Spatial-focused Multimodality Analysis in the City of Lisbon

Carlos Eduardo Rodrigues Lemonde de Macedo*
carlos.lemonde@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

Abstract

The Lisbon's City Council is establishing efforts to collect urban traffic data and their situational context for gaining more comprehensive views of the ongoing multimodal mobility changes and support decisions accordingly. The present work is anchored in the pioneer research and innovation project "Integrative Learning from Urban Data" (ILU), and describes a methodology to identify multimodal mobility patterns through the analysis of spatiotemporal indices of multimodality in passengers' public transport against the available situational context. The analysis was conducted by applying two social-economic indices (Gini Coefficient and Herfindahl index) in the city of Lisbon organized by a synthetic geographic unit, the Traffic Analysis Zones (TAZ). Results demonstrate that the center of the city, abundant in traffic generation and attraction poles, benefit from a more multimodal usage of the public transport system. Additionally, a software tool was built in order to aid specialists in the field, to find inconsistencies on the public transportation network of the city of Lisbon.

Keywords: Multimodality; Sustainable Mobility; Data Analysis; Spatio-temporal Data Analysis; Public Transportation.

1 Introduction

Multimodality, the use of different transport modes on the same journey, can offer more efficient transport solutions whilst contributing to a more sustainable and integrated transport system, by taking advantage of the benefits of the different modes, such as convenience, reliability, cost, speed and predictability.

The Lisbon's City Council is making efforts in becoming sensorized by collecting heterogeneous urban data for a better understanding of the city mobility patterns. Big data are currently being consolidated in the Intelligent Management Platform of the City of Lisbon (PGIL)[1] 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 work is anchored in the pioneer research and innovation project "Integrative Learning from Urban Data" (ILU)[2], a project that joins the Lisbon City Council and two research institutes (INESC/IST and LNEC) , bridging the ongoing research on urban mobility with recent advances from artificial intelligence. It proposes a methodology for the comprehensive understanding of multimodal synergies in demanding urban areas of Lisbon, using the available data collected from different sources, and to relate that knowledge with relevant situational context. The rest of the present work is organized as follows: Chapter 2 introduces concepts related to multimodality and spatial temporal analysis, also including a description of Lisbon's public transport network; Chapter 3 presents insights of multimodality behavior related studies, and related work on inequality measurement, and on multimodality performance measurement; In Chapter 4 a methodology to the assessment of multimodality at a spatial level is described; and Chapter 5 presents results of applying our approach in order to evaluate multimodal patterns of the public transportation in Lisbon; finally, Chapter 6 concludes this work and proposes some potential expansions for future work.

2 Background

2.1 Multimodality

The term *Multimodality* has been briefly described at the beginning of this document, but not clearly defined. **Multimodality** is commonly defined as the use of more than one transport mode to complete a trip within a certain time period, Not to mistake with **intermodality**, which refers refers to combining various modes of transportation in the course of one trip, it is a subset of multimodality. By contrast, **monomodality** generally refers to the exclusive use of one mode of transport[19]. 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[4]. 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[19]. This distinction relates to

*supervision by Rui Henriques and Elisabete Arsénio
[1]PGIL: https://https://lisboainteligente.cm-lisboa.pt.

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

MEIC-A, IST, UL
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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.

2.2 Spatial Temporal Analysis

Time Series: A time series represents a collection of values obtained from sequential measurements over time [3, 9]. A *time series* x is an ordered sequence of t real-valued observations from a random variable,

$$x = (x_1, ..., x_t), x_i \in \mathbb{R}. \tag{1}$$

Multivariate Time Series: The previous definition refers to univariate time series. When multiple variables are monitored along the same time range, the gathered observations form a multivariate time series [3, 9],

$$\vec{x}_t = [x_{1,t}, .., x_{m,t}], x_{i,t} \in \mathbb{R}. \tag{2}$$

Dynamic Time Warping: Dynamic Time Warping (DTW) is based on the Levenshtein distance, it finds the optimal alignment (or coupling) between two numeric time series, and captures flexible similarities by aligning the coordinates inside both series [21]. The cost of the optimal alignment can be recursively computed by,

$$D(A_i, B_j) = \delta(a_i, b_i) + min \begin{Bmatrix} D(A_{i-1}, B_{j-1}) \\ D(A_i, B_{j-1}) \\ D(A_{i-1}, B_j) \end{Bmatrix}, \tag{3}$$

where A_i is the subsequence $\langle a_1, ..., a_i \rangle$, B_j is the subsequence $\langle b_1, ..., b_j \rangle$, and δ is a distance between elements of two series. The overall similarity is given by:

$$D(A_{|A|}, B_{|B|}) = D(A_X, B_X). \tag{4}$$

Average Time Series (Barycenter): Given a set of time series $D = \{x_1, .., x_n\}$ in a space E induced by Dynamic Time Warping, the average time series $\overline{x}$ is the time series that minimizes [20]:

$$argmin \sum_{i=1}^{n} DTW^2(\overline{x}, x_i). \tag{5}$$

Pearson's Correlation: For numeric attributes, it's possible to evaluate the correlation between two attributes, A and B, by computing the Pearson's Correlation Coefficient (PCC) [11]. It can also be applied to time series by pairing observations by time points and ignoring time dependencies between observations,

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{n\sigma_A\sigma_B}, \tag{6}$$

where n is the number of tuples, a_i and b_i are the respective values of A and B in tuple i, $\overline{A}$ and $\overline{B}$ are the respective mean values of A and B, σ_A and σ_B are the respective standard deviations of A and B. And $-1 \leqslant r_{A,B} \geqslant +1$.

Spearman's Correlation: Spearman's Rank Correlation (SRC) is a non-parametric (distribution-free) correlation measure which equals the Pearson correlation computed from the ranks of the observations, only if all ranks are distinct integers, usually designated as r_S[25],

$$r_S = 1 - \frac{6\sum d_i^2}{n^3 - n}, \tag{7}$$

where $d_i = rank(X_i) - rank(Y_i)$. If each of the n measurements of one of the time series is denoted as X_i (i.e. $X_1, X_2, ..., X_n$), then $R(X_i)$ may represent the rank of X_i, where each rank is an integer, from 1 through n, indicating relative magnitude.

Detrended Cross-Correlation Analysis: Podobnik and Stanley (2008) proposed a modification of the above covariance equation, called Detrended Cross-correlation Analysis (DXA), in order to quantify long-range cross-correlations when non-stationarities are present[22]. In this procedure, two long-range cross-correlated time series y_i and y'_i of equal length N, are divided into Nn overlapping boxes, each containing $n + 1$ values. Defining two integrated signals $R_k = \sum_{i=1}^{k} y_i$ and $R_k = \sum_{i=1}^{k} y'_i$, where $k = 1, .., N$, a local trend to be the ordinate of a linear least-squares fit, $\tilde{R}_{i,k}$ where $i \leq k \leq i + n$, and the detrended walk as the difference between the original walk and the local trend. The covariance of the residuals in each box is calculated by:

$$f^2_{DXA}(n, i) = \frac{1}{n-1}\sum_{k=i}^{i+n}(R_k - \tilde{R}_{k,i})(R'_k - \tilde{R}'_{k,i}). \tag{8}$$

Finally, the detrended covariance is given by:

$$F^2_{DXA}(n) = \sum_{i=1}^{N-n} f^2_{DXA}(n, i). \tag{9}$$

3 Literature Review

3.1 Multimodality Patterns

Comparison of findings on multimodality analyses and initiatives 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 [5, 14, 19]; car availability is negatively correlated with multimodal behaviour, and positively correlated with monomodal driving [7, 14, 19]; and having a driver's license is negatively associated with multimodal users [13, 19]. A broad spectrum of studies on multimodality analysis can be found on the literature, tackling different perspectives.

3.2 Inequality Measurement

Assessing multimodality can be seen as a particular instance of a more general issue, the measure of diversity and inequality[8]. This section presents the properties of well-established inequality measures from the branch of socio-economics, and how to compute and model inequality.

The Lorenz curve is one of the simplest representations of inequality. Specialists use the Lorenz curve[16] to represent graphically the degree of inequality in the distribution of income in societies. Lorenz (1905) has pointed out the mathematical inaccuracy of certain commonly used methods, and has suggested a graphic solution. Individuals holding assets of varying size, are arranged in order, poorest to richest. The horizontal axis represents the percent of people and the vertical axis the percent of income those people receive. Equality of distribution would give a series of points in a straight line. The Lorenz curve is obtained by plotting the cumulative proportion of income against the cumulative proportion of population, represented in Figure 1.

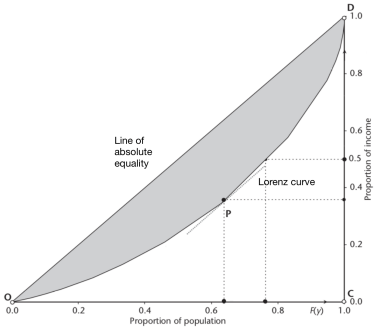


Figure 1. Lorenz curve.

Source: Economic Trends, November 1987

A common use of the Lorenz curve is to derive the Gini coefficient, expressed as the ratio of the shaded area in Figure 1 to the area OCD:

$$G = \frac{ODP}{OCD}. \tag{10}$$

The Gini coefficient was developed by the Italian statistician Corrado Gini[10] as a summary measure of income inequality in society. The Gini coefficient can be presented as a value between 0 and 1 or as a percentage. A coefficient of 0 reflects a perfectly equal society in which all income is equally shared; in this case the Lorenz curve would follow the line of equality. The more the Lorenz curve deviates from the line of equality, the higher will be the resulting value of the Gini coefficient. A coefficient of 1 (or 100%) represents a perfectly unequal society, where all income is earned by one individual in an infinite population. The equation 10 can be applied in practical terms as the mean of the difference between every possible pair of individuals, divided by the mean size[23]:

$$Gini = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu}. \tag{11}$$

The Gini coefficient as with many inequality measures, it is a synthetic index. Therefore, it does not contain all the information in the Lorenz curve, and it has been pointed out that different Lorenz curves can have the same Gini coefficient[24].

The inability of the Gini coefficient brought other inequality measures into the field of social welfare. One of these measures is the Atkinson index, which is used to evaluate the effectiveness and fairness of social distribution[23]. It ranges from 0 to 1, where 0 means the highest equality in distribution, while 1 means it is most unequal in distribution. Let T_i be the income in the ith income range, f_i the proportion of the population in ith group, and $\overline{T}$ the mean household income. The Atkinson equation is defined as follows[1]:

$$Atk = 1 - \left[\sum_{i=1}^{n}\left(\frac{T_i}{\overline{T}}\right)^{1-\epsilon} f_i(T_i)\right]^{\frac{1}{1-\epsilon}}. \tag{12}$$

Other interesting measure, slightly different from the previous mentioned inequality measures, is the Herfindahl-Hirschman (HH) index. The HH index is a measure of the size of firms in relation to the industry and an indicator of the amount of competition among them[12]. It can range from $1/n$ to 1 moving from a perfect competitive environment (all firms operate with equal market share) to a single monopolistic producer[2]. A commonly accepted measure for market concentration, the general form of the HH index is expressed as the sum of the squares of the market shares $s_i (i = 1, 2, ..., n)$ of all entities in the industry:

$$HH = \sum_{i=1}^{n}(s_i)^2, \tag{13}$$

if the following constraint holds:

$$\sum_{i=1}^{n} s_i = 1. \tag{14}$$

A modified version of HH is called the Normalized Herfindahl–Hirshman Index (NHH). Unlike the HH index, the NHH ranges between values from 0 to 1 and is calculated as follows[13]:

$$NHH = \frac{HH - 1/n}{1 - 1/n}. \tag{15}$$

3.3 Multimodality Traffic Performance Measurement

Multimodality is generally measured by considering the fraction of users that use a given number of travel modes. For example, Nobis (2007) showed 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[19]. 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[4]. 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[8]. They analysed some of the socio-economic measures, presented in the previous section, as multimodal indices, and in order to easily assess the different indices, they rewrote them as a function of a common set of parameters, reasoning at the end that there is not an index that outperforms the others, still, some measures give best results for different cases.

4 Methodology

4.1 Processing Multimodal Traffic Data

4.1.1 Data Preprocessing. The first step of the methodology comprises the collection, characterization, preprocessing and uniformization of the available traffic data from the different sources.

In the collection phase, the traffic sources are chosen and their corresponding data is collected. That choice can be based on the following factors:

- **Relevance**: the usage by the population, quantified by the number of validations per day, week or another temporal granularity (Section 4.1.2).
- **Accessibility**: the mode of transport must be available to a large number of the population. its usage must not be compromised by any type constraint, in other words, the access to its stations/stops must be straightforward and scattered across the chosen spatial granularity (Section 4.1.2).
- **Privacy Policy**: Even if working with modes of transport managed by public entities, that doesn't mean that the usage and disclosure of the respective traffic data can be easily available. Enforcement of privacy policies by those entities can be a hindrance for a complete and adequate analysis, by not sharing the requested urban data or just sharing a portion of the relevant information (i.e. incomplete data).
- **Quality of the Data**: The public transport entities can have open privacy policies, however, if the technology used to collect the data isn't reliable, the obtained information can be useless and not adequate for analysis. Another issue is if those entities only collect insufficient traffic data, like for example, collecting only the number of validations and not registering a timestamp associated to the validations. Low-quality data will lead to low-quality mining results, affecting accuracy, completeness, consistency, timeliness, believability, and interpretability[11].

The step of collecting relevant traffic data is followed by its profiling stage. This phase is about getting familiar with the data, obtain knowledge about the data before preprocessing it. After characterising the data, the next step is the preprocessing. Knowing basic statistics regarding each attribute, in the profiling phase, makes it easier to fill in missing values, smooth noisy values, and spot outliers during data preprocessing. Quantile plots, histograms, and scatter plots are graphic displays of basic statistical descriptions that can be useful during data preprocessing and can provide insight into areas for mining.

4.1.2 Selecting Spatial and Temporal Criteria. Multimodal pattern analysis can be conducted at different spatial and temporal granularities. In terms of spatial specifications, two major possibilities can be considered. One of them is to manually specify the target geographical region of interest using a polygon or a circular marking facility. The other, is to select predefined regions. We propose 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.
- Administrative zones: coarsest geographical unit for the city, it can range from municipalities to parishes, depending on the geographical organization of the target city.
- Sections: finest geographical unit, comprising small districts and neighborhoods.

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.

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. Multimodal patterns can be represented 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. 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.

4.1.3 Consolidating Traffic Data. Once these constraints are fixed, data mappings are applied to transform the retrieved spatiotemporal data structures into georeferenced multivariate time series structures[17]. These time structures can be aggregated at different granularities and DTW averaging can be applied for a more consistent analysis. Inspired by the work of Santos (2020), who studied the correlation between the demand for bicycles in Gira stations

and the weather[**?**], correlation between time series of different modes of transport can aid into understanding multimodal synergies, by using linear correlation coefficients (e.g. Pearson's, Spearman's or Kendall's) and also detrended cross-correlation analysis for correlating time series under non-stationarity.

4.2 Multimodality Index Data Analysis

4.2.1 Desired Properties. The properties of the inequality measures described in [6], must be adapted to the context of transportation before describing the multimodal indices. M. Diana & M. Pirra (2016) reformulate them as follows[8]:

1. *Weak Principle of Transfers*:
   Consider two travel modes whose intensities of use are I and $I\delta$, where $\delta > 0$. If the intensity of the most used mode decreases and that of the least used increases by the same quantity $I < 2\delta$ then the multimodality index should increase.
2. *Scale Independence*:
   If the frequency of use of each mode changes by the same proportion, the multimodality index should remain the same.
3. *Principle of Population*:
   The multimodality index should remain the same for any replication of the modes with their corresponding intensities of use. The choice set of the modes represents our population and 'replicating a mode' can be seen in our context simply as an increase in the population size due to the consideration of an additional number of modes with the same intensities of use of those already in the choice set.
4. *Decomposability*:
   Multimodality rankings of alternative distributions of intensities of use in the whole set of travel modes, should match the multimodality rankings of the corresponding distributions of intensities within any of the subgroups in which the whole set of travel modes can be composed.
5. *Strong Principle of Transfers*:
   Considering the following distance measure
   $$d = h(I_1/I_{total})h(I_2/I_{total}), \tag{16}$$
   for modes 1 and 2, with $I1 < I2$, where I_{total} is the sum of all intensities and h is a decreasing function defined as $h(I) = (1I\beta)/\beta$, with β a parameter. If the intensity of the most used mode I_2 decreases and the one of the least used I_1 increases, the variation of the index depends only on the variation of d. The ratios I_i/I_{total} are the 'intensity shares' of mode i; the larger the share, the more predominant is the use of that mode compared to others. The function h is introduced to decrease the distance, and therefore the effect on the index, when the modal transfer is taking place between two modes that are progressively more predominant, even if the difference in their relative intensity shares is constant.

4.2.2 Candidate Multimodality Indices. The Herfindahl–Hirschman Index is a typical measure of market concentration and is used to determine market competitiveness, as described in Section 3.2. This measure can be adapted to the context of urban mobility, where 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, while the value increases when the use of a smaller number of modes tends to dominate[8]. The index can be defined as follows:

$$HH = \frac{1}{n}\left[\frac{n\sum_{i=1}^{n}(f_i - \overline{f})^2}{(\sum_{i=1}^{n} f_i)^2} + 1\right]. \tag{17}$$

In order to distinguish between the set of available modes and the set of effectively used modes, M. Diana & M. Pirra (2016) proposed a variant of Equation 17 which takes into account only the m elements different from zero (the effectively used modes)[8]:

$$HH_m = \frac{1}{m}\left[\frac{n\sum_{i=1}^{n}(f_i - \overline{f})^2}{(\sum_{i=1}^{n} f_i)^2} + 1\right]. \tag{18}$$

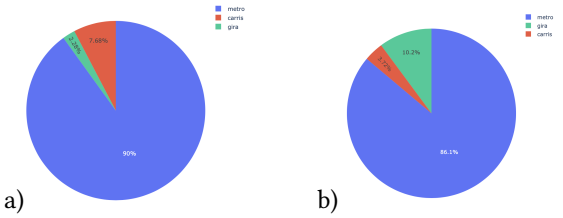
The Gini coefficient or Gini index, is a summary statistic of the Lorenz curve and is usually used as a measure of inequality in a population. It considers the differences among values of a frequency distribution. M. Diana & M. Pirra (2016) translated the usual formulation of the index, presented in Section 3.2, in the context of multimodality where f_i is the intensity of use of ith mode and n the total number of modes, formulated as[8]:

$$Gini = \frac{2}{n}\frac{\sum_{i=1}^{n} i \cdot f_i}{\sum_{i=1}^{n} f_i} - \frac{n+1}{n}. \tag{19}$$

The Gini coefficient ranges from a minimum value of zero to a maximum of one. The former one corresponds to an equal usage of all modes, while the latter refers to an infinite population of modes in which all of them except one are not used (monomodality).

4.3 Incorporating Situational Context

The analysis of multimodality indices can be complemented with 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 6 provides a map of the city with some poles with impact on the city traffic.



a) b)

Figure 2. Weekly mode share distribution of TAZ nº66. a) Week days. b) Weekends.

The combined analysis of the traffic generation/attraction poles maps with the computed multimodality indices, as well as station-route maps, providing a comprehensive and dynamic way of modelling the spatiotemporal distribution of traffic along the city. Additionally, 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.

5 Results

5.1 The Dataset

Three public transport modes were chosen for this study - Carris, Metro and Gira - according to the choice factors presented in Section 4.1. Carris and Metro are the two most used public transport modes and their stations can be found almost in all the Municipality of Lisbon. Gira, the biking sharing system, however, can only be found in the center axis of the city and in the neighborhood of *Parque das Nações*; and the validations during the week are fewer than the other two modes (Figure 2), not fulfilling all the factors in Section 4.1. The *Relevance* and *Accessibility* constraints may not be satisfied, but, compared to the other modes not included in this work, Gira traffic data was easily available and its attributes were relevant for this analysis.

Smart card technology (the VIVA card) was used to gather public transport traffic data. For Carris, the smart card data only monitors entries, estimators of existing validations can be used to infer the exits (not in the scope of this research). For the remaining modes of transport, Metro and Gira, we have access to both passengers' entry and exit records. The entities responsible for the modes Carris and Metro, allowed the use of their data from the month of October 2018, whereas for Gira, the records range from 13 December 2018 to 31 December 2018.

The data retrieved from Carris buses' smart card readers, is modeled as tabular data with 9865446 rows. Each row represents a smart card validation, which means that a passenger as entered the bus. Since the passengers can pay the bus ticket either with the smart card or with cash by buying a ticket at a Carris partner or directly to the bus driver, there is a portion of the validations that is not take into account. Yet, that small number of validations is not significant enough to influence the analysis, so that fact can be ignored. Figure 3 displays a sample (five rows) of the validations data gathered by the smart card readers.

	Date/Time	NºFleet	Route	Variant	Nº Plate	Nº Trip	Direction	Stop	NºSerial	NºCard	Description	Title Code	Stop ID	Designation
0	2018-10-26 18:26:12	179	32B	0	1.0	19	CIRC	19	2016706778	2885890.0	*L12 (Normal)	3142.0	89912,00	Esc. D. Dinis
1	2018-10-9 16:56:23	179	32B	0	1.0	24	CIRC	16	2016551987	2871032.0	*L12 (Reformado Pensionista)	3145.0	3606,00	Bela Vista (Centro Comercial)
2	2018-10-29 16:27:08	179	26B	0	2.0	14	ASC	7	2468590403	4001496.0	LX/BT-16	31373.0	79104,00	Rot. Oliveiras
3	2018-10-13 15:45:41	179	26B	0	1.0	18	ASC	6	2458020629	13492.0	LX/BT-08 (418/sub23(A))	NaN	79106,00	Av. Peregrinação
4	2018-10-28 08:53:08	179	31B	0	1.0	6	DESC	4	2461817389	1942792.0	*L12 (Normal)	3142.0	3715,00	Av. Paulo VI (Igreja)

Figure 3. Sample of Carris smart card's validations data.

Metro data presents its data in the form of an origin-destination matrix (Figure 4). Contrary to Carris, Metro subway is equipped with two smart card readers, one at the boarding and the other at the alighting, thus, the data retrieved is organized in two matrix (entries and exits). The rows and columns are labeled with the Metro stations, and each cell has a numerical value describing the number of validations (entries or exits) from those stations, in a time range of 15 minutes. Some cells contain missing values which will be filled with 0's.

		Linha de Oc	Nº viagens Linha Azul												
Linha de Oc	Estação de	1/4 hora/Es	Reboleira	Amadora E	Alfornelos	Pontinha	Carnide	Colégio Mil	Alto dos Mo	Laranjeiras	Jardim Zoo	Praça de E	São Sebas	Parque	Marquês de	Avenida
Linha Azul	Reboleira	06: 00 - 15														
		06: 15 - 30				8	1	5	3	3	3	5	7		13	1
		06: 30 - 45			3	4		3	1	4	3	1	4	2	9	4
		06: 45 - 60			2	15	3	27	3	2	13	9	12		23	5
		07: 00 - 15		2	5	18	5	15	5	3	9	4	11	6	38	7
		07: 15 - 30			5	14	1	19	8	1	27	8	22	10	56	11

Figure 4. Sample of Metro smart card's validations data (entries).

For the Gira bike sharing system, the collected data is modeled by tabular data with a total of 88747 rows (Figure 5). The bike stations update the number of available bikes (*num_bicycles*) and the number of empty bike slots (*num_empty_stations*), in their records, every time a bike has been picked-up or dropped-off.

	station_id	date	state	num_bicycles	num_empty_stations	num_stations
0	468	2018-12-13 18:44:57	1	4	17	21
1	468	2018-12-13 18:43:22	1	5	16	21
2	468	2018-12-13 22:19:50	1	5	16	21
3	468	2018-12-14 01:19:49	1	5	16	21
4	468	2018-12-14 13:00:14	1	4	17	21

Figure 5. Sample of Gira stations' data.

Additionally to the information about Carris, Metro and Gira, other types of urban data are available. We have access to the localization of the stations of every mode, in geographic coordinates, but also the localization of traffic generation poles for the city of Lisbon (i.e. commercial areas, enterprises, hospitals, schools and stadiums), as well as major city events such as large concerts, congresses and soccer matches.

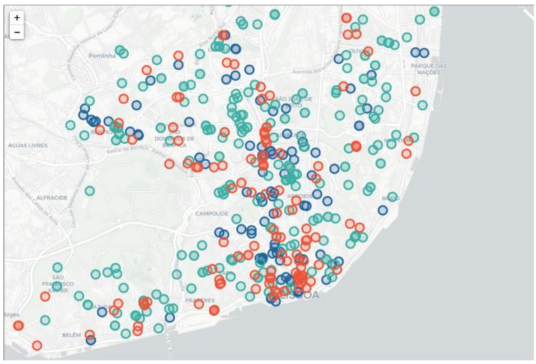


Figure 6. Major traffic generation poles: commercial (blue), schools and institutes (green), and health centres (red).

5.2 Public Transport Data Analysis

5.2.1 Spatial Granularity. The assessment of temporal data requires beforehand the specification of a spatial granularity. Among all the possibilities of spatial criteria defined in Section 4.1.2, we opted by choosing the Traffic Analysis Zones (TAZ) of Lisbon. The TAZ are not administrative divisions, these form of spatial modelling is derived from trip generation densities processed by delineation algorithms that use the peaks of densities as the centre of a zone[18]. Figure 7 illustrates all the 103 TAZs of the Municipality of Lisbon. Only eleven TAZs can be considered for analysis since these are the only TAZs that enclose stations from all three chosen modes of transport (Figure 8).

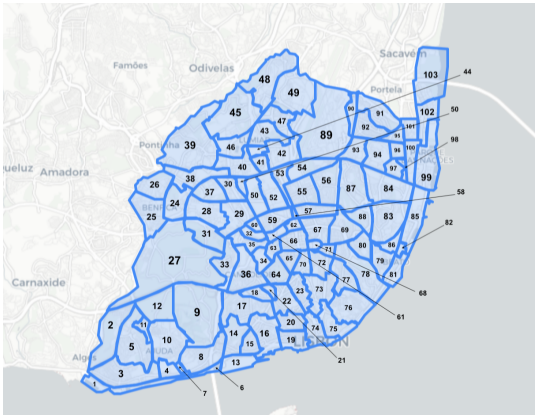


Figure 7. TAZs of the Municipality of Lisbon.

The location of the TAZs and parishes that contain stations of the three modes, Carris, Metro and Gira, reveals that the intermodal interfaces are located in the central axis of the city, and although there are other modes of transport besides those we are analyzing here, this demonstrates the urban planning priority given to the city's commercial zone and the high-density residential zone, and the lack of options in terms of public transport for the population in the medium and low-density resident zone. But these facts are already predictable

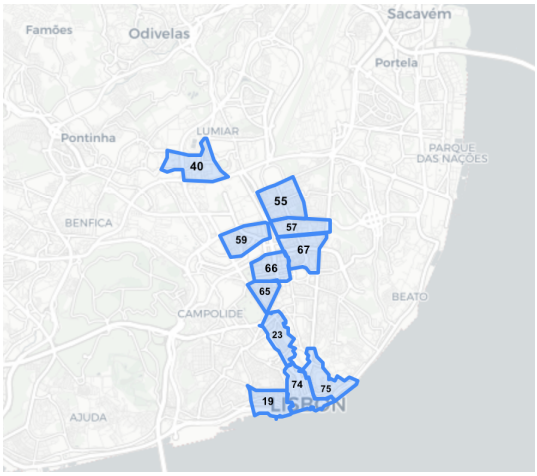


Figure 8. TAZs with three modes of transport (Metro, Carris, Gira).

considering the market share of each mode (Figure **??**), which is highly irregular.

5.2.2 Temporal Series. Among all the Traffic Analysis Zones, the TAZ of *Avenidas Novas (Avenidas Novas — Este)* (TAZ n°66), was the chosen geographical zone for analysis. There's no particular reason that we chose that TAZ. However, its a zone enclosing stations from all three public modes of transport under study; and secondly, its a zone adjacent to *Saldanha*, which is an influential multimodal interface encompassing multiple modes of transport and characterized by the presence of business and cultural traffic generation poles.

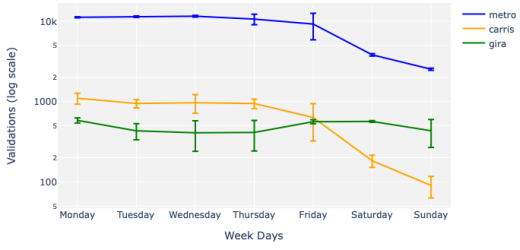


Figure 9. Weekly volume and variation of validations in TAZ n°66.

Since we are studying the behavior of urban multimodal demand, only the check-ins are considered (in the case of Metro and Gira, cause Carris only validates entries). Figure 9 shows the volume and variation of validations in TAZ n°66 during a week. In terms of volume, Metro has roughly ten times more validations than the other two modes. To better respond to the skewness towards large values, we applied

a logarithmic scale for the Y axis. Metro and Carris have a similar behavior during the week, however the dispersion of values for Metro is minimal, where for Carris, the dispersion of values is slightly more substantial. The observations during the week for these two modes are approximately constant during the week, and then at Friday it starts to decrease till the end of the weekend. Gira demand behavior is relatively the opposite: its decreasing during the week days, and it increases during the weekend. The standard deviation for Gira during the weekend is low, where during the week days, the dispersion of values is significant. This difference between Gira and the other two modes, can be explained by the fact that bicycles are mostly used as leisure and sport transport, and not so much as transport for commuting (home-work), like Carris and Metro, which can also explain the high dispersion of values during the week. Another cause could be he fact that the bicycle is more accessible to the younger population, as it's a human-powered mode of transport, while the Carris and the Metro are used by people of all ages.

5.2.3 Barycenter Averaging. The data used in the analysis carried out in the previous section was processed using the mean of validations per station, causing to be sensitive to noisy data[21]. A more accurate procedure is applied in this section, where euclidean averaging and DTW barycenter averaging (DBA) is applied to the sequences of each mode. These sequences or time series are built from each week of the available date range. For example, Carris data is available for the month of October (2018), so the DBA is computed from four sequences (i.e. a cluster) corresponding to the validations of the weeks of that month. Figure 10 shows the weekly volume of validations through barycenter averaging in TAZ nº66, where there are no significant differences in trend of the series, compared to the previous analysed plot in Figure 9. However it's interesting to notice that in the days where there is a high variability of the validations, the euclidean sequence is affected by it and deviates from the DBA sequence. Thus, it enables to visualize that the difference between the demand in the last day of the week (Friday) and the first day of the weekend (Saturday) is much more distinct with this type of plots.

5.2.4 Correlations. The final task of the public transport data analysis is to correlate the data of the available transport modes. Figures 11 and 12 show respectively the Pearson's, Spearman's and DXA correlation coefficients between the different modes weekly[3]. There is not a major difference between the used coefficients, the correlation varies in the same way between modes. Already perceived with the previous analysed line plots, but now we can confirm: Carris and Metro entry validations are highly correlated, whereas those volumes are negatively correlated with those from Gira.

[3]Results were computed for daily correlations, but they were very similar to the weekly correlations, so they were not included.

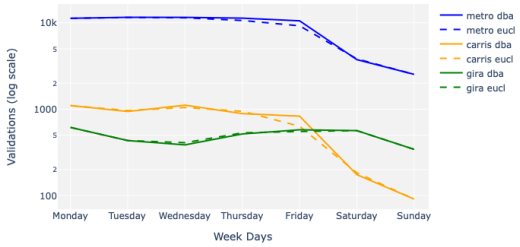


Figure 10. Barycenter averaging (DTW and Euclidean) for weekly volume of validations in TAZ nº66.

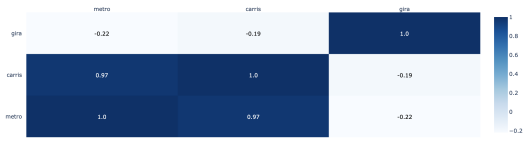


Figure 11. Weekly Pearson correlation heatmap between modes of TAZ nº66.

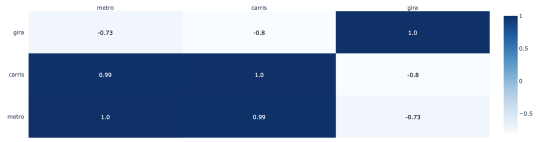


Figure 12. Weekly DCCA correlation heatmap between modes of TAZ nº66.

5.3 Multimodality Indices Analysis

The previous section analysed multimodality in one zone. This section will now assess multimodality among all available geographical units (TAZs), in order to have a global vision of multimodality at the spatial level of the city of Lisbon. Among all candidate multimodality indices presented in Section 4.2.2, the Gini coefficient and the Herfindahl–Hirschman index were the measures selected for this analysis, because they are simple to implement and to adapt to the context of transportation and they differ in their properties but range between the same values. Figure 13 displays four TAZ maps of Lisbon coloured with the values of the Gini index (green is 0 corresponding to multimodality and 1 is red corresponding to monomodality) at different hours. There isn't a major change in the index values among these hours. The TAZs with a higher degree of multimodality correspond mostly to the TAZs containing all the three modes (bus, subway and cycling) (see Figure 8) and encompass a large number of traffic generation poles (see Figure 6). Still there are TAZs containing all selected modes, but they have a medium index value (around 0.5); this is due to the fact that the Gini

index is sensible to the intensity of usage of the modes (scale dependence, see Section 4.2.1). The results for the Herfindahl–Hirschman index (HH) are similar to the ones of the Gini index, with the exception that the HH index is highly sensible to the number of used modes (principle of population, see Section 4.2.1), that justifies the color red over almost all TAZs, since most of the TAZs out of the center only have two or one mode of transport (only including bus, subway and cycling).

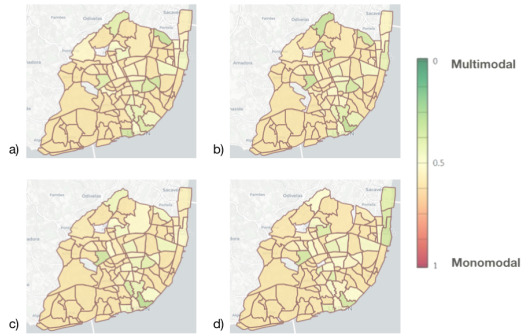


Figure 13. Gini index TAZ map (week days). a) 8h. b) 12h. c) 17h. d) 21h.

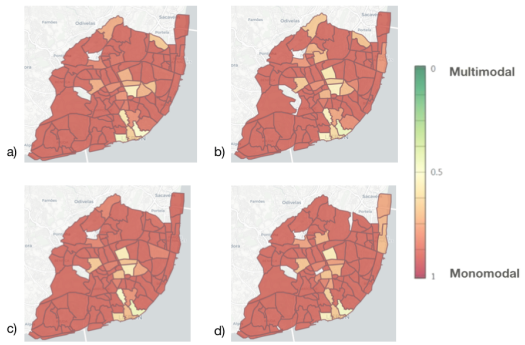


Figure 14. HH index TAZ map (week days). a) 8h. b) 12h. c) 17h. d) 21h.

The variation of both indices was plotted throughout the week (Figures 15 and 16). And for both indices, the variation is almost identical. Until the middle of the week (Wednesday) the indices values rise, then start to decrease until Saturday, and at Sunday they slightly increase. The standard deviation was calculated for both plots, however, the variability was too large to be included in the plots, which suggest that the weekly variation of the indices may not be significant enough to be considered in the analysis. This means that if the Y axis had coarser value ticks, both plots would look like a straight line.

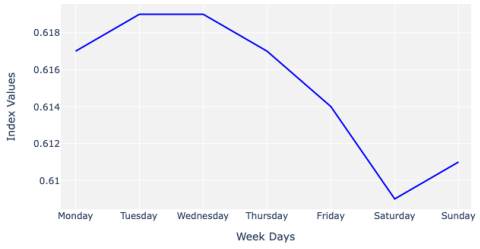


Figure 15. Weekly Gini index lineplot.

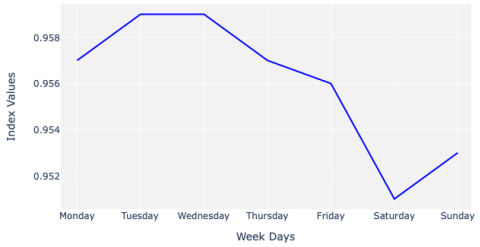


Figure 16. Weekly HH index lineplot.

5.4 Software Tool

The ILU project contributions are integrated in the ILU App web platform. The application was developed in Python with Dash, which is a framework for building machine learning and data science web applications, powered by Plotly[4]. This work contributes to the task *Descrição: Padrões de Mobilidade Urbana* in the application's home page. It contains two pages, one for the multimodal pattern analysis and other for the assessment of the multimodal indices.

6 Conclusion

6.1 Concluding Remarks

This Dissertation offered a structured view of the multimodal synergies from heterogeneous sources of urban data. To the best of our knowledge, the literature only focused on the measurement of multimodality at the individual level, in other words, it studied to what degree people (or public transports passengers) were multimodal. So this thesis aimed at studying multimodality at the spatial level, in order to aid specialists in the field of transportation engineering and urban planning. The inherent sensitivity of the properties of the used indices, is greatly influenced by the number of modes and their intensity, which could be attenuated if the other public transport operators shared their traffic

[4]Plotly: https://plotly.com/dash/.

data, providing the possibility to analyse more transport modes and to obtain a more extensive and complete view of the multimodal patterns of the city. The data received from Carris, Metro and Gira was also incomplete, relying only on information from the months of October and December 2018.

6.2 Community and Scientific Acceptance

The research pursued for this dissertation is being subject to international reviews through the submission of work to peer-review international Conferences in the field. the article – "Exploring multimodal mobility patterns with big data in the city of Lisbon", was submitted to the scientific committee of the 48th European Transport Conference (ETC 2020)[5] held online last September[15]. This latter article was presented by the former author at the ETC 2020 "Young Researchers' and Practioners' Forum" at the session "Mobility" and it received a very positive feedback from the international scientific community. It was also submitted to the European Transport Research Review Journal linked to the ETC 2020.

6.3 Future Work

This research presented an innovative method to evaluate traffic patterns, through the analysis of spatial multimodality data. The Traffic Analysis Zones were chosen as the geographical unit under study but other spatial granularities could equally be suggested. With more data, other types of temporal granularities could also be added to the research such as months and years. And, the use of the other mentioned inequality measures could provide more insights on the degree of spatial multimodality in the city. The multimodal data can also be correlated with other types of situational context, such as weather or public events (e.g. concerts), in order to evaluate the factors that influence variations in multimodality on the city. And finally, this study was applied to the city of Lisbon but would be equally interesting to explore multimodality behaviors in other cities.

References

[1] Atkinson, A. B., et al. On the measurement of inequality. *Journal of economic theory 2*, 3 (1970), 244–263.

[2] Brezina, I., Pekár, J., Čičková, Z., and Reiff, M. Herfindahl–hirschman index level of concentration values modification and analysis of their change. *Central European journal of operations research 24*, 1 (2016), 49–72.

[3] Brockwell, P. J., Davis, R. A., and Calder, M. V. *Introduction to time series and forecasting*, vol. 2. Springer, 2002.

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

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

[6] Cowell, F. *Measuring inequality*. Oxford University Press, 2011.

[7] Diana, M., and Mokhtarian, P. L. 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 (2009), 107–119.

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

[9] Esling, P., and Agon, C. Time-series data mining. *ACM Computing Surveys (CSUR) 45*, 1 (2012), 1–34.

[10] Gini, C. Variabilità e mutabilità (variability and mutability). *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955) ed. Bologna* (1912).

[11] Han, J., Pei, J., and Kamber, M. *Data mining: concepts and techniques*. Elsevier, 2011.

[12] Herfindahl, O. C. *Concentration in the steel industry*. PhD thesis, Columbia University New York, 1950.

[13] Khurshid, S., Singh, R., and Singh, G. Levels and trends of competition among mutual funds in india. *Research Journal of Business Management 3*, 2 (2009), 47–67.

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

[15] Lemonde C, Arsénio A, H. R. Exploring multimodal mobility patterns with big data in the city of lisbon. *Association for European Transport* (2020).

[16] Lorenz, M. O. Methods of measuring the concentration of wealth. *Publications of the American statistical association 9*, 70 (1905), 209–219.

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

[18] Martínez, L. M., Viegas, J. M., and Silva, E. A. A traffic analysis zone definition: a new methodology and algorithm. *Transportation 36*, 5 (2009), 581–599.

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

[20] Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., and Keogh, E. Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. *Knowledge and Information Systems 47*, 1 (2016), 1–26.

[21] Petitjean, F., Ketterlin, A., and Gançarski, P. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition 44*, 3 (2011), 678–693.

[22] Podobnik, B., and Stanley, H. E. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Physical review letters 100*, 8 (2008), 084102.

[23] Sen, A. *On Economic Inequality*. Oxford University Press, 1973.

[24] Weiner, J., and Solbrig, O. T. The meaning and measurement of size hierarchies in plant populations. *Oecologia 61*, 3 (1984), 334–336.

[25] Zar, J. H. Spearman rank correlation. *Encyclopedia of Biostatistics 7* (2005).

[5] https://aetransport.org