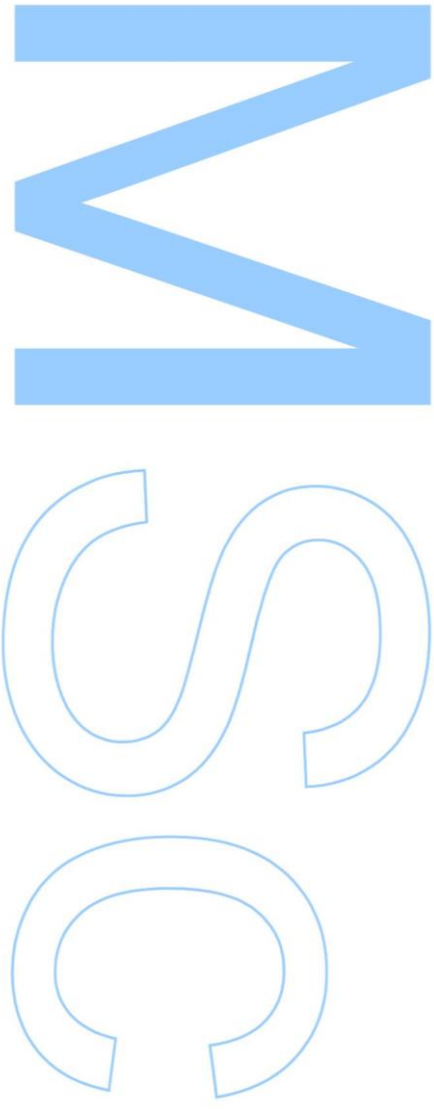
Propositional and Relational Approaches to Spatio-Temporal Data Analysis

Mariana Rafaela Oliveira
Master's degree in Computer Science
Computer Science Department
2015

Supervisor
Luís Torgo, Associate Professor, Faculty of Sciences, University of Porto

Co-supervisor
Vítor Santos Costa, Associate Professor, Faculty of Sciences, University of Porto



All corrections determined by the jury, and only those, were incorporated.

The President of the Jury,

Porto, _____/_____/_____

To Marvin

Acknowledgements

I would like to express my sincere gratitude to my advisors Prof. Luís Torgo and Prof. Vítor Santos Costa for all their guidance, support, patience and time. I am truly thankful for all the knowledge they have imparted to me, and the kindness they demonstrated along the way, even when halfway across the world.

I thank the members of the jury, Prof. Sandra Alves and Prof. Rui Camacho (Faculty of Engineering, University of Porto), for the careful reading of my work. Their corrections and suggestions greatly improved this dissertation.

I would also like to thank Dr. João Torres for all his help with our case study, which motivated this work. Many thanks to Prof. Rita Ribeiro and Paula Branco, MSc, for the helpful discussions on data mining under imbalanced domains. Thanks to Prof. Paulo Azevedo (University of Minho) and Prof. Alípio Jorge for their help with association rule mining.

I am also thankful to my professors, who taught me much about Computer Science. Specially so to Prof. Fernando Silva for also having encouraged me to enrol in this Master's program. Many thanks to my classmates and colleagues that considerably improved my journey in Physics and Computer Science, in particular, to my friend Miguel who accompanied me in both. I would also like to thank the administrative staff, in particular, Ms. Alexandra Ferreira.

Finally, I would like to express my deep gratitude to my family and friends. I am very lucky to have such an incredibly supportive and loving family, and I wouldn't be where I am today without my parents, siblings, aunts and uncle. My friends add so much to my life. Special thanks to Teresa, Pires and Catarina for their unwavering support this year. Last, but not least, I would like to thank Marvin for being such an inspiration, and always keeping me smiling even from miles away.

This work was financed by the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project UID/EEA/50014/2013.

Mariana Oliveira
Porto, 2015

Abstract

Understanding spatio-temporal phenomena is a fundamental challenge in the field of Data Mining with applications ranging over a myriad of domains. One such challenge arises from spatio-temporal databases of time-varying data that can be represented by an evolving thematic map. Several approaches aiming at describing spatio-temporal data and predicting future values have already been proposed. Among them, we find propositional approaches that work on a single table, and relational approaches with the ability to work on multiple related tables. We review both types of approaches to association rule learning and regression problems, and enumerate the challenges faced.

Our motivating application concerns wildfires in Portugal, which every year have a strong socio-economical and environmental impact in the country. We adapt a notion of spatio-temporal neighbourhood to include spatial direction, propose a concept of simplified border for heterogeneous spatial objects, build spatio-temporal indicators based on these notions, design relational predicates that deal with numerical attributes and include the temporal and spatial dimensions, and deploy a re-sampling tecnhique to improve regression under an imbalanced domain. We apply a relational and a propositional approach to the problems of understanding and predicting wildfires in mainland Portugal, and draw comparisons between the two. We are able to find strong association rules and accurately predict the yearly percentage of burnt area in each Portuguese civil parish in spite of the several challenges posed by this problem.

Keywords: spatio-temporal databases, relational data mining, propositional data mining, spatio-temporal association rule learning, spatio-temporal forecasting.

Resumo

A compreensão de fenómenos espácio-temporais é um desafio fundamental na área de Análise de Dados com aplicações numa quantidade enorme de domínios. Um destes desafios advém de bases de dados com dados variantes no tempo que podem ser representados pela evolução de um mapa temático. Diversas abordagens aos problemas de descrever dados espácio-temporais e prever futuros dados já foram propostos. Entre elas, encontram-se as abordagens proposicionais que trabalham sobre uma só tabela, e as abordagens relacionais que são capazes de trabalhar sobre múltiplas tabelas relacionadas entre si. Neste trabalho, revemos ambos os tipos de abordagens à aprendizagem de regras de associação e a problemas de regressão, enumerando os desafios enfrentados.

A nossa motivação vem da aplicação destas técnicas ao problema de fogos florestais em Portugal que todos os anos têm um forte impacto socio-ecónomico e ambiental no país. Adaptamos uma noção de vizinhança espácio-temporal para incluir direcção espacial, propomos um conceito de fronteiras simplificadas, construímos indicadores espácio-temporais baseados nestes conceitos, implementamos predicados relacionais que lidam com atributos numéricos e que incluem as dimensões temporal e espaciais, e empregamos uma metodologia de re-amostragem para melhorar regressão num domínio desequilibrado. Aplicamos abordagens proposicionais e relacionais aos problemas de compreender e prever fogos e fazemos comparações entre as duas abordagens. Fomos capazes de encontrar regras de associação fortes e prever adequadamente a percentagem anual de área ardida em cada freguesia portuguesa, mesmo tendo em conta os desafios postos por esta aplicação.

Palavras-chave: análise de dados espácio-temporais, análise de dados relacional, análise de dados proposicional, aprendizagem de regras de associação espácio-temporais, previsão espácio-temporal.

Contents

List of Tables

List of Figures

List of Acronyms

F_β F-measure 35, 37

SLD_{OI}**-deduction** Selective Linear Definite deduction under Object Identity 21

Aleph A Learning Engine for Proposing Hypotheses 18, 19, 45, 52, 53, 55, 57, 58, 60, 61, 60, 62, 65, 67, 71

ANN Artificial Neural Network 28

AOC Area Over the Curve 36

AR-HMM Auto-Regressive Hidden Markov Model 32

ARES Association Rules Extractor from Spatial data 20

ARIMA Auto-Regressive Integrated Moving Average 30

AUC Area Under the Curve 35

BN Bayesian Network 31

CAREN Class project Association Rule ENgine 50

CIBIO Research Centre in Biodiversity and Genetic Resources Associate Laboratory 39

CSV Comma Separated Values 53, 55, 68

EH-Apriori Extended Hash-based Apriori 13

EMA Exponential Moving Average 47, 48

FITI First Intra Then Inter 13

FN False Negative 34, 35

FP False Positive 34, 35

FP-growth Frequent Pattern Growth 17

FPR False Positive Rate 35

GIS Geographical Information System 1, 16, 20

GSP Generalized Sequential Patterns 11, 17

HMM Hidden Markov Model 32

IDW Inverse Distance Weighting 27, 49

ILP Inductive Logic Programming x, xi, 18, 19, 20, 23, 31, 32, 37, 45, 52, 58, 61, 76, 77

INE National Statistics Institute - Instituto Nacional de Estatística 39

INGENS INductive GEographic iNformation System 20

MAD Mean Absolute Deviation 35

MBR Minimum Bounding Rectangles 14

MRF Markov Random Field 32

MSE Mean Squared Error 4, 25, 35, 69

OLS Ordinary Least Squares 30

PCT Predictive Clustering Tree 31

PPV Positive Predictive Value 35

REC Regression Error Characteristic 36

RECS Regression Error Characteristic Surfaces 36

RF Random Forest 64, 68, 70, 71, 72, 73

ROC Receiver Operating Characteristic 35, 36

RPPI Reduced Prefix-Projected Itemsets 13

RROC Receiver Operating Characteristic for Regression 36

SPADA Spatial PAttern Discovery Algorithm 19, 20

SPADE Sequential PAttern Discovery using Equivalence classes 11, 15

SPC Specifity 35

SQL Structured Query Language 41

STRF Spatio-Temporal Random Field 32

SVM Support Vector Machine 64

SVR Support Vector Regression machine 64, 65, 68, 70, 71, 73

TN True Negative 34, 35

TNR True Negative Rate 35

TP True Positive 34, 35

TPR True Positive Rate 35

YAP Yet Another Prolog 45

List of Code

Chapter 1

Introduction

1.1 Context and problem definition

Humans inhabit an environment that changes in space and time. Understanding these changes is a crucial human endeavour. Geographical Information Systems (GIS) were the first software artifacts to represent and store this type of information. Originally these systems allowed for simple queries on geographical relationships. The next step is to try to understand what relations are important in the database and whether we can predict target variables based on the existing data. This will be the focus of this dissertation.

1.1.1 Spatio-temporal databases

There are two classes of spatio-temporal databases (Mamoulis, 2009). The first one consists of sequences of measurements generated over time by sensors located at fixed points across space (e.g., sensors measuring wind speed at different stations), and time-varying data that can be represented by thematic maps (e.g., land value maps). Our case study concerns wildfires in Portuguese civil parishes, and falls under this category. In this work, we will not investigate the second class of spatio-temporal databases where each instance records a moving object's trajectory (e.g., taxi movements in a city or object movement in a video). The interested reader is referred to Nanni *et al.*

(2008), Mamoulis (2009) or Aggarwal (2015) for an overview of spatio-temporal data mining tasks involving this second class of databases.

1.1.2 Propositional and relational modelling approaches

Traditional data mining methods work on a single table, therefore being categorised as propositional. Most propositional methods assume that each item in the data set has been obtained independently from the others, and that all the objects can be seen as sampled from the same underlying distribution. Multi-relational methods try to obtain further insight by explicitly considering the complex nature of the data. Often, relational approaches are obtained through upgrading, that is, by extending a corresponding propositional approach to be able to work on multiple tables from a relational database, keeping the single-table approach as a special case (see Džeroski (2003)). Malerba (2008) argues that relational approaches are particularly suited to spatial data mining tasks since they can: **i)** deal with heterogeneous spatial objects; **ii)** distinguish between *reference* (the main subject of a task) and *task-relevant* objects (objects spatially related to reference objects); **iii)** naturally represent a wide variety of spatial relationships between objects; and **iv)** accommodate spatial auto-correlation. A similar argument can be made for spatio-temporal data mining tasks, with additional consideration for the temporal dimension.

Throughout this dissertation we **i)** explore how both types of approaches can be applied to the analysis of spatio-temporal databases and **ii)** draw comparisons between them.

1.1.3 Descriptive and predictive data mining tasks

When trying to extract knowledge from a database, we usually have one of two main purposes in mind: **a)** to describe our data by finding human-interpretable patterns in it or **b)** to successfully predict unknown or future values based on a set of explanatory variables. Although the boundary between goals is not always clear, each of them usually demand a set of common data mining methods (Fayyad *et al.* , 1996).

Descriptive data mining methods include (generally unsupervised) learning approaches to anomaly detection, association rule learning and clustering. Predictive data mining methods fall under (generally supervised) regression and classification.

We aim at **i)** investigating methods that will allow us to achieve the two goals, and **ii)** applying the most promising to our case study which, as mentioned above, concerns wildfires in Portugal.

Considering there is such a vast array of methods with each goal in mind, we focus on association rule learning and regression. Association rule learning allows us to reach a human-interpretable understanding of the data in our case study while regression enables the prediction of not only whether a certain parish will be impacted by wildfires but also to what extent.

Since the mid-1990s, there has been an abundance of research dedicated to dealing with both spatial and temporal dimensions separately in a data mining context. Efforts to deal with both dimensions concurrently are fairly recent in comparison; the first workshops organised on the matter first appearing in the mid-2000's (Nanni *et al.* , 2008). We proceed to enumerate some of the challenges facing researchers working on spatio-temporal data mining.

1.1.4 Challenges

Working with both temporal and spatial dimensions presents several problems. The two dimensions have different properties posing a multitude of challenges to integrating and dealing with them in a data mining context as discussed below.

Time is generally considered to be one-dimensional, unidirectional and ordered while space is three-dimensional (although geographical data often just considers two spatial dimensions such as latitude and longitude). Spatial objects can be heterogeneous and have complex geometries. Temporal events can have different durations. Both temporal (e.g., before and after) and spatial metric (e.g. distance) and non-metric (e.g., topological and directional) relationships between spatio-temporal objects are often fuzzy or implicit (Andrienko *et al.* , 2006). Both dimensions can be seen at multiple levels of granularity and of abstraction, impacting results differently (Yao,

2003). Spatial and temporal auto-correlation can obfuscate important insights (Malerba, 2008). Additionally, scalability often is a concern.

Adding to the challenges already posed by tackling a spatio-temporal problem, we are also dealing with an imbalanced domain in our case study. Almost 90% of instances in our data set correspond to cases where the impact of wildfires was residual or null, but we are most interested in occurrences of major wildfires. Such domains create their own set of hurdles since **i)** standard evaluation metrics such as Mean Squared Error (MSE) for regression become inadequate (as discussed in Section 3.6), and **ii)** learning methodologies must be adapted to focus on a small but important subset of cases (Branco *et al.* , 2015).

1.2 Motivation and main goals

1.2.1 Real world applications

Spatio-temporal data mining techniques can be applied to several real-world problems including but not limited to meteorology (e.g., prediction of wind speed and temperature), biology (e.g., species relocation) and ecology (e.g., predicting wildfires – as in our case study).

Several propositional and relational approaches have been used to tackle the problems of finding patterns in spatio-temporal databases in order to describe them and predicting unknown values, as presented in Chapters 2 and 3 respectively. But not a lot of attention has been paid to uncovering the strengths and weaknesses of each kind of approach when dealing with spatio-temporal data.

1.2.1.1 Wildfires in Portugal

The problem of wildfires in Portugal is a particularly important one since wildfires severely affect the country every year, and any new knowledge found could prove to be actionable and have a beneficial impact in the country's forestry.

On average, from 1991 to 2010, more than 122 000 ha were burnt by wildfires per year. In 2003 alone, about 440 000 ha were burnt corresponding to almost 5% of the country's

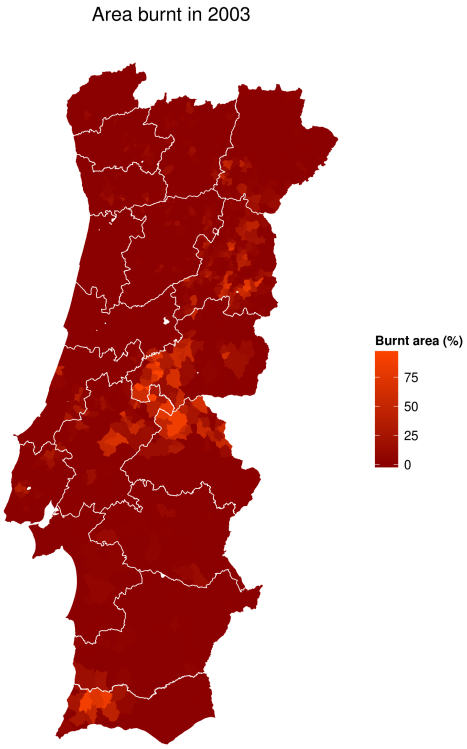


Figure 1.1: Percentage of burnt area in each Portuguese civil parish in 2003. Brighter red indicates a higher percentage. The boundaries pictured in white delineate Portuguese districts.

mainland area and almost 10% of the Portuguese forest (see Figure 1.1). Besides loss of forest area, carbon emission and deterioration of soil and downstream water quality are big environmental concerns. The economic damage caused is also significant. The average financial impact of wildfires between 2002 and 2006 is estimated to amount to 300 million euros per year, although in 2003 alone 1 billion euros were lost. But the repercussions are not just ecological and financial. Wildfires are responsible for the loss of civilians' and firefighters' lives. Twenty-one people died in 2003, followed by eighteen in 2005; both years registered more than a thousand injured. (Bassi *et al.* , 2008).

Given this motivating application, we aim at **a)** understanding wildfires in Portugal by learning spatio-temporal association rules and **b)** forecasting the fraction of each civil parish's area burnt yearly by wildfires.

1.2.2 Main goals

Thus, this dissertation aims at **i)** reviewing the state of the art and enumerating the main challenges facing both propositional and relational approaches to the data mining tasks of **a)** association rule learning (descriptive) and **b)** forecasting (predictive) in a spatio-temporal setting; **ii)** applying promising approaches to a case study involving wildfires in Portugal (an imbalanced domain); **iii)** comparing and providing insight on the strengths and weaknesses of the propositional and relational approaches used in the two very different tasks.

1.3 Dissertation outline

The dissertation is structured in seven chapters whose contents are described below:

Chapter 2 – Descriptive Spatio-Temporal Data Analysis presents an overview of the state-of-the-art propositional and relational approaches to spatio-temporal association rule learning, including a discussion of metrics for performance evaluation.

Chapter 3 – Predictive Spatio-Temporal Data Analysis presents an overview of the state-of-the-art propositional and relational approaches to spatio-temporal regression tasks followed by a discussion on performance metrics (including methods and metrics specially adapted to address imbalanced domains).

Chapter 4 – Wildfires in Portugal: A Case Study describes our motivating application, and details the computation of spatial relationships in the data set.

Chapter 5 – Describing Wildfires describes a propositional and a relational approach to association rule learning applied to our motivating application. Experimental results are presented, and their strengths and weaknesses discussed.

Chapter 6 – Predicting Wildfires describes a propositional and a relational approach to the problem of forecasting the fraction of each Portuguese parish's area

burnt yearly by wildfires (an imbalanced domain). Experimental results are presented and discussed.

Chapter 7 – Conclusion concludes the dissertation and outlines possible future research directions.

Chapter 2

Descriptive Spatio-Temporal Data Analysis

2.1 Introduction

As stated in Chapter 1, major descriptive data mining tasks include anomaly detection, clustering and association rule learning.

There has been much interesting work on spatio-temporal clustering (e.g., Neill *et al.* (2005); Camossi *et al.* (2008); Ciampi *et al.* (2010)) and anomaly detection (e.g., Janeja *et al.* (2010); Das *et al.* (2012); Telang *et al.* (2014)). We will focus on learning association rules in this study, as this is a widely used approach that can be naturally upgraded to perform multi-relational learning. Association rules are usually very interpretable, which can be a huge advantage, particularly when we want to **i)** understand the importance of attributes and relations in a certain problem, and **ii)** effectively communicate our results to others that can potentially find them useful and/or actionable. Note that this applies to our case study since we would like to gain a better understanding of the factors contributing to wildfires, and express the relationships we find in a way that could be easily interpretable by policy-makers.

In order to better understand common approaches to pattern mining in spatio-temporal databases, it is useful to first define the general problems of association rule learning and the related problems of sequential pattern mining and inter-transaction rule mining.

2.2 Problem definition

2.2.1 Association rules

Association rules as proposed by Agrawal & Srikant (1994) reveal regularities in large transactional databases. In order to understand them, let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of literals, called *items*. An item might be a literal item purchased at a store, an event or anything that can be abstracted as an attribute-value pair, depending on context. A set $A = \{i_1, \ldots, i_k\} \subseteq I$ is called an *itemset* (or k-itemset if $|A| = k$). A transactional database D consists of pairs $T = (TID, A_{TID})$ where TID is the unique identifier associated with each transaction and A_{TID} is an itemset. A transaction $T = (TID, A_{TID})$ is said to support an itemset B if $B \subseteq A_{TID}$.

An association rule can then be expressed by an implication of the form

$$A \Rightarrow C$$

where the antecedent (A) and consequent (C) are sets of items with $A \cap C = \emptyset$. The *support* of the rule is defined to be the percentage of transactions that support $A \cup C$. The *confidence* of the rule is defined as the proportion of transactions containing A that also contain C, i.e., the confidence is given by $support(A \cup C)/support(A)$.

Common propositional algorithms for association rule mining include Apriori (Agrawal & Srikant, 1994), eclat (Zaki, 2000) and FP-growth (Han *et al.* , 2004). WARMR (Dehaspe & Toivonen, 1999) upgrades the Apriori algorithm to a multi-relational setting.

An association rule, as defined above, is said to be spatial when at least one of the items in A or C expresses a spatial relationship (Koperski & Han, 1995). Spatial items represent:

- Topological relationships between spatial objects, e.g. disjoint, intersects, inside/outside, adjacent_to, covers/covered_by, equal;
- Spatial orientation or ordering, e.g. left, right, north, east;
- Information regarding distance, e.g. close_to, far_away.

We introduce the temporal dimension in this context by either **i)** designing temporal attributes that one finds important (e.g., expressing temporal order) or having a first step that extracts them (e.g., expressing temporal change or patterns), or **ii)** using an algorithm to find sequential patterns or inter-transaction rules which intrinsically respect temporal order.

2.2.2 Sequential patterns

Sequential pattern mining was first introduced by Agrawal & Srikant (1995) and it is closely related to association rule mining in databases. A *sequential pattern* is a maximal sequence that has at least a minimal support, where support for a pattern is defined as the percentage of sequences in the database that contain the pattern. When looking for sequential patterns, the order of the item-sets becomes significant.

A sequence, denoted by $< a_1, a_2, \ldots a_n >$, is an ordered list of itemsets a_i, as defined above. A sequence of k itemsets is a k-sequence. We will continue to denote itemsets by $\{x_1, x_2, \ldots x_m\}$ where x_j is an item. An item can occur only once in an itemset, but multiple times in a sequence. A sequence $< a_1, a_2, \ldots a_n >$ is contained in another sequence $< b_1, b_2, \ldots b_n >$ if there exist integers $i_1 < i_2 \cdots < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$,...,$a_n \subseteq b_{i_n}$. For example, $< \{a\}, \{b, c\}, \{d\} >$ is contained in $< \{x\}, \{a, y\}, \{w\}, \{b, c, z\}, \{d\} >$. As another example, $< \{a, b\} >$ is not contained in $< \{a\}, \{b\} >$ (or vice-versa); the former implies that a and b are concurrent while the latter implies that b occurs after a. Given a set of sequences, we say that a sequence is maximal if it is not contained in any other sequence.

A sequence can encompass several transactions. In a market-basket problem, a sequence can literally consist of sets of items purchased by a certain costumer at ordered times. In a spatio-temporal context, a sequence can consist of ordered sets of events co-occurring in the same location.

Common propositional algorithms for sequential pattern mining include Apriori-based GSP (Srikant & Agrawal, 1996), SPADE (Zaki, 2001) and prefixspan (Pei *et al.* , 2001, 2004). MDSL (Esposito *et al.* , 2009) is an Apriori-based algorithm for sequential pattern mining in a multi-relational setting.

2.2.3 Inter-transaction association rules

As we have stated, the classical concept of association rules is intra-transactional since it looks for correlations within transactions (examples of transaction include all items bought by the same costumer in a single trip or atmospheric events that happened at the same time).

Even though sequential pattern mining does encompass several transactions in each sequence, it can be argued that it is still intra-transactional in nature since it is possible to abstract the database so that **i)** each sequence is seen as one ordered transaction and **ii)** the mining process looks for similarities between sequences, i.e., patterns always respect the boundary imposed by the sequence identifier that works as a transaction identifier would in a classical association rule setting. But it is possible to extend this framework from intra-transactional to inter-transactional (Tung *et al.* , 1999), even including n-dimensional rule discovery (Lu *et al.* , 1998, 2000). An example might illustrate the relevance of this extension: while the intra-transactional classical setting might allow us to find a rule like *"When the temperatures in Braga and Aveiro increase, the temperature in Porto also increases (on the same day)"* and a sequential pattern like *"The temperature increases after decreasing three consecutive times"*, the inter-transaction framework allows the discovery of single-dimensional rules like *"If the temperatures in Braga and Aveiro increase, the temperature in Porto will increase the next day"* and two-dimensional rules like *"After wildfires occur in Rio Caldo and Valdosende, another wildfire will occur two months later less than twenty kilometers away"*.[1]

Let us again consider a set of literals $I = \{i_1, i_2, \ldots, i_n\}$ called items and a transactional database $T = \{t_1, t_2, \ldots, t_m\}$. A single-dimensional mining space can be represented by a dimensional attribute (e.g., time, latitude) with an ordinal domain that can be divided into equal-sized intervals. For example, time can be divided into days, weeks, or months; other continuous attributes can be discretised in order to define equal-sized intervals.

Let $n_i =< v >$ and $n_j =< u >$ be two points in the one-dimensional space. A relative distance between them is defined as $\Delta(n_i, n_j) =< u - v >$ and their reference point is defined as $n_0 = \min(u, v)$. We will use $\Delta(n_i)$ or simply Δ_i to refer to $\Delta(n_0, n_i)$. We

[1] Adapted from (Lu *et al.* , 1998)

call an item i_k at the point Δ_j an extended item and denote it as $\Delta_j(i_k)$. In general, $\Delta_i(i_k)$ and $\Delta_j(i_k)$ are not equal unless $\Delta_i = \Delta_j$. Similarly, we call a transaction t_k at the point Δ_j an extended transaction and denote it as $\Delta_j(t_k)$. The set of all possible extended items, I_e, is defined as the set of $\Delta_j(i_k)$ for any $i_k \in I$ at all points Δ_j in the one-dimensional space. The set of all extended transactions in the one-dimensional space is represented by τ_e. The reference point of an extended transaction subset is defined to be the minimum Δ_j among all $\Delta_j(t_k)$ in the subset.

A single-dimensional inter-transaction association rule is an implication of the form $A \Rightarrow C$, where $A \subset I_e$, $C \subset I_e$ and $A \cap C = \emptyset$. The support of such a rule is defined as $|T_{ac}|/|\tau_e|$ where T_{ac} is the set of extended transactions that contain $A \cup C$. Its confidence can be defined as $|T_{ac}|/|T_a|$ where T_a is the set of extended transactions that contain A.

Propositional algorithms proposed to mine inter-transactional association rules include EH-Apriori (Lu *et al.* , 2000; Feng *et al.* , 2001) based on Apriori, FITI designed specifically for this problem (Tung *et al.* , 1999, 2003) and RPPI (Huang *et al.* , 2008) based on prefixspan.

2.3 Propositional approaches to spatio-temporal descriptive data mining

Before presenting spatio-temporal pattern mining techniques, let us broach the related subject of spatial pattern mining. First, we define a few key concepts that will be used from this point on:

- A *reference* spatial object is the main subject of a description task;
- A *task-relevant* spatial object is an object that is spatially related to the reference object;
- A *concept hierarchy* is a tree structure that forms a taxonomy of concepts ranging from a single, most general concept at the root to all specialisations of them at the leaves.

2.3.1 Spatial patterns

Koperski & Han (1995) present a top-down, progressive refinement method to discover multi-level spatial association rules between objects based on different neighbourhood relations.

First, the task-relevant objects are extracted by the execution of a query. Then, neighbourhood relations between reference (e.g., town) and task-relevant objects (e.g., road) are computed at a coarse level using efficient spatial algorithms (such as R-trees or fast MBR technique and plane-sweep algorithm).

Town	Water	Road
Victoria	< adjacent to, J. Fuca Strait >	< intersects, highway 1 > < intersects, highway 17 >
Saanich	< adjacent to, J. Fuca Strait >	< intersects, highway 1 > < close to, highway 17 >
Prince George		< intersects, highway 97 >
...	...	...

Table 2.1: Finely computed neighbourhood relations adapted from Koperski & Han (1995)

Starting at the top-most spatial taxonomy level, the specific spatial objects in the table are abstracted (e.g., the J. Fuca Strait will just be considered a body of water) so item frequencies can be counted. Note that, in this context, items correspond to pairs of a spatial relation (in regard to a reference object) and a spatial (task relevant) object. These items, stored in a single double-entry table, are filtered by minimum support following an Apriori strategy and, if desired, a more refined computation of spatial relationships follows with results similar to Table 2.1 which follows an extended relational model (each cell of the table may contain more than one entry). The frequency of these refined items is again counted and filtered by minumum support, producing a result like Table 2.2.

Finally, strong association rules can be directly mined from this table using the Apriori algorithm. Following the progressive deepening process first presented by Han & Fu (1995) for non-spatial transaction-based databases, only descendants of the frequent 1-itemsets found at the topmost concept level are examined at a lower concept level and so forth.

k	frequent k-itemset	count
1	$<$ `adjacent to, water` $>$	32
1	$<$ `intersects, highway` $>$	29
2	$<$ `adjacent to, water` $>$, $<$ `intersects, highway` $>$	25
...	...	...

Table 2.2: Large itemsets found at the top concept level (for 40 large towns) as presented by Koperski & Han (1995)

From the example, a rule like

$$\{< \texttt{large town} >, < \texttt{intersects, highway} >\} \Rightarrow < \texttt{adjacent to, water} >$$

could be derived at the top-most concept level and one like

$$< \texttt{large town} > \Rightarrow < \texttt{adjacent to, sea} >$$

could be found at a lower concept level.

GeoMiner (Han *et al.* , 1997), a spatial data mining system prototype, integrates this procedure into its Geo-associator module. Ester *et al.* (1997) present a way to implement this algorithm using the concepts of neighborhood graphs and paths.

2.3.2 Spatio-temporal patterns

We have identified three classes of propositional approaches to association rule learning: **i)** intra-transaction pre-processing based, **ii)** intra-transaction context based and **iii)** inter-transaction based.

2.3.2.1 Intra-transaction pre-processing based

These approaches are heavily dependent on pre-processing steps, using standard out-of-the-box algorithms for association rule learning. After those steps, the items can correspond to **i)** attribute-value pairs or **ii)** anomalous events, being identified with a time and location pair or just a location when values represent attributes' temporal changes.

Item as attribute-value pair Tsoukatos & Gunopulos (2001) present DFS_MINE, a depth-first-search-like sequential pattern mining algorithm which allows fast discovery of maximally frequent sequences of events in spatio-temporal data sets (possibly with various levels of spatial granularity) by performing database scans. This algorithm is applicable to a market-basket table in which each transaction is identified by a time-stamp and location ID and each item is a spatio-temporal event defined as an attribute-value pair (e.g., indicating a temperature and/or atmospheric pressure). Frequent sequences of lower granularity are generated by joining several regions into a greater region in which the sequence of events of each sub-region is considered to occur (this allows a region to have more than one value assigned to a certain attribute at a given time, which is not usually permitted at a higher level of granularity). The efficiency of the method is tested on synthetic meteorological data sets outperforming SPADE (Zaki, 2001) on space efficiency.

Mennis & Liu (2005) explore socioeconomic and land cover change in a region of the USA using GIS-based pre-processing to integrate diverse data sets, discretise numerical data, extract spatio-temporal relationships and encode them in tabular format which is used by standard association rule mining software (Ma *et al.* , 1998; Agrawal & Srikant, 1994). In this case, the temporal dimension is incorporated by considering the percent changes in variables of interest, that is, each transaction is identified by a location only. Discretisation is based on Jenks natural breaks optimisation (Jenks, 1967). A hierarchy is determined by the number of classes established for a variable (the most coarse level having fewer categories) so that it becomes possible to mine for association rules on multiple concept-levels.

Item as anomalous event Tan *et al.* (2001) transform climate data into market-basket transactions in order to apply standard techniques for spatial association rule and sequential pattern mining after dealing with the data's temporal seasonality and auto-correlation. Both intra-zone and inter-zone sequential and non-sequential patterns are considered.

The data consisted of spatially-indexed time series of several climate indices. To convert this into a market-basket type transaction table for intra-zone pattern mining, lower and upper limits were defined for each variable so that values outside of that interval were considered anomalous events (distinguishing between lower or higher

than the normal range for the variable). Anomalous events are then attributed to the appropriate transactions (identified by a time-stamp and a location ID).

When searching for intra-zone non-sequential pattern mining (standard association rule mining problem), the Apriori algorithm or FP-growth are applied to the resulting table; for intra-zone sequential patterns (temporal problem), the sequential pattern mining algorithm GSP is used. More pre-processing is required before inter-zone non-sequential and sequential patterns are searched for. The approach used is based on the work of Koperski & Han discussed in Section 2.3.1, except the geographical landmarks are replaced with events over certain regions of interest.

2.3.2.2 Intra-transaction context based

Tang *et al.* (2008) systematically derive a set of contexts by combining the concept levels of user-defined time and location hierarchies. An efficient algorithm for context-based market basket analysis in a multiple-store and multiple-period environment is proposed. A section of the database is defined as the subset of transactions occurring at certain combinations of time and location, each with a specific level of granularity (e.g., a subset of transactions in Gulf Coast and March or East Coast and February). Contexts for itemsets, sections and itemsets in determined sections are defined in order to create time- and place- specific rule selection criteria considering that not every item is on-shelf at every store at all times. The algorithm uses Apriori or FP-growth to obtain context-large itemsets (i.e., itemsets that are frequent given their context) and hash-trees to derive count information on lower granularity levels from higher granularity counts.

2.3.2.3 Inter-transaction based

Inter-transaction rule mining methods can be applied to transactions identified by a time of measurement and including attribute-value pairs as items.

Feng *et al.* (2001) use single-dimensional inter-transactional association rule mining for a weather forecasting application involving meteorological data captured from six different spatially related stations over a period of four years. The six meteorological variables were first discretised, and missing values were imputed. Then, the data was

transformed into market-basket format where each transaction included attribute-value pairs for each station and meteorological variable, extended with a dimensional attribute representing time (i.e., each transaction is identified by the time the measurements were taken).

Later, Huang *et al.* (2008) would use a very similar approach to model abnormal changes in ARGO salinity and temperature data in Taiwan. Instead of working with data from a number of stations, they work with data from neighbouring regions defined by concentric circles centred on Taiwan (following the reference-centric model first introduced by Huang *et al.* (2004)). Each transaction is identified by its dimensional attribute (time of record) and includes the set of all events regarding temperature and salinity variations (which were also discretised) in each defined region (see Table 2.3).

Time	Salinity in A1	Temperature in A1	Salinity in A2	Temperature in A2	...
2001Jan	SDL	TRM	SDM	TDL	...
2001Feb	SDL	TRL	NOR	NOR	...
...	...	...	...	...	...

Table 2.3: Data format used by Huang *et al.* (2008). The items correspond to discretised intervals of variation in temperature and salinity (e.g., SDL means salinity dropped little, TRM means temperature rose much, NOR means no abnormal event happened). Each column is dedicated to variations in a certain spatial region (such as A1 and A2).

When we consider time-dimensional inter-transaction rules, the usual emphasis of association rule mining on description is somewhat shifted to prediction (Feng *et al.* , 2001), a topic that will be discussed in Chapter 3.

2.4 Relational approaches to spatio-temporal descriptive data mining

Relational approaches to association rule learning we found were based on Inductive Logic Programming (ILP) (see Muggleton & De Raedt (1994); Lavrac & Džeroski (1994)).

During routine use, A Learning Engine for Proposing Hypotheses (Aleph), the ILP system we will be using in Chapters 5 and 6, follows a four step procedure (Srinivasan, 2007):

Example selection A (positive) example for generalisation is selected. If none exist, Aleph stops.

Saturation A bottom-clause (or saturated clause) is constructed by building the most specific clause that entails the selected example, within language constraints (for example, respecting a maximum number of layers of new variables). This step follows the work of Muggleton (1995), and can be reproduced by the command `sat(example_number)`.

Reduction A search for a more general clause is conducted by looking for some subset of the literals in the bottom-clause corresponding to the best score (which can be defined in various ways). This is implemented by a (restricted) branch-and-bound algorithm which allows an intelligent enumeration of acceptable clauses under a range of different conditions. This step can be reproduced by using the command `reduce/0` after a saturation.

Redundancy removal The clause with the best score is added to the theory, and all examples covered by it are removed. After this step, Aleph returns to the first step.

Before moving on to methods to handle spatio-temporal data sets, we will present related work on spatial data sets.

2.4.1 Spatial patterns

The ILP system SPADA (Spatial PAttern Discovery Algorithm) for the discovery of multi-level spatial association rules from a deductive spatial database is introduced and built upon by Malerba & Lisi in 2001a; 2001b; 2004.

SPADA resorts to Datalog (Ceri *et al.* , 1989) as its data representation formalism. In this multi-relational setting, items are represented as first-order logic atoms, that is, n-ary predicates applied to n terms which can be either variables or constants. Most

often, ILP systems need to search a very large space of possible clauses, trying to generalise from individual examples in the presence of background knowledge in order to find patterns/hypothesis about yet unseen instances. A language bias is specified to indicate which predicates can be used in the patterns and to formulate constraints on the binding of variables.

The atom denoting the reference object is called the key atom and it must be included in the antecedent of a spatial association rule; and at least one spatial relation must be in the antecedent and/or the consequent. A spatial observation is the set of ground facts in the spatial database that can be uniquely identified by relating to a particular reference object. A pattern covers an observation if, when turned into a Datalog query, it is true in the union of the observation with the background knowledge. The support of the rule is defined as the percentage of spatial reference objects covered by both the antecedent and consequent of the rule.

SPADA starts by finding frequent patterns through a breadth-first search in the lattice of patterns spanned by a generality order based on θ-subsumption (Plotkin, 1970). From these patterns, highly confident spatial association rules are then generated (Malerba *et al.* , 2009).

SPADA has been interfaced to modules for the extraction of spatial features from a spatial database and for numerical attribute discretisation (Appice *et al.* , 2003). It has also been integrated into the spatial data mining distributed system ARES (Appice *et al.* , 2005) and into GIS prototypes with a geographic knowledge discovery engine, INGENS (Malerba *et al.* , 2003; Appice *et al.* , 2008) and INGENS 2.0 (Malerba *et al.* , 2009).

These relational approaches have been applied to geo-referenced UK census data (Malerba & Lisi, 2001a; Lisi & Malerba, 2004; Appice *et al.* , 2005), spatial data of an Italian province (Malerba & Lisi, 2001b; Malerba *et al.* , 2009) and urban accessibility of a UK hospital (Appice *et al.* , 2003).

2.4.2 Spatio-temporal patterns

Next, we present existing approaches to relational spatio-temporal association rule learning that, as previously mentioned, are ILP based.

2.4.2.1 ILP based

Multi-dimensional relational patterns are defined by Esposito *et al.* (2009) as a set of Datalog atoms, involving k events and concerning n dimensions. The atoms may be non-dimensional or dimensional. Non-dimensional atoms can be divided into *fluents* defined as functions whose domain is the space of situations (which, in turn, are defined as the complete state of the universe at a certain instant of time), explicitly referring to a given event (i.e., one of its arguments denotes an event), and *non-fluents* denoting relations between objects or characterising an object involved in the sequence. Each event may be related to another event by means of dimensional operators that refer dimensional relations between events involved in the sequence (such as *next step in dimension i* and *after n steps on dimension i*).

In order to define the support of a pattern, we first define pattern subsumption. A pattern P subsumes a sequence S if there exists a SLD_{OI}-deduction of P from $B \cup U$ where B is the background knowledge and U is the set of ground atoms in the sequence S. A SLD_{OI}-deduction is a Selective Linear Definite deduction under Object Identity, meaning that within a clause, terms that are denoted with different symbols must be distinct, i.e., they must represent different objects of the domain. Given a multi-dimensional relational pattern $P = (p_1, p_2, \ldots, p_n)$ and S a multi-dimensional relational sequence, the frequency of pattern P is equal to the number of different ground literals used in all the possible SLD_{OI}-deductions of P from $B \cup U$ that make true the literal p_1.

The proposed MDSL algorithm, based on Apriori, starts with the most general patterns and successively tries to specialize them using θ_{OI}-subsumption. Its performance compared well with WARMR when applied to synthetic relational data.

2.5 Performance metrics

In order to evaluate and compare association rule mining approaches, we need to define the most adequate performance metrics. When evaluating the quality of rules, there are several possible metrics that try to quantify their interestingness or predictive

ability, from the most standard (and widely used in the mining process) support and confidence to less used metrics such as improvement, lift, conviction and the χ^2-test.

The support, supp, of an association rule is the probability of finding an observation containing $A \cup C$, i.e.,

$$\text{supp}(A \Rightarrow C) = \Pr(A, C) \tag{2.1}$$

We discuss these metrics in terms of probability because it allows for a more generalised interpretation of support. In a classical setting, as mentioned in Section 2.2, this probability corresponds to the fraction of transactions in the database containing $A \cup C$. When working with inter-transaction rules (as in Section 2.3.2.3) or in a multi-relational setting (as in Section 2.4), it must be defined in a different way.

The confidence of the rule, conf, is the probability that an observation containing A will also contain C, i.e.,

$$\text{conf}(A \Rightarrow C) = \Pr(C|A) \tag{2.2}$$

Rules with high support and confidence are usually considered to be strong rules. Other measures of predictive ability (or interestingness) of rules include improvement, lift, conviction and Chi-square test statistics (χ^2).

Improvement of a rule, imp, is the minimum difference between its confidence and the confidence of any of its immediate simplifications (Bayardo *et al.* , 2000), that is,

$$\text{imp}(A \Rightarrow C) = \min(\text{conf}(A \Rightarrow C) - \text{conf}(As \Rightarrow C)|As \subset A) \tag{2.3}$$

Lift (Berry & Linoff, 1997) (also named interest by Brin *et al.* (1997b) or strength by Dhar & Tuzhilin (1993)) of a rule can be defined as

$$\text{lift}(A \Rightarrow C) = \frac{\Pr(C|A)}{\Pr(C)} \tag{2.4}$$

Conviction (conv), unlike confidence, is normalized based on both the antecedent and consequent of a rule and, unlike lift, is directed. It is defined by Brin *et al.* (1997b) as

$$\text{conv}(A \Rightarrow C) = \frac{\Pr(A) \cdot \Pr(\neg C)}{\Pr(A, \neg C)} \tag{2.5}$$

Pearson's χ^2-test is a widely used method for testing independence and/or correlation based on the comparison of observed frequencies with the corresponding expected

frequencies. The closer these two values are, the greater is the weight of evidence in favor of independence between antecedent and consequent, meaning lower interestingness of the rule.

$$\chi^2 = \sum \frac{(f_o - f)^2}{f} \tag{2.6}$$

where f_o is an observed frequency and f is an expected frequency. Brin *et al.* (1997a) use this measure (instead of the common support-confidence framework) in the mining process to find correlation rules, which are a generalization of classical association rules.

Other criteria can be just as important although sometimes harder to quantify. For example, Calargun & Yazici (2008) consider interpretability, utility and novelty of the discovered rules.

Association rule mining techniques can be evaluated by their computational performance when dealing with voluminous data, the number of associations found and the distribution of their quality metrics.

2.6 Summary

In Section 2.2, we have defined the problem of association rule learning and sequential pattern mining (which can be seen as a version of the former considering temporal order). We have identified and presented different approaches to these problems in a spatio-temporal setting, in Sections 2.3 and 2.4, which we summarise in Table 2.4.

Propositional	Intra-transaction	Pre-processing based	Mennis & Liu (2005); Tsoukatos & Gunopulos (2001); Tan *et al.* (2001)
		Context based	Tang *et al.* (2008)
	Inter-transaction		Feng *et al.* (2001); Huang *et al.* (2008)
Relational	ILP based		Esposito *et al.* (2009)

Table 2.4: Propositional and relational approaches to spatio-temporal association rule learning

Finally, we have cited in Section 2.5 measures and criteria to evaluate the quality of rules (see Table 2.5) and techniques used to mine them (such as computational performance).

Quantitative	Undirected	Support	(Eq. 2.1)
		Confidence	(Eq. 2.2)
		Improvement	(Eq. 2.3)
		Lift	(Eq. 2.4)
		χ^2-test	(Eq. 2.6)
	Directed	Conviction	(Eq. 2.5)
Qualitative		Interpretability	
		Utility	
		Novelty	

Table 2.5: Metrics to assess the quality of mined association rules

Chapter 3

Predictive Spatio-Temporal Data Analysis

3.1 Introduction

Predictive data analysis faces the problem of approximating an unknown function $Y = f(X_1, X_2, \ldots, X_p)$ mapping values of a set of predictors, X_i, into the values of a *target* variable, Y. The function approximation is usually called the *model*.

The model is built using a training set $D = \{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$. If the target variable is categorical, that is, if it takes its value from a finite set, we face a classification problem; if it is numerical, a regression problem. The learning process tries to optimise the model's parameters according to a criterion such as the Error Rate (for classification) or the Mean Squared Error (for regression problems).

Given the nature of our motivating application, which concerns the prediction of the extent of a parish's area burnt yearly by wildfires (numerical target variable), we mainly focus on regression. Regression can be used to predict future or otherwise unobserved numerical values, but it can also be used to fill in missing data.

Often, values in spatial data sets are measured at a limited number of geographical points and it is useful to estimate them at unobserved locations. The same is true for spatio-temporal data sets, with the added issue of measurements being taken at limited time points.

Missing data occurs when values are missing for some (but not all) variables and for some (but not all) cases (in a spatial data set, each case corresponding to a location; in a spatio-temporal data set, to a location and time). Data can be missing for a number of reasons, including sensor failures and mistakes in data collection or entry. Missing data can pose a serious problem since conventional (propositional) modelling methods presume that all variables in a specified model are measured for all cases. When dealing with missing data, different assumptions can be made: one might assume that the data is missing **a)** completely at random, the strongest assumption; **b)** at random, but where missingness can be accounted for by variables with complete information, a weaker assumption; or **c)** not at random, when both of the previous assumptions are violated, and the missing data mechanism cannot be ignored (but can usually be described in detail by a relational graphical model) (Allison, 2001).

In the following sections, we define the regression problem of spatio-temporal forecasting, present existing approaches to tackle it, and describe evaluation metrics to assess their performance.

3.2 Problem definition

At this point, and before we define the main problem of spatio-temporal forecasting, we would like to introduce the related problem of spatial interpolation, i.e., predicting missing or unobserved values in spatial data sets.

3.2.1 Spatial interpolation

The problem of spatial interpolation can be defined as the problem of forecasting missing or unobserved values in spatial data sets. The predictive task essentially consists of estimating unknown values of a target variable, Y, on certain locations, based on a spatial data set $D = \{y_1, y_2, \ldots, y_n\}$ where y_i corresponds to the value of variable Y at location i. Some of these problems can include other predictor variables, X_i, as background knowledge. In Sections 3.3.1 and 3.4.1, we introduce a few propositional and relational solutions to this very well researched problem.

3.2.2 Spatio-temporal forecasting

Working in a spatio-temporal setting, we want to predict values at different times and locations. Two main scenarios can be considered: **a)** forecasting of future values given information from spatially related locations in the past, and **b)** interpolation of missing or unobserved values given information on spatially and temporally related points, which can be viewed as an extension of the spatial interpolation problem mentioned above. That is, given a data set $D = \{y_1^1, y_2^1, \ldots, y_n^1, y_1^2, \ldots, y_n^m\}$ where y_i^j corresponds to the value of target variable Y at location i and time j, **(a)** aims to predict the value of Y at a location of interest, l (usually among the locations in D), at a future point in time, k, with $k > m$, while **(b)** aims to predict values within the time-frame of D, with $1 \leq k \leq m$. Again, the data set might also include other explanatory variables, X_i, as background knowledge.

In Sections 3.3.2 and 3.4.2, we present a few propositional and relational approaches to the spatio-temporal forecasting problem, mainly concerning the first scenario described.

3.3 Propositional approaches

3.3.1 Spatial interpolation

There has been extensive research on the topic of spatial interpolation, most of it motivated by the first law of geography (Tobler, 1970) stating that neighbouring points should have strongly correlated values.

Li & Shi (2010) divide techniques into three categories: **i)** non-geostatistical interpolators, **ii)** geostatistical interpolators and **iii)** a combination of both. Non-geostatistical interpolators are based on the distance between neighbours while geostatistical interpolators are based on Kriging (Krige, 1951).

The simplest example of a non-geostatistical interpolator is Inverse Distance Weighting (IDW) (Isaaks & Srivastava, 1989) which approximates the unknown value at a certain location as the weighted average of the known values at neighbouring locations, where the weights are inversely proportional to the distance from the target location.

Kriging also approximates the unknown value at a location by considering the known values at neighbouring locations. However, the weights are calculated considering the covariation between known data points at several locations. There are several variants of kriging which approximate these weights differently.

3.3.2 Spatio-temporal forecasting

We found four main types of propositional approaches to spatio-temporal forecasting which can be divided into solutions based on **a)** pre-processing, **b)** clustering, **c)** the combination of spatial and temporal methods, and **d)** the integration of the spatial and temporal dimensions.

3.3.2.1 Pre-processing based

Approaches heavily dependent on pre-processing use **a)** lagged temporal and/or spatial inputs or **b)** other spatio-temporal indicators.

Lagged temporal and/or spatial inputs Luk *et al.* (2000) transformed rainfall data so that the input, later fed to an Artificial Neural Network (ANN), becomes a vector $\mathbf{x}(t), \ldots, \mathbf{x}(t-k+1)$ where $\mathbf{x}(i)$ represents a vector of rainfall values at M locations at time i with k temporal lag. They investigated the effect of varying temporal lags and number of neighbouring spatial inputs included on the prediction accuracy of $\mathbf{x}(t+1)$, finding an apparent trade-off between the inclusion of temporal and spatial information. Note that the number of neighbour values included, M, does not necessarily correspond to the number of locations in the prediction, N. For their particular dataset, they found the best performing ANN used information from the eight nearest neighbouring sites lagged by only one time-step.

Bilgili *et al.* (2007) also used an ANN to predict the future monthly average wind speed of a target station using as inputs the month in question and the monthly average wind speed at reference locations, selected based on the correlation between them and the target.

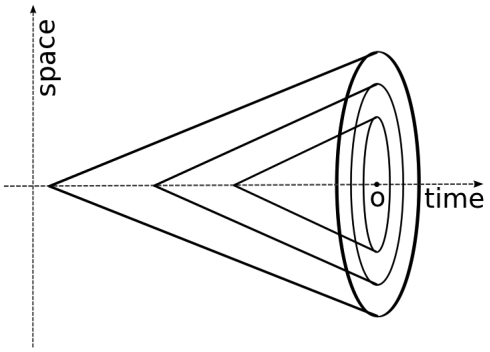


Figure 3.1: Spatio-temporal neighbourhoods of different sizes as defined by Ohashi & Torgo (2012)

Other spatio-temporal indicators Ohashi & Torgo (2012) tackled wind speed forecasting as a multiple-regression problem by building spatio-temporal indicators from historical wind speed data both at the target and neighbouring locations within a radius that is higher for more recent observations. Spatio-temporal conic neighbourhoods like the ones represented in Figure 3.1 are defined around a central location. The proposed approach considers conic neighbourhoods of different sizes for a spatio-temporal point O and calculates several indicators, including **i)** the average/standard deviation of values within each conic neighbourhood; **ii)** ratios between averages of different neighbourhoods; and **iii)** a weighted version of the averages where the weights of data points are inversely proportional to their spatio-temporal distance to O.

3.3.2.2 Spatio-temporal clustering based

We found approaches that follow spatio-temporal clustering by **a)** data pre-processing steps and a standard learning technique or **b)** a temporal forecasting technique.

Data pre-processing and standard learning technique Appice *et al.* (2013a) took a very similar approach to Ohashi & Torgo (2012), but the similarly calculated spatio-temporal indicators were built from automatically discovered spatio-temporal clusters instead of conic spatio-temporal neighbourhoods. The spatio-temporal clusters are discovered over a temporal sliding window in two steps. First, each temporal snapshot of the window is divided into regions based on both spatial location and attribute information; then, spatio-temporal clusters are obtained by grouping locations, which were classified into the same sequence of spatial clusters in the previous step.

Appice *et al.* (2013b) described a similar approach to the problem of spatio-temporal interpolation (instead of the forecasting of future values), which also operates in two phases. The exploration phase encompasses trend cluster discovery, determining data trends and geographically aware station interactions in a time window. The estimation phase uses inverse distance weighting both to approximate observed data and to estimate missing data.

Temporal forecasting technique Pravilovic & Appice (2014) described a two-stepped algorithm that accounted for the spatio-temporal correlation of geo-referenced time series. Firstly, spatio-temporal k-means clusters are computed. Secondly, a new inference procedure computes the best Auto-Regressive Integrated Moving Average (ARIMA) forecasting parameters valid for all time series in the cluster, and use it to produce forecasts.

3.3.2.3 Combined temporal and spatial methods

Li *et al.* (2003) proposed a spatio-temporal forecasting framework (STIFF) consisting of three steps. First, an ARIMA model is applied to the time series of each location, saving the temporally-influenced forecasts. Secondly, a neural network is trained with values at the target and neighbouring locations, and then fed with the ARIMA forecast of the neighbouring locations, outputting a spatially-influenced forecast for the target location. Finally, the two forecasts for the target location are combined via statistical regression to generate the overall forecast, in this case, of water flow rate.

Cheng & Wang (2008) aimed at improving the accuracy of this approach by substituting the static feed-forward neural network by a dynamic recurrent one (with feedback connections) for wildfire area prediction.

3.3.2.4 Integrated spatial and temporal dimensions

Pace *et al.* (1998) proposed a parsimonious auto-regressive model for housing prices estimation that accounts for both temporal and spatial dependence and obtains overall better results than a traditional Ordinary Least Squares (OLS) hedonic pricing model using indicator variables.

Lindström *et al.* (2014) presented a spatio-temporal framework that predicts ambient air pollution by combining data from different monitoring networks and deterministic air pollution models with geographic information system co-variates. The model has been implemented in an R package (Lindström *et al.* , n.d.). Accuracy in predicting long-term average concentrations is evaluated using an elaborate cross-validation setup that accounts for a sparse spatio-temporal sampling pattern in the data, and adjusts for temporal effects.

3.4 Relational approaches

3.4.1 Spatial interpolation

Stojanova *et al.* (2012) proposed the NCLUS algorithm for network regression that explicitly considers autocorrelation when building regression models from network data, based on the concept of Predictive Clustering Trees (PCTs). On networks obtained from spatial data, edges are defined for each pair of nodes and dissimilarities are computed according to the spatial distance between the nodes.

3.4.2 Spatio-temporal forecasting

Relational approaches to spatio-temporal forecasting can be divided into **i)** graphical models and **ii)** ILP based solutions.

3.4.2.1 Graphical models

Bayesian Networks (BNs) Cano *et al.* (2004) presented a local learning algorithm for BNs which takes advantage of the spatial character of the problem, applying the resulting graphical models to different meteorological problems including local weather forecasting using rainfall and maximum wind data.

Madadgar & Moradkhani (2014) developed a statistical forecasting model within BNs at each spatial grid cell using historical runoff data. A family of multivariate distribution

functions are applied to forecast future drought conditions given the drought status in the past.

Hidden Markov Models (HMMs) Thompson *et al.* (2007) proposed a HMM for daily rainfall observed over a network of stations. This model introduces a variable representing a local weather type at each location and establishes spatial dependence using copulas. The model accurately captures the persistence of rainfall occurrence, but not rainfall amounts.

The weather type can also be introduced as a hidden state variable that can better capture the stochastic properties of rainfall but will not necessarily be as interpretable. Ailliot *et al.* (n.d.) modelled temporal dependence using a regional (common to all locations) weather type HMM and, conditional on weather type, modeled the spatial dependence of rainfall occurrence and amount using censored, power transformed, Gaussian distributions. The marginal distributions and spatial structure of the data are well-described but the model cannot fully reproduce the local dynamics of rainfall.

Barber *et al.* (2010) described a regime-aware Auto-Regressive Hidden Markov Model (AR-HMM) and introduced a simple approximate inference method which tolerates missing data, applying it to short-term wind speed forecasting.

Markov Random Fields (MRFs) Piatkowski *et al.* (2013) proposed Spatio-Temporal Random Fields (STRFs), a discrete probabilistic graphical model based on MRFs with improved scalability. Model parameters are represented in a way that enables parameter storage compression and the optimization algorithm can be used in parallel in each graph node.

3.4.2.2 ILP based

Vaz *et al.* (2011) used an ILP engine coupled with a logic-based spatial database to predict whether a spatial polygon will catch fire based on two detailed spatial data sets: one describing the landscape mosaic and characterising it in terms of its use; and another describing polygonal areas where wildfires took place over several years in Portugal

McGovern *et al.* (2014) improved two spatio-temporal relational learning methods – the spatio-temporal relational probability tree and the spatio-temporal relational random forest – that increase their ability to learn from spatio-temporal data, applying them to hazardous weather prediction.

3.5 Regression under imbalanced domains

Our case study requires the prediction of the percentage of each Portuguese parish's area burnt yearly by wildfires (see Chapter 6). Most parishes do not burn at all most years. However, we are most interested in predicting accurately the cases where the percentage is higher than zero. This means our target variable has a very imbalanced distribution that does not correspond to our preference bias.

There is an abundance of strategies to deal with imbalanced domains, although many of them are geared towards classification. Methods focused on data pre-processing such as re-sampling, active learning and weighting the data space, have the advantage of permitting the use of standard learning techniques. Two of the simplest existing methods are the re-sampling strategies of random under- and over-sampling. In a two-class problem, the former removes a random set of majority class examples from the original data set, while the latter adds a random set of copies of minority class examples to the data. This can cause the removal of useful examples or exacerbate over-fitting problems, respectively. In both cases the ideal target distribution might not be easy to determine. However, these have still been proved to be efficient methods of dealing with the imbalance problem (Batuwita & Palade, 2010; Fernández *et al.* , 2008). A re-sampling technique geared towards regression was proposed by Torgo *et al.* (2013) working with a user-defined relevance function and threshold to determine the values to re-sample.

A multitude of measures to assess the performance of classification and regression models have been proposed. For a comprehensive survey on predictive modelling under imbalanced distributions, see Branco *et al.* (2015). Next, we discuss how to evaluate performance under this setting.

3.6 Performance metrics

3.6.1 Introduction

Most metrics to measure performance in regression tasks do not consider that prediction errors might have different costs. This becomes especially problematic when the distribution of the variable we are trying to predict does not correspond to the preference bias we assign to its domain. For example, in our case study we are most interested in instances where a high percentage of a parish's area was burnt but most parishes rarely burn at all (see Chapter 4).

The problem of finding metrics that work for imbalanced domains is already very well studied for classification tasks with many standard solutions available, some of which have inspired solutions for regression. We present a few of them next. For a more thorough discussion on this subject, see Branco *et al.* (2015).

3.6.2 Classification metrics

Considering a two-class problem, the confusion matrix of a classifier (see Table 3.1) reports **i)** the number of instances correctly classified as True Positives (TP) and True Negatives (TN), and **ii)** the wrongly classified instances as Type I errors or False Positives (FP) and Type II errors or False Negatives (FN).

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

Table 3.1: Confusion matrix for a two-class classification problem

Accuracy (Equation 3.1) and its complement error rate are standard classification performance metrics which can be extracted from this matrix.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

However, these metrics are not the most appropriate when the user prefers the least frequent class in an imbalanced domain since the minority class has a comparatively smaller impact on the results. In this case, other more appropriate metrics should be used, such as

$$\text{true positive rate (recall, sensitivity or hit rate): } TPR = \frac{TP}{TP + FN} \tag{3.2}$$

$$\text{true negative rate (specificity): } TNR = SPC = \frac{FP}{FP + TN} \tag{3.3}$$

$$\text{false positive rate (fall-out): } FPR = \frac{FP}{FP + TN} = 1 - TNR \tag{3.4}$$

$$\text{positive predictive value (precision): } PPV = \frac{TP}{TP + FP} \tag{3.5}$$

Since there is a trade-off between some of these measures and it is impractical to monitor more than one, alternative measures were proposed. The *F-measure* or F_β-score (based on Van Rijsbergen's effectiveness measure) is the harmonic mean of *precision* and *recall*, attaching β times as much importance to recall as precision (see Equation 3.6). The G-mean (Kubat *et al.* , 1998) is the geometric mean of specificity and sensitivity (see Equation 3.7).

$$F_\beta = \frac{(1 + \beta)^2 \cdot precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{3.6}$$

$$G - Mean = \sqrt{sensitivity \cdot specificity} \tag{3.7}$$

The area under a Receiver Operating Characteristic (ROC) curve (AUC) (Metz, 1978; Provost *et al.* , 1998) is yet another popular way to assess the performance of a classifier. Each point of the curve corresponds to the pair (TPR, FPR) obtained by using a different decision or threshold parameter for classifying examples.

$$AUC = \frac{1 + TPR - FPR}{2} = \frac{TPR + TNR}{2} \tag{3.8}$$

3.6.3 Regression metrics

Standard metrics for regression include Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) as defined in Equations 3.9 and 3.10, where y_i is a true value and $\hat{y}_i$ its prediction.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.9}$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.10}$$

Both of these are inadequate when dealing with imbalanced domains. One possible solution is to use the ROC space for regression (RROC space) (Hernández-Orallo, 2013) which is obtained by plotting the total under-estimation against the total over-estimation. The Area Over the Curve (AOC) is, in this case, equivalent to the error variance.

Another would be the Regression Error Characteristic (REC) curves (Bi & Bennett, 2003) that plot the accuracy and the error tolerance of a regression function which is defined as the percentage of points predicted within a certain tolerance ϵ. In this instance, the AOC corresponds to a biased estimate of the expected error. Torgo (2005) has proposed an extra dimension representing the cumulative distribution of the target variable thus allowing the study of the error's behaviour across different ranges of the target variable domain, which is important when the importance of the values in this domain is not uniform, as is the case of our target application's domain (Regression Error Characteristic Surfaces (RECS)).

Finally, we would like to introduce the precision/recall evaluation framework in the context of utility-based regression (Torgo & Ribeiro, 2007; Ribeiro, 2011). In utility-based regression the usefulness of a prediction is given by a function of both the numeric error of the prediction (given by some loss function $L(\hat{y}, y)$) and the importance of both the predicted $\hat{y}$ and true y values. The importance (relevance) is a user-specified continuous function ϕ mapping the target variable domain into a scale of relevance from 0 to 1. Ribeiro (2011) defines the notion of utility as

$$U_\phi^p(\hat{y}, y) = B_\phi(\hat{y}, y) - C_\phi^p(\hat{y}, y) = \phi(y) \cdot (1 - \Gamma_B(\hat{y}, y)) - \phi^p(\hat{y}, y) \cdot \Gamma_C(\hat{y}, y) \tag{3.11}$$

where ϕ^p is the joint relevance function, i.e., a weighted average of the relevance values of y and $\hat{y}$ where the penalisation factor p is the weight of the former, and Γ_B and Γ_C are bounded loss functions with respect to benefit and cost threshold functions.

Based on previous work by Torgo & Ribeiro (2009) and Ribeiro (2011), Branco (2014) proposed the following measures of precision and recall for regression

$$\text{precision}_R = \frac{\sum\limits_{\phi(\hat{y}_i)>t_R} (1+u_i)}{\sum\limits_{\phi(\hat{y}_i)>t_R} (1+\phi(\hat{y}_i))} \tag{3.12}$$

$$\text{recall}_R = \frac{\sum\limits_{\phi(y_i)>t_R} (1+u_i)}{\sum\limits_{\phi(y_i)>t_R} (1+\phi(y_i))} \tag{3.13}$$

These measures can be combined into an F-measure (F_β) for regression as defined by Equation 3.6. We will use this metric to evaluate our approaches in Chapter 6.

3.7 Summary

In Section 3.2, we defined the problem of spatio-temporal forecasting and the related spatial interpolation. We have identified and presented different approaches to tackle both problems in Sections 3.3 and 3.4, mentioning strategies to deal with imbalanced target domains in Section 3.5 and citing measures to evaluate them in Section 3.6. The main types of existing approaches to spatio-temporal forecasting and their references are summarised in Table 3.2.

Propositional	Pre-processing based	Luk *et al.* (2000); Bilgili *et al.* (2007); Ohashi & Torgo (2012)
	Spatio-temporal clustering based	Appice *et al.* (2013b); Li *et al.* (2003); Cheng & Wang (2008)
	Combined temporal and spatial methods	Li *et al.* (2003); Cheng & Wang (2008)
	Integrated spatial and temporal dimensions	Pace *et al.* (1998); Lindström *et al.* (2014)
Relational	Graphical models	Cano *et al.* (2004); Thompson *et al.* (2007); Ailliot *et al.* (n.d.); Barber *et al.* (2010); Piatkowski *et al.* (2013)
	ILP based	Vaz *et al.* (2011); McGovern *et al.* (2014)

Table 3.2: Propositional and relational approaches to spatio-temporal forecasting

Chapter 4

Wildfires in Portugal: A Case Study

Our motivating application is the evolution of wildfires across mainland Portugal from 1991 to 2010. Describing instances where civil parishes suffered from major wildfires and predicting to which extent they burnt yearly will be the focus of Chapters 5 and 6, respectively. In this chapter, we introduce the reader to the case study, and discuss the computation of spatial relationships in the data set.

4.1 Data set

The data for this case study (with the exception of census data used as additional background knowledge) was provided to us by Dr. João Torres, a researcher at CIBIO[1]. Details regarding data collection can be found in Torres (2014). The variable we are interested in is the percentage of burnt area for time periods of one year. The area burnt is non-cumulative, that is, even if a certain area burns multiple times during the year, it will be considered only once. Thus, the variable's domain ranges between 0% and 100%.

The background knowledge for the descriptive and predictive data mining tasks consists of explanatory variables with different temporal levels of granularity (see Table 4.1). We retrieved census data directly from the web portal maintained by the Portuguese National Statistics Institute - Instituto Nacional de Estatística (INE).[2]

[1] https://cibio.up.pt/

[2] https://www.ine.pt/

Land cover	*Eucalyptus* Natural Forest Tall scrubland Small scrubland Broad-leaved managed forest Pinewood Urban	(%)	Fixed
Terrain	Maximum altitude Mean altitude	(m)	
	Maximum slope Mean slope		
Road density	All roads Roads (>6m wide) Roads (<6m wide)		
Census data	Irrigable area Meadow area	(%)	Decennial (from 1989)
	Bovine population density Ovine population density Caprine population density	(ha^{-1})	
	Population density Population's mean age	(ha^{-1}) (years)	Decennial (from 1991)
	Population of age 65+ Housing density	(%) (ha^{-1})	Decennial (from 2001)

Table 4.1: Explanatory variables used as background knowledge for the wildfire case study.

The administrative boundaries shapefiles we used for delineating each civil parish (2014 version) are avaiable on the website for the General Directorate of Regional Planning (Direcção-Geral do Território).[3]

Since 2013, mainland Portugal is divided into civil 2882 parishes (from this point on, referred to as parishes), forming 278 municipalities which in turn constitute 18 districts. The total area of each parish is variable, ranging from 20 ha to more than 88000 ha with the median standing at about 1700 ha.

The values taken by our target variable range from 0% (when no wildfire occurred throughout the year in the parish) to an incredibly high 99.8%, and their distribution is highly imbalanced meaning that most instances record 0% of burnt area. Actually, only about a third of instances present a positive percentage of burnt area, and less

[3]http://www.dgterritorio.pt/

than a third of these (below 9% of the whole data set) amount to 5% or more of area burnt.

To better understand the problem at hand, let us focus our attention on Figure 4.1.

Subfigure (a) pictures the total area burnt in the country per year. Notably, in 2003, a year marked by an European heat wave, almost 5% of the country's area burnt at least once. Compare this with Subfigure (c), which depicts the number of parishes that suffered some wildfire (even if it corresponded to a very residual percentage of their area). Their distributions look considerably different, which is mostly due to the fact that, as previously stated, more than two thirds of instances of parishes suffering wildfires over the years burnt less than 5% per year. In fact, almost two thirds of instances of burnt parishes do not see more than 1% of their areas burnt and, therefore, do not contribute as much to the total area burnt at the end of the year.

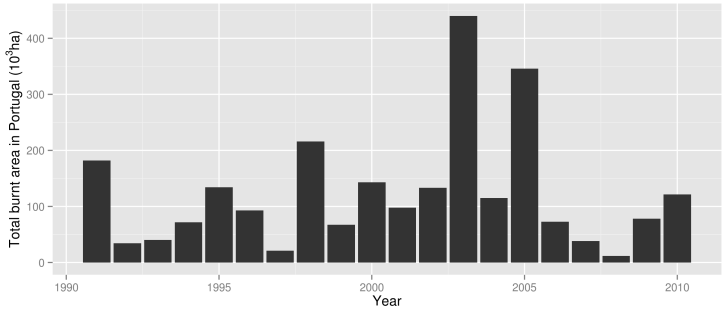
Subfigure (b) shows the number of parishes with 5% or more area burnt per year which has a distribution much more similar to that of Subfigure (a). However, there is still no direct correspondence and, once again, 2003 stands out as it resulted in a much higher burnt area relative to other years with comparable number of parishes suffering from major wildfires. This can be explained by the fact that the median area of parishes targeted by major wildfires in 2003 was much higher than the norm.

It is also important to understand that the spatial distribution of the percentage of yearly burnt area is not uniform as depicted in Figure 4.2. Some districts are comprised of parishes reaching much higher mean and maximum percentages of area burnt in the time period between 1991 and 2010.

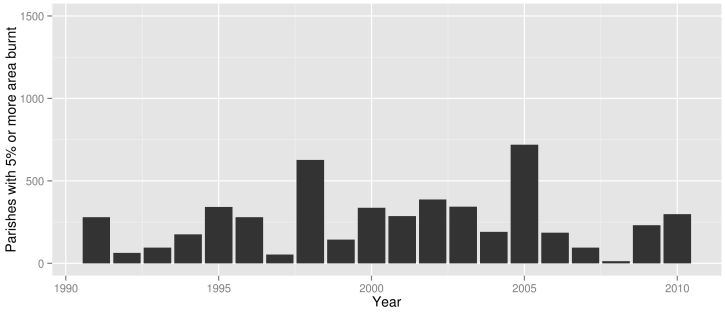
4.2 Computing spatial relationships: a common step

All the propositional and relational approaches tested require the computation of spatial relationships in the data set. We have decided to pre-compute them which is known as the eager approach (Andrienko *et al.* , 2006). Another option would have been to compute them on-the-fly as needed (lazy approach).

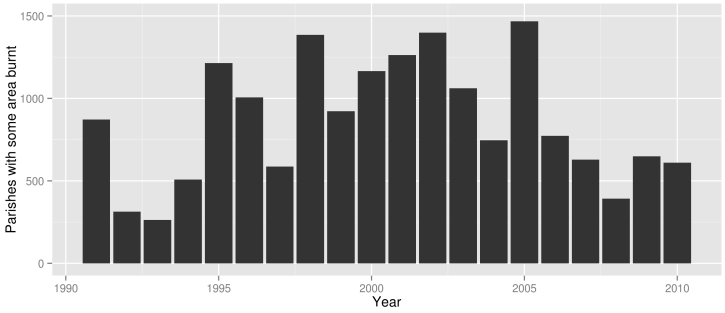
We loaded the data and shapefiles into a PostgreSQL database extended by PostGIS (Ramsey *et al.* , 2005), which adds support for geographic objects allowing location



(a) Total area burnt (non-cumulative) in Portugal per year

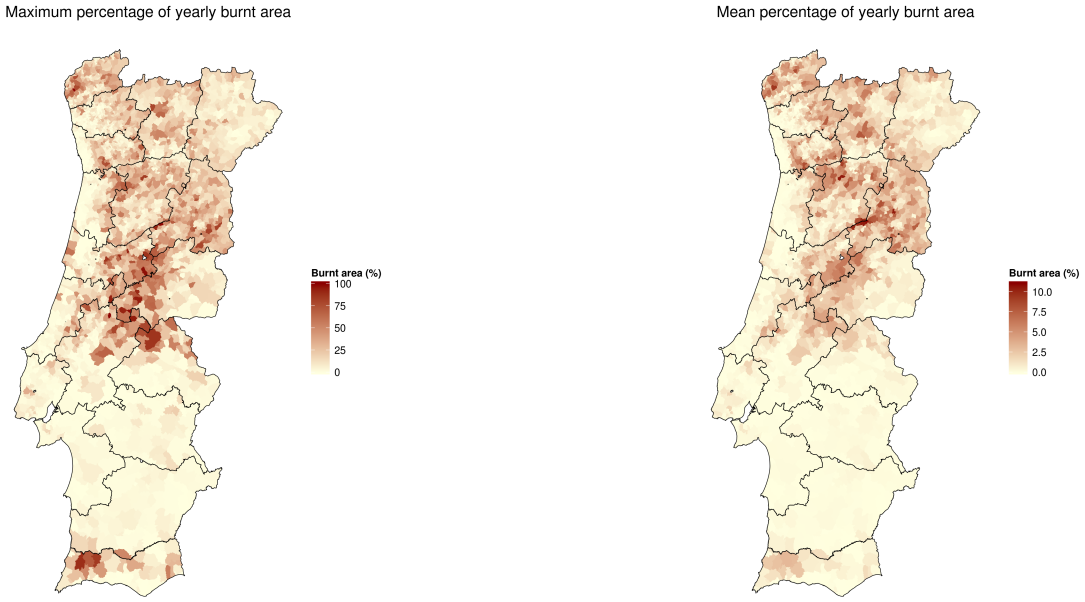


(b) Number of parishes with 5% or more of their area burnt over the years



(c) Number of parishes with more than 0% of area burnt over the years

Figure 4.1: Comparison of (a) total area burnt in Portugal, (b) number of parishes with 5% or more area burnt and (c) number of parishes with positive percentage of burnt area.



(a) Maximum burnt area (%)

(b) Mean burnt area (%)

Figure 4.2: Statistics on yearly burnt area per parish from 1991 to 2010. The boundaries depicted in black outline Portuguese districts. Dark red areas correspond to high percentages of burnt area. Note, however, that the two thematic maps have different scales.

queries to be run in SQL. PostGIS incorporates spatial data types (such as point, line, and polygon), uses multi-dimensional spatial indexing for efficient processing of spatial operations and implements a series of spatial functions for conversion, management, retrieval, comparison and generation of geometries (Ramsey & Columbia, 2005).

4.2.1 Neighbourhoods

Neighbourhoods for each parish consist of all intersecting parishes, calculated using the PostGIS function `ST_Intersect`. From this point on, we assume that all functions with prefix `ST` were provided by PostGIS.

Neighbour direction The relative direction of a neighbour in relation to a reference parish, O, was also taken into consideration. This is a less straight-forward problem, given the heterogeneous shapes presented by the parishes. Our solution, exemplified by Figure 4.3 and calculated using Code 4.1, is meant to be fast and easily computed. It revolves around the parishes' centroids, calculated with the function `ST_Centroid`.

Code 4.1: Calculation of neighbour direction

```
CREATE OR REPLACE FUNCTION CardinalDirection(azimuth float8)
RETURNS character varying AS
$BODY$SELECT CASE
  WHEN $1 < 0.0 THEN 'less_than_0'
  WHEN degrees($1) < 45.0 THEN 'N'
  WHEN degrees($1) < 135.0 THEN 'E'
  WHEN degrees($1) < 225.0 THEN 'S'
  WHEN degrees($1) < 315.0 THEN 'W'
  WHEN degrees($1) <= 360.0 THEN 'N'
END;$BODY$ LANGUAGE SQL IMMUTABLE COST 100;

SELECT Parish, Neighbour, Direction
FROM
    ( SELECT A.Parish AS Parish,
        B.Parish AS Neighbour,
        CardinalDirection(ST_Azimuth(ST_Centroid(A.geom),
            ST_Centroid(B.geom))) AS Direction
    FROM Parishes AS A, Parishes AS B
    WHERE A.gid!=B.gid and ST_Intersects(A.geom, B.geom) )
AS Temp;
```

Cartographic azimuths (`ST_Azimuth`) are calculated using the reference parish's and each neighbour's centroids. Note that the azimuth is clockwise relative to the north. Thus, azimuths in the interval $[45, 135[^{\circ}$ define eastern neighbours (C in Figure 4.3); $[135, 225[^{\circ}$, southern neighbours (E and D in Figure 4.3); $[225, 315[^{\circ}$, western neighbours; and $([315, 360] \cup [0, 45[)^{\circ}$, northern neighbours ($A$ and B in Figure 4.3). This example illustrates a problem raised by this proposal. Although O shares borders with neighbours A and (especially) E in the western direction, no neighbour is found in that direction since their centroids fall under the northern and southern space.

4.2.2 Parishes in the country's border

We have determined which parishes belong to the country's border by querying the data for parishes intersecting (`ST_Intersects`) the union of all parishes forming the country (`ST_Union`). A shapefile of Spain (Portugal's only physical neighbour) was downloaded from the GADM database of Global Administrative Areas, (version 2.0, December 2011) and loaded into our database.[4] Parishes intersecting with this spatial object were

[4] http://www.gadm.org/

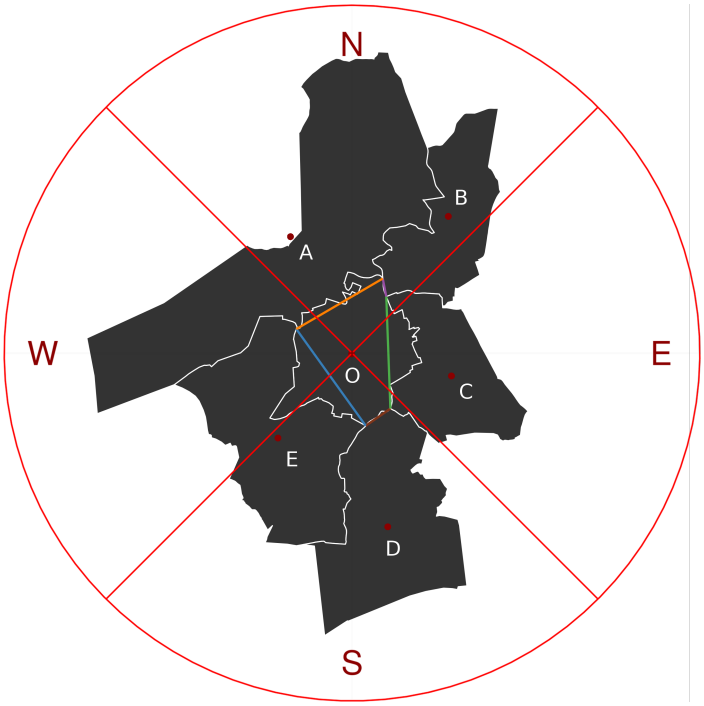


Figure 4.3: A parish's neighbourhood divided by cardinal directions. Red dots represent the parishes' centroids. The red lines centred at the reference parish's centroid divide the neighbourhood space in four directions. The remaining multi-coloured lines define the simplified borders between the reference parish, O, and each of its neighbours. According to this division, the reference parish has no western neighbours. A and B are northern neighbours; E and D, southern neighbours; C, an eastern neighbour.

noted as belonging to the border with Spain while the remaining border parishes were in contact with large bodies of water (the Atlantic Ocean and the Mediterranean Sea).

Other tools

All propositional approaches were implemented in R (R Core Team, 2015), with extensive use of packages DPLYR (Wickham & Francois, 2015) and LAZYEVAL (Wickham, 2015). Relational approaches were implemented using a mix of R and YAP PROLOG (Santos Costa *et al.* , 2010), with the ALEPH ILP system (see Section 2.4).

Most figures presented in this dissertation were generated using the R package GGPLOT2 (Wickham, 2009). Mapping of spatial objects was handled with additional packages, including RGDAL (Bivand *et al.* , 2015), SP (Pebesma & Bivand, 2005; Bivand *et al.* ,

2013) and GGMAP (Kahle & Wickham, 2013). Cleaning of spatial objects was performed using CLEANGEO (Blondel, 2015). Other R packages used will be mentioned as needed. Parallel programming in R was handled with packages FOREACH (Analytics & Weston, 2014) and DOMC (Analytics, 2014).

4.3 Summary

In Section 4.1 we introduced our case study: wildfires in mainland Portugal from 1991 to 2010. We presented the background knowledge that will be used in the next two chapters. We established that the percentage of yearly burnt area per parish is imbalanced, being concentrated at 0%, which does not correspond to our preference bias (we are most interested in parishes suffering from major wildfires). In Section 4.2 we explained the computation of spatial relationships in the data set, a pre-processing step common to all relational and propositional approaches described in the next two chapters. Finally, we mentioned the main tools we used to accomplish our goals.

Chapter 5

Describing Wildfires

In this chapter, we describe propositional (Section 5.1) and relational (Section 5.2) approaches applied to the problem of using association rules to describe instances of parishes whose percentage of burnt area reached 5% or more in a certain year. All the attributes enumerated in Table 4.1 were used for this problem. The results obtained are presented and discussed in Section 5.3.

5.1 Propositional approach

Our proposed propositional approach is pre-processing based, since the spatial and temporal aspects of the problem are considered by building indicators that express the temporal evolution of the target variable in the geographical neighbourhood of each instance. Thus, allowing the use of classical association rule learning methods to find spatio-temporal associations.

5.1.1 Pre-processing

The pre-processing stage of our approach can be divided into three steps: **i)** calculation of spatio-temporal indicators, **ii)** imputation of missing data and **iii)** categorisation of numerical attributes.

5.1.1.1 Building spatio-temporal indicators

We built a purely temporal indicator (self-indicator) by calculating the Exponential Moving Average (EMA) of past values of the target variable for the reference parish (EMA implemented in R package TTR (Ulrich, 2013)).

We also built spatio-temporal indicators considering the historic values of the target variable for neighbours located at each cardinal direction. Considering the notion of neighbourhood defined in Section 4.2, we compute the indicator for a particular direction in the following two steps. First, we calculate the EMA of the target variable for each neighbour whose centroid falls in that direction. Then, we calculate a weighted mean of these values, where the weights reflect a rough approximation of the fraction of the border exposed to each neighbour (in that direction).

The moving average of the self-indicator was calculated based on values for the previous nine years, while the directional indicators used only five years. Although these EMAs do not quite produce a conic spatio-temporal neighbourhood, our proposal did take inspiration from this concept introduced by Ohashi & Torgo and detailed in Section 3.3.2.1.

Weighing neighbours: simplified borders One way to measure the strength of connection between neighbors would be by the percentage of shared border. Unfortunately, considering the lengths of the actual intersections between a reference parish and its neighbours for this calculation is not ideal, since meandering borders can easily increase in length without proportionately increasing the exposure of the reference parish to fires originating in that particular neighbour. Instead, we consider the maximum distance between any two points of the intersection as a simplified border (resulting in the coloured lines pictured in Figure 4.3). These are easily computed with function `ST_MaxDistance` in Code 5.1 (they were pictured using `ST_LongestLine`). Finally, the weight of each neighbour is the length of its simplified border divided by the sum of the lengths of the simplified borders of all neighbours whose centroids are in that direction.

Issues Using these simplified borders to weigh the EMA of each neighbour in a certain direction raises a problem that was not previously mentioned, but is easy to imagine

Code 5.1: Calculation of simplified border length

```
SELECT Parish, Neighbour,
ST_MaxDistance(Intersection, Intersection) AS BorderLength
FROM (
    SELECT A.Parish AS Parish, B.Parish AS Neighbour,
    ST_Intersection(A.geom, B.geom) as Intersection
    FROM Parishes as A, Parishes as B
    WHERE A.gid!=B.gid AND ST_Intersects(A.geom,B.geom)
) as Temp;
```

from the example pictured in Figure 4.3. Assume that parish A was elongated in the western direction, its centroid moved to the western division. The orange simplified border between A and O exists mostly in the northern direction, but A would not be considered northern. A would now be O's only western neighbour, and E would still not be accounted for in the West even though about half of its blue simplified border with O belongs there and it corresponds to a much higher proportion of the western borders than A. Since we do not have information regarding which portion of a neighbour was burnt (whether it was close to the border with the reference parish or in the complete opposite direction), and the level of temporal and spatial granularity is not very high, these approximations are still reasonable.

The problem mentioned in Section 4.2.1 that no neighbour is found in the West in Figure 4.3 even when neighbour E is a big part of the Western border is partially solved by imputation as detailed below.

5.1.1.2 Handling missing data: imputation

Our data has a fair amount of missing values stemming from unavailability of data (about 2.8% of all background knowledge is missing for this reason), low level of temporal granularity of explanatory attributes (being only measured once or decennially and resulting in 20.6% of values missing), and by construction of our spatio-temporal directional indicators (about 6.8% of spatio-temporal indicators are unavailable due to no neighbours being found in particular directions). Each of these problems is solved in a different manner.

Values missing due to unavailability of data are filled in using independent spatial-only IDW as implemented in the R package GSTAT (Pebesma, 2004) (attributes

expressing percentages are forced into the range [0,100]). Then, missing values due to heterogeneous temporal granularity are filled in with the latest measurement (possibly previously calculated using IDW). More sophisticated ways to impute missing values by considering the temporal dimension were not used since the low level of temporal granularity of these variables lead us to have only two (or less) temporal data points for each parish.

Missing spatio-temporal indicators are filled in separately by either **i)** the value zero if the parish in question borders with the sea or the ocean or **ii)** the average of the two contiguous directions. Note that this ensures that neighbour E from the example in Figure 4.3 influences the Western indicator. There are seven parishes with neighbours in only one direction, but only one of them does not belong to the country's border. This last parish is Borba (São Bartolomeu), the smallest Portuguese parish amounting to only 20 ha. The indicators for this parish are filled in with the only value available.

5.1.1.3 Handling numerical attributes: categorisation

Association rule learning algorithms usually do not deal with numerical variables, so we have categorised them. The number of categories was set to four, and the breaks calculated using the Jenks natural breaks classification method (Jenks, 1967) as implemented in the R package BAMMtools (Rabosky *et al.* , 2015). This method seeks to minimise the variance within categories, while maximising the variance between categories. Each variable was categorised independently, with the exception of two groups of attributes whose breaks were calculated together: **i)** the self and directional spatio-temporal indicators, and **ii)** the road density variables. This process resulted in the categorisation intervals in Table 5.1

5.1.2 Modelling

Association rules were mined using R package carenR (Jorge, 2015) which interfaces with the Class project Association Rule ENgine (CAREN) (version 2.6.3), a Java based implementation of a depth-first algorithm for association rule mining.[1]

[1] carenR available at http://www.dcc.fc.up.pt/ amjorge/software/carenR/
and CAREN available at http://www4.di.uminho.pt/ pja/class/caren.html

Attribute		**Very low**	**Low**	**Medium**	**High**
Burnt Area	(%)	]0,5]	]5,20]	]20,40]	]40,100]
Self indicator Northern indicator Eastern indicator Southern indicator Western indicator	(%)	[0,2]	]2,5]	]5,10]	]10,30]
Eucalyptus		[0,6]	]6,20]	]20,40]	]40,70]
Natural Forest		[0,5]	]5,20]	]20,40]	]40,80]
Tall scrubland		[0,4]	]4,9]	]9,20]	]20,60]
Small scrubland	(%)	[0,7]	]7,20]	]20,40]	]40,80]
Broad-leaved managed forest		[0,3]	]3,9]	]9,20]	]20,50]
Pinewood		[0,10]	]10,30]	]30,50]	]50,80]
Urban		[0,10]	]10,30]	]30,60]	]60,100]
Maximum altitude	(m)	[10,300]	]300,600]	]600,1000]	]1000,2000]
Mean altitude		[2,200]	]200,400]	]400,700]	]700,1000]
Maximum slope		[4,30]	]30,40]	]40,60]	]60,100]
Mean slope		[0.6,8]	]8,10]	]10,20]	]20,40]
Road density Road (>6m wide) density Road (<6 wide) density		[0.5,500]	]500,900]	]900,2000]	]2000,5000]
Irrigable area	(%)	[0,10]	]10,20]	]20,40]	]40,100]
Meadow area		[0,10]	]10,40]	]40,200]	]200,400]
Bovine population density		[0,0.3]	]0.3,0.8]	]0.8,2]	]2,6]
Ovine population density	(ha^{-1})	[0,0.2]	]0.2,0.5]	]0.5,1]	]1,4]
Caprine population density		[0,0.06]	]0.06,0.2]	]0.2,0.4]	]0.4,2]
Population density	(ha^{-1})	[0.03,10]	]10,40]	]40,100]	]100,200]
Population's mean age	(years)	[30,40]	]40,40]	]40,50]	]50,60]
Population of age 65+	(%)	[6,20]	]20,30]	]30,40]	]40,60]
Housing density	(ha^{-1})	[0.02,7]	]7,20]	]20,50]	]50,100]

Table 5.1: Categorisation intervals for each attribute used in the wildfire case study.

This algorithm includes several pruning mechanisms to improve efficiency of both the extraction of frequent itemsets and the generation of rules (Azevedo, 2003; Azevedo & Jorge, 2010). It also allows the user to limit the attributes appearing in the antecedent and consequent of rules. Since we aim at describing parishes suffering from fires, we set the consequent to only allow percentage of burnt area in the consequent.

5.2 Relational approach

Our chosen relational approach is ILP based. The spatial and temporal dimensions are contemplated by the use of special predicates designed to explicitly express neighbourhood and temporal relations. An upgraded version of a classical association rule algorithm is then used to mine spatio-temporal associations.

5.2.1 Pre-processing

In order to use ALEPH, an ILP system written in Prolog, the data had to be converted into Prolog clauses.

5.2.1.1 Background knowledge

Each fixed attribute in Table 4.1 was converted into a binary predicate (i.e., a predicate of arity two, having two arguments) of type `numAttribute(Parish, Value)`. Attributes with temporal granularity (including burnt area percentages) were transformed into ternary predicates of type `numAttribute(Parish, Year, Value)`. Auxiliary binary and ternary predicates were designed to categorise the values of these attributes according to the boundaries on Table 5.1.

Predicates were also created to express spatial relationships computed as detailed in Section 4.2. A ternary predicate `neighbour(Parish, Neighbour, Direction)` was created where `Parish` is a reference parish, `Neighbour` takes the identifiers of each intersecting parishes, and direction takes one of four values (`north`, `east`, `south` or `west`) as defined in Section 4.2.1. A binary predicate `border(Parish, Object)` where

Code 5.2: Predicates designed to categorise background knowledge attributes

```
attribute(Parish,Category):-
        numAttribute(Parish,Value),
        % verylow in [Bound0,Bound1]
        ((Value=<Bound1, Category=verylow);
        % low in ]Bound1,Bound2]
        (Value>Bound1, Value=<Bound2, Category=low);
        % medium in ]Bound2,Bound3]
        (Value>Bound2, Value=<Bound3, Category=medium);
        % high in ]Bound3,Bound4]
        (Value>Bound3, Categ=high)).

    attribute(Parish,Year,Category):-
        numAttribute(Parish,Year,Value),
        ((Value=<Bound1, Category=verylow);
        (Value>Bound1, Value=<Bound2, Category=low);
        (Value>Bound2, Value=<Bound3, Category=medium);
        (Value>Bound3, Category=high)).
```

`Object` takes the value `spain` or `water` was generated to identify parishes in the country's border (see Section 4.2.2).

Files with grounded (i.e., containing no variables) facts (i.e., clauses with no bodies) expressing these relations were generated using a program written in Prolog for converting them from CSV format.

Auxiliary predicates

The predicates mentioned above (with the exception of `border/2`) are not used directly to build spatio-temporal association rules. Instead, they are used to define auxiliary predicates that categorise them or to better express temporal and spatial relationships.

Categorisation

Numerical attributes in the background knowledge are categorised by resorting to auxiliary predicates of type `attribute(Parish,Category)` or `attribute(Parish, Year, Category)`, as exemplified in Code 5.2.

Past fires

The number of years past since a wildfire last affected a certain parish is determined by a pair of predicates: `yearsSinceFireLE(Parish, Year, TimeDist)`

Code 5.3: Predicates expressing temporal distance to last occurrences of wildfire

```
yearsSinceFireLE(Parish,Year,TimeDist) :-
    var(TimeDist),!,
    lastFire(Year,YearLastFire),
    burntArea(Parish,YearLastFire,V),!, % V>0
    TimeDist is Year-YearLastFire.

yearsSinceFireLE(Parish,Year,TimeDist) :-
    lastFire(Year,YearLastFire),
    burntArea(Parish,YearLastFire,V),!, % V>0
    ThisTimeDist is Year-YearLastFire,
    ThisTimeDist=<TimeDist.

lastFire(Year1,Year2) :-
    between(1,20,Delta),
    Year2 is Year1-Delta.
```

and `yearsSinceFireGE(Parish, Year, TimeDist)`, which are true if by year `Year` the parish `Parish` has suffered a wildfire `TimeDist` or less years ago, or `TimeDist` or more years ago, respectively. Note that ALEPH always calls these predicates (as part of the procedure described in Section 2.4) with constants for `Parish` and `Year` obtained from the example being tested. Each of these predicates is defined by two clauses (see Code 5.3).

1. The first clause only applies if `TimeDist` is a variable. If it is, then the predicate is being called from the *saturation* step mentioned in Section 2.4 and we simply calculate the distance to the last wildfire before `Year` that happened in that `Parish`, and assign that value to `TimeDist`.
2. Otherwise, the predicate is being called from the *reduction* step, and we compare the assigned value of `TimeDist` with the distance (in years) of the last wildfire before `Year` that occurred in the particular `Parish` of that example. If the distance is lesser than `TimeDist`, then `yearsSinceFireLE/3` is true; if it is greater than `TimeDist`, then `yearsSinceGE/3` is true; if they are equal, then both are true.

Spatial relationships

Two spatial predicates are defined to deal with neighbourhood information. If we used the predicate mentioned in the previous section directly, we would have

no control over the values taken by `Neighbour`, so we could end up with a recursive situation where a reference parish itself is considered a neighbour of its neighbour in the same association rule. The predicate `fixedNeighbour(Parish, Neighbour)` accesses a global variable to check if a new prospective neighbour was already considered. The predicate `neighbourDirection(Parish, Neighbour, Direction)` depends on this new predicate to define pairs of `Parish`(es) and their `Neighbour`(s) and on the previously defined `neighbour/3` predicate to determine their `Direction`.

5.2.1.2 Examples

Besides requiring files with background knowledge, ALEPH usually also needs a file identifying positive examples, and a file containing negative examples. Although we do not need to categorise the explanatory attributes (thanks to the auxiliary predicates we discussed in the previous section), we still need to categorise the attribute we want to focus on for ALEPH to work. We will be using it in association rule mode, so we only need a file with positive examples for the learning step (since the search is based only on support (Equation 2.1), ignoring other metrics such as confidence). For this task, the positive examples are represented by a predicate `burntArea(Parish, Year, Category)` where `Parish` and `Year` identify instances where the percentage of burnt area is within the ranges specified in Table 5.1 for `low`, `medium` and `high` percentage of burnt area, and `Category` takes the respective categorical value.

A file containing positive examples was also generated by a Prolog program, converting data from a CSV file.

5.2.2 Modelling

In order to find associations, we used ALEPH's command, `induce/0`, which invokes the procedure described in Section 2.4. The default search strategy used by `induce/0` is `bf` which enumerates shorter clauses first, but this can be changed. To tackle the problem of association rule learning, the search strategy was set to `ar`, which implements a simplified version of WARMR. The number of layers of new variables, `i`, was set to 3 and the number of nodes, `nodes`, to 7500.

5.2.2.1 Modes and determinations

ALEPH needs the specification of mode types. So, for example, we specify

```
:- modeh(*,burntArea(+parish,+year,#category)).
```

The `h` means that we only allow the predicate expressing burnt area at the head of clauses. The + before arguments `parish` and `year` determine that when a literal with predicate symbol `burntArea` appears in a hypothesised clause, the corresponding arguments should be *input* variables of type `parish` and `year`, and the # before the `category` argument specifies that it should take a *constant* value (`low`, `medium` or `high`, in this case).

Modes for predicates of categorised explanatory attributes are defined with `modeb` (`b` for body of the clause).

```
:- modeb(*,attribute(+parish,+year,#category)).
:- modeb(*,attribute(+parish,#category)).
```

The first one represents cases with temporal granularity and the the second, cases with fixed values. Once again, the + specifies input variables and the # specifies that `category` should appear as a constant in clauses.

The modes of spatial predicates are defined the following way:

```
:- modeb(*,fixedNeighbour(+parish,-parish)).
:- modeb(*,neighbourDirection(+parish,-parish, #direction)).
:- modeb(*,'border'(+parish,#object)).
```

The − before the second parish in the first two modes means they are *output* variables.

Any input variable of a certain type (`parish`, `year`, `category`, `direction` or `object`) in a body literal appears as an output variable of that type in a body literal appearing before it, or as an input variable of that type in the head of the clause. Any output variable of a certain type appearing in the head of a clause, appears as an output variable of that type in some body literal. As previously mentioned, variables marked as constants only appear as ground terms.

Determination statements are also needed to declare the predicates that can be used to construct clauses. They take the form:

```
:- determination(HeadPredicate/Arity, BodyPredicate/Arity).
```

We have included a determination for all the auxiliary and spatial predicates discussed so far, with `burntArea/3` as the only head predicate.

5.3 Experimental analysis

5.3.1 Experimental setup

Both propositional and relational algorithms were applied to the data set consisting of all instances where a parish's area burnt more than 5%, with the same background knowledge categorisation. As mentioned above, the percentage of burnt area was divided in three categories: low, corresponding to values between 5% and 20%; medium, from 20% to 40% and high, for values above 40%. For a representation of the distribution of the categories, see Figure 5.1. We only searched for rules with this attribute in the consequent.

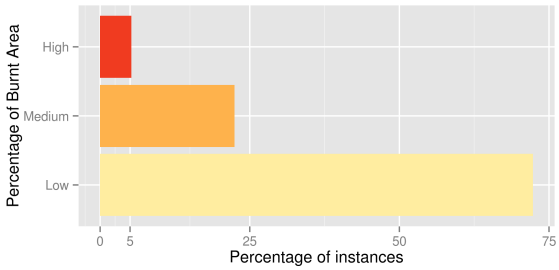


Figure 5.1: Distribution of (categorised) burnt area percentage for instances with more than 5% of area burnt.

Since ALEPH cannot enforce confidence boundaries during the process of association rule learning (in mode `ar`), we set the minimum confidence in CARENR to 0, so the comparison on the number of rules is fairer. Clause length/number of antecedents in the rule was limited to 5 in both tools.

We first set the minimum support to 0.01 (or 1%, corresponding to about 20% of the class with the least amount of cases). Then, we vary the minimum support in order

to determine how that affects the number of rules found and the time taken by the algorithms.

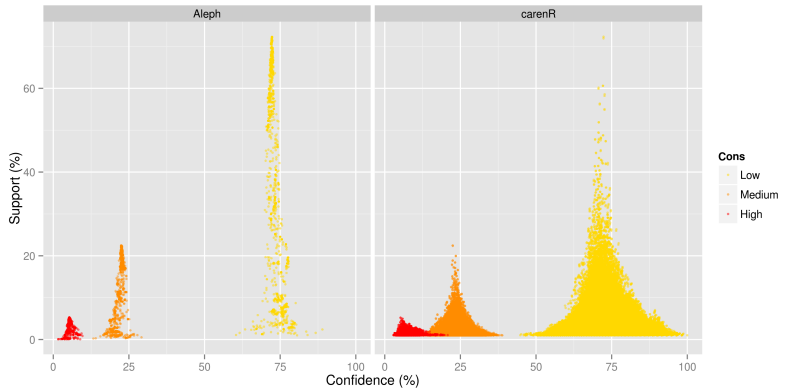
5.3.2 Results and discussion

When examining the results, it is important to keep in mind that the propositional approach required the effort of categorising the data, while the relational approach did not necessarily need categorisation of the data in order to work. However, if the categorisation step was skipped, the effort of building predicates that can deal with numerical data (see Section 6.2.1.1) would be required. The difficulty and time consumed by this alternative step depends on one's experience with ILP. Note that the propositional approach we used automatically calculates performance metrics of the rules, while the relational approach demands a programming effort in order to calculate any other metric besides support. When analysing the results, keep in mind that only instances where parishes had more than 5% of area burnt are being considered.

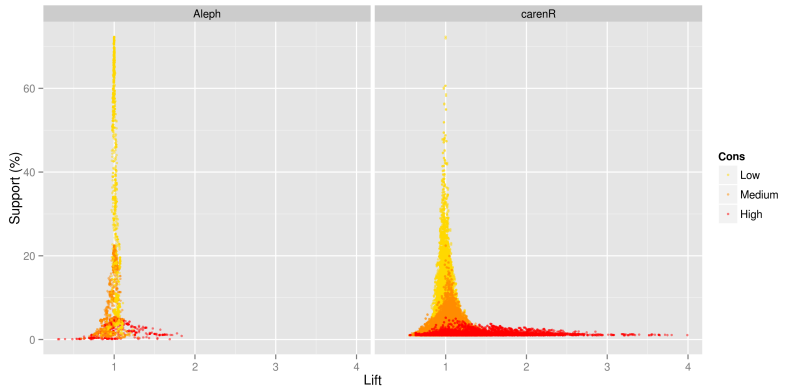
5.3.2.1 Fixed minimum support

First, we focus on the results obtained for a fixed minimum support of 0.01 (or 1%). Figure 5.2a plots the support of the rules against their confidence. The colour of the points represents the category of wildfires found in the consequent, which allows us to see how these evaluation metrics behave by category. From this graph, it is clear that **a)** ALEPH discovers a much lower total number of rules, and **b)** while ALEPH finds more rules with higher support, their confidence and lift falls within a narrower range than that of rules discovered by CARENR. The lower number of rules found by ALEPH was to be expected given that its learning algorithm strives to find a minimum amount of rules to cover each instance at least once, while CARENR tries to find every association rule within imposed restrictions such as minimum support. Besides, the number of rules searched by ALEPH was constrained by the nodes setting which limits the number of rules explored.

We present in Table 5.2 a few strong rules (i.e., rules with high support and confidence) that also present a good lift obtained with each approach. These rules shed some light on the characteristics of parishes with a significant percentage of burnt area. Note that



(a) Support vs. Confidence



(b) Support vs. Lift

Figure 5.2: Performance distribution of discovered association rules. Subfigure (a) pictures the support of rules against their confidence; Subfigure (b) represents the support of rules against their lift. Each point represents a rule, its colour indicating the percentage of burnt area found in the consequent.

the selected propositional rules include the spatio-temporal indicators we have built and the relational rules include auxiliary spatial and temporal predicates, which is an indication that the pre-processing to include the two dimensions was worthwhile in both cases.

For the selected rules with low percentage of burnt area as consequent, we calculated the spatial coverage, i.e., the number of years for which only the antecedent holds subtracted from the total number of years that the whole rule holds for each parish. Although the two rules are not necessarily representative of others found by each approach, it is still interesting to compare the different spatial distributions of this simple metric calculated for them. From Figure 5.3, it is apparent that while the propositional rule presents lower values of coverage across a larger number of parishes

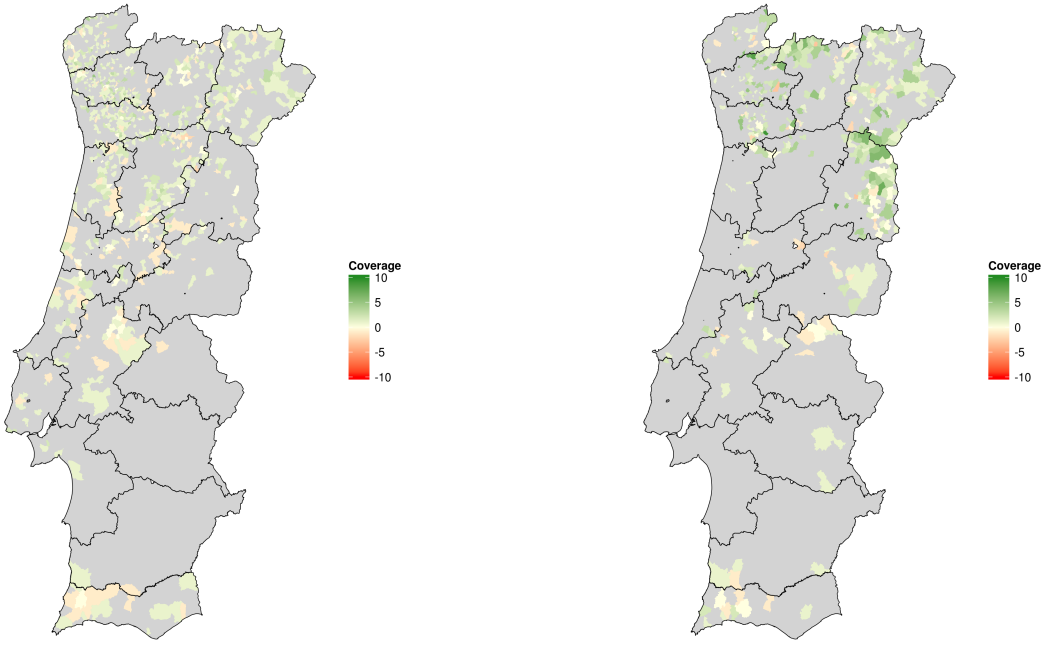
Approach	Antecedent	Consequent	Supp (%)		Conf (%)	Lift
Propositional	Self-indicator=Very Low, Caprine dens.=Very Low, Meadow area=Very Low	Burnt area=Low	15	(21)	80	1.1
	East-indicator=Low, Broad-leaved man. forest=Very Low, Eucalyptus=Very Low, Bovine dens.=Very Low	Burnt area=Medium	5.6	(25)	28	1.3
	Self-indicator=Very Low, Housing dens.=Very Low, Ovine dens.=Very Low, Road (>6m) dens.=Very Low	Burnt area=High	1.1	(20)	16	3.1
Relational	fixedNeighbour(Parish, Neib), yearsSinceFireLE(Neib, Year, 8), pinewood(Parish, verylow).	burntArea(Parish, Year, low)	18	(25)	78	1.1
	neighbourDirection(Parish, Neib1, west), neighbourDirection(Neib1, Neib2, south), tallScrubland(Neib2, medium).	burntArea(Parish,Year,medium)	6.6	(29)	24	1.1
	neighbourDirection(Parish, Neib, east), pinewood(Neib, high).	burntArea(Parish,Year,high)	1.4	(28)	9.3	1.8

Table 5.2: Selected association rules found using a propositional and a relational approach, their support (supp), confidence (conf) and lift. The support value between parenthesis indicates the support within the respective burnt area category.

in the general north and eastern centre of the country, the relational rule achieves much higher values of this metric in smaller regions concentrated in the north and northeast. Note that the maximum possible value of coverage would be twenty (the time-span of our data set), and this would only occur if a low percentage of a particular parish burnt every year and the antecedent of the rule also held true every year.

5.3.2.2 Varying minimum support

Figure 5.4 shows the results obtained using different values of minimum support. Figure 5.4a plots the number of rules found by each approach. Unsurprisingly, the number of rules found by CARENR increases almost exponentially with the decrease of minimum support. This is explained by the combinatorial nature of the expansion of frequent itemsets. In contrast, the number of rules found by ALEPH is much more stable. Figure 5.4b pictures the time taken by each approach. ALEPH is slower than CARENR across the board, its running time increasing with the increase of minimum support, always spending more time than its counterpart. CARENR shows the opposite tendency, decreasing the time spent very rapidly from the minimum support of 1% to



(a) Self-indicator=Very Low,
Caprine dens.=Very Low,
Meadow area=Very Low
⇒ BurntArea=Low

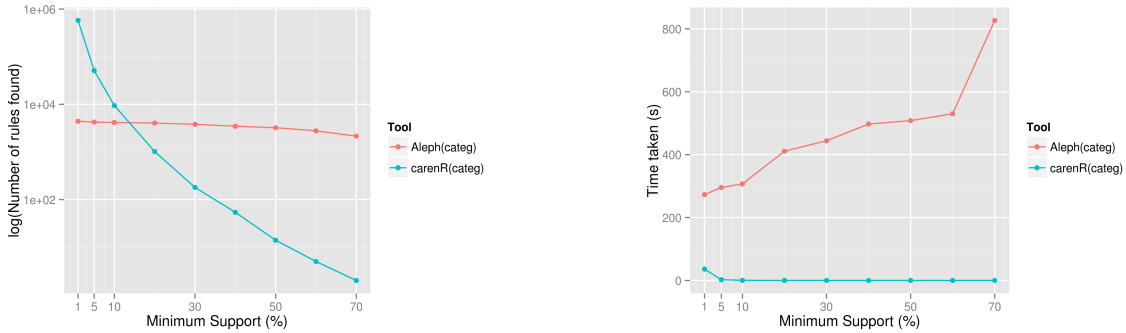
(b) fixedNeighbour(Parish, Neib),
yearsSinceFireLE(Neib, Year, 8),
pinewood(Parish, verylow)
⇒ burntArea(Parish, Year, low)

Figure 5.3: Spatial coverage of example association rules obtained using a) a propositional and b) a relational approach. Spatial coverage is defined as the number of years for which only the antecedent holds subtracted from the total number of years that the whole rule holds for each parish. Green values correspond to high values of coverage; red values to low. Grey areas picture parishes where the antecedent never holds.

10% (which, again, is explained by the combinatorial nature of the frequent itemset expansion) and maintaining very low running times from that point on.

5.4 Summary

We categorised the data set in order to apply a propositional (using CARENR) and a relational approach (using ALEPH, an ILP system) to association rule mining considering only instances where wildfires burnt more than 5% of a parish's area.



(a) Number of rules found by each tool with varying minimum support.

(b) Time taken by each tool to find rules with varying minimum support.

Figure 5.4: Information on the rule mining procedure with association rule learning tools CARENR and ALEPH (both using categorised data), for different values of minimum support. Note that the number of rules found is being shown with a logarithmic scale.

For the propositional approach, we built spatio-temporal indicators inspired by the work of Ohashi & Torgo (2012) and imputed missing data. For the relational approach, we converted the data into Prolog and defined auxiliary predicates to express spatial and temporal relationships. Numerical data was categorised in both cases.

Both approaches were capable of finding strong rules with low minimum support. With the increase of minimum support, the number of rules found and time spent by CARENR decreases almost exponentially, while the number of rules discovered by ALEPH keeps fairly constant with a slight increase in running time.

Chapter 6

Predicting Wildfires

In this chapter, we describe propositional (Section 6.1) and relational (Section 6.2) approaches applied to the problem of predicting the percentage of area burnt yearly by wildfires in Portuguese parishes. Results obtained are presented and discussed in Section 6.3.

6.1 Propositional approach

Our propositional approach to the predictive task is pre-processing based, as was the approach to the association rule learning problem. The idea is, once again, to rely on the construction of spatio-temporal indicators to contemplate the spatial and temporal dimensions. An extra step will be needed to deal with the imbalanced domain of the target variables. Then, standard out-of-the-box learning algorithms can be used.

6.1.1 Pre-processing

Pre-processing steps for this task are very similar to the steps taken to find association rules. That is, we have built spatio-temporal indicators as detailed in Section 5.1.1.1 and imputed missing data as described in Section 5.1.1.2. However, there was no need to categorise numerical variables, since the learning algorithms we use cope well with them.

At this point, we have already transformed our problem into a standard multiple regression problem, and we intend on using standard out-of-the-box learning algorithms to produce our predictions. However, since we are working with an imbalanced domain, and we are most interested in major wildfires, our methodology needs to be adapted to focus on these instances.

6.1.1.1 Handling an imbalanced domain: re-sampling

Several pre-processing techniques exist to tackle the problem of an imbalanced domain. Re-sampling techniques are quite effective and have already been proposed for both classification and regression (see Section 3.5). They also have the advantage of working equally well with numerical and categorical attributes and target variables. Besides balancing the domain, under-sampling also reduces the dimensionality of the data set, mitigating scalability issues. We used the under-sampling technique for regression proposed by Torgo *et al.* (2013) as implemented in R package UBL (Branco *et al.* , 2014). This method automatically calculates the amount of re-sampling needed to balance the domain, but it requires the specification of a relevance function for the target's variable domain and a threshold of relevance above which instances are considered relevant. After discussion with our domain expert (Dr. João Torres), we have settled on the function shown in Figure 6.1 with a relevance threshold of 0.5, corresponding to 5% of burnt area. This relevance function will also be used for performance evaluation in Section 6.3.1.

6.1.2 Modelling and post-processing

The following learning algorithms were used:

Random Forest (RF) Random Forest is an ensemble learning algorithm proposed by Breiman (2001). The original implementation was used, as available in R package randomForest (Liaw & Wiener, 2002).

Support Vector Regression machines (SVRs) Support Vector Machines (SVMs) were developed by Cortes & Vapnik (1995) for binary classification. A version of this method was later proposed for regression by Drucker *et al.* (1996). We used

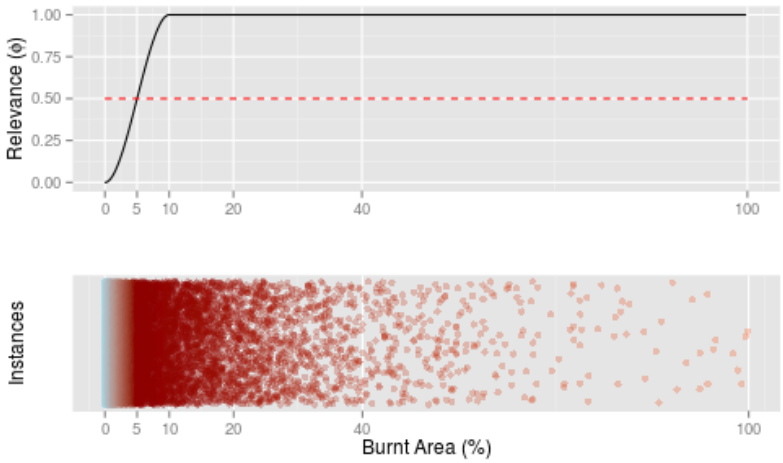


Figure 6.1: Relevance function, ϕ, used for re-sampling technique and performance metrics adapted to regression under imbalanced domains. Note that for a percentage of burnt area below 5%, the relevance is zero. Below the function, a representation of instances across the domain (most are concentrated at 0% of burnt area).

the interface with the LIBSVM implementation provided by R package E1071 (Meyer *et al.* , 2014) which refers to this algorithm as ϵ-regression.

The predictions produced were forced into the allowed range for the domain. That is, predictions below 0% were changed to 0%; predictions above 100%, to 100%.

6.2 Relational approach

For the relational approach to prediction, we decided to use ALEPH to search for clauses that we can then propositionalise, that is, convert each clause into a binary attribute with value 1 if it is true for the instance, and 0 if it is not. This kind of approach has been used before in diverse contexts, including spatial classification (Appice *et al.* , 2005; Ceci & Appice, 2006). Although ALEPH searches for clauses that are optimized to perform well for binary classification, we hypothesise that standard propositional prediction models trained with these Boolean attributes and a numerical variable as target (in our case, the percentage of burnt area) can yield an effective regression model. That is, standard regression algorithms such as SVRs can successfully use the binary features representing interesting clauses as predictors for approximating the numerical values of a target variable.

6.2.1 Pre-processing

The pre-processing procedure for this task includes most of the steps discussed in Section 5.2.1, excluding the auxiliary categorisation of attributes. Since the search for clauses is also part of this step, modes and determinations very similar to the ones mentioned in Section 5.2.2.1 are used. However, there are a few key differences in the clause search process, and the added step of propositionalisation.

6.2.1.1 Background knowledge

Once again, we do not use the numerical predicates defined in Section 5.2.1.1 directly. Instead, we define auxiliary predicates that can deal with numerical attributes without categorising them. The auxiliary predicates described below are added to the temporal and spatial auxiliary predicates described in Section 5.2.1.1.

Auxiliary predicates

In order to deal with numerical attributes, predicates similar to `yearsSinceFireLE/3` and `yearsSinceFireGE/3` (Code 5.3) are built for each explanatory attribute represented by a non-spatial ternary predicate defined in the background knowledge. That is, for each attribute, there is a predicate `attributeLE(Parish, Year, Value)` and `attributeGE(Parish, Year, Value)` meaning that the value of `attribute` in `Parish` measured in or before `Year` is lesser or equal (or greater or equal) to `Value`. Again, these are defined by two clauses but instead of calculating and assigning (or comparing) distances in time, they just assign or compare values directly (see Code 6.1). Another important difference is that, since we will be working with training and test sets, all the predicates of type `attributeLE(Parish, Year, Value)` and `attributeGE(Parish, Year, Value)` make sure not to include data beyond a pre-defined training size. This verification is also included in versions of `yearsSinceFireLE/3` and `yearsSinceFireGE/3` for this task. The two clauses work the following way:

Code 6.1: Predicates designed to handle numerical attributes without categorisation

```
attributeLE(Parish,Year,Value) :-
    var(Value),!,
    lastMeasure(Year,YearLastMeasure),
    numAttribute(Parish,YearLastMeasure,Value),
    YearLastMeasure>=MinBK.

attributeLE(Parish,Year,Value) :-
    lastMeasure(Year,YearLastMeasure),
    numAttribute(Parish,YearLastMeasure,ThisValue),!,
    ThisValue =< Value,
    YearLastMeasure>=MinBK.

lastMeasure(Year1,Year2) :-
    between(0,20,Delta),
    Year2 is Year1-Delta.
```

1. If `Value` is a variable, assign it a value by unification with the last measurement of the attribute in question. Verify that the measurement was taken after lower bound for the data subset.

2. Otherwise, compare the constant value previously assigned to `Value` with the value of the last measurement of `attribute` for that particular example (defined by a specific `Parish` and `Year`) in order to assess if `attributeLE/3` and/or `attributeGE/3` are true. Verify that the measurement was taken after lower bound for the data subset.

The same type of predicate is defined for the binary predicates (`attributeLE(Parish, Value)` and `attributeGE(Parish, Value)`), but with no concern for the time of the measurements since the values are considered fixed across time.

6.2.1.2 Examples

For this problem, we work with not only positive (as in the previous chapter) but also negative examples. Positive examples are considered to be the ones that previously qualified for the predicates `low`, `medium` and `high`; negative examples correspond to percentages of burnt area below 5%. The predicate on the head of clauses will then be `burnt(Parish, Year)`, which will be reflected on the Aleph modes.

6.2.1.3 Clause search and selection

The ALEPH search strategy is changed back to default (instead of `ar` mode for association rule learning), while the maximum number of layers of new variables and nodes stay at 3 and 7500, respectively. However, we do not use `induce/0` to search for a theory. Instead, we use our own method, and change the cost (used for generalisation on the reduction step) to the F-measure (Equation 3.6). The main differences between `induce/0` and our method is that **i)** we do not try to find a theory covering all examples, instead using random examples as seeds and skipping the step of redundancy removal, and **ii)** we store each and every clause that has been the best so far for each saturated example according to the F-measure, our chosen metric.

We used this method with a set of different values of β for the F-measure, trying 60 random seed examples for each $\beta \in \{0.75, 0.9, 1.0, 1.1, 1.25\}$. Note that this requires that the clause found to be the best so far be reset every time we change the value of β. By varying β, we hope to add some diversity to our discovered clauses, while keeping it around 1.0 assigns similar importance to their precision and recall.

6.2.1.4 Propositionalisation

After finding the clauses, a Prolog program converts the stored clauses into a CSV file with rows corresponding to instances and columns to clauses. This program is capable of filtering out clauses that are exact repetitions of others, but cannot filter clauses that are even extremely similar except for some minor change in a constant numeric literal, for example.

6.2.2 Modelling and post-processing

The same modelling techniques (Random Forest and Support Vector Regression machines) and re-sampling methodology used for the propositional approach (see Section 6.1.2) are applied to the data obtained after the pre-processing steps mentioned above. The predictions produced by the models are then forced into the domain range, as also described in Section 6.1.2 for the propositional approach.

6.3 Experimental analysis

6.3.1 Experimental setup

The main goal of our experimental setup is to compare the predictive performance of a propositional approach with a relational approach to a spatio-temporal regression problem.

As discussed in Section 3.6.3, standard metrics like MSE are not well equipped to deal with domains where the user preference bias does not correspond to the target domain distribution as in our case (see Section 4.1). Therefore, we evaluate our methodologies using precision_R and recall_R as defined by Equations 3.13 and 3.12 where the utility depends on the relevance function pictured in Figure 6.1 with a balanced penalisation factor of 0.5, as well as the resulting F_1-measure (β set to 1 in order to value precision and recall equally). A graphical representation of the possible values taken by utility depending on the quality of predictions can be found in Figure 6.2.

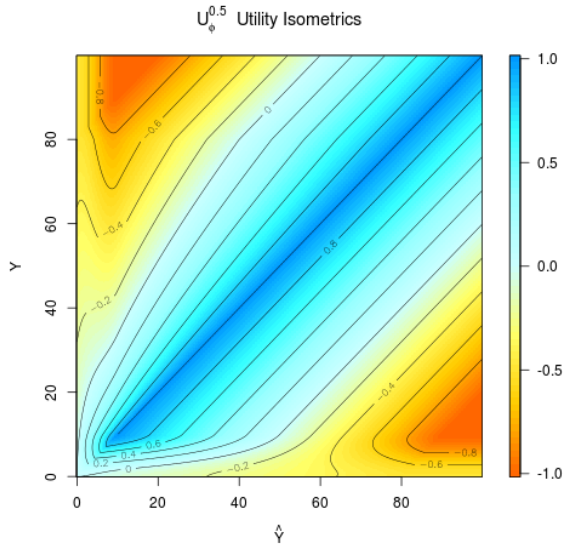


Figure 6.2: Contour map of regression utility. The x-axis shows predicted values ($\hat{Y}$) for true values (Y) in the y-axis. Colouring and contour lines map the utility of each prediction as defined by the relevance function ϕ pictured in Figure 6.1 (with a penalisation factor of 0.5, meaning the utility is symmetric about the $Y = \hat{Y}$ line).

In order to obtain reliable estimates of these metrics, we divided the data set in several training and test sets, and averaged them over 10 repetitions. Since the data is

temporal in nature, it is important to respect the natural order of the data, therefore, we always test our models in future data. We set each training set to consist of data for a stretch of eight years (23056 instances), and the corresponding test set to consist of data for the next three years (8646 instances). Thus, the first training set starts in 1991 and ends in 1998, with the corresponding test set starting in 1999 and ending in 2001; while the tenth and last training set starts in 2000 and ends in 2007, with the corresponding test set starting in 2008 and ending in 2010 (see Figure 6.3).

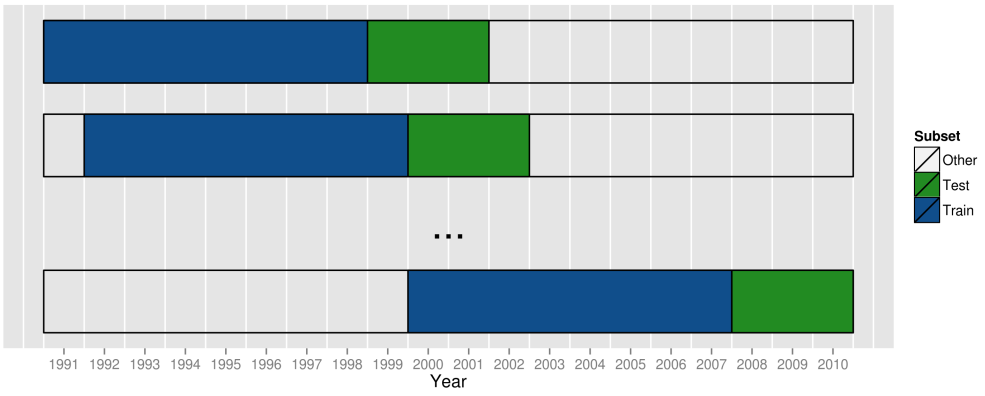


Figure 6.3: Graphical representation of the training and testing sets.

The experiments were carried out using the R package PERFORMANCEESTIMATION (Torgo, 2014), and repeated with and without the under-sampling step detailed in Section 6.1.2.

6.3.2 Results and discussion

Tables 6.1 and 6.2 summarise the results obtained and execution time for each setup described above. The difference in predictive performance between the propositional and relational approaches is rather small, both obtaining good results when the under-sampling step is performed. In every case, recall is at least somewhat higher than precision, which is interesting for this application given that the impact of wildfires often out-weights the costs of fire prevention, i.e., false negatives usually prove more costly than false positives. Note that the relational approach obtains good results despite the fact that the binary features were extracted by propositionalising clauses optimised for a two-class classification problem. This confirms our hypothesis that it is possible to build a good regression model by applying a standard algorithm to a table

with a numerical target variable and Boolean features optimised for classification (of the categorised target variable).

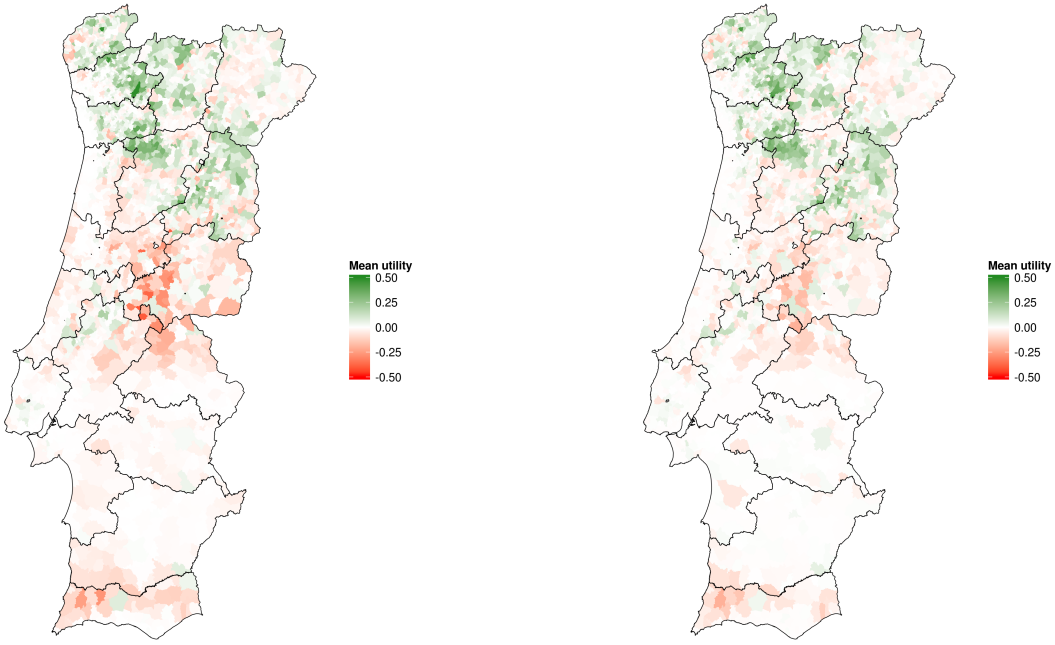
				Under-sampling	
		RF	SVR	RF	SVR
Propositional	$\text{Precision}_\text{R}$	0.3 ± 0.1	0.2 ± 0.4	*0.7* ± 0.1	0.6 ± 0.2
	Recall_R	0.69 ± 0.03	0.74 ± 0.06	*0.80* ± 0.02	0.78 ± 0.05
	$\mathbf{F_1\text{-measure}_R}$	0.4 ± 0.1	0.3 ± 0.4	**0.72** ± 0.07	0.6 ± 0.1
Relational	$\text{Precision}_\text{R}$	0.22 ± 0.09	0.008 ± 0.009	*0.6* ± 0.1	0.5 ± 0.1
	Recall_R	0.71 ± 0.03	0.6 ± 0.2	*0.80* ± 0.03	0.76 ± 0.02
	$\mathbf{F_1\text{-measure}_R}$	0.3 ± 0.1	0.02 ± 0.02	**0.67** ± 0.08	0.55 ± 0.07

Table 6.1: Average and standard deviation of results obtained with various setups for a regression task with spatio-temporal data.

				Under-sampling	
		RF	SVR	RF	SVR
Propositional	Pre-processing time (s)	1.4e-3	1.4e-3	1.4e-3	1.4e-3
	Training time (s)	2.8e-2 ± 2e-3	1.1e-2 ± 7e-3	2.3e-3 ± 8e-4	*3e-4* ± 1e-4
	Prediction time (s)	*1.5e-4* ± 1e-5	1.1e-3 ± 5e-4	8.0e-5 ± 8e-6	4.3e-4 ± 7e-5
	Total time (s)	3.1e-2 ± 3e-3	1.4e-2 ± 7e-3	3.7e-3 ± 9e-4	**2.2e-3** ± 2e-4
Relational	Pre-processing time (s)	1.7	1.7	1.7	1.7
	Training time (s)	5e-2 ± 3e-2	3e-2 ± 1e-2	5e-3 ± 1e-3	*3.0e-3* ± 4e-4
	Prediction time (s)	1.7e-4 ± 5e-5	6e-3 ± 1e-3	*1.3e-4* ± 4e-5	2.0e-3 ± 5e-4
	Total time (s)	1.75 ± 3e-2	1.73 ± 1e-2	1.706 ± 2e-3	**1.7049** ± 9e-4

Table 6.2: Average and standard deviation of time taken by various setups for a regression task with spatio-temporal data. The pre-processing time shown for propositional approaches includes time spent calculating spatio-temporal indicators and imputing missing data for propositional approaches; for relational approaches, it includes time spent finding clauses using the Aleph system and converting them to propositional form. In both cases, the time shown is the average time taken per observation.

Moreover, by examining the results, it becomes clear that under-sampling not only greatly improves the predictive ability of the models, but also decreases the training time needed to build the model. On average, the under-sampling methodology used reduces the training sets to 20% of their original size, and almost doubles the F_1-measure obtained by the regression models. This is a significant difference, indicating that the methodology chosen to deal with the fact that our preference bias did not correspond to the distribution of the percentage of burnt area over its domain has accomplished its goal.



(a) Propositional approach.

(b) Relational approach.

Figure 6.4: Mean prediction utility per parish averaged over ten repetitions and across ten test sets for a propositional and a relational approach. Both represent the results obtained by under-sampling the training set and using RF to model the data. The colour green indicates that, on average, the predictions were accurate and useful to the parish, while red evidences difficulty in providing good predictions.

Finally, RF usually perform at least slightly better than SVR, the difference being especially noticeable when considering the precision of the relational approach without under-sampling. This indicates that the SVR might be suffering with the high dimensionality of the data coupled with a higher dimensionality of the feature space. Note that the propositional feature space is the same for every training set, consisting of 29 explanatory variables, while the relational approach uses a variable number of features per training set (depending on the number of clauses found by our method) ranging from 39 to 89 binary features, with the median falling on 67 – more than double the amount of features used by the propositional approach. However, it is possible that a tuning effort would improve the results obtained with SVR.

Figure 6.4 shows the mean utility obtained for each parish by the best combination of methodologies for both propositional and relational approaches. In both cases, this corresponds to under-sampling of the training set followed by RF modelling. Both

approaches present similar spatial distribution of utility, performing worse in the countryside Centre (particularly in the mountainous regions of Castelo Branco district) and coastal South (especially in the also mountainous Monchique municipality of the Algarve district) than in the North of the country. Compare with Figure 4.2 for a notion of the target variable's spatial distribution. The utility obtained seems to be strongly (and positively) correlated with the average historic and neighbourhood values of the target variable itself, which is not too surprising considering its higher level of temporal granularity. This strong correlation is closely followed by positive correlations with the percentage area occupied by small scrubland, altitude and slope, as well as a negative correlation with the mean age of the resident population.

6.4 Summary

We tackled the regression problem of predicting percentages of yearly burnt area for each Portuguese parish. We used the same spatio-temporal indicators and imputation methods for the propositional approach as in Chapter 5. We defined auxiliary predicates for the relational approach that allow it to deal with numerical variables in the search for clauses, which we then propositionalised.

We used an under-sampling technique for regression and trained the same models (RF and SVR) on ten transformed training sets (repeating the experiments ten times). Both approaches resulted in good performances, with slighly better recall than precision (which can be beneficial in a scenario where false negatives are more costly than false positives). The under-sampling method doubles the F_1-measure obtained with each setup. The best propositional and relational approaches result in similar spatial distributions of regression utility, performing worse in the countryside Centre and coastal South.

Chapter 7

Conclusion

7.1 Summary

The problems of descriptive and predictive data mining in spatio-temporal databases have many applications. Several propositional and relational approaches have been proposed to tackle these problems in databases consisting of time-varying data that can be represented by an evolving thematic map. In this work, we were most interested in methods that both **a)** would be competitive with state-of-the-art, and **b)** would be interpretable by a domain expert. We therefore chose to focus on association rule learning and regression. We enumerated the main challenges posed by these problems, and provided a review of existing propositional and relational approaches to solve them.

Our main motivation was to understand and predict wildfires in mainland Portugal which every year have a strong socio-economical and environmental impact in the country. Furthermore, we wanted to understand the strengths and limitations of two different kinds of approaches – propositional and relational – often used to tackle the two problems. By comparing a specific approach of each type, we work toward obtaining a deeper insight on the methodological questions that arise when applying them to two different tasks. Besides the difficulty of including both the spatial and temporal dimensions in our approaches, we faced the added challenges raised by **a)** missing data, **b)** varying levels of temporal granularity, **c)** a low number of temporal data points, and **d)** an imbalanced target domain that did not correspond to our preference bias,

i.e., most instances corresponded to very residual burnt area percentages, when we were most interested in high values.

We define our space through a notion of spatio-temporal neighbourhood and neighbour direction with heterogeneous spatial objects. We used parishes as our atomic spatial object. Time is represented as several layers of labels over this space.

We opted to base our propositional approach on pre-processing, and our relational approach on ILP. For the propositional approach, we built spatio-temporal indicators based on the concept of spatio-temporal neighbourhood and of simplified borders we proposed for this setting. For the relational approach, we designed predicates expressing spatial and temporal relationships. Only the propositional approach required imputation of missing data, which was compensated for in the design of our relational predicates.

Both methodologies of association rule discovery required the categorisation of numerical attributes. In both cases, we applied standard association rule learning algorithms. The propositional approach is more time-efficient, and can find a larger number of rules with a wider range of confidence and lift. However, the relational approach allows the discovery of more interpretable and expressive rules.

The propositional approach to the prediction problem required only the application of standard regression algorithms to the transformed data. In contrast, the relational approach involved a few extra steps: **i)** the design of new predicates able to deal with numerical attributes (instead of categorising them), **ii)** the extraction of Boolean features through a methodology we developed to search for and select rules which optimises the F1-measure of a two-class classification problem, and **iii)** the propositionalisation of the selected rules. Only after these steps, were the same regression algorithms as the propositional approach applied to the data. The evaluation method we chose respected the temporal order of the data, and the precision and recall regression metrics dealt with the imbalance of the domain by employing a relevance function. The results confirmed that, in spite of the relational features having been optimised for classification, the relational approach is competitive in regression, presenting only a slight disadvantage in performance when compared to the propositional approach. However, the propositional approach has a clear advantage in pre-processing time. We compared the results with and without the use of an under-sampling method designed

for regression on the training set, and concluded that under-sampling greatly improves the predictive performance of both approaches.

In conclusion, we **i)** reviewed the state-of-the-art of propositional and relational spatio-temporal association rule learning and forecasting, **ii)** adapted a notion of spatio-temporal neighbourhood to include spatial direction, **iii)** proposed a concept of simplified border for heterogeneous spatial objects, **iv)** built spatio-temporal indicators based on these notions, **v)** designed relational predicates that deal with numerical attributes and include the temporal and spatial dimensions, **vii)** deployed a re-sampling technique to improve regression under an imbalanced domain, **viii)** developed and compared methodologies that relied on pre-processing (propositional) and ILP (relational) to the domain of associations discovery and prediction of wildfires.

7.2 Future research directions

Using our particular case study, some interesting directions worth exploring include the study of **a)** the changes in the importance of attributes with time and space, **b)** the impact of different settings for the spatio-temporal indicators, **c)** the difference in association rule learning results by changing categorisation parameters, **d)** the effects of different training and test sizes as well as different data set balances obtained by the re-sampling technique in forecasting performance.

Other interesting propositional approaches based on pre-processing that we would like to explore include the use of clustering to find neighbourhoods as proposed by Appice *et al.* (2013a), or an extension of the work proposed by Oliveira & Torgo (2014) to include spatial dimensions. In relational learning, we would be most interested in graphical modelling. In particular, we would like to experiment with tools such as Markov Logic Networks, as they seem naturally suitable to this task.

Finally, it would be interesting to compare propositional and relational approaches in other settings besides the one provided by our case study in order to generalise our findings.

Bibliography

Aggarwal, Charu C. 2015. Mining spatial data. *Chap. Trajectory mining, pages 544–554 of: Data Mining.* Springer International Publishing.

Agrawal, Rakesh, & Srikant, Ramakrishnan. 1994. Fast algorithms for mining association rules in large databases. *Pages 487–499 of: Proceedings of the 20th International Conference on Very Large Data Bases.* VLDB'94. Morgan Kaufmann Publishers Inc.

Agrawal, Rakesh, & Srikant, Ramakrishnan. 1995. Mining sequential patterns. *Pages 3–14 of: Proceedings of the 11th International Conference on Data Engineering.* ICDE'95.

Ailliot, Pierre, Thompson, Craig, & Thomson, Peter. Space–time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58**(3).

Allison, Paul D. 2001. *Missing data.* Vol. 136. Sage publications.

Analytics, Revolution. 2014. *doMC: Foreach parallel adaptor for the multicore package.* R package version 1.3.3.

Analytics, Revolution, & Weston, Steve. 2014. *foreach: Foreach looping construct for R.* R package version 1.4.2.

Andrienko, Gennady, Malerba, Donato, May, Michael, & Teisseire, Maguelonne. 2006. Mining spatio-temporal data. *Journal of Intelligent Information Systems*, **27**(3), 187–190.

Appice, Annalisa, Ceci, Michelangelo, Lanza, Antonietta, Lisi, Francesca A., & Malerba, Donato. 2003. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*, **7**(6), 541–566.

Appice, Annalisa, Berardi, Margherita, Ceci, Michelangelo, & Malerba, Donato. 2005. Mining and filtering multi-level spatial association rules with ARES. *Pages 342–353 of: Proceedings of the 15th International Conference on Foundations of Intelligent Systems*. ISMIS'05. Springer-Verlag.

Appice, Annalisa, Ciampi, Anna, Lanza, Antonietta, Malerba, Donato, Rapolla, Antonella, & Vetturi, Luisa. 2008. Geographic knowledge discovery in INGENS: An inductive database perspective. *Pages 326–331 of: Workshops Proceedings of the 8th IEEE International Conference on Data Mining*. ICDM'08. IEEE Computer Society.

Appice, Annalisa, Pravilovic, Sonja, Malerba, Donato, & Lanza, Antonietta. 2013a. Enhancing regression models with spatio-temporal indicator additions. *Pages 433–444 of: Proceedings of the XIIIth International Conference of the Italian Association for Artificial Intelligence*. AI*IA 2013. Springer.

Appice, Annalisa, Ciampi, Anna, Malerba, Donato, & Guccione, Pietro. 2013b. Using trend clusters for spatiotemporal interpolation of missing data in a sensor network. *Journal of Spatial Information Science*, **6**(1), 119–153.

Azevedo, Paulo J. 2003. CAREN – A java based apriori implementation for classification purposes.

Azevedo, Paulo J., & Jorge, Alípio M. 2010. Ensembles of jittered association rule classifiers. *Data Mining and Knowledge Discovery*, **21**(1), 91–129.

Barber, Chris, Bockhorst, Joseph, & Roebber, Paul. 2010. Auto-regressive HMM inference with incomplete data for short-horizon wind forecasting. *Pages 136–144 of: Advances in Neural Information Processing Systems 23*. NIPS'10. Curran Associates, Inc.

Bassi, Samuela, Kettunen, Marianne, Kampa, Eleftheria, & Cavalieri, Sandra. 2008. Forest fires: causes and contributing factors in Europe. *European Parliament, Brussels*.

Batuwita, Rukshan, & Palade, Vasile. 2010. Efficient resampling methods for training support vector machines with imbalanced datasets. *Pages 1–8 of: International Joint Conference on Neural Networks*. IJCNN'10. IEEE.

Bayardo, Roberto J., Agrawal, Rakesh, & Gunopulos, Dimitrios. 2000. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, **4**(2/3), 217–240.

Berry, Michael J., & Linoff, Gordon. 1997. *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons, Inc.

Bi, Jinbo, & Bennett, Kristin P. 2003. Regression error characteristic curves. *Pages 43–50 of: Proceedings of the 20th International Conference on Machine Learning*. ICML'03. AAAI Press.

Bilgili, Mehmet, Sahin, Besir, & Yasar, Abdulkadir. 2007. Application of artificial neural networks for the wind speed prediction of target station using reference stations data. *Renewable Energy*, **32**(14), 2350 – 2360.

Bivand, Roger S., Pebesma, Edzer J., & Gómez-Rubio, Virgilio. 2013. *Applied spatial data analysis with R - Second edition*. Use R!, vol. 10. Springer.

Bivand, Roger S., Keitt, Tim, & Rowlingson, Barry. 2015. *rgdal: Bindings for the geospatial data abstraction library*. R package version 1.0-7.

Blondel, Emmanuel. 2015. *cleangeo: Cleaning geometries from spatial objects*. R package version 0.1-1.

Branco, Paula. 2014. *Re-sampling approaches for regression tasks under imbalanced domains*. M.Phil. thesis, Computer Science Department, Faculty of Sciences – University of Porto.

Branco, Paula, Ribeiro, Rita P., & Torgo, Luís. 2014. *UBL: Utility-based learning*. R package version 0.0.1.

Branco, Paula, Torgo, Luís, & Ribeiro, Rita P. 2015. A survey of predictive modelling under imbalanced distributions. *Computing Research Repository*, **abs/1505.01658**.

Breiman, Leo. 2001. Random forests. *Machine Learning*, **45**(1), 5–32.

Brin, Sergey, Motwani, Rajeev, & Silverstein, Craig. 1997a. Beyond market baskets: Generalizing association rules to correlations. *SIGMOD Record*, **26**(2), 265–276.

Brin, Sergey, Motwani, Rajeev, Ullman, Jeffrey D., & Tsur, Shalom. 1997b. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Record*, **26**(2), 255–264.

Calargun, Seda Unal, & Yazici, Adnan. 2008. Fuzzy association rule mining from spatio-temporal data. *Pages 631–646 of: Proceedings of the International Conference on Computational Science and Its Applications, Part I*. ICCSA'08. Springer.

Camossi, Elena, Bertolotto, Michela, & Kechadi, M. Tahar. 2008. Mining spatio-temporal data at different levels of detail. *Pages 225–240 of: Proceedings of the 11th Agile Conference*. AGILE'08. Springer.

Cano, Rafael, Sordo, Carmen, & Gutiérrez, José M. 2004. Applications of Bayesian networks in meteorology. *Pages 309–328 of: Advances in Bayesian Networks*. Studies in Fuzziness and Soft Computing, vol. 146.

Ceci, Michelangelo, & Appice, Annalisa. 2006. Spatial associative classification: Propositional vs structural approach. *Journal of Intelligent Information Systems*, **27**(3), 191–213.

Ceri, Stefano, Gottlob, Georg, & Tanca, Letizia. 1989. What you always wanted to know about Datalog (and never dared to ask). *IEEE Transactions on Knowledge and Data Engineering*, **1**(1), 146–166.

Cheng, Tao, & Wang, Jiaqiu. 2008. Integrated spatio-temporal data mining for forest fire prediction. *Transactions in GIS*, **12**(5), 591–611.

Ciampi, Anna, Appice, Annalisa, & Malerba, Donato. 2010. Discovering trend-based clusters in spatially distributed data streams. *Pages 107–122 of: Proceedings of the ECML-PKDD'10 International Workshop on Mining Ubiquitous and Social Environments*.

Cortes, Corinna, & Vapnik, Vladimir. 1995. Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Das, Debasish, Kodra, Evan, Ganguly, Auroop R., & Obradovic, Zoran. 2012. Mining extreme values: Climate and natural hazards. *In: Proceedings of the KDD'12 Workshop on Data Mining Applications in Sustainability*. ACM.

Dehaspe, Luc, & Toivonen, Hannu. 1999. Discovery of frequent Datalog patterns. *Data Mining and Knowledge Discovery*, **3**(1), 7–36.

Dhar, Vasant, & Tuzhilin, Alexander. 1993. Abstract-driven pattern discovery in databases. *IEEE Transactions on Knowledge and Data Engineering*, **5**(6), 926–938.

Drucker, Harris, Burges, Christopher J. C., Kaufman, Linda, Smola, Alexander J., & Vapnik, Vladimir. 1996. Support vector regression machines. *Pages 155–161 of: Advances in Neural Information Processing Systems 9*. NIPS'96. MIT Press.

Džeroski, Sašo. 2003. Multi-relational data mining: An introduction. *SIGKDD Explorations Newsletter*, **5**(1), 1–16.

Esposito, Floriana, Di Mauro, Nicola, Basile, Teresa M. Altomare, & Ferilli, Stefano. 2009. Multi-dimensional relational sequence mining. *Fundamenta Informaticae*, **89**(1), 23–43.

Ester, Martin, Kriegel, Hans-Peter, & Sander, Jörg. 1997. Spatial data mining: A database approach. *Pages 47–66 of: Proceedings of the 5th International Symposium on Advances in Spatial Databases*. SSD'97. Springer-Verlag.

Fayyad, Usama M., Piatetsky-Shapiro, Gregory, & Smyth, Padhraic. 1996. From data mining to knowledge discovery in databases. *AI Magazine*, **17**(3), 37–54.

Feng, Ling, Dillon, Tharam S., & Liu, James. 2001. Inter-transactional association rules for multi-dimensional contexts for prediction and their application to studying meteorological data. *Data & Knowledge Engineering*, **37**(1), 85–115.

Fernández, Alberto, García, Salvador, del Jesus, María J., & Herrera, Francisco. 2008. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, **159**(18), 2378 – 2398.

Han, Jaiwei, Koperski, Krzysztof, & Stefanovic, Nebojsa. 1997. GeoMiner: A system prototype for spatial data mining. *SIGMOD Record*, **26**(2), 553–556.

Han, Jiawei, & Fu, Yongjian. 1995. Discovery of multiple-level association rules from large databases. *Pages 420–431 of: Proceedings of the 21st International Conference on Very Large Data Bases*. VLDB'95. Morgan Kaufmann Pub. Inc.

Han, Jiawei, Pei, Jian, Yin, Yiwen, & Mao, Runying. 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, **8**(1), 53–87.

Hernández-Orallo, José. 2013. ROC curves for regression. *Pattern Recognition*, **46**(12), 3395–3411.

Huang, Yan, Shekhar, Shashi, & Xiong, Hui. 2004. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering*, **16**(12), 1472–1485.

Huang, Yo-Ping, Kao, Li-Jen, & Sandnes, Frode-Eika. 2008. Efficient mining of salinity and temperature association rules from ARGO data. *Expert Systems with Applications*, **35**(1), 59–68.

Isaaks, Edward H., & Srivastava, R. Mohan. 1989. *Applied geostatistics*. Vol. 2. Oxford University Press New York.

Janeja, Vandana P., Adam, Nabil R., Atluri, Vijayalakshmi, & Vaidya, Jaideep. 2010. Spatial neighborhood based anomaly detection in sensor datasets. *Data Mining and Knowledge Discovery*, **20**(2), 221–258.

Jenks, George F. 1967. The data model concept in statistical mapping. *Pages 186+ of: International Yearbook of Cartography*, vol. 7. Rand McNally & Co.

Jorge, Alípio M. 2015. *carenR: CarenR - Classification and Association Rules ENgine for the R statistical package*. R package version 0.1-15-03.2.

Kahle, David, & Wickham, Hadley. 2013. ggmap: Spatial visualization with ggplot2. *The R Journal*, **5**(1), 144–161.

Koperski, Krzysztof, & Han, Jiawei. 1995. Discovery of spatial association rules in geographic information databases. *Pages 47–66 of: Proceedings of the 4th International Symposium on Advances in Spatial Databases*. SSD'95. Springer.

Krige, Danie G. 1951. *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. Ph.D. thesis, University of the Witwatersrand.

Kubat, Miroslav, Holte, Robert C., & Matwin, Stan. 1998. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, **30**(2-3), 195–215.

Lavrac, Nada, & Džeroski, Sašo. 1994. *Inductive Logic Programming – Techniques and applications*. Ellis Horwood series in artificial intelligence. Ellis Horwood.

Li, Gong, & Shi, Jing. 2010. On comparing three artificial neural networks for wind speed forecasting. *Applied Energy*, **87**(7), 2313–2320.

Li, Zhigang, Dunham, Margaret H., & Xiao, Yongqiao. 2003. STIFF: A forecasting framework for spatiotemporal data. *Pages 183–198 of: Revised papers of the PAKDD'02 Workshop on Mining Multimedia and Complex Data*. Springer.

Liaw, Andy, & Wiener, Matthew. 2002. Classification and regression by randomForest. *R News*, **2**(3), 18–22.

Lindström, Johan, Szpiro, Adam, Sampson, Paul D., Bergen, Silas, & Sheppard, Lianne. SpatioTemporal: An R package for spatio-temporal modelling of air-pollution.

Lindström, Johan, Szpiro, Adam A., Sampson, Paul D., Oron, Assaf P., Richards, Mark, Larson, Tim V., & Sheppard, Lianne. 2014. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and Ecological Statistics*, **21**(3), 411–433.

Lisi, Francesca A., & Malerba, Donato. 2004. Inducing multi-level association rules from multiple relations. *Machine Learning*, **55**(2), 175–210.

Lu, Hongjun, Han, Jiawei, & Feng, Ling. 1998. Stock movement prediction and n-dimensional inter-transaction association rules. *Page 12 of: Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

Lu, Hongjun, Feng, Ling, & Han, Jiawei. 2000. Beyond intratransaction association analysis: mining multidimensional intertransaction association rules. *ACM Transactions on Information Systems*, **18**(4), 423–454.

Luk, K. C., Ball, James E., & Sharma, Ashish. 2000. A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *Journal of Hydrology*, **227**(1), 56–65.

Ma, Yiming, Liu, Bing, & Hsu, Wynne. 1998. Integrating classification and association rule mining. *In: Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'98. AAAI Press.

Madadgar, Shahrbanou, & Moradkhani, Hamid. 2014. Spatio-temporal drought forecasting within Bayesian networks. *Journal of Hydrology*, **512**, 134–146.

Malerba, Donato. 2008. A relational perspective on spatial data mining. *International Journal of Data Mining, Modelling and Management*, **1**(1), 103–118.

Malerba, Donato, & Lisi, Francesca A. 2001a. Discovering associations between spatial objects: An ILP application. *Pages 156–163 of: Proceedings of the 11th International Conference on Inductive Logic Programming*. ILP'01. Springer.

Malerba, Donato, & Lisi, Francesca A. 2001b. An ILP method for spatial association rule mining. *Pages 18–29 of: First Workshop on Multi-Relational Data Mining*.

Malerba, Donato, Esposito, Floriana, Lanza, Antonietta, Lisi, Francesca A., & Appice, Annalisa. 2003. Empowering a GIS with inductive learning capabilities: The case of INGENS. *Computers, Environment and Urban Systems*, **27**(3), 265–281.

Malerba, Donato, Lanza, Antonietta, & Appice, Annalisa. 2009. Leveraging the power of spatial data mining to enhance the applicability of GIS technology. *Chap. 10, page 255 of: Geographic Data Mining and Knowledge Discovery*. CRC Press.

Mamoulis, Nikos. 2009. Spatio-temporal data mining. *Pages 2725–2730 of: Encyclopedia of Database Systems*. Springer US.

McGovern, Amy, Gagne Ii, David J., Williams, John K., Brown, Rodger A., & Basara, Jeffrey B. 2014. Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Machine Learning*, **95**(1), 27–50.

Mennis, Jeremy, & Liu, Jun Wei. 2005. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, **9**(1), 5–17.

Metz, Charles E. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**(4), 283–298.

Meyer, David, Dimitriadou, Evgenia, Hornik, Kurt, Weingessel, Andreas, & Leisch, Friedrich. 2014. *e1071: Misc functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4.

Muggleton, Stephen. 1995. Inverse entailment and Progol. *New Generation Computing*, **13**(3-4), 245–286.

Muggleton, Stephen, & De Raedt, Luc. 1994. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, **19**, 629–679.

Nanni, Mirco, Kuijpers, Bart, Körner, Christine, May, Michael, & Pedreschi, Dino. 2008. Spatiotemporal data mining. *Pages 267–296 of: Mobility, Data Mining and Privacy*. Springer Berlin Heidelberg.

Neill, Daniel B., Moore, Andrew W., Sabhnani, Maheshkumar, & Daniel, Kenny. 2005. Detection of emerging space-time clusters. *Pages 218–227 of: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'05. ACM.

Ohashi, Orlando, & Torgo, Luís. 2012. Wind speed forecasting using spatio-temporal indicators. *Pages 975–980 of: Proceedings of the 20th European Conference on Artificial Intelligence*. ECAI'12. IOS Press.

Oliveira, Mariana, & Torgo, Luís. 2014. Ensembles for time series forecasting. *Pages 360–370 of: Proceedings of the 6th Asian Conference on Machine Learning*. ACML'14. Journal of Machine Learning Research, Inc.

Pace, R. Kelley, Barry, Ronald, Clapp, John M., & Rodriquez, Mauricio. 1998. Spatiotemporal autoregressive models of neighborhood effects. *The Journal of Real Estate Finance and Economics*, **17**(1), 15–33.

Pebesma, Edzer J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683–691.

Pebesma, Edzer J., & Bivand, Roger S. 2005. Classes and methods for spatial data in R. *R News*, **5**(2), 9–13.

Pei, Jian, Han, Jiawei, Mortazavi-Asl, Behzad, Pinto, Helen, Chen, Qiming, Dayal, Umeshwar, & Hsu, Meichun. 2001. PrefixSpan: Mining sequential patterns by prefix-projected growth. *Pages 215–224 of: Proceedings of the 17th International Conference on Data Engineering*. ICDE'01. IEEE Computer Society.

Pei, Jian, Han, Jiawei, Mortazavi-Asl, Behzad, Wang, Jianyong, Pinto, Helen, Chen, Qiming, Dayal, Umeshwar, & Hsu, Mei-Chun. 2004. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, **16**(11), 1424–1440.

Piatkowski, Nico, Lee, Sangkyun, & Morik, Katharina. 2013. Spatio-temporal random fields: compressible representation and distributed estimation. *Machine Learning*, **93**(1), 115–139.

Plotkin, Gordon D. 1970. A note on inductive generalization. *Machine Intelligence*, **5**(1), 153–163.

Pravilovic, Sonja, & Appice, Annalisa. 2014. Application of spatio-temporal clustering in forecasting optimization of geo-referenced time series. *American Journal of Modeling and Optimization*, **2**(1), 8–15.

Provost, Foster J., Fawcett, Tom, & Kohavi, Ron. 1998. The case against accuracy estimation for comparing induction algorithms. *Pages 445–453 of: Proceedings of the 15th International Conference on Machine Learning*. ICML'98. Morgan Kaufmann.

R Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, Dan, Grundler, Mike, Title, Pascal, Anderson, Carlos, Shi, Jeff, Brown, Joseph, & Huang, Huateng. 2015. *BAMMtools: Analysis and visualization of macroevolutionary dynamics on phylogenetic trees*. R package version 2.0.5.

Ramsey, Paul, & Columbia, Victoria-British. 2005. *Introduction to PostGIS*.

Ramsey, Paul, *et al.* . 2005. *PostGIS manual*. Refractions Research Inc.

Ribeiro, Rita P. 2011. *Utility-based regression*. Ph.D. thesis, Computer Science Department, Faculty of Sciences – University of Porto.

Santos Costa, Vítor, Damas, Luís, Reis, Rogério, & Azevedo, Rúben. 2010. *YAP User's manual*. University of Porto.

Srikant, Ramakrishnan, & Agrawal, Rakesh. 1996. Mining sequential patterns: Generalizations and performance improvements. *Pages 3–17 of: Proceedings of the 5th International Conference on Extending Database Technology*. EDBT'96. Springer.

Srinivasan, Ashwin. 2007. *The Aleph manual*. University of Oxford.

Stojanova, Daniela, Ceci, Michelangelo, Appice, Annalisa, & Džeroski, Sašo. 2012. Network regression with predictive clustering trees. *Data Mining and Knowledge Discovery*, **25**(2), 378–413.

Tan, P., Steinbach, Michael, Kumar, Vipin, Potter, Christopher, Klooster, Steven, & Torregrosa, Alicia. 2001. Finding spatio-temporal patterns in earth science data. *In: Proceedings of the KDD'01 Workshop on Temporal Data Mining*.

Tang, Kwei, Chen, Yen-Liang, & Hu, Hsiao-Wei. 2008. Context-based market basket analysis in a multiple-store environment. *Decision Support Systems*, **45**(1), 150–163.

Telang, Aditya, Deepak, P., Joshi, Salil, Deshpande, Prasad, & Rajendran, Ranjana. 2014. Detecting localized homogeneous anomalies over spatio-temporal data. *Data Mining and Knowledge Discovery*, **28**(5-6), 1480–1502.

Thompson, Craig S., Thomson, Peter J., & Zheng, Xiaogu. 2007. Fitting a multisite daily rainfall model to New Zealand data. *Journal of Hydrology*, **340**(1), 25–39.

Tobler, Waldo R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 234–240.

Torgo, Luís. 2005. Regression error characteristic surfaces. *Pages 697–702 of: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'05. ACM.

Torgo, Luís. 2014. An infra-structure for performance estimation and experimental comparison of predictive models in R. *Computing Research Repository*, **abs/1505.01658**.

Torgo, Luís, & Ribeiro, Rita P. 2007. Utility-based regression. *Pages 597–604 of: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*. PKDD'07. Springer.

Torgo, Luís, & Ribeiro, Rita P. 2009. Precision and recall for regression. *Pages 332–346 of: Proceedings of the 12th International Conference on Discovery Science*. DS'09. Springer.

Torgo, Luís, Ribeiro, Rita P., Pfahringer, Bernhard, & Branco, Paula. 2013. Smote for regression. *Pages 378–389 of: Proceedings of the 16th Portuguese Conference on Artificial Intelligence*. EPIA'13. Springer.

Torres, João. 2014. *Patterns and drivers of wildfire occurrence and post-fire vegetation resilience across scales in Portugal*. Ph.D. thesis, Biology Department, Faculty of Sciences – University of Porto.

Tsoukatos, Ilias, & Gunopulos, Dimitrios. 2001. Efficient mining of spatiotemporal patterns. *Pages 425–442 of: Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*. SSTD'01.

Tung, A., Lu, Hongjun, Han, Jiawei, & Feng, Ling. 2003. Efficient mining of intertransaction association rules. *IEEE Transactions on Knowledge and Data Engineering*, **15**(1), 43–56.

Tung, Anthony K. H., Lu, Hongjun, Han, Jiawei, & Feng, Ling. 1999. Breaking the barrier of transactions: Mining inter-transaction association rules. *Pages 297–301 of: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD'99. ACM.

Ulrich, Joshua. 2013. *TTR: Technical Trading Rules*. R package version 0.22-0.

Van Rijsbergen, Cornelis J. 1979. *Information retrieval*. 2nd edn. Newton, MA, USA: Butterworth-Heinemann.

Vaz, David, Santos Costa, Vítor, & Ferreira, Michel. 2011. Fire! firing inductive rules from economic geography for fire risk detection. *Pages 238–252 of: Revised papers of the 20th International Conference on Inductive Logic Programming*. ILP'10. Springer.

Wickham, Hadley. 2009. *ggplot2: Elegant graphics for data analysis.* Springer NY.

Wickham, Hadley. 2015. *lazyeval: Lazy (non-standard) evaluation.* R package version 0.1.10.

Wickham, Hadley, & Francois, Romain. 2015. *dplyr: A grammar of data manipulation.* R package version 0.4.1.

Yao, Xiaobai. 2003. Research issues in spatio-temporal data mining. *Pages 18–20 of: A white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery.*

Zaki, Mohammed J. 2000. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, **12**(3), 372–390.

Zaki, Mohammed J. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, **42**(1-2), 31–60.