Exploring multimodal mobility patterns with big data in the city of Lisbon

Carlos Lemonde
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa & LNEC
Elisabete Arsénio
Laboratório Nacional de Engenharia Civil (LNEC), Department of Transport
Rui Henriques
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa

ABSTRACT
Most European cities such as Lisbon in Portugal are establishing efforts to collect urban traffic data and their situational context for gaining more comprehensive views of the ongoing mobility changes and support decisions accordingly. Hence, cities are becoming sensorized and heterogeneous sources data are being consolidated for monitoring multimodal traffic patterns. Multimodal traffic patterns encompass all major transportation modes (road, railway, inland waterway, and active transport modes such as walking and cycling including other shared schemes).
The research reported in this paper aims at bridging the existing literature gap on the integrative analysis of multimodal traffic data and its situational urban context. This work is anchored in the pioneer research and innovation project "Integrative Learning from Urban Data and Situational Context for City Mobility Optimization"(ILU), in the field of artificial intelligence applied to urban mobility that joins the Lisbon city Council and two national research institutes. The manuscript is focused on the analysis of spatiotemporal indices of multimodality in passengers' public transport, offering three major contributions. First, it provides a structured view on the scientific and technical opportunities and challenges for data-centric multimodal mobility decisions to support mobility planning decisions. Second, it outlines key principles for the discovery of multimodal patterns from heterogeneous sources of urban data. Finally, a case study is presented on the spatiotemporal analysis of multimodality indices from available urban data, followed by a discussion on the relevance of cross-modal pattern analysis for the cooperation of public transport operators along with its contribution to enable align supply with passengers' demand to fit the self-actualizing city dynamics.

Keywords: *Multimodality; Sustainable Mobility; Data Science; Smart Cities; Public Transportation*

1. INTRODUCTION

In the last decades, road traffic and mobility needs have increased significantly, especially in urban and metropolitan areas, as a result of the ongoing economic growth and other socioeconomic changes. These are being challenged by the needs to guarantee norms of social distancing, climate objectives to reach carbon neutrality and the decentralization of activities and services to the periphery of urban centers. The heavy use of cars as private transport compromises the sustainability of modern cities. The European Commission have already recognised the important of multimodal passenger transport to increase the use of public transport and other active modes such as walking and cycling, including shared mobility options.

© AET 2020 and contributors

Multimodality, the use of different modes of transport in a single trip, can support the shift to a low carbon economy by taking advantage of the benefits of using different transport modes, such as convenience, safety, speed, cost and reliability.

Mobility in major European capitals is not yet sustainable, prompting those capitals to reevaluate their public transport systems to meet societal goals. Lisbon's City Council is making efforts in collecting heterogeneous urban data for a better understanding of the city mobility patterns. Detection of mobility patterns can offer data-centric views ensuring:

- the city mobility plan is sensitive to changes (self-adaptation);
- fully transparent decisions to the citizens, enhancing the accountability of authorities;
- an objective coordination between the different authorities involved in planning and managing urban mobility.

Big data are currently being consolidated in the Intelligent Management Platform of the City of Lisbon (PGIL) to meet various purposes. Still, the potentialities of exploring the multiplicity of available urban data sources in an integrative manner for reaching sustainable mobility goals are still untapped.

This paper aims at bridging the existing gap on the integrative analysis of multimodal traffic data and its situational urban context. In particular, it uses Lisbon as a reference case study to identify major opportunities and current challenges around and towards a reinforced context-aware multimodal mobility. A set of principles are further identified to address existing challenges. Finally, a spatiotemporal analysis of multimodality indices is conducted for the city of Lisbon using the available urban data, highlighting the policy and planning relevance of cross-modal pattern analysis.

The remainder of this paper is organized as follows. *Section 2* presents the essential background on multimodality. *Section 3* explores emergent opportunities for multimodal mobility decisions, and identifies current major obstacles. *Section 4* introduces principles for multimodal data analysis, offering guidelines to overcome each obstacle. *Section 5* outlines the methodology and case study in the city of Lisbon, to be integrated in the Web Application ILU, where most of the work is being submitted. A final discussion is presented in *Section 6*.

2. BACKGROUND

2.1. Multimodality

Multimodality is commonly defined as the use of more than one transport mode to complete a trip within a certain time period. By contrast, monomodality generally refers to the exclusive use of one mode of transport. Buehler and Hamre (2016) state that multimodality is a subfield of a larger body of research on intrapersonal variability of travel behaviour, which consists of four dimensions: temporal, spatial, purpose and modal. Where the "modal" dimension describes the variability in the use of means of transport over time, referring to the multimodality research. Nobis (2007) emphasizes the fact that the general definition of multimodality must be observed along individual trips to ensure its separation from the monomodality concept. This distinction relates to the chosen time period, the longer the time period is, the higher is the probability that a person uses more than one mode of transportation. For instance, Nobis (2007) uses in her study a loose definition of multimodality, where any person who uses more than one mode of transportation within one week is a multimodal transport user.

© AET 2020 and contributors

Comparison of findings about multimodality across studies is challenging, because of different geographic locations, data sources, timing, and definitions of multimodality. However, some relevant results are common among studies: the percentage of multimodal persons decreases with advancing age (Kuhnimhof et al., 2006; Nobis, 2007; Chlond, 2012); car availability is negatively correlated with multimodal behaviour, and positively correlated with monomodal driving (Kuhnimhof et al., 2006; Nobis, 2007; Diana and Mokhtarian, 2009); and having a driver's license is negatively associated with multimodal users (Kuhnimhof et al., 2006; Nobis, 2007).

Multimodality is generally measured by considering the fraction of users that use a given number of travel modes. For example, Nobis (2007) shows that car and public transportation users tend to be between 10 and 25 years old, with the largest group consisting of people aged 18–25, in Germany. While Buehler and Hamre (2016) indicate that 87% of all trips in the United States are made by car and 90% of Americans use automobiles in their commuting trips for work purposes. Most of these works don't have in consideration the intensity of use of each mode. Diana and Pirra (2016) targeted the problem of measuring multimodality at the individual level, by finding a multimodality index that comprises both descriptive statistics on the number of travel means, and the intensity of use of each mode.

One of those measures is the Herfindahl–Hirschman index (HHI). It's a measure of market concentration and is used to determine market competitiveness. More specifically, it's a measure of the size of firms in relation to the industry, it can range from 0 to 1 moving from a perfectly competitive market with a high number of small firms to a monopoly. According to Diana and Pirra (2016), in the context of transportation, the value of the index is closer to zero when a lot of different travel means are used and no means is very intensively used, whereas the value increases when the use of a smaller number of modes tends to dominate. The index can be defined as follows (Rosenbluth, 1955):

$$HH = \frac{1}{n}\left[\frac{n\sum_{i=1}^{n}(f_i - \bar{f})^2}{(\sum_{i=1}^{n} f_i)^2} + 1\right], \qquad (1)$$

where f_i is the intensity of use of ith mode, and f is the mean value of intensities of all n modes. Diana and Pirra (2016) proposed a variant of (1) in order to distinguish between the set of available modes and the set of effectively used modes. This alternative equation, considers only the m elements different from zero, while the coefficient of variation and the variance are computed over all n modes:

$$HH_m = \frac{1}{m}\left[\frac{n\sum_{i=1}^{n}(f_i - \bar{f})^2}{(\sum_{i=1}^{n} f_i)^2} + 1\right]. \qquad (2)$$

Susilo and Axhausen (2014) used HHI to measure the repetitiveness of identical combinations of individual's spatial–activity–travel mode choices within an observed period. In their study, higher index values were associated with habitual behaviour and lower index values with less repetitive and variety-seeking behaviour. Susilo and Axhausen (2014) recommend the use of HHI to characterise the level of repetition of activity–travel patterns.
A comparable multimodal index is the Gini coefficient (Allison, 1978), which is usually used as a measure of income inequality in a population. A Gini coefficient of zero expresses perfect equality,

while a value of one expresses maximal inequality. In the context of multimodality, it behaves like the previous index. The Gini coefficient is defined as:

$$Gini = \frac{2}{n}\frac{\sum_{i=1}^{n} i \cdot f_i}{\sum_{i=1}^{n} f_i} - \frac{n+1}{n}. \quad (3)$$

where f_i is the intensity of use of the i^{th} mode and *n* the total number of modes. Tahmasbi et al. (2019) used the Gini coefficient to evaluate the equity of the distribution of urban public facilities and accessibility level of different groups of people; this work presents a similar methodology (see *section 5*).

Diana and Mokhtarian (2007) reinterpret the concept of Shannon Entropy (Shannon, 1948) by considering an hypothetical mode choice experiment, where the uncertainty of the outcome is proportional to past multimodality behaviours of the traveller, thus defining the following multimodality index:

$$OM_{PI} = \sum_{i=1}^{n}\left[\frac{f_i}{\sum_{j=1}^{n} f_j}\log_n\frac{\sum_{j=1}^{n} f_j}{f_i}\right]. \quad (4)$$

When OM_PI tends to 0 the individual uses only one mode among those being considered, whereas when OM_PI = 1 the individual uses all these modes with the same intensity. Diana and Mokhtarian (2007) proposed a variant of OM_PI that is sensitive to the mean mobility level of individuals. With M as the absolute maximum reported frequency of utilisation of any mode, and nM as the potential maximum total frequency across all considered modes, hence defining a mobility-level-sensitive multimodality index, given by:

$$OM\_MI = \sum_{i=1}^{n}\left[\frac{f_i}{nM}\left[1 + ln\left(\frac{M}{f_i}\right)\right]\right]. \quad (5)$$

Diana and Pirra (2016) established an analogy between income inequality and multimodality, where individuals and their income, respectively map into travel means and their intensities of use. So they adapted an inequality measure, the Dalton Index (Cowell, 2011), for the multimodal transportation problem, where the parameter ϵ represents the decreasing influence of more intensely used modes to determine the degree of multimodality of a traveler:

$$DAL_m = 1 - \frac{\frac{1}{n}\sum_{i=1}^{m}\left({f_i}^{1-\epsilon} - 1\right)}{\left(\frac{1}{m}\sum_{i=1}^{m} f_i\right)^{1-\epsilon}}. \quad (6)$$

In their study, Diana and Pirra (2016) showed that there is not an index that outperforms the others, still, some measures give best results in specific cases. For example, if the goal is comparing multimodal behaviors of different social groups, an index that is not replication invariant is recommended: HH_m, OM_PI or OM_MI. Otherwise, if the mean intensities of use of the different modes are different across respondents, but some modes in the set are never used, in this situation, the application of the DAL_m index is more appropriate.

© AET 2020 and contributors

Transfers affect the attractiveness of passenger transport. Therefore, examining transfer patterns can be beneficial for public transport management. Jang (2010) illustrates that transfer data can be used to locate the critical transfer points that need improvement. The dataset used in his research came from the automatic fare collection (AFC) system of Seoul in South Korea. Contrary to the fare collection method adopted in Lisbon (see Section 3.1), the AFC in Seoul is distance-based, where the fare is calculated on the basis of the total distance run by buses, subway trains, or both from boarding to alighting. Seoul's transports are equipped with two smart card readers located at the doors, for boarding and alighting, so it is possible to obtain the whole itinerary of each individual trip from the departing location to the final destination, including intermediate transfers. The data collected from the AFC allows to identify stops or stations that have heavy transfer demands, pointing out the areas that need improvement.

2.2. Multimodality in Lisbon: the study case

This work is anchored in the pioneer research and innovation project ILU[1], a project that joins the Lisbon city Council and two research institutes, bridging the ongoing research on urban mobility with recent advances from artificial intelligence. The available traffic data comes from various heterogeneous sources collected for the Lisbon Metropolitan Area (LMA). The LMA is an administrative division in Portugal centered in the municipality of Lisbon and covering more 17 municipalities (Figure 1). Although the reported research is directed towards the municipality of Lisbon, its contribution and results can be extended and applied to other nearby municipalities to enable more comprehensive analysis of inter-municipal commuting mobility patterns.

Source: https://www.vimeca.pt

Figure 1. Lisbon Metropolitan Area and its municipalities.

Among the available sources, we find transport network data disclosing updatable information on the routes, stops and trips of the major public transport operators (see Table 1). Passenger transport

[1] Integrative Learning from Urban Data and Situational Context for City Mobility Optimization (DSAIPA/DS/0111/2018)

© AET 2020 and contributors

run by the mentioned public transport operators are equipped with smart card readers for boarding and, depending on the type of vehicle, alighting. As such, in addition to network planning data, card validations are further considered.

Table 1 Public transport operators in the Lisbon Municipality

Operators	Type	Routes (n.º)	Stops (n.º)	Length (km)	Distance average between stops (m)
Carris	Bus	77	2.174	796	366
Rodoviária de Lisboa	Bus	101	2.238	1.316	588
TST - Transportes Sul do Tejo	Bus	193	5.247	3.927	748
CP - Urbanos de Lisboa	Rail transport	3	65	179	2.754
Fertagus	Rail transport	1	14	54	3.857
Metro - Metropolitano de Lisboa	Subway	4	50	43	860
Transtejo & Soflusa	Ferry	6	9	37	4.111

Sources:
Road operators - IMT (SIGGESC, February 11, 2016)
Railway operators - Gismédia (Transporlis) October, 2012.

In particular, this work considers trip records gathered from three distinct modes: the bus operator CARRIS, the subway (METRO), and the Lisbon's public bike sharing system (GIRA). For CARRIS, the smart card data only monitors entries. As such, we make use of estimators of existing validations. For the remaining modes of transport, METRO and GIRA, we have access to both passengers' entry and exit records.

In the context of GIRA, these records correspond to pick-up or drop-off acts of bikes. CARRIS and METRO further provide consolidated identifiers associated with each validation, thus offering the possibility to model origin-destination (OD) matrices and infer multimodal patterns. Aside from public transport data, sources of contextual/situational data are also available including urban city plans with its major traffic attractor-generator poles (land use specific trip generation), weather data, major events (sports, political, concerts, etc.) and road traffic data.

3. OPPORTUNITIES AND CHALLENGES

3.1. Opportunities

Integrated fare collection system

The public transport network in the Lisbon Metropolitan Area (LMA) is composed by 12 major carriers. With the exception of active modes of transportation, the providers of bus, subway, railway and inland waterway modes of transport are currently operating under an integrated fare collection system, enabled through the VIVA card initiative. The VIVA card initiative, firstly established between the subway operator (METRO) and the major bus operator (CARRIS), was in 2017 extended to further encompass railway operator, Comboios de Portugal (CP), and in 2019 extended towards the remaining major carriers operating within (or interfacing with) the city of Lisbon[2]. To this end, the early individual ticketing systems were consolidated into a unique ticketing system coordinated by OTLIS, the entity responsible to manage the information resources shared among carriers.

[2] https://www.portalviva.pt/pt/homepage/sobre-a-otlis/empresas-aderentes.aspx

© AET 2020 and contributors

The integrated fare collection offers the unprecedented possibility to trace the movements of each user throughout the modes of the public transportation system, providing an essential source of information to understand the true mobility dynamics in the city. In 2019, multimodal tariff plans were also released to create incentives towards a multimodal use of the public transportation system.

Urban data consolidation

Among the diverse initiatives established by the Lisbon City Council towards sustainable mobility, we find initiatives concerning the access and consolidation of numerous sources of urban data – covering areas such as mobility, security, decarbonisation, urban planning, local development and civil protection. In the context of mobility, the following sources of traffic data are currently being consolidated:

- road traffic data from three major types of sensors: 1) inductive loop detectors in major road junctions in the city, offering discrete views on traffic flow; 2) mobile devices with global positioning systems (GPS) and active applications such as WAZE[3] or TomTom[4], offering aggregated views of traffic congestion along specific road segments (geolocalized speed data); and 3) privacy-compliant cameras in major roads;
- aggregated views of public transport data, including passengers' card validations and the GPS positioning of public vehicles. Due to privacy and security aspects, only aggregated views of card validations along the public transport network are maintained by the city Council. The raw trip records are maintained separately by each operator and consolidated by OTLIS to collect statistics and ensure the sound interoperability of ticketing systems;
- bike sharing data from the Lisbon's public bike sharing system (GIRA), including trip records per user, user feedback on bicycle's condition, bike charging information, bike malfunction and repair status, among others;
- other sources: emerging modes of transportation, including private scooter traffic data, are being also consolidated. An entry requirement for new private operators is precisely the full disclosure of trip records.

Context data incorporation

The Lisbon city Council further established protocols to collect diversified sources of situational context with potential impact on traffic for guiding mobility decisions. Some of the available sources of context data include:

- public events, including conventions, festivals, concerts, and sport events. The historic and prospective events are currently sourced from the cultural city agenda and planned usage of halls, stadiums and open areas;
- urban planning of the city with the localization of traffic generator-attractor poles, including: commercial poles (malls, commercial permits, markets, terminals), education facilities (schools, universities, institutes), health-related facilities, sport facilities, cultural poles (concert halls, museums, theatres), recreational spaces, parks, or citizen spaces;
- ongoing and planned construction road works (traffic conditioning events) characterized by a set of trajectories with (possibly non-convex) interval of obstruction and accompanying details (including the number of affected ways and whether interruption is spasmodic);

[3] https://www.waze.com/en-GB/

[4] https://www.tomtom.com/en gb

© AET 2020 and contributors

- weather record data sourced from three meteorological stations maintained by Instituto Português do Mar e da Atmosfera (IPMA);
- other sources of interest including details on traffic and transport networks (mostly walking, road and cycling infrastructures), zoning information (including traffic analysis zones), city occurrences (including road accidents and incidents, medical emergencies, fires and floods, logistical help and falling structures, transport requests, conservation and complaints, and rescue and civil protection), and other calendric information with impact on traffic patterns (e.g. bank holidays).

3.2.Challenges

Multimodal data analysis

The integrative analysis of traffic data produced from heterogeneous modes of transportation is challenged by four major factors:

- the inherent spatiotemporal and multimodal nature of traffic data. The rich spatial, hourly, calendrical and modal content of traffic data should be properly explored, and the available sources of urban traffic data soundly processed and consolidated;
- the massive size of data. When we look to the Lisbon's major public carriers, over 50 million of trips are recorded per month. Analyzing such massive data comes with strict scalability requirements for the pursued processing and learning algorithms;
- the inherent traffic variability associated with non-pendular trips, changing mode preferences, and sporadic event-driven traffic. These factors, together with the inherent temporal and spatial stochasticity of pendular trips, create challenges for assessing multimodal traffic patterns;
- the context-dependent nature of mobility data. The presence of large-scale events create irregular picks of demand; road traffic congestion and interdictions condition mobility; weather impacts transportation mode decisions, specially active modes of transport; changes to the city urban planning affect the way traffic is generated and attracted to different parts of the city throughout the day.

Emerging traffic changes

Mobility patterns within a city are subjected to both technological and non-technological disruptive changes, such as those triggered by the emerging COVID-19 pandemic. The value of static studies is thus of limited relevance as their findings can easily become depreciated. Instead, multimodal traffic data analysis should be fully-automated and updatable once more recent data becomes available. In this context, there is the need to guarantee that the ongoing mobility changes are reflected in the computational models, as well as the ability to learn from traffic data streams and detect emerging traffic patterns at early stages in order to act in a timely manner.

Assessing multimodal decisions

A major difficulty is on promoting and testing the efficacy, actionability, and statistical significance of decisions driven by the multimodal traffic data analysis. In this context, the impact of decisions on user's waiting time and intra/inter-mode commutes, the exploitation of cost synergies, the incentives for eco-friendly modes of transport (walking and cycling), among many other criteria should be established and objectively assessed using ground truth from available data. Once mobility decisions are placed, the need for continuous monitoring their impact and adjusting reform programs using traffic data as the ground truth remains an additional challenge.

© AET 2020 and contributors

Multimodal planning
In addition to the above technical challenges, at the present moment, the public transport operators in Lisbon do not share their raw trip record data. Hence, each operator is only able to acquire a partial view of their passengers/users' movements along the city network without exploring synergies with other modes. In spite of the integrated fare collection system and joint tariffs, the partial access to trip record data prevents a comprehensive view of multimodal traffic patterns, including those pertaining to cross-modal commuting. As a result, passenger public transport operators need to make efforts to enable an objective and transparent ground to coordinate their planning, to explore schedule and route synergies for the benefit of the citizens.

4. MULTIMODAL BIG DATA ANALYSIS: PRINCIPLES/GUIDELINES

4.1. Addressing challenges

Multimodal data analysis
Numerous principles have been suggested in the literature for the integrative analysis of traffic data from heterogeneous modes of transport:

- *descriptive analysis*: 1) inference of multimodal origin-destination matrices by consolidating trip record data and tracing the complete movements of individual users throughout the public transport network (Munizaga and Palma, 2012; Wong et al., 2005); 2) mining of actionable traffic patterns, including frequent, periodic, emerging and anomalous patterns (Li et al., 2020; Liao et al., 2020; Yang et al., 2015); 3) discovery of bottlenecks to multimodal mobility (waiting times, number of commutes, walking distances within and outside commutes) from trip record data (Rempe et al., 2016; Munizaga and Palma, 2012); and 4) modelling traffic expectations by exploring the rich spatiotemporal content of the available traffic data and taking into consideration user-specific commutes in interface areas. State-of-the-art principles on spatiotemporal pattern mining, urban data fusion and analytics, and relational data mining can be pursued towards these ends (Atluri et al., 2018; Zheng et al., 2014; Dzeroski, 2009);
- *predictive analysis*: traffic forecasting is the predominant prediction task (Luo et al., 2019). Following breakthroughs from deep learning along the last decade, we observed a shift from classic statistical approaches towards recurrent neural networks (Fu et al., 2016) and graph neural networks (Wu et al., 2020) able to capture heterogeneous modes of transport along both short-term and long-term forecasts;
- *prescriptive analysis*: comprises advances on simulation, control and optimization to support decisions related with both individual and multimodal planning of the public transportation network (schedule-, vehicle- and route-wise) and urban traffic positive conditioning. Model-based multi-agent reinforcement learning (Wiering et al., 2004), hierarchical network agent structures (Choy et al., 2003) and the use of deep neural networks as the underlying representation of the control problem (Genders and Razavi, 2016) have been proposed towards these ends.

Emerging traffic changes
To account for ongoing urban mobility changes, traffic data analysis should be an automated process taking an arbitrary periods of monitored urban data as input. In this context, the following principles should be pursued:

© AET 2020 and contributors

- principles from incremental data mining and online learning should be placed to guarantee the ability to learn from data streams, where new records are continuously arriving. These principles guarantee the updatability of the models in the presence of more recent data without the need to compute descriptive and predictive models fully from scratch;
- an additional important principle is the early discovery of emerging mobility patterns, which are driven by mobility dynamics (demand at a given station/time that are gradually changing e.g. due to variations in contextual variables) to anticipate significant changes ahead (strategic and tactical planning). In addition, trends and periodicities should be further identified for a proper understanding of non-seasonal changes in the city traffic.

Assessing multimodal decisions

Objective assessments are necessary to guarantee the adequacy of decisions placed from multimodal models of urban traffic. In this context, assessment should be pursued at three major levels:

1. *data analysis level*: the aforementioned descriptive, predictive and prescriptive multimodal analyzes should be equipped with robust evaluation criteria to assess their proper decision translation. For instance, residual analysis and inference of upper and lower statistical bounds should be pursued in predictive models using a sound evaluation setting, such as cross-validation schema on a rolling basis;
2. *decision level*: a structured set of assumptions to model the impact of decisions on user behavior should be carefully identified, including receptivity for mode-commutes and tolerable walking distances on a user-by-user basis in accordance with historical data (Clark et al., 2003; Heinen and Bohte, 2014). Once these assumptions are defined, post-decision mobility dynamics can be estimated by inferring new patterns of multimodality, conducting simulation studies, or gathering user feedback for an objective assessment;
3. *post-decision level*: it is the easiest assessment level since the mobility dynamics before and after a decision can be objectively compared towards this aim. Illustrating, the new patterns of multimodality can be discovered in order to measure the impact of changes in the public transportation network for specific groups of users or the overall population in terms of waiting times, number of commutes, and adherence towards active and public modes of transport.

Multimodal planning

The data-centric analysis of the traffic demand and public transportation supply provides a ground truth for the transparent and objective coordination between carriers. In this context, it is important to satisfy the following principles:

- guarantee the interpretability of the learned models and the traceability of the recommendations. The models should be easily auditable in order to guarantee that there is no preference towards specific carries in detriment others;
- offer a robust statistical frame. Given the stochastic nature of mobility dynamics, it is essential to assess whether the found patterns of multimodality occur or not by chance in order to strictly guarantee statistically significant outputs (Henriques and Madeira, 2018). In this context, statistical tests can be placed to assess the trustworthy degree of decisions, and new heuristics incorporated within the learning process to minimize false positive and false negative discoveries;

© AET 2020 and contributors

- comprehensively compare alternative decisions (e.g. suboptimal routing and scheduling plans) in order to assess complementary scenarios and further validate the quality of the suggested recommendations.

4.2.Leveraging Opportunities

Integrated fare collection system

The recently consolidated ticketing system for the public passenger transport operators in LMA provides the unprecedented possibility to trace both cross-carrier and multimodal trips along the public network. This system offers an essential source of information to: 1) assess the efficacy of transport mode transfers in urban interfaces; 2) refine OD matrices in accordance with the complete (instead of partial) commuting travel patterns of individuals; 3) recover multimodal patterns to assess the needs and cross-modal preferences of the citizens; 4) understand demand; and 5) support the multimodal planning of routes and schedules to minimize commuting needs and waiting times.

Urban data consolidation

Urban traffic data can be consolidated by identifying shared dimensions between sources, including the user dimension (unique card identifiers for trip records), time-and-date dimensions, spatial dimension (point, origin-destination or trajectory geographical annotations), and carrier dimension. Considering a multidimensional schema, this modeling enables a coherent cross-modal navigation throughout the records of specific users, passenger transport operators, geographies and time periods.

Given the massive size of urban data, data extraction facilities should be able to adequately index spatial, temporal and modal information for the efficient retrieval of information (Mamoulis et al., 2004). In this context, the target data centric recommendation systems should be equipped with efficient slicing and dicing procedures. Particular attention should be further paid to avoid unnecessary inefficiencies – for example, the characteristics of the stations (or details of the cards) should be decoupled from the card validation records. In addition, data cleaning procedures should be available to ensure the absence of duplicates and gross errors, and further treat outlier and missing values whenever necessary. Finally, updating routines are necessary for the automatic extraction, transformation and loading of the continuously arriving data records into the consolidated database.

Context data incorporation and context-aware learning

Recent attention has been paid on how to incorporate context to enhance traffic data analysis (Leite et al., 2020). Different principles have been placed to incorporate and learn from different sources of context, namely weather records, planned events, and occurrences of potential relevance from social media data (Soua et al., 2016; Tang et al., 2019; Wibisono et al., 2012; Rodrigues et al., 2017; Kwoczek et al., 2014). Two major classes of context-sensitive approaches can be identified from the existing literature. First, approaches that aim to describe and predict traffic dynamics by segmenting data in slices according to the available situational context and using only context-resembling slices for understanding and forecasting demand (Li et al., 2015; Kwoczek et al., 2014). Second, approaches able to embed the context directly in the models by capturing correlations with the context and using these correlations as correction factors to automatically adjust descriptive and predictive models (Gallop et al., 2012; Rodrigues et al., 2017).

© AET 2020 and contributors

5. METHODOLOGY AND RESULTS

This section outlines the methodology for multimodal traffic data analysis using the previously introduced principles. Empirical evidence is collected along the major stages of the process and applied for the Lisbon's study case to integrate the application ILU project Web tool.

Processing multimodal traffic data. The first step of the methodology comprises the collection, preprocessing and uniformization of the available sources of traffic data. Figures 2A, 2B, 2C and 2D identifies the routes of the major public passenger transport operators in Lisbon. Five modes of transport are considered – road (CARRIS, TST, RodLisboa, Sulfertagus), subway (METRO), railway (CP

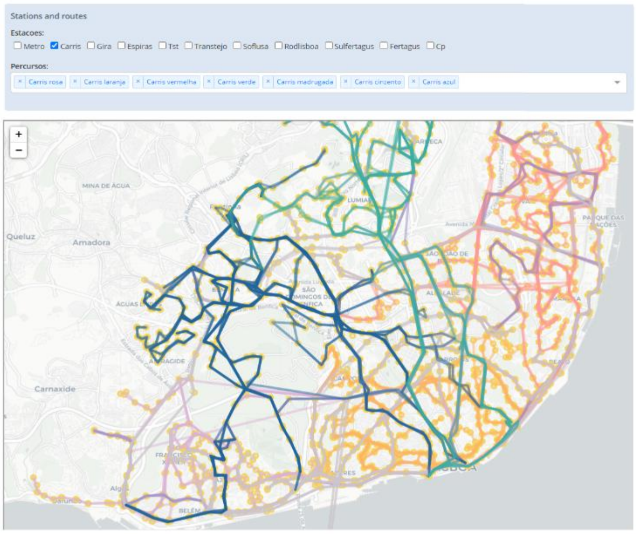


Figure 2A. Routes of CARRIS (major bus operator) by class: night, red, green, blue, yellow, orange, pink buses.

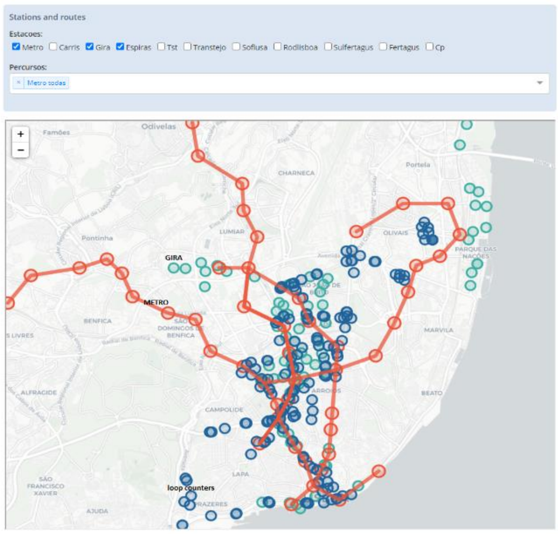


Figure 2B. Subway METRO stations and lines (red), GIRA bike stations (green) and road sensors (blue)

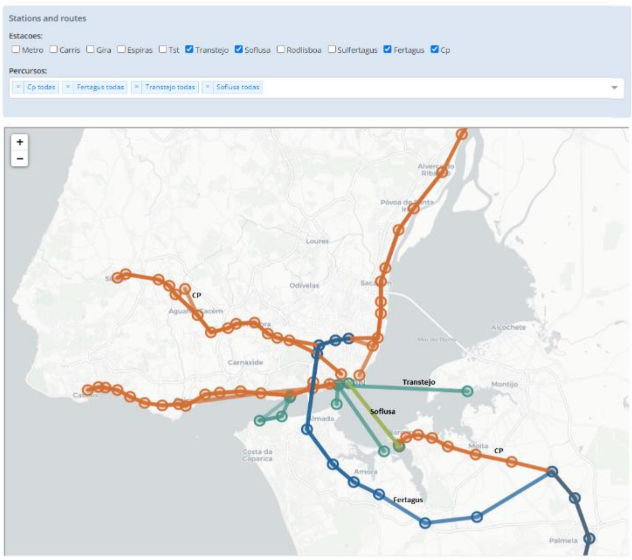


Figure 2C. Stations and routes of railway and inland waterway operators: CP (orange), Fertagus (blue), Transtejo (dark green) and Soflusa (lime)

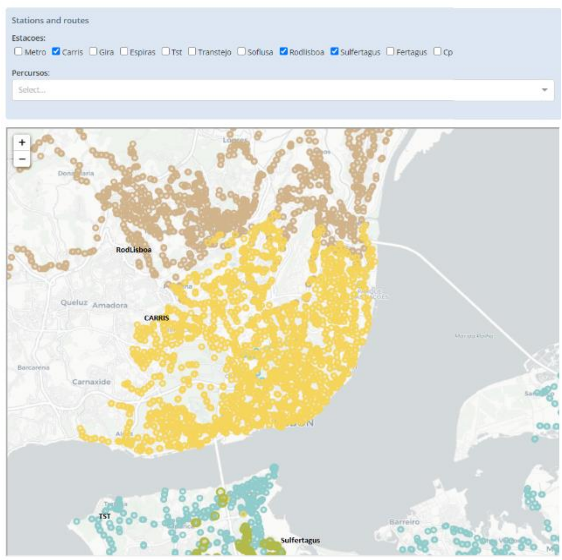


Figure 2D. Stations of four major bus operators: CARRIS (yellow), RodLisboa (brown), TST (light blue) and Sulfertagus (lime)

© AET 2020 and contributors

and Fertagus), inland waterways (Soflusa and Transtejo), and cycling (GIRA). Fixed road sensors (inductive loop counters) to assess private road traffic are also displayed in Figure 2B. In addition to real-time traffic data views, gathering routing and calendar planning information (using for instance GTFS protocols) is essential along this initial stage to guarantee a proper consolidation of the traffic data produced by the different carriers.

Figure 3 provides an overview on the traffic volume for the two largest public passenger transport operators in Lisbon -- METRO (subway) and CARRIS (bus) – the Entrecampos urban area throughout October 2018. Entrecampos is an interface area that encompasses multiple modes of transport and is further characterized by the presence of business and cultural traffic generation poles. Figure 3A depicts the volume distribution (given by card validations along October 2018) along the bus and subway stations situated along the Entrecampos region. Figure 3B shows the hourly volume of check-in validations on Entrecampos' bus stations, while Figure 3C shows both check-in and check-out validations for the two subway stations situated in this area for 15-minute intervals. Generally, the amount and pattern of card validations strongly vary across stations. There is a preference towards subway mode of transport in this area.

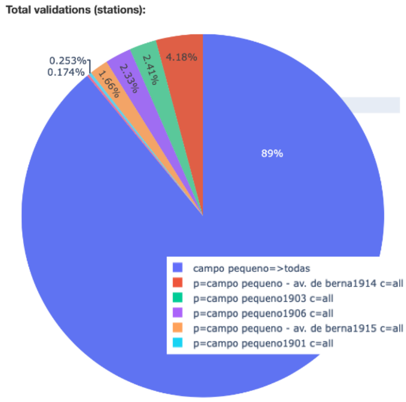


Figure 3A. Distribution of the volume of passengers for bus and subway stations at Entrecampos (October 2018)

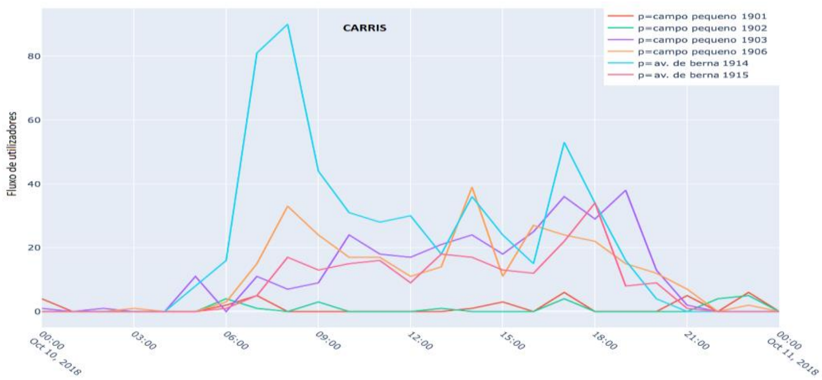


Figure 3B. Hourly check-in validations in CARRIS buses at Entrecampos on a regular weekday.

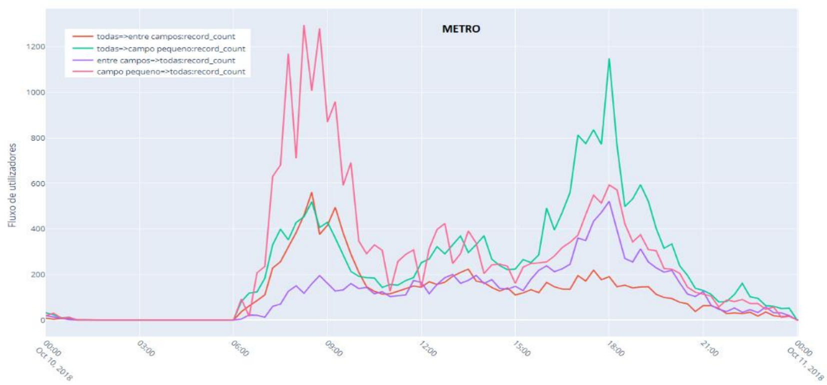


Figure 3C. Check-in and check-out validations at METRO stations in Entrecampos on a typical weekday.

© AET 2020 and contributors

A more comprehensive view of the market quota of these public passenger modes of transport -- bus, subway and cycling (bike sharing) -- along traffic analysis zones (TAZ) is provided in Figure 4. Figure 4 depicts the geographical limits of the target TAZ. It shall be noted that not all TAZ are covered by subway or bike stations, hence the predominance of the bus mode (CARRIS) for a significant number of zones. The adherence towards the public bike sharing system is smaller in magnitude for most of the covered zones.

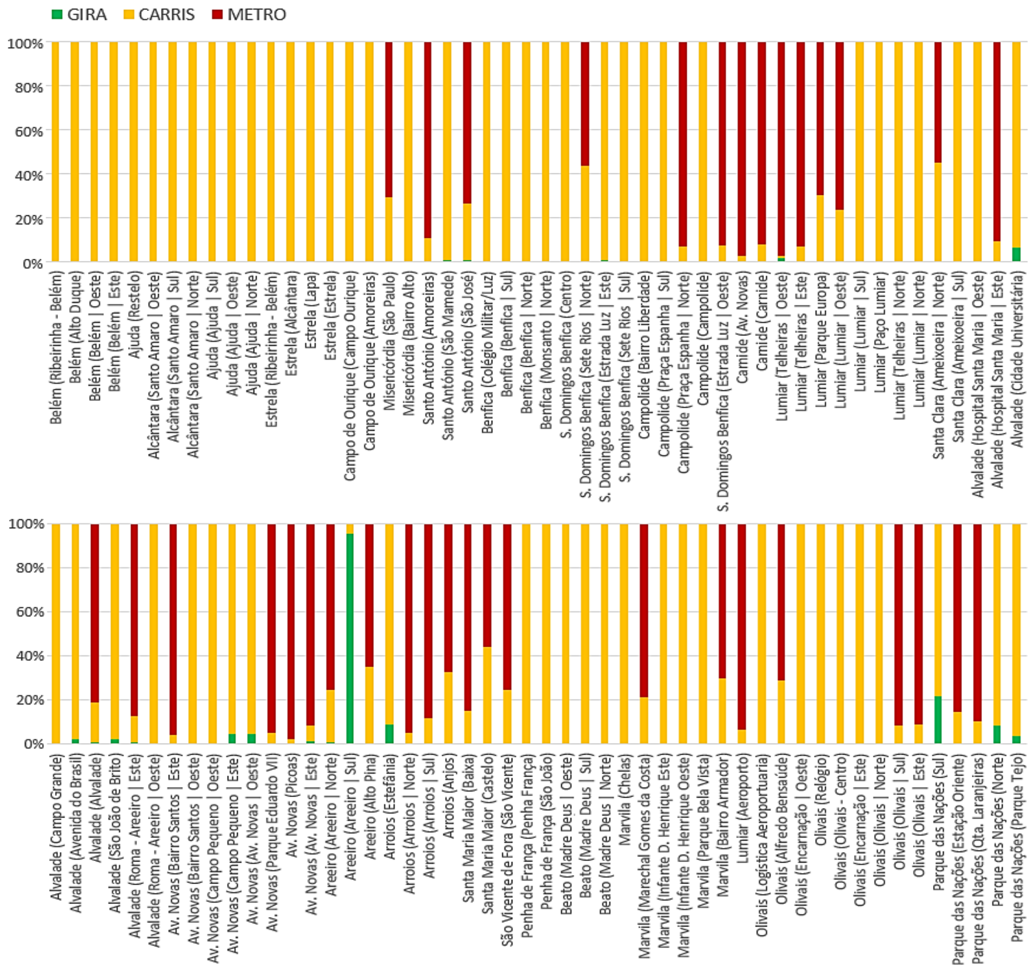
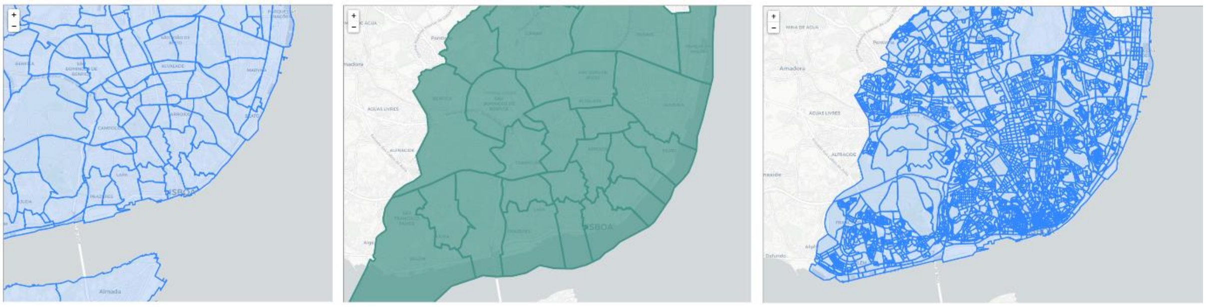


Figure 4. Cycling-bus-subway market quota (modal trip share) for major traffic analysis zones (TAZ) in Lisbon.

Selecting spatial criteria. Multimodal pattern analysis can be conducted at different spatial granularities. Two major possibilities are considered. First, the user can manually specify the target geographical region of interest using polygon and circular marking facilities. Second, the user can select predefined regions. We provide the following zoning maps for the Lisbon Metropolitan Area:

- Traffic Analysis Zones (TAZ): geographical unit used in transportation planning models to assess socio-economic indicators (Figure 5a);
- Municipalities: coarsest geographical unit for the city. Currently, this work uses city parishes as the administrative criterion of division (Figure 5b);
- Sections: finest geographical unit, comprising small districts and neighborhoods (Figure 5c);

© AET 2020 and contributors



a) TAZ zoning b) City municipalities c) Neighborhood sectioning

Figure 5. Zoning: geographical decomposition of the Lisbon city at different granularities.

Under the selected spatial granularity, traffic events, such as card validations and trajectories, as well as the accompanying situational context data, are then linked to one or more Lisbon's zones in accordance with their spatial extent.

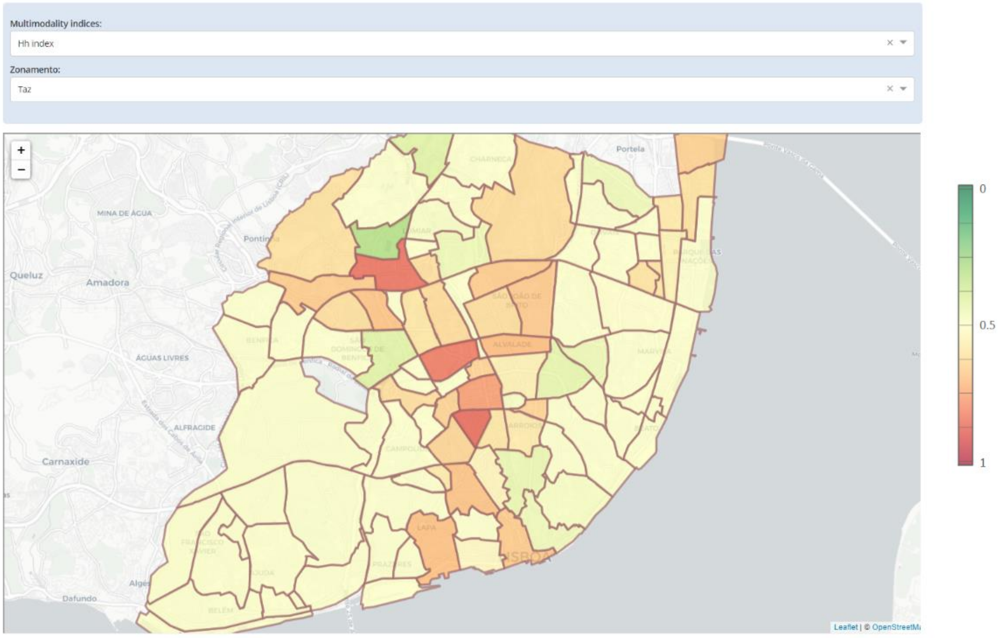
Selecting temporal criteria. Two major types of temporal constraints can be placed. First, calendrical constraints – such as day of the week (e.g. Mondays), weekdays, holidays or on/off-academic period calendars – can be placed to segment the available traffic data. The introduced principles for multimodal pattern analysis (section 4) can then be applied per calendar or, alternatively, correction factors can be learned from calendrical annotations in order to guide the target tasks. Second, time intervals (e.g. on/off-peak hour intervals) or a fixed time granularity (e.g. 15-minute) can be optionally specified to guide traffic data descriptors or predictors. For instance, passenger volume series in public transport can be resampled from card validations. In the absence of a minimum time granularity, the data analysis can be conducted at the raw event level or under multiple time aggregations.

Consolidating traffic data. Once these constraints are fixed, multi-dimensional querying can be automatically derived to produce the consolidated data. In addition, data mappings are generally further applied to transform the retrieved spatiotemporal data structures into georeferenced multivariate time series structures, more conducive to the subsequent traffic data analysis stages.

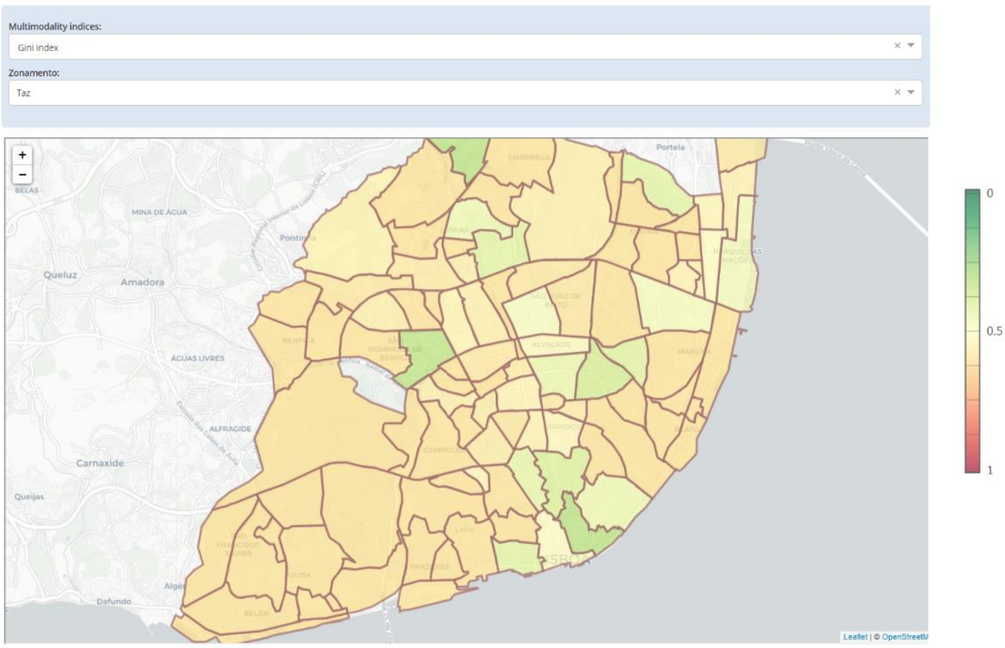
Multimodality index data analysis. For detecting vulnerabilities associated with multimodal transportation, two major options are made available. First, the inference of multimodal origin-destination matrices. The origin-destination matrices, currently provided for CARRIS and METRO operators, are inferred from shared card identifiers along the public passenger transport operators. Entries in these matrices are marked with statistics, including number of cross-carrier and cross-mode commutes necessary to accomplish a complete origin-destination trip, that support the analysis of commuting susceptibilities.

Second, the user can select one of the introduced indices of multimodality and compute them for different regions and time periods. Figure 6 (a and b) depicts a spatial distribution of the revised Herfindahl Hirschman index (equation 2) and multimodality Gini index (equation 3) for the traffic analysis zones (TAZ) of the city using the average daily volumes of passengers produced by the bus (CARRIS), subway (METRO) and cycling (GIRA) modes of transport along October 2018.

© AET 2020 and contributors



a) revised Herfindahl Hirschman index, HH_m (2)



b) Gini index as proposed by (Tahmasbi et al. 2019)
Figure 6. TAZ distribution of multimodality indices of bus-subway-cycling modes of transport in October 2018.

Generally, we can observe two major sources of multimodal penalties: the presence of many zones with only mode of transport (generally bus on the periphery), as well as the intense preference towards subway transport in the center of the city. The revised HH index is sensitive to the absence of traffic generated by modes without stations on a given zone, hence we can see that the peripherical zones of the city are not as penalized by this index as they are by the Gini index which is normally used for equity assessment. Despite the concordant views offer by the two indices in sign, the gathered results further underline the presence of some significant differences, highlighting the importance of selecting each multimodality index aligned with the end purpose of the study. Figure 7 extends the previous analysis for Lisbon municipalities, highlighting differences as the coarser zones are now able to encompass new stations and further suggesting the importance of identifying a proper spatial criterion for the analysis of multimodal indices. Multimodal index visuals offer a high-level visualization that can be further complemented with additional information, including traffic annotations and traffic generation poles (Figure 8).

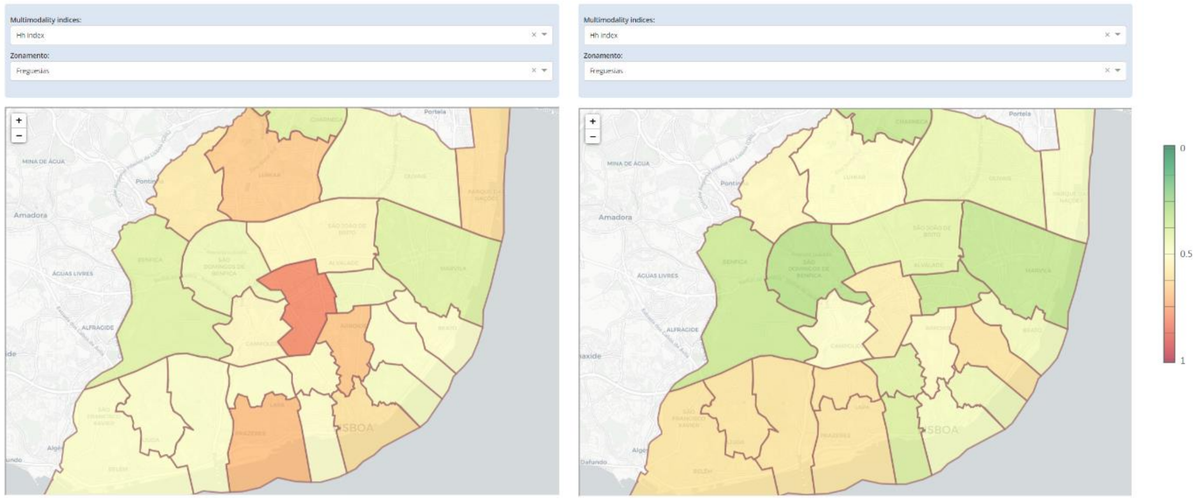


Figure 7. Distribution of multimodality indices (revised HH and Gini indices) along Lisbon municipalities considering bus-subway-cycling modes of transport in October 2018.

Incorporating situational context. The analysis of multimodality indices is only meaningful in the presence of situational context. The major constituent elements of such context are the traffic generation poles. The concept of traffic generation and attraction poles generally refers to commercial areas, employment centres such as business parks and enterprises, and collective equipment like hospitals, schools and stadiums, that generate or attract a significant volume of vehicle trips, either from contributors, visitors or providers. We currently maintain a complete localization of traffic generation poles for the city of Lisbon, as well as major city events (such as large concerts, congresses and soccer matches). Figure 8 provides maps of the city with some poles with impact on the city traffic.

The combined analysis of the above traffic generation/attraction poles' maps with the computed multimodality indices, as well as station-route maps, provide a comprehensive and dynamic way of modelling the spatiotemporal distribution of traffic along the city. Complementarily, the surveyed indices can be revised to further measure how the volume of passengers generated and attracted by nearby poles are being currently satisfied by the co-located modes of public transport.

© AET 2020 and contributors

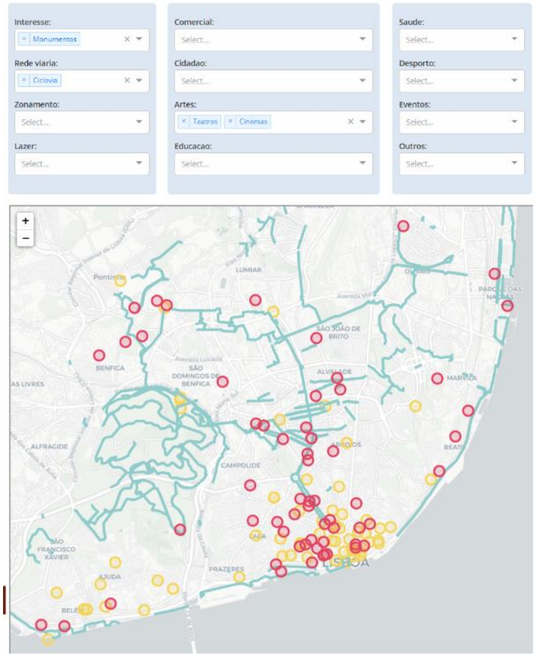


Figure 8a. Cycling roads (green), art and cultural poles (red), and tourist attraction poles (yellow)

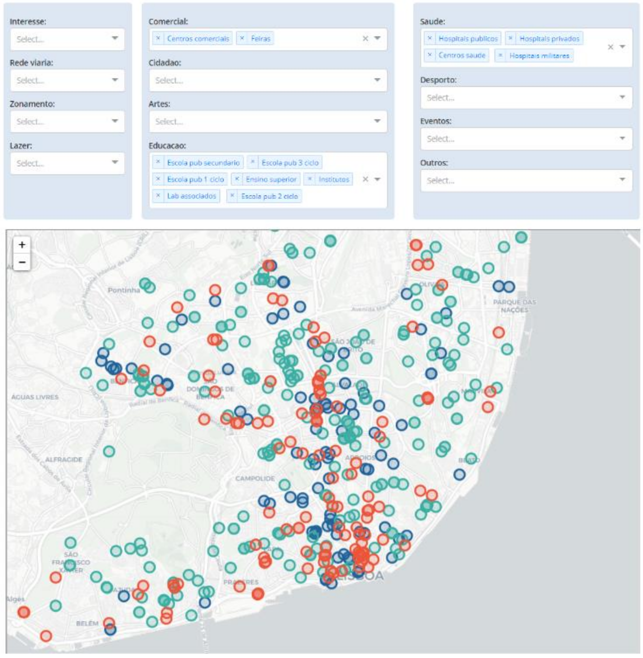


Figure 8b. Major traffic generation poles: commercial (blue), schools and institutes (green) and health centres (red)

6. CONCLUSIONS

The research work offers a structured view on the opportunities and challenges for the analysis of big traffic data produced from heterogeneous sources and passenger transport modes. A set of guidelines to address existing challenges, while leveraging on opportunities, were sourced from the ongoing advances in the fields of artificial intelligence and data science which were applied to urban mobility through a real-life case study engaging the City of Lisbon and its major public passenger transport operators.

The established initiatives by the Lisbon City Council towards the consolidation of relevant sources of urban data on its intelligent management platform, together with the integrative fare collection system and entry requirements for carriers operating in the Lisbon metropolitan area, offers unique opportunities for multimodal pattern analysis and cross-carrier coordination. Still, the inherent nature of multimodal traffic data – heterogeneous, massive in size, rich in spatiotemporal dynamics, subjected to variable aspects, and context-dependent – together with the increasing disruptive changes in urban traffic poses challenges towards the pursue of data-centric multimodal decisions. To tackle these challenges, the research outlined and started to apply a comprehensive set of principles from context-aware, spatiotemporal, distributed and relational data mining.

The conducted analysis of multimodal aspects pertaining to the Lisbon case suggest that decisions grounded in available traffic data provide an objective and transparent means to improve the cross-modal cooperation of public passenger transport operators and explore untapped synergies for multimodal and sustainable mobility planning.

© AET 2020 and contributors

Acknowledgments
The authors thank the support of Câmara Municipal de Lisboa, particularly the Centro de Operações Integrado and Gabinete de Mobilidade. This work is further supported by national funds through Fundação para a Ciência e Tecnologia under project ILU (DSAIPA/DS/0111/2018) and INESC-ID pluriannual (UIDB/50021/2020).

Bibliography

Allison, P.D., 1978. Measures of inequality. American sociological review, pp.865-880.

G. Atluri, A. Karpatne, and V. Kumar. Spatiotemporal data mining: A survey of problems and methods. ACM Computing Surveys (CSUR), 51(4): 1–41, 2018.

R. Buehler and A. Hamre. An examination of recent trends in multimodal travel behavior among american motorists. International journal of sustainable transportation, 10(4):354–364, 2016.

B. Chlond. Making people independent from the car–multimodality as a strategic concept to reduce co 2-emissions. In Cars and Carbon, pages 269–293. Springer, 2012.

M. C. Choy, D. Srinivasan, and R. L. Cheu. Cooperative, hybrid agent architecture for real-time traffic signal control. IEEE Transactions on Systems, Man, and Cybernetics-Part A: systems and humans, 33(5):597–607, 2003.

W. A. Clark, Y. Huang, and S. Withers. Does commuting distance matter?: Commuting tolerance and residential change. Regional Science and Urban Economics, 33(2):199–221, 2003.

F. Cowell. Measuring inequality. Oxford University Press, 2011.

M. Diana and P. L. Mokhtarian. The individual desire to change a “modal basket” and its relationship with the actual and perceived use of different transport means. In 11th World Congress on Transport Research, 2007.

M. Diana and P. L. Mokhtarian. Desire to change one’s multimodality and its relationship to the use of different transport means. Transportation research part F: traffic psychology and behaviour, 12(2):107–119, 2009.

M. Diana and M. Pirra. A comparative assessment of synthetic indices to measure multimodality behaviours. Transportmetrica A: Transport Science, 12(9):771–793, 2016.

S. Dzeroski. Relational data mining. In ˇ Data Mining and Knowledge Discovery Handbook, pages 887–911. Springer, 2009.

R. Fu, Z. Zhang, and L. Li. Using lstm and gru neural network methods for traffic flow prediction. In 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), pages 324–328. IEEE, 2016.

C. Gallop, C. Tse, and J. Zhao. A seasonal autoregressive model of vancouver bicycle traffic using weather variables. In Transportation Research Board 91st Annual Meeting, number 12-2119, 2012.

W. Genders and S. Razavi. Using a deep reinforcement learning agent for traffic signal control. arXiv preprint arXiv:1611.01142, 2016.

E. Heinen and W. Bohte. Multimodal commuting to work by public transport and bicycle: Attitudes toward mode choice. Transportation Research Record, 2468(1):111–122, 2014.

© AET 2020 and contributors

R. Henriques and S. C. Madeira. Bsig: evaluating the statistical significance of biclustering solutions. Data Mining and Knowledge Discovery, 32 (1):124–161, 2018.

W. Jang. Travel time and transfer analysis using transit smart card data. Transportation research record, 2144(1):142–149, 2010.

T. Kuhnimhof, B. Chlond, and S. Von Der Ruhren. Users of transport modes and multimodal travel behavior: Steps toward understanding travelers' options and choices. Transportation research record, 1985(1):40–48, 2006.

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, 25(6):973–980, 2014.

I. Leite, A. C. Finamore, and R. Henriques. Context-sensitive modeling of public transport data. 2020.

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, 114:225–240, 2020.

Y. Li, Y. Zheng, H. Zhang, and L. Chen. Traffic prediction in a bike-sharing system. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '15, pages 33:1–33:10, New York, NY, USA, 2015. ACM.

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, 10(1):1–12, 2020.

X. Luo, D. Li, Y. Yang, and S. Zhang. Spatiotemporal traffic flow prediction with knn and lstm. Journal of Advanced Transportation, 2019, 2019.

N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung. Mining, indexing, and querying historical spatiotemporal data. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 236–245, 2004.

M. A. Munizaga and C. Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. Transportation Research Part C: Emerging Technologies, 24:9–18, 2012.

C. Nobis. Multimodality: facets and causes of sustainable mobility behavior. Transportation Research Record, 2010(1):35–44, 2007.

F. Rempe, G. Huber, and K. Bogenberger. Spatio-temporal congestion patterns in urban traffic networks. Transportation Research Procedia, 15: 513 – 524, 2016. ISSN 2352-1465. International Symposium on Enhancing Highway Performance (ISEHP), June 14-16, 2016, Berlin.

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, 39(11):2113–2126, Nov 2017. ISSN 1939-3539.

G. Rosenbluth. Measures of concentration. In Business Concentration and Price Policy, pages 55–99. Princeton University Press, 1955.

© AET 2020 and contributors

C. E. Shannon. A mathematical theory of communication. Bell system technical journal, 27(3):379–423, 1948.

R. Soua, A. B. Koesdwiady, and F. Karray. Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory. 2016 International Joint Conference on Neural Networks (IJCNN), pages 3195–3202, 2016.

Y. O. Susilo and K. W. Axhausen. Repetitions in individual daily activity–travel–location patterns: a study using the herfindahl–hirschman index. Transportation, 41(5):995–1011, 2014.

B. Tahmasbi, M. H. Mansourianfar, H. Haghshenas, and I. Kim. Multimodal accessibility-based equity assessment of urban public facilities distribution. Sustainable Cities and Society, 49:101633, 2019

L. Tang, Z. Duan, and Y. Zhao. Toward using social media to support ridesharing services: challenges and opportunities. Transportation Planning and Technology, 42(4):355–379, 2019.

A. Wibisono, I. Sina, M. A. Ihsannuddin, A. Hafizh, B. Hardjono, A. Nurhadiyatna, W. Jatmiko, et al. Traffic intelligent system architecture based on social media information. In 2012 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pages 25–30. IEEE, 2012.

M. Wiering, J. Vreeken, J. Van Veenen, and A. Koopman. Simulation and optimization of traffic in a city. In IEEE Intelligent Vehicles Symposium, 2004, pages 453–458. IEEE, 2004.

K.-I. Wong, S. C. Wong, C. Tong, W. Lam, H. K. Lo, H. Yang, and H. Lo. Estimation of origin-destination matrices for a multimodal public transit network. Journal of advanced transportation, 39(2):139–168, 2005.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 2020.

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), pages 5577–5582. IEEE, 2015.

Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 5(3):1–55, 2014.

© AET 2020 and contributors