Caminha etal. *Journal of Internet Services and Applications* (2018) 9:22
https://doi.org/10.1186/s13174-018-0094-3

Journal of Internet Services and Applications

RESEARCH Open Access

Graph mining for the detection of overcrowding and waste of resources in public transport

Carlos Caminha*, Vasco Furtado, Vládia Pinheiro and Caio Ponte

Abstract

The imbalance between the quantity of supply and demand in public transport systems causes a series of disruptions in large metropolises. While extremely crowded vehicles are uncomfortable for passengers, virtually empty vehicles generate economic losses for system managers, and this usually comes back to passengers in the form of fare increases. In this article a new data processing methodology will be presented for the evaluation of collective transportation systems. It proposes the construction and mining of graphs that represent complex networks of supply and demand of the system to find such imbalances. In a case study with the bus system of a large Brazilian metropolis, it was shown that the methodology in question is capable of identifying global imbalances in the system based on an evaluation of the weight distributions of the edges of the supply and demand networks. It has also been shown that even in a scenario where information about the demand is incomplete, using community detection techniques it is possible to identify the stretches of the network that are potentially causing these imbalances on a global scale.

Keywords: Complex networks, Graph mining, Human mobility

1 Introduction

The era of digital information has created the means to generate knowledge from diverse and voluminous data and apply it effectively [1]. One of the areas that best exemplifies this context is urban mobility. Sensors and digital media record information daily on vehicle routes and the increasing availability of such data in open data portals enables studies to be carried out with the potential to be applied to the improvement of the mobility of people [2].

An example of a largely sensorized system and one, which has also increasingly been opened up by governments, is the bus system of large metropolises [3–5]. With electronic devices powered with GPS (Global Positioning System), which record where vehicles travel, and ticketing systems, which register passengers' payments and fares, databases are growing bigger every day [6] and a challenge is to extract information that has the potential to identify overcrowding.

*Correspondence: caminha@unifor.br
Programa de Pós Graduação em Informática Aplicada, Universidade de Fortaleza, Fortaleza, Ceará, Brasil

The problem of overcrowding occurs when vehicles exceed their recommended maximum capacity, however, in this article, we are also interested in situations where these buses are practically empty. In both cases it is possible to note dissatisfactions on the part of actors involved in the context of the bus system of a large metropolis. While crowded buses irritate public transport users, empty buses are detrimental to system managers.

Over the years, numerous studies have tried to analyze the supply and demand of bus systems [7–9] and there is a consensus that overcrowding is a problem that generates dissatisfaction for a large number of people, thus it is important to identify buses with this characteristic. The dynamics of the system, where people get on and off vehicles at all times, as well as the absence of sensors that accurately identify these arrivals and departures, make it difficult to analyze the data and specifically prevent us from knowing about the occupancy of the vehicles at each moment of the day. In other words, the challenge is to compare the supply of vehicles, which are fully sensorized by GPS devices, with their demand, which is usually only a sample of the use of the system. The lack of complete knowledge about the demand for bus systems is essentially

Springer Open

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

due to the fact that ticketing systems usually do not have sensors that record the passengers' ascent and descent, with only one of these two being registered moments. For this reason, it is often necessary to use heuristic methods to find the missing information [10, 11]. Another challenge observed when evaluating overcrowding in collective transport networks is that the complexity of the connections of these systems makes it difficult to perceive bottlenecks (stretches of the network with crowded buses). In this type of system it is common to evaluate weak supply connections as possible bottleneck points, however, only a topological evaluation of the subcomponents surrounding these edges can confirm the existence of the problem [12].

Given this problem, a strategy that is recommended for the understanding of this type of phenomenon is the use of complex networks. The potential of this type of instrument to treat problems that require the abstraction of aspects such as ignorance of part of the data, together with the wealth of metrics and algorithms already in existence in the state of the art, justify their application to solve a series of problems. A significant amount of work is currently being done in modeling networks for understanding complex systems in the context of air [13], rail [14], urban [5] and bicycle [15] transport. In general, these studies aim to characterize networks using metrics such as weight distribution, average route length and cluster coefficient. The work of [4], for example, compared the public transport system in 22 cities in Poland. Among other characteristics, the authors showed that the distribution of degrees and weights at the edges follow a hierarchically organized power law function, $Y = aX^{\alpha}$, where X is the degree, Y is a frequency of the degree in the distribution, a is a constant and α is the power law exponent [16, 17].

This paper proposes a data processing methodology that assesses the imbalance between supply and demand in transportation systems. Complex networks and mining in graphs are applied to find bottlenecks (places where supply is insufficient) or resource waste (stretches of the network where buses travel almost empty). An open source software was developed that implements this methodology and it was validated in a case study with real GPS data and ticketing of the bus system of Fortaleza, a Brazilian city with a population of over 2.5 million inhabitants. In this case study, a process of characterizing the weights of the edges of its supply and demand networks has revealed a global imbalance in the city's public transportation system, including the surprising finding that demand grows disproportionately in relation to the growth of the bus service supply. We also used community detection algorithms to identify bus lines that have overcrowding problems. In this way, the main contributions of our work are:

(a) A data processing methodology for evaluating collective transportation systems.
(b) A case study of the application of a methodology in a large metropolis.
(c) Open source software that implements this methodology via an algorithm based on complex network metrics.

In addition to this introduction, this article is composed of five more sections. In Section 2, the state of the art, and the main works and concepts that surround this research are presented. In Section 3, the characteristics of the studied city are explained, as well as the datasets used. Section 4 describes the proposed data processing methodology. Sections 5 and 6 discuss our experience of applying the methodology in a case study with the city of Fortaleza, in these sections our main findings in this research are described.

2 State of the art

The problem of the allocation of public transport systems has been extensively studied over the years [18–23]. Specifically, in 1991, Chang and Schonfeld [7] developed a model for the analysis of demand elasticity (the ability to increase and decrease over a day) considering supply characteristics. Several characteristics of the system were studied with the objective to construct the ideal transport system, in other words, a system with low fares, a small fleet, without convoy (buses of the same line running together), a balance between the cost of the management and cost of the user and finally, without crowded buses. In spite of the impact of the work, the limited access to the data of the time made it impossible to fully verify the model with real data of supply and demand.

Hasan et al. (2013) [21] considered data obtained from transactions using an intelligent subway travel card to characterize the patterns of urban mobility. Bottlenecks of the subway network structure were studied and a mobility model was presented that predicted the locations visited by the people using the popularity of the places of a great metropolis. This model was sufficient to reproduce several characteristics of observed travel behavior, such as the number of trips between different places in the city, the exploration of new places, the frequency of individual visits to a particular location and the number of people using a particular part of the network at the same time. In spite of the good predictability, the model in question does not point to the specific lines that cause bottleneck (overcrowding), in that only the section of the network where several lines pass was indicated.

Toole et al. (2015) [24] estimated the demand for travel in different cities around the world. The authors estimated source and destination matrices using mobile telephone data and reconstructed individual patterns of mobility by evaluating cases of overcrowding in the public roads of

the cities studied. Despite the strong relationship with the work presented in this paper, Toole et al. evaluated supply and demand at a macro level, in other words, the authors were interested in evaluating the supply and demand as a whole in a city, considering all modes of transport (private cars, buses and subways). This type of approach is substantially different from a micro approach, which can focus on only one mode of transport, with the potential, for example, of finding out where there is discomfort or waste of resources in this mode.

Even more recently, Caminha et al. (2016) [3] used information from GPS systems and ticketing to propose micro-interventions in a bus supply network. The authors evaluated the supply based on the demand measured from origins and destinations of bus users. Micro-interventions (creation of express lines) were suggested for regions with large passenger flows. Despite the consistent results, the work in question did not focus on a deep assessment of the relationship between supply and demand of bus systems, essentially because demand is estimated only from origins and destinations, in other words, the route that the users traveled on the buses was not considered.

Saberi et al. (2017) [25] used complex networks to characterize two origin and destination (OD) networks. The OD networks of Melbourne and Chicago were evaluated and compared. In the networks in question each node represented a location of the city (origin or destination of a citizen) and each edge informed that a citizen traveled from one node of origin to another of destination. Numerous complex network metrics have been calculated to assess the similarity between the mobility demands of cities. The authors concluded that networks of travel demand in these two cities exhibited similar properties, despite significant differences in topography and urban structure.

Considering the survey carried out in this section, it was possible to advance the state of the art by proposing a data processing methodology for evaluating collective transportation systems that can detect imbalances between supply and demand in these systems, even in situations where information about the demand was only partially available. It is supposed that the modeling of supply and demand as complex networks and the use of graph algorithms available in the literature may reveal patterns that indicate the locations in the system where these imbalances are more detrimental to the actors involved.

3 Characteristics of the city considered in the experiments

In total, Fortaleza has 4783 bus stops served by 2034 buses that run on 359 different routes (bus lines). Each bus in the city is equipped with a GPS system which registers the bus position at intervals of approximately thirty seconds. On a weekday, about four million geographic coordinates are recorded for the buses in Fortaleza.

The integrated transportation model adopted by the Town Hall of Fortaleza, named Bilhete Único (Single Ticket), allows registered users to make a free bus transfer anywhere in the city, as long as it is within two hours of the last validation of the travel card. The validation process of the Single Ticket is understood as the act of the user sliding his card into the electronic reader located near the bus turnstile or the turnstile of a bus terminal. Generally, this procedure happens at the beginning of the trip, since the turnstile is near the entrance of the bus in Fortaleza. In one weekday, there are on average 1.2 million validations of the Single Ticket. As can be seen in Fig. 1a, there are three characteristic peaks of validations during a working day in Fortaleza, the first occurring around 7 a.m. in the morning, the second near lunchtime, around 12 p.m.

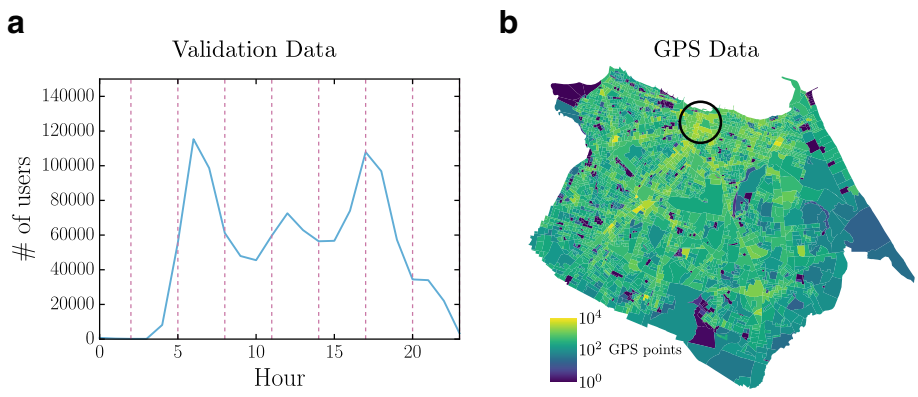

Fig. 1 Visualization of information contained in the analyzed data sets. **a** number of validations of the Single Ticket during a weekday. The blue line shows the number of Single Ticket validations per hour on a weekday, and the vertical red dashed lines show three hour periods. **b** Density map of bus GPS points in the city of Fortaleza. The number of GPS points in the 3043 census tracts of the city was calculated, the colors were attributed from the logarithmic scale of the GPS point density

and, finally, the last occurs at the time when most of the population returns home, specifically at 6 p.m.

Figure 1b shows a map of density of GPS points recorded by buses in Fortaleza on March 11, 2015. The map was calculated from a division of census tracts. Fortaleza has in total 3043 census tracts, distributed throughout the city, which is 313 km^2. The regions with the highest density of points, the areas shown in yellow, are basically the locations of terminals and the commercial center of the city, and are highlighted by the black circle in Fig. 1b. Within one business day, approximately 4 million GPS points are recorded in the city of Fortaleza's database.

The information mentioned in this section is available in three datasets: datasets of bus stops; GPS; and ticketing of buses (Single Ticket) related to the bus network of Fortaleza. All referring to the day of March 11, 2015, a Wednesday, a normal working day in the city. These data have been extensively studied in recent years [3, 26–30] and is available for download on the website of the city hall of Fortaleza [31].

Finally, even if it is only a day of data, recent studies have shown that there is no significant difference in the supply and demand of the Fortaleza bus network between working days [3, 26]. A significant change in the functioning of the system is observed only on Saturdays, Sundays and holidays. However, because of the substantial decrease in demand during these days, there are no problems of extreme occupation. As a consequence, the evaluation of these days is not a priority.

4 Data processing methodology

The proposed methodology aims at finding imbalances between supply and demand of a bus system and is composed of three phases. In the first phase, data from services offered on the Web or from open data portals are used and two networks (graphs) are constructed, one of which represents the bus system supply and the other represents demand. In the second phase distributions are analyzed relating to the weights of the edges of these networks, the main objective in this phase is to find clues that indicate global imbalances in the bus system. These clues do not point to bus lines that are crowded or empty, but reveal whether the resources spent by the bus system management is allocated proportionally to the needs of the population precisely, and if there is more supply where there is more demand for buses, which is desirable in balanced systems. Finally, in the third phase, the use of community detection algorithms is proposed to find bus lines that are system bottlenecks (crowded bus lines) and lines where there is a waste of resources (often empty buses). Algorithm 1 shows a pseudocode of the entire methodology. To simplify the application of the methodology proposed in other cities, we provide a open source software[1] that implements the three steps mentioned in this paragraph.

Algorithm 1 Finding imbalances between supply and demand in a bus system

```
S ← read supply network                                   ▷ Phase I
D ← read demand network
for all edge e_S of S and e_D of D do                     ▷ Phase II
    if e_S and e_D have the same node of source and target then
        records weight of e_S and e_D
    end if
end for
plot histogram of recorded weights
runs Louvain Modularity Method to find the communities    ▷ Phase III
selects the intercommunal edges and identifies all bus lines that pass through them
```

Algorithm 2 Louvain Modularity

```
G is the original network
end ← false
repeat
    a different community is assigned for each network node   ▷ First phase
    while some nodes are moved do
        for all node n of G do
            node n is placed to community for which the gain is maximal
        end for
    end while
    if the new modularity is higher than the prior modularity then   ▷ Second phase
        G ← construct a new network whose vertices are communities found during the first phase
    else
        end ← true
    end if
until end = true
```

The first phase of the methodology consists of the construction of the supply and demand networks of the transportation system. The supply network is constructed as proposed in [3], in which a bus supply network is represented as a directed graph $G(V, E)$, with vertices, v ($\in V$), representing the stops and edges of buses, e ($\in E$), between two bus stops. Formally the weight of the edge, $w_{v_i \to v_j}$, represents the bus supply between two bus stops,

v_i and v_j ($\in V$). This supply is calculated from the sum of the weight of bus lines, w_{L_k}, which pass through two bus stops. Formally $w_{v_i \to v_j} = \sum_{k=1}^{N} w_{L_k}$, where N is the total number of bus lines that visit, in turn, the stops v_i and v_j in their itinerary. Thus, the weight of the bus lines is calculated from the product of the number of vehicles allocated on the line L_k, V_{L_k}, by the number of trips each vehicle makes on a day in L_k, C_{L_k}. That is $w_{L_k} = V_k C_k$.

The demand network is built as proposed by Caminha et al. (2017) [26], who defined a heuristic method to estimate the real path made by users of the bus system. Formally, the bus demand network is a directed graph $G(V, E)$, where V and E are the sets of vertices v and edges e, respectively. An edge e between the vertices v_i and v_j is defined by the ordered pair (v_i, v_j). In this network, the vertices represent bus stops and the edges represent the demands of bus users between two consecutive bus stops. Each edge, e, which binds an ordered pair (v_i, v_j) is defined as a weight, $w_{e_{ij}}$, which sums up the total number of users that passed through e. In all, this network has the same 4783 vertices and 5876 edges that are in the supply network.

The only aspect that differs between the supply and demand networks and the weight of their edges, is that in the supply network this weight is measured by the quantity of vehicles that offer routes and in the demand network the same is measured by the number of people that have passed inside the buses. In this way, both networks have the same distribution of degrees. It is also worth mentioning that the demand network represents only a sample of the need of the population, even though [26] argue that the sample represents 40% of the city's bus demand, one of the biggest samples of demand of the bus system that is known. The authors further showed that the extracted sample is free of any spatial bias.

The second phase of the methodology consists of the evaluation of the distribution of the edges weights in order to identify global imbalances in the system. By evaluating the differences between these distributions it is possible to identify overcrowding problems in the system. Beyond that, the characterization of these distributions may reveal different manners to improve the system through public policies. For instance, the existence of a Pareto law [32] in the demand indicates that micro-interventions in small parts of the transport system may significantly improve the balancing of the system as a whole [3]. We suggest here the visualization of these distributions by a *Probability Density Function* and *Cumulative Density Function*, in which the variables in study must be the weights for the edges of the supply and demand networks.

At last, the third phase of the methodology proposes an evaluation of the modularity of the supply and demand networks in order to find bottlenecks within the system. For the supply network this may indicate the existence of areas where the bus supply is privileged (within communities) and also where there are bottlenecks (weak edges) between these areas. The algorithm of Blondel et al. (2008) [33] was used to estimate the modularity and the communities of the supply and demand networks. This algorithm makes use of a heuristic method based on the optimization of the graph modularity. The modularity of a set of vertices is measured by a real value between 0 and 1 and is calculated from the relation between the quantity of edges that connect the elements of the set to each other by the total edges of the set of vertices [34, 35]. The algorithm is divided into two phases that are repeated iteratively. In the first phase a different community is assigned for each network node. Thus, in this initial division there are as many communities as there are vertices. For each node i each one of its neighbors j is considered, the gain of modularity as i moves from its community to the community of j is evaluated. The node i is then placed in the community for which this gain is maximal, but only if that gain is positive. If no positive gain is possible, node i remains in its original community. This process is applied to all vertices repeatedly until no improvement can be achieved and thus the first phase is completed. The second phase of the algorithm consists of the construction of a new network whose vertices are communities found during the first phase. To do this, the weights of the connections between the new vertices are given by the sum of the weight of the connections between the vertices in the two corresponding communities. The connections between the vertices of the same community are represented by self-references. The phases are repeated until the maximum global modularity is found. The Algorithm 2 show the pseudocode of the community detection algorithm used in the propose methodology.

The methodology proposed in this section has linear spatial and temporal complexity. More precisely, in phase 1, the complexity is $O(E)$, where E is the amount of edges to be explored to built the graphs. The same complexity is obtained in phase 2, due to the need to iterate in the edges only once to compute the distribution of the weights of the edges. Finally, phase 3 also has linear complexity as a function of the number of edges. This complexity is achieved due to the execution of the community detection algorithm of Blondel et al. (2008) [33].

5 Analysis of the distributions of edge weights to find global imbalances in bus systems

In this section the process of characterizing the supply and demand networks of the Fortaleza bus system will be described. To allow comparisons, the edge weights of the two networks were normalized. In the demand network, for each edge, e_d, the weight, w_d, which measures the passenger demand between two bus stops, was normalized

by the ratio w_d/w_{dmax} where w_{dmax} is the largest weight recorded in the network. For the supply network the same strategy was used to normalize the weight, which in this case represents the number of available buses. It should be pointed out that both networks have the same distribution of degrees, that is, for each edge in the supply network there is a corresponding edge in the demand network, these networks differ only in relation to their edge weights. The networks have 4783 vertices and 5876 edges. In this article, we used the network constructed by Caminha et al. (2016) [3] to represent the Fortaleza bus supply. Figure 2 illustrates the nodes and edges dispositions in the studied networks.

Figure 3 illustrates the distribution of edge weights for the two networks. In a we observe a power law [16, 17] with exponent α = -2.90 for the supply network. This result suggests that this network was designed to offer a high number of resources in a few places and few resources in various locations. Likewise, a power law can be observed for the demand network in Fig. 3b. The exponent α = -1.97 indicates that there are few places with high demand and many places with low demand.

Although the distributions are similar in scale, this does not necessarily imply that the networks are balanced. The first suggestion of imbalance is the difference between the exponents, the volume of concentration in a few places is much more intense in the supply network. Figure 3c illustrates the difference in slope of the estimated lines. The orange line illustrates the estimated regression in the supply network and the yellow line is the estimated regression for the distribution of the edge weights of the demand network.

Moreover, in regard to the distribution of edge weights in the networks, Fig. 3d shows the cumulative distribution of these data. The yellow curve shows that the network of demand accumulates more edges with small weights than the supply network (orange curve), the probabilities are equal in the two networks and edges with a weight of 0.16. This is another suggestion of imbalance due to the differences between the probabilities accumulated in both networks.

There was also a need to analyze the relation between supply and demand of the bus system considering the modifications that it undergoes during different periods in a day. Each edge weight corresponds to an interval of three hours on the day. We have six supply/demand networks for one day. This interval was chosen due to the fact that the principal peaks of demand occur within this range of three hours as can be seen in Fig. 1a. For this reason, supply and demand networks were generated for the following time intervals: 5:01 to 8:00; 8:01 to 11:00, 11:01 to 14:00; 14:01 to 17:00; 17:01 to 20:00 and from 20:01 to 23:00. No networks were generated for the early morning

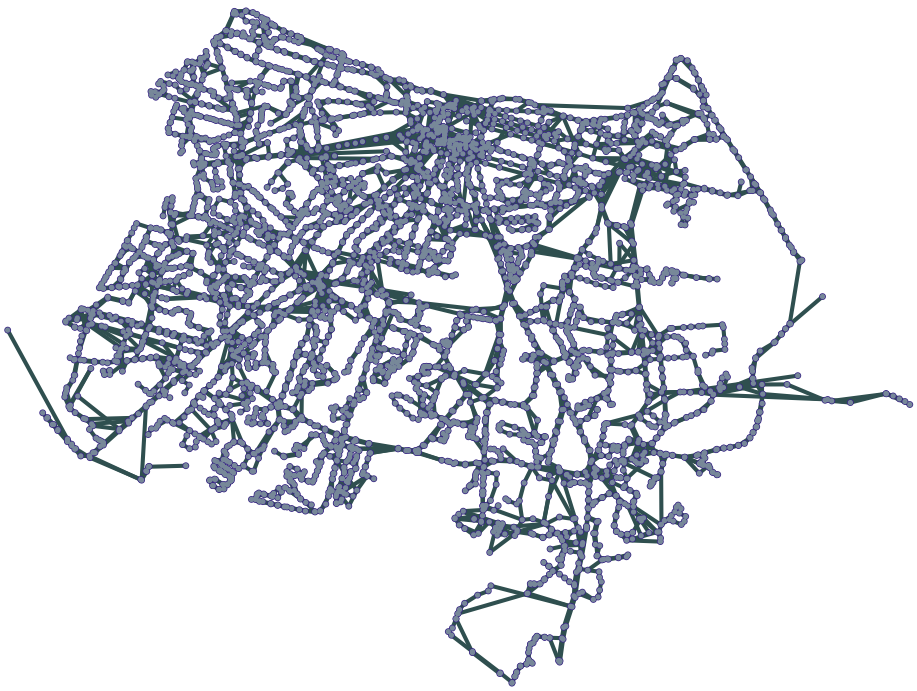

Fig. 2 Arrangement of the graph vertices representing the supply and demand of the Fortaleza bus system. The vertices represent bus stops and the edges connect possible routes within the bus network. The vertices were geo-referenced for better visualization

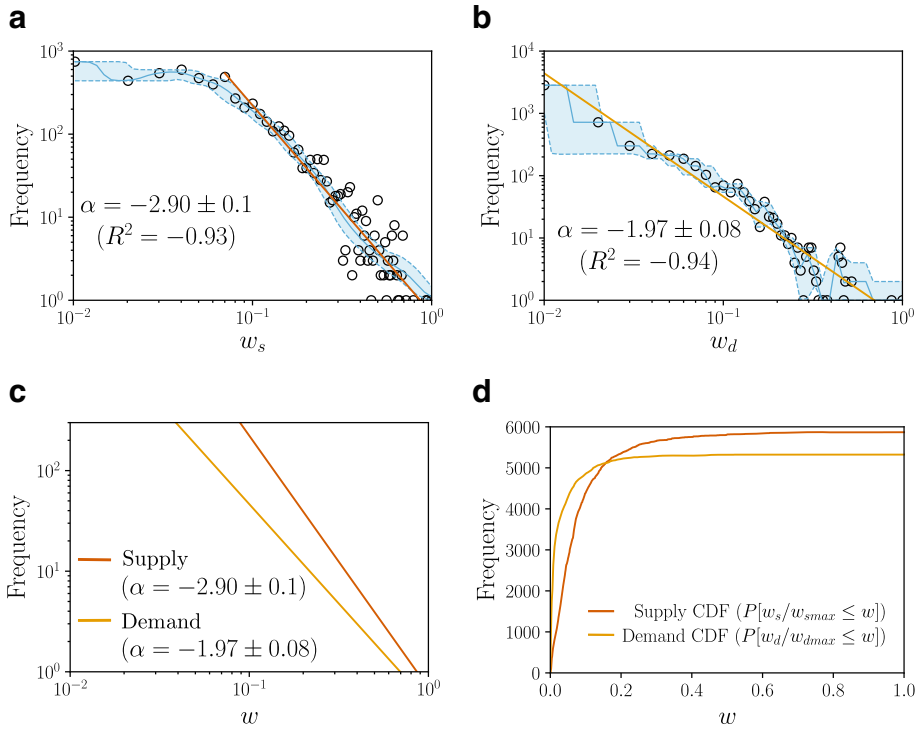

Fig. 3 Distribution of weights of the edges of the supply and demand networks. **a** the weight distribution of the edges of the supply network is illustrated. In **b** the same distribution is shown for the demand network. The straight lines represent linear regressions [36, 37] applied to the data, the blue contoured lines represent the Nadaraya-Watson method [38, 39] and the blue dashed lines delimit the estimated 95% configuration interval by bootstrap [40, 41]. Both distributions were generated in 50 bins. In **c** the inclinations of the lines estimated by the regressions in both districts can be compared in the same view. Finally, in **d**, we show the cumulated distributions of the edge weights of the supply and demand networks

period due to the lack of use of the network in that period.

In balanced networks, where supply grows in direct proportion to the growth of demand, it is expected that there will be an isometric relation between supply and demand volumes, [42]. An isometric relation between two variables is characterized by the proportional growth, of one variable according to the other [43, 44]. In other words, in the specific case of this work, we verify, for instance, if whenever the supply doubles in one specific region of the city the demand doubles too. Figure 4 shows the correlation between the weights of the edges of the supply and demand networks at three-hour intervals from 5:00 a.m. to 11:00 p.m. on March 11, 2015. In all the figures, each black circle represents an edge of the bus network. The x-axis illustrates the weight of edges in the supply network and the y-axis is the weight of edges in the demand network. The red line is the regression applied to the data. The blue dashed lines illustrate confidence intervals estimated with the Nadaraya Watson method [38, 39], the data are presented in a logarithmic scale. Assuming a better equation adjustment of type $Y = aX^{\beta}$, we found the exponent $1.36 \leq \beta \leq 1.46$ at different moments of a working day in the city, revealing a superlinear allometric growth ($\beta > 1$) [26, 45–49] of the demand as a result of the supply of the Fortaleza bus system. This superlinear allometric growth indicates that demand grows disproportionately in stretches of the network where the supply is greater, in other words, in the bus system of Fortaleza, when bus supply doubles, demand more than doubles.

The analyzes of the weight distributions of the edges of the two networks indicate a global imbalance of the bus system. In particular, the identification of a non-isometric relationship reveals the existence of stretches of the network with bottlenecks and waste of resources. These stretches are possibly responsible for the global imbalance of the system. In the following section of this article an approach will be presented to find the bus lines that are causing this unbalance.

6 Detection of communities to find bottlenecks and waste of resources

Based on the analysis of the edge distributions of the supply and demand networks, it was necessary to verify their modularity, in other words, we needed to verify if the networks have sub-networks with a high coefficient of grouping and low connectivity with other sub-networks. In order to compute the modularity of the supply and

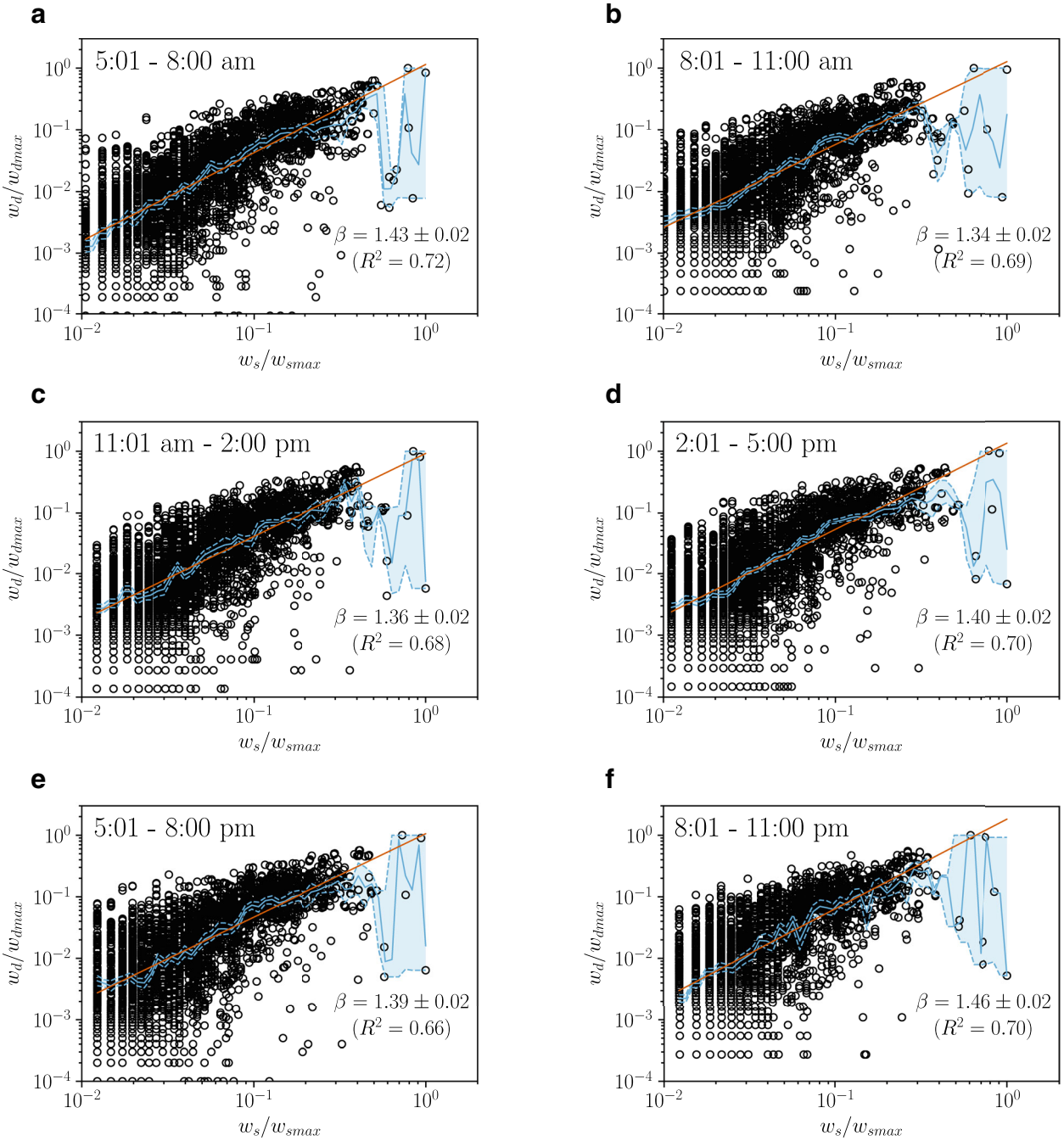

Fig. 4 Correlation between supply and demand of the bus network. The supply and demand of the Fortaleza bus system are correlated. It is possible to observe the correlations in the intervals: 5:01 to 8:00; 8:01 to 11:00, 11:01 to 14:00; 14:01 to 17:00; 17:01 to 20:00 and 20:01 to 23:00, respectively, in items (a-f). In all of the figures in this panel, the red lines represent linear regressions [36, 37] applied to the data, the blue contoured lines represent the Nadaraya-Watson method [38, 39], and the blue dashed lines delimit the interval of confidence 95% estimated by bootstrap [40, 41]

demand networks it was used the Blondel et al. (2008) [33] algorithm.

Figure 5 depicts the relation between the number of communities and their respective modularities for the supply and demand networks. The orange circles mark the modularity of the supply network. The yellow symbols mark the modularity of the demand network. The modularity supply network is larger than the demand network in virtually all configurations. This result reveals that the supply network seems to be designed to cater for shorter

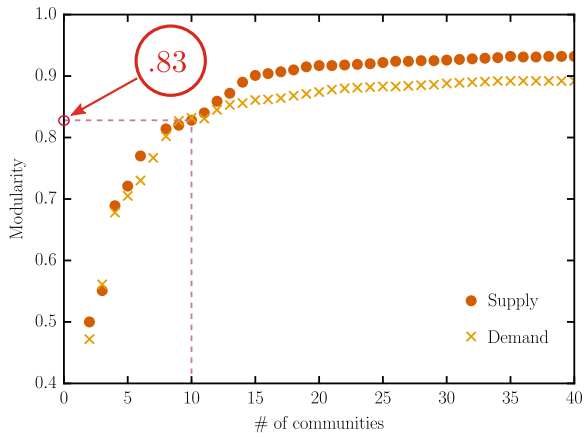

Fig. 5 Relation between the number of communities generated and their respective modularities in the supply and demand networks. The supply network has more modulation than the demand network in configurations with a larger number of communities. When the networks are divided into exactly 10 communities, their modularity converges to ≈ 0.83

movements than what in fact people need. The modularities of networks tend to converge when they are divided into fewer communities, with a greater approximation of values when the networks are divided into exactly ten communities, with a modularity of ≈ 0.83.

The modularity values found in the supply network indicate that the edges within the communities of this network are stronger than the edges found in the communities of the demand network, possibly because the users of the bus system travel longer stretches than expected (in the supply network), increasing the weight of local edges and preventing community detection algorithms from finding configurations with the level of modularity found in the supply network.

In order to identify the bus lines that are overloaded, lines that have a greater demand than they can support, and bus lines where there is loss of resources, which are lines that have a very large supply for their demand, the configurations with ten communities of both networks were used. These configurations have ≈ 0.83 of modularity and the algorithm used [33] removed 241 edges of the supply network and 380 edges in the demand network in the detection of the ten communities. An overload index and a waste index were verified to identify the problematic stretches in the network.

In Fig. 6a there are illustrated stretches of the supply network where it is overloaded, these stretches are traversed by bus lines that are possibly with crowded vehicles. These stretches were identified from the edges removed by the method of detecting communities proposed by Blondel et al. (2008) [33]. A weak edge removed (by the method of detecting communities) from the supply network can mean a supply bottleneck, since this edge connects intercommunal movements that have already been proved to be predominant [3]. In all the edges that were removed, the overload index was calculated $IS = w_s/w_{smax} - w_d/w_{dmax}$. Edges with $IS < 0$ are sections where supply is greater than demand, and potentially where bottlenecks are. It was also observed that bus lines that have a bottleneck usually take people to the city center (highlighted by the red circle).

In Fig. 6b, network stretches are shown where there are potentially a waste of resources, i.e., empty buses. These stretches were identified from the edges removed by the method of detecting communities, now in the demand network. Community detection methods remove weak

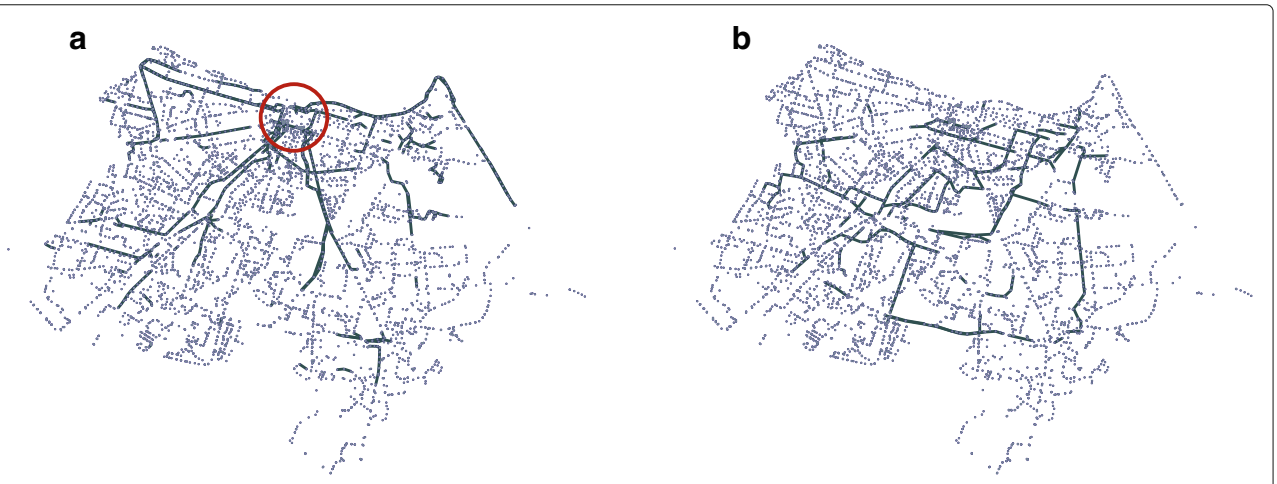

Fig. 6 Excerpts from the network where bottlenecks and waste of resources were found. Highlighted edges of the network are highlighted in **a** where the routes of lines of buses that have the potential to be overloaded was observed. In **b** we can see stretches of the network where there is potentially a waste of resources

connections between components with high cluster coefficient, which in the case of the demand network are sub-regions of intense passenger flow. In this context, the edges eliminated by the method in question represent stretches from the network of low passenger movement, where potential resource waste may occur. In a similar way to what was done in the supply, we calculate the waste index $ID = w_d/w_{dmax} - w_s/w_{smax}$, where edges with $ID < 0$ represent segments where proportionally demand is greater than supply, and consequently where resources are being wasted.

The proposed index, *IS*, identified that the lines 503, 029, 014, 605, 725, 379, 087, 325, 316, 013, 130, 815, 030, 077, 011, 650, 755, 361, 612, and 075 have stretches in their itineraries where the vehicles are crowded, while *ID* identified that lines 394, 315, 024, 220, 605, 709, 087, 013, 050, 311 and 317 are at some point in their itinerary virtually empty. While validating on location with professionals from ETUFOR (Urban Transport Company of Fortaleza) we were informed that it is known that on some days of the week the lines 503, 029, 014, 605, 725, 379, 087, 325, 316, 013, 130, 815, 030, 077, 075 and 011 are crowded in the highlighted stretches. As for the lines 650, 755, 361 and 612, the professionals were surprised that they were full, and informed us that they would be evaluated in future actions. Regarding to the lines with few passengers the professionals chose not to comment. This was justified by the fact that it is difficult to know if a line has been used by few passengers at a certain time, since this fact does not generate any complaints from the users.

7 Conclusion

This work explored the supply and demand of the bus system in Fortaleza, a large Brazilian metropolis. Algorithms and metrics of complex networks were used to explore the imbalance between what public office offers and what the population needs. At the macro level, the differences between the supply and demand networks in their respective distributions of edge weight density, and the cumulated distributions of edge weight and levels of modularity revealed a global imbalance in the system. It has been found that the networks of supply and demand of public transport have highly modular communities. The existence of this structural pattern motivated us, at the micro level, develop a model that makes use of community detection techniques to identify where the bottlenecks are and where resources are being wasted. This model has shown promise even in situations where it is not possible to gain complete information on passenger demand.

As a main result for the city, the proposed data processing methodology identified locations of the public transport system with either crowded buses or virtually empty buses. With this information, the city hall can balance the system, reducing supply in places where there are virtually empty vehicles and leading to places where the lines are full. This information can bring economic gains, since the literature shows that a good distribution of the supply in collective transport systems may even lower the price of the bus ticket [50].

As to future work, a need was identified to build a simulator [51] capable of reproducing the dynamics of human mobility, through the bus system, in a large metropolis, making use of data mining to estimate functions of probability which represent the actual demand for a bus system. With this simulator, it will be possible to validate the application of balancing techniques of the supply network, testing its equilibrium with the demand network. At the same time, it will be possible to synthetically alter demand in order to fit it in with supply, in a real scenario, the change in demand can be achieved through the creation of public policies, such as the building of terminals at optimal points for transfers, education campaigns in the use of the system and micro-interventions in the urban space (e.g., tax incentives for the creation of employment opportunities in areas where supply is underutilized).

Endnote

[1] Available in https://github.com/caiocponte/Graph-mining-for-the-detection-of-overcrowding-and-waste-of-resources-in-public-transport

Abbreviations

ETUFOR: Urban transport company of Fortaleza; GPS: Global positioning system; OD: Origin and destination

Acknowledgements

We thank the anonymous reviewers whose valuable comments and suggestions have significantly improved the presentation and the readability of this work.

Funding

VF is financed by CNPQ grants 454899/2014 and 307803/2014

Availability of data and materials

All data used in this study are available in http://dados.fortaleza.ce.gov.br/catalogo/dataset/dados-de-onibus-11-03-2015. The scripts used to analyze the bus system are available at https://github.com/caiocponte/Graph-mining-for-the-detection-of-overcrowding-and-waste-of-resources-in-public-transport.

Authors' contributions

All authors contributed equally. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 November 2017 Accepted: 16 August 2018
Published online: 15 November 2018

References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. The kdd process for extracting useful knowledge from volumes of data. Commun ACM. 1996;39(11): 27–34.
2. Zheng Y, Li Q, Chen Y, Xie X, Ma W-Y. Understanding mobility based on gps data. In: Proceedings of the 10th international conference on Ubiquitous computing. New York: ACM; 2008. p. 312–21.
3. Caminha C, Furtado V, Pinheiro V, Silva C. Micro-interventions in urban transportation from pattern discovery on the flow of passengers and on the bus network. In: Smart Cities Conference (ISC2), 2016 IEEE International. Trento: IEEE; 2016. p. 1–6.
4. Sienkiewicz J, Hołyst JA. Statistical analysis of 22 public transport networks in poland. Phys Rev E. 2005;72(4):046127.
5. Munizaga MA, Palma C. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. Transp Res C Emerg Technol. 2012;24:9–18.
6. Sagiroglu S, Sinanc D. Big data: A review. In: Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE; 2013. p. 42–47.
7. Chang SK, Schonfeld PM. Multiple period optimization of bus transit systems. Transp Res B Methodol. 1991;25(6):453–78.
8. Oppenheim N, et al. Urban travel demand modeling: from individual choices to general equilibrium. New York: Wiley; 1995.
9. Domencich TA, McFadden D. Urban travel demand-a behavioral analysis. Tech Rep. 1975.
10. Gordillo F. The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation. PhD thesis, Massachusetts Institute of Technology. 2006.
11. Hua-ling R. Origin-destination demands estimation in congested dynamic transit networks. In: Management Science and Engineering 2007, ICMSE 2007. International Conference on. Harbin: IEEE; 2007. p. 2247–52.
12. GAO Z-Y, Wu J-J, Mao B-H, Huang H-J. Study on the complexity of traffic networks and related problems. Commun Transp Syst Eng Inf. 2005; 2:014.
13. Wang J, Mo H, Wang F, Jin F. Exploring the network structure and nodal centrality of china's air transport network: A complex network approach. J Transp Geogr. 2011;19(4):712–21.
14. Lenormand M, Tugores A, Colet P, Ramasco JJ. Tweets on the road. PloS ONE. 2014;9(8):e105407.
15. Hamon R, Borgnat P, Flandrin P, Robardet C. Networks as signals, with an application to a bike sharing system. In: Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE. Austin: IEEE; 2013. p. 611–4.
16. Barabási A-L, Albert R, Jeong H. Scale-free characteristics of random networks: the topology of the world-wide web. Physica A Stat Mech Appl. 2000;281(1):69–77.
17. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. SIAM Rev. 2009;51(4):661–703.
18. Gerland HE, Sutter K. Automatic passenger counting (apc): Infra-red motion analyzer for accurate counts in stations and rail, light-rail and bus operations. In: 1999 APTA Bus Conference, Proceedings. Cleveland: TRB; 1999.
19. Yahiaoui T, Meurie C, Khoudour L, Cabestaing F. A people counting system based on dense and close stereovision. Image Sig Process. 2008;5099:59–66.
20. Boyle DK. Passenger counting technologies and procedures. Washington: TRB; 1998.
21. Hasan S, Schneider CM, Ukkusuri SV, González MC. Spatiotemporal patterns of urban human mobility. J Stat Phys. 2013;151(1-2): 304–18.
22. Yildirimoglu M, Kim J. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. Transp Res C Emerg Technol. 2018;89:254–67.
23. Hamedmoghadam-Rafati H, Steponavice I, Ramezani M, Saberi M. A complex network analysis of macroscopic structure of taxi trips. IFAC-PapersOnLine. 2017;50(1):9432–7.
24. Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, González MC. The path most traveled: Travel demand estimation using big data resources. Transp Res C Emerg Technol. 2015;58:162–77.
25. Saberi M, Mahmassani HS, Brockmann D, Hosseini A. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. Transportation. 2017;44(6):1383–402.
26. Caminha C, Furtado V, Pequeno THC, Ponte C, Melo HPM, Oliveira EA, Andrade Jr JS. Human mobility in large cities as a proxy for crime. PloS ONE. 2017;12(2):e0171609.
27. Caminha C, Furtado V. Impact of human mobility on police allocation. In: Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on. Beijing: IEEE; 2017. p. 125–7.
28. Sullivan D, Caminha C, Melo H, Furtado V. Towards understanding the impact of crime in a choice of a route by a bus passenger. Porto: Springer; 2017. arXiv preprint arXiv:1705.03506.
29. Ponte C, Caminha C, Furtado V. Busca de melhor caminho entre dois pontos quando múltiplas origens e múltiplos destinos são possíveis. Recife: ENIAC; 2016.
30. Furtado V, Furtado E, Caminha C, Lopes A, Dantas V, Ponte C, Cavalcante S. A data-driven approach to help understanding the preferences of public transport users. Boston: IEEE; 2017, pp. 1926–35.
31. Dados abertos F. http://dados.fortaleza.ce.gov.br/catalogo/dataset/dados-de-onibus-11-03-2015. Accessed Jan 6 2017.
32. Rosen KT, Resnick M. The size distribution of cities: an examination of the pareto law and primacy. J Urban Econ. 1980;8(2):165–86.
33. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10): P10008.
34. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci. 2002;99(12):7821–6.
35. Wilkinson DM, Huberman BA. A method for finding communities of related genes. Proc Natl Acad Sci. 2004;101(suppl 1):5241–8.
36. Rawlings JO, Pantula SG, Dickey DA. Applied regression analysis: a research tool. New York: Springer Science & Business Media; 2001.
37. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. New York: Wiley; 2015.
38. Nadaraya EA. On estimating regression. Theory Probab Appl. 1964;9(1): 141–2.
39. Watson GS. Smooth regression analysis. Sankhyā: Indian J Stat Ser A. 1964;26:359–72.
40. Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. J Econ. 2004;119(1):99–130.
41. Li Q, Racine J. Cross-validated local linear nonparametric regression. Stat Sin. 2004;14(2):485–512.
42. Banavar JR, Damuth J, Maritan A, Rinaldo A. Supply–demand balance and metabolic scaling. Proc Natl Acad Sci. 2002;99(16):10506–9.
43. Operti FG, Oliveira EA, Carmona HA, Machado JC, Andrade JS. The light pollution as a surrogate for urban population of the us cities. Physica A Stat Mech Appl. 2018;492:1088–96.
44. Melo HPM, Moreira AA, Batista É, Makse HA, Andrade JS. Statistical signs of social influence on suicides. Sci Rep. 2014;4:6239.
45. Kleiber M, et al. The fire of life. an introduction to animal energetics. The fire of life. An introduction to animal energetics. New York: Wiley; 1961, pp. 1–454.
46. Bettencourt LMA, Lobo J, Strumsky D, West GB. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. PloS One. 2010;5(11):e13541.
47. Makse HA, Havlin S, Stanley HE. Modelling urban growth patterns. Nature. 1995;377(6550):608.
48. Oliveira EA, Andrade Jr JS, Makse HA. Large cities are less green. Sci Rep. 2014;4:4235.
49. Melo HPM, Moreira AA, Batista É, Makse HA, Andrade JS. Statistical signs of social influence on suicides. Sci Rep. 2014;4:6239.
50. Currie G, Wallis I. Effective ways to grow urban bus markets–a synthesis of evidence. J Transp Geogr. 2008;16(6):419–29.
51. Santana EFZ, Bastista DM, Kon F, Milojicic DS. Scsimulator: An open source, scalable smart city simulator. In: Tools Session of the 34th Brazilian Symposium on Computer Networks (SBRC). Salvador: SBRC; 2016.