Discovery of Patterns in Urban Traffic

Francisco Miguel Consciência Neves

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Rui Miguel Carrasqueiro Henriques
Dr. Anna Carolina Nametala Finamore do Couto

Examination Committee

Chairperson: Prof. Alberto Manuel Rodrigues da Silva
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Sara Alexandra Cordeiro Madeira

September 2020

Acknowledgments

A conclusão deste trabalho reforça o sentimento de gratidão que me acompanhou não só durante a sua realização mas também durante todo o meu percurso académico. Às pessoas que me acompanharam durante este percurso, deixo os meus agradecimentos.

Começo por demonstrar o meu profundo agradecimento aos meus orientadores, o Professor Rui Henriques e a Professora Anna Carolina Finamore. O seu entusiástico apoio, motivação e confiança foram insubstituíveis para mim. Também a sua ajuda, ideias e visão científica foram essenciais para o desenvolvimento dos conteúdos desta tese. Para mim, vão ser sempre grandes modelos de humanidade e de trabalho.

Deixo também um agradecimento à equipa da DSI do IST. Em especial, ao Luis Cruz e ao Sérgio Silva, que foram meus coordenadores e que sempre demonstraram o seu apoio no meu sucesso académico. Eles partilharam comigo conhecimentos que foram essenciais para a conclusão desta etapa e que me vão ajudar para toda a minha vida.

Aos meus amigos mais próximos que me acompanharam ao longo de todos estes anos e que sempre foram um refúgio para mim, um obrigado mais uma vez por estarem presentes. Em especial, obrigado Carlota, Daniela, Joana, Mariana, Sofia, *Caju*, *Dani*, *Rafa* e Pedro.

Por fim, quero agradecer do fundo do coração aos meus pais, irmãos e avós. Mas, em especial, aos meus pais Fátima e Alfredo, com quem sempre pude contar. Eles estiveram sempre presentes para mim, oferecendo o seu incondicional apoio e amor, que me foi tão importante para a realização desta etapa. Agradeço também a ti Luna, por toda a companhia que me fizeste ao longo deste percurso.

Abstract

The comprehensive access to road traffic patterns in the continuously growing urban areas is key to achieve sustainable mobility. However, the inherent complexity of urban traffic data poses many challenges to achieve this goal, including: (i) the spatiotemporal intricacies of geolocated speed and loop count data; (ii) the need to integrate heterogeneous views of road traffic (such as speed limits, congestion size, delay, throughput); (iii) the need to mine jam patterns with varying degrees of severity; (iv) the inherent traffic variability and unexpected occurrence of events; (v) the need to guarantee the statistical significance, actionability and interpretability of the target patterns; (vi) the difficulty of detecting emerging patterns not yet markedly noticeable at early stages; (vii) and massive data size. This work proposes two methods for mining road traffic patterns from heterogeneous sources of spatiotemporal data, each one tackling the challenges presented previously in different ways. The first method explores the relevance of using biclustering for mining traffic patterns of road mobility. The second method proposes E2PAT, a scalable method to detect emerging patterns from heterogeneous sources of spatiotemporal data generated by large sensor networks. These contributions are comprehensively assessed in the context of the Lisbon's road traffic monitoring system, which features a large-scale network of mobile and fixed sensors that produce geolocalized speed data and loop counter data.

Keywords

sustainable mobility; spatiotemporal pattern mining; heterogeneous sensor networks; emerging patterns; biclustering; road traffic data; geolocalized time series; timestamped trajectory data.

Resumo

O acesso a padrões de tráfego rodoviário nas áreas urbanas é fundamental para alcançar uma mobilidade sustentável. No entanto, a complexidade inerente dos dados de tráfego urbano apresenta inúmeros desafios para a sua descoberta, incluindo: (i) a natureza espacial e temporal dos dados de velocidade geolocalizada e das espiras; (ii) a necessidade de integrar vistas heterogéneas do tráfego rodoviário (como limites de velocidade, tamanho de congestionamentos, atrasos, fluxo); (iii) a necessidade de extrair padrões de congestionamento com vários graus de severidade; (iv) a variabilidade inerente do tráfego e a ocorrência inesperada de eventos; (v) a necessidade de garantir a significância estatística, praticabilidade e interpretabilidade dos padrões; (vi) a dificuldade em detectar padrões emergentes, com mudanças ainda não marcadas o suficiente em estados iniciais; (vii) e o tamanho massivo dos dados. Este trabalho propõe dois métodos para a descoberta de padrões de tráfego rodoviário de fontes heterogéneas de dados espaço-temporais, cada um enfrentando os desafios apresentados anteriormente de maneira diferente. O primeiro método explora a relevância do uso de biclustering para extrair padrões de tráfego de mobilidade rodoviária. O segundo método propõe o E2PAT, um método escalável para detetar padrões emergentes de fontes heterogéneas de dados espaço-temporais gerados por grandes redes de sensores. Estas contribuições são avaliadas de forma exaustiva no contexto do sistema de monitorização de tráfego rodoviário de Lisboa, que inclui uma grande rede de sensores móveis e fixos que produzem dados de velocidade geolocalizada e dados de espiras.

Palavras Chave

mobilidade sustentável; descoberta de padrões espaço-temporais; redes de sensores heterogéneos; padrões emergentes; biclustering; dados de tráfego rodoviário; séries temporais geolocalizadas; dados de trajetória com registo de data e hora.

Contents

I Foundations 1

1 Introduction 3

1.1 Major Contributions 5

1.2 Organization of the Document 6

2 Background 7

2.1 Spatiotemporal data 9

2.1.1 Georeferenced time series 9

2.1.2 Trajectory 9

2.1.3 Spatiotemporal event data 9

2.2 Road traffic data 10

2.2.1 Inductive loop detector data 10

2.2.2 Geolocalized speed data 10

2.3 Biclustering 11

2.4 Emerging pattern mining 13

2.5 Road traffic patterns: qualities 14

3 Related Work 15

3.1 Classic approaches to traffic data analysis 17

3.2 Clustering-based approaches towards traffic data analysis 18

3.3 Pattern-centric approaches towards traffic data analysis 19

3.4 Context-sensitive analysis of urban traffic 20

3.5 Emerging pattern discovery 20

II Frequent Pattern Mining 23

4 Solution 27

4.1 Road traffic patterns using biclustering 29

4.1.1 Data mappings 29

4.1.2 Biclustering 30

4.2 On why, how and when to apply biclustering . . . 34

5 Results 37

5.1 Exploratory data analysis . . . 39

5.1.1 Loop detectors . . . 39

5.1.2 WAZE's geolocalized speed data . . . 42

5.2 Road traffic patterns . . . 44

5.2.1 ILD traffic patterns . . . 44

5.2.2 WAZE traffic patterns . . . 46

5.2.3 Integrative patterns of road traffic . . . 48

5.3 Statistical significance . . . 49

6 Situational Context 53

6.1 Context-incorporation mechanism . . . 55

6.2 Results . . . 56

III Emerging Pattern Mining 59

7 Solution 63

7.1 Spatiotemporal data mappings . . . 65

7.1.1 Spatial constraints . . . 65

7.1.2 Temporal constraints . . . 66

7.1.3 Data transformation . . . 66

7.2 Series differencing and emergence . . . 67

7.2.1 Comprehensive emerging pattern discovery . . . 67

7.2.2 Differencing . . . 68

7.2.3 Regression . . . 68

7.3 Integrative scoring . . . 69

7.4 Computation complexity of E2PAT . . . 70

8 Results 73

8.1 Emerging patterns of road traffic . . . 75

8.2 E2PAT visualization tool . . . 78

9 Situational Context 81

9.1 Context-aware filtering mechanism . . . 83

9.2 Results . . . 84

IV Conclusions 87

10 Concluding Remarks 89

10.1 Discussion 91

10.2 Future work 92

10.3 Scientific communication 92

List of Figures

2.1 Biclustering with varying homogeneity criteria: three biclusters were found under a constant, additive and order-preserving assumption. Illustrating, constant bicluster has pattern (value expectations) $\{c_1 = 1.05, c_2 = 0.45, c_3 = 0.9\}$ on x_2 and x_3 observations, while the order-preserving bicluster satisfies the $y_1 \geq y_2 \geq y_3$ permutation on $\{x_1, x_2, x_3\}$ observations. 11

4.1 ILD data mapping. 30

4.2 WAZE data mapping. 30

4.3 Illustrative road traffic patterns given by biclusters separately and integratively found in ILD and WAZE data (collected at Marquês do Pombal junction within the Lisbon city). . . 31

4.4 Effects of coherence strength and assumption on the resulting traffic patterns. 32

5.1 Location of loop detectors under analysis. Loop detectors will be mentioned by their street name: Av. Manuel da Maia(1) (top right), Av. António José de Almeida (top left), Av. Rovisco Pais (bottom left) and Av. Manuel da Maia(2) (bottom right). 39

5.2 Time Series plot of the four loop detectors under analysis. 40

5.3 Box plots of the four induction loops for weekdays and weekends. 41

5.4 Heatmap for the number of congested trajectories by day and hour in the city of Lisbon. . 42

5.5 Congestion snapshots on 27th September 2018, between 8-11 a.m. 43

5.6 Map visualization of the two sources of urban traffic data along the studied area (Marquês de Pombal): a) ILD sensor placement; b) WAZE jam events on peak hour (1/14/2020, 9AM). 44

5.7 Illustrative constant and order-preserving traffic patterns found in ILD data. 45

5.8 Three illustrative constant patterns of road traffic found in WAZE data. 47

5.9 Three illustrative order-preserving patterns of road traffic found in WAZE data. 48

5.10 Illustrative mobility patterns found from heterogeneous traffic data (event and time series traffic data), integrating views on traffic flow, speed and jam extension. 50

5.11 Statistical significance versus size of the collected constant patterns of road traffic ($|L|$=3 and 70% of quality). 51

6.1 Time series plot of normalized meteorological variables between 27th October of 2018 and 8th November of 2018. 55
6.2 Meteorological data mapping. 56
6.3 Illustrative patterns found using context-incorporation on geolocalized speed data. 58

7.1 Original and transformed spatiotemporal data structures. 67
7.2 Role of time series differencing operations for detecting emerging trends using linear models. . . . 68

8.1 Map visualization of the two sources of urban traffic data along the studied area: a) ILD sensor placement; b) WAZE events. 75
8.2 Illustrative set of patterns found by E2PAT in consolidated ILD-WAZE data. 77
8.3 Overview of the user dashboard for querying the road traffic data sources. 78
8.4 Map visualization of the found patterns from both ILD-WAZE data sources using score-based coloring of point-based and trajectory-based emerging patterns. 79

9.1 Overview of the interface for filtering road traffic data by situational context. 83
9.2 Map visualization of the geolocalized speed data: a) without situational context; b) filtered by road restrictions (visually depicted in black); c) filtered by air humidity superior to 80%. 84

List of Tables

5.1 Properties of biclustering solutions in ILD data using BicPAMS with varying homogeneity criteria. . . . 45

5.2 Properties of biclustering solutions in WAZE data using BicPAMS with varying homogeneity criteria. . . . 46

5.3 Biclustering results from consolidated ILD and WAZE data using BicPAMS with different homogeneity criteria. . . . 49

6.1 Illustrative example to demonstrate the concatenation between road traffic data and situational context mask. . . . 57

8.1 Top 20 congestion and decongestion traffic patterns in the studied area under a weekday calendar. 76

9.1 Top 5 congestion and decongestion traffic patterns without situational context. . . . 85

9.2 Top 5 congestion and decongestion traffic patterns with filtered road restrictions. . . . 85

9.3 Top 5 congestion and decongestion traffic patterns for high humidity conditions (>80%). . . . 85

Acronyms

ILD	Inductive Loop Detector
CML	Lisbon City Council
E2PAT	Emerging Event Pattern Miner
EP	Emerging Pattern
PSE	Planned Special Events
STCS	Space-Temporal Congestion Subgraphs
GPS	Global Positioning System
BicNET	Biclustering Networks
EDA	Exploratory Data Analysis
BicPAMS	Biclustering based on Pattern Mining Software
BSig	Biclustering Significance
CSV	Comma-separated Values

Part I

Foundations

1

Introduction

Contents

1.1 Major Contributions 5

1.2 Organization of the Document 6

Mobility in most capital cities is not yet sustainable. Road mobility is susceptible to significant externalities, causing daily congestions, in turn aggravating air pollution, accessibility problems, traffic noise, and safety hazards [1, 2]. Motivated by this observation, many cities are establishing initiatives to collect heterogeneous sources of urban data to comprehensively monitor road traffic [3, 4]. Among them, the Lisbon City Council (CML) is currently able to gather and consolidate different views on road traffic data along the city from mobile sensors, road cameras, and loop counters.

Despite the relevance of these heterogeneous views to understand road traffic dynamics, the comprehensive discovery of traffic patterns is hampered by numerous challenges: (i) The inability of traditional pattern mining methods to handle the spatiotemporal intricacies of geolocalized speed and loop count data; (ii) The need to combine multiple aspects of road traffic, including speed limits, congestion size, duration, as well as frequentist views on traffic flow; (iii) The need to discover road traffic patterns sensitive to varying jam levels; (iv) The need to mine patterns robust to the inherent traffic variability and sporadic occurrence of unexpected events; (v) The need to find comprehensive sets of road traffic patterns with guarantees of statistical significance, actionability and interpretability; (vi) The need for a robust and timely detection of emerging patterns [5]; (vii) The massive size of data produced by traffic monitoring systems.

1.1 Major Contributions

This work proposes two methods to comprehensively discover patterns from heterogeneous sources of road traffic data, each addressing several of the challenges stated previously. The first method explores the use of biclustering to unravel traffic patterns. In this method, a traffic pattern is defined as a recurring congestion profile, possibly spanning diverse locations and time periods within a day. Biclustering, the discovery of coherent subspaces within real-valued data, has unique properties of interest, thus being positioned to unravel such traffic patterns, while satisfying the aforementioned challenges. Despite its relevance, the potentialities of applying biclustering in the mobility domain remain unexplored.

The second method, referred as E2PAT (Emerging Event PATtern miner), proposes a scalable method to comprehensively detect emerging patterns from heterogeneous sources of spatiotemporal data generated by large sensor networks, in particular in the context of the Lisbon's road traffic monitoring system. We combine simplistic time differencing and spatial intersection principles to identify all emerging patterns distributed along geographies of interest. We show that the use of these principles guarantee a linear-time efficiency of E2PAT on the size of the input data. In addition, we propose an integrative score to measure the relevance of emerging patterns and show its role to support pattern retrieval, promote usability, and guarantee the actionability of the found patterns.

Road traffic dynamics are also largely influenced by situational context such as public events (e.g.

sport events and concerts), bottlenecks in the roads (e.g. accidents and maintenance works), and weather conditions. Integrating this view in the discovery of road traffic patterns gives us the unprecedented opportunity to further leverage the actionability and relevance of the discovered patterns.

In addition, given the inherent dependence of road traffic with different sources of situational context (e.g. weather; road traffic interdictions; public events) we extended our methods to encompass situational context in the pattern discovery task.

1.2 Organization of the Document

This thesis is organized in four parts, each part being divided in chapters. Part I is divided in three chapters. It starts on Chapter 1 with an introductory note on the thesis, then on Chapter 2 we present some essential concepts that are used throughout our work, and finally Chapter 3 reviews related work within urban traffic analysis. Part II and Part III expose the two introduced methods developed in the context of this thesis, each divided in three chapters:

1. Solution, where we detail the implementation of the method;
2. Results, where the results of the methods' application is gathered and discussed;
3. Situational Context, where we detail how we extended each method to integrate situational context in the pattern discovery task.

Part IV summarizes the achievements of this work and points out some possible future directions.

2

Background

Contents

2.1 Spatiotemporal data 9

2.2 Road traffic data 10

2.3 Biclustering 11

2.4 Emerging pattern mining 13

2.5 Road traffic patterns: qualities 14

This chapter provides essential background regarding the proposed methods to discover patterns from road traffic data. Section 2.1 describes different types of spatiotemporal data structures. Section 2.2 details the two types of road traffic data that were used in this work. Section 2.3 introduces some important aspects of biclustering. Section 2.4 introduces the concept of emerging pattern. Finally, Section 2.5 describes road traffic patterns, as well as the desirable properties to be pursued during their discovery.

2.1 Spatiotemporal data

2.1.1 Georeferenced time series

A time series is an ordered set of observations $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_T)$, each observation $\mathbf{x}_t$ being recorded at a specific time point t. Time series can be *univariate*, $\mathbf{x}_t \in \mathbb{R}$, or *multivariate*, $\mathbf{x}_t \in \mathbb{R}^m$, where $m > 1$ is the multivariate order (number of variables). Time series recorded at a particular location are mentioned as georeferenced. A *georeferenced time series* is a tuple $GT = (\phi, \mathbf{x})$, where ϕ is a pair $(latitude, longitude)$ describing the location where the series $\mathbf{x}$ is being recorded.

Time series can be decomposed into *trend*, *seasonal*, *cyclical*, and *irregular components* using additive or multiplicative models [6]. Classical approaches for time series analysis generally rely on statistical principles, including *auto-regression*, *differencing* and *exponential smoothing* operations [7]

2.1.2 Trajectory

A *trajectory* is a sequence $\langle\phi_1, \phi_2, \cdots, \phi_n\rangle$, where ϕ_i is a pair $(latitude, longitude)$. A timestamped trajectory, $\langle(\phi_1, t_1), \cdots, (\phi_n, t_n)\rangle$, has its coordinates, ϕ_i, annotated with a timestamp, t_i.

Floating car data are paradigmatic examples of timestamped trajectory data produced from mobile devices with active global positioning systems (GPS), gathering the position of vehicles along time. Methods for producing floating car data from GPS information generally produce rather sparse trajectories that need to be completed within the constraints of the road network mesh [8–10].

2.1.3 Spatiotemporal event data

An *event* is a tuple $e = (\mathbf{x}, s, \tau)$, where:

- $\mathbf{x} = (x_1, \cdots, x_m)$ is the observation, either *univariate* (m=1) or *multivariate* ($m > 1$) depending on the number of monitored variables. For instance, given speed (y_1) and throughput (y_2) variables, an illustrative observation is $\mathbf{x}$=(x_1=15km/h, x_2=10cars/min);

- s is the *spatial extent* of the observation $\mathbf{x}$. The spatial extent s can be any spatial representation associated with the event, such as a *geographic coordinate* or a *trajectory*;

- τ is the *temporal extent* of the observation $\mathbf{x}$, either given by a time instant or a time interval.

A spatiotemporal *event dataset* is a collection of events, $E = \{e_1, e_2, \cdots, e_n\}$, each event producing a (multivariate) observation recorded along specific spatial and temporal context.

2.2 Road traffic data

2.2.1 Inductive loop detector data

Inductive loop detectors (ILDs), also referred as loop detectors or induction loops, are equipment installed under roads pavements that detect vehicle passages. Depending on the type of ILD, these equipment are able to detect volume, speed and classify vehicles passing. ILDs are relatively susceptible to failure rates in their estimations. Martin et al. [11] provide a detailed summary on loop detectors. ILD raw data are often aggregated to provide frequentist views on the cumulative number or average speed of different classes of vehicles on a given road along specific time intervals, i.e. georeferenced multivariate time series data. In the city of Lisbon, ILDs are placed on the major road junctions within the city and are calibrated to stream the number of passing vehicles for every period of 15 minutes in real-time.

Aggregated ILD data are a collection $\langle gt_1, gt_2, \cdots, gt_n \rangle$, where $gt_k = (\phi, \mathbf{x})$ is a georeferenced time series with m variables being monitored (e.g. number of vehicles) and T periods (e.g. intervals of 15 minutes).

2.2.2 Geolocalized speed data

Individual trajectories produced by mobile devices can be aggregated as spatiotemporal events, by recording specific features of interest (such as speed) extracted from devices circulating throughout the same trajectory segments at similar time periods [12]. Applications, such as GOOGLEMAPS[1], WAZE[2] or TOMTOM[3], installed in some of the mobile devices, offer localization and navigation facilities, providing an aggregate view of the ongoing traffic dynamics within the city.

Geolocalized speed data is a common example of data produced by aggregating individual trajectories' features. Geolocalized speed data is a collection of events where each event, $e_i = (\mathbf{x}_i, s_i, \tau_i)$, is a traffic jam event that occurred at time t_i in a trajectory (road segment) s_i. The set of observations

[1] https://www.google.com/maps
[2] https://www.waze.com/en-GB/
[3] https://www.tomtom.com/en_gb

$\mathbf{x}_i$ contains traffic information – such as the recorded speed, delay, severity level or road type – that characterizes the occurring jam.

2.3 Biclustering

Given a dataset defined by a set of observations $X = \{x_1, .., x_n\}$, variables Y=$\{y_1, .., y_m\}$, and elements $a_{ij} \in \mathbb{R}$ observed for observation x_i and variable y_j:

- a **bicluster** B=(I,J) is a $n \times m$ subspace, where $I = (i_1, .., i_n) \subseteq X$ is a subset of observations and $J = (j_1, .., j_m) \subseteq Y$ is a subset of variables;
- the **biclustering** task aims at identifying a set of biclusters $\mathcal{B} = (B_1, .., B_s)$ such that each bicluster $B_k = (I_k, J_k)$ satisfies specific criteria of *homogeneity*, *dissimilarity* and *statistical significance*.

Homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster [13]. Merit functions are typically applied to guide the formation of biclusters in greedy and exhaustive searches. In stochastic approaches, a set of parameters that describe the biclustering solution are learned by optimizing a merit (likelihood) function.

The pursued homogeneity determines the coherence, quality and structure of a biclustering solution [14]. The *coherence* of a bicluster is determined by the observed form of correlation among its elements (coherence assumption) and by the allowed value deviations from perfect correlation (coherence strength). The *quality* of a bicluster is defined by the type and amount of accommodated noise. The *structure* of a biclustering solution is defined by the number, size, shape and positioning of biclusters. A flexible structure is characterized by an arbitrary number of (possibly overlapping) biclusters. These concepts, formalized below, are illustrated in Figure 2.1.

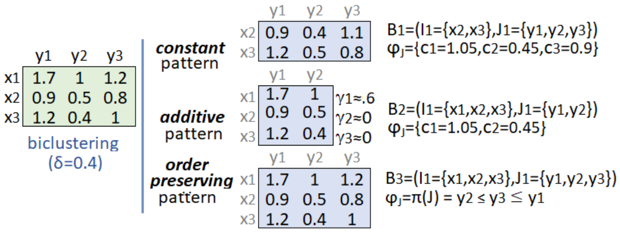


Figure 2.1: Biclustering with varying homogeneity criteria: three biclusters were found under a constant, additive and order-preserving assumption. Illustrating, constant bicluster has pattern (value expectations) $\{c_1 = 1.05, c_2 = 0.45, c_3 = 0.9\}$ on x_2 and x_3 observations, while the order-preserving bicluster satisfies the $y_1 \geq y_2 \geq y_3$ permutation on $\{x_1, x_2, x_3\}$ observations.

Given a dataset, the elements within a bicluster $a_{ij} \in (I, J)$ have coherence across variables (**pattern on observations**) if w_{ij}=c_j+γ_i+η_{ij}, where c_j is the expected value of variable y_j, γ_i is the adjustment for observation x_i, and η_{ij} is the noise factor of w_{ij}.

A bicluster has **constant coherence** when γ_i=0 (or γ_j= 0), and **additive coherence** otherwise, $\gamma_i \neq 0$ (or $\gamma_j \neq 0$).

Let r be the amplitude of values of the input data, **coherence strength** is a value $\delta \in [0, r]$ such that $a_{ij} = c_j + \gamma_i + \eta_{ij}$ where $\eta_{ij} \in [-\delta/2, \delta/2]$.

Given a real-valued dataset, a bicluster $B = (I, J)$ satisfies the **order-preserving coherence** assumption iff the values for each observation in I follow the same ordering π along the subset of variables in J.

Figure 2.1 instantiates the introduced concepts, illustrating biclusters with constant, additive and order-preserving coherence (right) found in real-valued data (left). The pattern of each bicluster is further provided.

The bicluster **pattern** φ_J is the set of expected values in the absence of adjustments and noise $\{c_j \mid y_j \in J\}$. Consider the illustrative biclusters B_1, B_2 and B_3 in Figure 2.1. Their patterns are respectively given by φ_{B_1}=$\{c_1$=1.05,c_2=0.45,c_3=0.9$\}$, φ_{B_2}=$\{c_1$=1.05,c_2=0.45$\}$ (assuming a_{ij}=$c_j + \gamma_i$ and additive factors γ_1=0.65, γ_2=0 and γ_3=0) and φ_{B_3}=$(y_2 \leq y_3 \leq y_1)$.

Statistical significance criteria, in addition to homogeneity, guarantee that the probability of a bicluster's occurrence (against a null data model) deviates from expectations [15].

Finally, ***dissimilarity*** criteria can be further placed to guarantee the comprehensive discovery of non-redundant biclusters [16].

Following Madeira and Oliveira's taxonomy [13], existing biclustering algorithms can be categorized according to the pursued homogeneity criteria and type of search. Hundreds of biclustering algorithms were proposed in the last decade, as shown by recent surveys [17, 18].

In recent years, a clearer understanding of the synergies between biclustering and pattern mining paved the rise of a new class of algorithms, generally referred to as **pattern-based biclustering** algorithms [14]. Pattern-based biclustering algorithms are inherently prepared to efficiently find exhaustive solutions of biclusters and offer the unprecedented possibility to affect their structure, coherency and quality [19]. This behavior explains why this class of biclustering algorithms are receiving an increasing attention in recent years [14]. BicPAMS (Biclustering based on PAttern Mining Software) consistently combines these state-of-the-art contributions on pattern-based biclustering [16].

Biclustering on traffic data. A traffic pattern produced by biclustering is a coherent form of traffic behavior that satisfies a specific criterion of frequency, where frequency is often represented by a form of temporal or spatial recurrence. An illustrative and self-explanatory road traffic pattern is:

$<$ (*jam extent in [1.5km,2km]* | *location* $\phi_1, [17h, 18h]) \wedge$
(*speed limit in [15km/h,20km/h]* | *trajectory* $T_A, [10h, 11h]) >$
with recurrence in [Mondays,Fridays].

In alternative to congestion extent and speed limits, patterns may further capture restrictions on vehicle passage flow, average traffic delay per distance, or severity.

Integrative patterns of road mobility combining heterogeneous traffic views should be also pursued. For instance, a low number of cars passing on a given road may be explained by a heightened speed limitation on that same road, which in turn may be explained by the spatial extent of traffic on a nearby location.

2.4 Emerging pattern mining

Emerging Patterns (EPs) were firstly introduced by Dong et al. [20] in the context of multivariate observations collected from two periods/datasets. An emerging pattern was in this context defined as a multivariate pattern whose support suffered a significant change between the two given periods.

This work extends this early notion of emerging pattern to encompass an arbitrary number of time periods and to further incorporate spatial information. Given a spatiotemporal dataset, an **emerging pattern** is a set of spatially correlated observations whose values satisfy specific *growth*, *fitness* and *support* criteria along time.

The *growth* criterion defines the rate at which observations change along time. For instance, given a specific location and periodicity, a growth rate of 1% indicates that the values of a given observation increase 1% on every period under assessment.

Given a specific growth rate, the *fitness* (error) criterion defines how well observations follow (deviate) from the given expectations. For instance, fluctuations of the observed values around the expected values produce residues that can be used to characterize the fitness (error) of a given emerging pattern. Emerging patterns below a given accuracy threshold may be spurious findings and should therefore be discarded.

Finally, *support* criterion defines the number of observations (temporal extent) satisfying the given growth and accuracy criteria. In this context, emerging pattern discovery can be applied under minimum growth, accuracy and support thresholds.

Emerging patterns on traffic data. An emerging pattern of road traffic is a coherent form of traffic behavior that satisfies a specific criterion of periodicity, frequency, or growth. An illustrative emerging pattern of road traffic is:

$$\{(\textit{speed limit decrease at weekly growth rate 2\%} \mid \textit{trajectory } \mathrm{s}_A),$$

$$(\textit{traffic throughput increase at weekly growth rate 3\%} \mid \textit{location } \phi_B)\}$$

$$\textit{where} = \text{Areeiro}, \ \textit{when} = (\ [10\text{h},11\text{h}[\ \wedge \text{Mondays})$$

$$\textit{satisfying} \ \ r^2 > 0.5 \wedge \text{support} > 10$$

where the coefficient of determination is used as the fitness criterion and at least 10 observations (support criterion) are necessary to infer the observed growth rates, 2% and 3% from congestions on a segment s_A and location ϕ_B within the Areeiro region at Mondays, 10h.

In alternative to speed limits and traffic flow, emerging patterns of road traffic may further capture growing mobility restrictions associated with the congestions' extent, recurrence, average delay per distance, and severity.

2.5 Road traffic patterns: qualities

Given a spatiotemporal dataset, a *pattern* is a spatially correlated set of frequent, periodic or coherently changing observations along time. Illustrating, periodic patterns describe recurrent behavior over regular time intervals at certain locations or trajectories. Graph patterns are sets of trajectories or locations within the target traffic monitoring network that are frequently co-associated with an event of interest (such as co-occurring congestions). Temporal association rules define hypothetical causal relationships between correlated frequent events at nearby locations or trajectories. Each of this pattern solutions are useful to gather different views that offer relevant opportunities to study road traffic patterns.
Patterns should satisfy a number of properties of interest, to ensure their quality and relevance:

- *non-triviality* (novelty) and *actionability* (support decisions);
- *robustness* (bounded noise tolerance);
- *statistical significance* (excluded spurious patterns that occur by chance);
- *interpretability*;
- *coverage* (complete solutions spanning different geographies and time periods);
- *efficiency* of the pattern retrieval process.

3

Related Work

Contents

3.1 Classic approaches to traffic data analysis 17

3.2 Clustering-based approaches towards traffic data analysis 18

3.3 Pattern-centric approaches towards traffic data analysis 19

3.4 Context-sensitive analysis of urban traffic 20

3.5 Emerging pattern discovery 20

The discovery of actionable patterns of urban mobility has received particular attention in recent years with the increased availability of urban data, advances on spatiotemporal data analysis, and global pressure towards sustainability [4]. Yang et al. [21] define mobility patterns as ”an abstraction of human movement's spatiotemporal regularity according to human's historical trajectories”. In addition to individual trajectories from mobile users data [21, 22], alternative sources of urban data are being unprecedentedly consolidated by world city Councils and subjected to pattern recognition, including: smart card data from integrated validation systems in public carriers [23]; aggregate event statistics from free GPS systems such as GoogleMaps and WAZE [12]; trajectories from GPS-equipped public bicycles and taxis [24]; and traffic data from ILD and cameras found along the major arteries of cities. Understanding the patterns of human motion, both globally and individually, is crucial for different purposes, among them urban planning [21], traffic forecasting [25], and monitoring the spread of an epidemic [26].

Although interest in mobility patterns dates back one century [27], their automated discovery is considered a recent research area [28]. Below, we group recent contributions on the discovery of urban traffic patterns along three major categories: classic/statistical approaches (Section 3.1), clustering-based approaches (Section 3.2) and pattern-centric approaches (Section 3.3) for understanding urban mobility patterns. Then, Section 3.4 surveys works that integrate context in urban traffic analysis. Finally, on Section 3.5 we show contributions on the discovery of emerging patterns.

3.1 Classic approaches to traffic data analysis

Classic approaches make use of statistics, parametric models and visualization principles to understand spatiotemporal traffic dynamics. Liao et al. [4] introduced a data fusion approach encompassing real-time traffic data and travel demand (estimated from Twitter data) that statistically assesses the difference in private versus public travel time for retrieving spatiotemporal patterns of time discrepancy. To this end, time-annotated origin-destination matrices are inferred for four cities: São Paulo, Stockholm, Sydney, and Amsterdam. Gonzalez et al. [22] analyzed trajectories of 10,000 mobile phone users for a six month period. Inspired by the work of Mantegna and Stanley [29], they identified prominent statistics, including returning peaks, to assess population's mobility patterns. Dozens of additional studies on traffic flow along major cities have been more recently conducted [30–34]. Li et al. [35] suggest categorization of traffic flow studies in microscopic-level studies (e.g. car-following models and lane-changing models), mesoscopic-level studies (e.g. headway/spacing distributions), and macroscopic-level studies (e.g. fundamental diagram and traffic wave models). They also highlight the changes in traffic flow models occurred from GPS-based and video-based trajectory data.

Guo et al. [24] proposed visualization principles to analyse a large point-based origin-destination dataset collected from taxi rides in Shenzhen, China. Unlike most taxi trajectory datasets, this study contains only the origin and destination points per trip. To this end, they apply spatial clustering to transform GPS points into meaningful regions, upon which they compute and plot statistics such as inflow, outflow, and flow ratio along different periods of the day. Hasan et al. [23] provided visualization facilities to understand spatiotemporal mobility patterns gathered from smart card transactions in London's public transportation system. Models for intermodal transportation networks proposed within previous works [36–38] can be used to extend this work for multiple carriers.

Research on traffic predictive models with guarantees of interpretability also offer the possibility of unraveling mobility patterns. Salamanis et al. [25] propose a method to predict traffic under normal and abnormal conditions differing in type, severity and duration. To tackle the issue of abnormalities, their method discovers traffic patterns that occur when an abnormal event of a specific class occurs using open traffic data from Performance Measurement System (PeMS) in California, spanning a period of 10 years.

3.2 Clustering-based approaches towards traffic data analysis

Clustering methods have the potential to unsupervisedly discover regions of interest, making them candidates to offer discrete views of urban traffic data. Necula et al. [12] applied clustering to identify statistically significant traffic patterns given by a contiguous road segments with similar traffic load over time from 10,000 GPS traffic traces of vehicles from New Haven County, Connecticut, USA. Rempe et al. [39] propose a graph-based approach to detect vulnerable parts of the road network, named by the authors as congestion clusters. To identify these vulnerable areas, the authors use spatial smoothing to compute areas with recurrent jams over time, termed congestion pockets. From the found time-dependent congestion pockets, congestion clusters are inferred, and their statistics computed (e.g. starting and ending time distributions) and visualized. Song et al. [3] propose the use of hierarchical clustering to mine spatiotemporal patterns of traffic congestion using multi-source data collected from Beijing, China. Once these patterns are discovered, geographical associations are retrived and assessed against influential factors (such as density, design, diversity, among others). Habtemichael et al. [40] introduce a short-term traffic forecaster based on clustering, winsorization, and rank exponent sensitive to traffic profiles over 36 freeway datasets from UK and USA.

Despite the relevance of the surveyed works, clustering-based approaches impose similarity to be assessed on a daily basis, preventing the discovery of non-trivial, statistically significant and time-sensitive associations.

3.3 Pattern-centric approaches towards traffic data analysis

Gowtham et al. [41] conducted a survey on spatiotemporal pattern mining algorithms. In the context of urban mobility, researchers have extended classic pattern mining algorithms to successfully discover co-occurring and sequential patterns in urban traffic data. According to Treiber and Kesting [42], the discovery of such traffic patterns can be used as features to improve descriptive and predictive mobility models. Contributions from alternative spatiotemporal data domains can provide important principles to this end, including research developed on the discovery of spatial dynamics of complex geographic phenomena. For instance, He et al. [43] proposed an event-based spatiotemporal association pattern mining approach that encompasses both point data representation and the geographic dynamics of events using air quality data from Beijing–Tianjin–Hebei regions.

Giannotti et al. [44] proposed new methods to find trajectory patterns (T-Patterns) – location precedences with timing constraints that occur frequently among trajectory instances – such as,

$$\textit{railway station} \xrightarrow{15min} \textit{town square} \xrightarrow{2h15min} \textit{museum}.$$

To this end, the authors propose methods based on the identification of timestamped sequences of regions of interest using a density-based spatial discretization of trajectory data, which are then used as an input to the temporally-annotated sequence mining algorithm [45]. Inoue et al. [46] proposed an extension of a classic pattern mining algorithm – FP-Growth – to mine patterns of daily congested traffic based on traffic sensor data, and build a representation of congestion propagation processes in the road network. The study separates weekdays and days with/without rainfall to identify differences in congestion patterns based on those variables. In contrast with our proposal, the extended FP-growth algorithm requires patterns to satisfy spatial and temporal contiguity. Chen et al. [47] proposed an approach to discover patterns in congested traffic from taxi trajectory data by identifying congested links at each time. Although resembling the idea proposed by Inoue et al. [46], the authors start by finding Space-Temporal Congestion Subgraphs (STCS) – corresponding to congested roads – using a moving sliding window, and then apply FP-Growth to mine frequent STCS. Yang et al. [21] study human mobility patterns by finding hotspots from trajectories of 3474 individuals collected from mobile internet data for 22 days in China. The authors also extend classic pattern mining searches – here Apriori – to find frequent hotspots, defined as ”the most significant locations along the human’s trajectories”.

Despite their relevance, the surveyed approaches are hampered by the discretization needs of classic pattern mining algorithms, and unable to handle event data or provide integrative views from heterogeneous sources of road traffic data.

3.4 Context-sensitive analysis of urban traffic

Most of the research conducted to understand and model human mobility in urban areas relies on observations of habitual behavior and historical data. Situational context such as public events, road restrictions and meteorology largely influences mobility. Yet, that influence is often disregarded when analyzing patterns and behaviors in urban traffic data. This section will provide an overview of works on human mobility that integrate situational context in their analysis.

Situational context includes factors such as public events (e.g. festivals and sport games), that are planned and known beforehand, as well as factors that aren't planned (e.g. weather conditions and incidents). The former factors are often mentioned as Planned Special Events (PSE), term that was introduced by Latoski et al. in [48].

Kwoczek et al. [49] proposed a method for predicting and visualizing traffic congestions caused by planned special events. The method is based on an observation of typical behavior in traffic due to PSEs, which normally have two subsequent waves of congestion: people arriving an event and people leaving it. Estimating the first wave of congestion is hard, because there are many variables that determine the popularity of the event. The authors recognize this difficulty and develop a clustering based solution to retrieve similar PSEs, and use that information to predict and visualize the second wave of traffic, based on the information of the first wave and on the category of the event (e.g. concert, sports game).

Rodrigues et al. [50] introduced a Bayesian additive model (BAM) for decomposing traffic time series into structural components – including routine behavior versus individual special events – in order to estimate the number of arrivals in a given area. The incorporation of public event information improved predictions. The proposed method has the additional advantage of disclosing each individual event's influence, making the model highly interpretable.

3.5 Emerging pattern discovery

The concept of emerging patterns (EP) can be traced back to two different research streams. In time series data analysis, the discovery of emerging behaviors generally corresponds to the modeling of non-linear trends within a time series [51]. In this field, emerging behaviors are generally approximated using non-linear regressive or auto-regressive models, including regime switching models and neural network models, approximated on the original time series or on a decomposed series after removing seasonal and cyclical components [52].

In the pattern mining field, emerging behaviors were in 1999 coupled with the pattern concept, implying the satisfaction of well-defined frequency criteria. An emerging pattern (EP), as firstly introduced by Dong et al. [53], is a set of data instances whose characteristics entail significant changes between two (or more) timestamped datasets. Since then, this original notion of EPs has been extended and mostly applied in bioinformatic domains [53–57]. Nevertheless, and to our knowledge, EPs have not yet been extended towards spatiotemporal data structures neither applied in the context of road traffic monitoring networks.

Dong et al. [53] observed that, given the nature of the early-formulated EPs (larger in size and small in support), naive algorithms are costly and the Apriori property does not hold for EPs, proposing dedicated EP searches, and further extending these searches to build a classifier. Similarly, a substantial number of following works proposed the use EPs in classification tasks [54, 54, 56, 58]. Liu et al. [54] monitored gene expression under varying conditions, aiming at detecting trends under these conditions across genes for cells of the same type in order to predict the class of cells from the underlying EPs. Li et al. [55] extended these contributions using more efficient search variants. Fan et al. [59] propose a hybrid version of previous EP classifiers and Naive Bayes, yielding interpretation facilities.

Novak et al. [57] presented a survey on supervised descriptive rule discovery, a framework combining contrast set mining (CSM), emerging pattern mining (EPM), and sub-group discovery (SD). They explain that ”while all these research areas aim at discovering patterns in the form of rules induced from labeled data, they use different terminology and task definitions, claim to have different goals, claim to use different rule learning heuristics, and use different means for selecting subsets of induced patterns”.

In terms of efficiency, Fan et al. [60] in an effort addressed the issue of a large number EPs being generated by EP mining approaches by proposing an algorithm which considers only interesting EPs. This interestingness score is based on: support, growth rate, and a relationship between EPs and statistical measures. Soulet et al. [61] further proposed condensed representations of EPs based on the classic concept of frequent closed pattern.

More recent contributions extend EP discovery towards large-scale data [62, 63] by combining evolutionary fuzzy systems with the MapReduce paradigm; as well as towards streaming data [64] by combining evolutionary algorithms with batch strategies.

Song et al. [65] developed a methodology to detect and assess emerging, unpexpected and added/perished changes in customer behavior taking into consideration customer profiles and sales data along time. In the same domain, Chen et al. [66] propose association rule discovery along different time periods. They extend the early Song et al. [65] concepts towards emerging, unexpected and added rules and propose corresponding evaluation measures of growth, difference, and modified difference. Li et al [67] considered data from online reviews to identify EPs of hotel features in order to give hotel managers’ insights about travellers’ interests and expectations. For additional contributions and applications on EP discovery, the reader is invited to consult the work of Garcia-Vico et al. [68]. Again, and despite the relevance of the surveyed contributions on EP discovery, its applicability towards spatiotemporal data structures and mobile sensor domains remains unexplored.

Part II

Frequent Pattern Mining

The Lisbon city Council (CML) is currently able to gather different views on road traffic data along the city. Despite the relevance of this data to understand road traffic dynamics, the comprehensive discovery of traffic patterns is hampered by numerous challenges highlighted in our introductory chapter.

To address these challenges, this part of the thesis details our proposal on the use of biclustering – the discovery of subspaces within real-valued data – to comprehensively find congestion patterns from heterogeneous sources of road traffic data. A traffic pattern is here defined as a recurring congestion profile, possibly spanning diverse locations and time periods within a day.

To this end, we first provide a discussion on what are actionable road traffic patterns. Second, we propose a structured view on why, when and how to use biclustering for their effective and efficient discovery. Finally, we show how each of the identified challenges can be addressed using integrative data mappings and state-of-the-art principles on pattern-based biclustering. Although biclustering has been largely used in the biomedical field [13, 16], its potential in the mobility domain remains untapped.

We focus our study on the discovery of jam patterns from two major sources of road traffic: 1) geolocalized speed data (WAZE data), and 2) loop detectors' data. WAZE data contains information relative to congestion events, where a congestion event is a road segment that, at some point in time, has an average traffic speed significantly lower than the regular flow speed for that segment. Loop detectors are commonly placed in city junctions to measure the number, speed and type of vehicle passages over time. Both data sources offer relevant complementary views to find patterns in road traffic, including speed limits, jam size, congestion duration, severity degree, and vehicle throughput. Considering the Lisbon city as the study case, the gathered results confirm the relevance of biclustering to unravel non-trivial, meaningful, actionable and statistically significant patterns able to combine heterogeneous road traffic aspects.

We finalize by detailing the developed context-incorporation mechanism, which allows us to use biclustering to discover patterns sensitive to situational context. As a case study, we use a dataset of meteorological conditions in the city of Lisbon to gather jam profiles that are recurrent under specific weather conditions. We also provide a small discussion on the several potentialities of integrating a context view.

4

Solution

Contents

4.1 Road traffic patterns using biclustering 29

4.2 On why, how and when to apply biclustering 34

As introduced, our work aims at discovering actionable patterns of road mobility from two heterogeneous sources of traffic data: georeferenced time series data from ILDs and multivariate event collections from GPS sensors. Given the spatiotemporal nature of road traffic data, as well as the desirable properties of the pursued patterns (a complete list is provided in Section 2.5), this is a challenging task. To solve this task, we propose a two-step methodology. First, transformation procedures are applied to consolidate the original data sources and map them into new data structures appropriate to the subsequent mining task. Second, the use of pattern-based biclustering to discover traffic patterns from the transformed data sources.

Accordingly, Section 4.1 describes the proposed data transformations and principles for biclustering traffic data. In addition, Section 4.2 provides a structured view on why, when and how-to biclustering road traffic data.

4.1 Road traffic patterns using biclustering

4.1.1 Data mappings

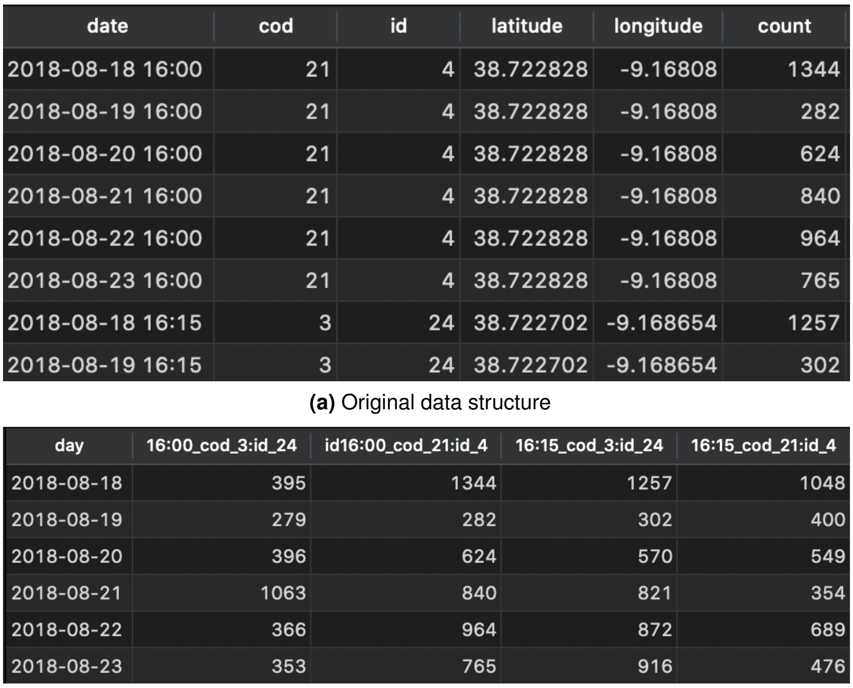
The first step of the discovery process is to fix spatial, temporal and calendric constraints, including the target geographies, date intervals, and weekday annotations. In addition, the time granularity (e.g. 15-minute, hour or on/off-peak intervals) can be optionally specified to guide road traffic data aggregation. In its absence, the proposed pattern discovery is iteratively performed using multiple time aggregations.

Once these constraints are fixed, data mappings are applied to transform the original spatiotemporal data structures into tabular data structures, more conducive to the subsequent pattern mining task. In the target structure, each observation/row represents a day and each variable/column measures some specific road traffic aspect on a specific location and time period of a day.

For the ILD data, each variable measures the number of cars passing over a single loop detector in a specific time interval of the day. Figure 4.1 shows the original structure of the ILD data and the corresponding data mapping.

For the geolocalized speed data (WAZE data), multiple measurements are taken per event, and events can occur on different roads. Here the columns correspond to a measurement on a single road for a specific time interval of the day. Figure 4.2 shows the original structure of WAZE data and the corresponding transformed data.

The integration of the previous mappings is a simple concatenation of the variables resulting from the transformation of each road traffic data source.



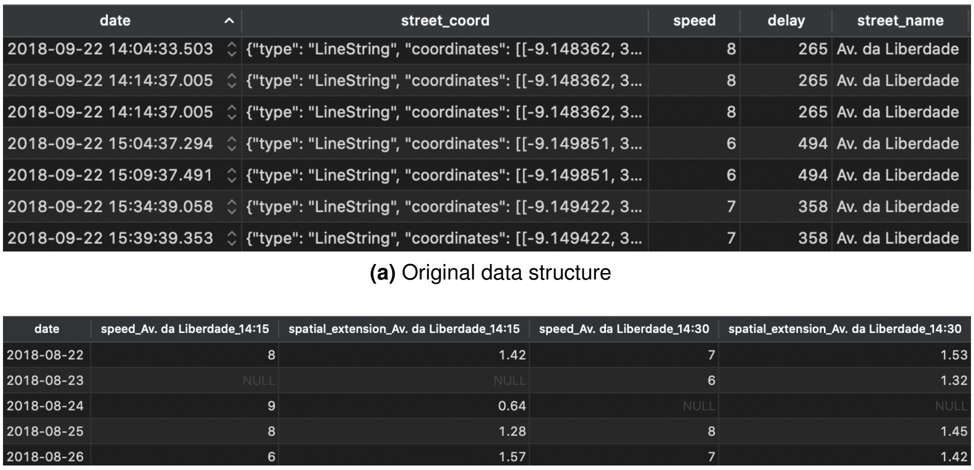
date	cod	id	latitude	longitude	count
2018-08-18 16:00	21	4	38.722828	-9.16808	1344
2018-08-19 16:00	21	4	38.722828	-9.16808	282
2018-08-20 16:00	21	4	38.722828	-9.16808	624
2018-08-21 16:00	21	4	38.722828	-9.16808	840
2018-08-22 16:00	21	4	38.722828	-9.16808	964
2018-08-23 16:00	21	4	38.722828	-9.16808	765
2018-08-18 16:15	3	24	38.722702	-9.168654	1257
2018-08-19 16:15	3	24	38.722702	-9.168654	302

(a) Original data structure

day	16:00_cod_3:id_24	id16:00_cod_21:id_4	16:15_cod_3:id_24	16:15_cod_21:id_4
2018-08-18	395	1344	1257	1048
2018-08-19	279	282	302	400
2018-08-20	396	624	570	549
2018-08-21	1063	840	821	354
2018-08-22	366	964	872	689
2018-08-23	353	765	916	476

(b) Transformed data structure

Figure 4.1: ILD data mapping.



date	street_coord	speed	delay	street_name
2018-09-22 14:04:33.503	{"type": "LineString", "coordinates": [[-9.148362, 3...	8	265	Av. da Liberdade
2018-09-22 14:14:37.005	{"type": "LineString", "coordinates": [[-9.148362, 3...	8	265	Av. da Liberdade
2018-09-22 14:14:37.005	{"type": "LineString", "coordinates": [[-9.148362, 3...	8	265	Av. da Liberdade
2018-09-22 15:04:37.294	{"type": "LineString", "coordinates": [[-9.149851, 3...	6	494	Av. da Liberdade
2018-09-22 15:09:37.491	{"type": "LineString", "coordinates": [[-9.149851, 3...	6	494	Av. da Liberdade
2018-09-22 15:34:39.058	{"type": "LineString", "coordinates": [[-9.149422, 3...	7	358	Av. da Liberdade
2018-09-22 15:39:39.353	{"type": "LineString", "coordinates": [[-9.149422, 3...	7	358	Av. da Liberdade

(a) Original data structure

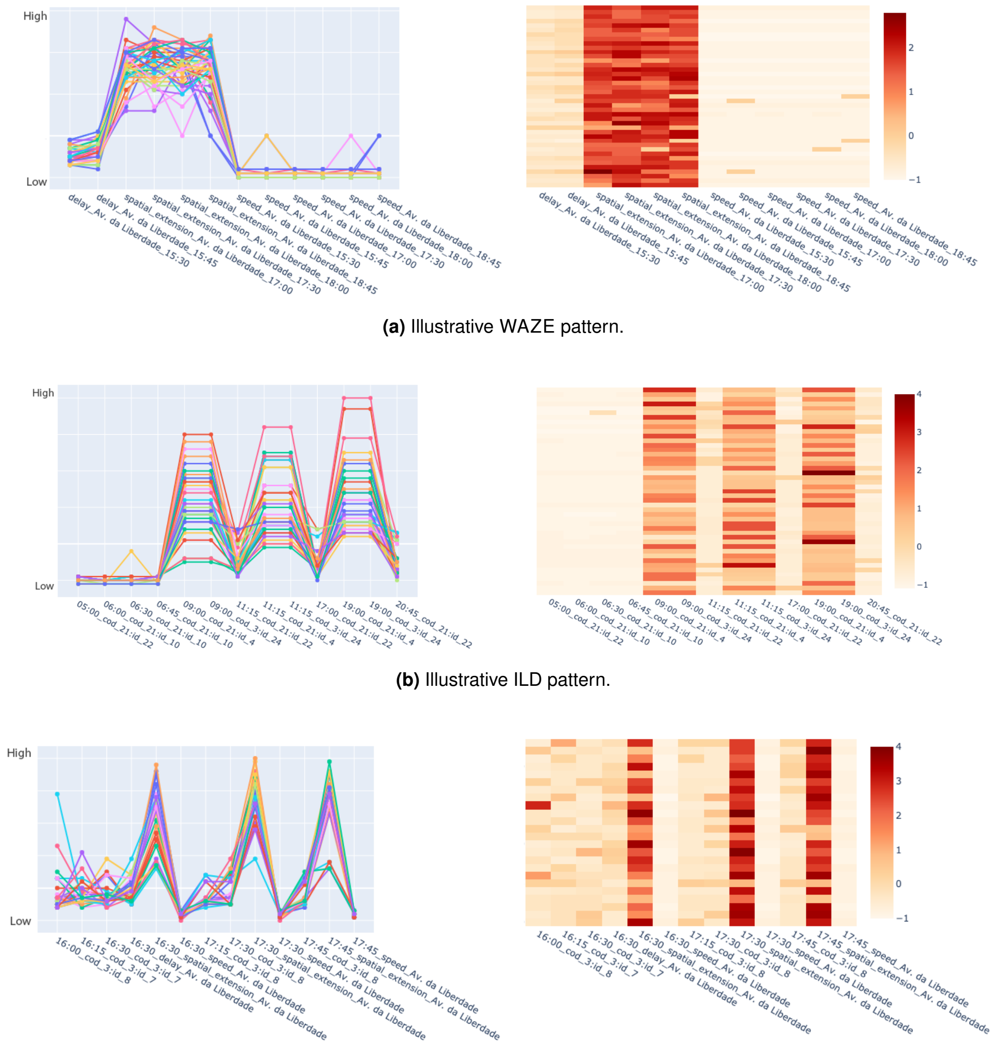
date	speed_Av. da Liberdade_14:15	spatial_extension_Av. da Liberdade_14:15	speed_Av. da Liberdade_14:30	spatial_extension_Av. da Liberdade_14:30
2018-08-22	8	1.42	7	1.53
2018-08-23	NULL	NULL	6	1.32
2018-08-24	9	0.64	NULL	NULL
2018-08-25	8	1.28	8	1.45
2018-08-26	6	1.57	7	1.42

(b) Transformed data structure

Figure 4.2: WAZE data mapping.

4.1.2 Biclustering

Under the previous mappings, traffic data still preserves their spatiotemporal content, yet denormalized within a tabular data structure, turning it a candidate for the application of biclustering. In fact, the specific properties of the introduced transformations were specifically proposed to this end. As a result, a *traffic pattern* is elegantly seen as a recurrent and coherent congestion profile (w.r.t. speed, volume, extent)



(a) Illustrative WAZE pattern.

(b) Illustrative ILD pattern.

(c) Illustrative integrative pattern.

Figure 4.3: Illustrative road traffic patterns given by biclusters separately and integratively found in ILD and WAZE data (collected at Marquês do Pombal junction within the Lisbon city).

that can span diverse locations and different time periods.

Pattern-based biclustering approaches provide the unprecedented possibility to comprehensively find patterns in real-valued data with parameterizable homogeneity and guarantees of statistical significance.

Biclustering aims at finding subsets of observations with values correlated on a subset of variables. In the context of our work, this means that the pattern of the bicluster corresponds to the jam profile, the pattern support (i.e. number of observations) corresponds to the number of days with the given jam profile (i.e. pattern recurrence), and the pattern length (i.e. number of variables) corresponds to the number of locations and time periods within a day associated with the given jam profile. Figure 4.3 provides an illustration of spatiotemporal traffic patterns given by the target biclusters using BicPAMS

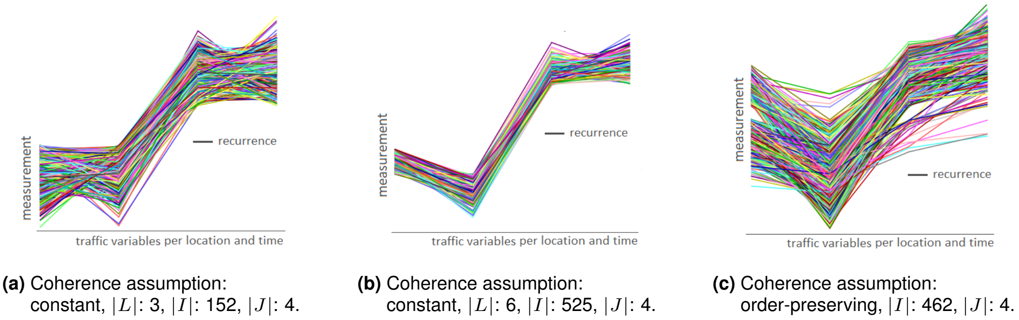
[16]. The instantiated road traffic patterns were obtained through the application of biclustering over ILD and WAZE data collected at the heart of the Lisbon city (Marquês de Pombal), Portugal.

To discover different jam profiles using biclustering, the *coherence strength* and *coherence assumption* of the target biclustering solutions can be customized in accordance with the desirable profiles of congestion.

Coherence strength. Biclustering also allows the calibration of *coherence strength* (Section 2.3) – e.g. how much speed limits (or car flow) need to differ to be considered dissimilar.

Patterns are inferred from similar (yet non-strictly identical) congestion properties, whether they are: 1) numerical (speed limits, spatial extent), 2) integer (number of vehicles), or 3) ordinal (congestion severity).

Figure 4.4a-b illustrates the impact that different coherence strength criteria can have on the found patterns. Considering $\delta=\frac{\bar{A}}{|L|}$ (Section 2.3), a looser coherence strength of $|L|$=3 allows the discovered traffic patterns to be sensitive to 3 profiles (e.g. low, medium and high volume car passage), while higher coherence strengths (such as $|L|$=7) indicates a greater sensitivity to traffic variability.



(a) Coherence assumption: constant, $|L|$: 3, $|I|$: 152, $|J|$: 4.

(b) Coherence assumption: constant, $|L|$: 6, $|I|$: 525, $|J|$: 4.

(c) Coherence assumption: order-preserving, $|I|$: 462, $|J|$: 4.

Figure 4.4: Effects of coherence strength and assumption on the resulting traffic patterns.

Allowing these strength-based deviations from pattern expectations in real-valued mobility data is key to prevent the item-boundaries problem associated with the discretization problems faced by classic pattern mining methods.

Constant mobility patterns. Depending on the goal, one or more *coherence assumptions* (Section 2.3) can be pursued. The classic binary coherence assumption is focused on patterns of congestion independently of the level of congestion. Such coherence assumption has severe problems because it is highly dependent on the criteria that determines what is a jam or not. This can be hard to identify given the heterogeneity of speed limits in accordance with road types. In addition, such option is unable to distinguish different levels of congestion, a necessary condition if we want to assess our traffic patterns and guarantee that they are actionable. The binary assumption can thus be replaced by a constant assumption. Figure 4.3 provides illustrative constant patterns of road traffic.

Non-constant mobility patterns. The constant assumption suffers from a problem: two days need to satisfy the same jam profile in order to count as supporting observations for a bicluster. However, congestion highly varies along days. Even when focusing on specific days (e.g. Tuesdays, Wednesdays and Thursdays; Fridays; holidays), there is a high traffic variability dependent on the presence of public events, weather context, or road traffic interdictions.

In this context, non-constant patterns should be pursued to guarantee a greater robustness to traffic variability, while still guaranteeing the coherence of the target traffic patterns. In particular, two types of traffic patterns are pursued:

- *additive* pattern: days with variations on the expected jam profile (along specific locations and time periods of the day), coherently explained by shifting factors;
- *order-preserving* pattern: days with preserved orderings of jam intensity over a set of locations and time periods (Figure 4.4c). Illustrating, if a specific location is always more congested than another with regards to speed limits, the same order is observed irrespectively of the absolute value associated with the speed limit. Illustrating, consider the measuring of jam extents (kilometers) between 9h-9h15 in three locations (corresponding to variables y_2, y_3 and y_7), days x_1 and x_2 $(y_2, y_3, y_7|x_1)$={0.32,0.50,0.47} and $p(y_2, y_3, y_7|d_2)$={0.29,0.97,0.55} are coherently associated since they preserve the permutation $a_{i2} \leq a_{i3} \leq a_{i7}$.

As a result, pattern-based biclustering allows the discovery of less-trivial yet coherent, meaningful and potentially relevant spatiotemporal associations that form the target traffic patterns.

Handling highly sparse traffic data. Road traffic data is inherently sparse, specially georeferenced speed data. After the proposed data mappings, an arbitrarily-high fraction of elements from the transformed data is empty due to the localized occurrence of jams in specific locations and time periods. This creates a new requirement for the target approach: ability to discover patterns in the presence of highly sparse data.

Since the proposal of BicNET [69], pattern-based biclustering approaches were enriched with principles to efficiently explore sparse data. In fact, pattern-based biclustering approaches further enable the discovery of biclusters with an upper bound on the allowed amount of missings. This is particularly relevant to guarantee that the sporadic absence of a jam on a specific time period does not impact the target road traffic patterns as can be shown in Figures 4.3 and 4.4.

4.2 On why, how and when to apply biclustering

On *WHY*. As motivated, biclustering of traffic data should be considered to:

- avoid the drawbacks of classic pattern mining methods, including: 1) their susceptibility to the item-boundaries problems[1], and 2) inability to comprehensively explore the spatiotemporal content of traffic data;
- discover non-trivial patterns of congestion given by constant, additive and order-preserving jam profiles;
- combine heterogeneous aspects of road traffic, including limited speed, vehicle volume, and spatial extent of jams;
- pursue patterns with parameterizable properties of interest by customizing the target coherence strength, quality (noise-tolerance), dissimilarity and statistical significance criteria.

On *WHEN*. Similarly, biclustering of traffic data should be applied when: 1) jam intensity/profile matters; 2) pursuing less-trivial forms of knowledge (including the introduced constant or order-preserving assumptions); 3) discretization drawbacks must be avoided; 4) heterogeneous sources of road traffic are available; and when 5) one seeks to find comprehensive solutions of traffic patterns with customizable homogeneity.

On *HOW*: comprehensive exploration of traffic data. Pattern-based biclustering offers principles to find complete solutions of traffic patterns by: 1) pursuing multiple homogeneity criteria, including multiple coherence strength thresholds, coherence assumptions and quality thresholds; and 2) exhaustively yet efficiently exploring different regions of the search space, preventing that regions with large patterns jeopardize the search [16]. As a result, non-trivial yet significant correlations within road traffic data are not neglected.

In addition, pattern-based biclustering does not require the input of support thresholds as it explores the search space at different supports [19], i.e. there is no need to place expectations on the minimum

[1] The possibility to allow deviations from value expectations (under limits defined by the placed coherence strength) together with multi-item assignments [19] are placed to prevent discretization problems from occurring

number of days for a jam profile to become relevant. The minimum number of locations and time periods within a day can be optionally inputted to guide the search. Dissimilarity criteria and condensed representations can be also placed [16] to prevent the delivery of redundant patterns.

On *HOW*: statistical significance. A sound statistical testing of road traffic patterns is key to guarantee the absence of spurious relations, and ensure the relevance of the given patterns to support mobility decisions. To this end, the statistical tests proposed in BSig [15] are suggested to minimize false positives (outputted patterns yet not statistically significant) without incurring on false negatives. This is done by approximating a null model of the target traffic data and statistically testing each bicluster against the null model in accordance with its underlying coherence.

On *HOW*: robustness to noise. Pattern-based biclustering can find biclusters with a parameterizable tolerance to noise [19]. Illustrating, a quality of 80% indicates that an upper limit given by 20% of entries within a bicluster may deviate from the target jam profile ($\mu_{ij} \notin [-\delta/2, \delta/2]$). This possibility ensures robustness to the inherent daily traffic fluctuations, as well as spontaneous jams caused by sporadic events which do not yield particular significance.

On *HOW*: other opportunities. Additional benefits of pattern-based biclustering that can be carried towards the analysis of traffic data include:

1. the possibility to remove uninformative elements in data to guarantee a focus, for instance, on non-trivial jam profiles (removal of entries denoting highly congested traffic) [69];
2. incorporation of domain knowledge to guide the task in the presence of background metadata [70];
3. support classification and regression task in the presence of labels (e.g. traffic conditioning modes, panel message recommendations, situational context) by guaranteeing the discriminative power of biclusters [14].

5

Results

Contents

5.1 Exploratory data analysis 39

5.2 Road traffic patterns 44

5.3 Statistical significance 49

In this chapter we start by showing some preliminary results on Section 5.1, that were gathered prior to the application of biclustering to get a better grasp on the data. Then, on Section 5.2 we apply the proposed approach to discover road traffic patterns using biclustering. Finally, on Section 5.3 we show that biclustering guarantees the statistical significance of the spatiotemporal associations found within road traffic data, providing a trustworthy means to support mobility reforms.

5.1 Exploratory data analysis

Exploratory data analysis (EDA) is the process of applying different analysis techniques to a dataset, to achieve a better understanding of its characteristics. The techniques applied are mostly done with the resource of visualizations, which reveal structures that help analysts maximize their insight of the dataset. We first conduct an EDA process on the loop detectors data and then on the WAZE's geolocalized speed data.

5.1.1 Loop detectors

The dataset provided by CML contains measurements of 76 loop detectors scattered around Lisbon's road network, which count the number of vehicles that cross them with a temporal resolution of 15 minutes. Compared to other works that use data collected from sensors similar to these, the temporal resolution of the ones under analysis is coarser, which makes deriving other variables such as traffic speed harder. The analysis is conducted on a set of 4 loop detectors that are placed on roads near Instituto Superior Técnico, Lisbon, as can be seen in Fig. 5.1.

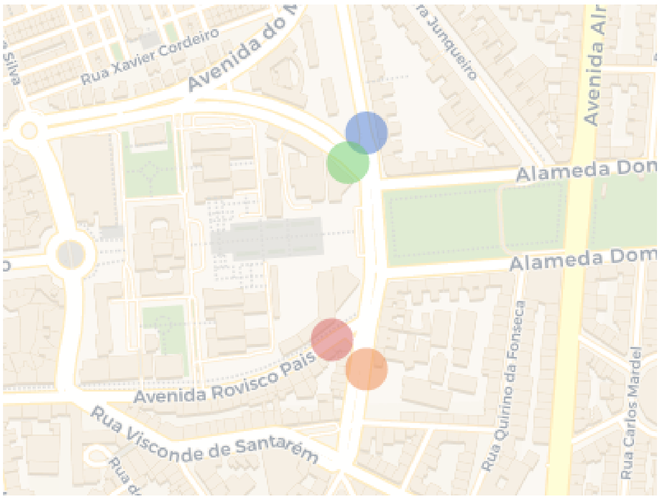


Figure 5.1: Location of loop detectors under analysis. Loop detectors will be mentioned by their street name: Av. Manuel da Maia(1) (top right), Av. António José de Almeida (top left), Av. Rovisco Pais (bottom left) and Av. Manuel da Maia(2) (bottom right).

The line chart on Fig. 5.2 shows the central tendency and standard deviation of the number of vehicle passages on the loop detectors under analysis, over the whole period of time available in the dataset. Cumulative counts over periods of one hour were considered, to show smoother tendencies. The figure depicts interesting seasonal patterns and characteristics, as well as evidences of problems in the data.

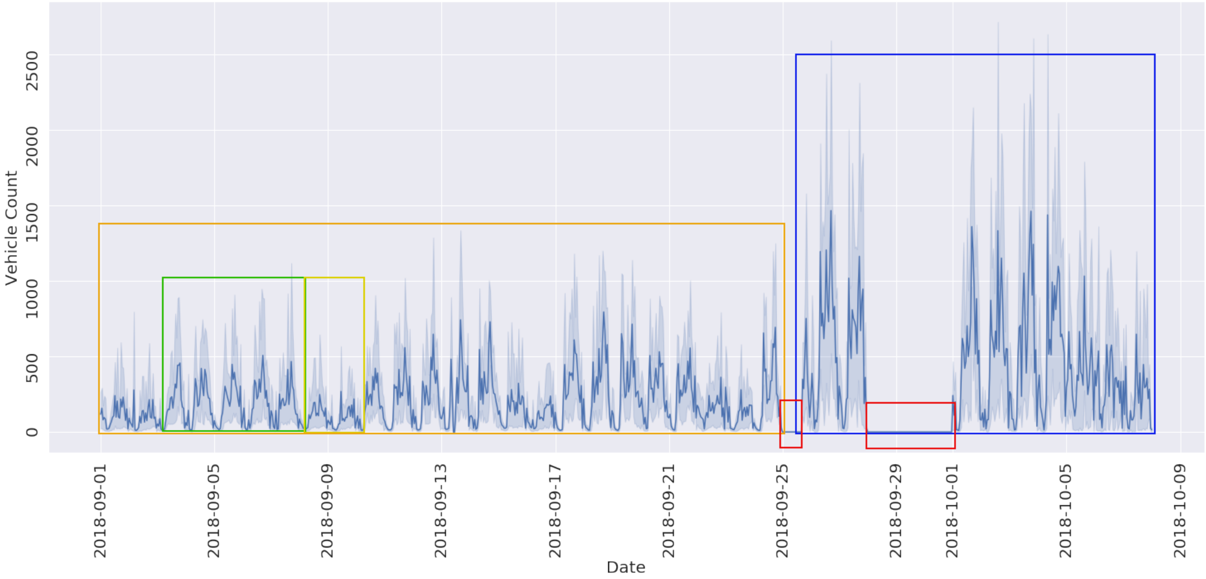


Figure 5.2: Time Series plot of the four loop detectors under analysis.

In particular:

- There is a high variability between the volume of cars passing the different loop detectors. This can be simply due to the road where each loop detector is located having different usage demands, or in a worse case due to wrong calibrations of some loop counters causing the registry of abnormally high/low values of car passages.
- The number of vehicle passages over the period of a day follows a clear pattern, where in early hours (0-6 a.m.) there are almost no cars passing over the induction loops.
- There is a clear distinction in the volume of traffic on weekdays, which is significantly higher than on weekends. The green box outlines an example of a workweek and the yellow box an example of a weekend.
- There are some gaps in the dataset with no data, which are depicted by the red boxes in the figure.
- Between the orange box and the blue box, the volume of vehicles is remarkably different. In particular, in the blue box the number of vehicle passages registered by the loop detectors is much higher. A hypothesis for this abrupt growth is the start of the academic year.

Box plots offer a way to visualize distribution of data based on statistical summary values: the lower and upper quartiles (respectively, the 25th and 75th percentile), the median and the extremes of the data's distribution. The values that lie outside of the extremes are represented as dots and denote outliers.

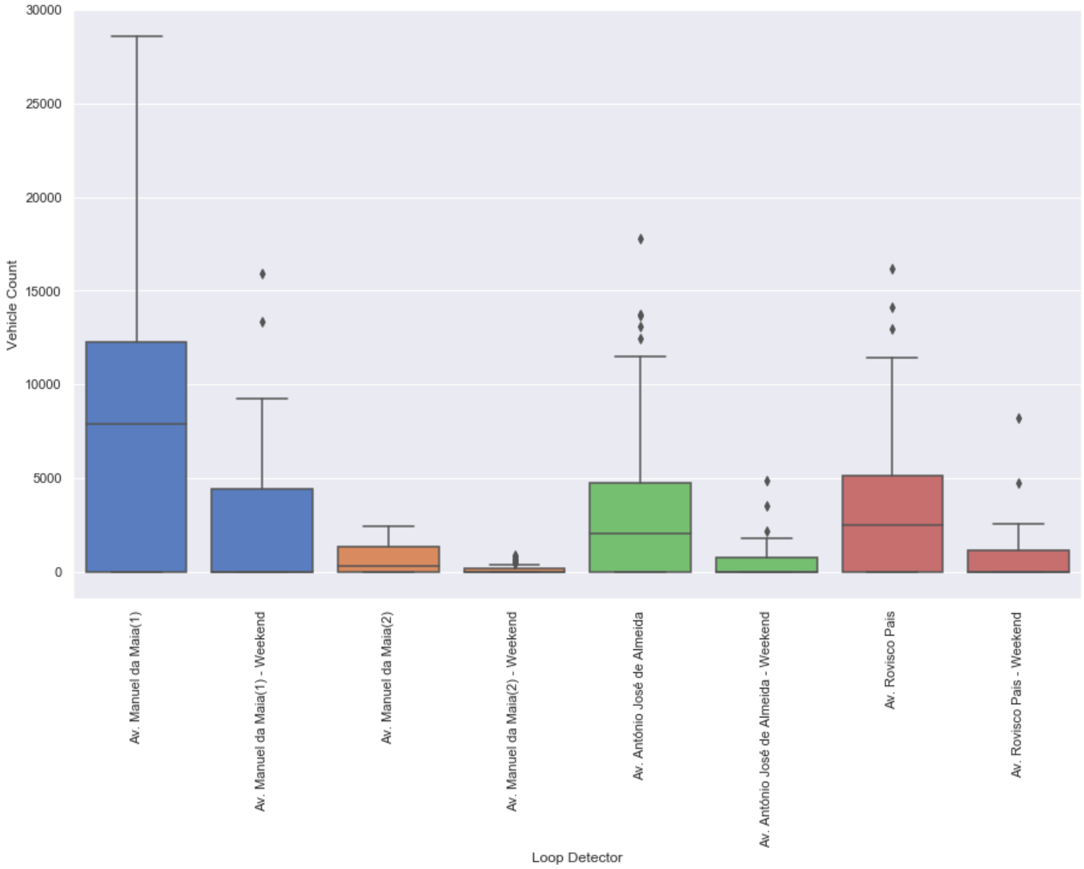


Figure 5.3: Box plots of the four induction loops for weekdays and weekends.

The box plot in Fig. 5.3 shows the distribution of cumulative vehicle passages on the four induction loops on day intervals, over the whole period of time of the dataset. Each induction loop has two box plots, corresponding to the distributions for the weekdays and the weekends, because of the difference in profiles for those periods. There are some properties derived from the figure that are worth mentioning:

- There is a clear difference in the profiles of each induction loop. In particular, the blue induction loop has a much higher volume of cars passing and a bigger variability in its measures than the others. As stated earlier for the time series plot, this can be due to errors in the sensor measures and the sensor being placed in a busier road. The variability can also be caused by the seasonality detected on the line chart presented before.
- The distribution for most of the loop detectors on weekdays is rather wide. This can be due to

the difference in volumes remarked on Fig. 5.2 by the orange and blue box. The loop detector *Av. Manuel da Maia(1)* in particular registers some skewness in it's distribution that resembles the proportion of data in the orange (less volume) and blue (more volume) boxes on Fig. 5.2.

- All the induction loops have significantly lower volume in the weekends, except for some outliers which are also present on weekdays. This could be due to the effect of situational context such as special planned events.

- The days that were missing are reflected on the distributions of the box plot, that reach zero for every induction loop.

5.1.2 WAZE's geolocalized speed data

The assessed data in this section contains results collected from the WAZE's API between 2018-09-03 and 2018-10-09 in the area of Lisbon. There are some missing days in the data probably due to downtimes in the API server or in the server running the script that collected data from it.

Fig. 5.4 shows a heatmap of the number of congested trajectories by day and hour. On a quick glance of the visualization, one can already notice some interesting characteristics of the data:

- The amount of missing days is relatively large. An example of a set of missing days is represented by the red box in the figure.

- From the beginning of September until the week of September 23rd, a vertical gradient is almost clear, meaning that the number of congested trajectories had a significant growth until that period.

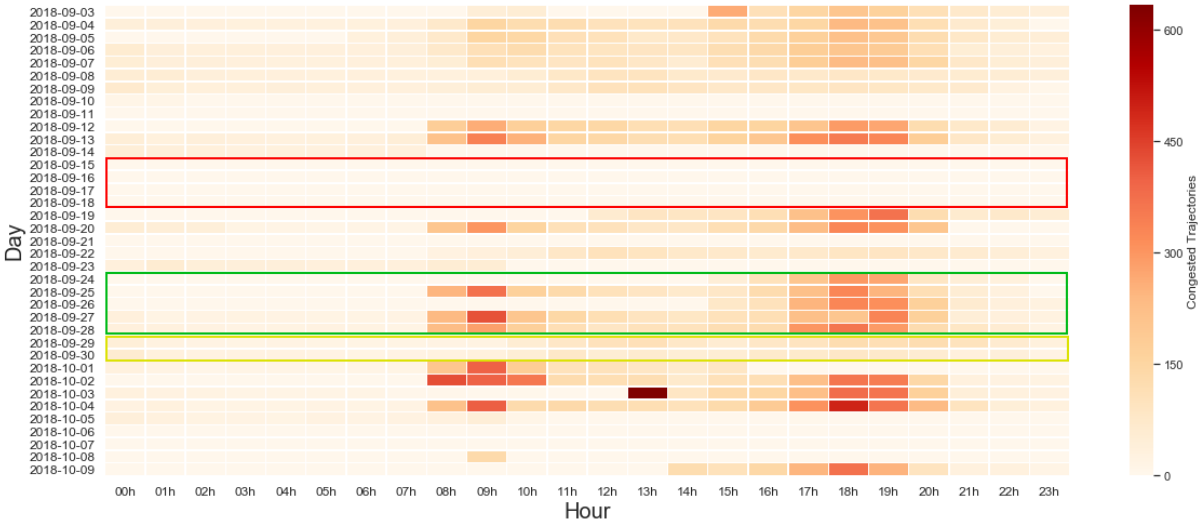


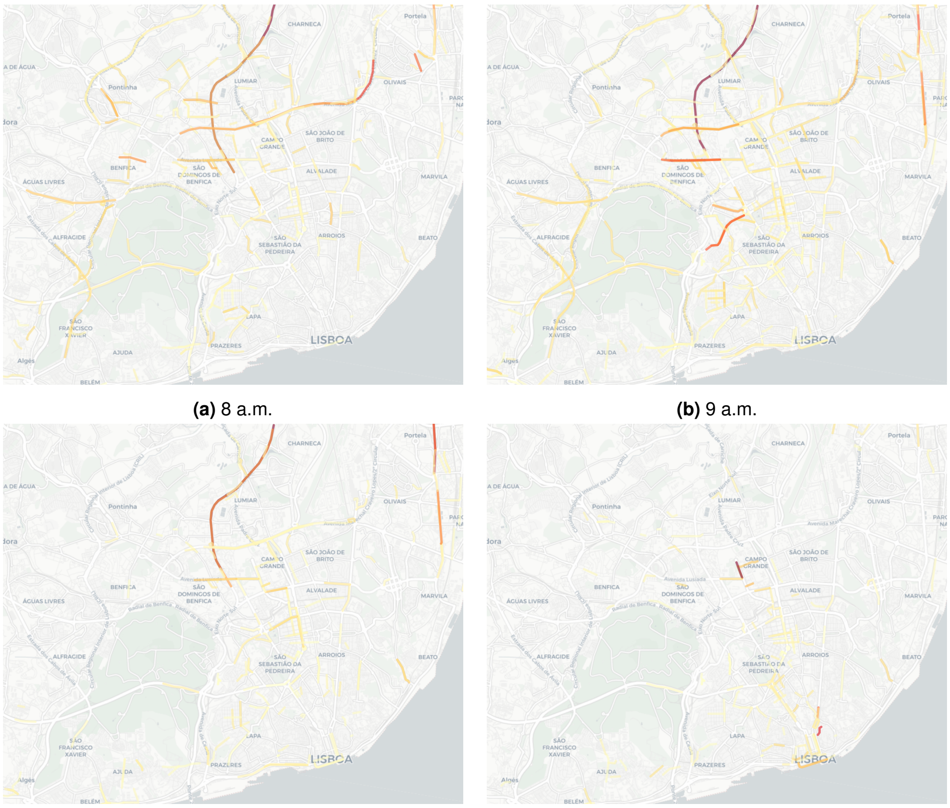
Figure 5.4: Heatmap for the number of congested trajectories by day and hour in the city of Lisbon.

We suspect that this might be an evidence of the effect of starting the school year and the ending of summer vacations for most people. This effect was also noticeable on the loop detectors data.

- The number of congested trajectories is much lower in the weekends than during weekdays. The green box highlights a weekday and the yellow box a weekend.

- During weekdays there are two peaks of congestions, on the morning between 8-10h and on the evening between 17-19h.

Fig. 5.5 shows a progression of the congestion state in Lisbon's road network, for the morning of September 27th, 2018. The color encodes the delay registered at a trajectory, redder means a bigger delay. It's noticeable the effect of commuting in the congestion progression. At 8 a.m. there is already a large amount of congested roads, and at 9 a.m. congested state reaches its peak. Coincidentally, most people start working near this time. At 10 a.m. congestions start to decrease, and at 11 a.m. most peripheral roads have no congestions while only some roads near the center of the city register delays in traffic flow.



(a) 8 a.m.

(b) 9 a.m.

(c) 10 a.m.

(d) 11 a.m.

Figure 5.5: Congestion snapshots on 27th September 2018, between 8-11 a.m.

5.2 Road traffic patterns

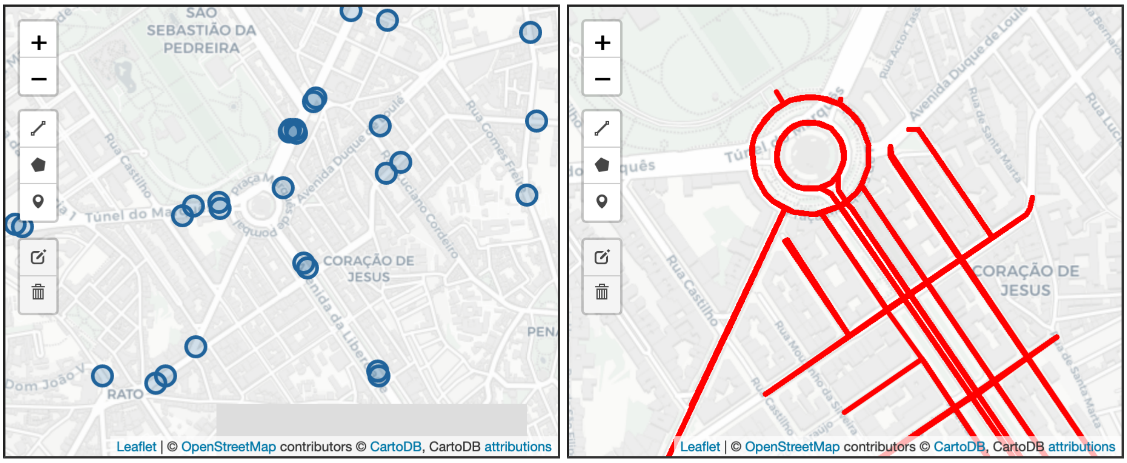
Considering the Lisbon city as a study case, we applied the proposed approach to comprehensively discover road traffic patterns from geolocalized speed data from WAZE and inductive loop detector (ILD) data collected during a two month period in central junctures of the city (Figure 5.6). To illustrate the enumerated potentialities, experiments are discussed in three major steps, corresponding to the analysis of the gathered results from ILD, WAZE, and consolidated ILD-WAZE data sources.

Experimental setting. BicPAMS [16] was the selected biclustering approach as it combines state-of-the-art principles on pattern-based biclustering. BicPAMS is used with default parameters: varying coherence strength ($\delta = \bar{A}/|\mathcal{L}|$ where $|\mathcal{L}| \in \{2,..,10\}$), decreasing support until 100 dissimilar biclusters are found, up to 30% noisy elements, 0.01 significance level, and constant and order-preserving coherence assumptions. Two search iterations were considered by masking the biclusters discovered after the first iteration to ensure a more comprehensive exploration of the data space and a focus on less-trivial patterns of road mobility.

Location-based distributions of speed, extent and frequency were approximated, and the statistical tests proposed in BSig [15] applied to compute each pattern's statistical significance.

5.2.1 ILD traffic patterns

Two months of observations produced from loop detectors placed at major junctures of the city were collected (Figure 5.6a). Table 5.1 synthesizes the results produced by biclustering ILD data with BicPAMS [16].


SÃO SEBASTIÃO DA PEDREIRA
CORAÇÃO DE JESUS
Túnel do Ma
Avenida Duque de Loulé
Rua Castilho
RATO
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions
CORAÇÃO DE JESUS
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions

(a) ILD locations

(b) WAZE events

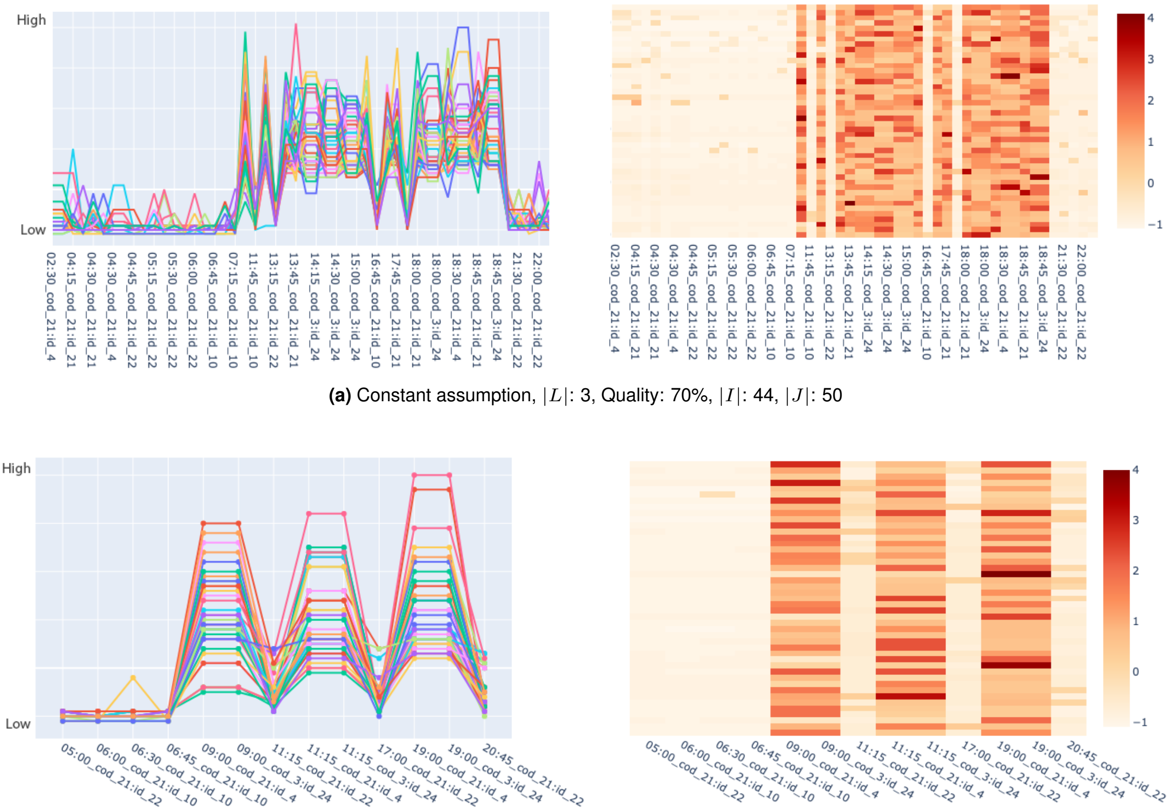
Figure 5.6: Map visualization of the two sources of urban traffic data along the studied area (Marquês de Pombal): a) ILD sensor placement; b) WAZE jam events on peak hour (1/14/2020, 9AM).

Confirming the potentialities listed in Chapter 4, BicPAMS was able to efficiently and comprehensively find homogeneous, dissimilar and statistically significant biclusters – recurrent variations on the flow of vehicles (throughput) spanning diverse locations and different time periods. Consider, for instance, traffic patterns given by constant biclusters sensitive to three degrees of volume ($|L|$=3) and 70% quality. These traffic patterns have an average of $\mu(|J|)$=20 features (corresponding to different city locations and time periods of a day) and occur on $\mu(|I|)$=43 days within a two month period (60 days). These initial results further show the impact of tolerating noise, placing different coherence assumptions (such as the order-preserving assumption) and parameterizing coherence strength ($\delta \propto \frac{1}{|\mathcal{L}|}$) on the biclustering solution.

Figure 5.7 visually depicts a constant and order-preserving patterns of road mobility using a line chart (where each line corresponds to a day when the traffic pattern was observed) and heatmap (where days

Query	Assumption	$\|L\|$	quality	#bics	$\mu(\|I\|)\ \pm\sigma(\|I\|)$	$\mu(\|J\|)\ \pm\sigma(\|J\|)$	p-value<1E-3
1	Constant	3	70%	71	43.4±1.7	19.7±16.1	71
2	Constant	4	70%	42	43.0±1.3	10.2±4.4	42
3	Constant	5	70%	10	43.0±0.7	10.6±3.2	10
4	Order-preserving	20	70%	1273	44.5±0.5	6.0±1.7	1273

Table 5.1: Properties of biclustering solutions in ILD data using BicPAMS with varying homogeneity criteria.



(a) Constant assumption, $|L|$: 3, Quality: 70%, $|I|$: 44, $|J|$: 50

(b) Order-preserving assumption, Quality: 70%, $|I|$: 45, $|J|$: 13

Figure 5.7: Illustrative constant and order-preserving traffic patterns found in ILD data.

correspond to rows). The traffic pattern captures coherent variations on the traffic flow across locations and time periods.

ILD data are in essence georeferenced multivariate time series data (Section 2.1). Understandably, biclustering can be as well applied over any alternative source of traffic data given by georeferenced time series, such as the average car speed.

5.2.2 WAZE traffic patterns

WAZE events associated with jam problems at the Marquês de Pombal area within Lisbon were collected for two months (Figure 5.6b). Table 5.2 synthesizes the biclustering results produced by the application of BicPAMS over WAZE data. Similarly to ILD, we observe an inherent ability of biclustering to efficiently retrieve a large number of robust, dissimilar and statistically significant patterns of road traffic. These patterns are reocurring speed limits and jam extent that span specific trajectories and time periods.

For this analysis we consider WAZE data in their whole richness, combining views on speed, jam extent, and perceived severity. Illustrating, traffic patterns given by constant biclusters with coherence strength determined by $|L|$=4 are sensitive to four levels of severity, speed and jam extension. We can, for instance, observe that biclusters with $|L|$=4 and 70% quality have a median of 6 features (corresponding to different city locations and time periods of a day) and occur on an average of $\mu(|I|)$=42 days within a two month period (60 days). These results further show the relevance of discovering patterns with different homogeneity criteria (coherence assumption, coherence strength and quality).

Figure 5.8 depicts three constant road traffic patterns (and the respective jam profile, spanned locations, time periods of the day) using BicPAMS with default parameters.

Each bicluster shows a unique traffic pattern. For instance, the first traffic pattern (Figure 5.8a) captures a congestion profile at the evening peak hour with locations where jam extensions are high and locations where speed is severely limited. These results motivate the relevance of finding constant biclusters to find patterns with coherent speed limits and congestion lengths for a statistically significant number of days.

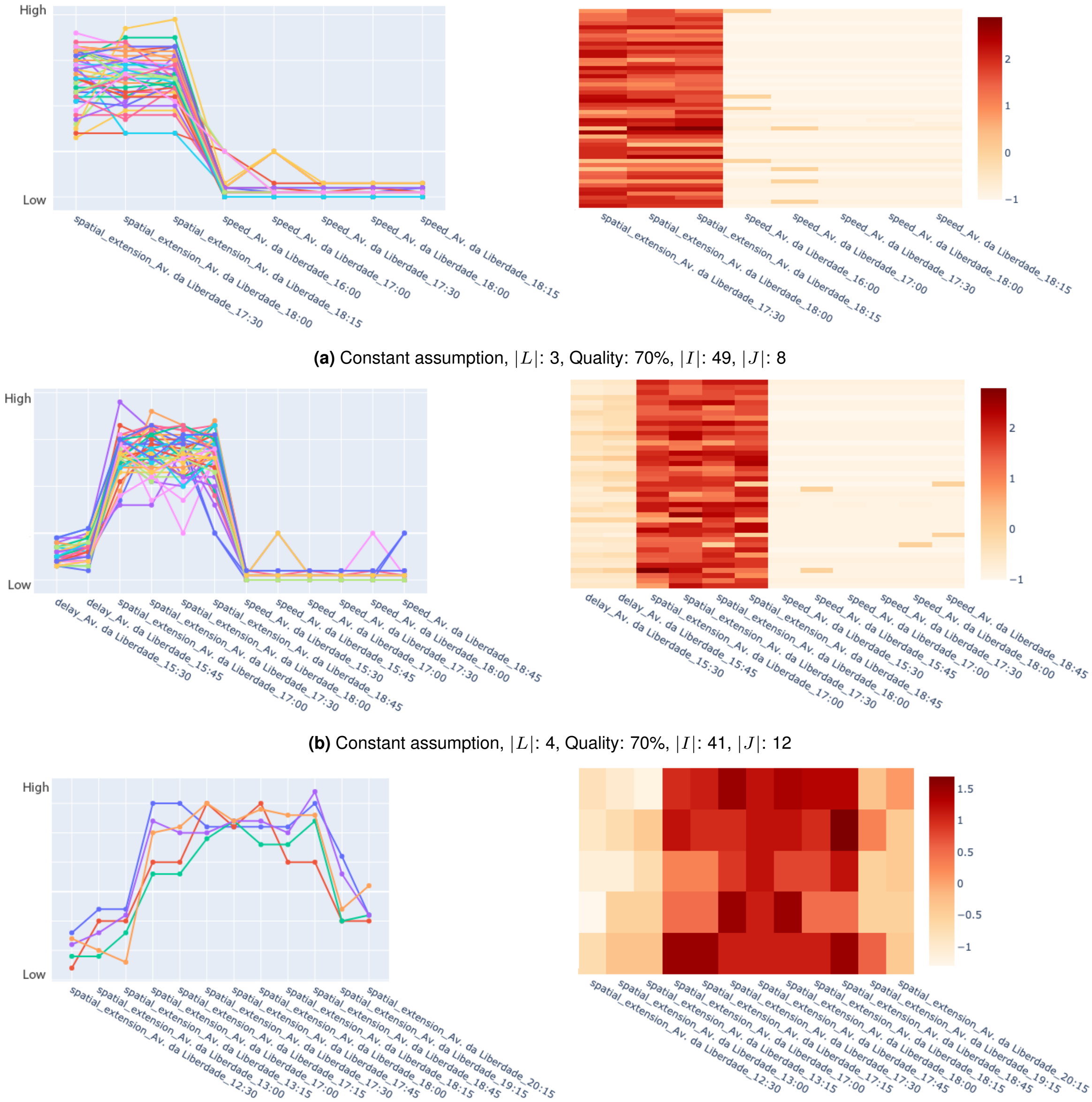
A closer analysis of the found road traffic patterns shows their robustness to the item-boundaries

Query	Assumption	$\|L\|$	quality	#bics	$\mu(\|I\|)$	$\mu(\|J\|)$	$p-$value
					$\pm\sigma(\|I\|)$	$\pm\sigma(\|J\|)$	<1E-3
1	Constant	3	70%	47	44.7±3.6	5.5±1.9	47
2	Constant	4	70%	79	42.1±3.6	5.7±2.0	79
3	Constant (only spatial extension)	3	100%	142	12.6±2.9	4.2±0.5	142
4	Order-preserving	20	70%	153	46.9±3.8	5.7±1.5	153
5	Order-preserving (only spatial extension)	20	70%	135	8.1±2.1	4.1±0.3	135

Table 5.2: Properties of biclustering solutions in WAZE data using BicPAMS with varying homogeneity criteria.

problem: slight deviations from the expect speed limit or jam extension are not excluded from the bicluster. The target patterns are thus not hampered by the drawbacks of discrete views on road traffic.

Non-constant patterns are in this work suggested to find more flexible patterns of road traffic, usually associated with less-trivial traffic associations. Figure 5.9 depicts two non-constant traffic patterns with an order-preserving assumption. This assumption is useful to capture coherent orders in jam profiles, thus being able to account for coherent differences in speed limits, jam extensions and expected delays across days. As one can clearly see on the heatmaps (Figure 5.9a and b), order-preserving patterns are characterized by a well-established permutation on the features associated with a congestion.

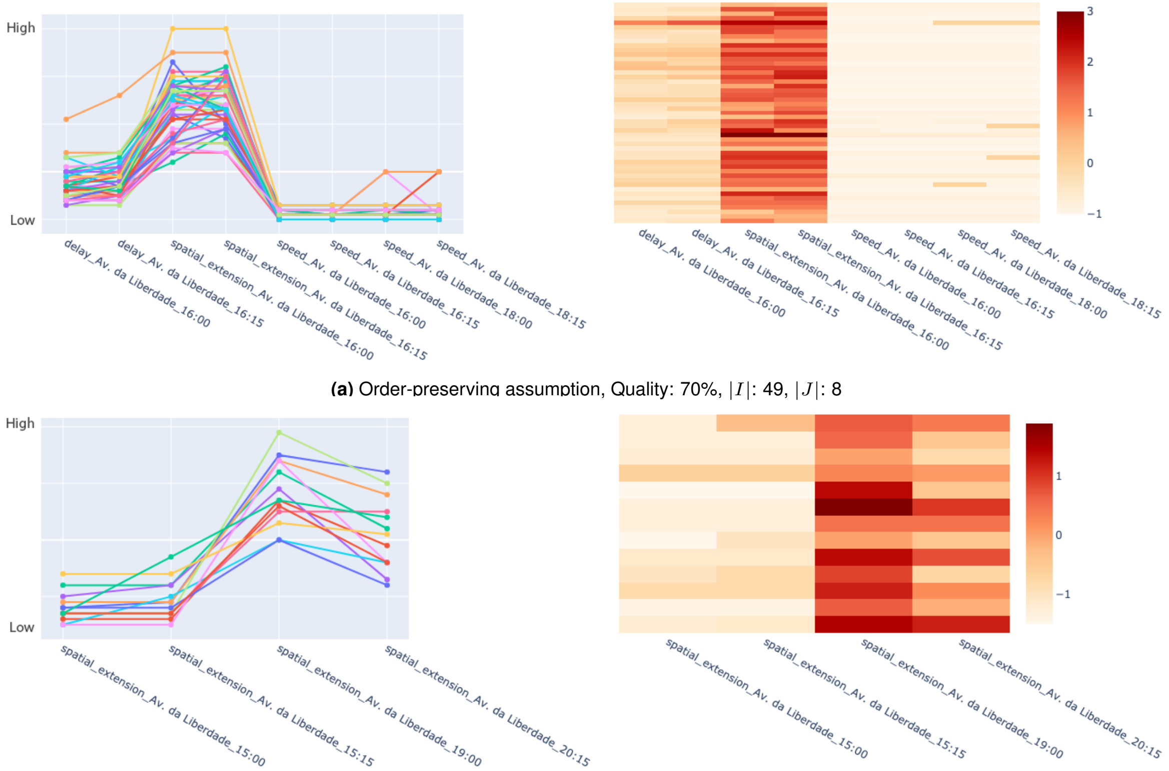


(a) Constant assumption, $|L|$: 3, Quality: 70%, $|I|$: 49, $|J|$: 8

(b) Constant assumption, $|L|$: 4, Quality: 70%, $|I|$: 41, $|J|$: 12

(c) Constant assumption, $|L|$: 3, Quality: 100%, $|I|$: 5, $|J|$: 12

Figure 5.8: Three illustrative constant patterns of road traffic found in WAZE data.



(a) Order-preserving assumption, Quality: 70%, $|I|$: 49, $|J|$: 8

(b) Order-preserving assumption, Quality: 70%, $|I|$: 13, $|J|$: 4

Figure 5.9: Three illustrative order-preserving patterns of road traffic found in WAZE data.

As introduced (Section 2.1), collections of WAZE events are characterized by an inherent structural sparsity – i.e. the mapped data structure can have an arbitrary-high amount of missing entries depending on the chose temporal granularity. In the conducted experiments, the amount of missing entries for the 15 minutes granularity surpasses 90%. This observation further confirms the robustness of pattern-based biclustering in discovering mobility patterns from highly sparse traffic data.

5.2.3 Integrative patterns of road traffic

Finally, we briefly show integrative traffic patterns from the consolidation of ILD and WAZE data sources. Table 5.3 describes the properties of the pattern solutions produced from specific biclustering searches. Given the need to account for cross-source relationships, we can observe that the resulting traffic patterns have in average either a lower number of supporting days (an average of 20 days from the monitored 60-day period) or a lower number of jam features (an average of approximately 10 features). A considerably high number of dissimilar and statistically significant patterns combining speed and volume views on road traffic was discovered. Tolerance to noise of these solutions can be easily customized in order to comprehensively find patterns with parameterizable degree of quality. In addition to noise-tolerance, $\eta_{ij} \notin [-\delta/2, \delta/2]$, coherence strength $\delta = \bar{\mathbf{A}}/|L|$ can be customized to comprehensively model

Query	Assumption	$\|L\|$	quality	#bics	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$p-$value <1E-3
1	Constant	3	70%	21	21.9±4.0	12.0±2.2	21
2	Constant	4	80%	56	20.5±2.1	11.4±1.3	56
3	Constant (speed limitation and ILD)	3	80%	77	5.3±2.7	10.6±1.0	77

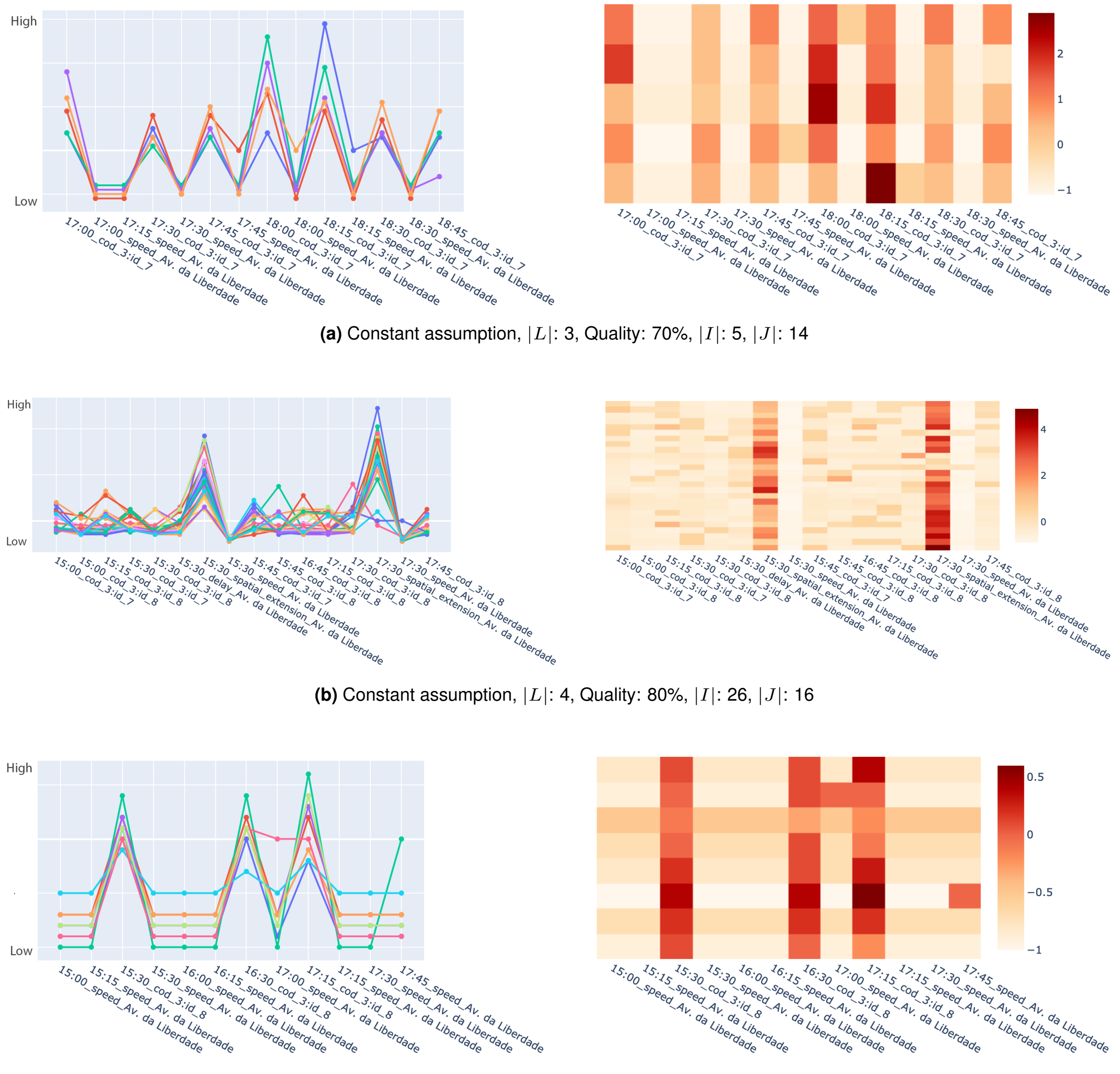
Table 5.3: Biclustering results from consolidated ILD and WAZE data using BicPAMS with different homogeneity criteria.

relations with slight-to-moderate deviations from traffic pattern expectations.

Figure 5.10 depicts three of the dozens of integrative traffic patterns found in Marquês de Pombal's junctures within the Lisbon city. The interesting aspects of all of these patterns is that they combine frequentist views pertaining to ILD data, as well as continuous views on speed and jam extension, pertaining to WAZE data. Considering the second depicted pattern (Figure 5.10b), it captures a traffic profile spanning different streets around Marquês de Pombal along different periods of the afternoon with a delineated jam profile in terms of flow, speed and spatial extent.

5.3 Statistical significance

Table 5.1 shows the ability of the target biclustering searches to find statistically significant relations within road traffic data. A bicluster is statistically significant if the number of days with a given congestion profile is unexpectedly low [15]. Figure 5.11 provides a scatter plot of the statistical significance (horizontal axis) and area $|I|$x$|J|$ (vertical axis) of constant biclusters with $|L|$=3 and $>$70% quality. This analysis suggests the presence of a soft correlation between size and statistical significance. We observe that a few biclusters from both ILD and WAZE data sources have low statistical significance (bottom right dots) and can therefore be discarded not to incorrectly bias mobility decisions.

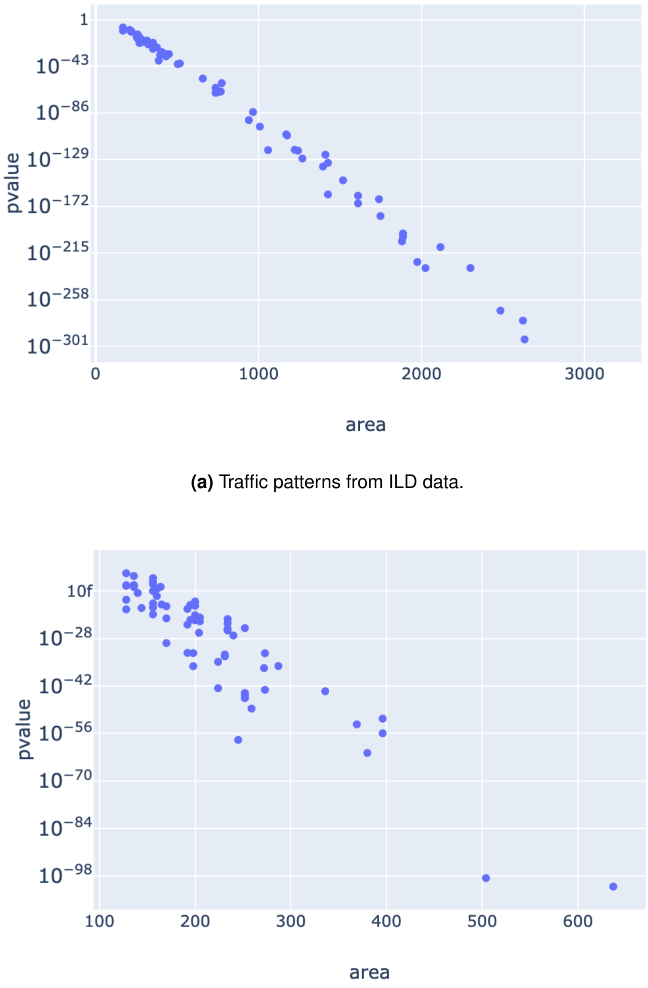


(a) Constant assumption, $|L|$: 3, Quality: 70%, $|I|$: 5, $|J|$: 14

(b) Constant assumption, $|L|$: 4, Quality: 80%, $|I|$: 26, $|J|$: 16

(c) Constant assumption, $|L|$: 3, Quality: 80%, $|I|$: 8, $|J|$: 12

Figure 5.10: Illustrative mobility patterns found from heterogeneous traffic data (event and time series traffic data), integrating views on traffic flow, speed and jam extension.



(a) Traffic patterns from ILD data.

(b) Traffic patterns from WAZE data.

Figure 5.11: Statistical significance versus size of the collected constant patterns of road traffic ($|L|$=3 and 70% of quality).

6

Situational Context

Contents

6.1 Context-incorporation mechanism 55
6.2 Results 56

Integrating situational context in the discovery of road mobility patterns allows us to further leverage our pattern mining solution. By integrating context in the discovery of patterns, we are able to find traffic profiles that are recurrent under specific context conditions. In this work we consider meteorological context, as this information can be crucial to reveal important aspects of the road infrastructure and personal travelling choices. For example, deficient water drainage can cause a recurrent profile of severe jams in a road, or high temperatures can cause people to prefer taking a road near the ocean, which in turn can cause abnormal congestions there. This section shows how we can use biclustering to find congestion profiles sensitive to situational context.

6.1 Context-incorporation mechanism

The first step to incorporate situational context in our solution is to apply appropriate data mappings to the context data, in our case meteorology data. These transformations will be similar to the ones applied to road traffic data, where temporal and calendric constraints are fixed, including date intervals, weekday annotations and time granularities to guide the context data aggregation. This will allow us to gather a structure that is conveniently designed to be integrated with road traffic data for the pattern discovery task.

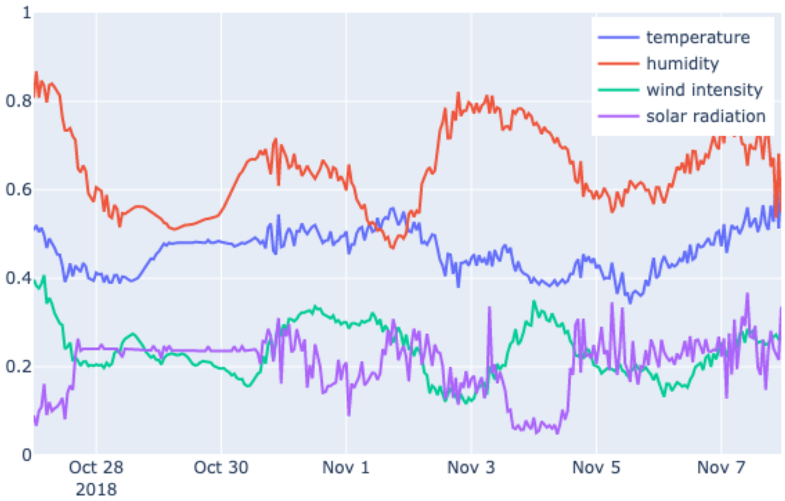
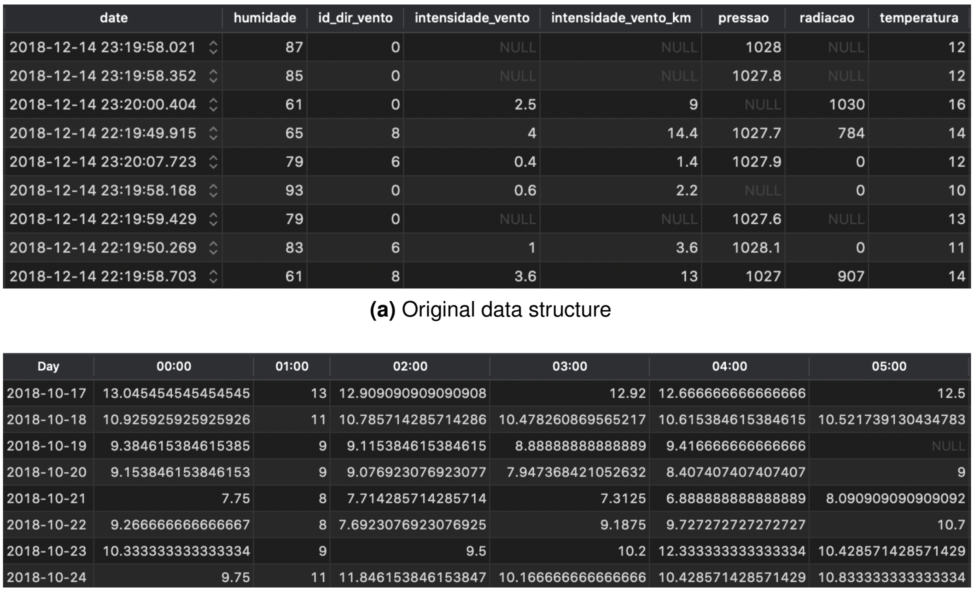


Figure 6.1: Time series plot of normalized meteorological variables between 27th October of 2018 and 8th November of 2018.

The used context dataset contains a diverse set of variables of meteorological conditions along time, namely temperature, humidity, wind intensity, and radiation levels (Figure 6.1). Figure 6.2 shows the original structure of the meteorological data and the corresponding data mappings for the temperature attribute.



date	humidade	id_dir_vento	intensidade_vento	intensidade_vento_km	pressao	radiacao	temperatura
2018-12-14 23:19:58.021	87	0	NULL	NULL	1028	NULL	12
2018-12-14 23:19:58.352	85	0	NULL	NULL	1027.8	NULL	12
2018-12-14 23:20:00.404	61	0	2.5	9	NULL	1030	16
2018-12-14 22:19:49.915	65	8	4	14.4	1027.7	784	14
2018-12-14 23:20:07.723	79	6	0.4	1.4	1027.9	0	12
2018-12-14 23:19:58.168	93	0	0.6	2.2	NULL	0	10
2018-12-14 22:19:59.429	79	0	NULL	NULL	1027.6	NULL	13
2018-12-14 22:19:50.269	83	6	1	3.6	1028.1	0	11
2018-12-14 22:19:58.703	61	8	3.6	13	1027	907	14

(a) Original data structure

Day	00:00	01:00	02:00	03:00	04:00	05:00
2018-10-17	13.045454545454545	13	12.909090909090908	12.92	12.666666666666666	12.5
2018-10-18	10.925925925925926	11	10.785714285714286	10.478260869565217	10.615384615384615	10.521739130434783
2018-10-19	9.384615384615385	9	9.115384615384615	8.88888888888889	9.416666666666666	NULL
2018-10-20	9.153846153846153	9	9.076923076923077	7.947368421052632	8.407407407407407	9
2018-10-21	7.75	8	7.714285714285714	7.3125	6.888888888888889	8.090909090909092
2018-10-22	9.266666666666667	8	7.6923076923076925	9.1875	9.727272727272727	10.7
2018-10-23	10.333333333333334	9	9.5	10.2	12.333333333333334	10.428571428571429
2018-10-24	9.75	11	11.846153846153847	10.166666666666666	10.428571428571429	10.833333333333334

(b) Transformed data structure

Figure 6.2: Meteorological data mapping.

Traffic-context consolidation. Under the previous mappings, situational context data can be matched to the road traffic data by day and time point. We discretize the road traffic data and the situational context mask separately, which allows us to calibrate independently the coherence strength for both data sources. As an example, consider $|L|$=3 (symbols $\{A, B, C\}$) for the road traffic data and $|L|$=2 (symbols $\{A, B\}$) for the situational context mask, where each symbol represents a value range that can overlap with others to prevent discretization problems. Table 6.1a illustrates road traffic data, Table 6.1b the situational context mask and Table 6.1c the result of the concatenation between both. The resulting concatenation between road traffic data and the context mask still preserves a structure adequate for the application of biclustering.

6.2 Results

To illustrate the potentialities of integrating situational context in the discovery of road traffic patterns, an experiment was conducted using the same experimental setting as in Section 5.2. For this analysis we consider only WAZE's geolocalized speed data since we want to focus on explore context's influence on congestion profiles, information that is given in a more confiable way in WAZE's data than on ILD data.

Day	ILD$_1$_00:00h	···	ILD$_n$_23:00h	Delay_Road$_1$_00:00h	···	Delay_Road$_n$_23:00h
D$_1$	A	···	B	C	···	A
D$_2$	C	···	A	C	···	B
···	···	···	···	···	···	···
D$_{n-1}$	A	···	B	C	···	B
D$_n$	B	···	C	A	···	B

(a) Illustrative road traffic data with $|L|$=3 (symbols $\{A, B, C\}$).

Day	ILD$_1$_00:00h	···	ILD$_n$_23:00h	Delay_Road$_1$_00:00h	···	Delay_Road$_n$_23:00h
D$_1$	D	···	E	E	···	D
D$_2$	E	···	E	E	···	D
···	···	···	···	···	···	···
D$_{n-1}$	E	···	E	D	···	D
D$_n$	E	···	D	D	···	E

(b) Illustrative situational context mask with $|L|$=2 (symbols $\{D, E\}$).

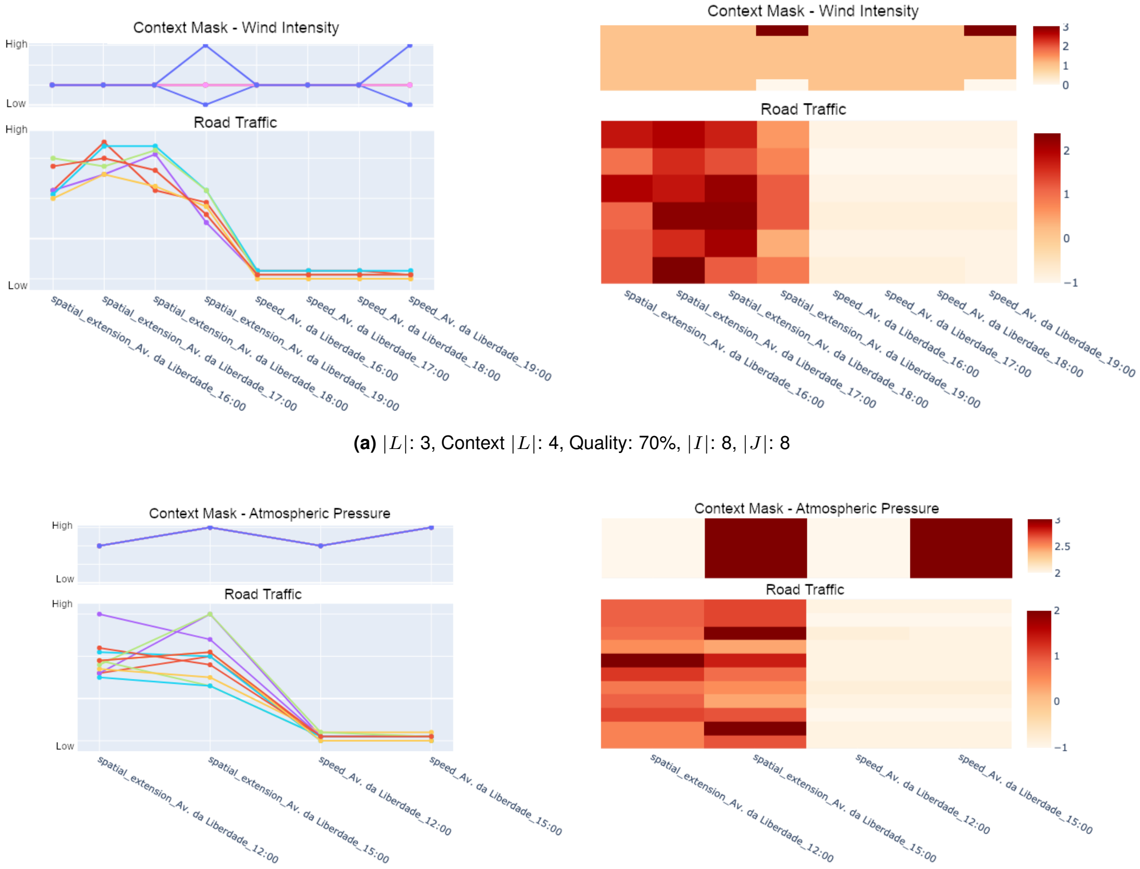
Day	ILD$_1$_00:00h	···	ILD$_n$_23:00h	Delay_Road$_1$_00:00h	···	Delay_Road$_n$_23:00h
D$_1$	A:D	···	B:E	C:E	···	A:D
D$_2$	C:E	···	A:E	C:E	···	B:D
···	···	···	···	···	···	···
D$_{n-1}$	A:E	···	B:E	C:D	···	B:D
D$_n$	B:E	···	C:D	A:D	···	B:E

(c) Result of concatenation between a) and b).

Table 6.1: Illustrative example to demonstrate the concatenation between road traffic data and situational context mask.

Figure 6.3 visually depicts two constant road traffic patterns from WAZE's geolocalized speed data and the corresponding situational context mask. Each context-incorporated traffic pattern capture coherent variations of traffic that are recurrent under specific context conditions, in this case meteorological conditions.

Figure 6.3a shows an evening peak hour congestion profile that is recurrent under medium-low wind intensity. It also shows that our solution preserves the robustness to the items-boundaries problem, allowing slight deviations in the context mask. Figure 6.3b depicts a profile with large jam extensions and severe speed limitations over a large period of time where the atmospheric pressure varies between medium-high and high values.



(a) $|L|$: 3, Context $|L|$: 4, Quality: 70%, $|I|$: 8, $|J|$: 8

(b) $|L|$: 3 Context $|L|$: 4, Quality: 80%, $|I|$: 13, $|J|$: 4

Figure 6.3: Illustrative patterns found using context-incorporation on geolocalized speed data.

Part III

Emerging Pattern Mining

This part of the thesis details E2PAT, a scalable method to comprehensively detect emerging patterns from heterogeneous sources of spatiotemporal data generated by large sensor networks.

Similarly to the method proposed on Part II of the thesis, we want to address the challenges posed by the inherent complexity of traffic data to comprehensively discover patterns of urban traffic. Here we consider the traffic data gathered by mobile sensors, road cameras, and loop detectors, to be a heterogeneous sensor network. Heterogeneous sensor networks include systems such as traffic monitoring systems and telemetry systems, which produce massive spatiotemporal data that offer the opportunity to acquire comprehensive views of systems' behavior along time.

Discovering emerging patterns in such heterogeneous sensor networks is essential to identify important changes that reveal needs for actuation [71]. Illustrating, increased utility, processing and communication needs along certain routes of a supply network may reveal future bottlenecks, propelling dynamic rebalancing initiatives. In urban mobility, emerging patterns reveal ongoing changes in city traffic dynamics, whose growth along time may indicate the establishment of new congestion trends with impact on the normal traffic flow [4, 46]. Those trends can evolve to create traffic bottlenecks if timely precautions are not taken [47]. As such, the early detection of emerging patterns offers urban planners the opportunity to make the necessary provisions to urban mobility.

The proposed method E2PAT, is able to discover emerging patterns from heterogeneous spatiotemporal data in linear-time. It combines three simplistic yet effective operations – time series differencing, spatial intersection and regression calculus – for the efficient discovery of all emerging patterns observed along geographies of interest. In addition, we propose an integrative score to measure the relevance of emerging patterns that yield statistical properties of interest. To mine emerging patterns sensible to situational context, we propose a context-aware filtering mechanism that filters data according to several context variables, guiding E2PAT's pattern discovery.

7

Solution

Contents

7.1 Spatiotemporal data mappings 65

7.2 Series differencing and emergence 67

7.3 Integrative scoring 69

7.4 Computation complexity of E2PAT 70

As introduced, our work aims at efficiently discovering emerging patterns from heterogeneous sources of spatiotemporal data produced by sensor networks. In particular, and as the motivating study case, we focus on emerging patterns of road mobility from georeferenced time series data produced by stationary loop counters, and multivariate event collections produced by GPS sensors. This task is challenged by the inherently complex spatiotemporal nature, heterogeneity, and massive size of the target sensor data. To address these challenges, we propose a linear-time method to comprehensively discover emerging patterns, termed E2PAT (Emerging Event PATtern miner).

E2PAT combines three simplistic yet effective principles: i) spatial intersection and time windowing operations for the comprehensive traversal of search space (Section 7.1); ii) combined use of time series differencing operations with linear regressors (Section 7.2); and iii) integrative scoring to measure the relevance of emerging patterns and control the amount of false positive and false negative discoveries (Section 7.3). E2PAT is available at Github.

7.1 Spatiotemporal data mappings

E2PAT is a two-step process. First, transformation procedures are applied to consolidate the original spatiotemporal data sources and map them into new data structures more conducive to the subsequent mining task. To this end, spatial and temporal constraints can be inputted at this stage to guide the discovery. Second, emerging patterns are discovered from the transformed data by combining differencing, regression and integrative scoring principles.

7.1.1 Spatial constraints

For handling trajectories of arbitrary length, there is the need to fix an adequate spatial granularity. E2PAT offers two major possibilities. First, E2PAT can rely on an already established categorization. For instance, street names in the context of road trajectory data or every segment between two junctures/nodes from a sensor network are supported criteria.

Second, the categorization can be automatically produced using a geographical mesh/grid for segmenting the set of all possible trajectories. The granularity of the input mesh can either create coarser or finer spatial views in comparison with the first option.

Under the selected spatial granularity, events are then linked to one or more segments in accordance with their spatial extent. To this end, simplistic yet efficient trajectory-mesh indexation and trajectory-segment intersections are applied to associate events to segments in linear time (Section 7.4).

Finally, the events associated with each one of the identified segments are temporally ordered to compose a *sparse (multivariate) time series* from each event's timestamp, t, and (multivariate) observation, $\mathbf{x}$.

7.1.2 Temporal constraints

Three major types of temporal constraints can be placed. First, calendrical constraints can be placed to segment the available data in (possibly overlapping) data chunks, and data transformations applied on each chunk. By default, the day of the week, weekdays, holidays, and on/off-academic period calendars are considered. Emerging patterns are discovered in linear-time for each one of these calendars (Section 7.4).

Second, a time granularity (e.g. 15-minute, hour or on/off-peak intervals) can be optionally specified to guide the discovery of emerging behaviors. In its absence, the proposed pattern discovery is iteratively performed using multiple time aggregations. Note that emerging patterns are not detected over a continuous timeline due the daily traffic cycles. Instead, they are discovered on these time windows throughout the days of the previously fixed calendar.

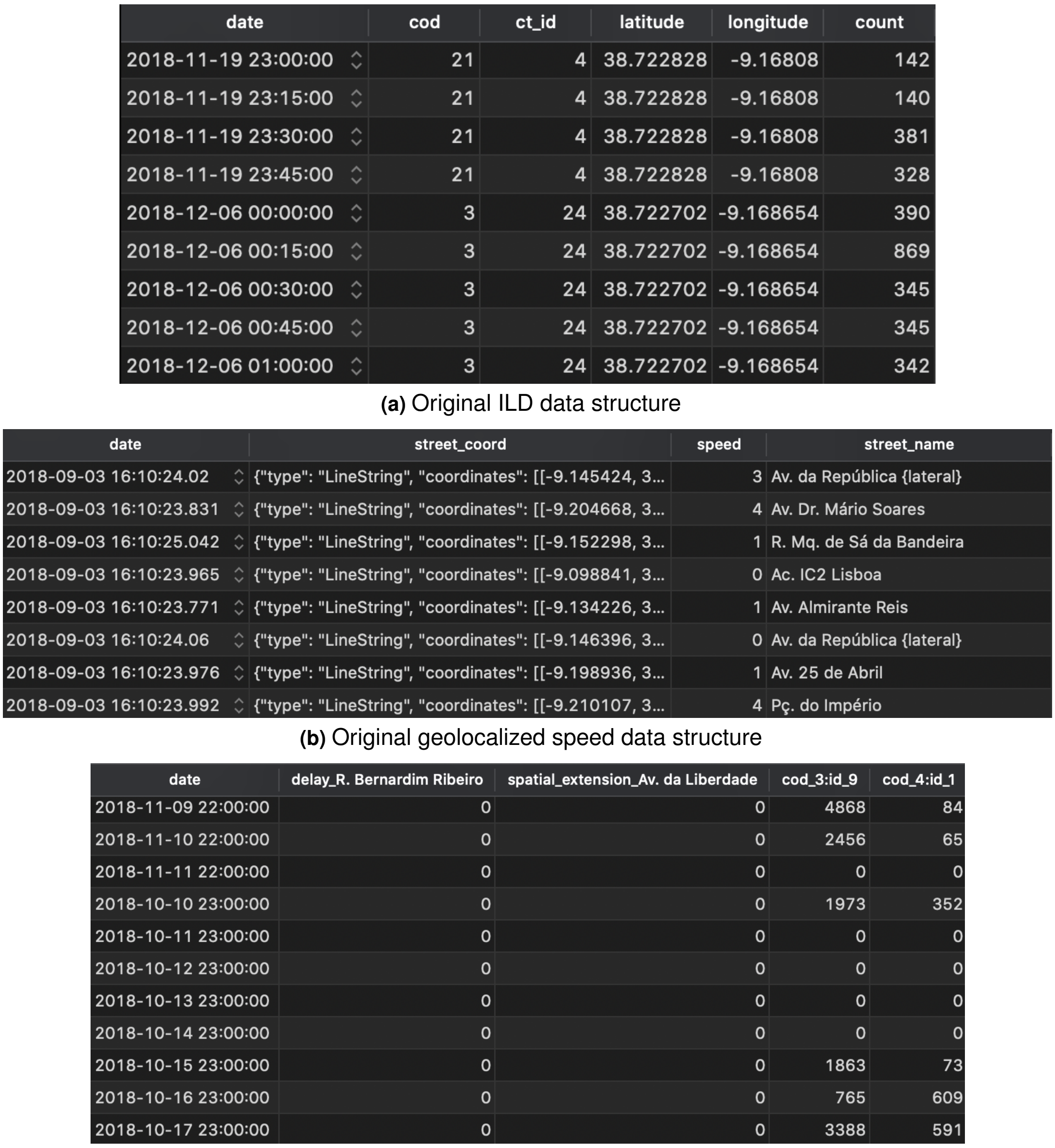
Given a specific time granularity, georeferenced time series can then be resampled using aggregators (e.g. *sum* of vehicles per time interval, *average* vehicle speed per time interval). In the context of spatiotemporal event data, the sparse series produced under the principles introduced in previous Section 7.1.1 are as well resampled using aggregation procedures in accordance with the target variables (e.g. event with *maximum* spatial extension per time interval, or severity *mode* from the occurring events).

Third, larger time windows, spanning a fixed number of days, can be optionally specified to guide the discovery of patterns whose emerging behavior is only recently elicited (spanning just a partial period). By default, a single window spanning all the available data is considered since late-occurring patterns can still be detected under the proposed differencing operations.

7.1.3 Data transformation

Once these constraints are fixed, data mappings are applied to transform the original spatiotemporal data structures into multivariate time series structures, more conducive to the subsequent step. In the target structure, each observation corresponds to a day from a specific calendar at a specific time (see Section 7.1.2) and each variable measures some aspect of the target system at a specific location. Considering road traffic monitoring systems, the transformed ILD data measures the number of cars passing over a single loop detector during a specific time interval for each calendar day. For the geolocalized speed data, multiple measurements are taken per event and principles from Section 7.1.1 applied to aggregate events per road segment. Figure 7.1a and 7.1b show the original structures of ILD and geolocalized speed data.

The integration of the previous mappings is a simple concatenation of the time series variables resulting from the transformation of each data source, as shown in Figure 7.1c.



date	cod	ct_id	latitude	longitude	count
2018-11-19 23:00:00	21	4	38.722828	-9.16808	142
2018-11-19 23:15:00	21	4	38.722828	-9.16808	140
2018-11-19 23:30:00	21	4	38.722828	-9.16808	381
2018-11-19 23:45:00	21	4	38.722828	-9.16808	328
2018-12-06 00:00:00	3	24	38.722702	-9.168654	390
2018-12-06 00:15:00	3	24	38.722702	-9.168654	869
2018-12-06 00:30:00	3	24	38.722702	-9.168654	345
2018-12-06 00:45:00	3	24	38.722702	-9.168654	345
2018-12-06 01:00:00	3	24	38.722702	-9.168654	342

(a) Original ILD data structure

date	street_coord	speed	street_name
2018-09-03 16:10:24.02	{"type": "LineString", "coordinates": [[-9.145424, 3...	3	Av. da República {lateral}
2018-09-03 16:10:23.831	{"type": "LineString", "coordinates": [[-9.204668, 3...	4	Av. Dr. Mário Soares
2018-09-03 16:10:25.042	{"type": "LineString", "coordinates": [[-9.152298, 3...	1	R. Mq. de Sá da Bandeira
2018-09-03 16:10:23.965	{"type": "LineString", "coordinates": [[-9.098841, 3...	0	Ac. IC2 Lisboa
2018-09-03 16:10:23.771	{"type": "LineString", "coordinates": [[-9.134226, 3...	1	Av. Almirante Reis
2018-09-03 16:10:24.06	{"type": "LineString", "coordinates": [[-9.146396, 3...	0	Av. da República {lateral}
2018-09-03 16:10:23.976	{"type": "LineString", "coordinates": [[-9.198936, 3...	1	Av. 25 de Abril
2018-09-03 16:10:23.992	{"type": "LineString", "coordinates": [[-9.210107, 3...	4	Pç. do Império

(b) Original geolocalized speed data structure

date	delay_R. Bernardim Ribeiro	spatial_extension_Av. da Liberdade	cod_3:id_9	cod_4:id_1
2018-11-09 22:00:00	0	0	4868	84
2018-11-10 22:00:00	0	0	2456	65
2018-11-11 22:00:00	0	0	0	0
2018-10-10 23:00:00	0	0	1973	352
2018-10-11 23:00:00	0	0	0	0
2018-10-12 23:00:00	0	0	0	0
2018-10-13 23:00:00	0	0	0	0
2018-10-14 23:00:00	0	0	0	0
2018-10-15 23:00:00	0	0	1863	73
2018-10-16 23:00:00	0	0	765	609
2018-10-17 23:00:00	0	0	3388	591

(c) Integrative series data structure

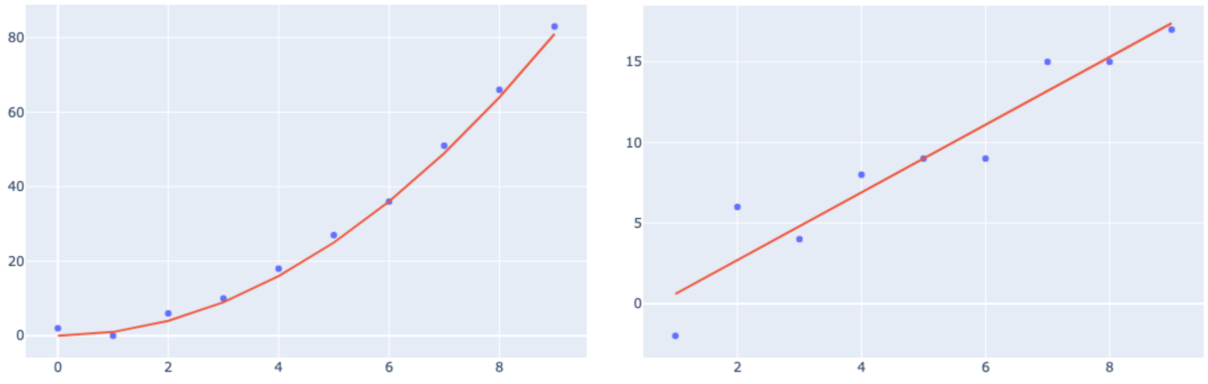
Figure 7.1: Original and transformed spatiotemporal data structures.

7.2 Series differencing and emergence

7.2.1 Comprehensive emerging pattern discovery

The proposed mapping generates as many multivariate time series as the number of calendars, time intervals per day, and fixed spatial granularity (Section 7.2), i.e. number of regions in the context of mobile sensors and sensorized locations in the context of stationary sensors.

E2PAT is then applied for each variable of each multivariate time series in order to detect isolated



(a) Original time series

(b) Differenced time series

Figure 7.2: Role of time series differencing operations for detecting emerging trends using linear models.

emerging behavior in accordance with differencing and regression principles (Sections 7.2.2 and 7.2.3). The found emerging behaviors are then comprehensively assessed using integrative scoring principles (Section 7.3) and combined to produce the target emerging pattern profiles.

7.2.2 Differencing

Given the fact that heterogeneous sensor networks produce massive data, learning non-linear (auto-)regressive models is a computationally expensive task, intractable to the target end. To tackle this observation, we make use of simple yet effective time series differencing operations.

Time series differencing is the act of subtracting consecutive observations from a time series, $\mathbf{x}_{t+1} - \mathbf{x}_t$. Time series differencing has been traditionally applied in time series analysis to stationarize non-stationary time series. In the context of our work, we use this principle with a different end, to approximate emerging trends using linear regressions as illustrated in Figure 7.2. When a time series follows an exponential trend explained by a certain growth factor (e.g. Figure 7.2a), the differenced time series will have a linear behavior (e.g. Figure 7.2b). This turns time series differencing a robust candidate for the targeted task.

7.2.3 Regression

E2PAT allows the search for three types of patterns: simple, emerging and abruptly changing patterns. *Simple patterns* are simple trends in the original time series data, approximated by a linear regression with the dependent variable being a variable (e.g. speed limit or jam spatial extent) and the independent variable the calendar day.

Emerging patterns are trends when the target series variable is differenced, allowing us to capture exponential trends.

Finally, *abruptly changing patterns* are trends observed in time series after two differencing operations (i.e. second-order differenced series).

To identify the linear trends on the (differenced) time series, a simple linear regression is estimated on each variable using the least squares method, with slope, d,

$$d = \frac{\sum_{i=1}^{n} x_i t_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} t_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i^2}{n}} \quad , \tag{7.1}$$

and a coefficient of determination, r^2,

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{x}_i - \bar{\hat{x}})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad . \tag{7.2}$$

Linear-time decomposition of the time series can be optionally applied to remove seasonal and cyclical components for a more correct approximation of the determination coefficient.

7.3 Integrative scoring

An integrative scoring, yielding statistical properties of interest, is proposed to quantify the relevance of patterns. The score is further used filter the outputted emerging patterns in order to minimize the presence of both false positive patterns (retrieved yet not relevant) and false negative patterns (not retrieved yet relevant).

The proposed score function is influenced by four major attributes:

- the slope of the linear regression, measuring the growth rate of the pattern;
- the r^2 of the regressive model, measuring the accuracy term of the pattern (1 when optimal and near 0 when random);
- the relative support of the pattern (i.e. the number of observations in the pattern divided by the maximum number of observations found within the same data source);
- the differencing order (Section 7.2.2), a term to favour emerging patterns against simple trends.

The proposed scoring function is given by:

$$\text{score}(d, r^2, sup, p) = (\alpha_1 \times d + \alpha_2 \times r^2 + \alpha_3 \times sup)^{\alpha_4(1+p)} \tag{7.3}$$

where d is the slope of the fitted linear regression (1), r^2 is the coefficient of determination (2), *sup* is the relative support, and p is differencing order (0 if absent). α_1, α_2, α_3 and α_4 are parameterizable weights for each factor that can obtain under a sensitivity analysis. Compelling empirical evidence from road monitoring traffic systems suggests α_1=0.3, α_2=0.5, α_3=0.2 and α_4=1 as default values.

The d term is bounded between -1 and 1, while r^2 and sup terms are bounded between 0 and 1. As a result, the proposed score function is also bounded between -1 and 1.

Because the sign of the score is dictated by the slope d, it easily informs whether we are in the presence of a congestion pattern (positive score) or a decongestion pattern (negative score). In the context of urban mobility, some of the monitored variables including speed limits, jam spatial extent, jam recurrence, jam delay, jam severity, and car frequencies. With the exception of speed limits, all the values measured for the remaining road traffic variables increase when congestion levels increase. As such, we change the sign of the slope of the speed limit variables, so that patterns with different road traffic variables can be interpreted seamlessly.

In addition to its easily interpretable bounds, the statistical distribution of the scores computed for the found patterns from spatiotemporal traffic data reveals the the observed values approximately follows a centered Gaussian distribution, passing tests of normality at α=0.05 significance level.

7.4 Computation complexity of E2PAT

Theorem. *E2PAT has linear time complexity on the input data size.*

Proof. First, considering georeferenced time series data. Let us assume the presence of r stationary devices, each measuring m variables along T steps. Input data in this context has $O(rmT)$ size. Now consider the presence of k_1 calendars and the presence of k_2 time periods per day. The proposed transformation process will lead to the formation of k_1k_2r time series, each with a m multivariate order. This leads to time series data of size $O(k_1k_2rm\frac{T}{k_2})$=$O(k_1rmT)$. Note that the number of calendars is always a small constant. For instance, considering the days of the week, where weekdays are further decomposed according to on and off-academic periods, k_1=2+5×2=12. As such, the produced time series have size $O(rmT)$. As the data transformation step is just based on linear-time segmentation and resampling operations, the computational complexity of this step is in fact $O(rmT)$.

Three differencing operations, k_3=3, are applied to allow the discovery of emerging and abruptly changing patterns. As differencing is a linear operation, this step takes $O(k_3rmT)$=$O(rmT)$ time. Finally, linear regressions are learn for each variable of each time series. Since the calculus of formulae (1) and (2) is also accomplished in linear time, the time complexity of this step is $O(k_3rmT)$=$O(rmT)$.

Let us now consider spatiotemporal event data, and the presence of massive number of q events with varying timestamps and trajectories spanning different geographies. Let us consider the presence of a total number of k_4 trajectory segments. The production of these segments against a user-defined mesh can be understandably performed in linear time using a simple spatial data structure. In addition, the indexation of the q events into these segments has time complexity of $O(k_4 q)$. Since k_4 is also a constant, the time complexity of the transformation stage is $O(q)$. Similarly as described for georeferenced time series data, the subsequent steps on the transformed event series with total size $O(qm)$ (where m is the multivariate order of the events) can be computed in linear time, yielding $O(qm)$ complexity. $\square$

8

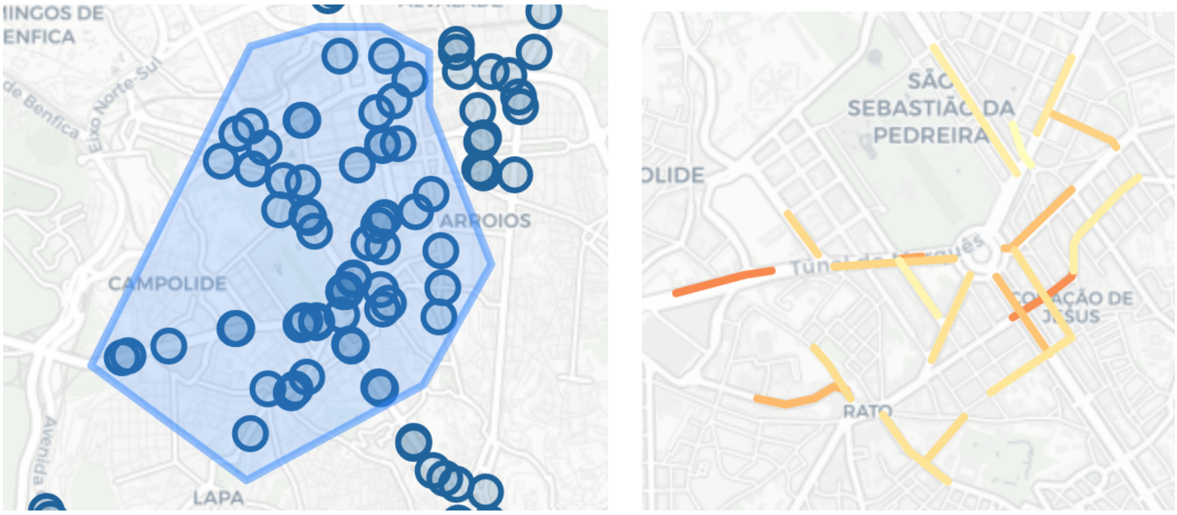
Results

Contents

8.1 Emerging patterns of road traffic 75

8.2 E2PAT visualization tool 78

Considering the Lisbon city as a study case, we applied the proposed approach to comprehensively discover emerging traffic patterns of road mobility from geolocalized speed data provided by WAZE and inductive loop detector (ILD) data collected during a two month period in central junctures of the city (Figure 8.1). To illustrate the enumerated potentialities, the gathered results from consolidated ILD-WAZE data sources are discussed.



(a) ILD locations

(b) WAZE events

Figure 8.1: Map visualization of the two sources of urban traffic data along the studied area: a) ILD sensor placement; b) WAZE events.

Experimental setting. The score parameters were chosen by empirically experimenting different values and validating the relevance of the best scored patterns with mobility experts from LNEC[1] and CML[2]. Upon experimental analysis, a higher weight was allocated to the r^2 term to ensure that high scored patterns corresponded to regressions with a good degree of fitness. Since we are analyzing consolidated sensor data from ILD and WAZE event data, a lower weight was given to the support terms so that patterns from the sparser event data are not penalized in relation to the more dense stationary ILD data. The fixed score parameter values were: $\alpha_1 = 0.3$; $\alpha_2 = 0.5$; $\alpha_3 = 0.2$; $\alpha_4 = 1$.

8.1 Emerging patterns of road traffic

Table 8.1 presents the best scored congestion and decongestion traffic patterns on the consolidated ILD-WAZE data. The results gathered from our solution capture a wide variety of simple and emerging patterns on different traffic variables, spanning different road segments at different periods of the day. By having a quick overview of the table results, we can remark some interesting aspects: (i) the discrepancy

[1] http://www.lnec.pt/en/departmental-units/transportation-department/

[2] https://www.lisboa.pt/cidade/urbanism

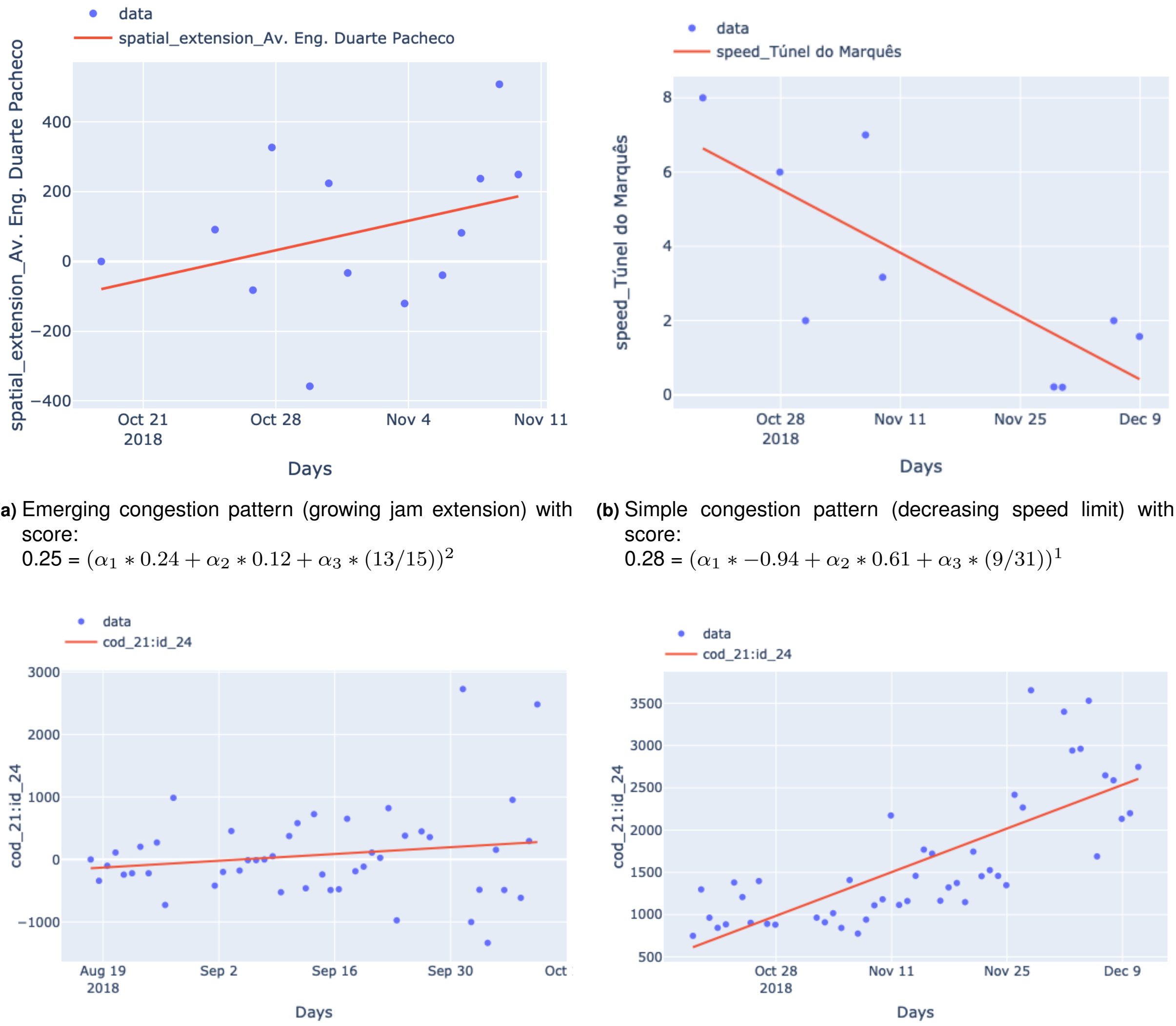
between the support of ILD and WAZE patterns did not prevent the discovery of varied patterns, showing the importance of a proper score parameterizations able to handle arbitrarily-high data sparsity levels; (ii) the r^2 of emerging patterns from the stationary ILD sensors is generally lower than those produced from mobile sensors. A minimum threshold on the r^2 term can therefore be place to guarantee the absence of false positive discoveries; (iii) among the highest and lowest scored patterns, we find an emerging sharing the same street (R. Castilho) and time (13:00) due to an increase of jam extensions but, simultaneously, a higher throughput of cars.

To guide the interpretation of the found patterns and the associated scores, Figure 8.2 visually depicts four of the found patterns (two emerging and two simple patterns from each data source). The red line corresponds to the resulting regression and the blue dots are the data points after the undertaken

Time	Location	Attribute	Pattern Type	Score	R^2	Slope	Support
13:00	R. Castilho	Spatial Extension	Emerging	0.51	0.81	0.50	5
22:00	Av. da Liberdade	Spatial Extension	Simple	0.39	0.75	1.00	6
08:00	R. Castilho	Spatial Extension	Emerging	0.36	0.68	0.31	5
13:00	R. Braamcamp	Spatial Extension	Simple	0.33	0.98	0.20	6
22:00	Túnel do Marquês	Speed	Simple	0.28	0.61	-0.94	9
03:00	Av. da Liberdade	Spatial Extension	Emerging	0.27	0.55	0.26	6
10:00	R. Sousa Martins	Delay	Emerging	0.24	0.60	0.10	5
18:00	R. Sol ao Rato	Spatial Extension	Simple	0.24	0.74	0.61	5
16:00	R. Joaquim António de Aguiar	Delay	Emerging	0.24	0.54	0.20	5
20:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	0.22	0.01	0.07	48
19:00	ILD cod$_{21}$:id$_{12}$	Car Frequency	Emerging	0.22	0.02	0.05	48
19:00	ILD cod$_{3}$:id$_{23}$	Car Frequency	Emerging	0.22	0.02	0.05	48
12:00	ILD cod$_{3}$:id$_{14}$	Car Frequency	Emerging	0.22	0.03	0.06	47
12:00	ILD cod$_{21}$:id$_{24}$	Car Frequency	Emerging	0.22	0.03	0.06	47
13:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	0.21	0.00	0.06	47
19:00	ILD cod$_{21}$:id$_{24}$	Car Frequency	Emerging	0.21	0.01	0.04	48
19:00	ILD cod$_{3}$:id$_{14}$	Car Frequency	Emerging	0.21	0.01	0.04	48
16:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	0.21	0.00	0.05	47
22:00	ILD cod$_{21}$:id$_{24}$	Car Frequency	Simple	0.21	0.58	0.31	48
22:00	ILD cod$_{3}$:id$_{14}$	Car Frequency	Simple	0.21	0.58	0.31	48
18:00	R. Sol ao Rato	Delay	Emerging	-0.21	0.39	-0.34	5
07:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.22	0.00	-0.07	47
12:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.22	0.00	-0.07	47
08:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.22	0.00	-0.08	47
22:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.22	0.00	-0.08	48
18:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.22	0.00	-0.08	48
11:00	R. Silva Carvalho	Speed	Emerging	-0.23	0.24	0.56	5
10:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.23	0.00	-0.11	47
11:00	R. Viriato	Spatial Extension	Simple	-0.24	0.89	-0.17	6
09:00	ILD cod$_{3}$:id$_{9}$	Car Frequency	Emerging	-0.24	0.00	-0.15	47
04:00	Av. da Liberdade	Speed	Simple	-0.27	0.78	0.57	6
23:00	Av. da Liberdade	Spatial Extension	Emerging	-0.31	0.15	-0.90	5
03:00	Av. da Liberdade	Speed	Simple	-0.34	0.76	0.87	6
08:00	R. das Amoreiras	Speed	Emerging	-0.34	0.14	1.00	5
13:00	R. Castilho	Speed	Simple	-0.35	0.72	1.00	5
08:00	R. Castilho	Speed	Simple	-0.36	0.83	0.75	5
18:00	R. Sol ao Rato	Speed	Simple	-0.39	0.77	0.99	5
22:00	Av. da Liberdade	Speed	Simple	-0.45	0.81	1.00	6
08:00	R. Castilho	Speed	Emerging	-0.49	0.76	0.56	5
13:00	R. Castilho	Speed	Emerging	-0.65	0.82	0.94	5

Table 8.1: Top 20 congestion and decongestion traffic patterns in the studied area under a weekday calendar.

transformation procedures. Fig.8.2a presents an emerging pattern associated with an increased traffic queue. The pattern has a considerably good score, which is visually justified by the accentuated slope and moderate fitness to the data points. Fig.8.2b depicts a speed congestion pattern, which for this variable has negative slope (decreasing limit). Fig.8.2c shows an ILD emerging pattern with a relatively low r^2, but with a small moderately positive tendency from many available data points. Finally, Fig.8.2d is one of the top simple ILD patterns, showing an accentuated slope and a good fitness term.



(a) Emerging congestion pattern (growing jam extension) with score:
0.25 = $(\alpha_1 * 0.24 + \alpha_2 * 0.12 + \alpha_3 * (13/15))^2$

(b) Simple congestion pattern (decreasing speed limit) with score:
0.28 = $(\alpha_1 * -0.94 + \alpha_2 * 0.61 + \alpha_3 * (9/31))^1$

(c) Emerging traffic throughput trend with score:
0.21 = $(\alpha_1 * 0.06 + \alpha_2 * 0.03 + \alpha_3 * (47/48))^2$

(d) Simple traffic throughput trend with score:
0.21 = $(\alpha_1 * 0.31 + \alpha_2 * 0.58 + \alpha_3 * (48/48))^1$

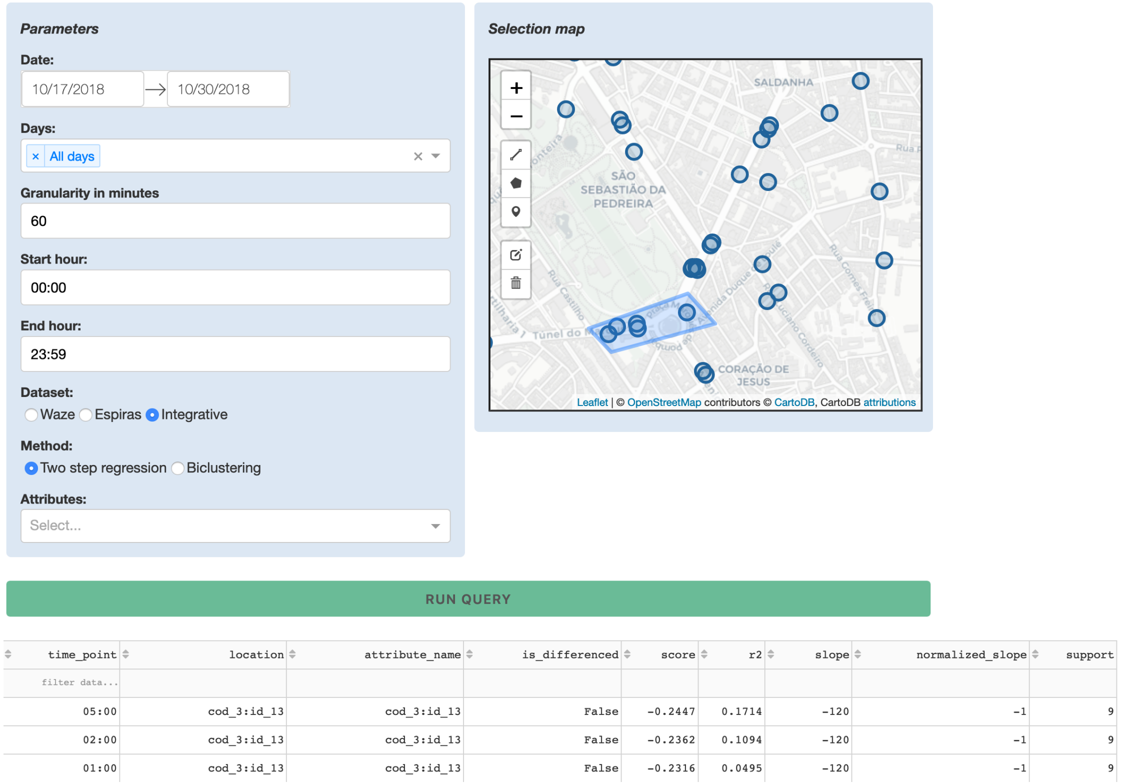
Figure 8.2: Illustrative set of patterns found by E2PAT in consolidated ILD-WAZE data.

8.2 E2PAT visualization tool

A visualization tool was developed to support the analysis and guide the navigation throughout the outputted pattern solutions. The E2PAT tool is integrated within a decision support system that is currently being deployed in the Lisbon city Council to support urban mobility reforms.

An overview of the tool is shown in Figures 8.3 and 8.4. The tool provides a user friendly interface for querying the desirable sources of spatiotemporal data by selecting the desirable types of sensors spread across the Lisbon's city and variables of interest. The data can be queried by date, under different calendrical and time granularity constraints, as well as filtered spatially using the geometric selection tool in the map.

The pattern solutions are presented using both interactive tables and interactive maps. The listed emerging patterns can be sorted by spatial and temporal criteria in order to aggregate potentially correlated patterns; as well as by the final score or by each one of the constituent terms (growth, fitness and support). The E2PAT further supports different importation-exportation facilities, allowing the queried data and pattern solutions to be exported into a CSV format to perform further analyzes.



Parameters

Date: 10/17/2018 → 10/30/2018

Days: × All days

Granularity in minutes: 60

Start hour: 00:00

End hour: 23:59

Dataset: ○ Waze ○ Espiras ● Integrative

Method: ● Two step regression ○ Biclustering

Attributes: Select...

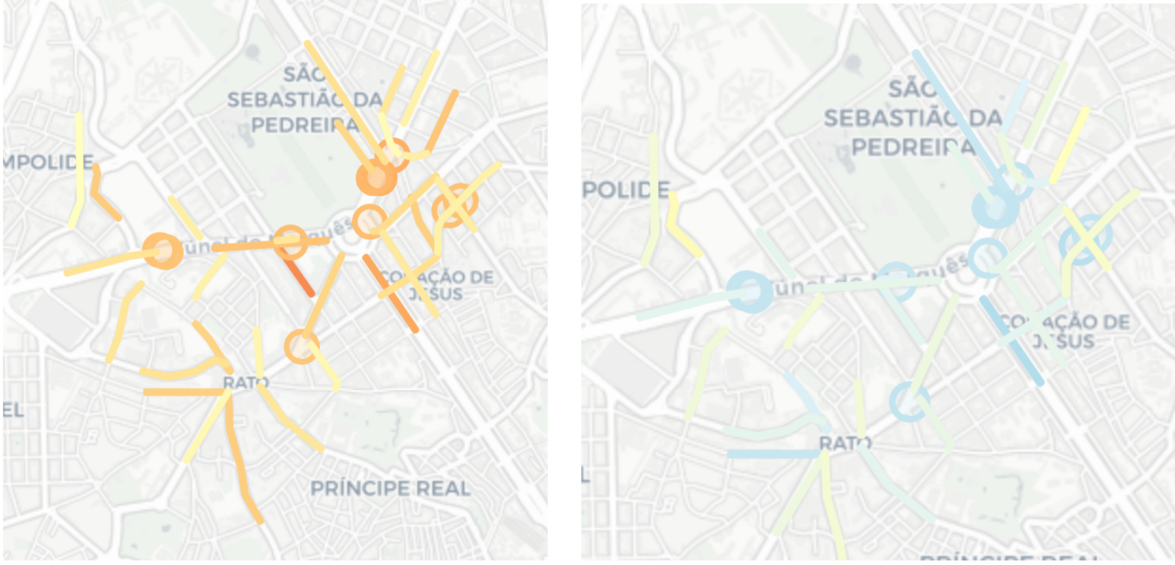
Selection map

RUN QUERY

time_point	location	attribute_name	is_differenced	score	r2	slope	normalized_slope	support
filter data...								
05:00	cod_3:id_13	cod_3:id_13	False	-0.2447	0.1714	-120	-1	9
02:00	cod_3:id_13	cod_3:id_13	False	-0.2362	0.1094	-120	-1	9
01:00	cod_3:id_13	cod_3:id_13	False	-0.2316	0.0495	-120	-1	9

Figure 8.3: Overview of the user dashboard for querying the road traffic data sources.

The outputted maps offer a wide-range of possibilites for users to comprehensively explore emerging patterns in accordance with their relevance and spatiotemporal properties. The map has a time selector which allows the user to select specific time points or aggregate results produced over a time range. The visualizations can also vary in accordance with the selected variables of higher interest. The count, which represents the number of times a trajectory was congested between a certain period, is summed, and for the speed and delay the mean is calculated.



(a) Congestion patterns

(b) Decongestion patterns

Figure 8.4: Map visualization of the found patterns from both ILD-WAZE data sources using score-based coloring of point-based and trajectory-based emerging patterns.

9

Situational Context

Contents

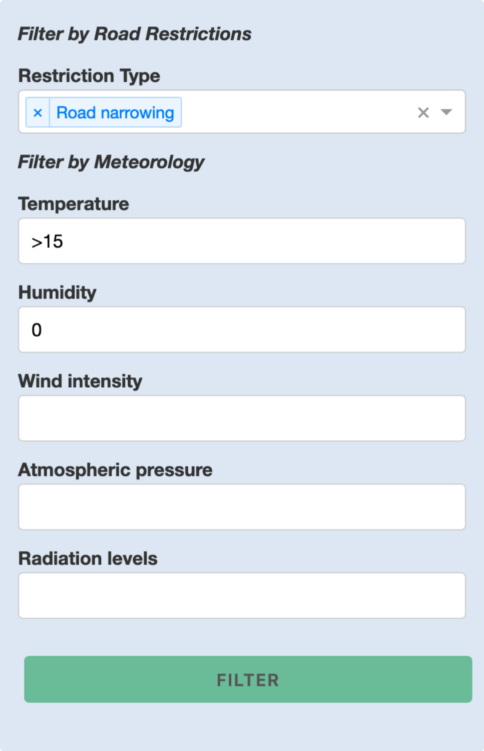
9.1 Context-aware filtering mechanism 83

9.2 Results 84

Likewise to the work done on Chapter 6 of the biclustering method, we developed a solution to mine emerging patterns sensible to context for E2PAT. The developed solution could also be applied to biclustering as it consists on applying additional filters in the data consolidation step, where spatial and temporal constraints are fixed to guide the pattern discovery.

9.1 Context-aware filtering mechanism

In addition to the querying abilities stated in Section 8.2, the tool also provides a user friendly interface (Figure 9.1) for querying data by situational context. The data can be queried by meteorological conditions allowing the guidance of the pattern discovery algorithm for specific weather states. This functionality can leverage our pattern mining solution by granting the ability to discover emerging patterns under specific meteorological conditions. Data can also be queried to filter out observations spatially and temporally close to road restrictions (e.g. road works, road narrowings). This functionality can potentially leverage the quality of the discovered emerging patterns, since we can remove observations that could possibly skew the results. Consider for example a road restricted by renovation works, this would likely cause a series of abnormal congestions in a certain area, that in turn would be discovered as emerging patterns. However, these congestions do not necessarily represent a change in traffic behavior that reveal needs for actuation, and the discovered patterns can be considered false positives.



Filter by Road Restrictions

Restriction Type: × Road narrowing

Filter by Meteorology

Temperature: >15

Humidity: 0

Wind intensity:

Atmospheric pressure:

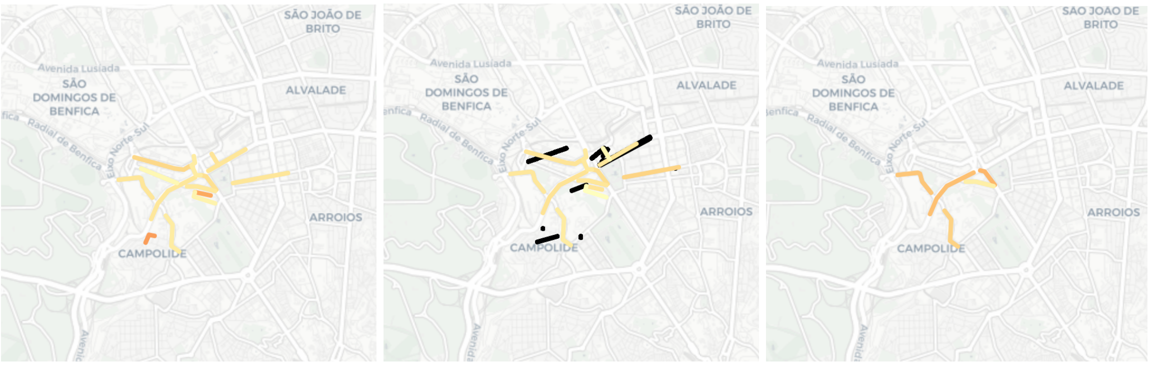
Radiation levels:

Figure 9.1: Overview of the interface for filtering road traffic data by situational context.

9.2 Results

To illustrate the potentialities of integrating situational context in the discovery of emerging road traffic patterns, an experiment was conducted using the same experimental settings as in Chapter 8 and data from geolocalized speed data provided by WAZE collected during a two month period. We analyze the results by presenting the discovered emerging patterns without situational context, and comparing them to the results with filtered road restrictions and under specific weather conditions (Figure 9.2).

Road restrictions. Table 9.1 presents the best scored congestion and decongestion traffic patterns on WAZE's geolocalized speed data without situational context, capturing a wide variety of simple and emerging patterns spanning different road segments at different periods of the day. By comparing the results to the ones gathered after removing observations that happened under the same time span as the captured restrictions (Table 9.2), we can remark some interesting aspects: (i) any of the captured congestion patterns captured without situational context were caught when filtering by road restrictions. This hints that most of the congestions weren't caused by emerging traffic behaviors on those locations and are a consequence of road restrictions in the area; (ii) the scores of the decongestion patterns are generally higher after filtering by road restrictions, hinting that some patterns weren't being captured due to the presence of observations that were skewing the results; (iii) the patterns on Av. Calouste Gunbenkian that were caught after filtering by road restrictions are also present in the results without situational context, however their score is different because of the use of relative support.



(a) Original **(b)** Filtered by road restrictions **(c)** Filtered by weather conditions

Figure 9.2: Map visualization of the geolocalized speed data: a) without situational context; b) filtered by road restrictions (visually depicted in black); c) filtered by air humidity superior to 80%.

Meteorological conditions. Table 9.3 presents the results gathered by filtering the road traffic data by observations recording while air humidity was superior to 80%. We can note that the captured patterns were different when applying a restriction on weather condition, proofing that we can guide the pattern discovery to be sensitive to contextual variables.

Time	Location	Attribute	Pattern Type	Score	R^2	Slope	Support
19:00	Av. Miguel Bombarda	Speed	Emerging	0.24	0.47	-0.29	5
15:00	Av. Ant. Aug. de Aguiar	Speed	Emerging	0.20	0.00	-0.01	23
15:00	Av. Ant. Aug. de Aguiar	Delay	Emerging	0.20	0.00	0.00	23
15:00	Av. Ant. Aug. de Aguiar	Spatial Extension	Emerging	0.20	0.00	0.00	23
10:00	Av. Ant. Aug. de Aguiar	Spatial Extension	Emerging	0.18	0.00	0.01	20
15:00	Av. Miguel Bombarda	Spatial Extension	Simple	-0.25	0.90	-0.11	7
13:00	Av. Calouste Gulbenkian	Delay	Emerging	-0.25	0.09	-0.68	5
15:00	Av. Calouste Gulbenkian	Speed	Emerging	-0.30	0.16	0.73	8
13:00	Av. Calouste Gulbenkian	Spatial Extension	Emerging	-0.31	0.13	-0.87	5
19:00	Av. Miguel Bombarda	Spatial Extension	Emerging	-0.33	0.65	-0.26	5

Table 9.1: Top 5 congestion and decongestion traffic patterns without situational context.

Time	Location	Attribute	Pattern Type	Score	R^2	Slope	Support
08:00	R. Prof. Lima Basto	Delay	Emerging	0.38	0.74	0.21	5
14:00	R. Pinheiro Chagas	Speed	Simple	0.34	0.84	-0.26	17
12:00	R. Pinheiro Chagas	Speed	Simple	0.34	0.83	-0.24	19
14:00	Al. Card. Cerejeira	Spatial Extension	Simple	0.28	0.92	0.19	6
16:00	R. Pinheiro Chagas	Speed	Simple	0.25	0.85	-0.08	13
13:00	Av. Calouste Gulbenkian	Spatial Extension	Emerging	-0.32	0.13	-0.90	5
14:00	R. Pinheiro Chagas	Spatial Extension	Simple	-0.33	0.88	-0.16	17
12:00	R. Pinheiro Chagas	Spatial Extension	Simple	-0.37	0.89	-0.14	19
14:00	Al. Card. Cerejeira	Speed	Simple	-0.39	0.96	0.38	6
15:00	Av. Calouste Gulbenkian	Speed	Emerging	-0.40	0.33	0.94	7

Table 9.2: Top 5 congestion and decongestion traffic patterns with filtered road restrictions.

Time	Location	Attribute	Pattern Type	Score	R^2	Slope	Support
12:00	Av. Ant Aug. de Aguiar	Speed	Simple	0.37	0.74	-0.81	5
18:00	Av. Calouste Gulbenkian	Speed	Emerging	0.32	0.62	-0.14	5
18:00	R. de Campolide	Speed	Emerging	0.30	0.46	-0.29	6
17:00	Av. Miguel Torga	Delay	Emerging	0.26	0.10	0.53	5
17:00	Av. Miguel Torga	Spatial Extension	Emerging	0.25	0.11	0.51	5
18:00	R. de Campolide	Spatial Extension	Emerging	-0.24	0.42	-0.16	6
17:00	Av. Ant Aug. de Aguiar	Delay	Emerging	-0.25	0.43	-0.24	5
15:00	Av. Calouste Gulbenkian	Speed	Emerging	-0.28	0.42	0.34	5
12:00	Av. Ant Aug. de Aguiar	Spatial Extension	Emerging	-0.30	0.42	-0.39	5
17:00	Av. Ant Aug. de Aguiar	Spatial Extension	Emerging	-0.41	0.65	-0.36	5

Table 9.3: Top 5 congestion and decongestion traffic patterns for high humidity conditions (>80%).

Part IV

Conclusions

10

Concluding Remarks

Contents

10.1 Discussion 91
10.2 Future work 92
10.3 Scientific communication 92

10.1 Discussion

The present thesis proposes two distinct methods to address the problem of mining actionable patterns of road mobility from heterogeneous sources of traffic data while addressing the influences of situational context. The first method proposes the combined use of data transformations and pattern-based biclustering searches to comprehensively explore spatiotemporal associations within road traffic data. Pattern-based biclustering holds unique properties of interest making it a great candidate for mining patterns in traffic data: efficient yet exhaustive searches; non-trivial traffic patterns with parameterizable coherence; tolerance to noise and missing data; ability to incorporate domain knowledge; and sound statistical testing. To integrate situational context into the pattern mining task we developed a context-consolidation mechanism that make use of data mappings to create a mask over the road traffic data, enabling the discovery of traffic profiles that are recurrent under specific context conditions.

Results from geolocalized speed and loop counter data confirm the unique role of biclustering in finding relevant patterns given by recurrent jam profiles spanning diverse locations and time periods within the day in accordance with inputted spatial and temporal constraints. Non-constant road traffic patterns can be further pursued to guarantee a greater robustness to traffic variability while still guaranteeing the coherence of the target traffic patterns. The target traffic patterns can combine different jam-related aspects, such as speed limits, vehicle passage frequencies, and the spatial extent of congested road segments. Results evidence the ability to unveil actionable, interpretable and statistically significant patterns of road mobility, thus providing a trustworthy context with enough feedback to support mobility reforms.

The second method proposes E2PAT, a method to discover emerging patterns from heterogoeneous sensor networks in linear time. E2PAT combines spatiotemporal data mappings with simple yet effective time series differencing operations to find emerging behaviors. Differencing orders are explored to further find regular trends and emerging behaviors. E2PAT further provides statistical guarantees of pattern growth, support and accuracy, as well as visualization and navigation facilities, to safeguard the soundness and usability of the pattern analysis process. An integrative score is also proposed to measure the relevance of emerging patterns, offering a sound criterion to control the false positive and negative discovery rates. Remarkably, we show that the proposed score yields statistical properties of interest: bounded, easily interpretable, and passes normality tests for the found pattern solutions. A context-aware filtering mechanism was also developed to enable the discovery of emerging patterns sensitive to context attributes, such as road restrictions and meteorological conditions.

Results from geolocalized speed and loop counter data confirm the ability to fully retrieve all the emerging congestions, spanning diverse city regions and time periods of the day in accordance with the inputted spatial criteria and calendrical constraints. The found emerging patterns of urban mobility explore the multivariate nature of the gathered data, covering different jam-related views, such as speed limits, vehicle passage frequencies, and the spatial extent of congested road segments. Results further evidence the ability to unveil actionable, interpretable of road mobility, thus providing a trustworthy context with enough feedback to support mobility reforms.

10.2 Future work

Starting with the biclustering method, we first intend to provide spatiotemporal navigation facilities among the multiplicity of traffic patterns present within a city at a certain time, as well as more usable visual representations of each pattern. Second, we expect to extend this analysis to other modalities of transport within the city of Lisbon, and then apply the proposed approach to urban data collected from other cities. Finally, we aim to extend the proposed approach to discover patterns sensitive to situational context, to other sources of contextual data such as road restrictions and large-scale events.

For E2PAT, we first expect to extend the conducted analysis towards other sources of urban data. Second, we intend to find new pattern abstractions from emerging behaviors that are spatially and temporally related. Third we look forward to provide other pattern navigation facilities. Finally, we expect to extend E2PAT to combine other sources of situational context such as public events.

10.3 Scientific communication

The development of this thesis resulted in the creation of two articles:

- Mining actionable traffic patterns of road mobility using biclustering (under review);
- Efficient discovery of emerging patterns in heterogeneous spatiotemporal data from mobile sensors (accepted).

Bibliography

[1] C. K. Gately, L. R. Hutyra, S. Peterson, and I. S. Wing, “Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone gps data,” *Environmental pollution*, vol. 229, pp. 496–504, 2017.

[2] J. Ma, Y. Tao, M.-P. Kwan, and Y. Chai, “Assessing mobility-based real-time air pollution exposure in space and time using smart sensors and gps trajectories in beijing,” *Annals of the American Association of Geographers*, vol. 110, no. 2, pp. 434–448, 2020.

[3] J. Song, C. Zhao, S. Zhong, T. A. S. Nielsen, and A. V. Prishchepov, “Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques,” *Computers, Environment and Urban Systems*, vol. 77, p. 101364, 2019.

[4] Y. Liao, J. Gil, R. H. Pereira, S. Yeh, and V. Verendel, “Disparities in travel times between car and transit: Spatiotemporal patterns in cities,” *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[5] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, “Unsupervised real-time anomaly detection for streaming data,” *Neurocomputing*, vol. 262, pp. 134–147, 2017.

[6] P. J. Brockwell, R. A. Davis, and S. E. Fienberg, *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991.

[7] W. W. Wei, “Time series analysis,” in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, 2006.

[8] B. Y. Chen, H. Yuan, Q. Li, W. H. Lam, S.-L. Shaw, and K. Yan, “Map-matching algorithm for large-scale low-frequency floating car data,” *International Journal of Geographical Information Science*, vol. 28, no. 1, pp. 22–38, 2014.

[9] F. Chen, M. Shen, and Y. Tang, “Local path searching based map matching algorithm for floating car data,” *Procedia Environmental Sciences*, vol. 10, pp. 576 – 582, 2011, 2011 3rd International Conference on Environmental Science and Information Application Technology ESIAT 2011.

[10] M. Bierlaire, J. Chen, and J. Newman, “A probabilistic map matching method for smartphone gps data,” *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 78 – 98, 2013.

[11] P. T. Martin, Y. Feng, X. Wang *et al.*, “Detector technology evaluation,” Mountain-Plains Consortium Fargo, ND, Tech. Rep., 2003.

[12] E. Necula, “Analyzing traffic patterns on street segments based on gps data using r,” *Transportation Research Procedia*, vol. 10, pp. 276–285, 2015.

[13] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: A survey,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[14] R. Henriques, C. Antunes, and S. C. Madeira, “A structured view on pattern mining-based biclustering,” *Pattern Recognition*, vol. 4, no. 12, pp. 3941—-3958, 2015.

[15] R. Henriques and S. C. Madeira, “Bsig: evaluating the statistical significance of biclustering solutions,” *Data Mining and Knowledge Discovery*, vol. 32, no. 1, pp. 124–161, 2018.

[16] R. Henriques, F. L. Ferreira, and S. C. Madeira, “Bicpams: software for biological data analysis with pattern-based biclustering,” *BMC bioinformatics*, vol. 18, no. 1, p. 82, 2017.

[17] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, “Biclustering on expression data: A review,” *Journal of biomedical informatics*, vol. 57, pp. 163–180, 2015.

[18] V. A. Padilha and R. J. Campello, “A systematic comparative evaluation of biclustering techniques,” *BMC bioinformatics*, vol. 18, no. 1, p. 55, 2017.

[19] R. Henriques and S. Madeira, “Bicpam: Pattern-based biclustering for biomedical data analysis,” *Alg. for Molecular Biology*, vol. 9, no. 1, p. 27, 2014.

[20] G. Dong and J. Li, “Efficient mining of emerging patterns: Discovering trends and differences,” in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’99. New York, NY, USA: ACM, 1999, pp. 43–52.

[21] J. Yang, X. Zhang, Y. Qiao, Z. Fadlullah, and N. Kato, “Global and individual mobility pattern discovery based on hotspots,” in *2015 IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 5577–5582.

[22] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” *nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[23] S. Hasan, C. Schneider, S. Ukkusuri, and M. C. Gonzalez, "Spatiotemporal patterns of urban human mobility," *Journal of Statistical Physics*, vol. 151, pp. 1–15, 04 2012.

[24] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris, "Discovering spatial patterns in origin-destination mobility data," *Transactions in GIS*, vol. 16, no. 3, pp. 411–429, 2012.

[25] A. Salamanis, G. Margaritis, D. D. Kehagias, G. Matzoulas, and D. Tzovaras, "Identifying patterns under both normal and abnormal traffic conditions for short-term traffic prediction," *Transportation research procedia*, vol. 22, pp. 665–674, 2017.

[26] J.-A. Yang, M.-H. Tsou, C.-T. Jung, C. Allen, B. H. Spitzberg, J. M. Gawron, and S.-Y. Han, "Social media analytics and research testbed (smart): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages," *Big Data & Society*, vol. 3, no. 1, p. 2053951716652914, 2016.

[27] W. Sloan, "Discussion, report of committee on highway traffic analysis," in *Highway Research Board Proceedings*, vol. 7, 1928.

[28] C. Yang, K. Clarke, S. Shekhar, and C. V. Tao, "Big spatiotemporal data analytics: a research and innovation frontier," 2019.

[29] R. N. Mantegna and H. E. Stanley, "Stochastic process with ultraslow convergence to a gaussian: the truncated lévy flight," *Physical Review Letters*, vol. 73, no. 22, p. 2946, 1994.

[30] S. Ahn, B. Coifman, V. Gayah, M. Hadi, S. Hamdar, L. Leclercq, H. Mahmassani, M. Menendez, A. Skabardonis, and H. van Lint, "Traffic flow theory and characteristics," *Centennial Papers*, 2019.

[31] Z. Zheng, "Recent developments and research needs in modeling lane changing," *Transportation research part B: methodological*, vol. 60, pp. 16–32, 2014.

[32] L. Li and X. M. Chen, "Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 76, pp. 170–188, 2017.

[33] X. Wang, R. Jiang, L. Li, Y.-L. Lin, and F.-Y. Wang, "Long memory is important: A test study on deep-learning based car-following model," *Physica A: Statistical Mechanics and its Applications*, vol. 514, pp. 786–795, 2019.

[34] S. Wang, L. Li, W. Ma, and X. Chen, "Trajectory analysis for on-demand services: A survey focusing on spatial-temporal demand and supply patterns," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 74–99, 2019.

[35] L. Li, R. Jiang, Z. He, X. M. Chen, and X. Zhou, “Trajectory data-based traffic flow studies: A revisit,” *Transportation Research Part C: Emerging Technologies*, vol. 114, pp. 225–240, 2020.

[36] H. K. Lo, C. Yip, and K. Wan, “Modeling transfer and non-linear fare structure in multi-modal network,” *Transportation Research Part B: Methodological*, vol. 37, no. 2, pp. 149 – 170, 2003.

[37] A. Loder, L. Ambühl, M. Menendez, and K. W. Axhausen, “Empirics of multi-modal traffic networks – using the 3d macroscopic fundamental diagram,” *Transportation Research Part C: Emerging Technologies*, vol. 82, pp. 88 – 101, 2017.

[38] K. Abdelghany and H. Mahmassani, “Dynamic trip assignment-simulation model for intermodal transportation networks,” *Transportation Research Record Journal of the Transportation Research Board*, vol. 1771, pp. 52–60, 01 2001.

[39] F. Rempe, G. Huber, and K. Bogenberger, “Spatio-temporal congestion patterns in urban traffic networks,” *Transportation Research Procedia*, vol. 15, pp. 513 – 524, 2016, international Symposium on Enhancing Highway Performance (ISEHP), June 14-16, 2016, Berlin.

[40] F. G. Habtemichael and M. Cetin, “Short-term traffic flow rate forecasting based on identifying similar traffic patterns,” *Transportation research Part C: emerging technologies*, vol. 66, pp. 61–78, 2016.

[41] G. Atluri, A. Karpatne, and V. Kumar, “Spatio-temporal data mining: A survey of problems and methods,” *CoRR*, vol. abs/1711.04710, 2017.

[42] M. Treiber and A. Kesting, “Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns,” *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 31 – 41, 2012.

[43] Z. He, M. Deng, J. Cai, Z. Xie, Q. Guan, and C. Yang, “Mining spatiotemporal association patterns from complex geographic phenomena,” *International Journal of Geographical Information Science*, pp. 1–26, 2019.

[44] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, “Trajectory pattern mining,” 01 2007, pp. 330–339.

[45] F. Giannotti, M. Nanni, and D. Pedreschi, “Efficient mining of temporally annotated sequences,” vol. 2006, 04 2006.

[46] R. Inoue, A. Miyashita, and M. Sugita, “Mining spatio-temporal patterns of congested traffic in urban areas from traffic sensor data,” 11 2016, pp. 731–736.

[47] Z. Chen, Y. Yang, L. Huang, E. Wang, and D. Li, "Discovering urban traffic congestion propagation patterns with taxi trajectory data," *IEEE Access*, vol. 6, pp. 69 481–69 491, 2018.

[48] S. P. Latoski, W. M. Dunn, B. Wagenblast, J. Randall, M. D. Walker *et al.*, "Managing travel for planned special events," United States. Joint Program Office for Intelligent Transportation Systems, Tech. Rep., 2003.

[49] S. Kwoczek, S. Di Martino, and W. Nejdl, "Predicting and visualizing traffic congestion in the presence of planned special events," *Journal of Visual Languages & Computing*, vol. 25, no. 6, pp. 973–980, 2014.

[50] F. Rodrigues, S. S. Borysov, B. Ribeiro, and F. C. Pereira, "A bayesian additive model for understanding public transport usage in special events," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2113–2126, Nov 2017.

[51] S. Gokcan, "Forecasting volatility of emerging stock markets: linear versus non-linear garch models," *Journal of forecasting*, vol. 19, no. 6, pp. 499–504, 2000.

[52] P. H. Franses, D. Van Dijk *et al.*, *Non-linear time series models in empirical finance*. Cambridge university press, 2000.

[53] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 43–52.

[54] J. Li and L. Wong, "Emerging patterns and gene expression data," *Genome Informatics*, vol. 12, pp. 3–13, 2001.

[55] ——, "Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns," *Bioinformatics*, vol. 18, no. 5, pp. 725–734, 2002.

[56] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome informatics*, vol. 13, pp. 51–60, 2002.

[57] P. K. Novak, N. Lavrač, and G. I. Webb, "Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining." *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

[58] J. Li, G. Dong, and K. Ramamohanarao, "Making use of the most expressive jumping emerging patterns for classification," *Knowledge and Information systems*, vol. 3, no. 2, pp. 131–145, 2001.

[59] H. Fan and K. Ramamohanarao, “A bayesian approach to use emerging patterns for classification,” in *ADC*, 2003.

[60] ——, “Efficiently mining interesting emerging patterns,” in *International Conference on Web-Age Information Management*. Springer, 2003, pp. 189–201.

[61] A. Soulet, B. Crémilleux, and F. Rioult, “Condensed representation of emerging patterns,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2004, pp. 127–132.

[62] A. M. Garcıa-Vico, P. González, C. J. Carmona, and M. J. del Jesus, “A big data approach for extracting fuzzy emerging patterns,” *Cognitive Computation*, vol. 11, no. 3, pp. 400–417, 2019.

[63] A. M. Garcıa-Vicoa, F. Chartea, P. Gonzáleza, D. Elizondob, and C. J. Carmonaa, “E2pamea: A fast evolutionary algorithm for extracting fuzzy emerging patterns in big data environments,” 2020.

[64] A. M. G. Vico, C. Carmona, P. Gonzalez, H. Seker, and M. J. Del Jesus, “Fepds: A proposal for the extraction of fuzzy emerging patterns in data streams,” *IEEE Transactions on Fuzzy Systems*, 2020.

[65] H. S. Song, J. kyeong Kim, and S. H. Kim, “Mining the change of customer behavior in an internet shopping mall,” *Expert Systems with Applications*, vol. 21, no. 3, pp. 157–168, 2001.

[66] R.-C. Wu, R.-S. Chen, and C.-C. Chen, “Data mining application in customer relationship management of credit card business,” in *29th Annual International Computer Software and Applications Conference (COMPSAC’05)*, vol. 2. IEEE, 2005, pp. 39–40.

[67] G. Li, R. Law, H. Q. Vu, J. Rong, and X. R. Zhao, “Identifying emerging hotel preferences using emerging pattern mining technique,” *Tourism management*, vol. 46, pp. 311–321, 2015.

[68] A. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto, and M. J. del Jesus, “An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 1, p. e1231, 2018.

[69] R. Henriques and S. C. Madeira, “Bicnet: Flexible module discovery in large-scale biological networks using biclustering,” *Algorithms for Molecular Biology*, vol. 11, no. 1, pp. 1–30, 2016.

[70] ——, “Bic2pam: constraint-guided biclustering for biological data analysis with domain knowledge,” *Algorithms for Molecular Biology*, vol. 11, no. 1, p. 23, 2016.

[71] G. Atluri, A. Karpatne, and V. Kumar, “Spatio-temporal data mining: A survey of problems and methods,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–41, 2018.