		contains_amenities
Cidade 5		contains_hospital	contains_roads		
Cidade 6	contains_school	contains_hospital	contains_roads	crosses_waterways	

um predicado espacial.

2.11 Extração de conhecimento

O processo de extração de conhecimento em bases de dados geográficas pode consistir na procura de regras de associação espacial interessantes. Estas regras são, na sua grande maioria, compostas por relações topológicas, *e.g.*, *touches*, *contains*, *...*, entre recursos espaciais, *e.g.*, *school*, *hotel*, *road*, *...*, que existem num determinado espaço geográfico, *e.g.*, *city*, também recurso espacial.

A extração de conhecimento de dados, segundo a técnica de regras de associação, tem a característica de gerar uma enorme quantidade de regras, muitas delas redundantes ou irrelevantes, o que dificulta a interpretação de resultados. Quando se trata de bases de dados geográficas, este problema aumenta, devido ao fato dos dados geográficos incluírem associações naturais intrínsecas aos próprios dados.

Este problema foi tratado em (Bogorny et al., 2008), onde é proposta uma solução para a redução das regras de associação espaciais geradas, a qual constitui pressuposto para este trabalho. A solução apresentada contempla a construção de uma geo-ontologia, composta pelo conhecimento formado, pelas restrições de integridade espacial existentes nos dados, assim como conhecimento semântico das disposições naturais das entidades geográficas, *e.g.*, uma estrada não contém um rio.

As restrições de integridade espacial permitem manter a consistência dos dados, numa base de dados geográfica. Estas restrições podem ser obrigatórias, proibidas ou possíveis. Uma restrição de integridade espacial obrigatória representa uma cardinalidade de um-para-um ou um-para-muitos, no modelo conceptual que representa os dados. As restrições de integridade espacial obrigatórias geram padrões conhecidos e regras com confiança alta, mas não acrescentam conhecimento, porque representam a forte correlação natural que existe entre dados de natureza geográfica. Somente as relações possíveis podem produzir padrões

interessantes. São, por via disso, as mais relevantes para mostrar novo e útil conhecimento. Num outro nível, as relações topológicas são igualmente uma forma de garantir a consistência e a qualidade dos dados, e podem ser classificadas em obrigatórias, possíveis ou proibidas (Cockcroft, 1997). Algumas relações topológicas permitem validar restrições de integridade. Uma relação espacial obrigatória representa uma restrição de integridade espacial.

A geo-ontologia proposta inclui informação sobre as restrições de integridade espacial, obrigatórias e proibidas, e restrições semânticas (Bogorny et al., 2008). As restrições de integridade obrigatórias são extraídas da base de dados, enquanto que as relações proibidas são identificadas pelo utilizador. Estas restrições referem-se a dois tipos de relações entre objetos espaciais. Um, espelha relações entre tipos de objetos espaciais representados nas linhas do *dataset*, e tipos de objetos espaciais representados nas colunas. O outro, refere-se a relações entre tipos de objetos espaciais existentes nas colunas. O objetivo da geo-ontologia é integrar conhecimento no processo de extração de regras, de forma a que, não sejam extraídas regras de associação conhecidas. Acresce que, pelo fato desta informação ser conhecida *'a priori'*, permite diminuir o número de junções espaciais a realizar, com vista à constituição do *dataset, input* do processo de extração. Este repositório de informação, a geo-ontologia, será considerado em duas fases do processamento. A primeira fase ocorre antes da constituição do *dataset*, e a segunda, depois da formação dos *itemsets* frequentes de dimensão 2. O objetivo da primeira fase é reduzir o número de junções espaciais e retirar da formação de *itemsets* predicados redundantes. Na segunda fase, o objetivo é impedir a formação de regras de associação, que contenham dependências conhecidas.

Embora o conhecimento *'a priori'* seja uma mais valia, num processo de descoberta de conhecimento, a utilização de uma geo-ontologia, com estas características, só é possível se o modelo de dados estiver dotado de restrições de integridade referencial.

Todas as relações topológicas devem à partida ser consideradas, no entanto, nem todas fazem sentido ser testadas (Bogorny et al., 2008). A análise semântica pode diminuir o número de relações topológicas a considerar no processo de descoberta de conhecimento. Na tabela 2.6 encontram-se assinalados os operadores topológicos que, depois de uma análise semântica, devem ser considerados entre os recursos indicados em linha.

No entanto, embora as relações topológicas assinaladas sejam semanticamente possíveis, nem todas trazem novo conhecimento. Assim se forem consideradas

Tabela 2.6: Relações topológicas possíveis e obrigatórias.

Relações topológicas	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Combinções semânticas							
Local_government and County					x		
River and Road	x		x			x	

Fonte: Extraído de (Bogorny et al., 2008).

as restrições de integridade espacial, o número de relações a considerar para descoberta de conhecimento decresce, como se mostra na tabela 2.7.

Tabela 2.7: Relações topológicas possíveis.

Relações topológicas	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Combinções semânticas							
Local_government and County							
River and Road	x		x			x	

Fonte: Extraído de (Bogorny et al., 2008).

2.12 Aprendizagem supervisionada e aprendizagem não supervisionada

A aprendizagem automática (*Machine learning*) consiste no desenho de algoritmos eficientes de predição precisa, (Mohri, Rostamizadeh, & Talwalkar, 2012). Em aprendizagem automática os dados desempenham um papel fundamental e os algoritmos são utilizados para descobrir e aprender conhecimento ou propriedades sobre os dados. Aprendizagem supervisionada e aprendizagem não supervisionada são dois tipos de aprendizagem automática, cujos conceitos se descrevem sumariamente, em seguida.

- **Aprendizagem supervisionada** - A aplicação desta técnica desenvolve-se em duas fases: fase de treino e fase de teste. Uma linha de dados do *dataset* designa-se por indivíduo. As ações realizadas em cada uma das fases apresentam-se nos pontos seguintes:
 1. **Fase de treino** - Nesta fase, cada indivíduo fornecido ao sistema é acompanhado da sua classe correta de classificação. O emparelhamento indivíduo - classe permite ao sistema inferir a função de classificação e construir o modelo.

2. **Fase de teste** - Na fase de teste é avaliado o desempenho do modelo construído.

Na aprendizagem supervisionada o sistema constrói um modelo durante a fase de treino, com base num conjunto de indivíduos que estão acompanhados da sua classificação. Este fato permite que na fase de teste seja possível avaliar o desempenho do modelo construído.

- **Aprendizagem não supervisionada** - Neste tipo de aprendizagem automática, não existe conjunto de treino, pelo que, o sistema precisa de reconhecer os padrões. O resultado final é um conjunto de descrições de classes. Uma descrição para cada classe descoberta no ambiente, i.e., na base de dados. O objetivo da aprendizagem não supervisionada é encontrar padrões nos dados.

3

Modelo proposto

Neste capítulo apresenta-se o modelo conceptual para desenvolvimento de uma plataforma de descoberta de conhecimento, na forma de regras de associação espacial, extraídas em zonas geográficas. Uma zona geográfica define-se através de uma geometria, do tipo ponto ou do tipo polígono (cf., secção 2.5), que caracteriza um fenómeno, adicionada por um *buffer* espacial variável.

Nas zonas geográficas seleccionadas ocorreram fenómenos, *e.g.*, acidentes rodoviários, fogos, doenças em animais. O objetivo da descoberta de conhecimento é encontrar coincidências entre a ocorrência de fenómenos e a existência de padrões espaciais. A variação do *buffer* espacial imposto, à geometria que caracteriza o fenómeno, permite comparar os padrões espaciais obtidos nas diferentes áreas resultantes.

O modelo conceptual deste trabalho assenta na MBA estendida para o '*mundo-geo*'. Esta generalização foi aplicada por (Bogorny et al., 2008), a qual constituiu pressuposto para este trabalho.

3.1 Requisitos

O sistema a desenvolver deve responder aos requisitos a seguir enunciados.

- Gerar, a partir de um ficheiro, uma base de dados com informação geográfica, com amplitude suficiente para permitir a análise requerida pelo utilizador.
- Gerar, numa base de dados espacial, informação que defina a georreferenciação das ocorrências do fenómeno.
- Gerar um *dataset* que será o *input* do processo de descoberta de conhecimento (cf., secção 2.10).
- Gerar uma geo-ontologia a partir da análise do *dataset*. A geo-ontologia requerida deve conter informação sobre dois tipos de dependências geográficas, que se descrevem em seguida.
 1. Dependências entre recursos espaciais representados em linha e recursos espaciais representados em colunas.
 2. Dependências entre os recursos espaciais existentes nas colunas.
- Aplicar um algoritmo de pesquisa que permita percorrer todo o espaço de *suporte e confiança*. O objetivo deste algoritmo é detetar pontos, no espaço analisado, onde o número de regras de associação extraídas não seja superior a um valor definido pelo utilizador, retornando as respetivas regras.
- Aplicar diferentes valores de *buffer* espacial às geometrias que correspondem a ocorrências do fenómeno.

3.2 Dataset - Modelo teórico

Considerem-se as seguintes premissas:

- $T = \{t_1, t_2, t_3, \dots, t_M\}$, um conjunto de recursos geográficos onde, sobre cada instância $t \in T$, se pretende determinar a relação espacial existente com tipos de recursos geográficos, que tenham alguma relação espacial com T . Uma instância t corresponde à geometria espacial que caracteriza o fenómeno acrescida por um *buffer* espacial.
- $S = \{O_1, O_2, O_3, \dots, O_N\}$, um conjunto de tipos de recursos geográficos com uma relação espacial com T .
- $O = \{o_1, o_2, o_3, \dots, o_P\}$, um conjunto de instâncias de um tipo de recurso espacial existente em T .

- $O \subset S$.
- Um predicado espacial denota uma relação topológica entre uma instância t e um conjunto de instâncias O_N .

O cálculo de um predicado espacial obtém-se realizando a junção espacial de uma instância t_M , de T , *e.g.*, Lisboa, com todas as instâncias de O_N , *e.g.*, roads, que têm alguma relação topológica, *e.g.*, *contains*, com a instância t_M . De um modo geral, a extração dos predicados espaciais implica o cálculo das junções espaciais de todas as instâncias de T , com todas as instâncias de O_N , $\forall O_N \in S$, que tenham alguma relação espacial com T . Todas as instâncias de T devem ser comparadas com todas as instâncias representadas por S .

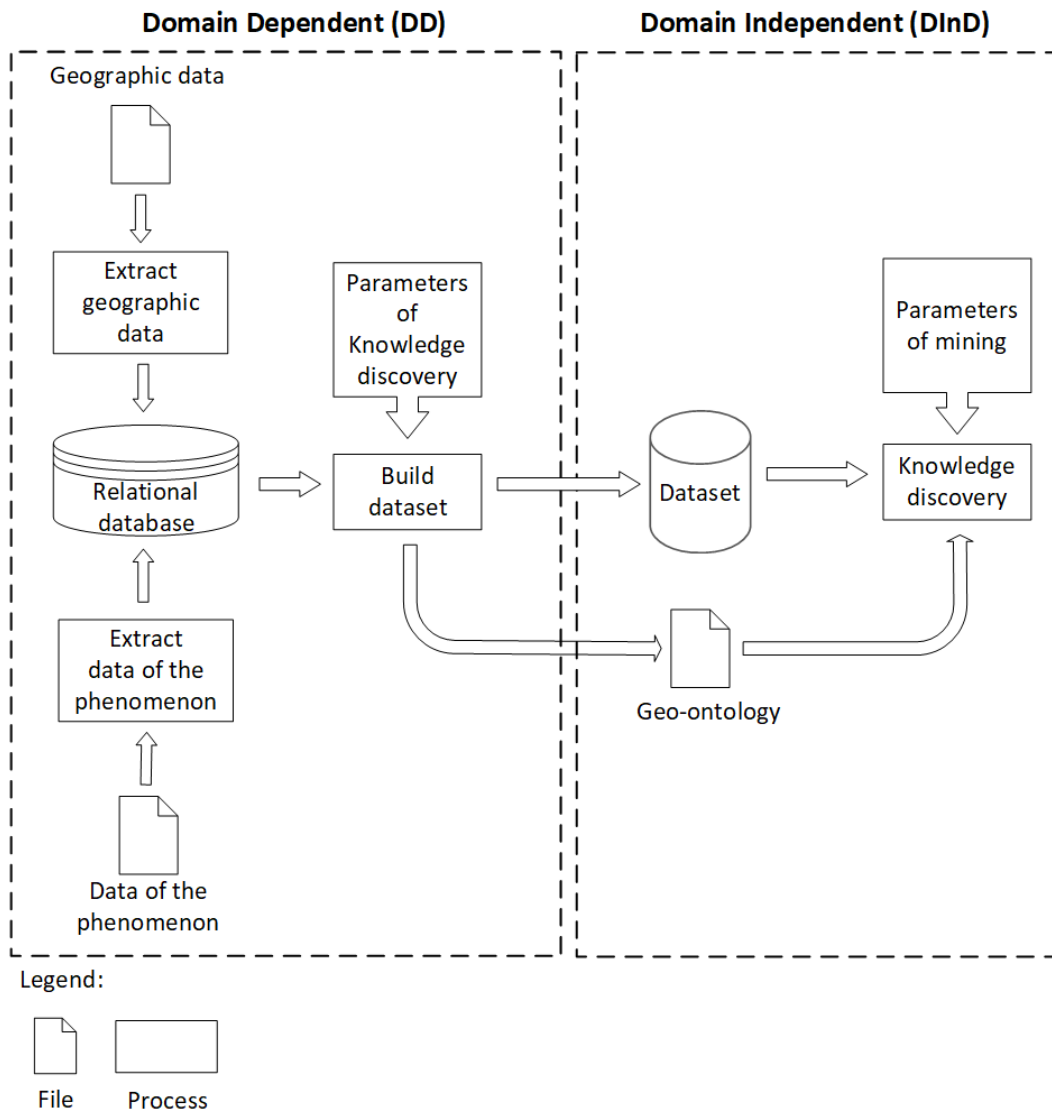
3.3 Arquitetura do modelo conceptual

O modelo a desenvolver deverá permitir a sua utilização em vários domínios. Na figura 3.1 apresenta-se a arquitetura do modelo conceptual. A arquitetura apresentada está dividida em dois blocos designados por *Domain Dependent* e *Domain Independent*. No bloco *Domain Dependent* são realizadas operações dependentes do domínio de aplicação. O conjunto de operações realizadas no bloco *Domain Independent* é independente do domínio de aplicação. A razão para esta divisão prende-se com o fato de existir um conjunto de operações que pode variar, dependendo do domínio de aplicação, e outro que se mantém inalterável. O conjunto de operações englobadas no *Domain Dependent* pode apresentar diferenças, devido aos seguintes fatos:

- As fontes de dados de informação geográfica não utilizam todas o mesmo formato, o que implica um conjunto diferente de operações, e eventualmente tecnologias, para serem armazenados num SBDE.
- Os dados geográficos que caracterizam o fenómeno a estudar podem apresentar-se em formatos e organizações diversas.

Os dois blocos que compõem a arquitetura apresentada implementam duas fases do processamento. Na primeira fase, o objetivo é construir um *dataset* e analisar as dependências geográficas existentes, com vista à construção da geo-ontologia. Depois de obtidos, os ficheiros com informação geográfica, acerca das zonas geográficas a analisar e de caracterização do fenómeno a estudar, estes devem ser

Figura 3.1: Arquitetura do modelo conceptual.



incluídos numa base de dados geográfica. A informação geográfica contida em 'Geographic data' deve ser armazenada segundo o modelo de dados espacial (cf., secção 2.6). O conjunto de parâmetros, indicado em 'Parameters of knowledge discovery', na figura 3.1, tem o objetivo de definir as linhas e as colunas do *dataset*, na forma de matriz (cf., secção 2.10). No final desta fase, o processo *Build dataset*, figura 3.1, desencadeia o processo de construção do *dataset*, assim como a geo-ontologia, com as características definidas (cf., secção 3.1). O objetivo da segunda fase é obter conhecimento, na forma de regras de associação espacial. No início desta fase, é solicitada a intervenção do utilizador, caso entenda, com a finalidade de analisar as dependências geográficas detetadas, presentes na geo-ontologia. Depois da tomada de decisão, por parte do utilizador, acerca das dependências

que devem ser consideradas no processo de extração de regras de associação, o processo *Knowledge discovery*, figura 3.1, reconstrói o *dataset*. Ainda este processo, *Knowledge discovery*, depois do *dataset* reconstruído, executa o processo de extração de conhecimento. Para a execução deste processo, são tomados em consideração, os parâmetros *Parameters of mining*, que configuram uma função filtro do processo de extração de conhecimento. A secção 3.4 contém detalhes sobre a função de cada elemento constituinte da arquitetura do modelo conceptual.

3.4 Elementos do modelo conceptual

A função de cada elemento presente, na arquitetura conceptual, apresentada na figura 3.1, é apresentada nesta secção.

- *Geographic data* - Este elemento tem a função de repositório de dados geográficos. Os elementos geográficos devem ter atributos que permitam a segmentação da informação geográfica em camadas. Por outro lado, esses atributos devem corresponder aos objetivos da descoberta de conhecimento. A área abrangida por essa informação deve permitir satisfazer a análise do utilizador. O projeto OSM reúne características que satisfazem estes requisitos.
- *Extract geographic data* - Processo que segmenta a informação geográfica em camadas, guardando-a numa base de dados espacial.
- *Data of the phenomenon* - Ficheiro que contém informação espacial, e eventualmente, não espacial sobre o fenómeno a analisar.
- *Extract data of the phenomenon* - Processo que extrai informação espacial e não espacial do ficheiro '*Data of the phenomenon*', guardando-a na base de dados espacial.
- *Relational database* - Base de dados espacial que armazena os dados necessários à construção do *dataset*.
- *Parameters of knowledge discovery* - Parâmetros que definem as linhas e as colunas da matriz (cf., secção 2.10).

– *Linhas da matriz*

* Seleção das ocorrências do fenómeno.

- * A medida do *buffer* espacial a considerar na figura geométrica que caracteriza o fenómeno.
- *Colunas da matriz*
 - * Indicação dos recursos geográficos que se pretendem incluir na análise.
 - * Seleção dos operadores a aplicar.
- **Build dataset** - Processo responsável pela construção do *dataset*. O modelo teórico deste processo está apresentado na secção 3.2.
- **Dataset** - *Dataset* construído com os princípios enunciados (cf., secção 2.10).
- **Geo-ontology** - Geo-ontologia cujos requisitos se descrevem na secção 3.1. A função da geo-ontologia é permitir a identificação de padrões de regras, com o objetivo de excluir, as regras que correspondem a estes padrões, do resultado a apresentar ao utilizador. Os padrões identificados contêm dependências conhecidas '*a priori*'. A seleção das dependências geográficas, que identificam padrões de regras é realizada pelo utilizador.
- **Parameters of mining** - Este conjunto de parâmetros configura uma função filtro no processo de extração de conhecimento. Esta função tem os seguintes objetivos:
 - Determinar se um *itemset* corresponde à parametrização definida pelo utilizador.
 - Avaliar se uma regra de associação extraída, está de acordo com a parametrização definida pelo utilizador.
 - Determinar o cancelamento do processo de extração de regras de associação, num dado ponto caracterizado por (*suporte, confiança*), pelo fato de se ter ultrapassado o número máximo de regras definido pelo utilizador.

Para atingir os objetivos descritos, a função de filtragem recebe os parâmetros descritos em seguida.

- Comprimento máximo e mínimo do *itemset*.
- Comprimento máximo e mínimo do *antecedente*.
- Comprimento máximo e mínimo do *consequente*.
- Valor mínimo de *lift*.

- Valor de salto de *suporte*.
 - Valor de salto de *confiança*.
 - Número máximo de regras a extrair em cada ponto analisado.
- **Knowledge discovery** - Este processo implementa o algoritmo de pesquisa que se descreve na secção 3.6.

3.5 Geo-ontologia

O objetivo da geo-ontologia é diminuir o número de regras geradas com base no conhecimento '*a priori*', obtido do *dataset* construído pelo processo *Build dataset* (cf., secção 3.3). Esta secção mostra as consequências das dependências geográficas na geração de regras de associação espacial.

3.5.1 Análise de dependências

A análise da informação do *dataset* pode identificar tipos de dependências que originarão padrões conhecidos e conseqüentemente regras não interessantes. A **dependência entre instâncias representadas em linhas e predicados espaciais**, representados em colunas, pode ser avaliada pela análise do *dataset*. Tomando como exemplo o *dataset* representado na tabela 2.5, observa-se que o predicado *contains_roads* tem um suporte de 100%. Um predicado com 100% de suporte aparece em pelo menos, metade do total de padrões, e gera um enorme número de regras não interessantes (Bogorny et al., 2008). Como exemplo, a regra *contains_school* \Rightarrow *contains_roads*, expressa que cidades que contêm *school* também contêm *roads*. Parece óbvio que esta regra não extrai conhecimento útil, pois será razoável admitir que uma cidade que tenha uma *school*, terá pelo menos uma *road*.

A obtenção de regras de associação é antecedida pela extração de *itemsets* frequentes. Estes *itemsets* são obtidos através de um processo de combinação de, neste caso, predicados espaciais. Este processo pode ser realizado pelo algoritmo *Apriori* (Agrawal et al., 1993). Este algoritmo executa várias passagens pelo *dataset*, sendo que na primeira dessas passagens calcula o suporte de cada *item* e obtém o *k-itemset*. São eleitos para o próximo passo, os *itemsets* que tenham *suporte* superior a um *suporte* mínimo pretendido. Nos passos seguintes, considerando *k* o número do passo corrente, ($k \geq 2$), os *itemsets* obtidos do passo anterior são combinados com *k* elementos, formando o conjunto de *itemsets* candidatos. O *suporte*

de cada *itemset* candidato é calculado, e se for maior que o *suporte* pretendido é considerado *itemset* frequente. O processo repete-se até o número de *itemsets* frequentes ser zero. Este processo está exemplificado na tabela 3.1, tomando como *itens* os predicados espaciais do *dataset* mostrado na tabela 2.5.

Tabela 3.1: *Itemsets* frequentes com suporte 50%.

k	<i>itemsets frequentes com suporte mínimo 50%</i>
1	{'contains_waterways'}, {'contains_school'}, {'contains_hospital'}, {'contains_amenities'}, {'contains_roads'}
2	{'contains_hospital', 'contains_waterways'}, {'contains_school', 'contains_hospital'}, {'contains_school', 'contains_waterways'}, {'contains_school', 'contains_roads'}, {'contains_hospital', 'contains_roads'}, {'contains_amenities', 'contains_roads'}, {'contains_waterways', 'contains_roads'}
3	{'contains_school', 'contains_hospital', 'contains_roads'}, {'contains_school', 'contains_waterways', 'contains_roads'}, {'contains_school', 'contains_hospital', 'contains_waterways'}, {'contains_hospital', 'contains_waterways', 'contains_roads'}
4	{'contains_school', 'contains_hospital', 'contains_waterways', 'contains_roads'}

Fonte: Adaptado de (Bogorny et al., 2008).

A **dependência entre predicados espaciais** aparece pela primeira vez com *itemsets* frequentes de dimensão 2, e são replicados em *itemsets* frequentes com $k > 2$ que alcançam o valor mínimo de *suporte*. A replicação de uma dependência entre predicados espaciais, e.g., {'contains_school', 'contains_hospital'}, pode ser visualizada na tabela 3.1. A tabela 3.2 apresenta algumas regras geradas com a dependência assinalada na tabela 3.1.

Tabela 3.2: Algumas regras geradas com a dependência estudada.

k	<i>Regras possíveis</i>
2	{'contains_hospital'} \Rightarrow {'contains_school'}
2	{'contains_school'} \Rightarrow {'contains_hospital'}
3	{'contains_waterways'} \Rightarrow {'contains_school', 'contains_hospital'}
3	{'contains_school'} \Rightarrow {'contains_hospital', 'contains_waterways'}
3	{'contains_hospital'} \Rightarrow {'contains_school', 'contains_roads'}
4	{'contains_waterways', 'contains_roads'} \Rightarrow {'contains_school', 'contains_hospital'}
4	{'contains_hospital', 'contains_roads'} \Rightarrow {'contains_school', 'contains_waterways'}
4	{'contains_hospital', 'contains_waterways'} \Rightarrow {'contains_school', 'contains_roads'}
4	{'contains_school', 'contains_roads'} \Rightarrow {'contains_hospital', 'contains_waterways'}

Fonte: Adaptado de (Bogorny et al., 2008).

Na tabela 3.2, observa-se que o *itemset* { contains_school, contains_hospital }, que exprime uma dependência geográfica, é reproduzido nas regras apresentadas.

3.5.2 Análise formal

A análise de dependências (cf., secção 3.5.1), mostra que as regras geradas a partir de dependências contêm padrões conhecidos. Como forma de descrever de modo explícito regras que integram dependências conhecidas procedeu-se à construção de uma geo-ontologia. As características do conhecimento a incluir na geo-ontologia, que cumprem os requisitos enunciados (cf., secção 3.1), podem ser descritas por dois axiomas que a seguir se apresentam. Considere-se o *dataset* Ω , gerado pelo processo *Build dataset* (cf., secção 3.4), e P e Q dois quaisquer predicados nele contidos.

1. $\forall P \in \Omega : support(P) = 1$
2. $\forall P, Q \in \Omega : P \sqsubseteq Q$

O axioma 1 identifica todos os predicados com suporte igual a 100% e cumpre o primeiro tipo de dependência enunciado nos requisitos (cf., secção 3.1). O axioma 2 contém os predicados cujo conjunto de valores está subsumido ao conjunto de valores de outro predicado. Este axioma satisfaz o segundo tipo de dependência enunciado nos requisitos (cf., secção 3.1).

Para identificar as dependências que satisfazem o axioma 1, vão ser procurados, no *dataset*, predicados espaciais com 100% de *suporte*. Os predicados que cumprem esta condição serão removidos do *dataset*.

Para verificar o axioma 2, analisou-se o *dataset*, com o objetivo de detetar predicados, cujo conjunto de valores estivesse subsumido, ao conjunto de valores de outro predicado. Esta condição indica uma dependência com 100% de *confiança*.

$$p_1 \Rightarrow p_2, confiança = 1 \tag{3.1}$$

Os predicados do padrão, representados na dependência 3.1, atribuem um papel mais geral ao predicado p_2 do que ao predicado p_1 . Por esta razão, o padrão $p_1 \Rightarrow p_2, confiança = 1$, pode substituir-se pelo axioma de inclusão, $P_1 \sqsubseteq P_2$.

Conforme descrito na secção 3.5.1 um processo de geração de regras de associação, pode passar, por anteriormente, serem gerados todos os *itemsets* frequentes. Considerando uma dependência, $p_1 \Rightarrow p_2$, vão analisar-se os *itemsets* de dimensão K que contenham p_1 e p_2 , tal que foi detetado 'a priori' que $p_1 \Rightarrow p_2, confiança = 1$.

- $K = 2$, *itemset* $\{p_1, p_2\}$; A regra onde p_1 está no lado esquerdo e p_2 no lado direito, pode eliminar-se.
- $K = 3$, *itemset* $\{p_1, p_2, p_3\}$; Como $p_1 \sqsubseteq p_2$, então pode eliminar-se p_2 deste *itemset*, passando a ter-se $k = 2$, com $\{p_1, p_3\}$. Agora caso $\{p_1, p_3\}$, seja um *itemset* frequente então já terá sido tratado com $K = 2$. Caso não seja um *itemset* frequente então não dá origem a regras.
- $K = 4$, *itemset* $\{p_1, p_2, p_3, p_4\}$; Como $p_1 \sqsubseteq p_2$, então pode eliminar-se p_2 deste *itemset*, passando a ter-se $k = 3$, com $\{p_1, p_3, p_4\}$. Agora caso $\{p_1, p_3, p_4\}$, seja um *itemset* frequente então já terá sido tratado com $K = 3$. Caso não seja um *itemset* frequente então não dá origem a regras.
- e assim sucessivamente.

Em síntese, na análise de um *itemset* J de dimensão K que contenha p_1 e p_2 , tendo-se detetado 'a priori' que $p_1 \Rightarrow p_2$, *confiança* = 1, tem-se $p_1 \sqsubseteq p_2$, pelo que, p_2 é redundante em J . Eliminando p_2 de J fica-se com um *itemset* J' de dimensão $k-1$, tal que:

1. Se J' for um *itemset* frequente, então as suas regras já foram anteriormente geradas (aquando do tratamento dos *itemsets* de dimensão $K - 1$).
2. Se J' não for um *itemset* frequente, então não dá origem a regras.

Considerando, agora, regras, tem-se que, qualquer regra que contenha simultaneamente p_1 e p_2 no *antecedente* ou no *consequente*, ou p_1 no *antecedente* e p_2 no *consequente*, terá sido originada a partir de um *itemset* de dimensão K onde p_2 deveria ter sido eliminado, passando o *itemset* para dimensão $K-1$ e apenas com p_1 . Logo estes tipos de regras podem ser eliminados.

Na demonstração realizada para $p_1 \Rightarrow p_2$, *confiança* = 1, em cada k -*itemset* que contivesse p_1 e p_2 , podia-se eliminar p_2 , porque $p_1 \sqsubseteq p_2$. No acaso de se verificar uma dependência, tal que, $p_1 \Rightarrow p_2$, *confiança* = 1 e $p_2 \Rightarrow p_1$, *confiança* = 1, como $p_1 \sqsubseteq p_2 \wedge p_2 \sqsubseteq p_1 \Rightarrow p_1 \equiv p_2$, em cada k -*itemset* que contenha p_1 e p_2 , pode eliminar-se p_1 e p_2 .

Analisando, agora, as regras geradas a partir de *itemsets*, com p_1 e p_2 tal que, $p_1 \equiv p_2$, conclui-se que, podem eliminar-se regras que contenham p_1 ou p_2 , no *antecedente* ou no *consequente*.

3.6 Processo de pesquisa de regras de associação

O objetivo do algoritmo de pesquisa é encontrar medidas de *suporte* e de *confiança* que correspondam a um número de regras de associação espacial não superior a um valor definido pelo utilizador. Para atingir este objetivo, o algoritmo deve considerar todo o espaço constituído por estas medidas, que tomam valores no intervalo]0, 1]. O varrimento do espaço é conseguido através da análise em pontos, que se encontram distanciados por um salto no valor de *confiança* ou um salto no valor do *suporte*, considerando genericamente dois pontos consecutivos. Cada ponto neste espaço é caracterizado por duas coordenadas, (*suporte*, *confiança*). Como exemplo, a tabela 3.3 ilustra a sequência dos pontos analisados para valores de salto de *suporte* e de *confiança*, iguais a 0.25.

Tabela 3.3: Sequência de pontos analisados.

<i>Suporte mínimo</i>	1	1	2	3	4
	0.75	5	6	7	8
	0.5	9	10	11	12
	0.25	13	14	15	16
		1	0.75	0.5	0.25
	<i>Confiança mínima</i>				

Como se mostra na tabela 3.3, com os valores de salto indicados para *suporte* e *confiança* seriam analisados 16 pontos, pela ordem indicada. Em cada ponto analisado, o algoritmo vai contabilizando o número de regras extraídas. Quando o número de regras extraídas ultrapassar o valor definido pelo utilizador, o algoritmo cancela a extração de regras. As regras de associação são devolvidas pelo algoritmo, somente se, o número de regras extraídas não for superior ao valor definido pelo utilizador. A função de filtragem é utilizada por este algoritmo, para avaliar as regras extraídas de forma a que estas estejam de acordo com a parametrização do utilizador (cf., secção 3.4).

4

Implementação

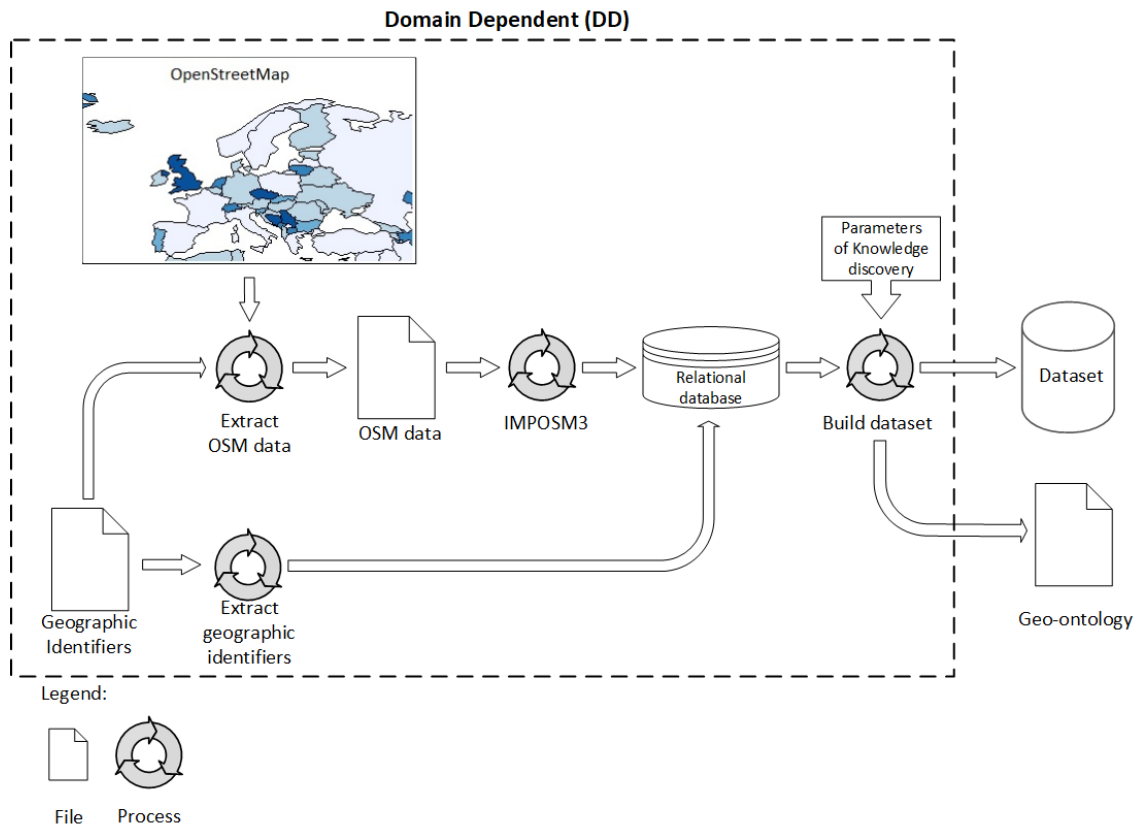
A partir do modelo proposto apresentado no capítulo 3, foi implementada uma solução com o objetivo de descoberta de conhecimento, na forma de regras de associação espacial, geradas a partir de zonas geográficas onde foram detetados surtos de gripe aviária. As ocorrências de surtos de gripe aviária definem-se no processo através de pontos geográficos. As zonas geográficas analisadas são constituídas por áreas concêntricas nos pontos geográficos de surto, definidas por um *buffer* espacial de valor definido pelo utilizador. Pretende-se também analisar a variação das regras de associação geradas segundo a variação da medida do *buffer* espacial que define a zona geográfica.

4.1 Arquitetura do modelo implementado

De acordo com o modelo proposto (cf., secção 3.3), a arquitetura do modelo desenvolvido foi dividida em dois blocos: *Domain Dependent* e *Domain Independent*.

4.1.1 Domínio dependente (DD)

A arquitetura do bloco *Domain Dependent* é apresentada na figura 4.1. A informação geográfica utilizada nesta implementação foi coletada pelo projeto OSM, e disponibilizada pela organização *Geofabrik*. Esta informação encontra-se simbolizada por um mapa na figura 4.1. O processo com vista à extração dos elementos

Figura 4.1: Arquitetura do bloco *Domain Dependent*.

geográficos pretendidos e ao seu posterior armazenamento numa base de dados espacial encontra-se descrito na secção 4.2.2. O conteúdo da base de dados está apresentado na secção 4.2.3.

Os dados constituídos por informação espacial e não espacial, que caracterizam os surtos de gripe aviária, encontram-se no ficheiro *Geographic identifiers*. O processo *Extract geographic identifiers* tem a função de armazenar na base de dados espacial os dados existentes no ficheiro *Geographic identifiers*.

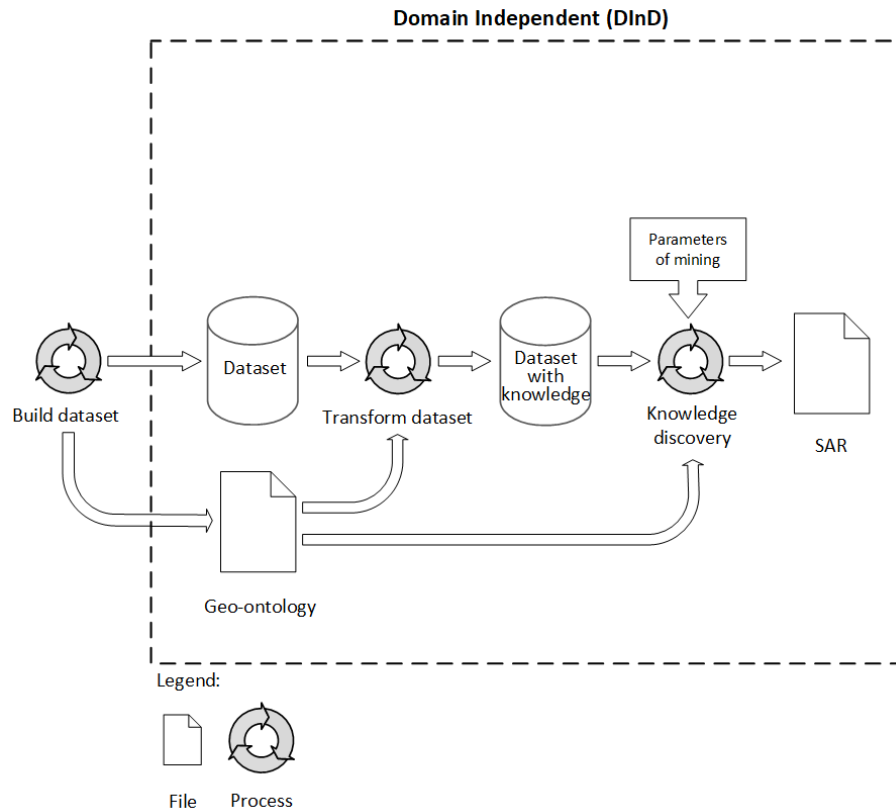
O conjunto de parâmetros, *Parameters of knowledge discovery*, determinam as linhas e as colunas da matriz (cf., secção 2.10). Este conjunto de parâmetros encontra-se descrito na secção 4.1.3.

O processo *Build dataset* realiza a construção do *dataset* segundo o modelo enunciado (cf., secção 3.2), utilizando os parâmetros definidos em *Parameters of knowledge discovery*. Depois da construção do *dataset* é realizada uma análise de dependências geográficas, com o objetivo de construir uma geo-ontologia com os requisitos enunciados na secção 3.1.

4.1.2 Domínio independente (DInD)

Na figura 4.2 apresenta-se a arquitetura do bloco *Domain Independent*.

Figura 4.2: Arquitetura do bloco *Domain Independent*.



O processo *Transform dataset* produz um *dataset* denominado na figura 4.2 por *Dataset with knowledge*. Este *dataset* é o resultado da integração do conhecimento existente na geo-ontologia, que foi selecionado pelo utilizador, no *dataset* construído pelo processo *Build dataset*.

O conjunto de parâmetros *Parameters of mining* encontra-se descrito na secção 3.4.

O processo *Knowledge discovery* implementa o algoritmo de pesquisa (cf., secção 3.6). A descrição desta implementação está apresentada na secção 4.1.9.

4.1.3 Definição de linhas e colunas da matriz

A definição das linhas e colunas da matriz que constitui o *dataset* é realizada através do conjunto de parâmetros *Parameters of knowledge discovery*. De acordo com o enunciado na secção 3.4, esta implementação tem o conjunto de parâmetros a seguir descrito.

- ***Linhas da matriz***

- ***Pontos de surto*** - A especificação de pontos de surto encontra-se descrita na secção 4.1.5.
- ***Buffer*** - A medida do *buffer* espacial. Nesta implementação esta medida tem o significado de um raio pelo fato das localizações dos surtos serem pontos geográficos.

- ***Colunas da matriz***

- ***Seleção dos tipos de recursos espaciais sobre os quais incide a descoberta de conhecimento*** - A informação geográfica está separada em camadas espaciais. Cada camada contém objetos espaciais com o mesmo tipo de geometria, *e.g.*, *roads* contém geometrias do tipo *linestring*. A camada *roads* é composta por vários subtipos, *e.g.*, *motorway*, *pedestrian*, ..., que permitem definir uma granularidade diferente, *e.g.*, composta somente por *roads* do tipo *pedestrian*. Em seguida, apresenta-se um exemplo de seleção de recursos espaciais que contém granularidades diferentes para diferentes camadas:

$$[('roads', ['pedestrian']), ('transport_areas',)]$$

No primeiro tuplo da lista são considerados os recursos espaciais da camada *roads* que sejam do tipo *pedestrian*. No segundo tuplo, a camada é considerada independentemente dos tipos que estejam definidos nessa camada.

- ***Operadores*** - Os operadores selecionados encontram-se descritos na secção 4.1.4.

4.1.4 Operadores topológicos e de distância

O princípio enunciado por (Koperski & Han, 1995) constitui base para delimitação de uma estratégia para seleção dos operadores topológicos a aplicar (cf., secção 2.8).

O operador de distância *close* foi incluído nos operadores possíveis de serem selecionados. Este operador fornece uma noção de proximidade entre dois recursos geográficos. Nesta implementação, foi configurado para detetar distâncias inferiores a metade da medida do *buffer* que limita cada geometria.

Neste trabalho, foi adotada a estratégia de seleção de operadores de acordo com o tipo de geometria existente em cada camada. Assim, foi implementada a seguinte relação, na aplicação de operadores topológicos e de distância:

- Geometria espacial do tipo *point* - {*within, close*}.
- Geometria espacial do tipo *linestring* - {*contains, crosses, close*}.
- Geometria espacial do tipo *polygon* - {*contains, overlaps, close*}.

No entanto, a aplicação desenvolvida encontra-se aberta a outras estratégias de seleção e adição de operadores.

4.1.5 Especificação de pontos de surto

Na seleção de pontos de surto são utilizados atributos alfanuméricos. Os pontos de surto registados referem-se a dois tipos de aves: aves do tipo *poultry* e aves do tipo *wild birds*. Sobre cada registo, além do atributo espacial, existem os seguintes atributos alfanuméricos:

- Tipo de ave.
- Ano em que foi detetado o surto.
- País.
- Região.
- Subtipo de infeção.

Além dos atributos descritos, podem ser utilizados para seleção de pontos de surto, interseções ou disjunções espaciais entre áreas infetadas, que variam de acordo com o *buffer* espacial. Estas possibilidades de seleção descrevem-se a seguir.

- Áreas disjuntas.
- Áreas que intercetam outras áreas infetadas pelo mesmo tipo de aves.
- Áreas que intercetam outra áreas infetadas pelo outro tipo de aves.
- Áreas que intercetam outra áreas infetadas independentemente do tipo de ave.

A seleção de pontos de surto realiza-se de acordo com a sintaxe que a seguir se apresenta.

```
{ 'global': ( <year>, <country>, <region>, <type> ),
  'habitat': ( <habitat>, <intersect_same>, <intersect_other> )+ }
```

Este *template* de seleção está dividido em duas partes, *global* e *habitat*. A parte *global* refere-se a uma condição geral sobre os registos do fenómeno. A parte *habitat* refere-se a um tipo de ave que se pretenda seleccionar. Em seguida apresenta-se um exemplo de aplicação que utiliza esta sintaxe.

```
{ 'global': ( [], [], [], [] ),
  'habitat': ( ( 'poultry', None, None ), ( 'wild_birds', None, None ) ) }
```

Neste exemplo, na parte global pretendem-se seleccionar pontos de surto independentemente do ano, país, região e tipo de surto. Na parte específica a cada tipo de ave, não se pretende filtragem alguma, i.e., pretende-se considerar todos os surtos registados.

4.1.6 Exemplo de *dataset*

Nesta secção é apresentado um exemplo de *dataset* com indicação dos valores de parâmetros a utilizar. Assim, para definição das linhas e das colunas da matriz (cf., secção 4.1.3), foram utilizados os seguintes valores:

- **Linhas da matriz**

- Ocorrências do fenómeno

```
{ 'global': ( [], ['ITALY'], [], [] ),
  'habitat': ( ( 'wild_birds', None, None ), ) }
```

De acordo com a sintaxe apresentada (cf., secção 4.1.5), pretendem-se seleccionar, pontos de surto registados no país *ITALY*, referentes ao tipo de aves *wild birds*.

- *buffer* = 5 Km.

- **Colunas da matriz**

- [('roads', ['pedestrian']), ('transport_points',)]

Os valores atribuídos a este parâmetro selecionam para análise os recursos geográficos, *roads* do tipo *pedestrian*, e todos os tipos de *transport_points*.

– Operadores

Os operadores selecionados, de acordo com as figuras geométricas que representam os recursos geográficos representados nas colunas (cf., secções 4.1.4, 4.4), são apresentados na tabela 4.1.

Tabela 4.1: Operadores selecionados

Tipo de recurso	Forma geométrica	Operadores
<i>roads</i>	<i>linestring</i>	{ <i>contains, crosses, close</i> }
<i>transport_points</i>	<i>point</i>	{ <i>within, close</i> }

Na tabela 4.2 é mostrado o *dataset* obtido.

Tabela 4.2: Exemplo de *dataset*.

Áreas \ Predicados	contains	crosses	close	within	close
	roads_pedestrian	roads_pedestrian	roads_pedestrian	transport_points	transport_points
Italy_wild_bird_1	1	0	0	1	1
Italy_wild_bird_2	1	0	1	1	0
Italy_wild_bird_3	1	0	1	0	0
Italy_wild_bird_4	1	1	0	1	1
Italy_wild_bird_5	1	1	1	0	0

4.1.7 Construção da geo-ontologia

A construção da geo-ontologia foi dividida em duas fases. Na primeira fase, é obtido conhecimento que verifica o axioma 1 (cf., secção 3.5.2), i.e., identificação de colunas com 100% de *suporte*, e.g., '*contains_roads_pedestrian*', predicado retirado do *dataset* apresentado na tabela 4.2. O algoritmo 1 implementa esta fase. Na segunda fase são analisadas as dependências geográficas entre todas as colunas do *dataset*. A análise de dependências geográficas entre colunas do *dataset* consiste em encontrar colunas, cujo conjunto de valores que lhes está associado, esteja subsumido, ao conjunto de valores que representa outra coluna. Esta análise, realizada pelo algoritmo 2, deteta as dependências que verificam o axioma 2 (cf., secção 3.5.2).

Algorithm 1 Algoritmo de pesquisa de predicados com 100% de *suporte*.

```

1: function AXIOM_1( D )
2: // D: dataset
3:
4:   predicates  $\leftarrow$  [ ]
5:   for all columns c  $\in$  D do
6:     predicate  $\leftarrow$  c[0]
7:     if All_ones(c[ 1: ]) then
8:       predicates += predicate
9:     end if
10:  end for
11:  return predicates
12: end function

```

Algorithm 2 Algoritmo de pesquisa de dependências geográficas.

```

1: function AXIOM_2( D )
2: // D: dataset
3:
4:   predicates  $\leftarrow$  D[ 0, : ]
5:   values  $\leftarrow$  D[ 1, : ]
6:   number_col  $\leftarrow$  length( predicates )
7:   for left = 0; left < number_col; ++ left do
8:     for right = number_col; right >= 0; - right do
9:       if left = right then
10:        continue
11:       end if
12:       predicate_left  $\leftarrow$  predicates[left]
13:       predicate_right  $\leftarrow$  predicates[right]
14:       collumn_left  $\leftarrow$  D[ 1 :, left ]
15:       collumn_right  $\leftarrow$  D[ 1 :, right ]
16:       dependence  $\leftarrow$  true
17:       for i = 0; i > length(collumn_left); ++ i do
18:         if collumn_left[i] > collumn_right[i] then
19:           dependence = false
20:           break
21:         end if
22:       end for
23:       if dependence then
24:         dependencies += ( predicate_left, predicate_right )
25:       end if
26:     end for
27:   end for
28:   return dependencies
29: end function

```

4.1.8 Filtro do processo de extração

A função do filtro do processo de extração de conhecimento é regular o processo de pesquisa, no espaço de *suporte* e de *confiança*, assim como os processos de extração de *itemsets* frequentes e regras de associação espacial (cf., secção 3.4). Esta função é parametrizada no início do processo e colabora nas suas várias fases. Na função de pesquisa do espaço de *suporte* e de *confiança* parametriza os saltos nestes espaços. Na extração de padrões frequentes delimita as dimensões admissíveis de *itemsets*. Na extração de regras de associação espacial, determina as regras admitidas no resultado final, verificando as dimensões do *antecedente* e do *consequente*. Ainda no processo de extração de regras de associação, pode determinar o cancelamento deste processo, pelo fato do número de regras extraídas ter ultrapassado o valor definido pelo utilizador. Uma implementação deste filtro pode ser encontrada no apêndice A.

4.1.9 Algoritmo de pesquisa

A função de pesquisa de regras de associação, no espaço de *suporte* e *confiança*, foi implementada pelo algoritmo 3 que cumpre os requisitos enunciados na secção 3.6. O algoritmo 3 analisa o número de regras de associação, obtidas em pontos do espaço de *suporte* e de *confiança*, e compara com o número de regras obtidas, na análise do ponto anterior. Quando entre dois pontos consecutivos, se deteta uma variação significativa, no número de regras obtidas, procede-se a um ajuste de *suporte* ou de *confiança*. As funções *Adjust_Support* e *Adjust_Confidence* realizam um ajuste, no valor de *suporte* ou de *confiança*, respetivamente. A descrição de situações que caracterizam uma variação significativa, no número de regras obtido, é apresentada na secção 4.1.10. Em cada ponto analisado, é chamada a função *Get_Rules*, que executa as seguintes ações:

1. Extração de *itemsets* frequentes. Os *itemsets* extraídos têm *suporte* mínimo não inferior ao *suporte* do ponto analisado. A extração de *itemsets* frequentes pode ser realizada pelo algoritmo *FP-growth* (Han, Pei, Yin, & Mao, 2004).
2. A partir dos *itemsets* frequentes obtidos, no passo anterior, e com recurso a um algoritmo de extração de regras de associação, e.g., (Agrawal et al., 1996), são obtidas as regras de associação com *confiança* mínima não inferior à *confiança* do ponto analisado.

Algorithm 3 Algoritmo de pesquisa de regras de associação.

```

1: function SEARCH_RULES( D, mining_filter,  $\phi$ )
2: // D: dataset
3: // mining_filter: extraction process filter function
4: //  $\phi$ : geo-ontology
5:
6:   step_sup  $\Leftarrow$  mining_filter('step_support')
7:   step_conf  $\Leftarrow$  mining_filter('step_confidence')
8:   N  $\Leftarrow$  mining_filter('max_number_rules')
9:   rules_acc  $\Leftarrow$  [ ]
10:  last_nbr_rules  $\Leftarrow$  nothing
11:  for sup = 1; sup > 0; step = -step_sup do
12:    for conf = 1; conf > 0; step = -step_conf do
13:      rules  $\Leftarrow$  Get_Rules(D, mining_filter, sup, conf,  $\phi$ )
14:      nbr_rules  $\Leftarrow$  #rules
15:      if ( last_nbr_rules > N and nbr_rules  $\leq$  N ) or
16:          ( last_nbr_rules  $\leq$  N and nbr_rules > N ) then
17:        if conf = 1 then
18:          new_sup  $\Leftarrow$  Adjust_Support(D, mining_filter, sup, sup-1,
19:                                     sup, conf,  $\phi$ )
20:          rules  $\Leftarrow$  Get_Rules(D, mining_filter, new_sup, conf,  $\phi$ )
21:        else
22:          new_conf  $\Leftarrow$  Adjust_Confidence(D, mining_filter, conf,
23:                                             conf-1, sup, conf,  $\phi$ )
24:          rules  $\Leftarrow$  Get_Rules(D, mining_filter, sup, new_conf,  $\phi$ )
25:        end if
26:      end if
27:      last_nbr_rules  $\Leftarrow$  nbr_rules
28:      if #rules  $\leq$  N then
29:        rules_acc  $\Leftarrow$  rules_acc + rules
30:      end if
31:    end for
32:  end for
33:  return rules_acc
34: end function

```

3. Caso o número de regras, obtidas no passo anterior, não seja superior a um valor definido pelo utilizador, estas são devolvidas.

O algoritmo da função *Get_Rules* é apresentado no algoritmo 4.

Algorithm 4 Algoritmo de obtenção de regras de associação.

```

1: function GET_RULES (D, mining_filter, support, confidence,  $\phi$ )
2: // D : dataset
3: // mining_filter: extraction process filter function
4: //  $\phi$ : geo-ontology
5:
6:   rules  $\leftarrow$  [ ]
7:   stop  $\leftarrow$  False
8:   for itemset in frequent_itemsets( D, support ) do
9:     if not mining_filter( 'itemset', itemset ) then
10:       continue
11:     end if
12:     for rule, lift in association_rules( itemset, confidence ) do
13:       if not mining_filter( 'rule', rule ) then
14:         continue
15:       end if
16:       if not mining_filter( 'lift', lift ) then
17:         continue
18:       end if
19:       if discard_rule( $\phi$ , rule) then
20:         continue
21:       end if
22:       rules += rule
23:       if #rules > mining_filter('max_number_rules') then
24:         stop = True
25:         break
26:       end if
27:     end for
28:     if stop then
29:       break
30:     end if
31:   end for
32:   return rules
33: end function

```

4.1.10 Variação significativa do número de regras

Uma variação significativa do número de regras de associação obtidas, entre pontos consecutivos, é detetada quando surge uma das situações explanadas na tabela 4.3. Na tabela 4.3, N representa o número máximo de regras de associação requerido pelo utilizador, em cada ponto analisado. Na tabela 4.3, na primeira situação passou-se de um número de regras superior a N, para zero, e na segunda, o número de regras encontrava-se no intervalo indicado e passou para um número superior a N. Em cada ponto analisado, além de calculado o número

de regras, este é comparado com o número de regras obtido no ponto anterior. Quando é detetada uma das situações presentes na tabela 4.3, procede-se a um ajuste do valor de *suporte* ou de *confiança*. O objetivo do ajuste do valor de *suporte* ou de *confiança* é encontrar um ponto, (*suporte*, *confiança*), tal que, o número de regras desse ponto esteja no intervalo] 0, N], o que pode não ser possível. Este processo é realizado, recorrendo a uma pesquisa dicotómica entre dois valores de *suporte* ou de *confiança*. A pesquisa dicotómica é realizada entre o valor atual e o valor anterior, na medida *suporte* ou na medida *confiança*. Quando se procede a um ajuste no valor de *suporte*, mantém-se constante o valor de *confiança*, e vice-versa.

Tabela 4.3: Caracterização de situações de análise.

Número de regras do ponto anterior	Número de regras do ponto atual
> N	0
] 0, N]	> N

Considerando um ajuste no valor de *suporte*, este processa-se da seguinte forma:

1. Consideram-se dois valores de suporte, o atual, denominado *limite_inf* e o anterior, denominado *limite_sup*;
2. Calcula-se um novo suporte, *new_support*, que tem valor situado no centro da diferença dos dois limites anteriores;
3. Com *new_support*, calcula-se o número de regras. Caso o valor obtido seja:
 - > N - Repete-se o algoritmo a partir de 1, alterando o valor de *limite_sup* para o valor de *new_support*;
 - 0 - Repete-se o algoritmo a partir de 1, alterando o valor de *limite_inf* para o valor de *new_support*;
 -] 0, N] - O processo termina e está encontrado um valor de *suporte* que permite extrair um número de regras inferior a N.

O algoritmo termina quando for encontrado um número de regras que esteja no intervalo] 0, N], ou quando a diferença entre *limite_sup* e *limite_inf* é inferior a um infinitésimo, ϵ , neste ultimo caso não é possível realizar o ajuste da medida.

Os algoritmos de ajuste de *suporte* ou de *confiança* não garantem encontrar um número de regras não superior ao valor definido pelo utilizador. Quando esta

situação ocorre, existe uma variação do número de regras, de 0 a um número superior ao valor definido pelo utilizador, ($>N$), num espaço de variação inferior a um ϵ da medida, *suporte* ou *confiança*, que estiver a ser ajustada. O infinitésimo ϵ é a sensibilidade máxima dos algoritmos de ajuste de *suporte* e de *confiança*. O algoritmo 5 mostra o pseudocódigo do algoritmo de ajuste de suporte.

Algorithm 5 Algoritmo de ajuste do valor de suporte.

```

1: function ADJUST_SUPPORT( D, mining_filter, iS, supp-1, supp, conf,  $\phi$ )
2: // D: dataset
3: // mining_filter: extraction process filter function
4: // iS: initial support
5: // supp-1: last support
6: // supp: support
7: // conf: confidence
8: //  $\phi$ : geo-ontology
9:
10:    $N \leftarrow \text{mining\_filter}('max\_number\_rules')$ 
11:    $limit\_sup \leftarrow supp_{-1}$ 
12:    $limit\_inf \leftarrow supp$ 
13:    $\epsilon \leftarrow 1 \times 10^{-15}$ 
14:    $new\_support \leftarrow limit\_inf + ((limit\_sup - limit\_inf)/2)$ 
15:   if  $limit\_sup - limit\_inf < \epsilon$  then
16:     return iS
17:   end if
18:    $rules \leftarrow \text{Get\_Rules}(D, \text{mining\_filter}, new\_support, conf, \phi)$ 
19:    $nbr\_rules \leftarrow \#rules$ 
20:   if  $nbr\_rules > N$  then{
21:      $limit\_sup \leftarrow limit\_sup$ 
22:      $limit\_inf \leftarrow new\_support$ }
23:   else if  $nbr\_rules > 0$  then
24:     return  $new\_support$ 
25:   else{
26:      $limit\_sup \leftarrow new\_support$ 
27:      $limit\_inf \leftarrow limit\_inf$ }
28:   end if
29:   return  $Adjust\_Support(D, \text{mining\_filter}, iS, limit\_sup, limit\_inf, conf, \phi)$ 
30: end function

```

O funcionamento do algoritmo *Adjust_Confidence* é idêntico ao funcionamento do algoritmo utilizado para a função *Adjust_Support*. A função *Adjust_Support* é evocada em pontos com *confiança* igual a 1, na situação de variação significativa do número de regras. Nos restantes pontos, com *confiança* inferior a 1, é evocada a função *Adjust_Confidence*.

4.1.11 Utilização da geo-ontologia

O conhecimento selecionado pelo utilizador, na geo-ontologia, com o objetivo de diminuir o número de regras de associação geradas é utilizado em duas fases.

Na primeira fase, o processo *Transform dataset* (cf., secção 4.1.2), toma como *input* o *dataset* produzido pelo processo *Build dataset* (cf., secção 4.1.1), e constrói o *dataset* designado por *Dataset with knowledge* (cf., secção 4.1.2). O processo *Transform dataset* remove do *dataset* de *input* os predicados espaciais que verificam o axioma 1 (cf., secção 3.5.2). O algoritmo 6 implementa este processo.

Algorithm 6 Algoritmo de remoção de predicados com 100% de *suporte*.

```

1: function TRANSFORM_DATASET(  $D, \phi$ )
2: //  $D$  : dataset
3: //  $\phi$ : geo-ontology
4:
5:    $predicates \leftarrow D[1, :]$ 
6:   for all  $predicate$  in  $predicates$  do
7:     if  $predicate \in \phi$  then
8:        $D = D - D[:, predicate]$ 
9:     end if
10:  end for
11:  return  $D$ 
12: end function

```

Na segunda fase, são descartadas as regras de associação geradas que contêm dependências identificadas pelo axioma 2 (cf., secção 3.5.2). O algoritmo 7 identifica regras que contêm as dependências selecionadas. Cada regra gerada no algoritmo 4 é analisada com o objetivo de detetar a existência de dependências. Caso existam, o algoritmo 4 descarta a regra.

4.2 Informação geográfica

4.2.1 Modelo de dados do OpenStreetMap

O modelo de dados do OSM, utiliza três tipos de objetos: *node*, *way* e *relation*, (*Elements - OpenStreetMap@ONLINE*, n.d.). Estes tipos de objetos, que expressam conceitos, são etiquetados com *tags*, para expressar pontos, linhas ou polígonos, na forma que os Sistemas de Informação Geográfica (SIG) normalmente expressam. As *tags* são pares nome-valor anexadas a um recurso geográfico e

Algorithm 7 Algoritmo de identificação de regras que contêm dependências.

```

1: function DISCARD_RULE ( $\phi$ , rule )
2: //  $\phi$ : geo-ontology
3:
4:   LHS_itens  $\Leftarrow$  LHS(rule)
5:   RHS_itens  $\Leftarrow$  RHS(rule)
6:   for all LHS_dependency, RHS_dependency in  $\phi$  do
7:     if LHS_dependency  $\in$  LHS_itens and RHS_dependency  $\in$  RHS_itens then
8:       return True
9:     else if ( { LHS_dependency, RHS_dependency }  $\subseteq$  LHS_itens ) or
10:    ( { LHS_dependency, RHS_dependency }  $\subseteq$  RHS_itens ) then
11:       return True
12:     end if
13:   end for
14:   return False
15: end function

```

destinam-se a descrevê-lo. O *node* é caracterizado por uma latitude e uma longitude e representa, *e.g.*, *place*. Uma *way* refere-se a uma lista ordenada de *nodes*, e destina-se a representar, *e.g.*, *roads*. Uma *relation* é uma lista ordenada de *node*, *way* ou *relation* e exprime uma relação próxima, local, entre objetos geográficos. Qualquer elemento geográfico possui um *id* numérico único. Os ficheiros de dados OSM são normalmente distribuídos em formato *eXtensible Markup Language* (XML) compactado. O formato *Protocolbuffer Binary Format* (PBF) é um formato binário que melhora a velocidade de análise, assim como diminui o tamanho do ficheiro, em relação ao formato XML compactado.

4.2.2 Processamento de dados *OpenStreetMap*

Considerando a área de dispersão geográfica ocupada, pelos pontos geográficos disponíveis para teste do modelo, tornou-se necessário importar todo o continente europeu. A organização *Geofabrik* disponibiliza dados geográficos OSM abrangendo todo o continente.

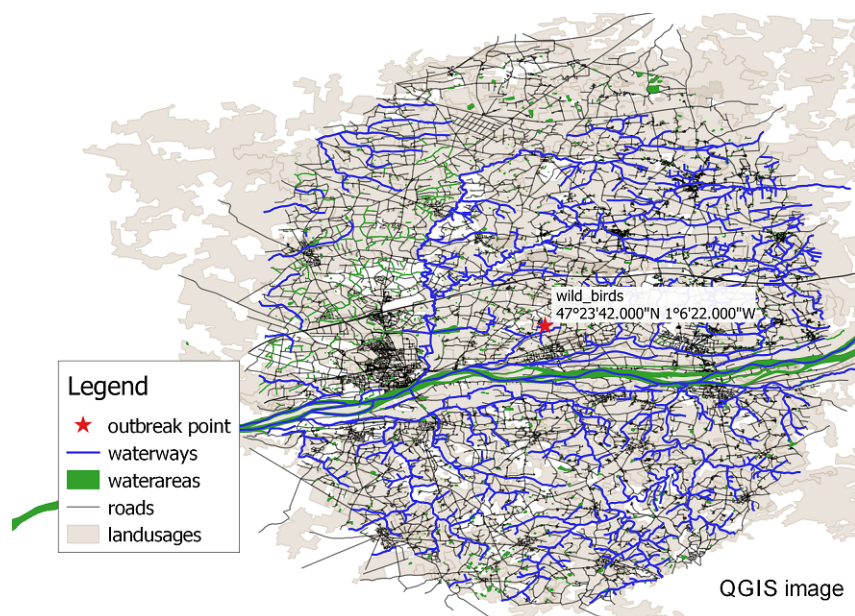
A *Framework Imposm3* (*Imposm3 - Imposm3 3.0.0a documentation@ONLINE*, n.d.), é um importador de dados OSM, executável em ambiente *Linux*, capaz de ler dados no formato PBF, para uma base de dados *PostgreSQL*. A base de dados deve ter instalado um extensor *PostGIS*, que dota a base de dados da capacidade de gerir informação espacial. A *Framework Imposm3* permite definir um ficheiro de configuração¹, contendo o nome das tabelas a criar na base de dados. Permite

¹ver apêndice B

também definir os pares chave/valor, que identificam *nodes*, que se pretendem em cada tabela.

A dimensão do ficheiro PBF, que contém o continente europeu, torna quase impraticável a sua importação para uma base dados. A título de exemplo, foram gastas 12 horas para ler cerca de 4.5% do ficheiro, num computador pessoal, utilizando uma máquina virtual com 5 GB de RAM. Devido a esta limitação, tornou-se necessário definir um processo no qual somente seriam importados os dados necessários. O processo de importação tem duas etapas, uma de leitura do ficheiro e uma de escrita na base de dados, de acordo com o ficheiro de configuração. O utilitário *osmconvert* (*Ubuntu Manpage: osmconvert - converter of OSM files@ONLINE, n.d.*) permite extrair partes de um ficheiro PBF, definidas por polígonos, representados num ficheiro de formato *Polygon Filter File Format* (POLY) (*OsmosisPolygon Filter File Format@ONLINE, n.d.*). Foi desenvolvida uma aplicação que produz um ficheiro de formato POLY. O processo que cria este ficheiro toma, como *input*, as coordenadas geográficas dos pontos onde foram localizados surtos, e produz um ficheiro, que tem definidos círculos concêntricos, nesses pontos. Neste trabalho foi assumido um raio de 16 km, constituindo este valor, o raio máximo para descoberta de conhecimento. A figura 4.3 mostra as camadas geográficas indicadas relativas a um ponto geográfico de surto.

Figura 4.3: Camadas geográficas extraídas relativas a um ponto geográfico de surto.



4.2.3 Base de dados espacial

A base de dados que contém informação geográfica, obtida do projeto OSM, contém as seguintes tabelas: *aeroways*, *amenities*, *barrierpoints*, *barrierways*, *buildings*, *landusages*, *railways*, *roads*, *transport_areas*, *transport_points*, *waterareas* and *waterways*. O nome e o conteúdo destas tabelas estão de acordo com o ficheiro de configuração utilizado na *Framework IMPOSM3*. Cada tabela contém um tipo de recurso espacial. Estes recursos espaciais referem-se a elementos etiquetados por uma relação chave/valor, como se mostra na tabela 4.4.

Tabela 4.4: Relação de tabelas existentes na base de dados espacial.

Nome da tabela	Tipo de geometria	Chave	Valor
aeroways	linestring	aeroway	runway, taxiway
amenities	polygon	amenity	university, school, college, library, fuel, parking, cinema, theatre, place_of_worship, hospital, swimming_pool
		leisure	swimming_pool
barrierpoints	point	barrier	block, bollard, cattle_grid, chain, cycle_barrier, entrance, horse_stile, gate, spikes, lift_gate, kissing_gate, fence, yes, wire_fence, toll_booth, stile
barrierways	linestring	barrier	city_wall, fence, hedge, retaining_wall, wall, bollard, gate, spikes, lift_gate, kissing_gate, embankment, yes, wire_fence
buildings	polygon	building	any
		barrier	hedge
landusages	polygon	leisure	park, garden, playground, golf_course, sports_centre, pitch, stadium, common, nature_reserve
		tourism	zoo
		natural	wood, land, scrub, wetland, heath
		place	island
		landuse	park, forest, residential, retail, commercial, industrial, railway, cemetery, grass, farmyard, farm, farmland, orchard, vineyard, wood, meadow, village_green, recreation_ground, allotments, quarry
railways	linestring	railway	rail, tram, light_rail, subway, narrow_gauge, preserved, funicular, monorail, disused
roads	linestring	man_made	pier, groyne
		highway	motorway, motorway_link, trunk, trunk_link, primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, road, path, track, service, footway, bridleway, cicleway, steps, pedestrian, living_street, unclassified, residential, raceway
transport_areas	polygon	railway	station, platform
		aeroway	aerodrome, terminal, helipad, apron
transport_points	point	railway	station, halt, tram_stop, crossing, level_crossing, subway_entrance
		aeroway	aerodrome, terminal, helipad, gate
		highway	motorway_junction, turning_circle, bus_stop
waterareas	polygon	waterway	riverbank
		landuse	basin, reservoir
		natural	water
waterways	linestring	waterways	stream, river, canal, drain, ditch
		barrier	ditch

5

Estudo de caso

Nos capítulos anteriores justificaram-se os passos de concepção de uma plataforma de descoberta de conhecimento na forma de regras de associação espacial. Apresentou-se uma implementação orientada para a extração de conhecimento em áreas concêntricas em pontos geográficos onde foram detetadas ocorrências de um fenómeno. A plataforma concretizada vai ser aplicada na análise de um problema epidémico, a gripe aviária. Com base em ocorrências deste fenómeno no espaço geográfico europeu, vai proceder-se, neste capítulo, à análise de resultados produzidos pela aplicação desenvolvida.

5.1 A infeção da gripe aviária

O vírus da gripe aviária denominado H5N1 é um sub-tipo do vírus *Influenza A*. Este vírus é muito contagioso entre aves, mas raramente se transmite a humanos ([Organization, 2007b](#)), embora existam alguns casos registados. As aves aquáticas selvagens são consideradas o reservatório natural do vírus e, quando infetadas, não apresentam sintomas, mas perpetuam a transmissão do vírus às aves domésticas. No entanto, as aves aquáticas domésticas, *e.g.*, patos, também podem ser portadoras da transmissão bidirecional do vírus entre aves aquáticas selvagens e aves de capoeira, *e.g.*, galinhas. A hospedagem do vírus em algumas espécies aviárias pode causar mortalidade significativa.

A ocorrência de vírus de alta patogenicidade designado por *Highly Pathogenic Avian Influenza* (HPAI), em aves selvagens e a probabilidade de infecção de explorações de aves de capoeira, é apoiada pela associação entre detecções do vírus, em aves selvagens, e aves de capoeira no campo (Authority, 2017). Na Europa, os surtos de HPAI em aves de capoeira deveram-se a incursões de aves selvagens migratórias infetadas (Authority, 2017). O vírus transmitido por aves aquáticas selvagens também pode ter baixa patogenicidade designando-se por *Low Pathogenic Avian Influenza* (LPAI). Contudo, este vírus pode sofrer mutações durante a replicação em aves de capoeira e transformar-se em cepas de gripe aviária altamente patogénica, i.e., HPAI.

Na opinião de especialistas, a prevenção do acesso de aves de capoeira, a superfícies de água, pode resultar numa redução para um terço, da probabilidade de contágio (Authority, 2017). As aves infetadas derramam grandes quantidades de vírus através das fezes, na saliva e nas secreções nasais, predominando a primeira (*Avian Influenza*, 2016). É provável que estes excrementos se juntem em ambientes de água utilizados por outras aves.

Além da deposição direta de fezes de aves aquáticas migratórias, em ambientes de água, é sugerida, no documento (Organization, 2007b), a possibilidade de resíduos fecais de aves domésticas entrarem nesses ambientes, via vento ou escoamento superficial, ou entrarem em águas subterrâneas através da compostagem de resíduos de aves de capoeira.

O vírus da gripe aviária pode persistir na água, por longos períodos de tempo (Organization, 2007a), dependendo da temperatura, *pH* e salinidade da água. Para o vírus H5N1, não existem dados quantitativos sobre a sua permanência, no entanto, a título de exemplo, refira-se que, num estudo realizado com o subtipo H3N6, no rio *Mississippi* (EUA), o vírus foi detetado por mais de 30 dias a 0°C e foi indetetável após 4 dias a 22°C (Organization, 2007a).

Devido ao fato do vírus poder permanecer ativo em ambientes de água, estas constituem um fator de risco, quando são partilhadas por aves aquáticas selvagens e aves domésticas. Este risco foi sugerido por (Stallknecht, Shane, Kearney, & Zwank, 1990).

O fato do vírus da gripe aviária ser transmitido e mantido na população de aves selvagens pela via fecal oral indiretamente através de água contaminada está documentado em (Stallknecht & Brown, 2009). Num estudo realizado em dois aviários de patos, foram examinados mensalmente, durante um período de um ano, a ocorrência e persistência de vírus nas comunidades de patos. O vírus da gripe

foi isolado mensalmente ao longo do ano a partir das fezes ou água da lagoa ou ambos, indicando um ciclo de transmissão mantido pela água (Markwell & Shortridge, 1982).

Uma exploração com aves infetadas gera um ambiente capaz de alojar e disseminar o vírus devido às atividades próprias e características físicas destes locais (Sharkey, Bowers, Morgan, Robinson, & Christley, 2008). As superfícies porosas ou não porosas que podem ser contaminadas com microrganismos patogénicos e servir de veículo de transmissão do vírus, designam-se por *fomites* (Boone & Gerba, 2007). Alguns materiais, *e.g.*, vidro, metal galvanizado, solo superficial, que constituem estas superfícies permitem o alojamento e a disseminação do vírus (Wood, Choi, Chappie, Rogers, & Kaye, 2010).

Na maioria dos casos, a LPAI permanece restrita a uma única exploração, embora a disseminação horizontal tenha sido observada em diversas ocasiões. A propagação de vírus HPAI entre explorações é altamente provável na ausência de medidas de controle (Authority, 2017). O vírus pode ser transmitido entre explorações através de movimentos de pessoas, *e.g.*, equipas de vacinação, trabalhadores. O movimento de meios de transporte de aves para abate é igualmente causador da disseminação do vírus. Explorações geograficamente próximas são suscetíveis de ser contaminadas por poeira soprada, *e.g.*, vento, de uma exploração para outra, ou vida selvagem carregando poeira contaminada.

Trabalhos recentes, nesta área, apontam como fatores de risco de disseminação de HPAI movimentos associados ao comércio de sémen dentro dos países da União Europeia e com países terceiros e movimentos intra comunitários de estrume provenientes de explorações com espécies de Anseriformes (ordem de aves aquáticas) (Authority, 2017).

Para além da via de aves selvagens, na disseminação do vírus LPAI, foram considerados como tendo um risco de introdução não negligenciável movimentos intra países, da União Europeia, de aves de capoeira vivas e pintos do dia, assim como movimentos intra comunitários de estrume provenientes de explorações com qualquer espécie (Authority, 2017). Na disseminação do vírus entre fronteiras podem intervir fatores naturais, *e.g.*, migração de aves selvagens e vento, e fatores acidentais, *e.g.*, embarque de produtos infetados (Monke & Corn, 2007).

Depois de detetada uma zona de surto, não existem evidências científicas que conduzam à definição de um perímetro de monitorização e controlo. Este perímetro, depende das espécies de aves infetadas, do habitat e do alcance do voo, no caso das aves selvagens (Authority, 2017).

5.2 Dados do fenômeno

A base de dados geográfica constituída permite responder ao conjunto de parâmetros *Parameters of knowledge discovery* (cf., secção 4.1.3). Estes parâmetros definem as linhas e as colunas da matriz a constituir. Para a definição espacial das linhas é necessária a informação espacial dos pontos onde foram localizados surtos e a medida de *buffer* imposto pelo utilizador. Para a definição espacial das colunas contribui a seleção de tipos de recursos espaciais (cf., tabela 4.4), que estão relacionados com os objetivos da descoberta de conhecimento (Bogorny et al., 2008).

O universo de ocorrências do fenômeno é constituído por 1019 ocorrências relativas a aves do tipo *poultry* e 1365 ocorrências relativas a aves do tipo *wild birds*. Os anos destas ocorrências situam-se em 2016 e 2017. Os países onde ocorreram abrangem uma parte significativa do continente europeu. A distribuição de ocorrências por países e por tipo de aves está indicada na tabela 5.1. O posicionamento relativo das áreas referentes a cada ponto geográfico de surto, de acordo com o *buffer* espacial imposto, está indicado na tabela 5.2.

5.3 Experiências

O objetivo das experiências realizadas é identificar as camadas relacionadas com padrões espaciais relevantes na análise efetuada. Os padrões espaciais constituem regras de associação espacial extraídas em áreas concêntricas em pontos geográficos, formadas por um *buffer* espacial variável, onde foram localizados surtos de gripe aviária.

O estudo realizado sobre o problema da gripe aviária (cf., secção 5.1), mostrou uma forte relação entre o elemento água e a disseminação desta doença. A utilização da plataforma desenvolvida nestas experiências foi direcionada para avaliar a relevância de elementos geográficos, que contêm água, nos padrões espaciais extraídos em áreas onde ocorreram surtos.

Os habitats, naturais e não naturais, das aves, florestas (*landusages*) para aves do tipo *wild birds* e explorações (*landusages*) para aves do tipo *poultry* justificam a inclusão desta camada nas experiências realizadas. A natural movimentação de pessoas e veículos fazem com que a camada *roads* assumam importância na análise do problema da gripe aviária.

Tabela 5.1: Número de ocorrências por países.

Country	poultry	wild_birds	Total
GERMANY	74	615	689
FRANCE	454	48	502
HUNGARY	234	58	292
POLAND	64	68	132
ROMANIA	16	82	98
SWITZERLAND	0	87	87
BULGARIA	67	13	80
CZECH REPUBLIC	37	33	70
SLOVAKIA	8	57	65
NETHERLANDS	9	45	54
AUSTRIA	2	50	52
DENMARK	1	45	46
SLOVENIA	0	40	40
UNITED KINGDOM	10	22	32
SWEDEN	3	23	26
REPUBLIC OF SERBIA	4	20	24
CROATIA	7	12	19
GREECE	5	9	14
ITALY	9	5	14
FINLAND	0	13	13
SPAIN	10	2	12
IRELAND	0	10	10
UKRAINE	3	1	4
LITHUANIA	0	3	3
BELGIUM	0	2	2
THE FORMER YUGOSLAV RE- PUBLIC OF MACEDONIA	1	1	2
BOSNIA AND HERZEGOVINA	1	0	1
PORTUGAL	0	1	1
Total of outbreak points	1019	1365	2384

Nas experiências realizadas, que a seguir se enumeram, foram considerados os grupos de pontos de surtos com as características indicadas.

1. **Aves do tipo *poultry* e do tipo *wild birds*.**
2. **Aves do tipo *poultry*.**
3. **Aves do tipo *wild birds*.**
4. **Aves do tipo *poultry* e do tipo *wild birds* - O conjunto de pontos de surto considerados nesta experiência cumprem o seguinte critério: a área relativa a cada ponto de surto considerado, interceta pelo menos, uma área relativa a**

Tabela 5.2: Posicionamento relativo das áreas concêntricas em pontos de surto.

		Intersect			Disjoint
		poultry	wild_birds	poultry and wild_birds	
1	poultry	585	12	22	400
	wild_birds	18	565	5	777
3	poultry	727	29	56	207
	wild_birds	33	748	22	562
5	poultry	691	35	146	147
	wild_birds	44	826	57	438
8	poultry	619	44	255	101
	wild_birds	48	908	102	307
12	poultry	529	53	374	63
	wild_birds	52	933	195	185

um ponto de surto do outro tipo de ave, i.e., uma área relativa a um surto do tipo *wild birds* será considerada, se interceder pelo menos, uma área relativa a um ponto de surto do tipo *poultry*.

Nas experiências desenvolvidas, foram selecionados recursos espaciais relacionados com os objetivos da descoberta de conhecimento. Em seguida, apresentam-se os tipos de recursos espaciais considerados e as respectivas características:

- *roads* - Tipo de recurso espacial que inclui qualquer estrada, rota, via ou passagem terrestre ligando um local a outro, podendo ser utilizada por veículos e pessoas.
- *landusages* - Este tipo de recurso espacial identifica, entre outras, áreas associadas a atividade humana e áreas de utilização primária da terra. Citam-se como exemplos, áreas de produção agrícola, pastagens ou florestas.
- *waterareas* - Tipo de recurso espacial que identifica áreas de ambientes aquosos.
- *waterways* - Este tipo de recurso espacial inclui cursos de água, rios, ribeiros ou canais.

A justificção para a seleção dos tipos de recursos espaciais, *roads*, *landusages*, *waterareas* e *waterways*, apresenta-se a seguir.

- *roads* - As vias de comunicação utilizadas por pessoas e veículos podem ser um fator de disseminação do vírus.

- *landusages* - As explorações de aves podem estar situadas em explorações agrícolas e, no caso das aves estarem contaminadas, a exploração agrícola pode ser identificada como um foco de infecção. As florestas são o principal habitat natural das aves do tipo *wild birds*.
- *waterareas* e *waterways* - As camadas geográficas *waterareas* e *waterways* assumem especial relevância devido à presença, em ambas, do elemento água. A água tem um papel importante na transmissão do vírus (cf., secção 5.1).

A sequência de valores a atribuir ao *buffer* espacial, tem como objetivo possibilitar a comparação de padrões, extraídos nas diferentes áreas consideradas. Ao mesmo tempo, fornece uma noção de distância, destes padrões, no caso de incluírem relações de distância, aos pontos geográficos que localizam surtos. Em todas as experiências, foram utilizados os valores de 1 km, 3 km, 5 km, 8 km e 12 km, para *buffer* espacial.

As geo-ontologias a apresentar, ao longo das experiências realizadas, contêm, quando for o caso, axiomas do tipo 1 e do tipo 2 (cf., secção 4.1.7). No início de cada ontologia são representados os axiomas do tipo 1, na forma de lista. A lista é formada por tuplos. Cada tuplo contém o nome do predicado e um valor que indica se, o predicado será considerado, valor 1, ou não, valor 0, na extração de regras de associação espacial. Os tuplos representados na lista 5.1 são exemplos de indivíduos que verificam o axioma do tipo 1.

$$[(\text{'contains_landusages'}, 0), (\text{'close_landusages'}, 1)] \quad (5.1)$$

Na lista 5.1, o predicado *close_landusages* será considerado, enquanto que, o predicado *contains_landusages*, não será considerado no processo de extração de regras de associação espacial. A seguir, na geo-ontologia, apresentam-se os axiomas de tipo 2. Cada axioma do tipo 2 tem a forma do *template* mostrado no tuplo 5.2.

$$((\text{valores de suporte}), \text{antecedente}, \text{consequente}, 0|1) \quad (5.2)$$

O primeiro elemento, 'valores de *suporte*', contém um ou dois valores, relativos ao *suporte* da dependência geográfica identificada. Se contiver um valor, esse valor indica o *suporte* da dependência geográfica (cf., secção 2.3.2). Se contiver dois valores, significa que existe uma dependência nos dois sentidos, entre *antecedente* e *consequente*, e esses valores de *suporte* serão iguais. No último elemento do tuplo existe a informação para a dependência ser considerada, valor 1, ou não, valor 0,

no processo de extração de regras de associação espacial. O tuplo 5.3, representa um individuo que verifica um axioma do tipo 2.

$$((0.977), 'contains_waterareas', 'close_roads', 1) \quad (5.3)$$

O tuplo 5.3, descreve uma dependência geográfica, $contains_waterareas \Rightarrow close_roads$, com *suporte* igual a 0.977, *confiança* igual a 100% e será considerado no processo de extração de regras.

O número máximo de regras a obter foi escolhido com base numa quantidade de informação aceitável para análise de um utilizador.

Os valores de salto de *suporte* e de *confiança* foram selecionados com base em testes realizados, tendo-se considerado que seriam aceitáveis para captar variações no número de regras extraídas.

O valor mínimo de *lift* empregue garante que as regras de associação extraídas têm um *consequente* com maior probabilidade de ocorrer no conteúdo da regra do que sem a presença do *antecedente*.

Tabela 5.3: Configuração da função filtro do processo de extração.

Extração de padrões frequentes	Mínimo comprimento do <i>itemset</i>	<i>None</i> ¹
	Máximo comprimento do <i>itemset</i>	<i>None</i>
Regras de associação	Mínimo comprimento do <i>antecedente</i>	<i>None</i>
	Máximo comprimento da <i>antecedente</i>	<i>None</i>
	Mínimo comprimento do <i>consequente</i>	<i>None</i>
	Máximo comprimento do <i>consequente</i>	<i>None</i>
	<i>Lift</i> mínimo	1
Algoritmo de pesquisa	Número máximo de regras a extrair	10
	<i>step_support</i>	0.1
	<i>step_confidence</i>	0.1

Nas experiências realizadas foram utilizados para *Parameters of mining* (cf., secção 3.4), os valores indicados na tabela 5.3, exceto se outra indicação existir. Estes conjunto de parâmetros configuram a função filtro do processo de extração de conhecimento (cf., secção 4.1.8). A metodologia empregue em cada experiência integra um processo interativo que se descreve nos passos a seguir indicados:

1. Indicação dos valores utilizados em *Parameters of knowledge discovery* (cf., secção 4.1.3), que definem as linhas e as colunas do *dataset*.

¹O valor *None* é uma máscara para qualquer comprimento.

2. Na primeira fase, de cada experiência, são extraídas regras de associação, obtidas a partir do *dataset*, sem integração de conhecimento presente na geo-ontologia.
3. Análise dos resultados obtidos, no passo anterior, e seleção na geo-ontologia, das dependências geográficas consideradas convenientes, de acordo com os objetivos da descoberta. Neste passo, o utilizador pode incluir, na geo-ontologia, dependências resultantes da análise das regras obtidas ou de conhecimento do domínio.
4. Na segunda fase da experiência, procede-se à extração de regras de associação considerando as dependências selecionadas no passo anterior.
5. Análise dos padrões espaciais extraídos. Caso os resultados não satisfaçam os objetivos da descoberta, o processo pode ser repetido a partir do ponto 3.

Os resultados produzidos serão colocados em apêndices.

5.3.1 Experiência 1

Nesta experiência foram selecionadas 2384 ocorrências de surtos em aves dos tipos *poultry* e *wild birds* (cf., tabela 5.1). Os valores utilizados para *Parameters of knowledge discovery* são referidos na tabela 5.4.

Tabela 5.4: Configuração de linhas e colunas da matriz.

Linhas	Ocorrências	{ 'global': ([], [], [], []), 'habitat': (('poultry', None, None), ('wild_birds', None, None)) }
	Buffer	[1, 3, 5, 8, 12] Km
Colunas	Recursos espaciais	[('landusages',), ('roads',), ('waterareas',), ('waterways',)]
	Operadores	(cf., secção 4.1.4)

Na primeira fase desta experiência, foram obtidas regras de associação com predominância de predicados espaciais referentes às camadas indicadas na tabela 5.5 (cf., tabelas C.1, C.2).

Tabela 5.5: Camadas relacionadas com padrões espaciais relevantes extraídos na primeira fase da experiência 1.

Buffer [km]	Camadas	(suporte, confiança)
1, 3, 5, 8, 12	roads, landusages	(> 91%, ≈ 100%)

No passo seguinte, foram selecionadas na geo-ontologia, as dependências geográficas que verificam o axioma 2 (cf., secção 3.5.2), de acordo com os objetivos da descoberta. Os resultados obtidos, para os valores de *buffer*, 1 km, 3 km e 5 km, não foram satisfatórios tendo em conta os objetivos da descoberta. Devido a este fato e considerando a análise das regras obtidas, foram inseridos na geo-ontologia, os axiomas do tipo 1 (cf., secção 3.5.2) marcados a estilo **bold** na tabela C.3. Em seguida, considerando as geo-ontologias mostradas nas tabelas C.3 e C.4, foram extraídas regras de associação com relevância em predicados espaciais obtidos de camadas relacionadas com os objetivos da descoberta de conhecimento, indicadas na tabela 5.6 (cf., tabelas C.5, C.5).

Tabela 5.6: Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 1.

Buffer [km]	Camadas	(suporte, confiança)
1	<i>waterways, waterareas</i>	(> 50%, ≈ 100%)
3	<i>waterways, waterareas</i>	(> 80%, ≈ 100%)
5, 8, 12	<i>waterways, waterareas</i>	(> 90%, 100%)

Os resultados e geo-ontologias relativas a esta experiência podem ser consultados no apêndice C.

5.3.2 Experiência 2

Na experiência 2 foram consideradas as 1019 ocorrências relativas a aves do tipo *poultry* (cf., tabela 5.1). Os valores utilizados para *Parameters of knowledge discovery* indicam-se na tabela 5.7.

Tabela 5.7: Valores de configuração de linhas e colunas da matriz.

<i>Linhas</i>	Ocorrências	{ 'global': ([], [], [], []), 'habitat': (('poultry', None, None),) }
	Buffer	[1, 3, 5, 8, 12]
<i>Colunas</i>	Recursos espaciais	[('landusages',), ('roads',), ('waterareas',), ('waterways',)]
	Operadores	(cf., secção 4.1.4)

Os resultados obtidos na primeira fase da experiência são semelhantes aos resultados obtidos, na mesma fase da experiência 1 (cf., tabela D.1), para os valores de *buffer* 1 km e 3 km. Nos restantes valores de *buffer* não foi possível ao algoritmo devolver regras de associação. Seguindo um processo de seleção e inclusão de dependências geográficas, semelhante ao realizado na experiência 1, foram consideradas as dependências indicadas nas geo-ontologias mostradas na tabela D.2.

Na segunda fase da experiência foram obtidas regras de associação com relevância de padrões espaciais relacionados com as camadas *waterways* e *waterareas*, como se indica na tabela 5.8.

Tabela 5.8: Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 2.

Buffer [km]	Camadas	(suporte, confiança)
1	<i>waterways, waterareas</i>	(> 33%, ≈ 100%)
3	<i>waterways, waterareas</i>	(> 74%, ≈ 100%)
5, 8, 12	<i>waterways, waterareas</i>	(> ≈ 90%, 100%)

No apêndice D apresentam-se as geo-ontologias e os resultados desta experiência.

5.3.3 Experiência 3

Nesta experiência foram considerados os pontos de surto relativos a aves do tipo *wild birds*. Na base de dados estão registadas 1365 ocorrências (cf., tabela 5.1). Os valores utilizados para *Parameters of knowledge discovery* indicam-se na tabela 5.9.

Tabela 5.9: Valores de configuração de linhas e colunas da matriz.

<i>Linhas</i>	Ocorrências	{ 'global': ([],[], [], []), 'habitat': (('wild_birds', None, None),) }
	Buffer	[1, 3, 5, 8, 12]
<i>Colunas</i>	Recursos espaciais	[('landusages'), ('roads'), ('waterareas'), ('waterways'),]
	Operadores	(cf., secção 4.1.4)

Na primeira fase da experiência obtiveram-se resultados semelhantes aos obtidos nas experiências anteriores (cf., tabelas E.1, E.2). Continuando a seguir um processo iterativo, na utilização da plataforma, foram selecionadas e incluídas dependências geográficas nas geo-ontologias, como se mostra nas tabelas E.3 e E.4. Também com este subconjunto de ocorrências de surtos de gripe aviária, foram extraídos padrões espaciais relevantes relacionados com as camadas *waterareas* e *waterways*, como se mostra na tabela 5.10.

Tabela 5.10: Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 3.

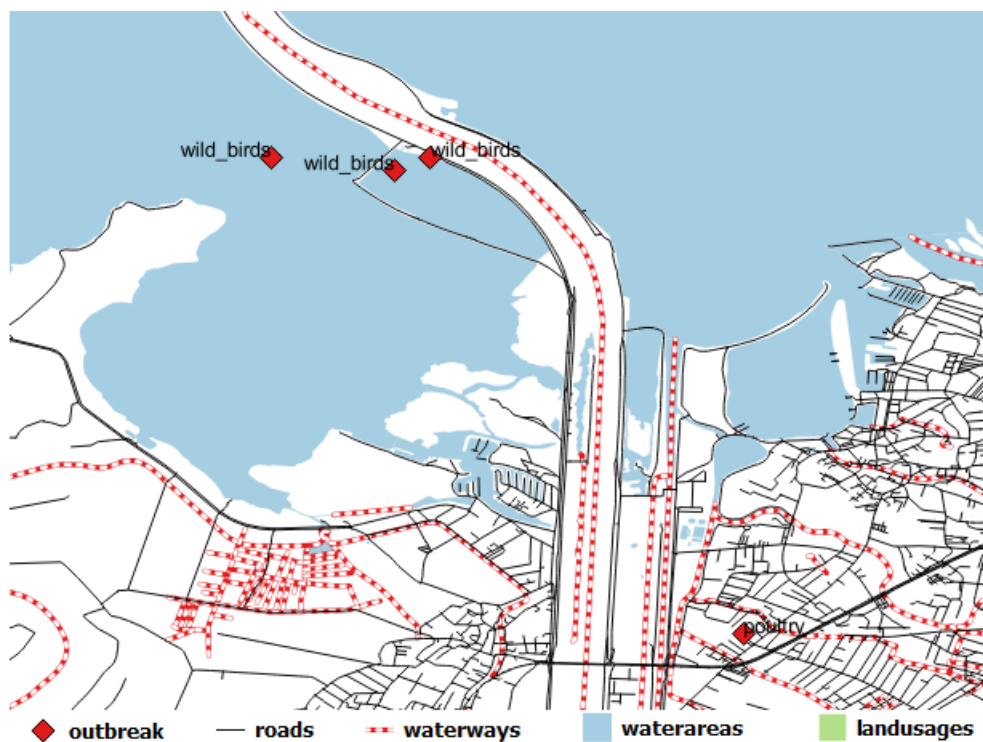
Buffer [km]	Camadas	(suporte, confiança)
1	<i>waterways, waterareas</i>	(> 70%, > ≈ 98%)
3	<i>waterways, waterareas</i>	(> 80%, ≈ 100%)
5, 8, 12	<i>waterways, waterareas</i>	(> 95%, 100%)

As geo-ontologias e os resultados obtidos nesta experiência são mostrados no apêndice E.

5.3.4 Experiência 4

Nesta experiência foram selecionados pontos de surto onde a área concêntrica relativa a cada ponto de surto interceta a área concêntrica relativa a um ponto de surto do outro tipo de ave. A figura 5.1 ilustra um exemplo de um surto localizado num ponto geográfico nestas condições. A figura 5.1 mostra quatro pontos de localização de surtos de tipos de aves diferentes. O fato de se visualizar todo o mapeamento geográfico, entre pontos de surto, significa que as áreas concêntricas, em cada ponto, não são disjuntas (cf., secção 4.2.2).

Figura 5.1: Áreas não disjuntas concêntricas em pontos de surto.



O número de ocorrências, segundo o valor atribuído ao *buffer* espacial variável, mostra-se na tabela 5.11 (cf., tabela 5.2). Na tabela 5.9, apresentam-se os valores utilizados para *Parameters of knowledge discovery*.

Na primeira fase da experiência, não foi possível ao algoritmo de pesquisa extrair regras de associação. Depois de selecionadas as dependências geográficas indicadas nas geo-ontologias, mostradas na tabela F.2, foram extraídas regras de associação espacial que contêm uma forte predominância de padrões espaciais

Tabela 5.11: Número de ocorrências utilizado na experiência 4.

Buffer (Km)	Poultry	Wild_birds	Total
1	34	23	57
3	85	55	140
5	181	101	282
8	299	150	449
12	427	247	674

Tabela 5.12: Valores de configuração de linhas e colunas da matriz.

<i>Linhas</i>	Ocorrências	{ 'global': ([] [], [], []), 'habitat': (('poultry', None, True), ('wild_birds', None, True)) }
	Buffer	[1, 3, 5, 8, 12]
<i>Colunas</i>	Recursos espaciais	[('landusages'), ('roads'), ('waterareas'), ('waterways')]
	Operadores	(cf., secção 4.1.4)

relacionados com as camadas *waterareas* e *waterways*, como se mostra na tabela 5.13.

Tabela 5.13: Camadas relacionadas com padrões espaciais relevantes extraídos na segunda fase da experiência 4.

Buffer [km]	Camadas	(suporte, confiança)
1	<i>waterways, waterareas</i>	(> 56%, 100%)
3	<i>waterways, waterareas</i>	(> 90%, > 91%)
5, 8, 12	<i>waterways, waterareas</i>	(> 93%, 100%)

No apêndice F, apresentam-se as geo-ontologias utilizadas e os resultados obtidos nesta experiência.

5.4 Validação dos resultados

Os resultados obtidos estão alinhados com os fatores de disseminação do vírus apresentados no estudo realizado sobre a gripe aviária (cf., secção 5.1). Em todas as experiências, foram obtidas regras de associação constituídas por padrões espaciais relevantes relacionados com as camadas espaciais selecionadas, *roads*, *landusages*, *waterareas* e *waterways*.

A partir do estudo realizado sobre a gripe aviária foram identificados os principais fatores de disseminação do vírus. Esses fatores apresentam-se a seguir, de forma resumida.

- O contacto de aves infetadas com aves não infetadas. As aves do tipo *wild birds* são o repositório natural do vírus embora as aves do tipo *poultry* possam também ser transmissoras do vírus.
- A água é um elemento natural que permite a transmissão e a sobrevivência do vírus.
- As explorações infetadas são um meio de disseminação do vírus por via das *fomites*, e.g., roupa de pessoas, movimento de pessoas, circulação de veículos que visitaram estas explorações, ...

As camadas geográficas constituídas pelo elemento água, e.g., *waterareas* e *waterways*, dão origem a padrões espaciais relevantes, nas regras de associação extraídas, em áreas concêntricas, em pontos geográficos, onde foram localizados surtos. Nas experiências realizadas obtiveram-se regras de associação constituídas por padrões espaciais obtidos das camadas *waterareas* e *waterways*, com um valor de *suporte* superior a 99%. Os padrões espaciais relevantes, relacionados com estas camadas, constituem regras de associação espacial de valor de *suporte* mais elevado, para valores de medida de *buffer* superiores a 1 km. Este resultado não contradiz a informação obtida sobre determinação de perímetros de monitorização e controlo, no estudo realizado sobre a gripe aviária (cf., secção 5.1). Os valores de medida de *buffer* utilizados devem merecer a análise de um perito.

As regras de associação obtidas em cada *dataset* considerado, mostram um padrão de variação semelhante em função do valor de medida de *buffer* imposto. Em todas as experiências, o valor de *suporte* das regras de associação extraídas, relacionadas com as camadas *waterareas* e *waterways*, vai aumentando à medida que aumenta o valor do *buffer* espacial considerado.

Nas experiências realizadas constatou-se que o *suporte* das regras de associação que contêm padrões espaciais relacionados com as camadas *waterareas* e *waterways*, tem o valor mais baixo em pontos identificados por aves do tipo *poultry* para o valor de *buffer* de 1 km. Por outro lado, para o mesmo valor de *buffer*, considerando as áreas relativas a aves do tipo *wild birds*, o valor de *suporte* das regras obtidas, contendo padrões espaciais relativos às mesmas camadas, é superior ao dobro do valor obtido para aves do tipo *poultry*. Esta constatação aparenta estar alinhada com o fato das aves do tipo *wild birds* serem aves aquáticas.

A experiência 4 considera áreas concêntricas, em pontos de surto, que intercetam, pelo menos, uma área relativa, a um ponto de surto do outro tipo de aves. A análise dos resultados obtidos, nesta experiência, mostra que as regras de associação

extraídas que envolvem padrões espaciais relacionados com as camadas *waterraes* e *waterways* possuem o valor de *suporte* máximo mais elevado (99.3 %), em comparação com o valor máximo de *suporte* obtido, nas restantes experiências, para as regras que contêm padrões relacionados com as mesmas camadas. No estudo realizado sobre a gripe aviária, foi recolhida a informação de que, as aves do tipo *wild birds* são o reservatório natural do vírus e perpetuam a transmissão às aves do tipo *poultry* através, mas não exclusivamente, da partilha de ambientes de água, podendo também as aves do tipo *poultry* serem transmissoras bidirecionais do vírus.

A análise da tabela de interseções de áreas relativas a pontos de surto, em função do raio de *buffer* considerado (cf., tabela 5.11), permite concluir que os pontos geográficos relativos a surtos de cada tipo de aves estão mais próximos entre si do que aos pontos de surto relativos ao outro tipo de aves. Como exemplo, considerando o valor de *buffer* de 12 km, observa-se que as áreas relativas a aves do tipo *wild birds* intercetam-se em 1128 ocorrências relativas ao seu tipo de aves e com 247 áreas relativas a ocorrências do tipo de ave *poultry*.

Os padrões espaciais relacionados com as camadas *roads* e *landusages* constituem regras de associação com valor de *suporte* acima de 90% e valor de *confiança* de 100% ou muito próximo. O fato da camada *roads* estar relacionada com padrões espaciais relevantes que constituem regras de associação extraídas, em áreas onde ocorreram surtos de gripe aviária, mostra que os pontos de surto localizam-se em áreas onde existem vias de circulação de pessoas e veículos. A camada *landusages* está também relacionada com padrões relevantes que constituem regras de associação extraídas em áreas onde ocorreram surtos, o que mostra que os pontos de surto estão localizados em áreas onde existe atividade humana, *e.g.*, exploração agrícola, pastagens, ou onde se desenvolve um uso primário da terra, *e.g.*, floresta.

Não foi possível obter as condições das aves, aquando da captura dos pontos geográficos, *i.e.*, se se tratavam de aves mortas ou simplesmente detetado o surto.

A validação de resultados efetuada teve como base o estudo realizado (cf., secção 5.1), sobre o problema da gripe aviária. O estudo teve como base documentação específica do problema da gripe aviária constituída por artigos de investigação, relatórios produzidos para apoiar a produção de orientações governamentais e relatórios da Organização Mundial de Saúde.

6

Conclusões e trabalho futuro

Neste trabalho, desenvolveu-se uma plataforma de descoberta de conhecimento, na forma de regras de associação espacial, em áreas onde ocorreram fenómenos. As áreas consideradas, na extração de regras de associação espacial, são constituídas pelas geometrias que caracterizam o fenómeno adicionadas por um *buffer* espacial variável.

A análise do objetivo da plataforma permitiu isolar, no processo global, um conjunto de operações independentes do domínio de aplicação. Esta definição no processo levou à criação de dois blocos de operações designadas por *Domain Dependent* e *Domain Independent*. Desta forma, construiu-se uma plataforma dotada de um arquitetura aberta, no sentido em que, é capaz de extrair conhecimento relativo a diferentes tipos de fenómenos. Um tipo de fenómeno é caracterizado espacialmente por um ponto ou por um polígono. A adaptação da plataforma para extração de conhecimento relativo a um tipo de fenómeno envolve reduzi-das alterações no *Domain Dependent*.

No mundo geográfico, as entidades geográficas possuem relações naturais entre si. A ocorrência de um fenómeno pode ter uma relação com o contexto geográfico onde acontece. Esta possibilidade levou à adoção de um *buffer* espacial variável que ‘contorna’ a geometria que define o fenómeno. Desta forma, torna-se possível capturar as relações naturais intrínsecas entre elementos geográficos que possam existir nas imediações da localização geográfica do fenómeno.

O modelo de dados espacial implementado permite extração de conhecimento

considerando níveis de granularidade nas camadas geográficas selecionadas.

O *OpenStreetMap* é a fonte de dados geográficos utilizada, o que confere a esta plataforma, uma abrangência mundial, quanto à possibilidade de extração de conhecimento em zonas onde ocorreram fenômenos. Esta fonte tem a vantagem de ser gratuita, contudo, pelo fato de ser um projeto *Volunteered Geographic Information* pode apresentar algumas imprecisões ou omissões.

A extensão geográfica necessária para alguns processos de extração de conhecimento pode tomar a dimensão de um continente, fato que praticamente impossibilita que o processo decorra num computador pessoal. O processo de conversão de dados do modelo OSM (cf., secção 4.2.1), para uma base de dados relacional, é um processo lento. O processo desenvolvido, neste trabalho, restringe a extração de dados geográficos às áreas que irão ser objeto de análise (cf. secção 4.2.2).

O algoritmo de pesquisa desenvolvido explora o espaço de soluções focando a procura e limitando o número de regras de associação espacial a apresentar ao utilizador, facilitando a análise das regras obtidas.

O algoritmo de extração de padrões frequentes utilizado nesta plataforma não tem fase de geração de candidatos, o que diminui consideravelmente o tempo de geração das regras de associação.

O elevado número de regras de associação gerado, num processo de extração de conhecimento, é um problema porque dificulta a análise de resultados pelo utilizador. Para atenuar este problema, nesta plataforma, é utilizada uma geo-ontologia constituída, por dependências geográficas, que foram identificadas no *dataset* construído (cf., secção 4.1.1). A geo-ontologia permite eliminar regras de associação geradas com as dependências geográficas identificadas, retirando regras triviais ou resultados já conhecidos, do resultado da extração. Devido a este processo, a geo-ontologia permite tornar visível conhecimento que estava oculto devido ao método de pesquisa. A geo-ontologia permite também, ao utilizador, incluir conhecimento do domínio ou ser utilizada de forma interativa com base nos resultados obtidos.

A plataforma de descoberta de conhecimento desenvolvida adequa-se a uma descoberta interativa de conhecimento.

O estudo de caso realizado consistiu na utilização da plataforma desenvolvida, na extração de conhecimento sobre as camadas geográficas *roads*, *landusages*, *waterareas* e *waterways*, em zonas onde ocorreram surtos de gripe aviária. Os resultados obtidos, em quatro experiências realizadas, mostraram-se alinhados com os

principais fatores de disseminação deste tipo de vírus. Estes fatores foram identificados num estudo desenvolvido sobre o problema da gripe aviária, baseado em documentação específica.

O modelo utilizado, nesta plataforma, é um modelo de aprendizagem não supervisionado (cf., secção 2.12). Neste tipo de modelo, ao contrário do modelo supervisionado, não existem respostas corretas. A validação de um modelo de aprendizagem não supervisionada faz-se recorrendo à confrontação dos resultados obtidos com documentação específica (cf., secção 5.1), ou a um perito que avalie a relevância desses resultados.

Durante o processo de desenvolvimento, desta plataforma, e do conhecimento obtido, neste percurso, surgiram ideias de melhoria que a seguir se descrevem.

- A avaliação é uma etapa essencial num processo de descoberta de conhecimento. Este fato gera a necessidade de desenvolvimento de uma abordagem de avaliação automatizada que seja baseada num conjunto de critérios aceites pelos intervenientes no processo. A avaliação da qualidade dos padrões obtidos é facilitada se for realizada por uma abordagem visual. Uma forma de representação destes padrões consiste na elaboração de uma rede, onde os nós representam *itens* e as arestas representam associações.
- O presente trabalho percorre o espaço de soluções e devolve, se for possível, um número máximo de regras definido pelo utilizador. A extração de regras de associação para os valores mais baixos de *suporte*, normalmente, excede o número máximo de regras definido pelo utilizador. No entanto, existem situações em que o conhecimento inesperado surge com conjuntos de regras de baixo *suporte* e *confiança* máxima. A deteção dessas 'bolsas' de conhecimento pode conduzir a resultados inesperados e interessantes.
- O operador de distância desenvolvido neste projeto tem um comportamento independente dos objetos operados. No entanto, é admissível desenvolver operadores espaciais de distância que tenham comportamentos diferentes de acordo com os tipos de objetos operados e com possibilidade de serem parametrizados para uma adequada adaptação ao domínio de aplicação.
- O projeto VGI, OSM, é constituído, por informação geográfica, recolhida por voluntários, aos quais não são exigidas qualificações. Este fato pode conduzir a inconsistências na informação produzida. Os resultados produzidos por esta plataforma podem ser acompanhados de um índice que permita aferir sobre a qualidade dos dados geográficos utilizados.

Referências

- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, pp. 207–216). (pp. 8, 9, 11, e 27)
- Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I., et al. (1996). Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1), 307–328. (p. 41)
- Appice, A., Ceci, M., Lanza, A., Lisi, F. A., & Malerba, D. (2003). *Discovery of spatial association rules in geo-referenced census data: A relational mining approach*. (p. 1)
- Asmai, S. A., Roslin, N. I. A., Abdullah, R. W., & Ahmad, S. (2014). Predictive crime mapping model using association rule mining for crime analysis. *Age*, 12, 21. (p. 1)
- Authority, E. F. S. (2017, September). *Avian influenza @ONLINE*. Retrieved from <http://onlinelibrary.wiley.com/doi/10.2903/j.efsa.2017.4991/epdf> (Acedido em 18/10/2017) (pp. 52 e 53)
- Avian Influenza*. (2016, February). The Center for Food Security & Public Health. Retrieved from www.cfsph.iastate.edu/Factsheets/pdfs/highly_pathogenic_avian_influenza.pdf (Acedido em 03/09/2017) (p. 52)
- Bogorny, V., Engel, P. M., & Alavares, L. O. (2008). Enhancing the process of knowledge discovery in geographic databases using geo-ontologies. *IGI Global, New York, USA*. (pp. 13, 14, 16, 17, 18, 19, 21, 27, 28, e 54)
- Boone, S., & Gerba, C. (2007, 3). Significance of fomites in the spread of respiratory and enteric viral disease. *Applied and Environmental Microbiology*, 73(6), 1687–1696. doi: 10.1128/AEM.02051-06 (p. 53)
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0 step-by-step data mining guide*. (p. 6)
- Cockcroft, S. (1997). A taxonomy of spatial data integrity constraints. *GeoInformatica*, 1(4), 327–343. (p. 18)
- CORRÊA, Â. M. J., & Sferra, H. (2003). Conceitos e aplicações de data mining. *Revista de ciência & tecnologia*, 11, 19–34. (p. 5)
- Duan, M. Y., Zang, Y., Pu, P. X., & Shi, L. (2010). Geo-info graph spectrum analysis for representing distance relations in gis. , XXXVIII-4/C7. (p. 13)
- Elements - openstreetmap@ONLINE*. (n.d.). Retrieved from <http://wiki>

- [.openstreetmap.org/wiki/Elements](https://www.openstreetmap.org/wiki/Elements) (Acedido em 10/09/2017) (p. 46)
- Ester, M., Frommelt, A., Kriegel, H.-P., & Sander, J. (2000, Jul 01). Spatial data mining: Database primitives, algorithms and efficient dbms support. *Data Mining and Knowledge Discovery*, 4(2), 193–216. Retrieved from <https://doi.org/10.1023/A:1009843930701> doi: 10.1023/A:1009843930701 (p. 13)
- Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal—The International Journal on Very Large Data Bases*, 3(4), 357–399. (pp. 11 e 12)
- Han, J., & Fu, Y. (1999). Mining multiple-level association rules in large databases. *IEEE Transactions on knowledge and data engineering*, 11(5), 798–805. (pp. 15 e 16)
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier. (p. 10)
- Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1), 53–87. (p. 41)
- Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N., & Topi, H. (2003). A review of software packages for data mining. *The American Statistician*, 57(4), 290–309. (p. 8)
- Imposm3 - imposm3 3.0.0a documentation@ONLINE*. (n.d.). Retrieved from <https://imposm.org/docs/imposm3/latest/> (Acedido em 10/09/2017) (p. 47)
- Inc, O. G. C. (2011, May). *Opengis, implementation standard for geographic information - simple feature access - part 1: Common architecture*. (Version: 1.2.1) (p. 13)
- Jackson, J. (2002). Data mining; a conceptual overview. *Communications of the Association for Information Systems*, 8(1), 19. (p. 5)
- Karasová, V., Krisp, J. M., & Virrantaus, K. (2005). Application of spatial association rules for improvement of a risk model for fire and rescue services. *Proceedings on the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS) Stockholm*, 183–194. (p. 1)
- Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In *Advances in spatial databases* (pp. 47–66). (pp. 14, 16, e 36)
- Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data*

- mining* (Vol. 454). Springer Science & Business Media. (p. 8)
- Markwell, D. D., & Shortridge, K. F. (1982). Possible waterborne transmission and maintenance of influenza viruses in domestic ducks. *Applied and Environmental Microbiology*, 43(1), 110–115. (p. 53)
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT press. (p. 19)
- Monke, J., & Corn, M. L. (2007, March). *Avian Influenza in Poultry and Wild Birds*. Congressional Research Service - Report for Congress. (p. 53)
- Openstreetmap@ONLINE. (n.d.). Retrieved from <https://www.openstreetmap.org/> (Acedido em 04/06/2017) (p. 1)
- Organization, W. H. (2007a, April). *Questions & Answers on potential transmission of avian influenza (H5N1) through water, Sanitation and Hygiene and ways to reduce the risks to human health*. (p. 52)
- Organization, W. H. (2007b, October). *Review of latest available evidence on potential transmission of avian influenza (H5N1) through water and sewage and ways to reduce the risks to human health*. (pp. 51 e 52)
- Osmosispolygon filter file format@ONLINE. (n.d.). Retrieved from http://wiki.openstreetmap.org/wiki/Osmosis/Polygon_Filter_File_Format (Acedido em 10/09/2017) (p. 48)
- Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the Association for Information Science and Technology*, 47(1), 1–4. (p. 7)
- Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1), 217–222. (p. 5)
- Sharkey, K., Bowers, R., Morgan, K., Robinson, S., & Christley, R. (2008). Epidemiological consequences of an incursion of highly pathogenic h5n1 avian influenza into the british poultry flock. *Proceedings of the Royal Society B: Biological Sciences*, 19–28. (p. 53)
- Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13–22. (p. 5)
- Shekhar, S., & Chawla, S. (2003). *Spatial databases: a tour*. prentice hall Upper Saddle River, NJ. (pp. 11 e 12)
- Stallknecht, D., & Brown, J. (2009). Tenacity of avian influenza viruses. *Revue scientifique et technique (International Office of Epizootics)*, 28(1), 59–67. (p. 52)
- Stallknecht, D., Shane, S., Kearney, M., & Zwank, P. (1990). Persistence of avian influenza viruses in water. *Avian diseases*, 406–411. (p. 52)

- Ubuntu manpage: osmconvert - converter of osm files@ONLINE.* (n.d.). Retrieved from <http://manpages.ubuntu.com/manpages/zesty/man1/osmconvert.1.html> (Acedido em 10/09/2017) (p. 48)
- Vinnakota, S., & Lam, N. S. (2006). Socioeconomic inequality of cancer mortality in the united states: a spatial data mining approach. *International Journal of Health Geographics*. Retrieved from <https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-5-9> (Acedido em 17/06/2017) (p. 1)
- Wood, J. P., Choi, Y. W., Chappie, D. J., Rogers, J. V., & Kaye, J. Z. (2010). Environmental persistence of a highly pathogenic avian influenza (h5n1) virus. *Environmental Science & Technology*, 44(19), 7515-7520. Retrieved from <http://dx.doi.org/10.1021/es1016153> (PMID: 20815380) doi: 10.1021/es1016153 (p. 53)



Função do processo de filtragem

```
1
2 class Filter:
3
4     def __init__(self, len_itemset_min, len_itemset_max, len_ant_min,
5                 len_ant_max, len_cons_min, len_cons_max, lift, ss, sc, n ):
6         self.len_itemset_min = len_itemset_min
7         self.len_itemset_max = len_itemset_max
8         self.len_antecedent_min = len_ant_min
9         self.len_antecedent_max = len_ant_max
10        self.len_consequent_min = len_cons_min
11        self.len_consequent_max = len_cons_max
12        self.lift = lift
13        self.step_support = ss
14        self.step_confidence = sc
15        self.max_number_rules = n
16
17    def mining_filter1(self, key, *args ):
18
19        if len( args ) > 1: raise ValueError('More than one value, error.
20            ')
21        k = key
22        v = args[ 0 ] if len( args ) > 0 else None
23        if k == 'itemset':
24            len_itemset = len( v )
25            if not( self.len_itemset_min is None) and ( len_itemset < self.
26                len_itemset_min ):
```

```
24         return False
25     if not (self.len_itemset_max is None) and ( len_itemset > self.
26         len_itemset_max ):
27         return False
28     return True
29 elif k == 'rule':
30     antecedent, consequent = v
31     len_antecedent = len( antecedent )
32     if not (self.len_antecedent_min is None) and ( len_antecedent <
33         self.len_antecedent_min ):
34         return False
35     if not (self.len_antecedent_max is None) and ( len_antecedent >
36         self.len_antecedent_max ):
37         return False
38     len_consequent = len( consequent )
39     if not (self.len_consequent_min is None) and ( len_consequent <
40         self.len_consequent_min ):
41         return False
42     if not (self.len_consequent_max is None) and ( len_consequent >
43         self.len_consequent_max ):
44         return False
45     return True
46 elif k == 'lift':
47     return v > self.lift
48 elif k == 'step_support':
49     return self.step_support
50 elif k == 'step_confidence':
51     return self.step_confidence
52 elif k == 'max_number_rules':
53     return self.max_number_rules
54 else:
55     raise ValueError('Key error.')
```



Ficheiro de configuração da *Framework IMPOSM3*

```
{
  "tables": {
    "landusages": {
      "fields": [{
        "type": "id",
        "name": "osm_id",
        "key": null
      }, {
        "type": "geometry",
        "name": "geometry",
        "key": null
      }, {
        "type": "string",
        "name": "name",
        "key": "name"
      }, {
        "type": "mapping_value",
        "name": "type",
        "key": null
      }
    ],
    "type": "polygon",
    "mapping": {
```

```

    "barrier": ["hedge"],
    "leisure": ["park", "garden", "playground", "golf_course", "
        sports_centre", "pitch", "stadium", "common", "nature_reserve
        "],
    "tourism": ["zoo"],
    "natural": ["wood", "land", "scrub", "wetland", "heath"],
    "place": ["island"],
    "landuse": ["park", "forest", "residential", "retail", "
        commercial", "industrial", "railway", "cemetery", "grass", "
        farmyard", "farm", "farmland", "orchard", "vineyard", "wood",
        "meadow", "village_green", "recreation_ground", "allotments
        ", "quarry"]
    }
},
"waterways": {
    "fields": [{
        "type": "id",
        "name": "osm_id",
        "key": null
    }, {
        "type": "geometry",
        "name": "geometry",
        "key": null
    }, {
        "type": "string",
        "name": "name",
        "key": "name"
    }, {
        "type": "mapping_value",
        "name": "type",
        "key": null
    }
    ],
    "type": "linestring",
    "mapping": {
        "waterway": ["stream", "river", "canal", "drain", "ditch"],
        "barrier": ["ditch"]
    }
},
"roads": {
    "fields": [{
        "type": "id",
        "name": "osm_id",
        "key": null
    }, {

```

```
    "type": "geometry",
    "name": "geometry",
    "key": null
  }, {
    "type": "mapping_value",
    "name": "type",
    "key": null
  }, {
    "type": "string",
    "name": "name",
    "key": "name"
  }, {
    "type": "boolint",
    "name": "tunnel",
    "key": "tunnel"
  }, {
    "type": "boolint",
    "name": "bridge",
    "key": "bridge"
  }, {
    "type": "direction",
    "name": "oneway",
    "key": "oneway"
  }, {
    "type": "string",
    "name": "ref",
    "key": "ref"
  }, {
    "type": "wayzorder",
    "name": "z_order",
    "key": "layer"
  }, {
    "type": "string",
    "name": "access",
    "key": "access"
  }, {
    "type": "string",
    "name": "service",
    "key": "service"
  }, {
    "type": "mapping_key",
    "name": "class",
    "key": null
  }
],
```

```

    "type": "linestring",
    "filters": {
      "exclude_tags": [["area", "yes"]]
    },
    "mapping": {
      "man_made": ["pier", "groyne"],
      "highway": ["motorway", "motorway_link", "trunk", "trunk_link", "
        primary", "primary_link", "secondary", "secondary_link", "
        tertiary", "tertiary_link", "road", "path", "track", "service
        ", "footway", "bridleway", "cycleway", "steps", "pedestrian",
        "living_street", "unclassified", "residential", "raceway"]
    }
  },
  "waterareas": {
    "fields": [{
      "type": "id",
      "name": "osm_id",
      "key": null
    }, {
      "type": "geometry",
      "name": "geometry",
      "key": null
    }, {
      "type": "string",
      "name": "name",
      "key": "name"
    }, {
      "type": "mapping_value",
      "name": "type",
      "key": null
    }
  ],
  "type": "polygon",
  "mapping": {
    "waterway": ["riverbank"],
    "landuse": ["basin", "reservoir"],
    "natural": ["water"]
  }
}
...
}
}

```



Resultados da experiência 1

Tabela C.1: Regras de associação espacial obtidas em zonas de surto do tipo *poultry* e do tipo *wild_birds*.

Regras de associação espacial			
<i>Buffer = 1 km</i>			
minimun support: 0.900 - minimun confidence: 0.994 - discarded rules: 0 (0.0%)			
<i>supp conf lift</i>			
1	0.959	0.999	1.019 ('close_roads',) ->('crosses_roads',)
2	0.945	0.996	1.016 ('overlaps_landusages',) ->('crosses_roads',)
3	0.927	1.000	1.020 ('overlaps_landusages', 'close_roads') ->('crosses_roads',)
4	0.923	0.999	1.019 ('contains_roads',) ->('crosses_roads',)
5	0.922	0.995	1.015 ('close_landusages',) ->('crosses_roads',)
6	0.913	0.999	1.019 ('close_roads', 'contains_roads') ->('crosses_roads',)
7	0.911	0.996	1.016 ('overlaps_landusages', 'close_landusages') ->('crosses_roads',)
8	0.910	0.999	1.019 ('close_roads', 'close_landusages') ->('crosses_roads',)

Tabela C.2: Regras de associação espacial obtidas em zonas de surto do tipo *poultry* e do tipo *wild_birds* (continuação da tabela C.1).

Regras de associação espacial			
Buffer = 3 km			
minimun support: 0.900 - minimun confidence: 0.999 - discarded rules: 0 (0.0%)			
supp conf lift			
1	0.974	1.000	1.015 ('contains_landusages', 'contains_roads') ->('close_roads',)
2	0.974	1.000	1.015 ('contains_landusages', 'crosses_roads', 'contains_roads') ->('close_roads',)
3	0.967	1.000	1.015 ('contains_landusages', 'close_landusages', 'contains_roads') ->('close_roads',)
4	0.966	1.000	1.015 ('contains_landusages', 'close_landusages', 'crosses_roads', 'contains_roads') ->('close_roads',)
5	0.965	1.000	1.015 ('contains_landusages', 'overlaps_landusages', 'contains_roads') ->('close_roads',)
6	0.964	1.000	1.015 ('contains_landusages', 'overlaps_landusages', 'crosses_roads', 'contains_roads') ->('close_roads',)
7	0.961	1.000	1.015 ('contains_landusages', 'overlaps_landusages', 'close_landusages', 'contains_roads') ->('close_roads',)
8	0.960	1.000	1.015 ('contains_landusages', 'overlaps_landusages', 'close_landusages', 'contains_roads', 'crosses_roads') ->('close_roads',)
Buffer = 5 km			
minimun support: 0.988 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
supp conf lift			
1	0.990	1.000	1.008 ('contains_roads',) ->('crosses_roads',)
2	0.989	1.000	1.008 ('contains_landusages', 'contains_roads') ->('crosses_roads',)
Buffer = 8 km			
minimun support: 0.994 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
supp conf lift			
1	0.995	1.000	1.005 ('contains_roads',) ->('crosses_roads',)
2	0.995	1.000	1.005 ('crosses_roads',) ->('contains_roads',)
3	0.994	1.000	1.005 ('contains_landusages',) ->('crosses_roads',)
4	0.994	1.000	1.005 ('contains_landusages',) ->('contains_roads',)
5	0.994	1.000	1.005 ('contains_landusages', 'contains_roads') ->('crosses_roads',)
6	0.994	1.000	1.005 ('contains_landusages',) ->('crosses_roads', 'contains_roads')
7	0.994	1.000	1.005 ('contains_landusages', 'crosses_roads') ->('contains_roads',)
Buffer = 12 km			
minimun support: 0.997 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
supp conf lift			
1	0.997	1.000	1.003 ('contains_roads',) ->('overlaps_landusages',)
2	0.997	1.000	1.003 ('overlaps_landusages',) ->('contains_roads',)
3	0.997	1.000	1.003 ('contains_roads',) ->('crosses_roads', 'overlaps_landusages')
4	0.997	1.000	1.003 ('overlaps_landusages',) ->('crosses_roads', 'contains_roads')
5	0.997	1.000	1.003 ('crosses_roads', 'contains_roads') ->('overlaps_landusages',)
6	0.997	1.000	1.003 ('crosses_roads', 'overlaps_landusages') ->('contains_roads',)
7	0.997	1.000	1.002 ('overlaps_landusages',) ->('crosses_roads',)
8	0.997	1.000	1.002 ('contains_roads',) ->('crosses_roads',)
9	0.997	1.000	1.002 ('overlaps_landusages', 'contains_roads') ->('crosses_roads',)

Tabela C.3: Geo-ontologias utilizadas na extração de regras de associação espacial em zonas de surto do tipo *poultry* e do tipo *wild_birds*.

Geo-ontologias
<i>Buffer = 1 km</i>
[('close_roads', 1), ('crosses_roads', 1), ('close_landusages', 1), ('overlaps_landusages', 1)]
<i>Buffer = 3 km</i>
[('contains_landusages', 1), ('contains_roads', 1), ('close_roads', 1)]
<i>Buffer = 5 km</i>
[('close_roads', 1)]
((0.99,), 'contains_roads', 'crosses_roads', 1)
((0.987,), 'close_roads', 'crosses_roads', 0)
((0.984,), 'close_landusages', 'contains_landusages', 1)
((0.93,), 'contains_waterareas', 'crosses_roads', 0)
((0.911,), 'crosses_waterways', 'crosses_roads', 0)
((0.885,), 'contains_waterways', 'crosses_roads', 0)
((0.848,), 'close_waterareas', 'crosses_roads', 0)
((0.848,), 'close_waterways', 'crosses_roads', 0)
((0.728,), 'overlaps_waterareas', 'crosses_roads', 0)
<i>Buffer = 8 km</i>
[]
((0.995, 0.995), 'contains_roads', 'crosses_roads', 1)
((0.994,), 'contains_landusages', 'contains_roads', 1)
((0.994,), 'contains_landusages', 'crosses_roads', 1)
((0.991,), 'overlaps_landusages', 'contains_roads', 1)
((0.991,), 'overlaps_landusages', 'crosses_roads', 1)
((0.99,), 'close_landusages', 'contains_landusages', 1)
((0.99,), 'close_landusages', 'contains_roads', 1)
((0.99,), 'close_landusages', 'crosses_roads', 1)
((0.99,), 'close_roads', 'contains_landusages', 1)
((0.99,), 'close_roads', 'contains_roads', 1)
((0.99,), 'close_roads', 'crosses_roads', 1)
((0.977,), 'contains_waterareas', 'contains_roads', 0)
((0.977,), 'contains_waterareas', 'crosses_roads', 0)
((0.951,), 'crosses_waterways', 'contains_roads', 0)
((0.951,), 'crosses_waterways', 'crosses_roads', 0)
((0.948,), 'contains_waterways', 'contains_roads', 0)
((0.948,), 'contains_waterways', 'crosses_roads', 0)
((0.919,), 'close_waterareas', 'contains_landusages', 0)
((0.919,), 'close_waterareas', 'contains_roads', 0)
((0.919,), 'close_waterareas', 'crosses_roads', 0)
((0.913,), 'close_waterways', 'contains_landusages', 0)
((0.913,), 'close_waterways', 'contains_roads', 0)
((0.913,), 'close_waterways', 'crosses_roads', 0)
((0.813,), 'overlaps_waterareas', 'contains_landusages', 0)
((0.813,), 'overlaps_waterareas', 'contains_roads', 0)
((0.813,), 'overlaps_waterareas', 'crosses_roads', 0)

Os predicados espaciais em estilo **bold** foram inseridos pelo utilizador.

Tabela C.4: Geo-ontologia utilizada na extração de regras de associação espacial em zonas de surto do tipo *poultry* e do tipo *wild_birds* (continuação da tabela C.3).

Geo-ontologia
<i>Buffer = 12 km</i>
[]
((0.997,)), 'contains_landusages', 'contains_roads', 1)
((0.997,)), 'contains_landusages', 'crosses_roads', 1)
((0.997,)), 'contains_landusages', 'overlaps_landusages', 1)
((0.997,)), 'contains_roads', 'crosses_roads', 1)
((0.997,0.997)), 'overlaps_landusages', 'contains_roads', 1)
((0.997,)), 'overlaps_landusages', 'crosses_roads', 1)
((0.993,)), 'close_roads', 'contains_landusages', 1)
((0.993,)), 'close_roads', 'contains_roads', 1)
((0.993,)), 'close_roads', 'crosses_roads', 1)
((0.993,)), 'close_roads', 'overlaps_landusages', 1)
((0.992,)), 'close_landusages', 'close_roads', 1)
((0.992,)), 'close_landusages', 'contains_landusages', 1)
((0.992,)), 'close_landusages', 'contains_roads', 1)
((0.992,)), 'close_landusages', 'crosses_roads', 1)
((0.992,)), 'close_landusages', 'overlaps_landusages', 1)
((0.992,)), 'contains_waterareas', 'contains_landusages', 0)
((0.992,)), 'contains_waterareas', 'contains_roads', 0)
((0.992,)), 'contains_waterareas', 'crosses_roads', 0)
((0.992,)), 'contains_waterareas', 'overlaps_landusages', 0)
((0.989,)), 'contains_waterways', 'contains_landusages', 0)
((0.989,)), 'contains_waterways', 'contains_roads', 0)
((0.989,)), 'contains_waterways', 'crosses_roads', 0)
((0.989,)), 'contains_waterways', 'overlaps_landusages', 0)
((0.978,)), 'crosses_waterways', 'contains_roads', 0)
((0.978,)), 'crosses_waterways', 'crosses_roads', 0)
((0.978,)), 'crosses_waterways', 'overlaps_landusages', 0)
((0.961,)), 'close_waterareas', 'close_landusages', 0)
((0.961,)), 'close_waterareas', 'close_roads', 0)
((0.961,)), 'close_waterareas', 'contains_landusages', 0)
((0.961,)), 'close_waterareas', 'contains_roads', 0)
((0.961,)), 'close_waterareas', 'contains_waterareas', 0)
((0.961,)), 'close_waterareas', 'crosses_roads', 0)
((0.961,)), 'close_waterareas', 'overlaps_landusages', 0)
((0.948,)), 'close_waterways', 'close_landusages', 0)
((0.948,)), 'close_waterways', 'close_roads', 0)
((0.948,)), 'close_waterways', 'contains_landusages', 0)
((0.948,)), 'close_waterways', 'contains_roads', 0)
((0.948,)), 'close_waterways', 'crosses_roads', 0)
((0.948,)), 'close_waterways', 'overlaps_landusages', 0)
((0.877,)), 'overlaps_waterareas', 'contains_landusages', 0)
((0.877,)), 'overlaps_waterareas', 'contains_roads', 0)
((0.877,)), 'overlaps_waterareas', 'crosses_roads', 0)
((0.877,)), 'overlaps_waterareas', 'overlaps_landusages', 0)

Tabela C.5: Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela C.3.

Regras de associação espacial			
<i>Buffer = 1 km</i>			
minimun support: 0.500 - minimun confidence: 0.900 - discarded rules: 0 (0.0%)			
<i>supp conf lift</i>			
1	0.780	0.988	1.069 ('contains_landusages',) ->('contains_roads',)
2	0.612	0.976	1.056 ('crosses_waterways',) ->('contains_roads',)
3	0.546	0.998	1.079 ('contains_landusages', 'crosses_waterways') ->('contains_roads',)
4	0.511	0.971	1.051 ('close_waterareas',) ->('contains_roads',)
5	0.510	0.986	1.067 ('contains_waterareas',) ->('contains_roads',)
<i>Buffer = 3 km</i>			
minimun support: 0.800 - minimun confidence: 0.998 - discarded rules: 0 (0.0%)			
<i>supp conf lift</i>			
1	0.841	0.999	1.012 ('crosses_waterways',) ->('crosses_roads',)
2	0.834	0.999	1.012 ('contains_waterareas',) ->('crosses_roads',)
3	0.831	0.999	1.013 ('overlaps_landusages', 'crosses_waterways') ->('crosses_roads',)
4	0.831	0.999	1.013 ('close_landusages', 'crosses_waterways') ->('crosses_roads',)
5	0.828	0.999	1.013 ('overlaps_landusages', 'contains_waterareas') ->('crosses_roads',)
6	0.826	0.999	1.013 ('close_landusages', 'contains_waterareas') ->('crosses_roads',)
7	0.825	0.999	1.013 ('overlaps_landusages', 'close_landusages', 'crosses_waterways') ->('crosses_roads',)
8	0.823	0.999	1.013 ('overlaps_landusages', 'close_landusages', 'contains_waterareas') ->('crosses_roads',)
<i>Buffer = 5 km</i>			
minimun support: 0.925 - minimun confidence: 1.000 - discarded rules: 24 (75.0%)			
<i>supp conf lift</i>			
1	0.930	1.000	1.008 ('contains_waterareas',) ->('crosses_roads',)
2	0.928	1.000	1.010 ('contains_landusages', 'contains_waterareas') ->('contains_roads',)
3	0.928	1.000	1.010 ('contains_landusages', 'contains_waterareas', 'crosses_roads') ->('contains_roads',)
4	0.928	1.000	1.008 ('contains_landusages', 'contains_waterareas') ->('crosses_roads',)
5	0.927	1.000	1.008 ('overlaps_landusages', 'contains_waterareas') ->('crosses_roads',)
6	0.926	1.000	1.010 ('overlaps_landusages', 'contains_landusages', 'contains_waterareas') ->('contains_roads',)
7	0.926	1.000	1.010 ('contains_waterareas', 'contains_landusages', 'overlaps_landusages', 'crosses_roads') ->('contains_roads',)
8	0.926	1.000	1.008 ('overlaps_landusages', 'contains_landusages', 'contains_waterareas') ->('crosses_roads',)
<i>Buffer = 8 km</i>			
minimun support: 0.950 - minimun confidence: 1.000 - discarded rules: 271 (98.5%)			
<i>supp conf lift</i>			
1	0.977	1.000	1.005 ('contains_waterareas',) ->('contains_roads',)
2	0.977	1.000	1.005 ('contains_waterareas',) ->('crosses_roads',)
3	0.951	1.000	1.005 ('crosses_waterways',) ->('contains_roads',)
4	0.951	1.000	1.005 ('crosses_waterways',) ->('crosses_roads',)

Tabela C.6: Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas na geo-ontologia mostrada na tabela C.4 (continuação da tabela C.5)

Regras de associação espacial			
<i>Buffer = 12 km</i>			
minimun support: 0.991 - minimun confidence: 1.000 - discarded rules: 360 (97.6%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.992	1.000	1.003 ('contains_waterareas',) ->('contains_landusages',)
2	0.992	1.000	1.003 ('overlaps_landusages', 'contains_waterareas') ->('contains_landusages',)
3	0.992	1.000	1.003 ('contains_waterareas', 'contains_roads') ->('contains_landusages',)
4	0.992	1.000	1.003 ('crosses_roads', 'contains_waterareas') ->('contains_landusages',)
5	0.992	1.000	1.003 ('contains_waterareas',) ->('overlaps_landusages',)
6	0.992	1.000	1.003 ('contains_waterareas',) ->('contains_roads',)
7	0.992	1.000	1.003 ('crosses_roads', 'contains_waterareas') ->('overlaps_landusages',)
8	0.992	1.000	1.003 ('crosses_roads', 'contains_waterareas') ->('contains_roads',)
9	0.992	1.000	1.002 ('contains_waterareas',) ->('crosses_roads',)



Resultados da experiência 2

Tabela D.1: Regras de associação espacial obtidas em zonas de surto do tipo *poultry*.

Regras de associação espacial			
<i>Buffer = 1 km</i>			
minimun support: 0.900 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.982	1.000	1.006 ('close_roads',) ->('crosses_roads',)
2	0.935	1.000	1.006 ('overlaps_landusages', 'close_roads') ->('crosses_roads',)
3	0.913	1.000	1.006 ('close_roads', 'close_landusages') ->('crosses_roads',)
4	0.904	1.000	1.006 ('contains_roads',) ->('crosses_roads',)
5	0.902	1.000	1.006 ('close_roads', 'contains_roads') ->('crosses_roads',)
<i>Buffer = 3 km</i>			
minimun support: 0.988 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.997	1.000	1.001 ('close_roads',) ->('crosses_roads',)
2	0.994	1.000	1.001 ('contains_roads',) ->('crosses_roads',)
3	0.993	1.000	1.001 ('close_roads', 'contains_roads') ->('crosses_roads',)
4	0.990	1.000	1.001 ('contains_landusages',) ->('crosses_roads',)
5	0.989	1.000	1.001 ('contains_landusages', 'close_roads') ->('crosses_roads',)
<i>Buffer = 5 km</i>			
O algoritmo de pesquisa não conseguiu devolver regras.			
<i>Buffer = 8 km</i>			
O algoritmo de pesquisa não conseguiu devolver regras.			
<i>Buffer = 12 km</i>			
O algoritmo de pesquisa não conseguiu devolver regras.			

Tabela D.2: Geo-ontologias utilizadas na extração de regras de associação espacial em zonas de surtos do tipo *poultry*.

Geo-ontologias
<i>Buffer = 1 km</i>
[('close_roads', 1), ('crosses_roads', 1)] ((0.982,), 'close_roads', 'crosses_roads', 0) ((0.904,), 'contains_roads', 'crosses_roads', 0) ((0.683,), 'contains_landusages', 'crosses_roads', 0) ((0.24,), 'contains_waterways', 'overlaps_landusages', 0)
<i>Buffer = 3 km</i>
[('contains_landusages', 1), ('close_roads', 1), ('contains_roads', 1)] ((0.997,), 'close_roads', 'crosses_roads', 0) ((0.994,), 'contains_roads', 'crosses_roads', 0) ((0.99,), 'contains_landusages', 'crosses_roads', 1) ((0.777,), 'crosses_waterways', 'crosses_roads', 0) ((0.761,), 'contains_waterareas', 'crosses_roads', 0) ((0.664,), 'contains_waterways', 'close_roads', 0) ((0.664,), 'contains_waterways', 'contains_roads', 0) ((0.664,), 'contains_waterways', 'crosses_roads', 0) ((0.644,), 'close_waterways', 'crosses_roads', 0) ((0.578,), 'close_waterareas', 'crosses_roads', 0) ((0.461,), 'overlaps_waterareas', 'crosses_roads', 0)
<i>Buffer = 5 km</i>
[('contains_landusages', 1)] ((0.999, 0.999), 'contains_roads', 'close_roads', 1) ((0.999, 0.999), 'contains_roads', 'crosses_roads', 1) ((0.999, 0.999), 'crosses_roads', 'close_roads', 1) ((0.899,), 'contains_waterareas', 'close_roads', 0) ((0.899,), 'contains_waterareas', 'contains_roads', 0) ((0.899,), 'contains_waterareas', 'crosses_roads', 0) ((0.887,), 'crosses_waterways', 'close_roads', 0) ((0.887,), 'crosses_waterways', 'contains_roads', 0) ((0.887,), 'crosses_waterways', 'crosses_roads', 0) ((0.829,), 'contains_waterways', 'close_roads', 0) ((0.829,), 'contains_waterways', 'contains_roads', 0) ((0.829,), 'contains_waterways', 'crosses_roads', 0) ((0.765,), 'close_waterways', 'close_roads', 0) ((0.765,), 'close_waterways', 'contains_roads', 0) ((0.765,), 'close_waterways', 'crosses_roads', 0) ((0.742,), 'close_waterareas', 'close_roads', 0) ((0.742,), 'close_waterareas', 'contains_roads', 0) ((0.742,), 'close_waterareas', 'crosses_roads', 0)
<i>Buffer = 8 km</i>
[('contains_landusages', 1), ('close_landusages', 1), ('contains_roads', 1), ('crosses_roads', 1)] ((0.977,), 'contains_waterareas', 'close_roads', 0) ((0.866,), 'close_waterareas', 'close_roads', 0) ((0.866,), 'close_waterways', 'close_roads', 0)
<i>Buffer = 12 km</i>
[('contains_landusages', 1), ('overlaps_landusages', 1), ('close_landusages', 1), ('contains_roads', 1), ('crosses_roads', 1), ('close_roads', 1)] ((0.948,), 'close_waterareas', 'contains_waterareas', 0) ((0.827,), 'overlaps_waterareas', 'contains_waterareas', 0)

Os predicados espaciais em estilo **bold** foram inseridos pelo utilizador.

Tabela D.3: Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela D.2.

Regras de associação espacial			
<i>Buffer = 1 km</i>			
minimun support: 0.300 - minimun confidence: 0.994 - discarded rules:,0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.468	0.992	1.044 ('crosses_waterways', 'close_landusages') ->('overlaps_landusages',)
2	0.457	0.991	1.044 ('crosses_waterways', 'close_landusages', 'contains_roads') ->('overlaps_landusages',)
3	0.397	0.998	1.104 ('contains_landusages', 'crosses_waterways') ->('contains_roads',)
4	0.394	0.998	1.104 ('contains_landusages', 'overlaps_landusages', 'crosses_waterways') ->('contains_roads',)
5	0.392	0.998	1.104 ('contains_landusages', 'close_landusages', 'crosses_waterways') ->('contains_roads',)
6	0.388	0.997	1.104 ('contains_landusages', 'overlaps_landusages', 'close_landusages', 'crosses_waterways') ->('contains_roads',)
7	0.345	0.997	1.103 ('contains_landusages', 'contains_waterareas') ->('contains_roads',)
8	0.340	0.997	1.103 ('contains_landusages', 'close_landusages', 'contains_waterareas') ->('contains_roads',)
9	0.339	0.997	1.103 ('contains_landusages', 'overlaps_landusages', 'contains_waterareas') ->('contains_roads',)
10	0.334	0.997	1.103 ('contains_landusages', 'overlaps_landusages', 'close_landusages', 'contains_waterareas') ->('contains_roads',)
<i>Buffer = 3 km</i>			
minimun support: 0.700 - minimun confidence: 1.000 - discarded rules: - 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.777	1.000	1.001 ('crosses_waterways',) ->('crosses_roads',)
2	0.764	1.000	1.001 ('overlaps_landusages', 'crosses_waterways') ->('crosses_roads',)
3	0.764	1.000	1.001 ('close_landusages', 'crosses_waterways') ->('crosses_roads',)
4	0.761	1.000	1.001 ('contains_waterareas',) ->('crosses_roads',)
5	0.754	1.000	1.001 ('overlaps_landusages', 'close_landusages', 'crosses_waterways') ->('crosses_roads',)
6	0.753	1.000	1.001 ('overlaps_landusages', 'contains_waterareas') ->('crosses_roads',)
7	0.750	1.000	1.001 ('close_landusages', 'contains_waterareas') ->('crosses_roads',)
8	0.743	1.000	1.001 ('overlaps_landusages', 'close_landusages', 'contains_waterareas') ->('crosses_roads',)
<i>Buffer = 5 km</i>			
minimun support: 0.897 - minimun confidence: 1.000 - discarded rules: 96 (91.4%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.899	1.000	1.001 ('contains_waterareas',) ->('contains_roads',)
2	0.899	1.000	1.001 ('contains_waterareas',) ->('crosses_roads',)
3	0.899	1.000	1.001 ('contains_waterareas',) ->('close_roads',)
4	0.898	1.000	1.001 ('close_landusages', 'contains_waterareas') ->('contains_roads',)
5	0.898	1.000	1.001 ('close_landusages', 'contains_waterareas') ->('crosses_roads',)
6	0.898	1.000	1.001 ('close_landusages', 'contains_waterareas') ->('close_roads',)
7	0.897	1.000	1.001 ('overlaps_landusages', 'contains_waterareas') ->('contains_roads',)
8	0.897	1.000	1.001 ('overlaps_landusages', 'contains_waterareas') ->('crosses_roads',)
9	0.897	1.000	1.001 ('overlaps_landusages', 'contains_waterareas') ->('close_roads',)

Tabela D.4: Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela D.2.

Regras de associação espacial			
<i>Buffer = 8 km</i>			
minimun support: 0.900 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.977	1.000	1.001 ('contains_waterareas',) ->('close_roads',)
2	0.975	1.000	1.001 ('overlaps_landusages', 'contains_waterareas') ->('close_roads',)
3	0.922	1.000	1.001 ('contains_waterareas', 'crosses_waterways') ->('close_roads',)
4	0.921	1.000	1.001 ('overlaps_landusages', 'contains_waterareas', 'crosses_waterways') ->('close_roads',)
<i>Buffer = 12 km</i>			
minimun support: 0.900 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.948	1.000	1.006 ('close_waterareas',) ->('contains_waterareas',)
2	0.938	1.000	1.006 ('crosses_waterways', 'close_waterareas') ->('contains_waterareas',)
3	0.937	1.000	1.006 ('close_waterareas', 'contains_waterways') ->('contains_waterareas',)
4	0.929	1.000	1.006 ('crosses_waterways', 'close_waterareas', 'contains_waterways') ->('contains_waterareas',)



Resultados da experiência 3

Tabela E.1: Regras de associação espacial obtidas em zonas de surto do tipo *wild_birds*.

Regras de associação espacial			
<i>Buffer = 1 km</i>			
minimum support: 0.900 - minimum confidence: 0.997 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.941	0.998	1.029 ('close_roads',) ->('crosses_roads',)
2	0.938	0.998	1.029 ('contains_roads',) ->('crosses_roads',)
3	0.922	0.998	1.029 ('close_roads', 'contains_roads') ->('crosses_roads',)
4	0.920	0.999	1.030 ('overlaps_landusages', 'close_roads') ->('crosses_roads',)
5	0.919	0.999	1.030 ('overlaps_landusages', 'contains_roads') ->('crosses_roads',)
6	0.908	0.998	1.028 ('close_roads', 'close_landusages') ->('crosses_roads',)
7	0.905	0.999	1.030 ('overlaps_landusages', 'close_roads', 'contains_roads') ->('crosses_roads',)
8	0.903	0.999	1.030 ('overlaps_landusages', 'close_roads', 'close_landusages') ->('crosses_roads',)
<i>Buffer = 3 km</i>			
minimum support: 0.900 - minimum confidence: 0.999 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.968	0.999	1.024 ('contains_landusages',) ->('close_roads',)

Tabela E.2: Regras de associação espacial obtidas em zonas de surto do tipo *wild_birds*.

Regras de associação espacial			
<i>Buffer = 5 km</i>			
minimun support: 0.981 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.983	1.000	1.014 ('contains_landusages',) ->('crosses_roads',)
2	0.983	1.000	1.014 ('contains_roads',) ->('crosses_roads',)
3	0.982	1.000	1.014 ('contains_landusages', 'contains_roads') ->('crosses_roads',)
<i>Buffer = 8 km</i>			
minimun support: 0.988 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.990	1.000	1.010 ('contains_roads',) ->('crosses_roads',)
2	0.990	1.000	1.010 ('crosses_roads',) ->('contains_roads',)
3	0.990	1.000	1.010 ('contains_landusages',) ->('crosses_roads',)
4	0.990	1.000	1.010 ('contains_landusages',) ->('contains_roads',)
5	0.990	1.000	1.010 ('contains_landusages', 'contains_roads') ->('crosses_roads',)
6	0.990	1.000	1.010 ('contains_landusages',) ->('crosses_roads', 'contains_roads')
7	0.990	1.000	1.010 ('contains_landusages', 'crosses_roads') ->('contains_roads',)
<i>Buffer = 12 km</i>			
minimun support: 0.995 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
	<i>supp</i>	<i>conf</i>	<i>lift</i>
1	0.996	1.000	1.004 ('contains_roads',) ->('overlaps_landusages',)
2	0.996	1.000	1.004 ('overlaps_landusages',) ->('contains_roads',)
3	0.996	1.000	1.004 ('contains_roads',) ->('crosses_roads', 'overlaps_landusages')
4	0.996	1.000	1.004 ('overlaps_landusages',) ->('crosses_roads', 'contains_roads')
5	0.996	1.000	1.004 ('crosses_roads', 'contains_roads') ->('overlaps_landusages',)
6	0.996	1.000	1.004 ('crosses_roads', 'overlaps_landusages') ->('contains_roads',)
7	0.996	1.000	1.004 ('overlaps_landusages',) ->('crosses_roads',)
8	0.996	1.000	1.004 ('contains_roads',) ->('crosses_roads',)
9	0.996	1.000	1.004 ('overlaps_landusages', 'contains_roads') ->('crosses_roads',)

Tabela E.3: Geo-ontologias utilizadas na extração de regras de associação espacial em zonas de surtos do tipo *wild_birds*.

Geo-ontologias
<i>Buffer = 1 km</i>
[('close_roads', 1), ('crosses_roads', 1), ('contains_roads', 1)]
<i>Buffer = 3 km</i>
[('contains_landusages', 1), ('close_roads', 1), ('overlaps_landusages', 1), ('contains_roads', 1), ('close_landusages', 1)]
<i>Buffer = 5 km</i>
[]
((0.983,), 'contains_landusages', 'crosses_roads', 1)
((0.983,), 'contains_roads', 'crosses_roads', 1)
((0.981,), 'overlaps_landusages', 'crosses_roads', 1)
((0.979,), 'close_roads', 'crosses_roads', 1)
((0.973,), 'close_landusages', 'close_roads', 1)
((0.973,), 'close_landusages', 'contains_landusages', 1)
((0.973,), 'close_landusages', 'crosses_roads', 1)
((0.952,), 'contains_waterareas', 'crosses_roads', 0)
((0.929,), 'crosses_waterways', 'crosses_roads', 0)
((0.927,), 'close_waterareas', 'crosses_roads', 0)
((0.927,), 'contains_waterways', 'crosses_roads', 0)
((0.91,), 'close_waterways', 'crosses_roads', 0)
((0.842,), 'overlaps_waterareas', 'crosses_roads', 0)
<i>Buffer = 8 km</i>
[]
((0.99,), 'contains_landusages', 'contains_roads', 1)
((0.99,), 'contains_landusages', 'crosses_roads', 1)
((0.99, 0.99), 'contains_roads', 'crosses_roads', 1)
((0.986,), 'overlaps_landusages', 'contains_roads', 1)
((0.986,), 'overlaps_landusages', 'crosses_roads', 1)
((0.982,), 'close_landusages', 'contains_landusages', 1)
((0.982,), 'close_landusages', 'contains_roads', 1)
((0.982,), 'close_landusages', 'crosses_roads', 1)
((0.982,), 'close_roads', 'contains_landusages', 1)
((0.982,), 'close_roads', 'contains_roads', 1)
((0.982,), 'close_roads', 'crosses_roads', 1)
((0.977,), 'contains_waterareas', 'contains_roads', 0)
((0.977,), 'contains_waterareas', 'crosses_roads', 0)
((0.974,), 'contains_waterways', 'contains_roads', 0)
((0.974,), 'contains_waterways', 'crosses_roads', 0)
((0.958,), 'close_waterareas', 'contains_landusages', 0)
((0.958,), 'close_waterareas', 'contains_roads', 0)
((0.958,), 'close_waterareas', 'crosses_roads', 0)
((0.957,), 'crosses_waterways', 'contains_roads', 0)
((0.957,), 'crosses_waterways', 'crosses_roads', 0)
((0.948,), 'close_waterways', 'contains_landusages', 0)
((0.948,), 'close_waterways', 'contains_roads', 0)
((0.948,), 'close_waterways', 'crosses_roads', 0)
((0.884,), 'overlaps_waterareas', 'contains_landusages', 0)
((0.884,), 'overlaps_waterareas', 'contains_roads', 0)
((0.884,), 'overlaps_waterareas', 'crosses_roads', 0)

Os predicados espaciais em estilo **bold** foram inseridos pelo utilizador.

Tabela E.4: Geo-ontologia utilizada na extração de regras de associação espacial em zonas de surtos do tipo *wild_birds* (continuação da tabela E.3).

Geo-ontologia
<i>Buffer = 12 km</i>
[]
((0.996,),'contains_roads','crosses_roads',1)
((0.996,0.996),'overlaps_landusages','contains_roads',1)
((0.996,),'overlaps_landusages','crosses_roads',1)
((0.995,),'contains_landusages','contains_roads',1)
((0.995,),'contains_landusages','crosses_roads',1)
((0.995,),'contains_landusages','overlaps_landusages',1)
((0.99,),'contains_waterareas','contains_landusages',0)
((0.99,),'contains_waterareas','contains_roads',0)
((0.99,),'contains_waterareas','crosses_roads',0)
((0.99,),'contains_waterareas','overlaps_landusages',0)
((0.99,),'contains_waterways','contains_landusages',0)
((0.99,),'contains_waterways','contains_roads',0)
((0.99,),'contains_waterways','crosses_roads',0)
((0.99,),'contains_waterways','overlaps_landusages',0)
((0.988,),'close_roads','contains_landusages',1)
((0.988,),'close_roads','contains_roads',1)
((0.988,),'close_roads','crosses_roads',1)
((0.988,),'close_roads','overlaps_landusages',1)
((0.987,),'close_landusages','close_roads',1)
((0.987,),'close_landusages','contains_landusages',1)
((0.987,),'close_landusages','contains_roads',1)
((0.987,),'close_landusages','crosses_roads',1)
((0.987,),'close_landusages','overlaps_landusages',1)
((0.971,),'crosses_waterways','contains_roads',0)
((0.971,),'crosses_waterways','crosses_roads',0)
((0.971,),'crosses_waterways','overlaps_landusages',0)
((0.97,),'close_waterareas','close_landusages',0)
((0.97,),'close_waterareas','close_roads',0)
((0.97,),'close_waterareas','contains_landusages',0)
((0.97,),'close_waterareas','contains_roads',0)
((0.97,),'close_waterareas','contains_waterareas',0)
((0.97,),'close_waterareas','crosses_roads',0)
((0.97,),'close_waterareas','overlaps_landusages',0)
((0.964,),'close_waterways','close_landusages',0)
((0.964,),'close_waterways','close_roads',0)
((0.964,),'close_waterways','contains_landusages',0)
((0.964,),'close_waterways','contains_roads',0)
((0.964,),'close_waterways','contains_waterways',0)
((0.964,),'close_waterways','crosses_roads',0)
((0.964,),'close_waterways','overlaps_landusages',0)
((0.914,),'overlaps_waterareas','contains_landusages',0)
((0.914,),'overlaps_waterareas','contains_roads',0)
((0.914,),'overlaps_waterareas','crosses_roads',0)
((0.914,),'overlaps_waterareas','overlaps_landusages',0)

Tabela E.5: Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas nas tabelas E.3 e E.4.

Regras de associação espacial	
Buffer = 1 km	
minimun support: 0.700 - minimun confidence: 0.975 - discarded rules: 0 (0.0%)	
<i>supp conf lift</i>	
1	0.714 0.979 1.033 ('close_waterareas',) ->('overlaps_landusages',)
2	0.709 0.981 1.035 ('crosses_waterways',) ->('overlaps_landusages',)
3	0.700 0.996 1.050 ('close_waterareas', 'close_landusages') ->('overlaps_landusages',)
4	0.700 0.982 1.059 ('overlaps_landusages', 'close_waterareas') ->('close_landusages',)
Buffer = 3 km	
minimun support: 0.800 - minimun confidence: 0.998 - discarded rules: 0 (0.0%)	
<i>supp conf lift</i>	
1	0.890 0.999 1.022 ('close_waterareas',) ->('crosses_roads',)
2	0.850 0.999 1.022 ('contains_waterareas', 'close_waterareas') ->('crosses_roads',)
3	0.834 0.999 1.022 ('close_waterways',) ->('crosses_roads',)
4	0.831 0.999 1.022 ('contains_waterareas', 'crosses_waterways') ->('crosses_roads',)
5	0.829 0.999 1.022 ('crosses_waterways', 'close_waterareas') ->('crosses_roads',)
6	0.818 0.999 1.022 ('contains_waterways', 'contains_waterareas') ->('crosses_roads',)
7	0.812 0.999 1.022 ('crosses_waterways', 'close_waterways') ->('crosses_roads',)
8	0.806 0.999 1.022 ('contains_waterways', 'close_waterareas') ->('crosses_roads',)
9	0.801 0.999 1.022 ('contains_waterareas', 'crosses_waterways', 'close_waterareas') ->('crosses_roads',)
Buffer = 5 km	
minimun support: 0.950 - minimun confidence: 1.000 - discarded rules: 100 (96.2%)	
<i>supp conf lift</i>	
1	0.974 1.000 1.017 ('overlaps_landusages', 'close_roads') ->('contains_landusages',)
2	0.973 1.000 1.017 ('overlaps_landusages', 'close_roads', 'contains_roads') ->('contains_landusages',)
3	0.952 1.000 1.014 ('contains_waterareas',) ->('crosses_roads',)
4	0.950 1.000 1.017 ('contains_landusages', 'contains_waterareas') ->('contains_roads',)
Buffer = 8 km	
minimun support: 0.975 - minimun confidence: 1.000 - discarded rules: 139 (98.6%)	
<i>supp conf lift</i>	
1	0.977 1.000 1.010 ('contains_waterareas',) ->('contains_roads',)
2	0.977 1.000 1.010 ('contains_waterareas',) ->('crosses_roads',)
Buffer = 12 km	
minimun support: 0.990 - minimun confidence: 1.000 - discarded rules: 84 (90.3%)	
<i>supp conf lift</i>	
1	0.990 1.000 1.005 ('contains_waterways',) ->('contains_landusages',)
2	0.990 1.000 1.005 ('overlaps_landusages', 'contains_waterways') ->('contains_landusages',)
3	0.990 1.000 1.005 ('contains_waterways', 'contains_roads') ->('contains_landusages',)
4	0.990 1.000 1.005 ('crosses_roads', 'contains_waterways') ->('contains_landusages',)
5	0.990 1.000 1.004 ('contains_waterways',) ->('overlaps_landusages',)
6	0.990 1.000 1.004 ('contains_waterways',) ->('contains_roads',)
7	0.990 1.000 1.004 ('crosses_roads', 'contains_waterways') ->('overlaps_landusages',)
8	0.990 1.000 1.004 ('crosses_roads', 'contains_waterways') ->('contains_roads',)
9	0.990 1.000 1.004 ('contains_waterways',) ->('crosses_roads',)



Resultados da experiência 4

Tabela F.1: Regras de associação espacial obtidas de zonas que intercetam pelo menos uma zona de outro tipo de surto.

Regras de associação espacial
<i>Buffer = 1 km</i>
O algoritmo de pesquisa não conseguiu devolver regras.
<i>Buffer = 3 km</i>
O algoritmo de pesquisa não conseguiu devolver regras.
<i>Buffer = 5 km</i>
O algoritmo de pesquisa não conseguiu devolver regras.
<i>Buffer = 8 km</i>
O algoritmo de pesquisa não conseguiu devolver regras.
<i>Buffer = 12 km</i>
O algoritmo de pesquisa não conseguiu devolver regras.

Tabela F.2: Geo-ontologias utilizadas na extração de regras de associação espacial em zonas que intercetam pelo menos uma zona de outro tipo de surto.

Geo-ontologias
<i>Buffer = 1 km</i>
[('crosses_roads', 1), ('close_roads', 1)] ((0.947,), 'close_landusages', 'overlaps_landusages', 1) ((0.947,), 'contains_roads', 'overlaps_landusages', 1) ((0.86,), 'contains_landusages', 'close_landusages', 1) ((0.86,), 'contains_landusages', 'contains_roads', 1) ((0.86,), 'contains_landusages', 'overlaps_landusages', 1) ((0.579,), 'crosses_waterways', 'close_landusages', 0) ((0.579,), 'crosses_waterways', 'overlaps_landusages', 0) ((0.456,), 'contains_waterareas', 'overlaps_landusages', 0) ((0.368,), 'close_waterareas', 'overlaps_landusages', 0) ((0.351,), 'close_waterways', 'close_landusages', 0) ((0.351,), 'close_waterways', 'crosses_waterways', 0) ((0.351,), 'close_waterways', 'overlaps_landusages', 0) ((0.281,), 'contains_waterways', 'close_landusages', 0) ((0.281,), 'contains_waterways', 'contains_landusages', 0) ((0.281,), 'contains_waterways', 'contains_roads', 0) ((0.281,), 'contains_waterways', 'contains_waterareas', 0) ((0.281,), 'contains_waterways', 'crosses_waterways', 0) ((0.281,), 'contains_waterways', 'overlaps_landusages', 0)
<i>Buffer = 3 km</i>
[('contains_landusages', 1), ('contains_roads', 1), ('crosses_roads', 1), ('close_roads', 1)] ((0.814,), 'contains_waterways', 'overlaps_landusages', 0) ((0.779,), 'close_waterareas', 'close_landusages', 0) ((0.686,), 'overlaps_waterareas', 'crosses_waterways', 0)
<i>Buffer = 5 km</i>
[('contains_roads', 1), ('crosses_roads', 1), ('close_roads', 1)] ((0.996, 0.996), 'contains_landusages', 'close_landusages', 1) ((0.993,), 'overlaps_landusages', 'close_landusages', 1) ((0.993,), 'overlaps_landusages', 'contains_landusages', 1) ((0.94,), 'close_waterareas', 'contains_waterareas', 0) ((0.844,), 'overlaps_waterareas', 'contains_waterareas', 0)
<i>Buffer = 8 km</i>
[('contains_landusages', 1), ('contains_roads', 1), ('crosses_roads', 1), ('close_roads', 1)] ((0.96,), 'close_waterareas', 'contains_waterareas', 0) ((0.902,), 'overlaps_waterareas', 'close_landusages', 0) ((0.902,), 'overlaps_waterareas', 'contains_waterareas', 0)
<i>Buffer = 12 km</i>
[('contains_landusages', 1), ('overlaps_landusages', 1), ('close_landusages', 1), ('contains_roads', 1), ('crosses_roads', 1), ('close_roads', 1)] ((0.993,), 'crosses_waterways', 'contains_waterways', 0) ((0.979,), 'close_waterareas', 'contains_waterareas', 0) ((0.973,), 'close_waterways', 'contains_waterways', 0) ((0.948,), 'overlaps_waterareas', 'contains_waterareas', 0)

Tabela F.3: Regras de associação obtidas depois de consideradas as dependências geográficas selecionadas nas geo-ontologias mostradas na tabela F.2.

Regras de associação espacial			
Buffer = 1 km			
minimun support: 0.500 - minimun confidence: 1.000 - discarded rules: 47 (92.2%)			
<i>supp conf lift</i>			
1	0.579	1.000	1.056 ('crosses_waterways',) ->('close_landusages',)
2	0.579	1.000	1.056 ('crosses_waterways', 'overlaps_landusages') ->('close_landusages',)
3	0.579	1.000	1.018 ('crosses_waterways',) ->('overlaps_landusages',)
4	0.561	1.000	1.056 ('crosses_waterways', 'contains_roads') ->('close_landusages',)
Buffer = 3 km			
minimun support: 0.900 - minimun confidence: 0.900 - discarded rules: 0 (0.0%)			
<i>supp conf lift</i>			
1	0.979	0.993	1.007 ('close_landusages',) ->('overlaps_landusages',)
2	0.979	0.993	1.007 ('overlaps_landusages',) ->('close_landusages',)
3	0.921	0.992	1.007 ('close_landusages', 'crosses_waterways') ->('overlaps_landusages',)
4	0.921	0.992	1.007 ('overlaps_landusages', 'crosses_waterways') ->('close_landusages',)
5	0.921	0.935	1.007 ('close_landusages',) ->('overlaps_landusages', 'crosses_waterways')
6	0.921	0.935	1.007 ('overlaps_landusages',) ->('close_landusages', 'crosses_waterways')
7	0.900	0.992	1.007 ('contains_waterareas',) ->('overlaps_landusages',)
8	0.900	0.992	1.007 ('contains_waterareas',) ->('close_landusages',)
9	0.900	0.913	1.007 ('overlaps_landusages',) ->('contains_waterareas',)
10	0.900	0.913	1.007 ('close_landusages',) ->('contains_waterareas',)
Buffer = 5 km			
minimun support: 0.925 - minimun confidence: 1.000 - discarded rules: 81 (95.3%)			
<i>supp conf lift</i>			
1	0.940	1.000	1.018 ('close_waterareas',) ->('contains_waterareas',)
2	0.936	1.000	1.018 ('contains_landusages', 'close_waterareas') ->('contains_waterareas',)
3	0.936	1.000	1.018 ('close_waterareas', 'close_landusages') ->('contains_waterareas',)
4	0.933	1.000	1.018 ('overlaps_landusages', 'close_waterareas') ->('contains_waterareas',)
Buffer = 8 km			
minimun support: 0.950 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
<i>supp conf lift</i>			
1	0.960	1.000	1.009 ('close_waterareas',) ->('contains_waterareas',)
2	0.958	1.000	1.009 ('overlaps_landusages', 'close_waterareas') ->('contains_waterareas',)
3	0.958	1.000	1.009 ('close_waterareas', 'close_landusages') ->('contains_waterareas',)
4	0.955	1.000	1.009 ('overlaps_landusages', 'close_landusages', 'close_waterareas') ->('contains_waterareas',)
5	0.955	1.000	1.009 ('close_waterareas', 'contains_waterways') ->('contains_waterareas',)
6	0.953	1.000	1.009 ('overlaps_landusages', 'contains_waterways', 'close_waterareas') ->('contains_waterareas',)
7	0.953	1.000	1.009 ('close_landusages', 'close_waterareas', 'contains_waterways') ->('contains_waterareas',)
8	0.951	1.000	1.009 ('contains_waterways', 'overlaps_landusages', 'close_landusages', 'close_waterareas') ->('contains_waterareas',)
9	0.951	1.000	1.009 ('close_waterareas', 'crosses_waterways') ->('contains_waterareas',)
Buffer = 12 km			
minimun support: 0.975 - minimun confidence: 1.000 - discarded rules: 0 (0.0%)			
<i>supp conf lift</i>			
1	0.993	1.000	1.001 ('crosses_waterways',) ->('contains_waterways',)
2	0.987	1.000	1.001 ('contains_waterareas', 'crosses_waterways') ->('contains_waterways',)
3	0.979	1.000	1.006 ('close_waterareas',) ->('contains_waterareas',)
4	0.978	1.000	1.006 ('close_waterareas', 'contains_waterways') ->('contains_waterareas',)

