Mining Actionable Patterns of Road Mobility From Heterogeneous Traffic Data Using Biclustering

Francisco Neves, Anna C. Finamore, Sara C. Madeira, and Rui Henriques

Abstract—**The comprehensive access to road traffic patterns in the continuously growing urban areas is key to achieve a sustainable mobility. However, the inherent complexity of urban traffic poses many challenges to achieve this goal, including: i) the need to integrate heterogeneous views of road traffic (such as speed limits, jam size, delay, throughput) from available sources; ii) the complex spatiotemporal intricacies of geolocalized speed and loop counter data; iii) the need to mine congestion patterns robust to the inherent traffic variability and unexpected occurrence of events, taking also into consideration the varying degrees of congestion severity; and iv) the need to guarantee the statistical significance and interpretability of the target patterns. In the context of our work, a road traffic pattern is a recurrent congestion profile (w.r.t. speed limits, jam extent and flow) that can span multiple locations and time periods within a day. Biclustering, the discovery of coherent subspaces (local patterns) within real-valued data, has unique properties of interest, being positioned to unravel such traffic patterns, while satisfying the aforementioned challenges. Despite its relevance, the potentialities of applying biclustering in mobility domains remain unexplored. This work proposes a structured view on why, when and how to apply biclustering for mining traffic patterns of road mobility, a subject remaining largely unexplored up to date. Using the city of Lisbon as a guiding case, we illustrate the relevance of biclustering geolocalized speed data and loop counter data. The gathered results confirm the role of biclustering in comprehensively finding statistically significant and actionable spatiotemporal associations of road mobility.**

Index Terms—**Sustainable mobility, spatiotemporal pattern mining, biclustering, road traffic data.**

I. Introduction

MOBILITY in most capital cities is not yet sustainable. The COVID-19 pandemic crisis is exposing new vulnerabilities as urban mobility patterns are rapidly changing. In particular, road mobility is susceptible to significant externalities causing daily congestions, in turn aggravating air pollution, accessibility problems, traffic noise, and safety susceptibilities [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 of road congestion is hampered by five major challenges. First, the inability of traditional pattern mining methods to handle the spatiotemporal intricacies of road traffic data sources, such as geolocalized speed data and loop counter data. Second, the need to mine patterns robust to the inherent traffic variability and sporadic occurrence of unexpected events. Third, the need to combine multiple aspects of road traffic, including speed limits, congestion size, duration, as well as frequentist views on traffic flow. Fourth, the need to discover patterns sensitive to jams with varying levels of sensitivity (flexible coherence to go beyond the focus on trivial congestions). Finally, the need to find comprehensive sets of road traffic patterns with guarantees of statistical significance, actionability and interpretability.

To address the introduced limitations, this paper proposes the combined use of spatiotemporal data transformations and biclustering to comprehensively find congestion patterns from heterogeneous sources of road traffic data. In contrast with clustering, biclustering – the discovery of subspaces within real-valued data – provides the possibility to search for traffic patterns on arbitrarily-sized geographical and temporal extents, offering local and modular views. 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 [5], [6], its potential in the mobility domain remains untapped. To the best of our knowledge, this is the first work aiming at comprehensively mining road traffic patterns with non-trivial forms of coherence.

Manuscript received 7 June 2020; revised 17 November 2020 and 26 January 2021; accepted 1 February 2021. Date of publication 18 February 2021; date of current version 8 July 2022. This work was supported in part by the *Fundação para a Ciência e Tecnologia* (FCT) under ILU project (DSAIPA/DS/0111/2018), INESC-ID (UIDB/50021/2020) and LASIGE (UIDB/00408/2020, UIDP/00408/2020) pluriannuals. The Associate Editor for this article was P. Wang. *(Corresponding author: Rui Henriques.)*

Francisco Neves and Rui Henriques are with the INESC-ID, 1000-029 Lisboa, Portugal, and also with the Instituto Superior Técnico, Universidade de Lisboa, 1649-004 Lisboa, Portugal (e-mail: rmch@tecnico.ulisboa.pt).

Anna C. Finamore is with the COPELABS, University Lusofona, 1749-024 Lisboa, Portugal, also with the the INESC-ID, 1000-029 Lisboa, Portugal, and also with the Instituto Superior Técnico, Universidade de Lisboa, 1649-004 Lisboa, Portugal.

Sara C. Madeira is with the LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1649-004 Lisboa, Portugal.

Digital Object Identifier 10.1109/TITS.2021.3057240

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) inductive loop detectors' data. WAZE data contain 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.

The paper is structured as follows. Section II provides essential background on road traffic data analysis and biclustering. Section III surveys important contributions from related work. Section IV offers a comprehensively list of road traffic patterns, and further describes why, when and how to apply biclustering for their discovery. Section V gathers and discusses results from biclustering heterogeneous sources of road traffic data within the city of Lisbon. Finally, concluding remarks and future directions are synthesized.

II. Background

A. Road Traffic Data

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 equipments are able to detect volume, speed and classify vehicles passing. ILDs are relatively susceptible to failure rates in their estimations. Martin *et al.* [7] 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. 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.

To formalize ILD data, consider a *time series* to be an ordered set of data points $\boldsymbol{x}_{it}$, where $i\in 1..m$ is the index of the variable y_i being recorded and $t\in 1..T$ the corresponding time point or time interval. Time series are referred as *univariate* when only one variable is recorded and *multivariate* when $m > 1$. Time series recorded at a particular location are referred as *georeferenced* time series. More formally, a georeferenced time series is a tuple $gt = \left(\phi, \{\boldsymbol{x}_{it}\}^{i=1..m,t=1..T}\right)$, where ϕ is a pair *(latitude, longitude)* describing the location where time series $\boldsymbol{x}$ was recorded.

Aggregated **ILD** data are a collection $\langle gt_1, gt_2, \cdots, gt_n\rangle$, where $gt_k = (\phi, \boldsymbol{x})$ 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) Geolocalized Speed Data: The use of mobile devices with active global positioning systems (GPS) is pervasive nowadays. Applications installed in some of these devices offer localization and navigation facilities, providing a comprehensive view of the ongoing traffic dynamics within the city. For instance, WAZE[1] is a free to use community-driven GPS application able to monitor world-wide traffic dynamics. In Lisbon, WAZE partnered with the city Council to provide real-time statistics on traffic jams.[2] The streaming data comes in the form of events, where an event corresponds to a significant change to the regular speed along a given road segment. To formalize WAZE jam data, the concepts of trajectory and spatial event need to be introduced.

A **trajectory** is a sequence $T = \langle\phi_1, \phi_2, \cdots, \phi_i\rangle$, where ϕ_k is a pair *(latitude,longitude)*. Trajectories are a common type of traffic data. Floating car data (FCD) are a common example of trajectory data produced from GPS devices, which gather vehicles' sequential positions. More details on FCD can be read in [8]. Methods for constructing FCD from GPS information produce rather sparse trajectories that need to be completed within the constraints of the road network mesh [9]–[11].

In the context of our work, an **event** is a tuple $E = (\mathbf{x}, s, t)$, 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 $\boldsymbol{x}$=(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;
- t is the *temporal extent* of the observation $\mathbf{x}$, either given by a time instant or a time interval.

Geolocalized speed data can thus be seen as a **collection of events** $E = \{e_1, e_2, \cdots, e_n\}$, where $e_k = (\mathbf{x}, s, t)$ is a traffic jam event that occurred at time t in a trajectory (road segment) s. The set of observations $\mathbf{x}$ contains traffic information – such as the recorded speed, delay, severity level or road type – that characterizes the occurring jam.

3) Integrating Heterogeneous Sources: Considering the introduced data structures, our work aims at finding non-trivial yet relevant road traffic patterns from:

1) georeferenced time series data from inductive loop detectors (frequentist view);
2) spatiotemporal event data made available by navigation applications (offering perspectives on speed limits, expected delays, severity levels and spatial extent of congestions); and
3) heterogeneous traffic data combining previous sources.

Understandably, these two sources of spatiotemporal data offer distinct yet complementary views on road traffic. On one hand, ILD data analysis alone is insufficient to distinguish whether low vehicle throughput is driven by a lack of

[1] https://www.waze.com

[2] Data can be freely accessed at https://emel.city-platform.com/opendata/

circulating vehicles or by traffic congestion. Geolocalized speed data can thus be used to augment frequentist views, thus supporting the characterization of low traffic flow scenarios. Similarly, the analysis of geolocalized speed data alone is insufficient to comprehensively characterize traffic dynamics, it is insensitive to the number of circulating cars and largely dependent on coherent communications from active GPS devices.

B. Road Traffic Pattern Mining

1) Patterns of Road Mobility: A traffic pattern 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:

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

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

The aforementioned patterns should satisfy certain properties of interest:

- *non-triviality* (novelty) and orientation towards *mobility* problems (congestions);
- *heterogeneity* (integrate multiple aspects of road traffic);
- *interpretability*;
- *actionability* (aid mobility decisions);
- *statistical significance* (road traffic patterns should not be spurious/occurring by chance);
- *robustness* (adequate tolerance to noise);
- guarantees of comprehensive (*complete* solutions) and efficient pattern retrieval.

2) Target Problem: Given the introduced sources of road traffic data (section 2.1), as well as desirable patterns of road mobility (section 2.2), the problem targeted in this work is to comprehensively discover road mobility patterns in an efficient and effective way.

C. 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*.

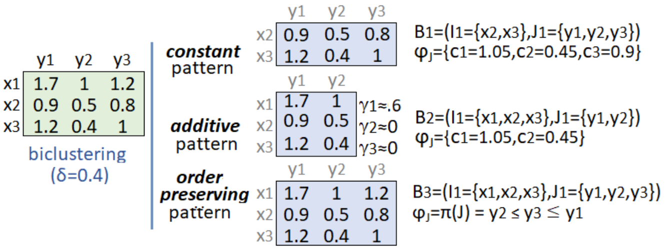


Fig. 1. Biclustering with varying homogeneity criteria: three biclusters were found under a constant, additive and order-preserving assumption. Illustrating, constant bicluster has pattern (value expectations) $\{c_1 = 1.05, c_2 = 0.45, c_3 = 0.9\}$ on x_2 and x_3 observations, while the order-preserving bicluster satisfies the $y_1 \geq y_2 \geq y_3$ permutation on $\{x_1, x_2, x_3\}$ observations.

Homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster [6]. 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 [12]. 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 1.

Given a dataset, the elements within a bicluster $a_{ij} \in (I, J)$ have coherence across variables (*pattern on observations*) if $a_{ij} = c_j + \gamma_i + \eta_{ij}$, 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 a_{ij}.

A bicluster has **constant coherence** when $\gamma_i = 0$ (or $\gamma_j = 0$), 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 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 1. Their patterns are respectively given by $\varphi_{B_1} = \{c_1 = 1.05, c_2 = 0.45, c_3 = 0.9\}$, $\varphi_{B_2} = \{c_1 = 1.05, c_2 = 0.45\}$ (assuming

$a_{ij} = c_j + \gamma_i$ and additive factors $\gamma_1 = 0.65, \gamma_2 = 0$ and $\gamma_3 = 0$) and $\varphi_{B_3} = (y_2 \leq y_3 \leq y_1)$.

Statistical significance criteria, in addition to homogeneity, guarantee that the probability of a bicluster's occurrence (against a null data model) deviates from expectations [13].

Finally, ***dissimilarity*** criteria can be further placed to guarantee the comprehensive discovery of non-redundant biclusters [5].

Following Madeira and Oliveira's taxonomy [6], 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 [14], [15].

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 [12]. 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 [16]. This behavior explains why this class of biclustering algorithms is receiving an increasing attention in recent years [12]. BicPAMS (Biclustering based on PAttern Mining Software) consistently combines these state-of-the-art contributions on pattern-based biclustering [5].

III. Related Work

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.* [17] 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 [17]–[19], 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 [20]; aggregate event statistics from free GPS systems such as Google Maps and WAZE [21]; trajectories from GPS-equipped public bicycles and taxis [22]; and traffic data from ILD and cameras found along the arteries of major cities. Understanding the patterns of human motion, both globally and individually, is crucial for different purposes, among them urban planning [17], traffic forecasting [23], providing notifications or choices to the travelers [24], and monitoring epidemic traffic responses to events and disasters [25].

Although interest in mobility patterns dates back one century [26], their automated discovery is considered a recent research area [27]. Below, we group recent contributions on this field along three major categories: classic/statistical approaches (section III-A), clustering-based approaches (section III-B) and pattern-centric approaches (section III-C) for understanding urban mobility patterns.

A. Classic Approaches to Traffic Data Analysis

Classic approaches make use of statistics, parametric models and visualization principles to understand spatiotemporal traffic dynamics. In contrast with pattern-centric stances on traffic, generative stances describe traffic flow dynamics with parametric models, offering the possibility to predict relevant state variables of a transportation system. Data-assimilation techniques are commonly combined to relate sensor observations to the system state. In this context, Yuan *et al.* [28] propose an extended Kalman filter able to combine ILD and floating car data, showing that discrete Lagrangian kinematics of traffic dynamics are preferred over the Eulerian counterpart.

Treiber *et al.* [29] introduce interpolation principles for estimating spatiotemporal distributions of traffic flow, speed and density. The approach is applied over ILD data and generalized to be enriched in the presence of floating car data or other traffic information. The relevance of this work for the targeted pattern discovery resides on the possibility to leverage data quality by handling sources of noise, including sensor failures and miscalibration of loop detectors.

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.* [19] analyzed trajectories of 10,000 mobile phone users for a six month period. Inspired by the work of Mantegna and Stanley [30], 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 [18], [31]–[35]. Li *et al.* [36] 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.* [22] 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.* [20] 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 [37]–[39] 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.* [23] 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. Rodrigues *et al.* [40] 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.

B. Clustering-Based Approaches to 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.* [21] 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.* [41] 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.* [42] 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.

C. Pattern-Centric Approaches to Traffic Data Analysis

Gowtham *et al.* [43] conducted a survey on spatiotemporal pattern mining algorithms. Some of these principles are further instantiated by Xiao *et al.* [44] for traffic pattern mining in maritime traffic service networks. 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 [45], 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.* [46] 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.

Huang *et al.* [47] proposed an architecture for traffic flow description and prediction consisting of two components: a deep belief network (DBN) and a multitask regression layer. The DBN is employed unsupervisedly and shown to be effective in extracting traffic features that support predictive tasks.

Naveh and Kim [48] propose the use of tensor factorization to extract spatiotemporal movement patterns from large-scale urban trajectory data obtained from public transport smart card systems and roadside Bluetooth detectors. To this end, traffic data is represented as a dynamic graph capturing region-to-region flow interactions across time-of-day and day-of-week.

Giannotti *et al.* [49] propose approaches to find trajectory patterns (T-Patterns) – location precedences with frequent time constraints among trajectory instances – such as,

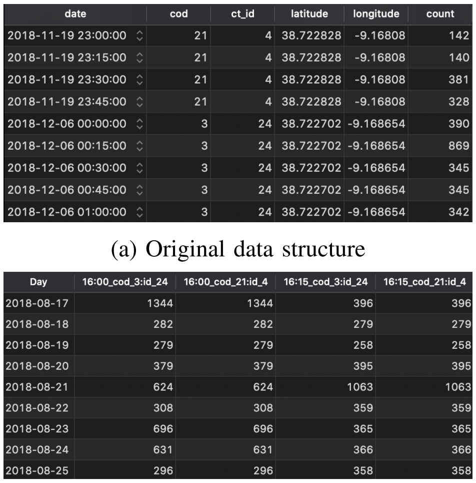
$$railway\ station \xrightarrow{15\ min} town\ square \xrightarrow{2h15\ min} museum.$$

To this end, the authors propose temporally-annotated sequence mining approaches using a density-based spatial discretization of trajectory data [50]. Inoue *et al.* [51] 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.* [52] 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.* [51], 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.* [17] 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 the relevance of the surveyed approaches, they are generally hampered by discretization needs (e.g. classic pattern mining algorithms), loose forms of coherence (e.g. tensor decomposition), and either unable to handle event data or to provide integrative views from heterogeneous sources of road traffic data.

IV. Solution

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 II-B), 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



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 data structure

Day	16:00_cod_3:id_24	16:00_cod_21:id_4	16:15_cod_3:id_24	16:15_cod_21:id_4
2018-08-17	1344	1344	396	396
2018-08-18	282	282	279	279
2018-08-19	279	279	258	258
2018-08-20	379	379	395	395
2018-08-21	624	624	1063	1063
2018-08-22	308	308	359	359
2018-08-23	696	696	365	365
2018-08-24	631	631	366	366
2018-08-25	296	296	358	358

(b) Transformed data structure

Fig. 2. ILD data mapping.

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

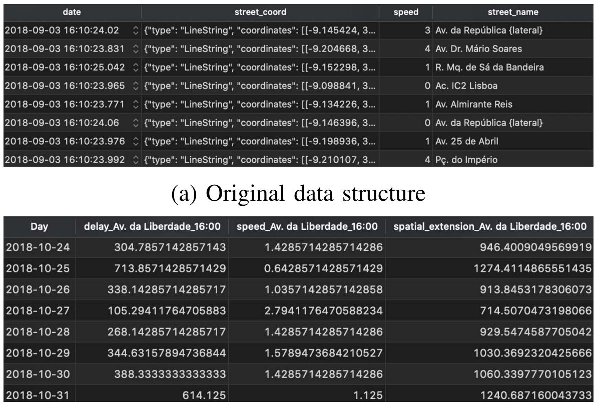
A. Road Traffic Patterns Using Biclustering

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. As default, the discovery process considers all available geographies and dates as part of the search space, and uses calendars: day-specific (e.g. all Mondays excluding holidays) and weekday calendars along academic and off-academic periods (similarly to [53]).

In addition, the time granularity (e.g. minute, hour or on/off-peak intervals) can be optionally specified to guide road traffic data aggregation. This choice is dependent on the targeted end: fine granularities are suggested for real-time notifications from online pattern analysis over traffic data streams, while coarser granularities (e.g. 15 minute) suggested for mobility planning or long-term traffic forecasting. In its absence, according to principles proposed in [53], the proposed pattern discovery is iteratively performed at different 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 2 shows the original structure of the ILD data and the corresponding data mapping.



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

(a) Original data structure

Day	delay_Av. da Liberdade_16:00	speed_Av. da Liberdade_16:00	spatial_extension_Av. da Liberdade_16:00
2018-10-24	304.7857142857143	1.4285714285714286	946.4009049569919
2018-10-25	713.8571428571429	0.6428571428571429	1274.4114865551435
2018-10-26	338.14285714285717	1.0357142857142858	913.8453178306073
2018-10-27	105.29411764705883	2.7941176470588234	714.5070473198066
2018-10-28	268.14285714285717	1.4285714285714286	929.5474587705042
2018-10-29	344.63157894736844	1.5789473684210527	1030.3692320425666
2018-10-30	388.3333333333333	1.4285714285714286	1060.3397770105123
2018-10-31	614.125	1.125	1240.687160043733

(b) Transformed data structure

Fig. 3. WAZE data mapping.

For the geolocalized speed data (WAZE data), multiple measurements are taken per event, and events are associated with a specific road trajectory. Here the columns correspond to a measurement on a single road for a specific time interval of the day. Figure 3 shows the original structure of WAZE data and the corresponding transformed data.

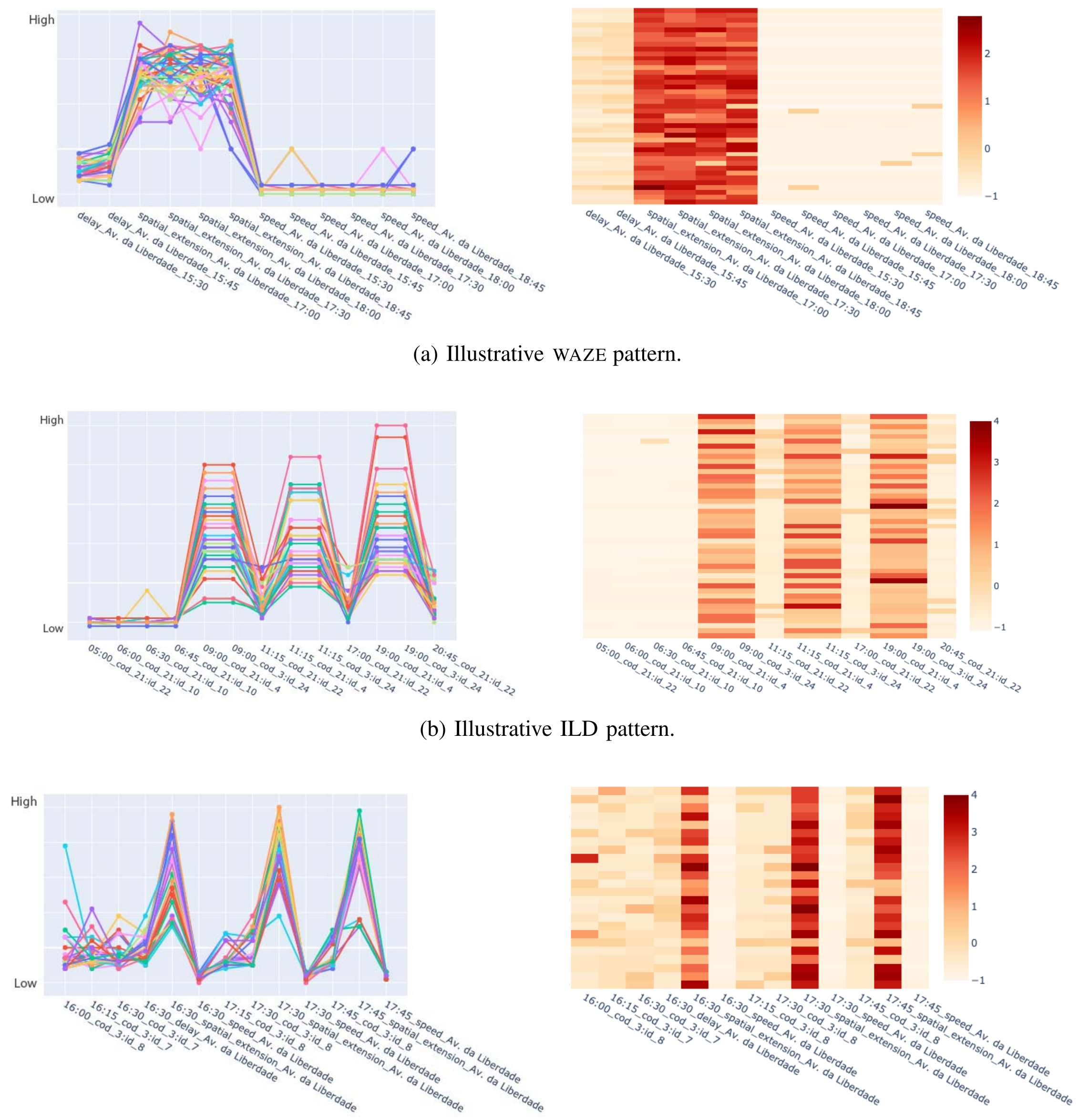
The integration of the previous mappings is a simple concatenation of the variables resulting from the transformation of each road traffic data source.

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) that can span diverse locations and different time periods.

As surveyed in the previous section, 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 provides an illustration of spatiotemporal traffic patterns given by the target biclusters using BicPAMS [5]. 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.



(a) Illustrative WAZE pattern.

(b) Illustrative ILD pattern.

(c) Illustrative integrative pattern.

Fig. 4. 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).

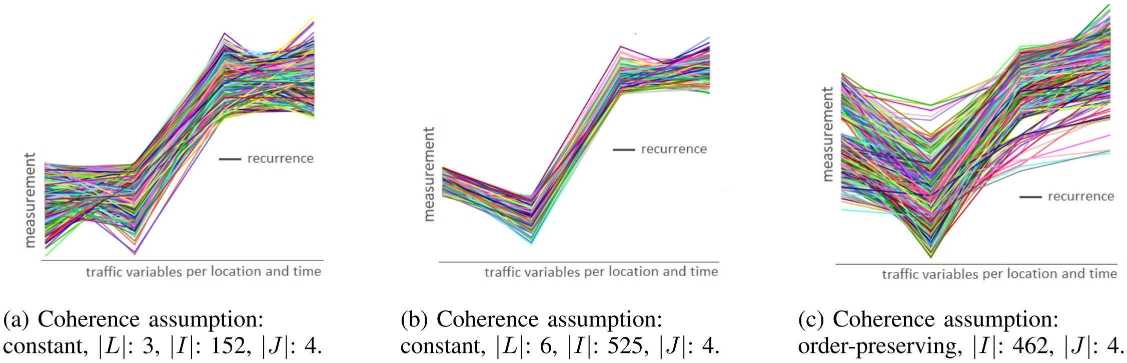
Coherence Strength: Biclustering also allows the calibration of *coherence strength* (section II.C) – e.g. how much speed limits (or car flow) need to differ to be considered dissimilar.

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 5a-b illustrates the impact that different coherence strength criteria can have on the found patterns. Considering $\delta = \frac{\bar{A}}{|L|} = 3$ (section II-C), 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.

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 II-C) 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 provides illustrative constant patterns of road traffic.



(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.

Fig. 5. Effects of coherence strength and assumption on the resulting traffic patterns.

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 5c). 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 $a_{i2} \leq a_{i3} \leq a_{i7}$.

Non-constant mobility patterns are a superset of constant mobility patterns. Their search is suggested for the discovery of road mobility patterns able to better tolerate the inherent traffic variability due to unexpected occurrences and situational context (e.g. road interventions, cultural and sport events). Although a comprehensive discovery of road mobility patterns can be pursued using the non-constant coherent assumption alone, constant road mobility patterns provide a simpler interpretation as they offer value expectations (e.g. speed limit in [22km/h,29km/h]).

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 are 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.

In fact, since the proposal of BicNET [54], 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 and 5.

B. On Why 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[3] 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

[3] The possibility to allow deviations from value expectations (under limits defined by the placed coherence strength) together with multi-item assignments [16] are placed to prevent discretization problems from occurring

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 [5]. 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 [16], i.e. there is no need to place expectations on the minimum 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 [5] 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 [13] 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.

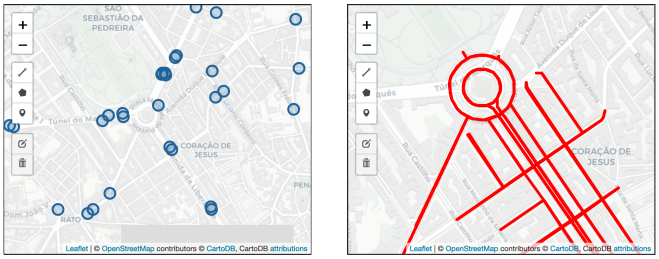
On HOW: Robustness to Noise: Pattern-based biclustering can find biclusters with a parameterizable tolerance to noise [16]. Illustrating, a quality of 80% indicates that an upper limit given by 20% of entries within a bicluster may deviate from the target jam profile ($\eta_{ij} \notin [-\delta/2, \delta/2]$). 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) [54];
2) incorporation of domain knowledge to guide the task in the presence of background metadata [55];
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 [12].

V. Results

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 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. Finally, 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.


São Sebastião da Pedreira
Túnel do Marquês
Coração de Jesus
Rato
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions

(a) ILD locations (b) WAZE events

Fig. 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).

Experimental Setting: BicPAMS [5] 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.

Both ILD and WAZE data sources are subjected to different forms of noise, which were carefully profiled before conducting the undertaken study. Inductive loop detectors that were statistically found to be poorly calibrated were removed.[4] WAZE events were found to be structurally sparser than initially expected, thus potentially missing less severe jams due to an hypothesized lack of coherent GPS communications.

Finally, location-based distributions of speed, extent and frequency were approximated, and the statistical tests proposed in BSig [13] applied to compute each pattern's statistical significance.

A. ILD Traffic Patterns

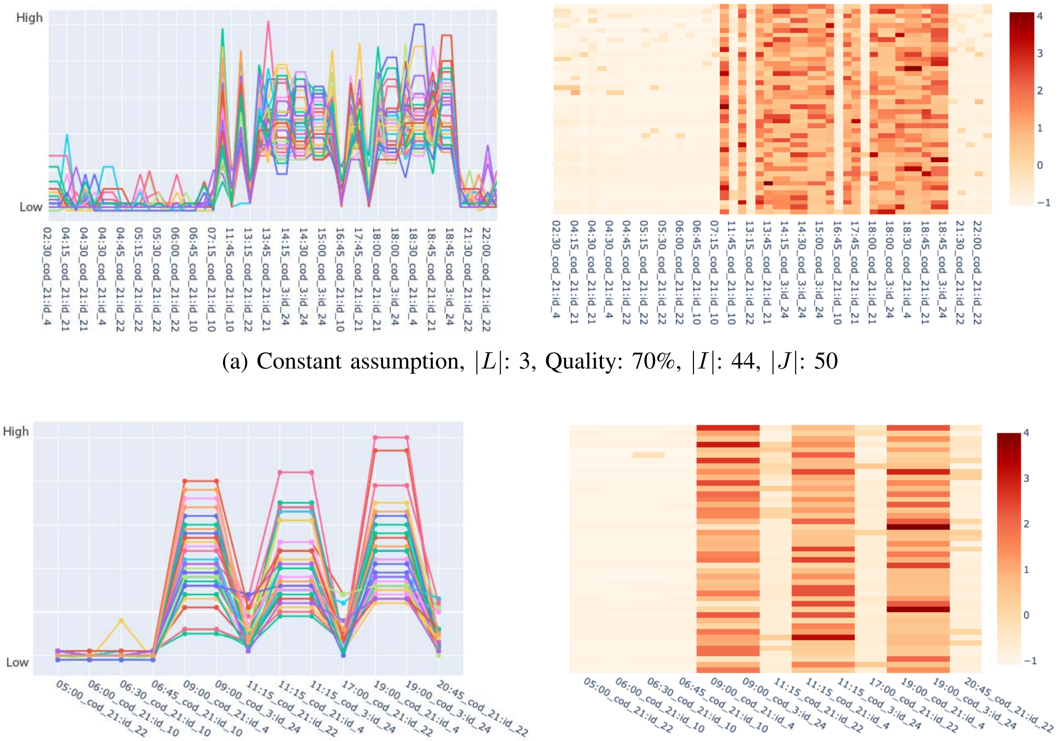
Two months of observations produced from loop detectors placed at major junctures of the city were collected (Figure 6a). Table I synthesizes the results produced by biclustering ILD data with BicPAMS [5].

Confirming the potentialities listed in Section IV, 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

[4] https://web.ist.utl.pt/rmch/ilu/

TABLE I
PROPERTIES OF BICLUSTERING SOLUTIONS IN ILD DATA USING BICPAMS WITH VARYING HOMOGENEITY CRITERIA

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



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

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

Fig. 7. Illustrative constant and order-preserving traffic patterns found in ILD data.

TABLE II
PROPERTIES OF BICLUSTERING SOLUTIONS IN WAZE DATA USING BICPAMS WITH VARYING HOMOGENEITY CRITERIA

Query	Assumption	$\|L\|$	quality	#bics	$\mu(\|I\|)$ $\pm\sigma(\|I\|)$	$\mu(\|J\|)$ $\pm\sigma(\|J\|)$	$p-$value <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

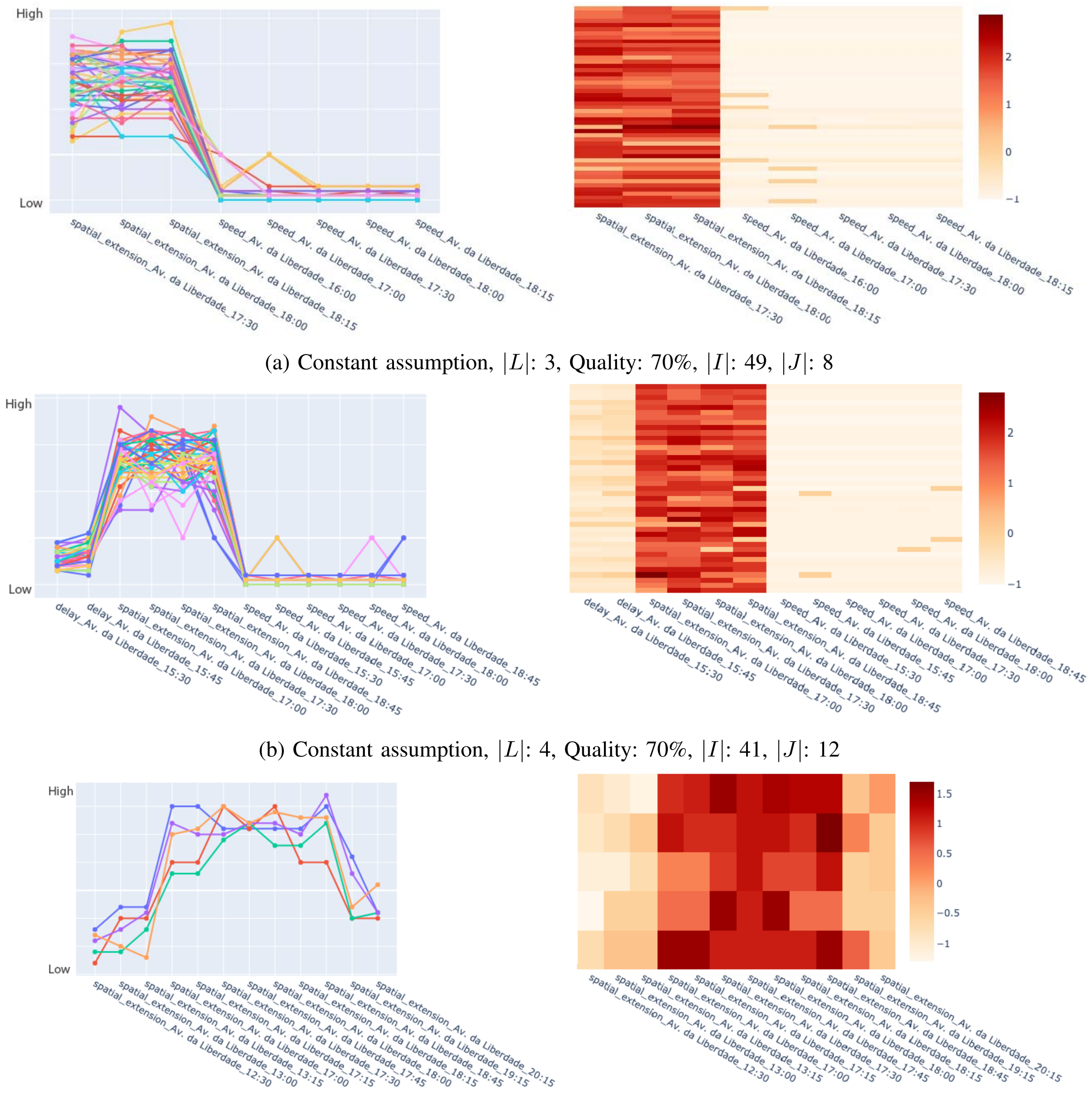
of a day) and occur on $\mu(|I|) = 43$ days within a two month period (60 days). These initial results further show the impact of tolerating noise, placing different coherence assumptions (such as the order-preserving assumption) and parameterizing coherence strength ($\delta \propto \frac{1}{|\mathcal{L}|}$) on the biclustering solution.

Figure 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 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 II-A). 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.

B. WAZE *Traffic Patterns*

WAZE events associated with jam problems at the Marquês de Pombal area within Lisbon were collected for two months



(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

Fig. 8. Three illustrative constant patterns of road traffic found in WAZE data.

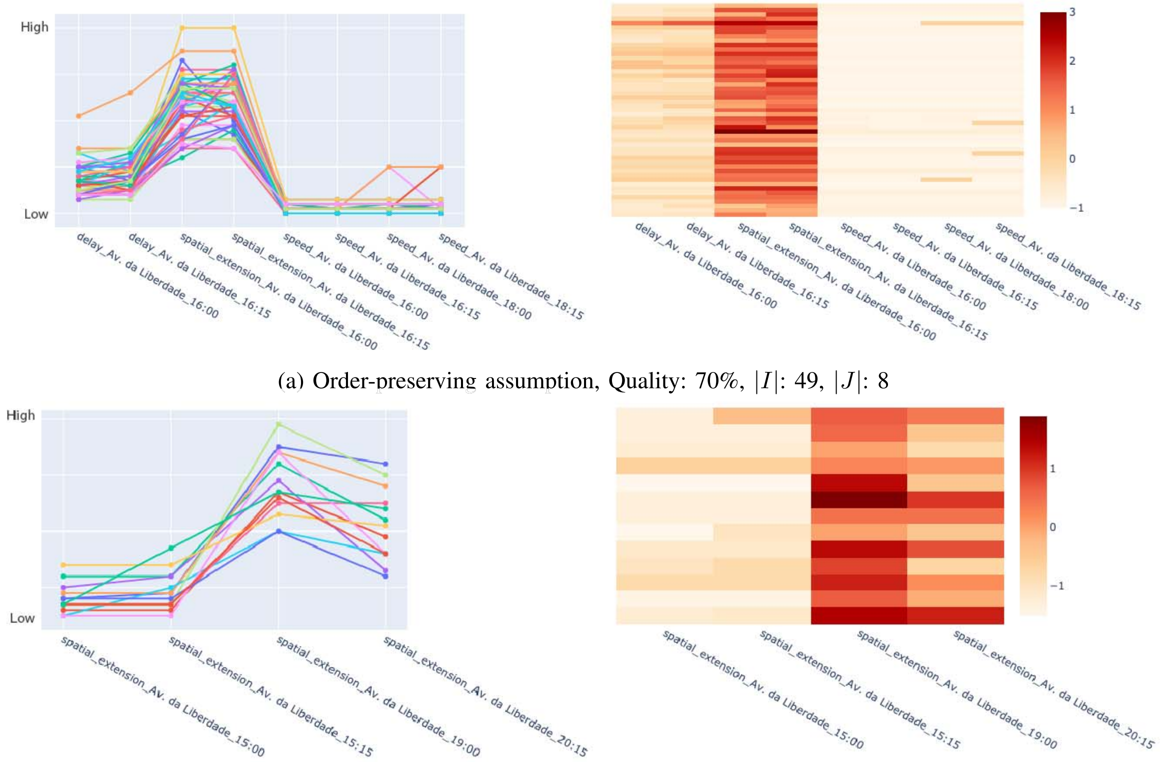
(Figure 6b). Table II synthesizes the biclustering results produced by the application of BicPAMS over WAZE data. Similarly to ILD, we observe an inherent ability of biclustering to efficiently retrieve a large number of robust, dissimilar and statistically significant patterns of road traffic. These 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 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 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 problem: slight



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

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

Fig. 9. Three illustrative order-preserving patterns of road traffic found in WAZE data.

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 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 9a and b), order-preserving patterns are characterized by a well-established permutation on the features associated with a congestion.

As introduced (section III-A), 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.

C. Integrative Patterns of Road Traffic

Finally, we briefly show integrative traffic patterns from the consolidation of ILD and WAZE data sources. Table III 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 $\delta = \boldsymbol{A}/|L|$ can be customized to comprehensively model relations with slight-to-moderate deviations from traffic pattern expectations.

Figure 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 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.

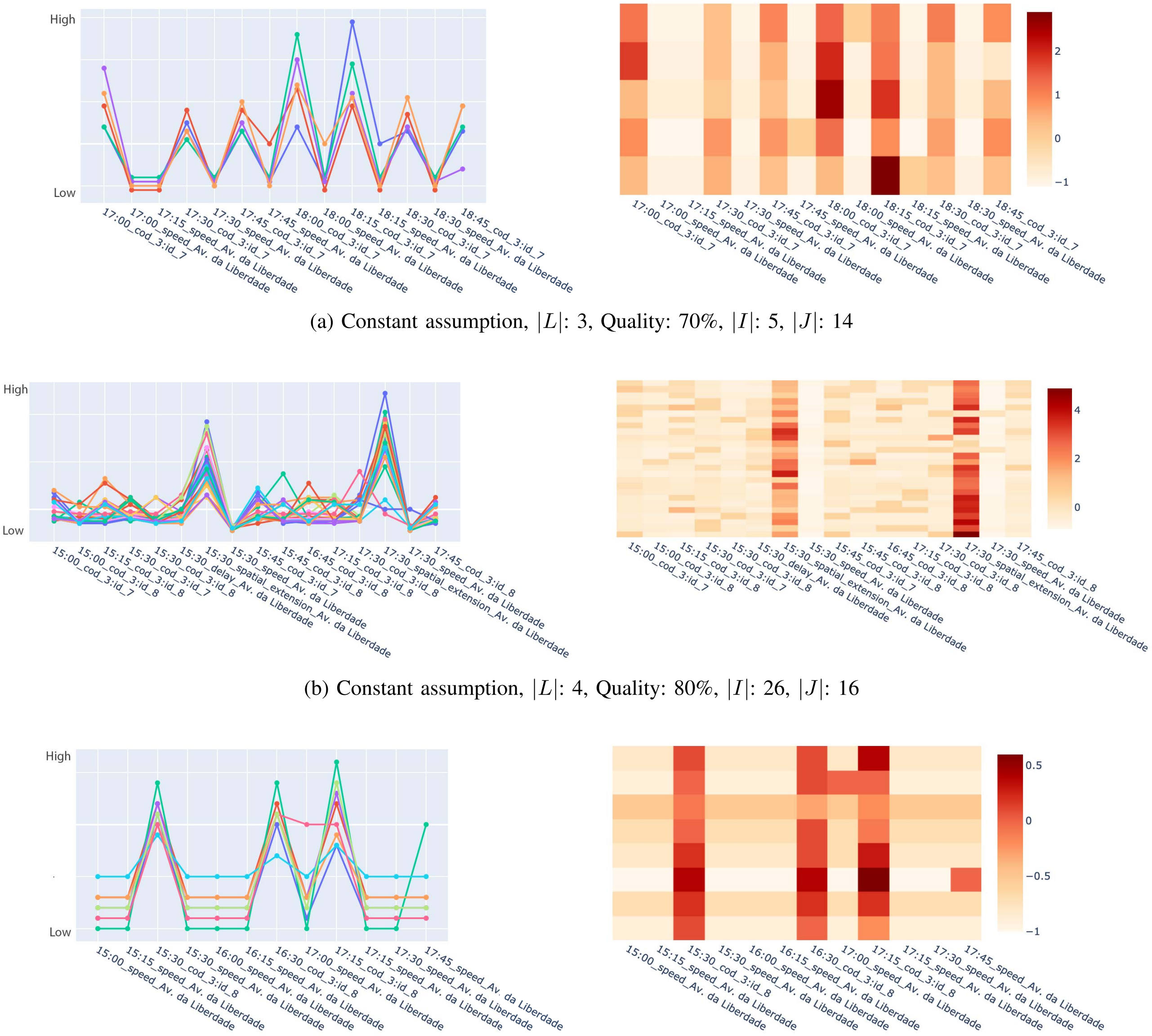
D. Statistical Significance

Table I 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

TABLE III

BICLUSTERING RESULTS FROM CONSOLIDATED ILD AND WAZE DATA USING BICPAMS WITH DIFFERENT HOMOGENEITY CRITERIA

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



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

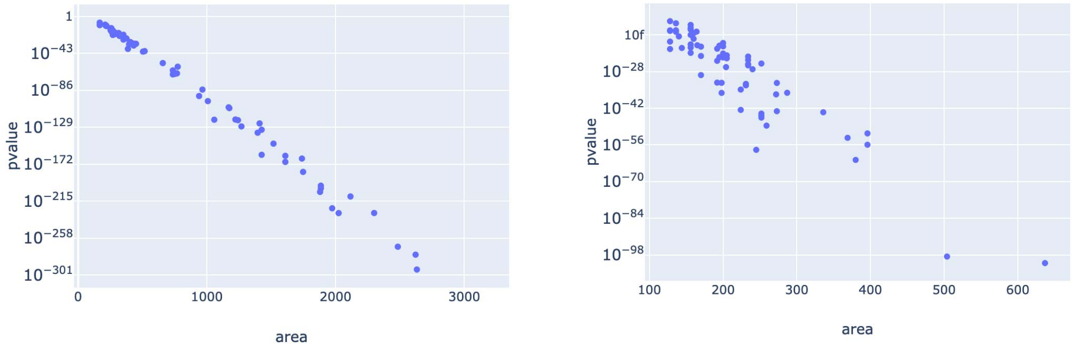
(b) Constant assumption, $|L|$: 4, Quality: 80%, $|I|$: 26, $|J|$: 16

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

Fig. 10. Illustrative mobility patterns found from heterogeneous traffic data (event and time series traffid data), integrating views on traffic flow, speed and jam extension.

a given congestion profile is unexpectedly low [13]. Figure 11 provides a scatter plot of the statistical significance (horizontal axis) and area $|I|\text{x}|J|$ (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 (top left dots) and can therefore be discarded not to incorrectly bias mobility decisions. Two major observations explain the different levels of statistical significance for the retrieved mobility patterns from ILD and WAZE data sources. First, the structurally



(a) Traffic patterns from ILD data.

(b) Traffic patterns from WAZE data.

Fig. 11. Statistical significance versus size of the collected constant patterns of road traffic ($|L|$=3 and 70% of quality).

sparser nature of WAZE data when compared with ILD data (georeferenced time series) leads to road mobility patterns with a generally lower number of supporting days (smaller patterns), hampering their statistical significance. Hence, the lower significance levels and absence of patterns on the bottom right part of Figure 11b. Second, not all road mobility patterns found from ILD data correspond to congestion profiles. Instead, they correspond to expectations well-detected by the null model and, therefore, generally have lower levels of statistical significance (upper left marks in Figure 11a).

VI. Concluding Remarks

This work addresses the problem of mining actionable patterns of road mobility from heterogeneous sources of traffic data. To this end, it proposes the combined use of data transformations and pattern-based biclustering searches to comprehensively explore spatiotemporal associations within road traffic data. Pattern-based biclustering searches are suggested to this end as they hold unique properties of interest: 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.

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.

Future Work: As future work, 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. Third, an incremental version of the proposed pattern discovery process can be considered by parameterizing BicPAMS [5] with FCFPIM [56] to provide strict real-time guarantees in the presence of road traffic streaming data. Finally, we aim to extend the proposed approach to discover patterns sensitive to sources of situational context, including weather records, interdictions, and large-scale events.

Acknowledgment

The authors thank *Câmara Municipal de Lisboa* for data provision, support and valuable feedback, particularly *Gabinete de Mobilidade* and *Centro de Operações Integrado*.

References

[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," *Environ. Pollut.*, vol. 229, pp. 496–504, Oct. 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," *Ann. Amer. Assoc. Geographers*, vol. 110, no. 2, pp. 434–448, Mar. 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," *Comput., Environ. Urban Syst.*, vol. 77, Sep. 2019, Art. no. 101364.

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

[5] R. Henriques, F. L. Ferreira, and S. C. Madeira, "BicPAMS: Software for biological data analysis with pattern-based biclustering," *BMC Bioinf.*, vol. 18, no. 1, p. 82, Dec. 2017.

[6] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 1, no. 1, pp. 24–45, Jan. 2004.

[7] P. T. Martin *et al.*, "Detector technology evaluation," Mountain-Plains Consortium Fargo, Fargo, ND, USA, Tech. Rep., 2003. [Online]. Available: https://trid.trb.org/view/740400

[8] R.-P. Schäfer, K.-U. Thiessenhusen, E. Brockfeld, and P. Wagner, "A traffic information system by means of real-time floating-car data," in *Proc. 9th World Congr. Intell. Transp. Syst.*, Chicage, IL, USA, vol. 11, Jan. 2002. [Online]. Available: https://trid.trb.org/view/662128

[9] B. Y. Chen, H. Yuan, Q. Li, W. H. K. Lam, S.-L. Shaw, and K. Yan, "Map-matching algorithm for large-scale low-frequency floating car data," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 1, pp. 22–38, Jan. 2014.

[10] F. Chen, M. Shen, and Y. Tang, "Local path searching based map matching algorithm for floating car data," *Procedia Environ. Sci.*, vol. 10, pp. 576–582, Jan. 2011.

[11] M. Bierlaire, J. Chen, and J. Newman, "A probabilistic map matching method for smartphone GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 78–98, Jan. 2013.

[12] R. Henriques, C. Antunes, and S. C. Madeira, "A structured view on pattern mining-based biclustering," *Pattern Recognit.*, vol. 4, no. 12, pp. 3941–3958, 2015.

[13] R. Henriques and S. C. Madeira, "BSig: Evaluating the statistical significance of biclustering solutions," *Data Mining Knowl. Discovery*, vol. 32, no. 1, pp. 124–161, Jan. 2018.

[14] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *J. Biomed. Informat.*, vol. 57, pp. 163–180, Oct. 2015.

[15] V. A. Padilha and R. J. G. B. Campello, "A systematic comparative evaluation of biclustering techniques," *BMC Bioinf.*, vol. 18, no. 1, p. 55, Dec. 2017.

[16] R. Henriques and S. C. Madeira, "BicPAM: Pattern-based biclustering for biomedical data analysis," *Algorithms Mol. Biol.*, vol. 9, no. 1, p. 27, Dec. 2014.

[17] J. Yang, X. Zhang, Y. Qiao, Z. Fadlullah, and N. Kato, "Global and individual mobility pattern discovery based on hotspots," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 5577–5582.

[18] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Mar. 2011.

[19] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[20] S. Hasan, C. Schneider, S. Ukkusuri, and M. C. Gonzalez, "Spatiotemporal patterns of urban human mobility," *J. Stat. Phys.*, vol. 151, pp. 1–15, Apr. 2012.

[21] E. Necula, "Analyzing traffic patterns on street segments based on GPS data using R," *Transp. Res. Procedia*, vol. 10, pp. 276–285, Jan. 2015.

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

[23] 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," *Transp. Res. Procedia*, vol. 22, pp. 665–674, Jan. 2017.

[24] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[25] J.-A. Yang *et al.*, "Social media analytics and research testbed (smart): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages," *Big Data Soc.*, vol. 3, no. 1, 2016, Art. no. 2053951716652914.

[26] W. Sloan, "Discussion, report of committee on highway traffic analysis," in *Proc. 7th Annu. Meeting Highway Res. Board, Part I*, vol. 7, 1928, pp. 259–268. [Online]. Available: https://trid.trb.org/view/120813

[27] C. Yang, K. Clarke, S. Shekhar, and C. V. Tao, "Big spatiotemporal data analytics: A research and innovation frontier," *Int. J. Geographical Inf. Sci.*, vol. 34, no. 6, pp. 1075–1088, 2019.

[28] Y. Yuan, H. Van Lint, F. Van Wageningen-Kessels, and S. Hoogendoorn, "Network-wide traffic state estimation using loop detector and floating car data," *J. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 41–50, Jan. 2014.

[29] M. Treiber, A. Kesting, and R. E. Wilson, "Reconstructing the traffic state by fusion of heterogeneous data," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 26, no. 6, pp. 408–419, Aug. 2011.

[30] R. N. Mantegna and H. E. Stanley, "Stochastic process with ultraslow convergence to a Gaussian: The truncated Lévy flight," *Phys. Rev. Lett.*, vol. 73, no. 22, p. 2946, 1994.

[31] L. Li and X. Chen, "Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 76, pp. 170–188, 2017, doi: 10.1016/j.trc.2017.01.007.

[32] Z. Zheng, "Recent developments and research needs in modeling lane changing," *Transp. Res. B, Methodol.*, vol. 60, pp. 16–32, Feb. 2014.

[33] L. Li and X. Chen, "Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 76, pp. 170–188, Mar. 2017.

[34] 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," *Phys. A, Stat. Mech. Appl.*, vol. 514, pp. 786–795, Jan. 2019.

[35] 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," *Transp. Res. C, Emerg. Technol.*, vol. 108, pp. 74–99, Nov. 2019.

[36] L. Li, R. Jiang, Z. He, X. Chen, and X. Zhou, "Trajectory data-based traffic flow studies: A revisit," *Transp. Res. C, Emerg. Technol.*, vol. 114, pp. 225–240, May 2020.

[37] H. K. Lo, C. W. Yip, and K. H. Wan, "Modeling transfer and non-linear fare structure in multi-modal network," *Transp. Res. B, Methodol.*, vol. 37, no. 2, pp. 149–170, Feb. 2003.

[38] A. Loder, L. Ambühl, M. Menendez, and K. W. Axhausen, "Empirics of multi-modal traffic networks—Using the 3D macroscopic fundamental diagram," *Transp. Res. C, Emerg. Technol.*, vol. 82, pp. 88–101, Sep. 2017.

[39] K. F. Abdelghany and H. S. Mahmassani, "Dynamic trip assignment-simulation model for intermodal transportation networks," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1771, no. 1, pp. 52–60, Jan. 2001.

[40] F. Rodrigues, S. S. Borysov, B. Ribeiro, and F. C. Pereira, "A Bayesian additive model for understanding public transport usage in special events," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2113–2126, Nov. 2017.

[41] F. Rempe, G. Huber, and K. Bogenberger, "Spatio-temporal congestion patterns in urban traffic networks," *Transp. Res. Procedia*, vol. 15, pp. 513–524, Jan. 2016.

[42] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 61–78, May 2016.

[43] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Comput. Surv.*, vol. 51, no. 4, Sep. 2018, Art. no. 83.

[44] Z. Xiao, X. Fu, L. Zhang, and R. S. M. Goh, "Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1796–1825, May 2020.

[45] M. Treiber and A. Kesting, "Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 21, no. 1, pp. 31–41, Apr. 2012.

[46] Z. He, M. Deng, J. Cai, Z. Xie, Q. Guan, and C. Yang, "Mining spatiotemporal association patterns from complex geographic phenomena," *Int. J. Geographical Inf. Sci.*, vol. 34, no. 6, pp. 1–26, 2019.

[47] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2191–2201, Oct. 2014.

[48] K. S. Naveh and J. Kim, "Urban trajectory analytics: Day-of-week movement pattern mining using tensor factorization," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 7, pp. 2540–2549, Jul. 2019.

[49] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 330–339.

[50] F. Giannotti, M. Nanni, and D. Pedreschi, "Efficient mining of temporally annotated sequences," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2006, pp. 348–359.

[51] R. Inoue, A. Miyashita, and M. Sugita, "Mining spatio-temporal patterns of congested traffic in urban areas from traffic sensor data," in *Proc. 19th Int. Conf. Intell. Transp. Syst.*, Nov. 2016, pp. 731–736.

[52] 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. 69481–69491, 2018.

[53] F. Neves, A. Finamore, and R. Henriques, "Efficient discovery of emerging patterns in heterogeneous spatiotemporal data from mobile sensors," in *Proc. 17th EAI Int. Conf. Mobile Ubiquitous Syst., Comput., Netw. Services (MobiQuitous)*. New York, NY, USA: Association for Computing Machinery, 2020.

[54] R. Henriques and S. C. Madeira, "BicNET: Flexible module discovery in large-scale biological networks using biclustering," *Algorithms Mol. Biol.*, vol. 11, no. 1, pp. 1–30, Dec. 2016.

[55] R. Henriques and S. C. Madeira, "BiC2PAM: Constraint-guided biclustering for biological data analysis with domain knowledge," *Algorithms Mol. Biol.*, vol. 11, no. 1, p. 23, Dec. 2016.

[56] J. Sun, Y. Xun, J. Zhang, and J. Li, "Incremental frequent itemsets mining with FCFP tree," *IEEE Access*, vol. 7, pp. 136511–136524, 2019.

Francisco Neves received the M.Sc. degree in information systems and computer engineering from Instituto Superior Técnico. He has authored scientific contributions on the ILU project, and a R&D project aiming at integratively mining heterogeneous sources of urban traffic data to guide the city of Lisbon becoming a world reference in sustainable mobility.

Anna C. Finamore received the Ph.D. degree in statistics and stochastic processes from the Instituto Superior Técnico, Universidade de Lisboa, Portugal. She is currently an Assistant Professor with University Lusofona and a Post-Doctoral Researcher with COPELABS and INESC-ID. She lectures courses on data engineering and intelligent systems. She develops research in applied statistics, and her research interests span the areas of pattern mining and generative learning, with applications in the education domain and smart cities.

Sara C. Madeira is currently an Associate Professor with the Department of Informatics of the Faculty of Sciences, University of Lisbon (FCUL), where she lectures courses on machine learning and data science. She is also a Senior Researcher at FCiencias-ID/LASIGE, where she coordinates the Data and Systems Intelligence Research Line of Excellence (RLE), and a member of the Health and Biomedical Informatics RLE. Since 2018, she has been coordinating the Data Science Graduation at FCUL. Her major research topics include biclustering and triclustering algorithms, together with their applications.

Rui Henriques graduated and received the Ph.D. degree from Instituto Superior Técnico, in 2008 and 2016, respectively. He has wide exposure to projects in the transportation and healthcare sectors as a Business Analyst at McKinsey and as the Head of the AI area at Tekever. He has been an Associate Researcher at INESC-ID since 2018, where he researches on how to learn from high-dimensional, heterogeneous, and temporal data. He is an Assistant Professor with the Department of Computer Science and Engineering, Instituto Superior Técnico, where he lectures courses in the domains of data science, artificial intelligence, and biomedical engineering. He is currently the PI of the large-scale ILU project for the context-senstive analysis of urban data in Lisbon. In the mobility domain, his research focuses on context-aware and integrative mining of heterogeneous traffic data.