Propositional and Relational Approaches to Spatio-Temporal Data Analysis

AUTHOR: MARIANA RAFAELA OLIVEIRA

SUPERVISOR: PROF. LUÍS TORGO

CO-SUPERVISOR: PROF. VÍTOR SANTOS COSTA

COMPUTER SCIENCE DEPARTMENT

FACULTY OF SCIENCES, UNIVERSITY OF PORTO

DECEMBER 2015

Outline

Introduction

Wildfires in Portugal: a case study

Describing wildfires
- Introduction
- Propositional and relational methods
- Experimental results

Predicting wildfires
- Introduction
- Propositional and relational methods
- Experimental results

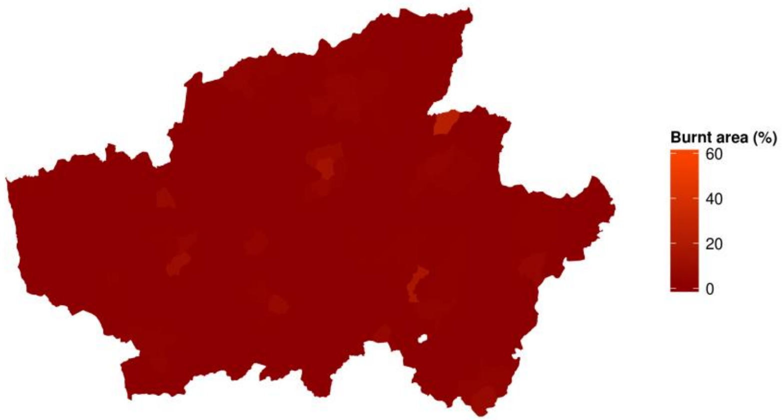
Conclusion

Introduction

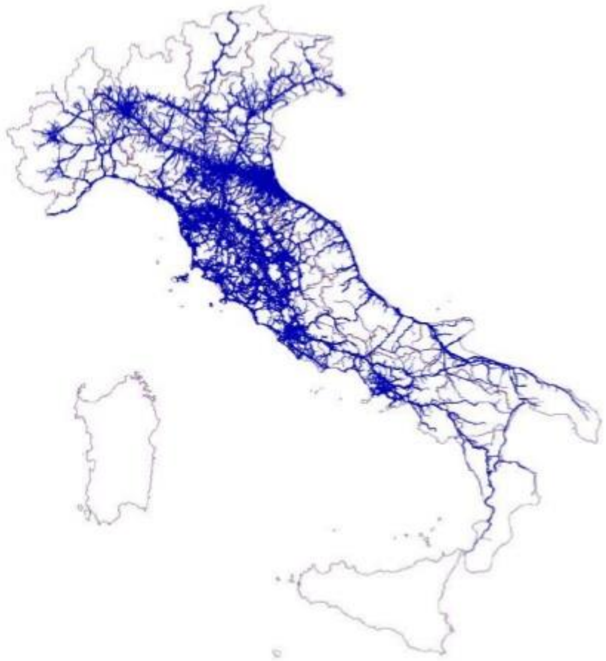
Spatio-temporal databases

EVOLVING THEMATIC MAPS

AND SENSOR NETWORKS



MOVING OBJECTS



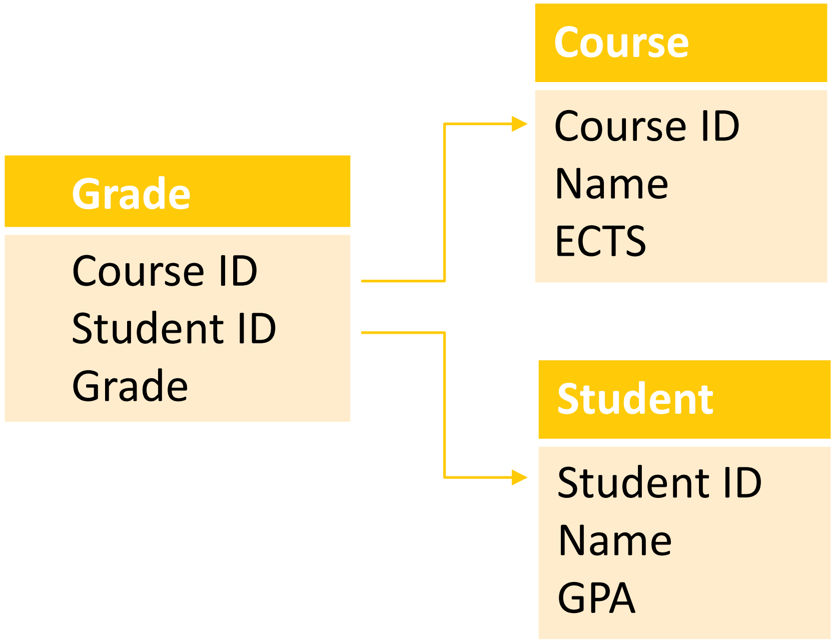
Source: Visualization in Mobility Data Mining , S. Rinzivillo

Propositional and relational approaches

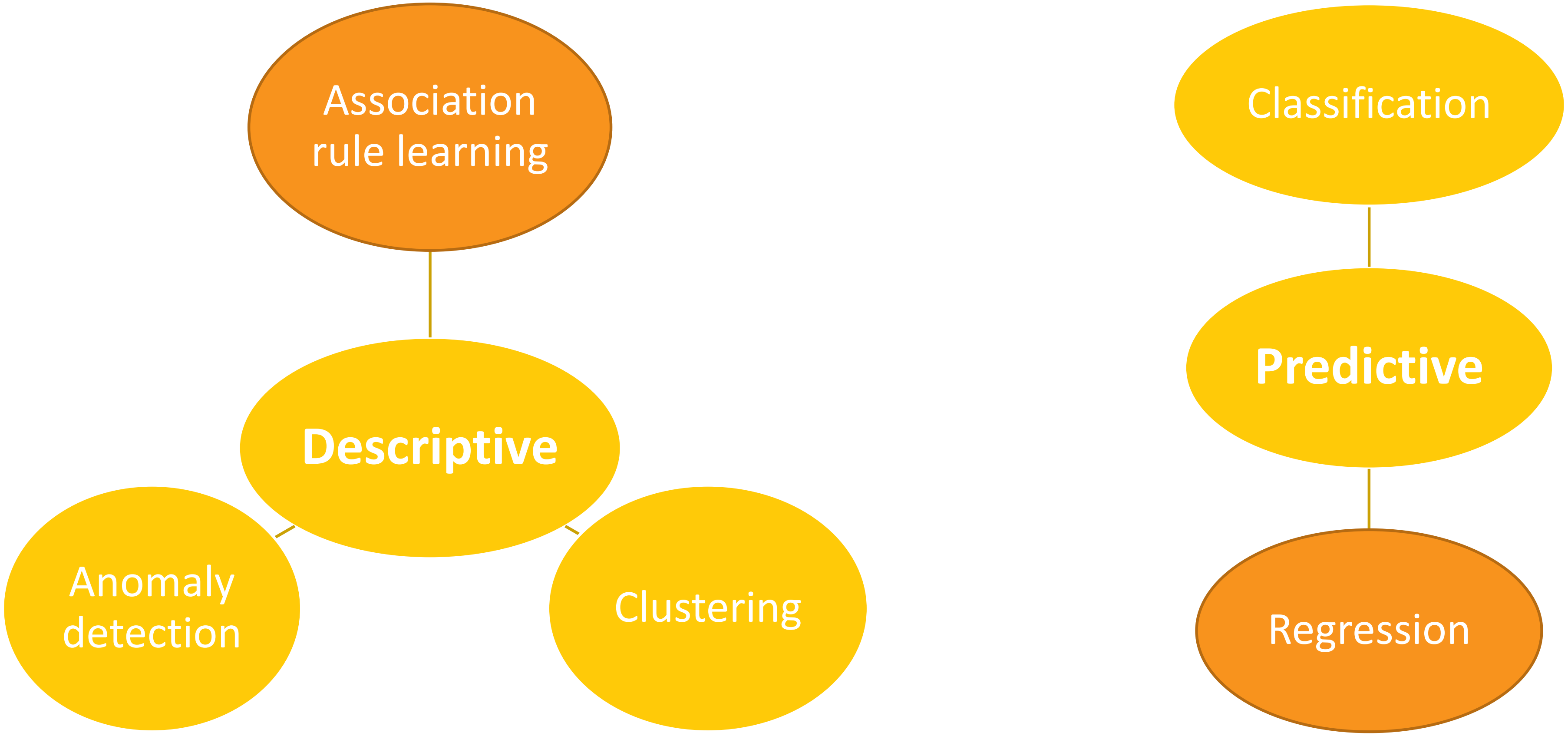
PROPOSITIONAL

Course ID	Student ID	Grade

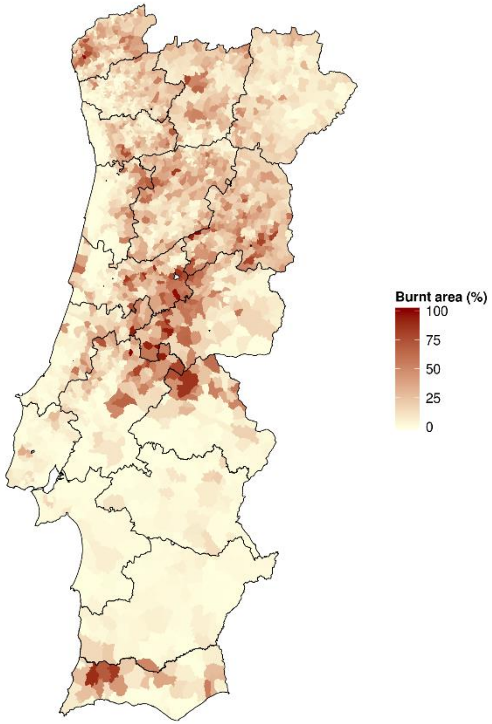
RELATIONAL



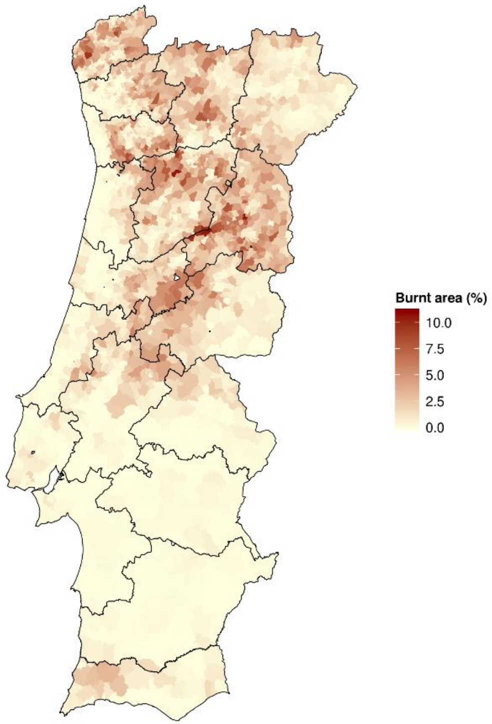
Descriptive and predictive methods



Motivation and main goals

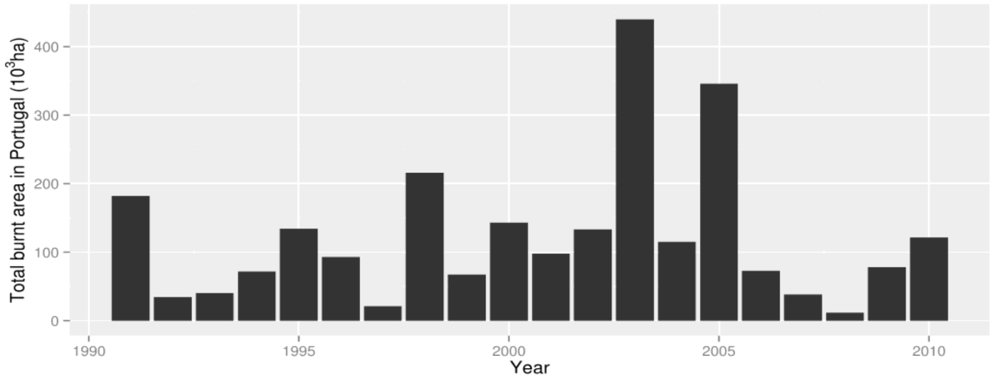


Maximum burnt area (%)



Mean burnt area (%)

Total burnt area in Portugal (10^3ha)



- Review the state-of-the-art
- Apply propositional and relational methods
- Compare approaches

Wildfires in Portugal

A CASE STUDY

Wildfires in Portugal

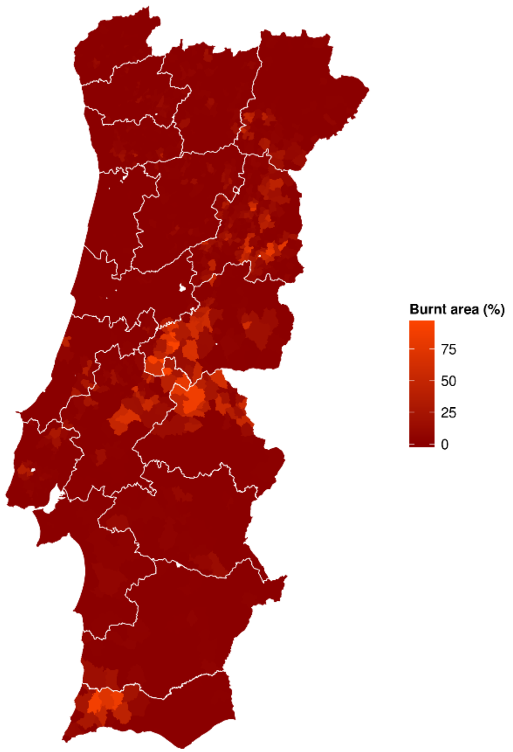
Portugal

- 2882 civil parishes
- 278 municipalities
- 18 districts

Parish total area

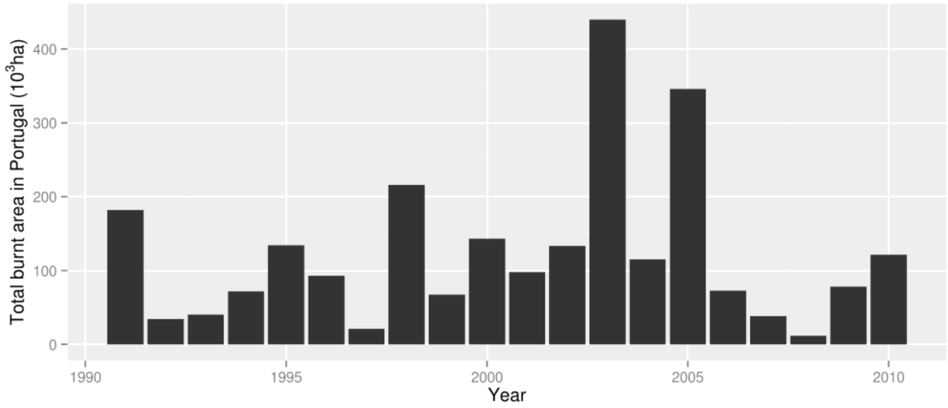
- 20 ha – 88 000 ha
- Median: 1700 ha

Area burnt in 2003 (%)

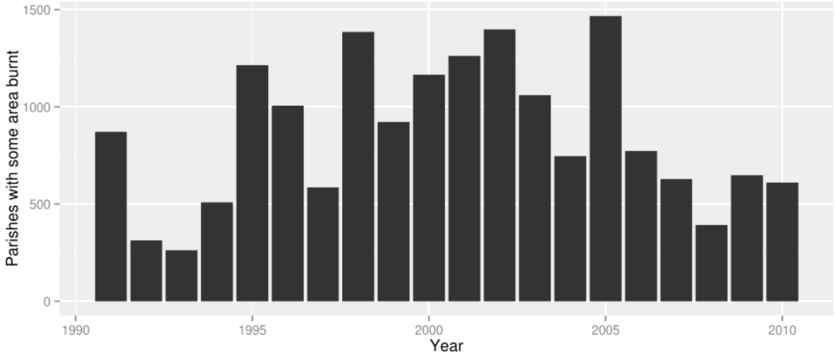


Percentage of burnt area (yearly)

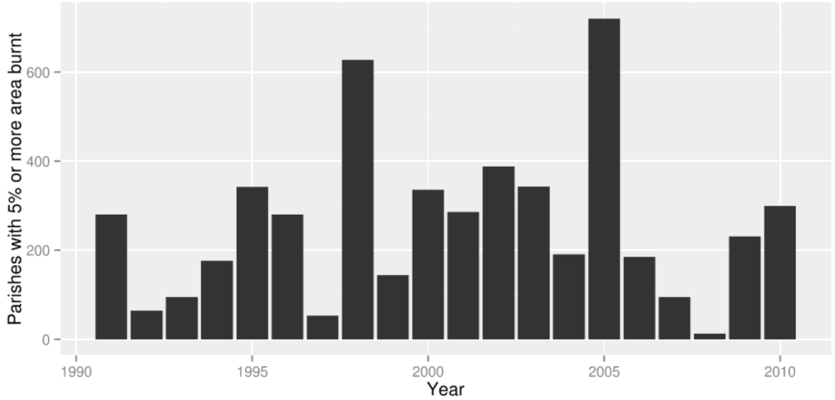
Total burnt area (10^3ha)



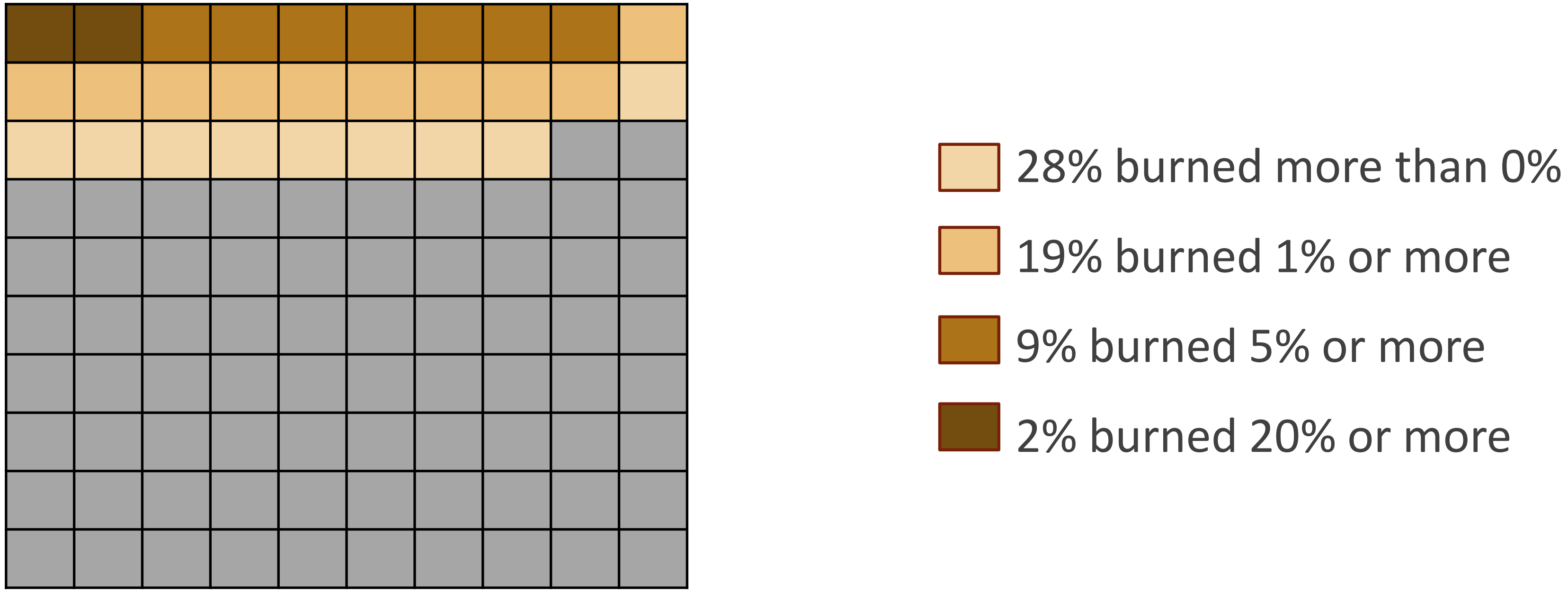
Number of parishes with >0% burnt area



Number of parishes with ≥5% burnt area



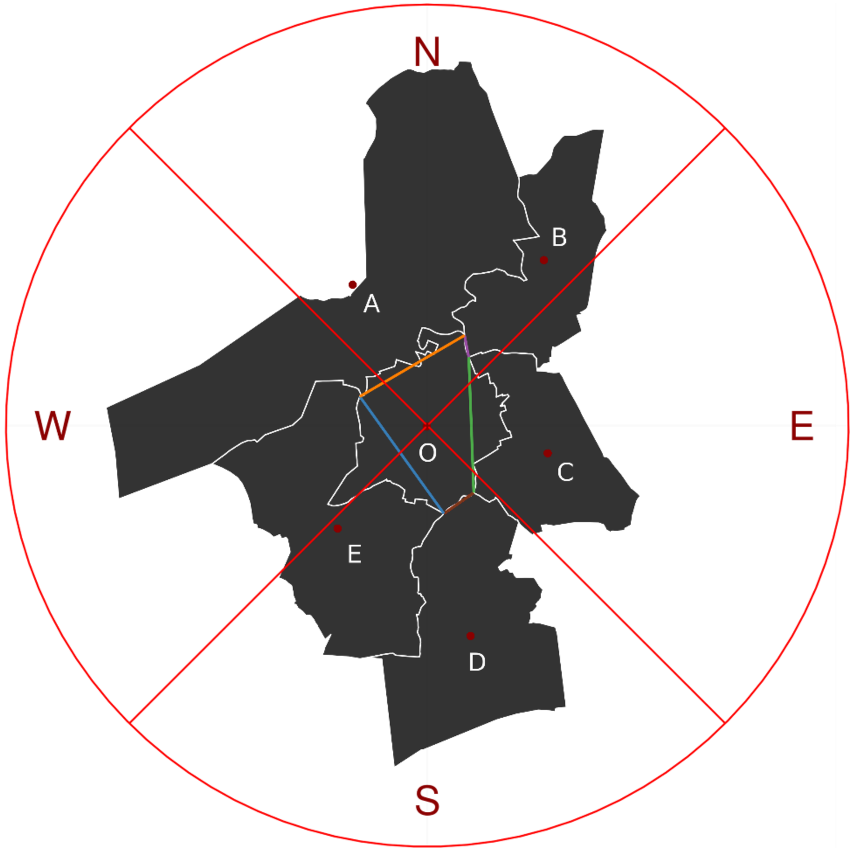
Imbalanced domain



Background knowledge

Land cover	Terrain	Road density	Census data (1989,1999,2009)	Census data (1991,2001)	Census data (1991)
• Eucalyptus • Tall scrubland • Small scrubland • Broad-leaved managed forest • Pinewood • Urban	• Maximum altitude • Mean altitude • Maximum slope • Mean slope	• All roads • Roads (>6m wide) • Road (<6m wide)	• Irrigable area • Meadow area • Bovine dens. • Ovine dens. • Caprine dens.	• Population density • Population's mean age	• Population aged 65+ • Housing density

Computing spatial relationships



Using PostGIS

- Find neighbours with *ST_Intersect*
- Calculate neighbour direction with *ST_Azimuth* and *ST_Centroid*
  - East – $[45,135[^{º}$
  - South – $[135,225[^{º}$
  - West – $[225,315[^{º}$
  - North – $([315,360[U[0,45[)^{º}$
- Calculate border parishes with *ST_Union* and *ST_Intersects*

Describing wildfires

INTRODUCTION

Association rules

TID	Basket
1	milk, bread
2	butter
3	beer, diapers
4	milk, bread, butter
5	bread

$$A \Rightarrow C$$

$$support = \Pr(A, C)$$

$$confidence = \Pr(C|A)$$

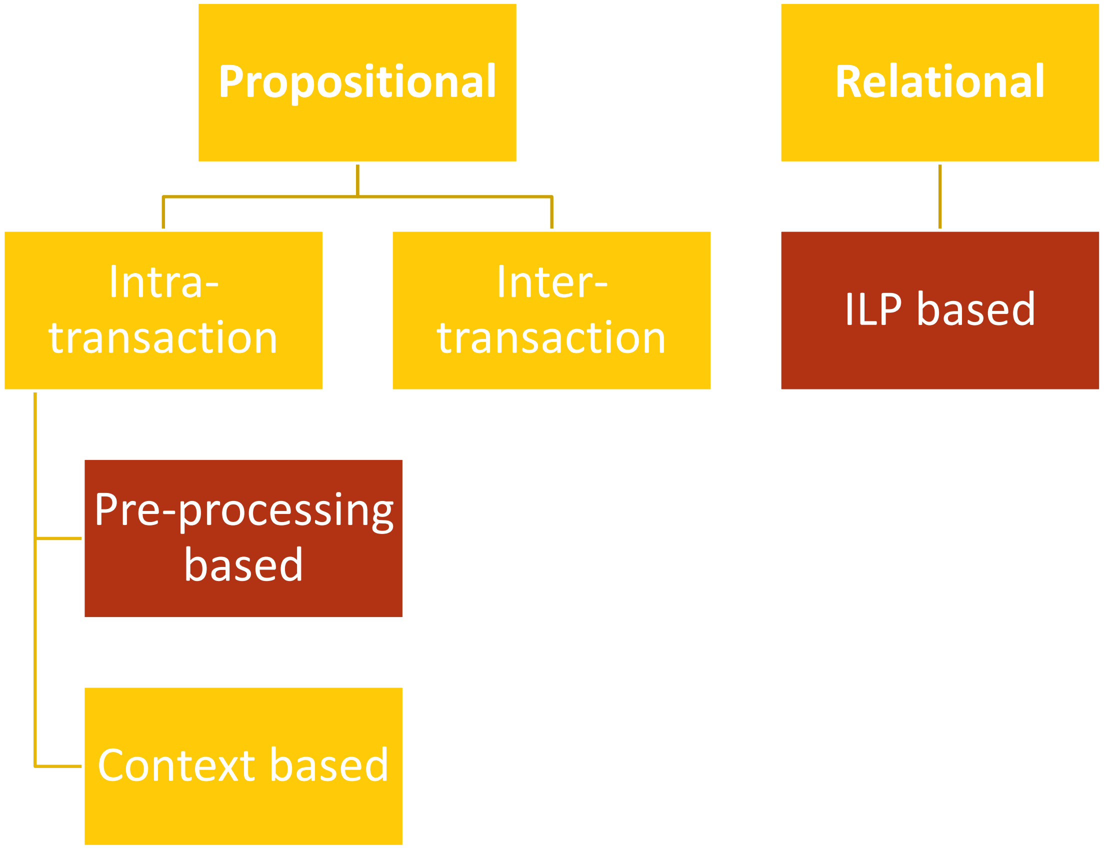
$$lift = \frac{\Pr(C|A)}{\Pr(C)}$$

Example:

$$\{butter, bread\} \Rightarrow \{milk\}$$

$$supp = 0.2; \quad conf = 1.0; \quad lift = 2.5$$

Spatio-temporal association rule learning

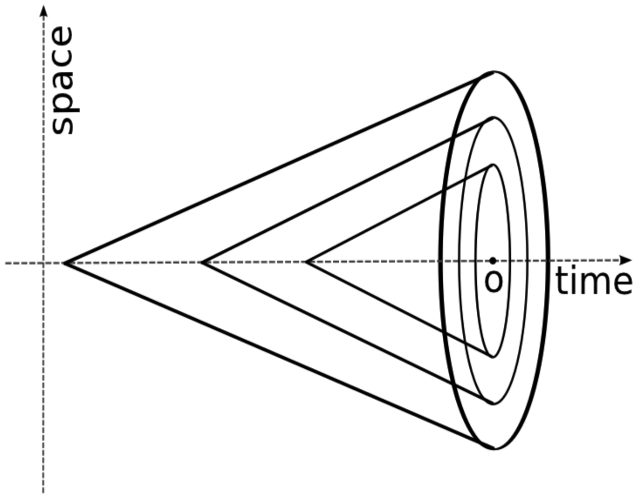


Describing wildfires

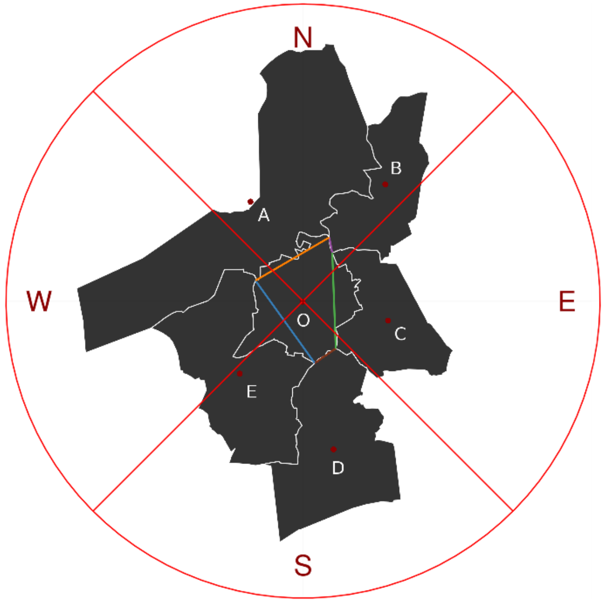
PROPOSITIONAL AND RELATIONAL METHODS

Propositional approach

Build spatio-temporal indicators → Impute missing data → Categorise numerical variables → Model with carenR



Source: Ohashi & Torgo (2012)



Temporal indicator

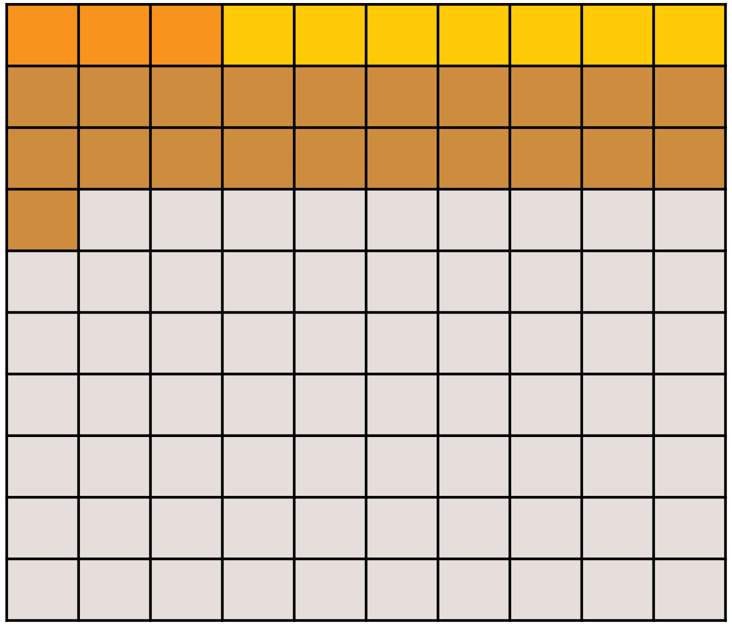
- EMA with k = 9

Spatio-temporal indicator

- Directional
- EMA with k = 5
- Weighted average with simplified borders (with *ST_MaxDistance*)

Propositional approach

Build spatio-temporal indicators → Impute missing data → Categorise numerical variables → Model with carenR



Missing data	Solution
3% due to unavailable data	Independent spatial-only IDW
7% due to ST-indicators construction	Fill with 0 if borders with sea, average of the two nearest directions otherwise
21% due to low temporal granularity	Use most recent measurement

Propositional approach

Build spatio-temporal indicators → Impute missing data → **Categorise numerical variables** → Model with carenR

- Jenk's natural breaks classification method
- 4 categories per variable
  - Very Low
  - Low
  - Medium
  - High

Relational approach

Convert data to Prolog → Design temporal and spatial predicates → Design categorising predicates → Model with Aleph (ar mode)

BACKGROUND KNOWLEDGE

- *numAttribute(Parish, Value).*
- *numAttribute(Parish, Year, Value).*
- *neighbour(Parish, Parish).*
- *border(Parish, Object).*

POSITIVE EXAMPLES

- *burntArea(Parish, Year, Category).*

Relational approach

Convert data to Prolog → **Design temporal and spatial predicates** → Design categorising predicates → Model with Aleph (ar mode)

TEMPORAL PREDICATE

- *yearsSinceLastFireLE(Parish, Year, TimeDist)*
  if last fire was TimeDist or less years ago
- *yearsSinceLastFireGE(Parish, Year, TimeDist)*
  if last fire was TimeDist or more years ago

SPATIAL PREDICATE

- *fixedNeighbour(Parish, Neighbour)*
  prevents neighbour recursion

Relational approach

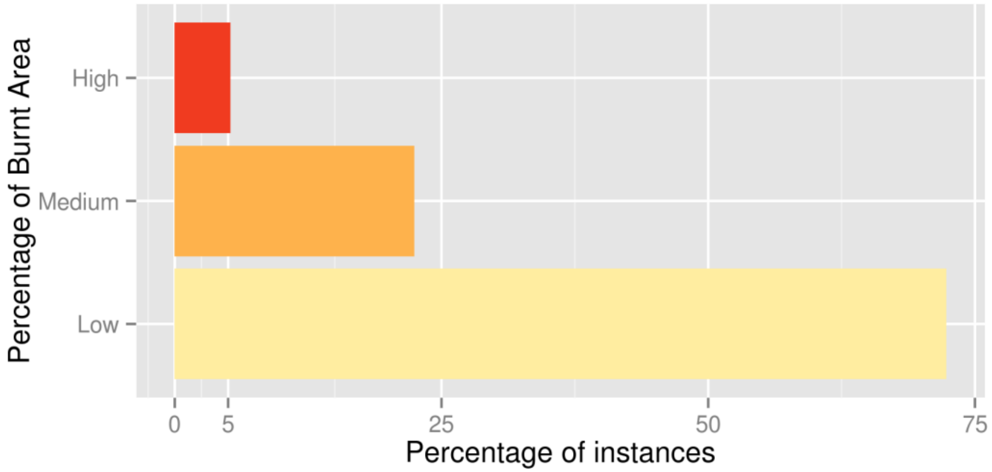
Convert data to Prolog → Design temporal and spatial predicates → **Design categorising predicate** → Model with Aleph (ar mode)

- *attribute(Parish, Category)* depending on *numAttribute(Parish, Value)*
- *attribute(Parish, Year, Category)* depending on num*Attribute(Parish, Year, Value)*

Describing wildfires

EXPERIMENTAL RESULTS

Experimental setup



Percentage of burnt area categories

- [0,5[- Very low
- [5,20[- Low
- [20, 40[- Medium
- [40, 100] – High

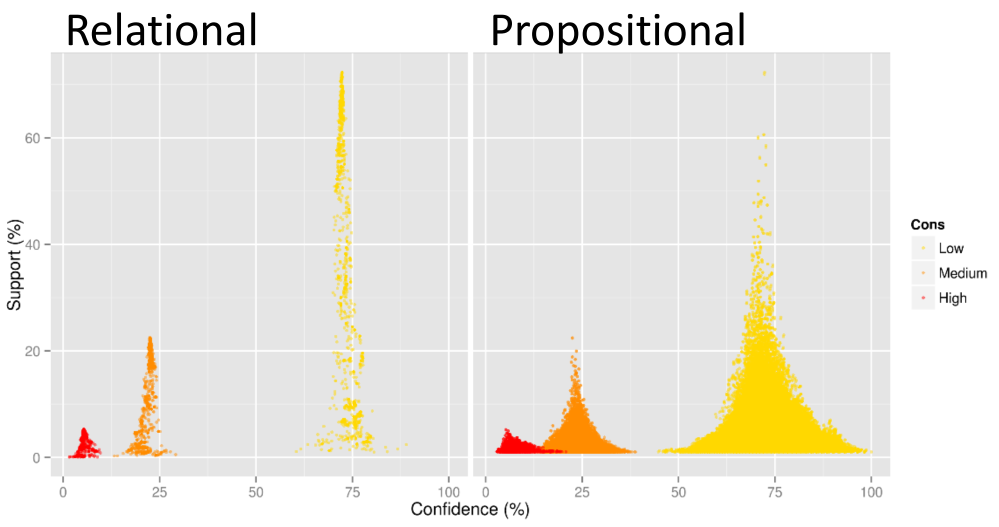
Consider only categories ≥ Low!

Minimum confidence set at 0.

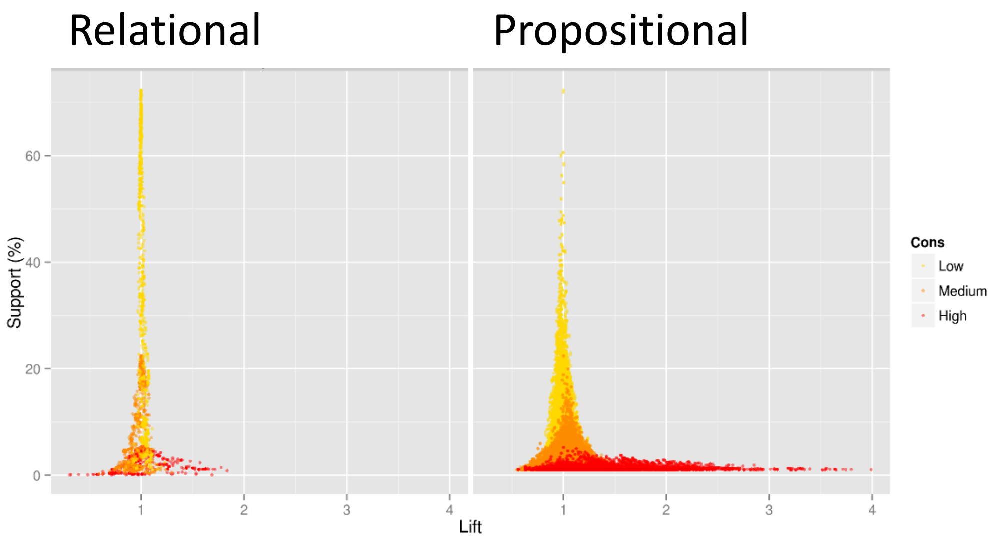
Consequent must be burnt area percentage.

Fixed minimum support (0.01)

SUPPORT VS CONFIDENCE



SUPPORT VS LIFT



Examples

PROPOSITIONAL

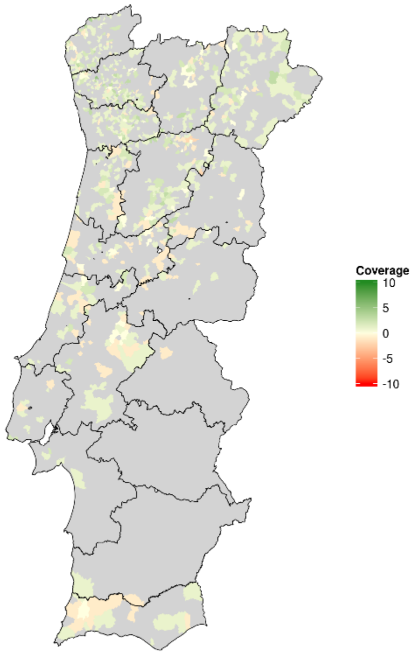
$\{Self = Very\ Low,$

$Caprine\ dens. = Very\ Low,$

$Meadow\ area = Very\ Low\}$

$\Rightarrow Burnt\ Area = Low$

$$supp = 0.15$$
$$conf = 0.8$$
$$lift = 1.1$$



RELATIONAL

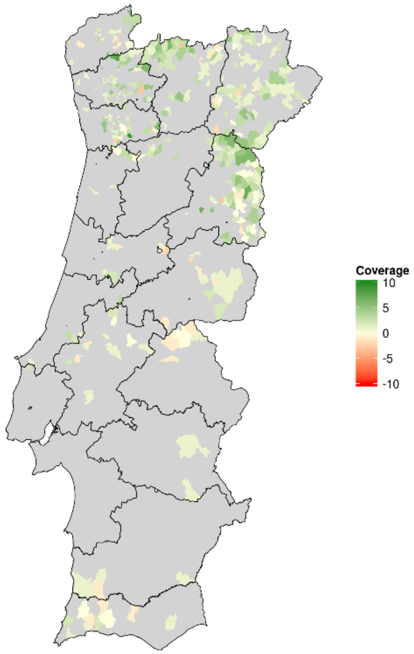
$pinewood(Parish, verylow),$

$fixedNeighbour(Parish, Neib),$

$yearsSinceFireLE(Neib, Year, 8)$

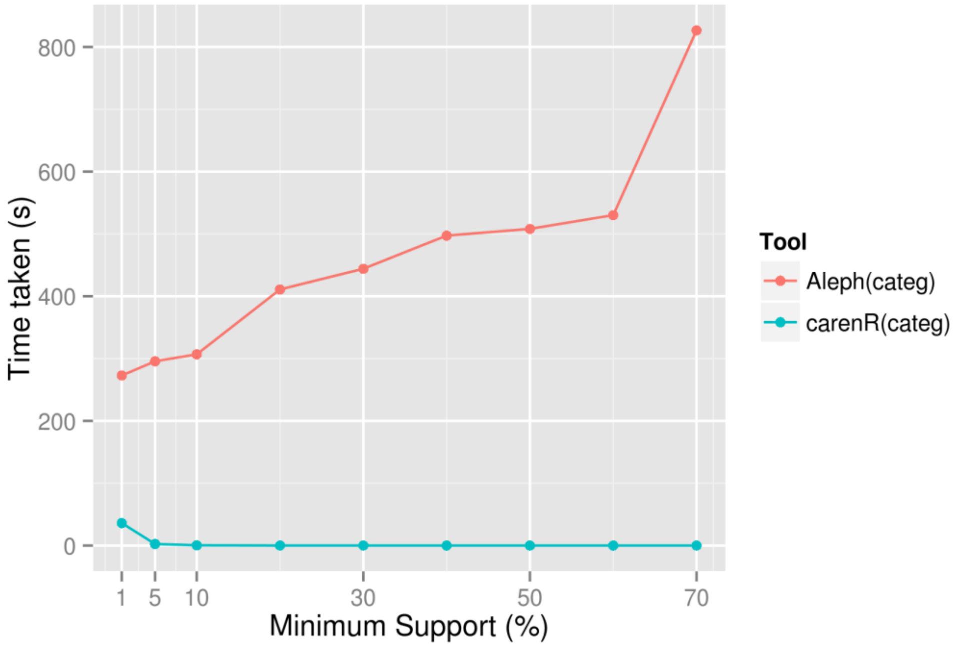
$\Rightarrow burntArea(Parish, Year, Low).$

$$supp = 0.18$$
$$conf = 0.78$$
$$lift = 1.1$$

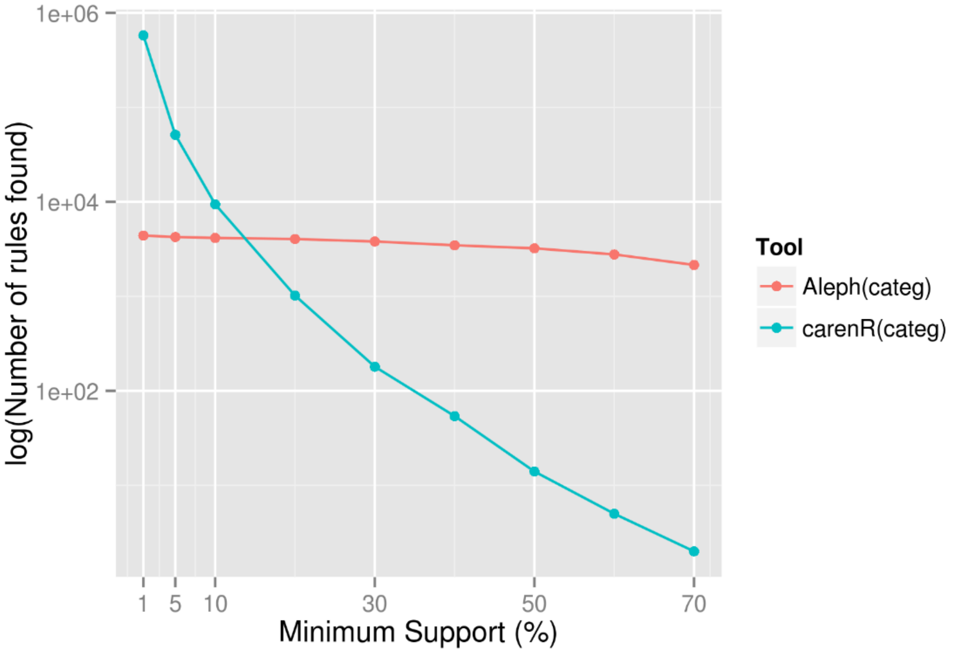


Varying minimum support

TIME TAKEN



RULES FOUND



Summary

PROPOSITIONAL APPROACH

- more time-efficient
- larger number of rules
- wider range of confidence and lift for rules with low support

RELATIONAL APPROACH

- more interpretable
- more expressive

Predicting wildfires

INTRODUCTION

Spatio-temporal forecasting (regression)

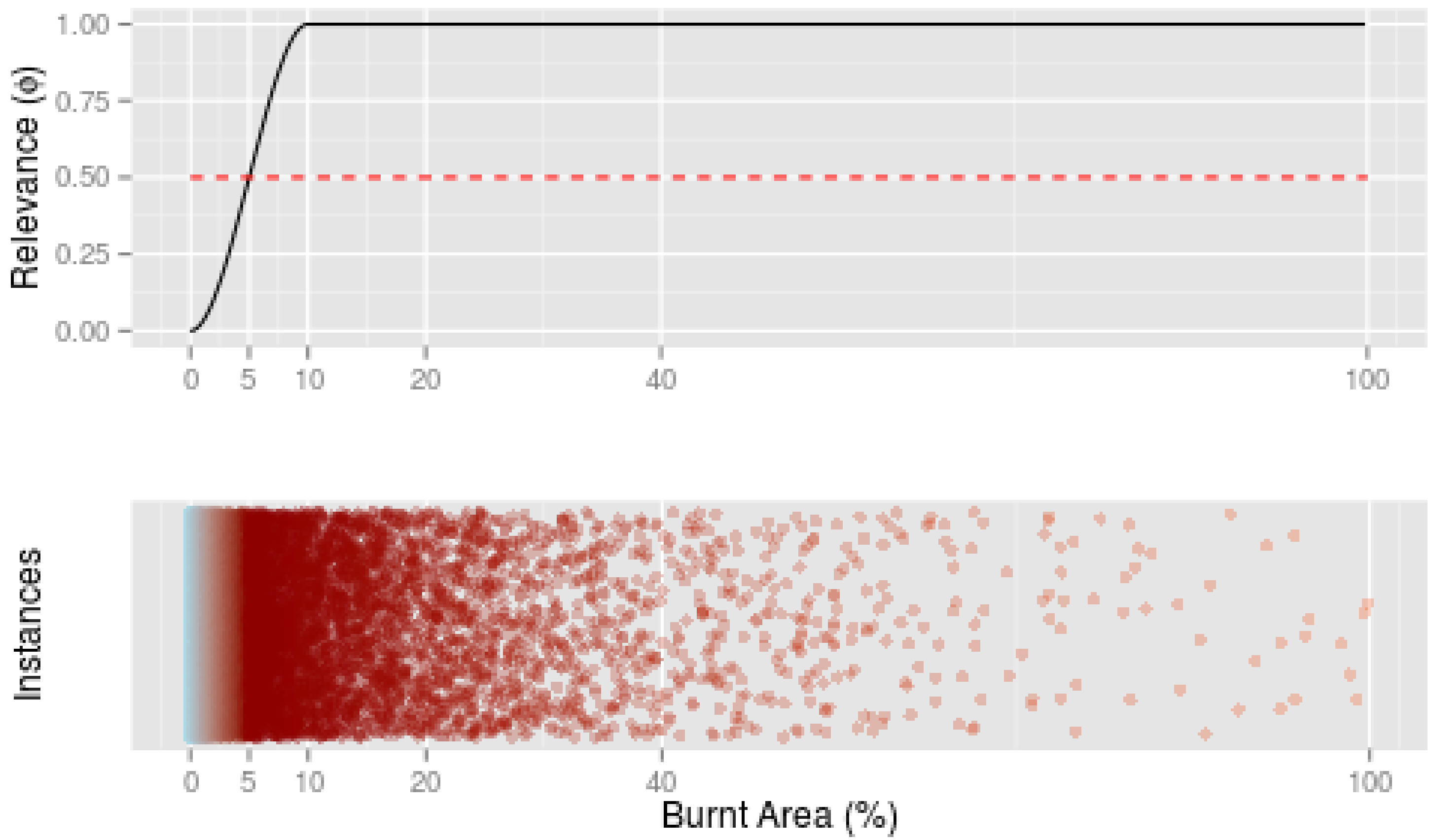
- **Propositional**
  - Pre-processing based
  - Spatio-temporal clustering based
  - Combined temporal and spatial methods
  - Integrated spatial and temporal dimensions
- **Relational**
  - Graphical models
  - ILP based

Predicting wildfires

PROPOSITIONAL AND RELATIONAL METHODS

Propositional approach

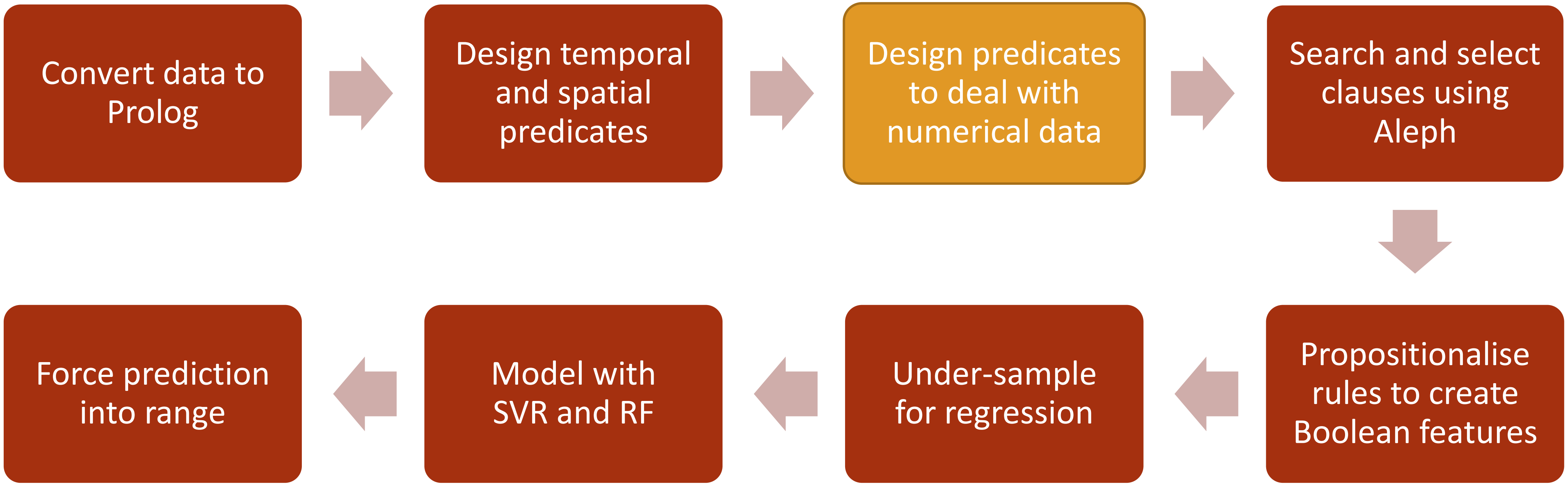
Build spatio-temporal indicators → Impute missing data → Under-sample for regression → Model with SVR and RF → Force prediction into range



Propositional approach

Under-sampling for regression proposed by Torgo *et al.* (2013) and implement in package UBL

Relational approach



Relational approach

Predicates for numerical data

- *attributeLE(Parish, Year, Value)* if *attribute* measured before or in *Year* was lesser or equal to *Value*
- *attributeGE(Parish, Year, Value)* if *attribute* measured before or in *Year* was larger or equal to *Value*

Relational approach

1. Convert data to Prolog
2. Design temporal and spatial predicates
3. Design predicates to deal with numerical data
4. Search and select clauses using Aleph
5. Propositionalise rules to create Boolean features
6. Under-sample for regression
7. Model with SVR and RF
8. Force prediction into range

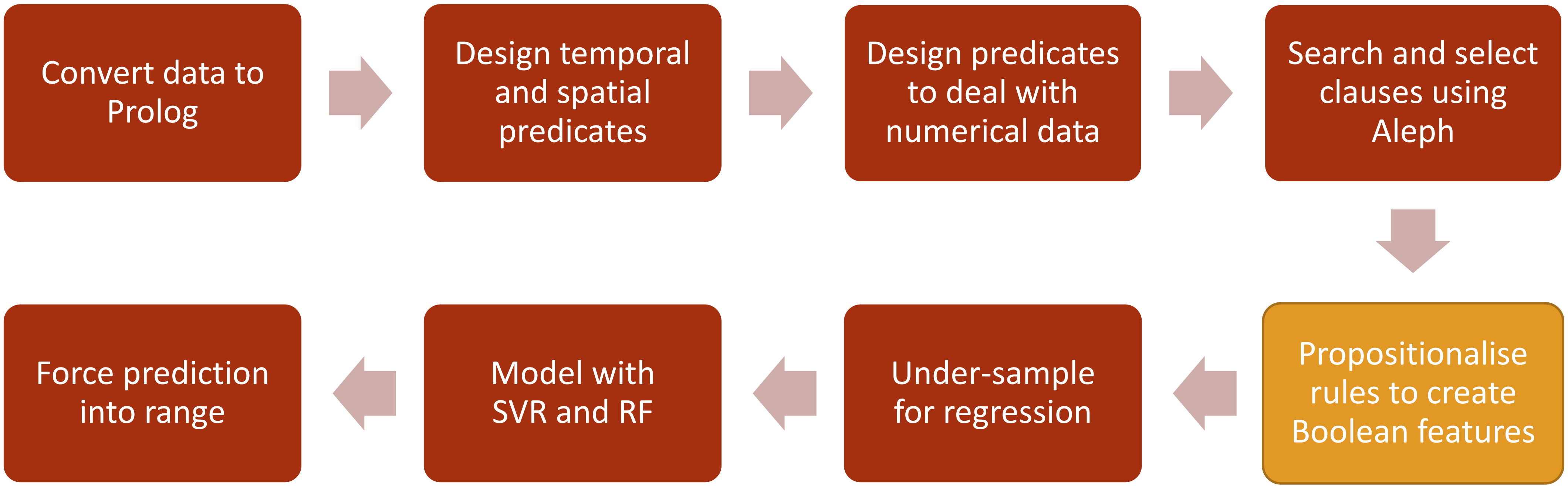
Relational approach

SEARCH AND SELECT CLAUSES

- Use random example as seed
- Saturate and reduce using $\boldsymbol{F_\beta - measure}$
- Save and select **best so far**
- Repeat 60 times for each $\beta \in \{0.75, 0.9, 1.0, 1.1, 1.25\}$

- $F_\beta = \frac{(1+\beta^2).precision.recall}{\beta^2.precision+recall}$
- $precision = \frac{TP}{TP+FP}$
- $recall = \frac{TP}{TP+FN}$

Relational approach

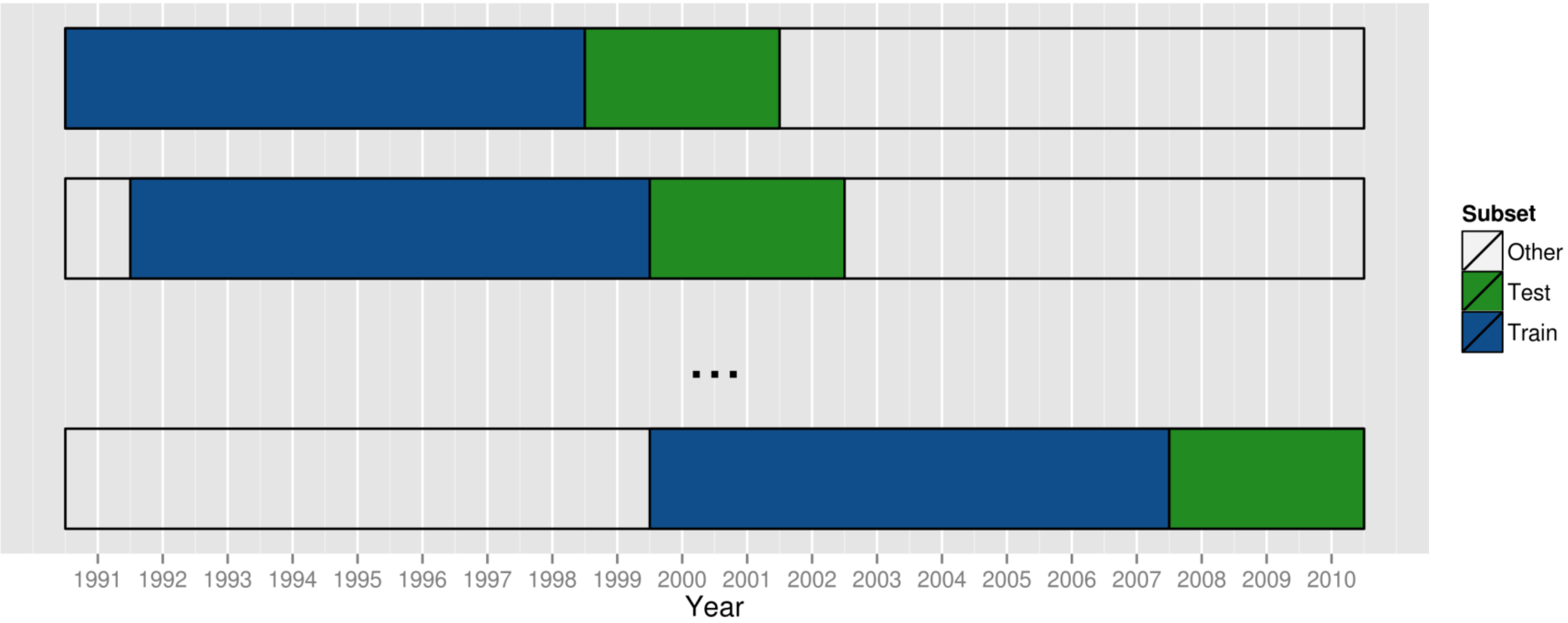


Predicting wildfires

EXPERIMENTAL RESULTS

Experimental setup

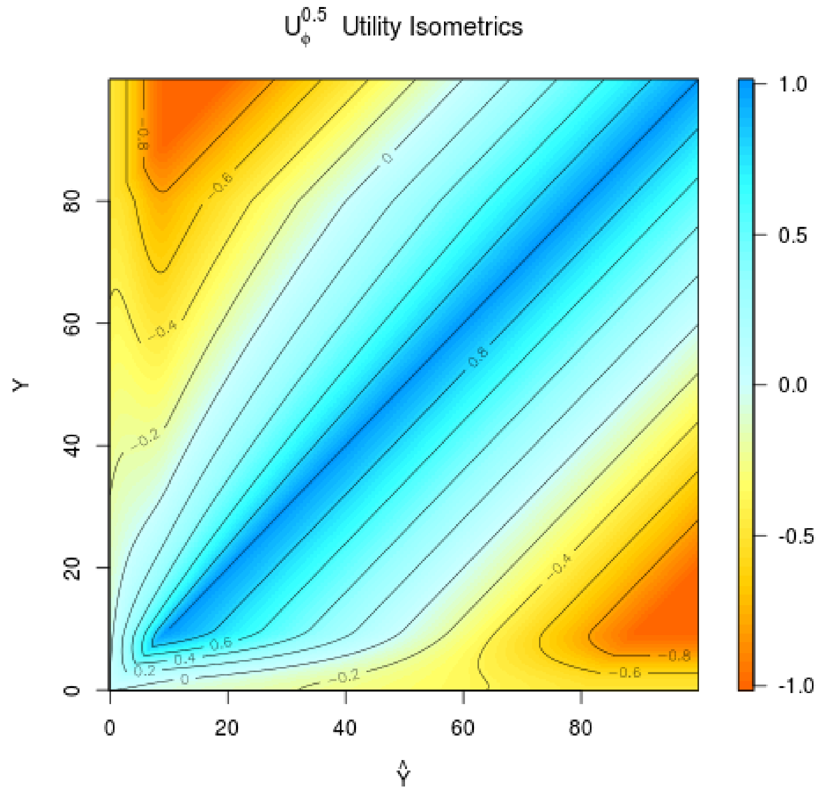
10 REPETITIONS



Performance metrics

$$precision_R = \frac{\sum_{\phi(\widehat{y_i})>t_R}(1+u_i)}{\sum_{\phi(\widehat{y_i})>t_R}(1+\phi(\widehat{y_i}))}$$

$$recall_R = \frac{\sum_{\phi(y_i)>t_R}(1+u_i)}{\sum_{\phi(y_i)>t_R}(1+\phi(y_i))}$$



$$U_\phi^p(\hat{y}, y) = B_\phi(\hat{y}, y) - C_\phi^p(\hat{y}, y)$$

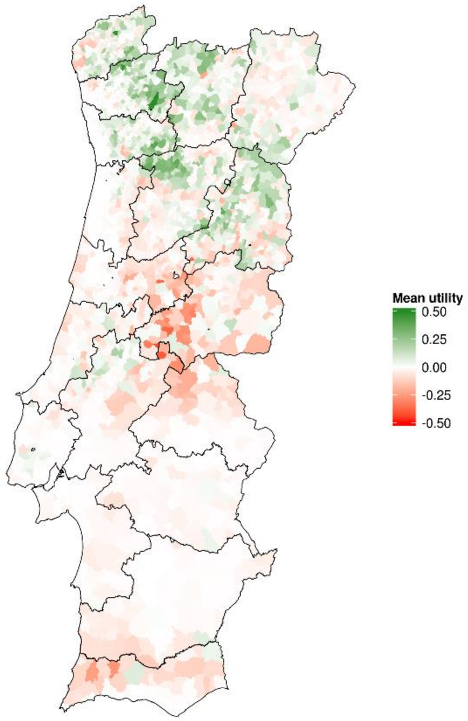
Results

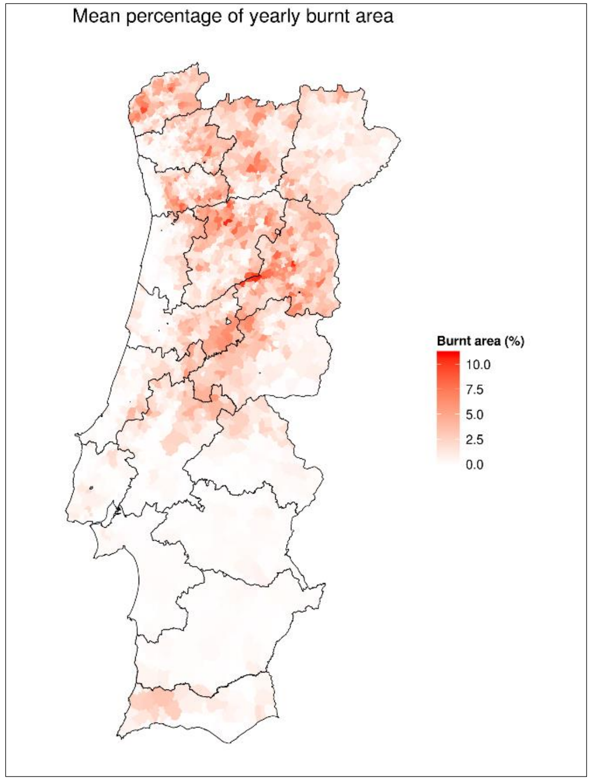
	Propositional				Relational			
			Under-sampling				Under-sampling	
	RF	SVR	RF	SVR	RF	SVR	RF	SVR
$\text{Precision}_\text{R}$	0.26	0.25	*0.65*	0.56	0.22	0.0082	*0.58*	0.45
Recall_R	0.69	0.74	*0.80*	0.78	0.71	0.65	*0.80*	0.76
$\mathbf{F_1\text{-measure}_R}$	0.38	0.37	**0.72**	0.65	0.34	0.016	**0.67**	0.57
Pre-processing time (s)	1.4e-3	1.4e-3	1.4e-3	1.4e-3	1.7	1.7	1.7	1.7
Training time (s)	2.8e-2	1.1e-1	2.2e-3	*3.2e-4*	5.4e-2	3.3e-2	5.4e-3	*3.0e-3*
Prediction time (s)	*1.5e-04*	1.1e-3	8.0e-5	4.3e-4	1.7e-4	5.6e-3	*1.3e-4*	2.0e-3
Total time (s)	3.1e-2	1.1e-1	3.7e-3	**2.2e-3**	1.8	1.7	**1.7**	1.7

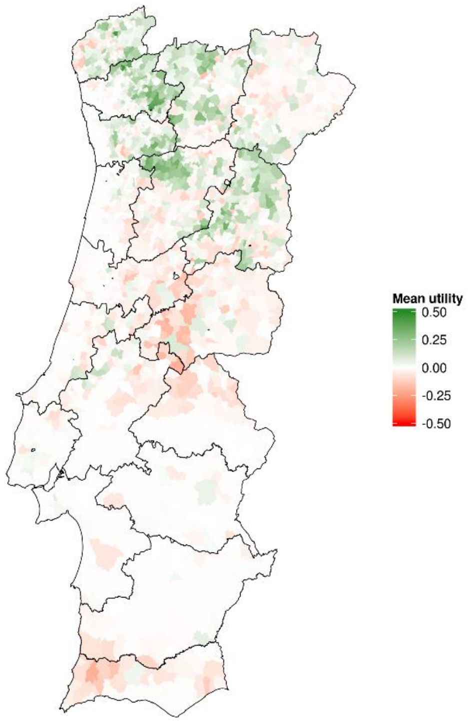
Example: under-sampling + RF

PROPOSITIONAL

RELATIONAL







Summary

- Comparable results, in spite of relational feature extraction optimised for classification
- Propositional approach more time-efficient again
- Under-sampling greatly improves results

Conclusion

Summary

- **Reviewed** the state-of-the-art;
- Developed and compared
  - a **propositional** methodology based on **pre-processing**, and
  - a **relational** methodology based on **ILP**

for

  - spatio-temporal **association rule learning**, and
  - spatio-temporal **forecasting** (regression).

Future research directions

- Explore other propositional approaches:
  - Extend the work of Oliveira & Torgo (2014) to include spatial dimensions;
  - Use clustering to select neighbourhoods as proposed by Appice *et al.* (2013).
- Explore other relational approaches:
  - Use graphical models such as Markov Logic Networks.
- Compare results in different domains to generalise our findings.

Thank you!

ACKNOWLEDGEMENTS

- PROF. LUÍS TORGO
- PROF. VÍTOR SANTOS COSTA
- DR. JOÃO TORRES
- PROF. RITA RIBEIRO
- PAULA BRANCO
- PROF. ALÍPIO JORGE
- PROF. PAULO AZEVEDO

Propositional and Relational Approaches to Spatio-Temporal Data Analysis

AUTHOR: MARIANA RAFAELA OLIVEIRA

SUPERVISOR: PROF. LUÍS TORGO

CO-SUPERVISOR: PROF. VÍTOR SANTOS COSTA

COMPUTER SCIENCE DEPARTMENT

FACULTY OF SCIENCES, UNIVERSITY OF PORTO

DECEMBER 2015