Journal of Internet Services and Applications

RESEARCH Open Access

Using bundling to visualize multivariate urban mobility structure patterns in the São Paulo Metropolitan Area

Tallys G. Martins[1], Nelson Lago[1], Eduardo F. Z. Santana[1], Alexandru Telea[2], Fabio Kon[1] and Higor A. de Souza[1]*

*Correspondence:
hamario@ime.usp.br
[1]Department of Computer Science, University of São Paulo, São Paulo, Brazil
Full list of author information is available at the end of the article

Abstract

Internet-based technologies such as IoT, GPS-based systems, and cellular networks enable the collection of geolocated mobility data of millions of people in large metropolitan areas. In addition, large, public datasets are made available on the Internet by open government programs, providing ways for citizens, NGOs, scientists, and public managers to perform a multitude of data analysis with the goal of better understanding the city dynamics to provide means for evidence-based public policymaking. However, it is challenging to visualize huge amounts of data from mobility datasets. Plotting raw trajectories on a map often causes data occlusion, impairing the visual analysis. Displaying the multiple attributes that these trajectories come with is an even larger challenge. One approach to solve this problem is trail bundling, which groups motion trails that are spatially close in a simplified representation. In this paper, we augment a recent bundling technique to support multi-attribute trail datasets for the visual analysis of urban mobility. Our case study is based on the travel survey from the São Paulo Metropolitan Area, which is one of the most intense traffic areas in the world. The results show that bundling helps the identification and analysis of various mobility patterns for different data attributes, such as peak hours, social strata, and transportation modes.

Keywords: Bundling, Urban mobility, Data visualization, Travel surveys, Smart cities, Open data

1 Introduction

Governments around the world are providing open datasets about city resources, such as citizen surveys, data collected with the help of Internet of Things devices, accountability and transparency reports, and so on. The Internet facilitates the gathering and the availability of a large variety of data sources, which allows citizens, researchers, companies, non-profit organizations, and public agents to perform analyses on their matters of interest regarding city-related issues. In the public administration realm, these data sources can be used to provide better services for citizens, to improve the management of urban infrastructure, or to reduce bureaucracy costs, supporting evidence-based policymaking [1, 2].

Springer Open

Urban mobility is a great concern for citizens and governments. It directly affects people's quality of life, causes prejudices to the environment, and has a high economical impact. For example, the traffic congestion in the São Paulo Metropolitan Area (SPMA) is estimated to affect 89% of work-related commuting trips[1], causing monetary losses of seven billion Brazilian reals (∼US$1.8 billion) every year [3]. Thus, the development of more efficient transportation systems is a critical issue that cities should tackle.

Several data sources can be used for urban mobility analysis such as data captured with IoT devices such as traffic cameras, GPS tracking, bike-sharing systems, as well as censuses and trip surveys [4, 5]. In the SPMA, every ten years since 1967, the São Paulo Metropolitan Company (Metrô), which manages the subway system in the city of São Paulo, conducts a travel study called *Origin–Destination (OD) survey*. This survey is performed by interviewing citizens about their life and commuting activities on a typical working day, resulting in a comprehensive panorama of the mobility behavior of the population over the SPMA. The last OD survey (2017) shows that there are around 42 million trips over the 24 hours of a regular working day. Beyond a trip's origin (O) and destination (D) data itself, the survey also gather a wide number of trip-related and socioeconomic aspects, or data attributes, such as transportation modes used, trip reasons, age, gender, and household income. Hence, the OD survey generates a large and multivariate dataset.

While the above-mentioned OD survey is very comprehensive and accurate, urban planners need proper tools to analyze this large amount of multivariate data. Spreadsheets and statistics programs help producing tables and charts with aggregated information showing, *e.g.*, the number of users of the public transportation system over the years or the number of women *vs* men that commute for work. Visual techniques, such as density maps [6], can help to answer geolocated data-related questions, such as finding regions that concentrate most of the mobility flow during the day or the most common origin–destination pairs. However, considering the huge amount of information (and attributes) that is produced every day in any city, translating geolocated data into meaningful images is very challenging.

A recent survey about traffic visualization methods indicates the common use of *line-based* visualization techniques to study the structure of mobility data, which often includes pattern discovery and clustering tasks [7]. Still, drawing OD lines can only handle thousands of trajectories. Visualizing 42 million trajectories, as recorded daily in the last OD survey, would only generate a fully cluttered image (as can be seen in Fig. 5). *Bundling techniques* improve upon line-based techniques and can depict high volumes of trajectory data by essentially simplifying trajectories in the image space, grouping spatially close and data-wise similar trajectories together [8, 9].

In this paper, we use bundling to visualize the structure and patterns of urban mobility in the São Paulo Metropolitan Area (SPMA). For this, we adapt the recent CUBu framework [10] – Compute Unified Device Architecture (CUDA) bundling, which has proven to be useful in other scenarios using large movement datasets (e.g., flight traffic, eye tracking) [11–13]. CUBu generates different bundling styles, grouping trips per distance, density, or direction. However, to the best of our knowledge, CUBu has not yet been used on OD data of sizes comparable to the SPMA dataset. We adapt CUBu to use bundling to find and visually explore mobility patterns created by different combinations

[1] We use the terms trip, travel, trail, and trajectory interchangeably in the paper

of the attributes present in the OD survey, such as traffic on peak hours, trips per social strata, different transportation modes, and trips made for distinct reasons. We show how the bundled visualizations help identify different structural patterns of urban mobility in São Paulo, providing insights into the data characteristics that were analyzed.

This paper extends an earlier work presented in the IV Brazilian Workshop of Urban Computing (CoUrb 2020) [14] with more detailed analyses of the findings obtained by exploring the bundled visualizations of the SPMA data as well as a better coverage of related work and more detailed explanation about the techniques we used. It is organized as follows. Section 2 outlines related work with a focus on OD data visualization, trajectory bundling, and investigation of mobility patterns. Section 3 presents the main characteristics of the SPMA and the OD survey data. Section 4 describes our visualization approach. Section 5 presents the results of our analysis and discusses the use of bundling in urban mobility visualization. In addition to the original explorations presented in [14], Section 5 shows the visual exploration of several other attribute combinations: density per social strata (Section 5.5), mobility of young students per social strata (Section 5.6), directions at peak hours (Section 5.7), density by transportation mode (Section 5.8), and trip distance per trip reasons (Section 5.9). Section 6 concludes the paper.

2 Background and related work

We next discuss related work on trajectory bundling with a focus on OD data and studies that investigate mobility patterns in urban scenarios.

2.1 Trajectory bundling

Trajectories (or for short, trails) are typically sets of spatial points $\mathbf{t} = \{\mathbf{x}_i\}$ recorded at consecutive time moments t_i that describe the motion of an object over time. Origin–Destination (OD) data are particular cases of trails which contain only the first and last recorded point. Trails can have additional data attributes $\mathbf{a}_j$, recorded either at each sample point, *e.g.*, speed $\mathbf{a}_j(\mathbf{x}_i)$; or globally for the entire trail, *e.g.*, vehicle type $\mathbf{a}_j$. As such, an OD dataset can be seen as a multivariate dataset with trails being the observations and the measured attributes being the dimensions or variables. Trail data can be directly plotted atop of 2D maps, with selected data attributes encoded into visual variables such as color, line thickness, and line style. Besides statically drawing entire trail-sets, these can be also shown by animating particles along their points $\mathbf{x}_i$ [15, 16].

Visualizing trail-sets having thousands of trails or more, each with multiple attributes, is challenging. Simply drawing the trails yields significant clutter which makes finding even the simplest spatial patterns very hard. Particle techniques also do not scale well – with over roughly 50 thousand points, the resulting visual pattern resembles Brownian motion. *Bundling* methods aim to solve this visual scalability issue by grouping trail fragments which are spatially close and, optionally, have similar data attributes. This creates more *overdraw* between (similar) trails, but also generates visually empty space between trail groups that reduces overall *clutter*.

Bundling is used in several other studies on movement data. Graser et al. [17] created a bundled visualization to study the characteristics of bird migrations. Klein et al. [16] developed a dynamic visualization for aircraft flight data where the trail-set itself varies over time. Willems et al. [11] analyzed maritime vessel traffic using bundling techniques. Separately, bundling was used to simplify eye-tracking data to infer reading patterns

[12, 13]. Lhuillier et al. [9] compiled a survey on the state of the art on bundling techniques. This survey distinguishes between bundling graph data (straight-line drawings of graph layouts) and trail data (such as our OD data), highlighting the best methods for both cases.

Image based bundling (IBB) techniques are a particular class. They essentially exploit mean shift clustering [18], well known in image processing. We outline this process for the Kernel Density Estimation Edge Bundling (KDEEB) technique [19], although all other IBB techniques we are aware of work similarly and share the same properties (see also Fig. 1). Let $T = \{\mathbf{t}_i\}$ be the trail set to bundle, and let $S = \{\mathbf{p}_i\}$ be a set of points that densely sample T. First, a density map $\rho : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ is computed from S by kernel density estimation (KDE), *i.e.* by convolving the points in S with an Epanechnikov kernel of radius k. This parameter essentially specifies the scale of visual simplification – trails farther than k units apart will not be bundled together. Hence, ρ will be high in regions of high density of the sampling points – and by implication, of trails – and low elsewhere. Second, the points in S are advected upwards in the density gradient $\nabla \rho$ with a small step. Terminal (O and D) points are kept fixed so that one can still identify the origin and destination of the bundled trails next. Finally, a 1D Laplacian filter is used to smooth the trails in T, thereby removing inherent numerical noise in the advection. The process is repeated for N iterations, whereby k is continuously decreased so as to 'slow down' the advection and make trails converge around the local density maxima. The main advantages of IBB are simplicity of implementation, control over the simplification induced by the bundling (given by the parameter k), and, notably, speed, as image-space operations map well to GPU functionality.

Zeng et al. [20] adapted the KDEEB technique into Road Aware Edge Bundling (RAEB), which constraints the bundles along the road network on which the respective trails are recorded. RAEB was demonstrated on 166K taxi trajectories from New York City. RAEB is the only work that uses bundling with urban mobility data that we know of. However, RAEB requires detailed trails data (not OD data) and the layout of the city road network.

Geometric-based edge bundling (GBEB) [21] partitions the trail-set T into clusters of trails $\mathbf{t}_i$ that are spatially close and, optionally, have similar attributes. Next, a geometric guideline is extracted from each cluster to bundle its trails along. GBEB was demonstrated on two real world OD datasets, US Airlines and US Migrations. Both have just a few thousand trails, being thus orders of magnitude smaller than the SPMA dataset. Moreover, the O's and D's of these datasets are cities on a map. In SPMA, the O's and D's are the start and end points of trails followed by individual persons. The distribution of O's and D's is thus very different: For the US Migrations and Airlines, the O's and

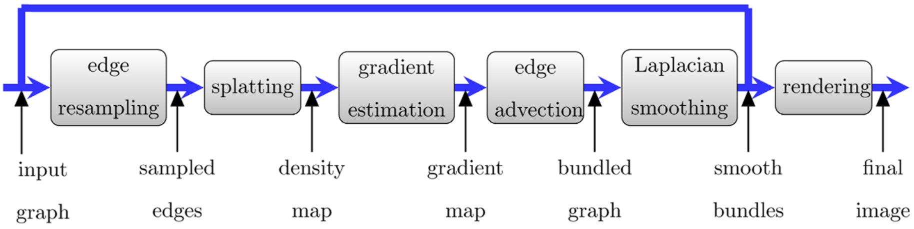


Fig. 1 Pipeline of the KDEEB algorithm

D's are city locations separated by large amounts of whitespace – specifically, 235 O/D's for US Air and 1790 O/D's for US Migrations. Bundling OD trails for such datasets is far easier than for a dataset like SPMA, where the number of O's and D's is massively larger.

Skeleton-based edge bundling (SBEB) [22] follows a similar approach to GBEB by partitioning the trail-set into clusters. Next, each cluster is reduced to its geometric skeleton, or medial axis, which is used to attract trails to form bundles. In addition to US Airlines and Migrations, SBEB was demonstrated on the France Airlines dataset. This dataset has the same characteristics as the first two – its O's and D's are airports in France (a few hundred).

Both SBEB and GBEB rely strongly on their clustering step to partition the trail-set into elongated clusters containing spatially close trails which are quite similar in direction. This works far easier for sparse OD datasets (like the ones discussed above) than for OD datasets where the O's and D's can virtually be anywhere (like the SPMA dataset). Clustering has other challenges too, as it requires careful parameter control. Too coarse clusters will bundle together trails which are far away from each other; too fine-grained clusters will bundle very little, thereby leaving the OD dataset cluttered. Clustering-based bundling methods such as SBEB and GBEB need, also, to control both clustering and bundling parameters. In contrast, in KDE-based methods, one needs to control only a single parameter with a clear geometrical meaning – the KDE radius k discussed above for KDEEB. Finally, both GBEB and SBEB cannot scale computationally to more than roughly 10K trails.

Clustering can be used as a visual simplification means for OD datasets also independent on bundling. One can replace each trail cluster by any suitable (simplified) representation, *e.g.*, a centerline (for an overview of such techniques, we refer to [23]). However, this exposes another problem: Clustering is a *discrete* process, which *explicitly* partitions the trail-set into distinct groups. This can easily lead to false insights in the data groups being strongly separated. In contrast, bundling – and in particular the one created by KDE-based methods – *implicitly* partitions the trail-set into fuzzy groups (the bundles) which can be *continuously* adjusted by the bundling parameters. As such, in cases where cluster identities are not known a priori, like in the case of the SPMA dataset, bundling is preferred to clustering for visual simplification.

CUBu, which stands for CUDA bundling, is an improved implementation of KDEEB [10]. It leverages CUDA[2] to implement all IBB steps (trail sampling, KDE computation, advection, smoothing, and final rendering) on the GPU, thereby surpassing KDEEB and older IBB methods [24, 25] in speed by one to two orders of magnitude. Additionally, CUBu supports several bundling styles, most notably *directional* bundling [24], which separates spatially close trails running in opposite directions in different bundles. Rendering-wise, CUBu supports a large palette of options, such as pseudo-shading to emphasize high-density bundles [26], encoding bundle importance into opacity, and color mapping trail attributes (*e.g.*, direction, time, length).

- *Density map:* We use the underlying density map ρ computed by the KDE process implemented in CUBu to show the local edge density in the produced bundles. This

[2] developer.nvidia.com/cuda-zone

way, one can easily separate visually high-density (important) traffic flows from less important ones;

- *Attribute filtering and coloring:* We use CUBu's ability to color code trail directions or trail lengths to study mobility behavior. We also extended CUBu to filter specific attribute ranges, and select specific combinations of attributes to display. This allows the analyst to search for different types of patterns present in the data;
- *City map:* We modified CUBu to blend its bundled result atop of a map of the SPMA where traffic is analyzed. This allows correlating the bundles with actual locations in the city.

2.2 Mobility patterns in urban areas

Discovering patterns of mobility can be useful to provide relevant information about the local dynamics for public managers, which can propose better public policies using evidence-based information. There are several studies that have investigated patterns of mobility flows in large cities.

Guo et al. [27] used spatial clustering to discover mobility patterns. They used OD pairs from GPS records of taxi trails from the city of Shenzen, China. First, their methodology identifies meaningful potential places through the use of spatial clustering. These meaningful places are clusters that could contain massive flows. After obtaining the clusters, they computed flow measures (e.g., inflows, outflows, net flow) for each cluster in different time periods to map spatial distribution and temporal trends. The mapped distribution is presented as a choropleth map showing city regions that receive or produce more flows in different time periods of a day. As in our work, the data used in [27] is based on OD pair trails. However, the dataset used in their study is related to a single transportation mode, which could present different patterns when compared to other means of transportation. Contrary to their approach, our study does not use the origin and destination places as a weight to search for movement patterns. Instead, bundling uses the proximity of flows to aggregate closer trajectories. Also, our approach does not depend on finding clusters to identify the spatial distribution of the movements.

Moreira and Ceccato [28] investigated space-time patterns of mobility with a focus on gender differences in violent victimization in São Paulo's metro train stations. The data used in this study are official crime records from the São Paulo's police, land-use data from Google Street View images, São Paulo's 2012 OD data, and census data from regions close to the metro train stations. The land-use features considered in this study include commercial establishments, parks, bus stops, number of employees per station, etc. They applied negative binomial regression modeling to identify patterns of mobility and land-use feature factors that differ between men and women. Their results showed that there are different space-time patterns for women and men regarding violent victimization. The approach proposed in [28] was based on a model that was built specifically to assess the violence by gender in the metro train transport mode, including other related data sources. Differently, our study applies bundling in the most recent OD dataset using time, space, and socioeconomic features to generate mobility flow patterns, which can be used to highlight mobility differences for several features available in the OD dataset.

Slovic et al. [29] studied the relationship between job accessibility, infrastructure, and socioeconomic factors in the city of São Paulo. They used data from General Transit

Feed Specification (GTFS) and Automatic Vehicle Location (AVL) from the São Paulo bus system to compute job accessibility. The Municipal Human Development Index (MHDI) was used for the comparison, from which they selected the extreme values (below the 10th percentile and above the 90th percentile). Then, they applied a spatial clustering technique, known as bivariate Local Indicator of Spatial Association (LISA), to measure job accessibility and MHDI. The resulting clusters were plotted in a map used to visualize overlapping patterns through the city. The results showed that areas with lower rates of job accessibility overlap with areas with lower MHDI rates. Contrary to [29], our approach explores multimodal transportation data along with socioeconomic features. In [29], origins (from MHDI data) and destinations (from GTFS and AVL data) are associated through LISA to produce the cluster patterns. In our study, both the origins and destinations came from the same dataset, which is bundled to produce mobility patterns for the whole population and also for distinct social strata.

Moreno-Monroy et al. [30] investigated the relationship between public secondary schools and public transport to assess social inequality access to these schools. They proposed an accessibility index that considers the spatial distribution of adolescents, school locations, and public transport access in school areas. They used the SPMA as their case study. Based on the proposed index, they simulated the impact of the redistribution of these schools on school accessibility. This simulation was based on a non-implemented policy proposed by the São Paulo State Government, which intended to reduce the number of school facilities and keep them more concentrated in central districts. The results showed that the redistribution would have negative impacts on school accessibility. The accessibility index was built using data from the Brazilian census, São Paulo's 2007 OD travel survey, the Brazilian School Census, geocoded public schools, and the Google Distance Matrix API. This study is focused on secondary schools, while our analysis regarding education addresses to identify flow patterns of students between 6 and 18 years of age. We are also using the most recent OD survey, from 2017, in our study. While their analysis is focused on accessibility for public transportation, our study also considers the other means of transportation.

3 Mobility in the São Paulo Metropolitan Area

To provide a better understanding of the scenario of our analysis, this section describes the São Paulo Metropolitan Area and the last travel survey data performed in that region. To avoid misunderstanding between the São Paulo Metropolitan Area (SPMA) and the city of São Paulo, we will refer to the city as capital or simply São Paulo. We will refer to the metropolitan area as SPMA or metro area.

3.1 São Paulo Metropolitan Area

The state of São Paulo is located in the Southeastern coast of Brazil. The São Paulo Metropolitan Area is the most populated region of South America. According to the Brazilian Institute of Geography and Statistics [31], the SPMA has 21.9 million citizens in 2020, which represents around 10% of the Brazilian population. The SPMA is composed of 39 cities in an area of 7946.84 km^2. The city of São Paulo is the capital of the state of São Paulo and it is placed at the center of the metro area. Figure 2 shows the 39 cities that form the SPMA. The capital is the most populated city in Brazil, with 12.3 million inhabitants. In the SPMA, the other cities with more inhabitants are Guarulhos (1.4 million),

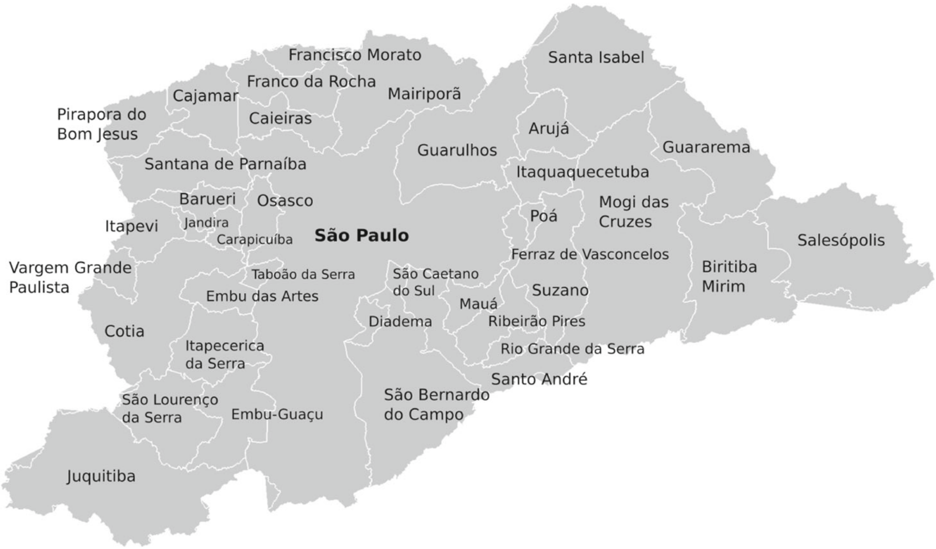


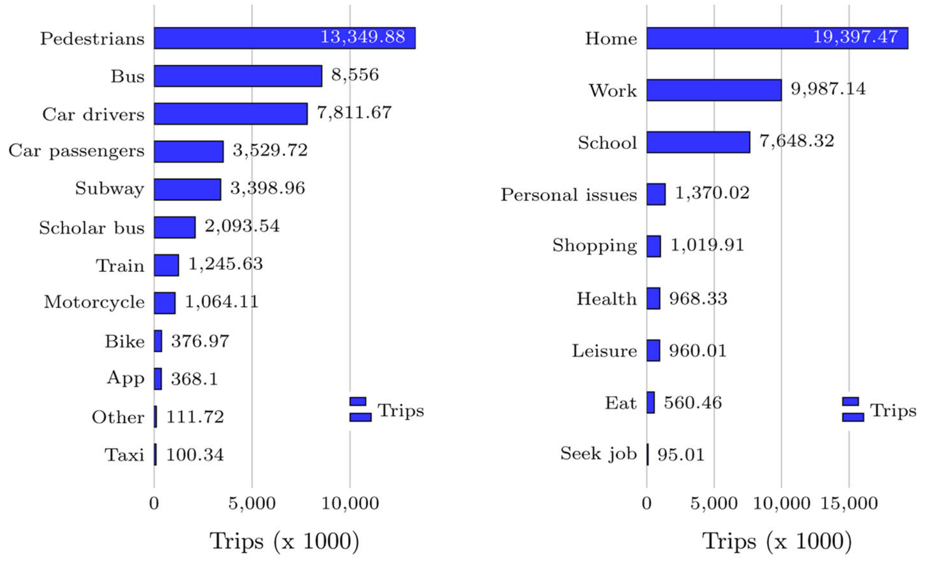
Fig. 2 Municipalities of the SPMA

São Bernardo do Campo (844 thousand), Santo André (721 thousand), and Osasco (699 thousand) [31].

The capital and its nearest cities concentrate most of the job opportunities, public facilities and services, universities, museums, and entertainment options [32]. Thus, there is huge daily commuting to the capital downtown and its surrounding neighboring. In São Paulo, neighborhoods close to the city center are highly valued, so living in these parts of the city is expensive. People with less financial conditions usually live in the peripheral areas of the capital or in other cities in the SPMA. Indeed, some cities in the SPMA are *dormitory towns* for those who cannot afford to buy or rent a house in the capital [33, 34].

Regarding the transportation infrastructure, the capital has a subway system that serves the northern, southern, eastern, western, southwest, and southeast regions of the city. There are a few subway lines, most of them crossing the center of the capital, which limits access to the subway system to some regions of the city. The other cities do not have subway systems. There is a metropolitan railway system that serves several surrounding cities and also the capital. The railway system is integrated with the subway system.

Each city has its own bus system and there is an intercity bus system to link the neighboring cities. During the 20th Century until the late 1990s, most investments in transportation were guided by a car-centric approach in detriment of public transportation [35]. The aims were to enlarge roads and streets to attend the increasing demand for private cars. In the last two decades, the local governments have invested more in policies to incentivize public transportation, such as the implantation of bus corridors and the replacement of bus fleet, building new subway lines, and modernizing the railway system. Also, there is an increasing cycling infrastructure in the capital that is being expanded in the last decade. Although these investments have increased in recent years, the SPMA still suffers from traffic congestion, especially during peak hours [3, 35]. Thus, there is a need for a better understanding of the traffic behavior in the SPMA to propose new policies to improve mobility for its citizens.



(a) Trip mode

(b) Trip reason

Fig. 3 Trips by mode and by reason

3.2 Origin–Destination survey

The Origin–Destination (OD) survey is the primary source of mobility information in the SPMA. It is performed by the São Paulo Metropolitan Company (Metrô) every ten years since 1967. The last OD survey of 2017 (OD17) has information about 157992 trips of people randomly sampled in the SPMA, which is a representative sample of the total population with an error margin of 6% and a confidence interval of 92% [36, p.34]. Trips occur for different reasons such as work, house, study, and leisure, and different trip modes such as by walk, car, subway, and train. Figure 3 shows the distribution of trips by transportation mode and by motivation. Figure 4 shows the distribution of trips by the hours of the day.

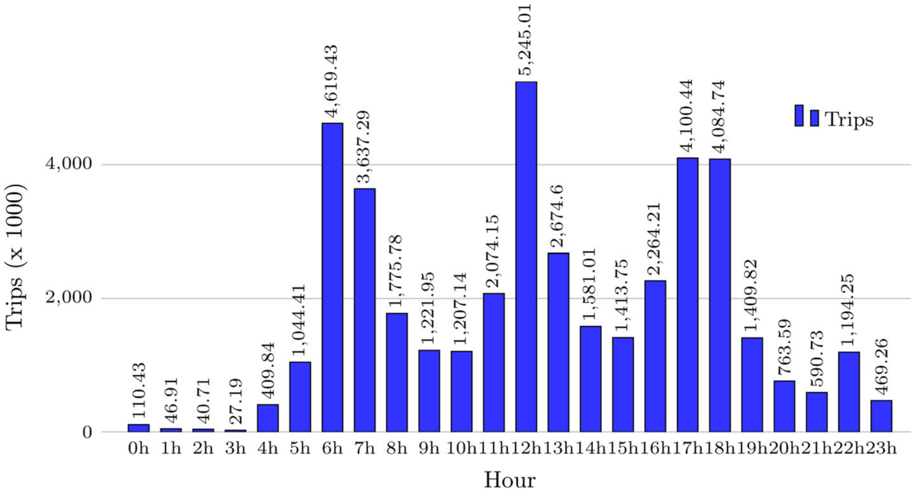


Fig. 4 Trips by hour of the day

As Fig. 3 shows, most trips in the city are made by pedestrians. Usually, pedestrian trips cover small distances (624m on average) and stay inside a single district of the city. However, there is also a significant number of trips by car and public transportation (bus, subway, and train), most of them longer (7.7 km on average), especially by train and subway (13.5 km on average). Looking at trips by hour (Fig. 4), we see that traffic in the SPMA has three main peaks: in the morning (6 AM to 8 AM), at lunchtime (noon), and in the early evening (5 PM to 7 PM).

The 157992 entries in the OD survey have several important attributes for our study, as follows (attribute type indicated between brackets). The most important ones are the *origin O* and *destination D* coordinates (quantitative), and the *expansion factor E* of the trip (quantitative), which is the statistical extrapolation for the population size that each surveyed trip represents. That is, a trip $\mathbf{t}_i$ in the OD survey with an expansion factor value E_i models a number of E_i actual trips that closely follow the trail $\mathbf{t}_i$. The expansion factor summed over all $N = 157992$ survey entries yields the $T = 42$ million trips on a typical working day, *i.e.*, $\sum_{i=1}^{N} E_i = T$. Other attributes are the *time* of departure (quantitative), the transportation *mode* (categorical, 17 values: train, subway, car driver, car passenger, bus from the São Paulo city and from other cities, intercity bus, monorail, chartered vehicle, school bus, regular taxi, non regular taxi, motorcycle driver, motorcycle passenger, pedestrians, bicycle, and others), and the *reason* for the trip (categorical: work, house, study, and leisure).

Figure 5 shows the OD survey trajectories plotted over the map of the SPMA, where each line represents an OD trail. The extreme clutter in this figure precludes the visualization of individual trajectories, traffic patterns, or connections between the regions of the map. From this image, we can only infer the existence of high traffic over the major part of the metropolitan area and a significant concentration of this traffic at São Paulo downtown. Moreover, this image does not show any of the available data attributes, except O and D. Hence, filtering and aggregation techniques are essential to simplify the visualization and make it understandable for the human eye.

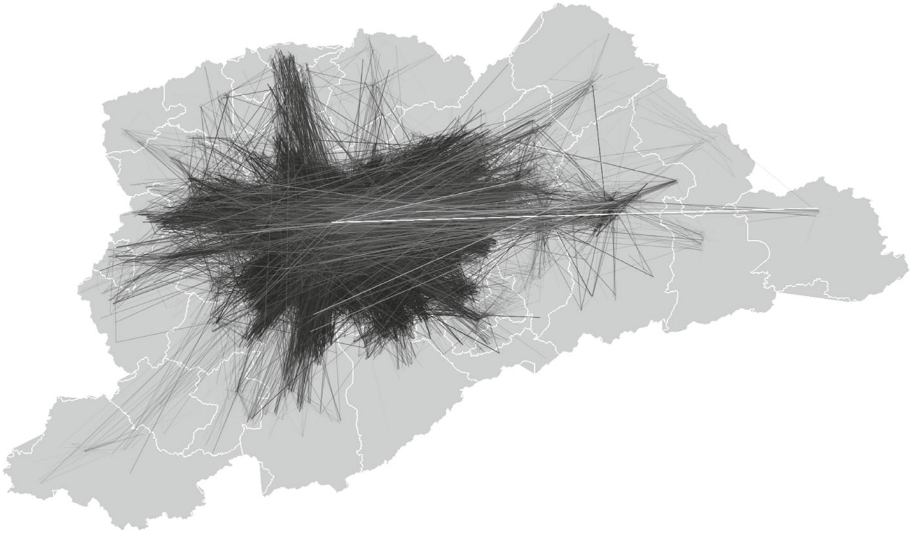


Fig. 5 Naïve straight-line drawing of OD trails from the SPMA

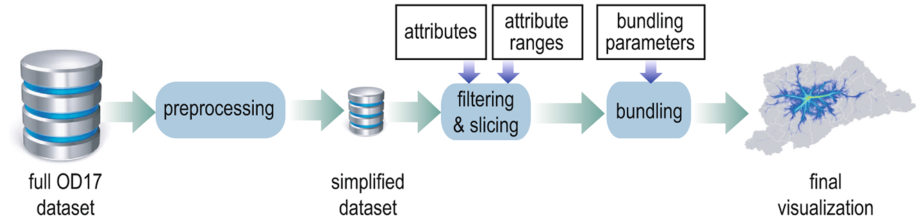


Fig. 6 Pipeline of the OD17 visual data exploration

4 Methodology

We used the OD17 data to explore how bundling can be used to visualize mobility patterns over the São Paulo Metropolitan Area (SPMA). The OD17 contains data that represents trips of a regular working day over the SPMA. We parameterized and adapted CUBu to explore several properties of this data with different bundled visualizations. Figure 6 shows our entire pipeline. We detail each of its steps next.

4.1 Data representation

The OD trajectory data we use as input for bundling is described by a table with six columns: trail ID, transportation mode, departure time, arrival time, origin coordinates, and destination coordinates. We extract this data from the OD17: the trail ID is the identifier obtained directly from the OD17 dataset. The origin and destination coordinates are transformed to the latitude/longitude system (required to geolocate these coordinates on a map). The transportation mode is stored as an integer in the range 1 to 17. The expansion factor of each OD record (introduced in Section 3.2) is used to replicate trails – for an OD17 trail with expansion factor E, we create E copies of that trail as input for bundling. This yields a complete dataset of 42 million trails. For all such replicated trails, we keep the ID of the underlying OD17 trail which generated them. This way, we can trace back which bundled trails correspond to an OD17 record. To this base data, extra attributes from the OD17 survey can be added such as trip reason (see Section 3.2) and personal data (age, income). Overall, this yields a dataset with over ten attributes per trail.

Table 1 shows a sample of the generated trails. Here, trails corresponding to the top two rows were created from the same OD17 trail with ID 50 and $E = 2$.

4.2 Data preprocessing

Although CUBu is – according to its authors and also to the best of our knowledge – the fastest existing solution for trail bundling, it still cannot process 42 million trails at interactive rates, which is required if one wants to explore a dataset by changing the visualization parameters and see the changes' effects within several tens of milliseconds. CUBu's original implementation, which uses a dual-GPU NVidia GTX 690 card, can process around 1 million trails interactively. Using newer GPUs can push this to several million. However, large trail sets create another problem, namely that these would not fit in the Video RAM (VRAM) memory of the card, as 1 million trails requires roughly 1GB VRAM in CUBu. To deal with these scalability problems, we reduced the OD17 dataset based on the expansion factor E.

Consider the distribution of the expansion values E_i over the sample records from OD17. As explained in Section 3.2, these generate by expansion $T = 42$ million trails.

Table 1 Format of data used as input for bundling

ID	Mode	Depart time	Arrive time	Origin	Destination
50	1	6:45	7:10	-46.62809376987491[a] -23.551691865840347[b]	-47.00348104352116[g] -23.39356328288028[h]
50	1	6:45	7:10	-46.62809376987491[c] -23.551691865840347[d]	-47.00348104352116[i] -23.39356328288028[j]
51	4	8:30	9:05	-47.00187231236886[e] -23.39846860627696[f]	-47.00348104352116[k] -23.39356328288028[l]

[a] Origin: Longitude
[b] Origin: Latitude
[c] Origin: Longitude
[d] Origin: Latitude
[e] Origin: Longitude
[f] Origin: Latitude
[g] Destination: Longitude
[h] Destination: Latitude
[i] Destination: Longitude
[j] Destination: Latitude
[k] Destination: Longitude
[l] Destination: Latitude

We argue that, if we are able to reduce the trail-dataset T to a smaller one T', which has the same distribution of E_i values, then T' roughly captures the same insights as T, but is faster to bundle. A simple way to obtain T' is to downscale the expansion factors as $E'_i = E_i/K$, where $K > 1$ is a downscaling factor, and next construct T' by expanding the factors E'_i. However, for $E_i < K$, $E'_i < 1$, which would mean the respective records would expand into less than one trail. Since the number of trails is an integer, this means those records would practically not influence T'. Hence, we simplify the OD17 record set by

- removing all records where $E_i < E_{min}$ (for a given E_{min} value, discussed next);
- setting $E'_i = E_i/E_{min}$ for the remaining records;
- expanding these records, each into E'_i trails.

To find a suitable E_{min} value, we analyze the accumulated percentage of trips that are removed for different threshold values E_{min} (see Fig. 7). For these values, *i.e.* where $E_{min} < 75$, the removed trails would represent less than 4% of the total population, *i.e.*, $\sum_{E_i<E_{min}} E_i < 0.04T$. Given that the error margin of OD17 was below 6% (see Section 3.2), we argue that removing such records would not affect the key insights captured by the dataset. In practice, we used the even more conservative value $E_{min} = 55$, which removes 2.35% of the total population (trail-set). Following this, we computed the downscaled distribution $E'_i = E_i/E_{min}$. This effectively reduces the size of the trail-set by a factor E_{min}, or, in practice, reduces T from 42 million to a trail-set T' of 685115 trails. This value is easily within the range of interactive exploration by CUBu.

4.3 Filtering and slicing the dataset

Beyond the data reduction process explained above, we also applied two other simplification methods, *filtering* and *slicing*, to investigate specific characteristics of the dataset.

To see the relations between the different transportation modes and the traffic directions over the peak hours, we used visual filters that make undesired trajectories fully transparent in the visualization. This is particular helpful when we want to visualize relations between different data, like how the buses from SP and from other neighboring

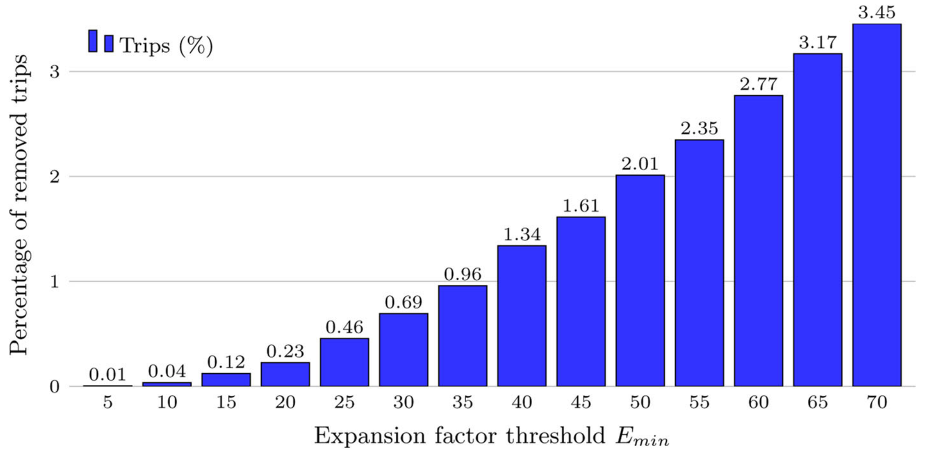


Fig. 7 Accumulated percentage of trip records removed by thresholding the expansion factor E_i. See Section 4.2

cities relates to each other, so it is suggestive to apply bundling in the entire data and then differentiate them somehow in the visualization, in this case we used different colors and filter options.

To perform the analyses related to trip reason, household income, and age of commuters, we sliced the whole dataset for each of these features separately and applied bundling to those subsets. This brings more details to the observed feature because it removes the interference of other data. This is also more straightforward to execute, instead of implementing filters to all the features, albeit it is desired to have both options in the analysis process.

4.4 Parameter setting

The literature is not clear on how to choose good bundling parameters [9]. This was observed and explicitly studied in Zeng et al. [20], which also proposed ways to compute good parameter settings. However, as they also mention, these settings are valid for their method (RAEB) and would not generalize directly to other methods, such as CUBu. Hence, we had to find good parameters for our study empirically. The obtained parameter set in this way was: image resolution $R = 512 \times 512$ pixels; trail sampling step: 10 pixels (in line with the recommendation in Van Der Zwan et al. [10] of using for this about 1% of the image diagonal dimension); kernel size $k = 18$ pixels; and number of bundling iterations $N = 15$. In our biggest dataset, represented by all 685115 trips, this configuration yielded around 3.2 million sample points, which were bundled in around 52 milliseconds per iteration on a PC with a 4GB NVidia GeForce 940MX graphics card GPU. We used this parameter set to create most of the bundled visualizations shown next in Section 5, except for the ones from Section 5.5 where we show the commuting patterns of different social strata classes. For these particular subsets, we lowered the sampling step to 5 pixels and set alpha transparency of the trails to the fraction of 0.15. In practice, half of the sampling step parameter doubles the number of points in the visualization, and also the changes in the transparency of trails reduces the amount of details in the visualization. But we observed that these variations are helpful to highlight the density spots that we seek in our analysis with no impact in the underlying structure of the visualization.

4.5 Visualization enhancements

To enrich our visualizations of São Paulo urban mobility, we made two important additions to CUBu. First, we added the map of the metropolitan area as a background image and plotted the subway and train lines atop of the bundled image (see Section 5.2). This way, one can use the map to reason about where certain bundles are; and can use the rails to see how bundles correspond to specific train infrastructure. Next, we implemented filters to select the subset of the 17 transportation modes to display in the visualization, using different categorical colors for each of them. We used this feature to explore relations between the different modes and their impact on urban mobility (see Section 5.4).

5 Results

This section presents the results of several analyses of the OD17. We started by exploring the different visual encodings provided by CUBu (density, distance, direction, and coloring) to visualize the distinct characteristics of the OD17 dataset. These aspects are covered next in Sections 5.1, 5.2, 5.3, and 5.4. Based on these visual encodings, we next designed several visualizations for studying specific mobility patterns. These are detailed in Sections 5.5, 5.6, 5.7, 5.9, and 5.8.

5.1 Adjusting density to visualize bundled trails

As explained in Section 2, standard bundling trades off clutter for overdraw. However, such a visualization does not tell us how many trails have been grouped in a bundle. The typical solution for this, pioneered by Holten [37] is to draw semi-transparent trails, each having a fixed transparency $\alpha < 1$. Thus, blending will show high-density trails as more opaque and low-density ones as more transparent, respectively. In addition to that (since transparency is not a strong quantitative visual variable [38]), we also encode trail density, directly estimated by the KDE density ρ, into color. Figure 8a shows a visualization obtained using color encoding for the entire OD17 dataset. We can see a few dense paths but the image still presents much clutter. This fact occurs because on consumer-grade GPUs, the transparency α is modeled, during blending, as an 8-bit integer value. Hence, only 255 different transparency levels are possible, *i.e.*, only 255 different trail-density levels can be displayed. Setting α too high would immediately saturate the transparency channel where higher trail densities occur – all densities above 255 are clamped to 255. Setting α lower than 1/255 would result in no image, since this would correspond to zero opacity on the 8-bit representation.

We address this problem by actually mapping the density ρ in both transparency and color. Since ρ is computed with floating-point accuracy on the GPU during KDE, no clamping or rounding-off issues occur. This density-based transparency modulation is useful to highlight even more the high-density areas and reduce general clutter – an alternative to the use of higher kernel sizes that would create stronger bundles but also cause more edge distortion. Figure 8b shows the resulting visualization on the same data as in Fig. 8a. Now high-density bundles appear more salient – there is more dynamic range in this image. The image suggests that the traffic network of the metropolitan area can be divided into a few core branches that are strongly connected to the central area, where the city of São Paulo is placed. Indeed, this makes sense because this is the most populous part of the metropolitan area (see Section 3.1). Moreover, most

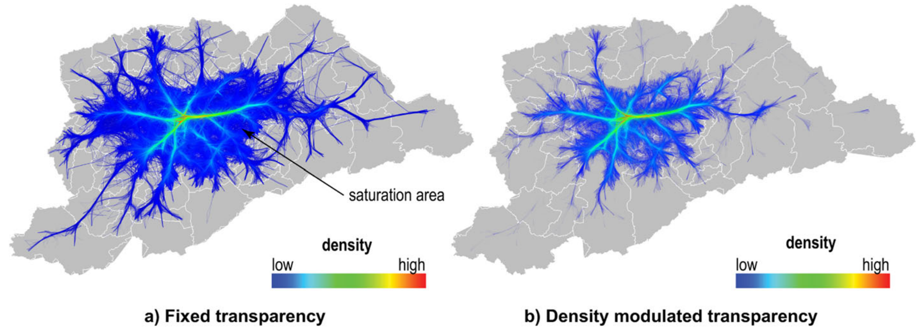


Fig. 8 Bundled trajectories colored by density in (**a**) fixed and (**b**) modulated transparency modes

transportation systems cross the capital downtown area, including subway and bus lines, and the main expressways.

5.2 Overlapping subway and train infrastructure *vs* bundled trails

The public transportation system is the most used by the citizens of the SPMA, responsible for 36.4% of the total daily trips [36, p.56]. The impact of the rail mesh over the commuting of people is clear when we plot the rail lines over the bundled trajectories, as regions with the highest flows correspond to the paths of the rail lines (see Fig. 9, where the rail lines are the black lines). This is an expected result since according to the OD survey, about 44% of daily trips by public transportation involve the subway or trains. More interestingly, one may wonder whether the rail system was accurately planned to supply the demand, as the bundled visualization suggests, or whether the availability of this transportation option influenced the existence of such dense flows. While we may not know how to answer this question, traffic managers might use this kind of visualization to devise policies for public transportation. The high correlation between the paths and the rail lines also indicate good parameter settings for the bundled visualization in the metropolitan scale.

Note that this type of correlation (of bundles with rails) is not the same as RAEB [20]. In RAEB, the bundling was *explicitly* made to follow roads. In our case, roads are *superimposed* by a bundling that uses only the OD data. One may argue that RAEB, in this sense, produces more 'correct' bundles since these are constrained to follow the roads. However, upon a closer look, we can see that RAEB *cannot* have all bundles precisely follow the roads their trails went along – doing so would basically block any bundling. Moreover, RAEB requires registering OD trails with an accurate road network to function, and is significantly more complex to implement and more expensive to run than our CUBu-based solution. Hence, we argue that, by using relatively small kernels k, thus limiting the distortion of the original trails, our CUBu-based solution is an overall better alternative to RAEB.

5.3 Mapping length and direction attributes to bundled trails

To explore urban mobility from different perspectives, we need means to visualize its multiple data attributes. However, this variety of attributes requires distinct visualization strategies. Two important trail attributes for the study of mobility patterns are trail *length*

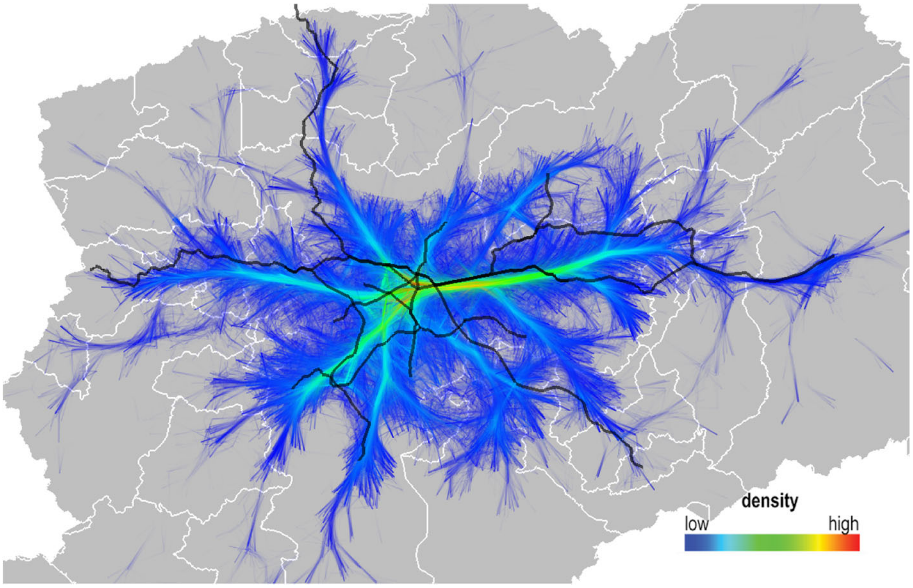


Fig. 9 Bundled trails colored by density and rail lines

and trail *direction*. Figures 10 and 11 show the visualization of the entire OD17 dataset using length and direction, respectively.

Figure 10 displays color-coded trail lengths with density-modulated transparency as explained in Section 5.1 using the same rainbow colormap as in Fig. 9. We see a single roughly horizontal red curve in this image, but with maximal opacity. This implies there are many long trips, all which perfectly map to this trajectory between the same origin and destination (if they did not, we would see a fanning-out bundle rather than a precise curve). This is an interesting finding that, we argue, could not be easily found using non-visual methods. Besides this outlier, other trails, in general, run over regular distances. Long-distance edges can indicate lack of services or resources that do not satisfy local

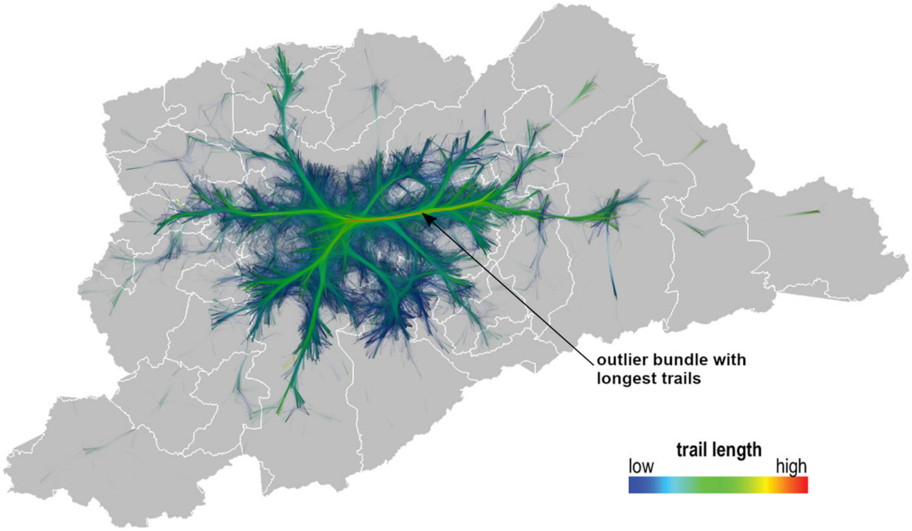


Fig. 10 Mapping trail length to colors

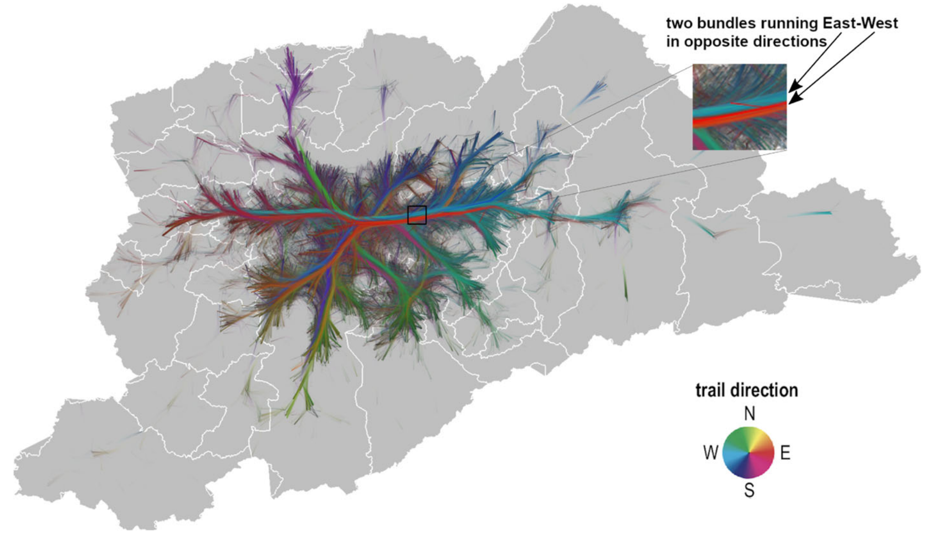


Fig. 11 Mapping trail direction to colors

regions, forcing people to commute long distances to access them. The OD17 contains more information that may help to investigate the reason behind these long trips.

Figure 11 shows the same data as in Fig. 10, but with color encoding trail directions. Also, we used here the directional bundling mode of CUBu which separates different-direction, close-location, trails into two roughly parallel bundles. We can clearly see the existence of parallel trajectories over the bundles, which is not surprising because the OD survey records the typical two-way commuting of people going from and then coming back to their origins. However, this symmetry would possibly not be seen if we analyzed a short time period of the day.

5.4 Coloring transportation modes: local *vs* intercity buses

The OD17 dataset contains 17 transportation modes. While it would be ideal to be able to see the 17 categories all at once in our bundled visualization, that would not be easy to do, since it would require the simultaneous encoding of 17 different categorical attributes. Instead, we use transparency to hide trails according to a user-set selector that filters them by transportation mode. Figure 12 shows how we can use these filters to visualize the integration between buses from the city of São Paulo (local buses) and intercity buses. Each transportation mode has a distinct color – olive for local buses and blue for intercity buses.

Figure 12 also highlights that these different transportation systems appear to complement each other. The central region of São Paulo has a very active commerce and industry (it accounts for 64% of jobs in the SPMA [36, p.48]) so that many people from neighboring cities work there. Thus, the availability of public transportation and its integration is very important for these people. This kind of filtering along with bundling helps to better understand correlations between data attributes – in this case, transportation modes.

5.5 Density per social strata

We used our bundle visualization to study how citizens with different economical conditions commute in the SPMA. The Brazilian Economical Classification Criterion

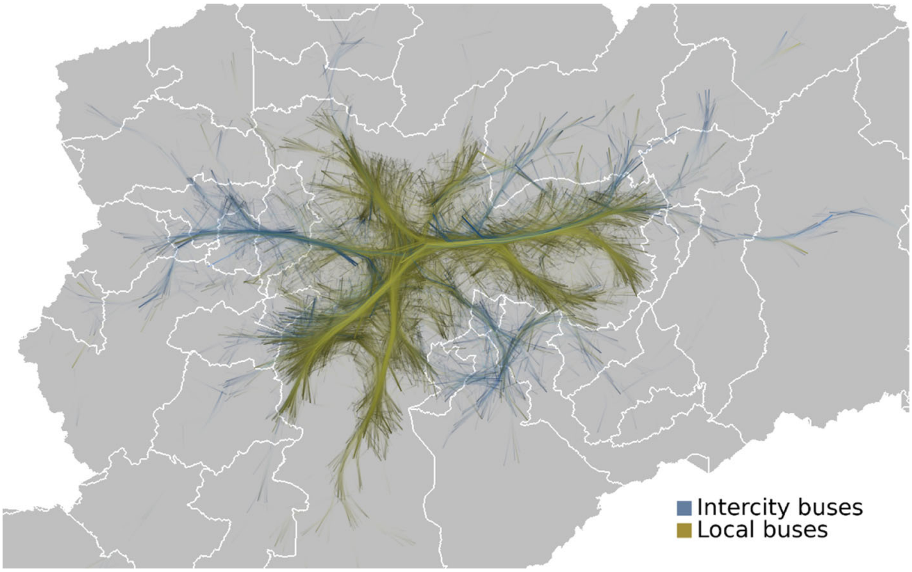


Fig. 12 Edges filtered by transportation modes: bundled trails of local and intercity buses

(BECC) [39] is the official socioeconomic index used in the Brazilian Demographic Census, which is performed by the Brazilian Institute of Geography and Statistics. It measures the purchasing power of the Brazilian society. The BECC is divided into six levels or strata (Table 2). This index is used in the OD17 survey to complement the mobility data. Table 2 also shows the average monthly income in the local currency (Brazilian reals) and in US dollars, and the number of trips in the SPMA for each BECC level considering the whole population and only citizens with age between 6 and 18 years that commute for study (see Section 5.6).

To compare the mobility patterns of different BECC social strata, we bundled the trails in each stratum separately, as shown in Figs. 13, 14, 15, 16, 17, 18 and 19. We see significant differences in the mobility patterns between the highest and lowest income levels as shown in Fig. 13. The *A* level (on the left-hand side) has a high density in the center of SPMA, which includes the capital downtown surroundings. The highest density is located in the west, southwest, and northeast neighborhoods near downtown. There are density flows between the capital and the cities of Barueri and Cotia, which have high-income residential areas. There are other high dense flows linking the capital to the cities of São Bernardo do Campo and Santo André. Comparing *A* to the *D*–*E* level (on the right-hand side of Fig. 13), we see that *D*–*E* has the highest dense flows in the capital eastern region.

Table 2 Trips grouped per BECC income level, social stratum, and traveler age

BECC level	Monthly income (Brazilian reals (R$))	Monthly income (US dollars (~US$))	Trips	Trips of 6 to 18 years old students
A	23,345	4,245	3,062,892	184,772
B1	10,386	1,888	3,854,040	260,652
B2	5,363	975	12,856,182	963,242
C1	2,965	539	11,277,159	976,745
C2	1,691	307	7,852,806	721,218
D-E	708	128	2,233,801	219,612

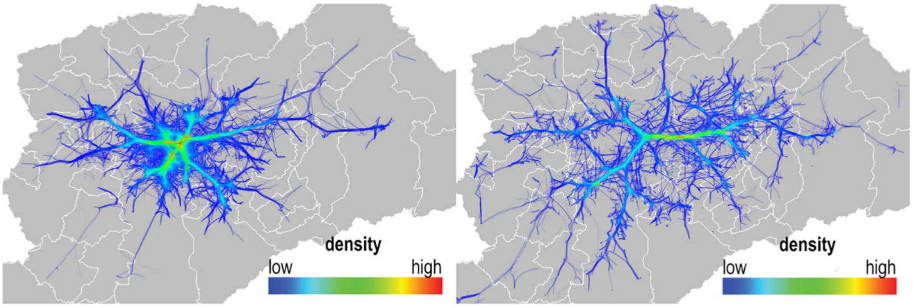


Fig. 13 Comparing density of trails between social strata *A* (left-hand side) and *D*–*E* (right-hand side)

In the *D*–*E* level map, we can see the absence of high-dense flows in regions that are nearest to the capital downtown; in contrast, these are present in the *A* level map. We can see more details of *A* and *D*–*E* strata in Figs. 14 and 19

When we compare all maps from the *A* to the *D*–*E* level (Figs. 14, 15, 16, 17, 18 and 19), we see that the densest flows (red) tend to displace from the capital downtown to the eastern region of the city. The concentration of high-density flows is increasingly spreading from the center to the peripheral regions of the SPMA. Even the less dense flows are increasing and spreading over the SPMA. However, the *D*–*E* map shows that those flows diminish considerably for these social strata. This may indicate that low-income citizens have less access to the urban mobility system. As a consequence, these people would have less access to the social, educational, health, and cultural services of the SPMA, as those facilities are concentrated in the center regions of the cities. For example, data from the city of São Paulo shows that in 2017, 68% of all cultural facilities were concentrated in the center areas of São Paulo, from which 30% were located in the downtown (Sé district) and 14% were located in the Pinheiros district [40]. It is worthy to note that those central regions also have more job opportunities. Looking at the *D*–*E* map, we can see a “hole" in the capital west downtown. This region (Pinheiros district) concentrates a large number

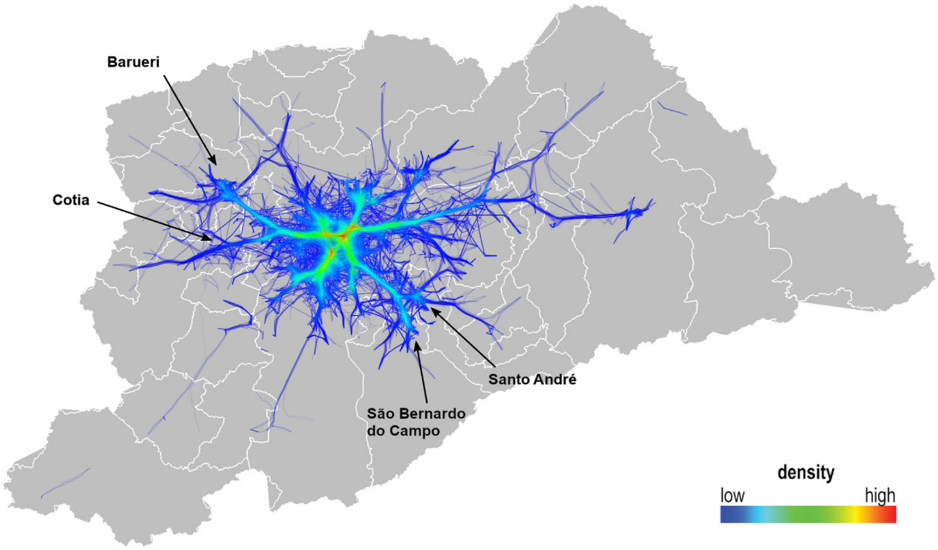


Fig. 14 Density of trails of social stratum *A*

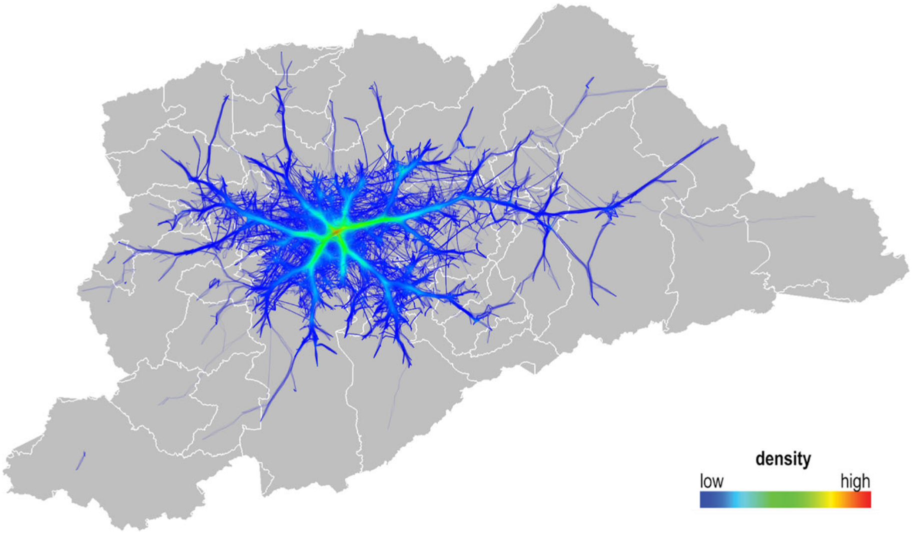


Fig. 15 Density of trails of social stratum *B*1

of jobs related to information technology and financial services, which requires workers with high and medium education levels. Thus, the map shows that low-income citizens are not going to that region, which reflects the inequality of opportunities that these citizens face. Indeed, data from the city of São Paulo [41] shows that in 2018, Pinheiros had 19.8% of all jobs for medium to high education levels against 7% of all jobs for those with lower education levels.

5.6 Mobility of young students from different social strata

To explore even more the mobility patterns showed by bundling visualizations, we compared the trips of students from different social strata. We filtered citizens with age between 6 and 18 years whose commuting reason is study. We split them into two groups,

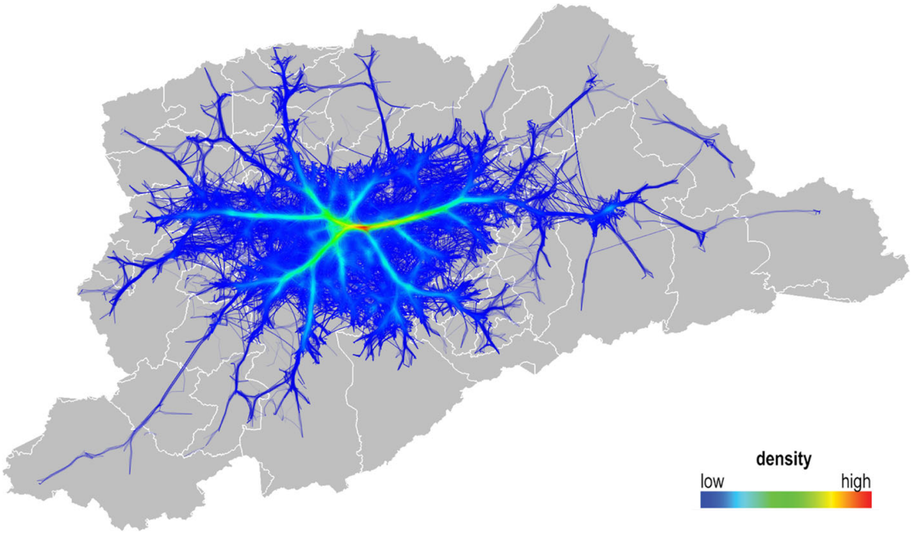


Fig. 16 Density of trails of social stratum *B*2

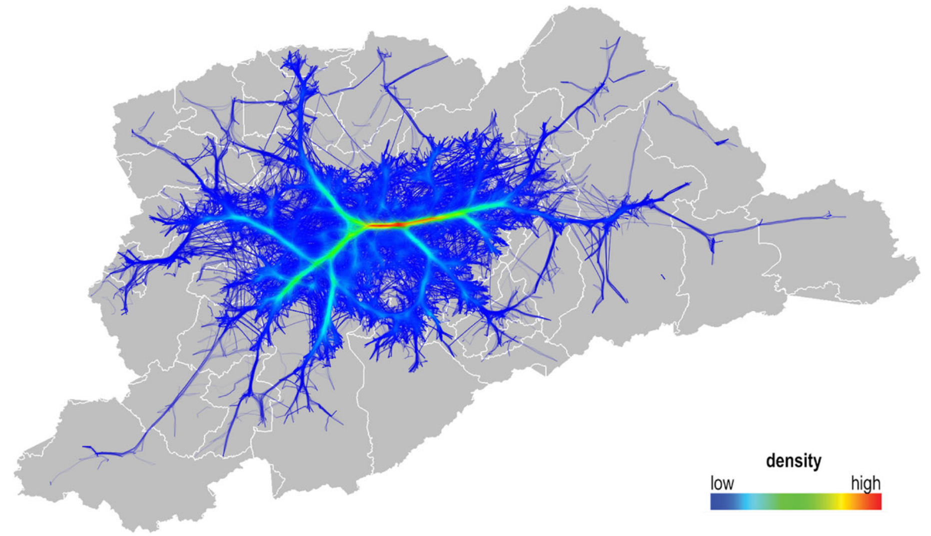


Fig. 17 Density of trails of social stratum *C*1

the high- to moderate-income, which includes the BECC levels *A*, *B*1, *B*2, and *C*1; and the low-income, which includes levels *C*2 and *D*–*E*. The *C*2 and *D*–*E* strata represent the population with family income of up to 4 minimum wages. These populations cannot pay for private schools for their children. Generally, students from private schools in Brazil have a higher performance compared to those from public schools [42]. Also, the socioeconomic characteristics of both the student and the student's peers correlate with academic performance [43]. Moreover, the average monthly expenses with education of E, D, and C2 strata are up to 25% of 1 minimum wage [44], which is not enough to pay a private school. For the C1 stratum, this value is at least 67% of 1 minimum wage. Figures 20 and 21 show the density maps for the high- to moderate-income and low-income groups.

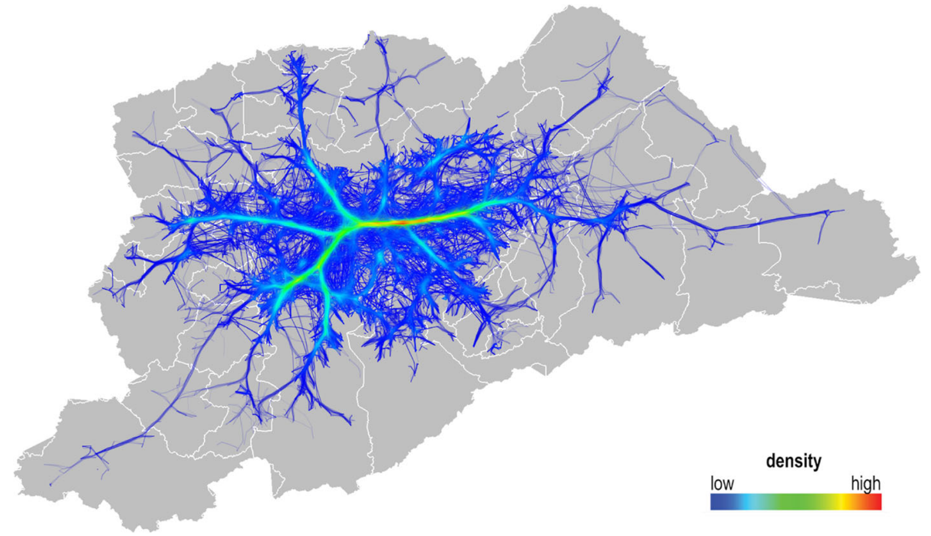


Fig. 18 Density of trails of social stratum *C*2

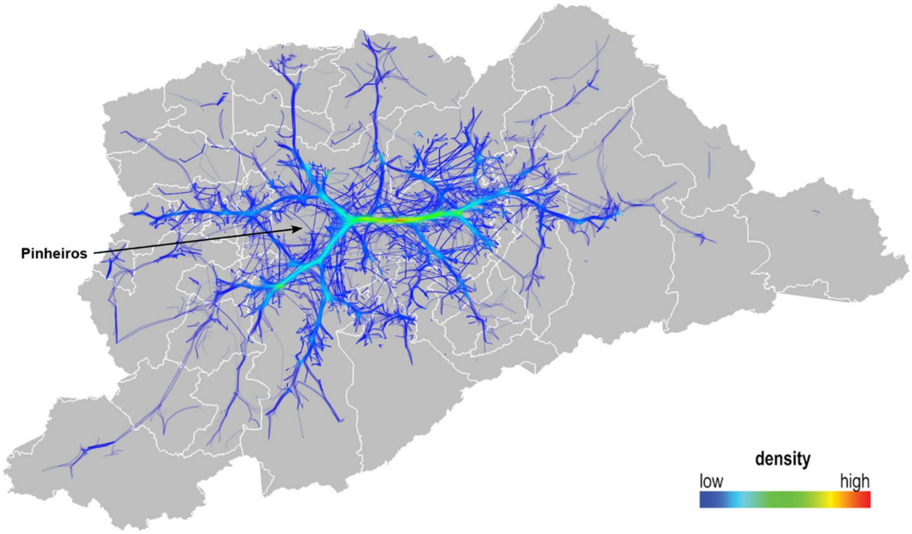


Fig. 19 Density of trails of social strata *D–E*

The density map of the high- to moderate-income students (Fig. 20) shows a large number of dense flows spread across the central region of SPMA. This part of SPMA concentrates most private schools, universities, and complementary colleges. In addition, high density flows are not as long as flows from other maps with all the data (e.g., Fig. 8). This indicates that trips to study are shorter than trips to work.

The density map of low-income students (Fig. 21) shows that their mobility is very limited compared to the higher-income students. There are a few dense flows, most of them outside of the capital downtown. The high density flows of low-income students are more present in the peripheral regions of the city and also in the neighboring cities. There is a concentration of both groups in the southwest region, where the neighborhoods of the Campo Limpo district are.

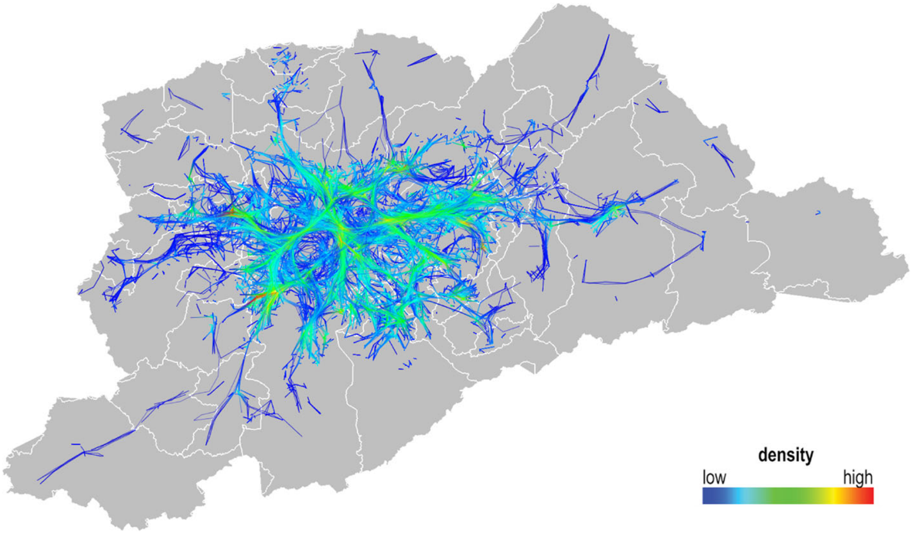


Fig. 20 Density of trails of young students from high-income households

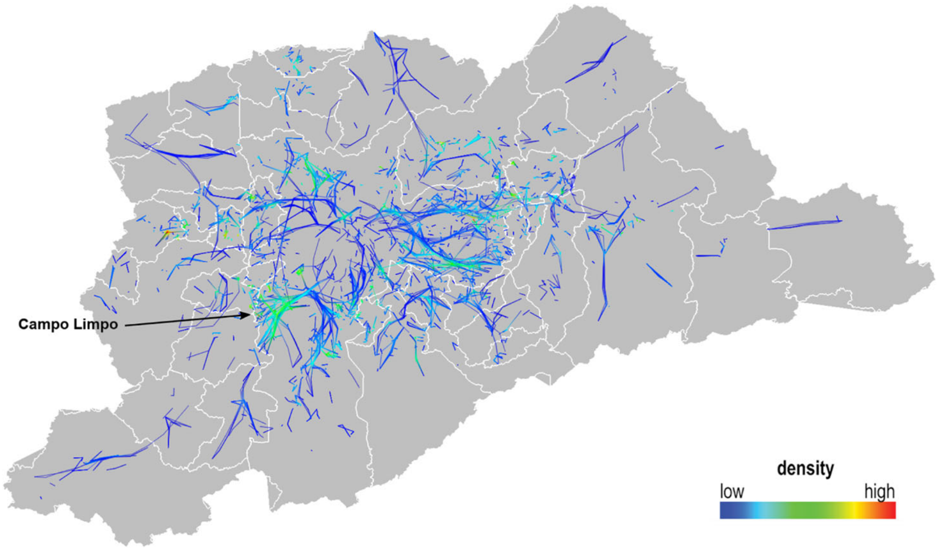


Fig. 21 Density of trails of young students from low-income households

It is worth noting that the public schools in the SPMA are spread across the central and peripheral parts of the cities. Generally, the students are enrolled in these schools according to the proximity of their residences [30]. Thus, they do not have to travel long distances to reach their schools. On the other hand, public schools have lower educational performance than private schools in São Paulo. Thus, citizens with better financial conditions usually enroll their children in private schools.

The scarcity of flows from low-income students may indicate that they have less school choices available, as they are enrolled in schools in their neighborhood. This also mean that their school peers also live nearby and, presumably, belong to a similar socioeconomic group – which, as mentioned before, is correlated with academic performance. They also do not use to go to the central region of the city and, thus, have less access to universities and complementary colleges. This inequality will probably impact these students' jobs and economical conditions.

We also see that there are many more trails for the high- to moderate-income students (Fig. 20) than for low-income students (Fig. 21). The high-to moderate-income students also travel larger distances to study, which indicates that they can choose more flexibly where to study. This fact is corroborated by urban mobility studies that indicate that people with better financial conditions have more mobility than those with poorest conditions [45, 46].

5.7 Directions at peak hours

As discussed earlier in Section 3.2, Fig. 4 shows the distribution of trips by hour of the day, with two main rush-hour peaks (6–9 AM and 5–8 PM). However, this aggregated table does not give us insights in how the rush-hour patterns may differ. To see this, we selected the two rush-hour time intervals and visualized them separately, using directional bundling and color-coding.

Comparing the peak hours, we can see that morning flows going to the SPMA center (Fig. 22, cyan bundle) are overall denser and longer than the flows coming from the SPMA

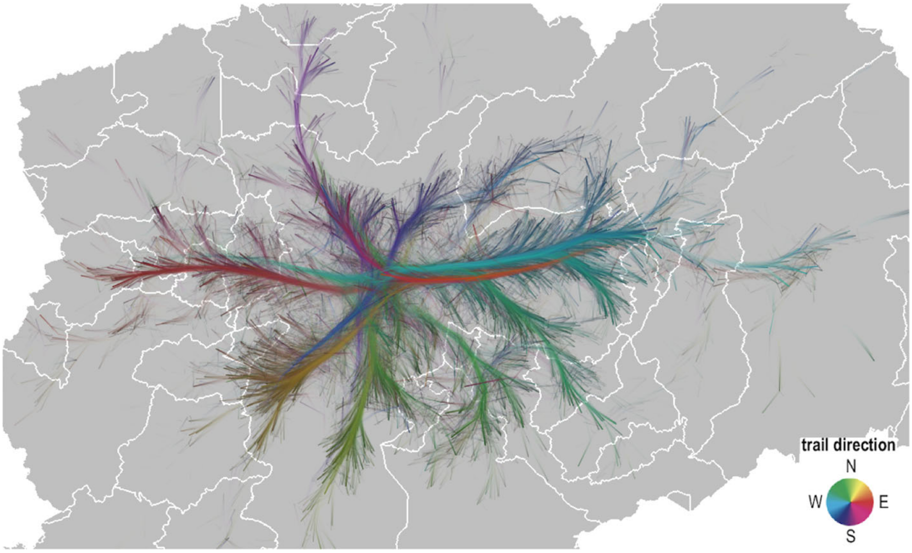


Fig. 22 Directions of trips between 6 to 9 AM

center during the afternoon/evening peak (Fig. 23, red). This suggests that in the morning people are in a hurry to reach their work, while they are less in a hurry to go back home (or to other destinations like schools or the gym) in the afternoon/evening.

In Fig. 22, we can see that flows in the morning peak going to the capital downtown (cyan bundle coming from the east) are denser than opposite flows (red bundle going to the east). Although flows leaving the capital downtown in the morning are thinner than their opposite ones, they also concentrate a large number of trips, especially to the east and southwest. In Fig. 23, the opposite flows seem more equally distributed.

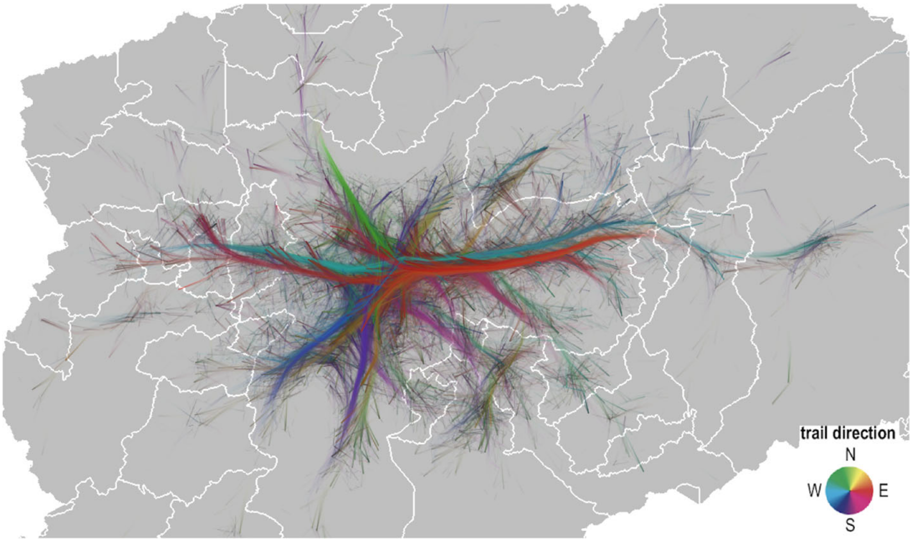


Fig. 23 Directions of trips between 5 to 8 PM

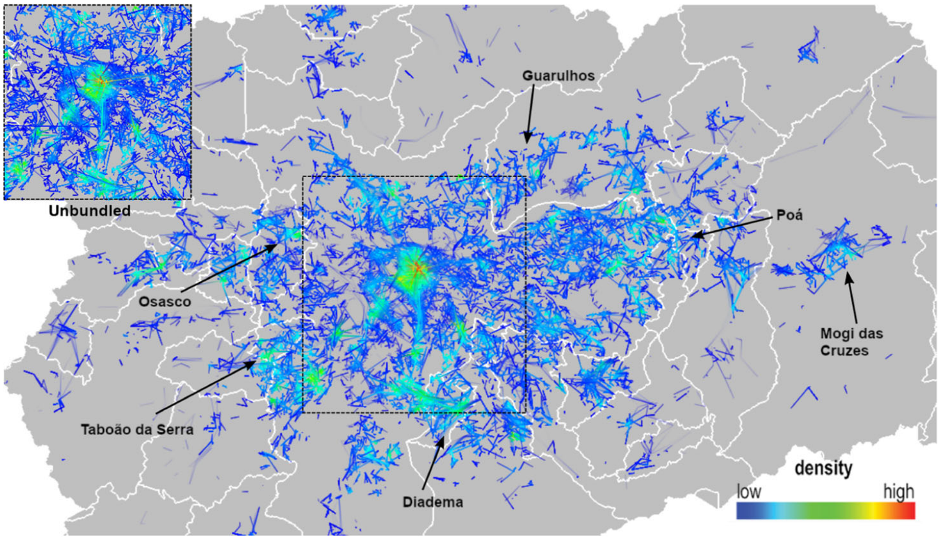


Fig. 24 Density of pedestrian trips

5.8 Density by transportation mode

We next split the OD17 data by transportation mode to compare the flow patterns for four different transportation modes: pedestrians, bicycles, cars, and subway. Figures 24, 25, 26 and 27 show the respective visualizations.

Pedestrian trails (Fig. 24) form several low-density 'islands' spread across the SPMA, with the densest one (red in figure) being in the capital downtown. Most trails are quite short, which is expected (pedestrians). However, we see a few longer bundles between the capital downtown and the south and north regions of the city. Dense flows are also present in the neighboring cities of Diadema, Taboão da Serra, Osasco, Guarulhos, Poá, and Mogi das Cruzes. Upon examination, we found these dense flows to match the cities' downtown

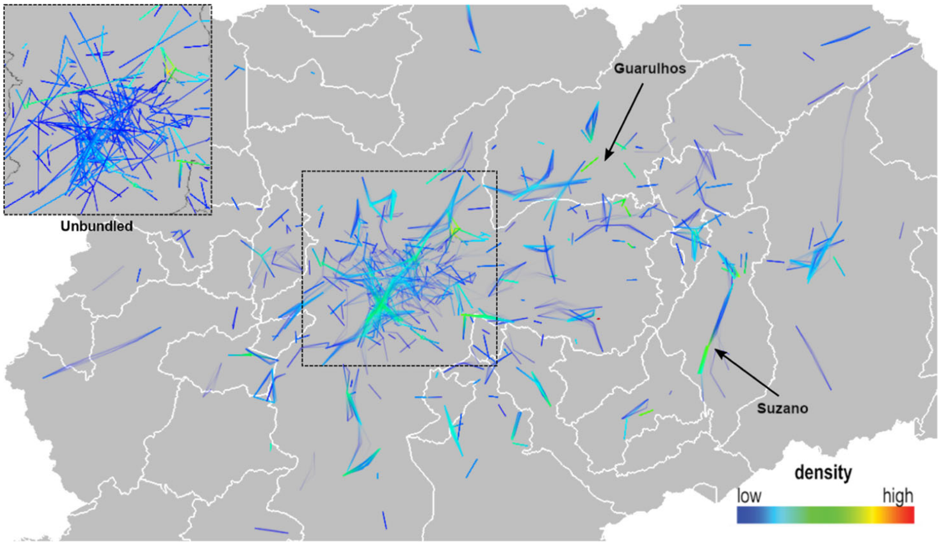


Fig. 25 Density of bicycle trips

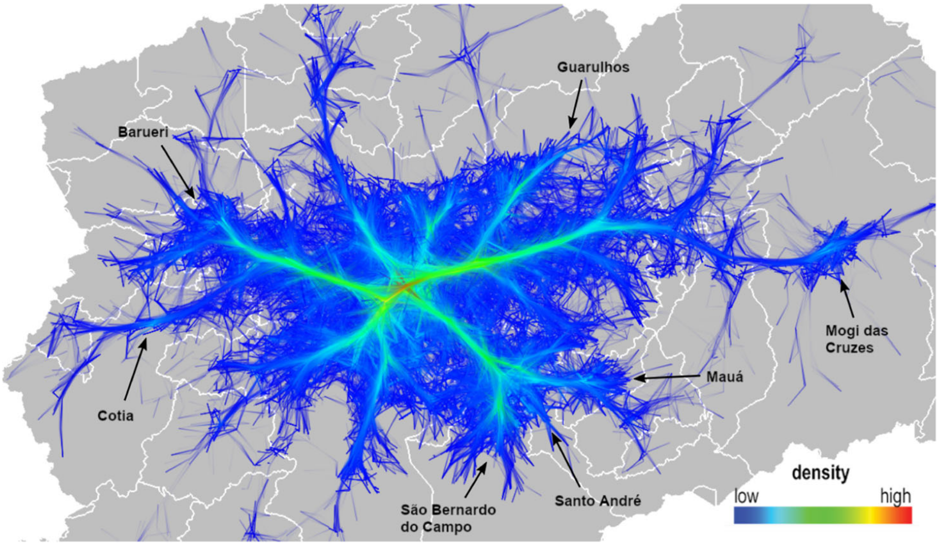


Fig. 26 Density of car trips

and commercial areas. This information could be useful to find places that could deserve the attention of local governments to provide improvements for pedestrians.

As most of the pedestrian trips are short, the bundling technique forms a few flows over the SPMA. Using bundling for those short trips result in low-density trails, which is less useful compared to long trips. Thus, in these cases it may not be necessary to use bundling. In the upper left area of Fig. 24, we can see the OD trails without using bundling, which are near identical to the main bundled area.

Bicycle trips (Fig. 25) exhibit similar patterns to pedestrian ones. They are shorter than three kilometers on average. In this figure, we see some thin flows in the capital downtown area. There are also some more salient flows in the capital northeast and in the

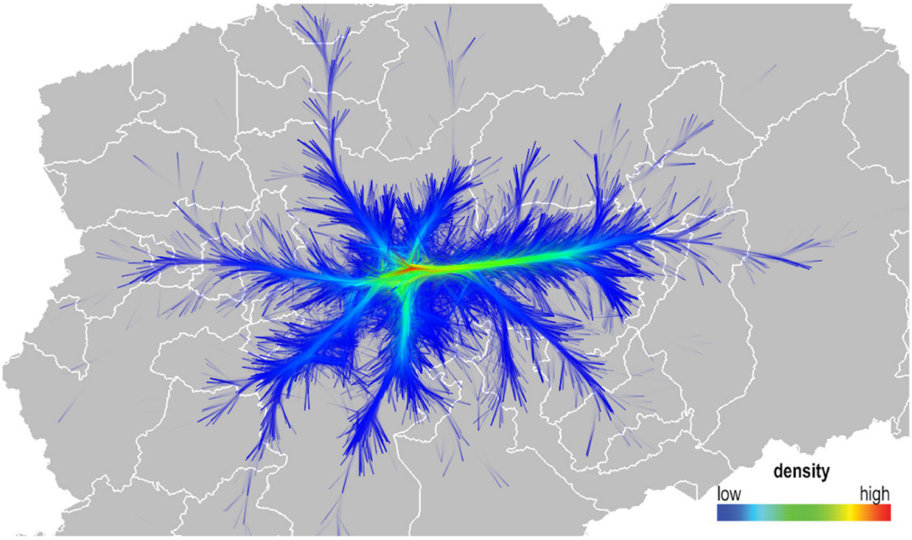


Fig. 27 Density of subway trips

neighboring cities of Suzano and Guarulhos. However, comparing Fig. 25 with all other transportation means, we immediately see that bicycle trips are by far the least numerous, and exhibit a far sparser pattern, with few star-shaped 'hubs' where many trails meet. This suggests that the cycling infrastructure is quite limited, and fragmented. Figure 25 also shows trails without using bundling in the upper left corner.

The car trips (Fig. 26) show a pattern similar to the one displaying the entire dataset, *i.e.*, all transportation modes (see *e.g.* Fig. 8). For a start, this tells that cars are *the* dominant form of transportation in the SPMA, accounting for the main traffic patterns. The highest-density flows occur in the capital downtown. There are several high-density flows linking the downtown area to the other regions of the capital, and also coming and going from the cities of Guarulhos, Barueri, Cotia, São Bernardo do Campo, Santo André, Mauá, and Mogi das Cruzes. Compared to all other transportation modes, cars show a far more 'spread out' pattern that covers very large areas, indicating that cars are the prevalent transportation mode in most parts of the SPMA.

Finally, subway trips (Fig. 27) show a strong star-shaped pattern, with very high density bundles that connect the capital with the neighboring cities, due to the integration of the subway system with the train system. Compared to all other transportation modes, subways show a clearer, simpler, trip pattern structure.

5.9 Different trip reasons

We next aim to study whether trips done for different reasons exhibit distinct trip patterns. For this, we create bundled visualizations from the OD17 data with trips grouped by work, health, education, and shopping. Figures 28, 29, 30 and 31 show the results.

Work-related trips (Fig. 28) are overall longer than the other trip reasons, and also cover a larger area (see the central agglomeration in the figure). Interestingly, the longest trips, between the east side and the city center (red bundle), are similar in pattern to the longest trips for health and education. Trips for health reasons are sparser than work-related ones, and also show a more star-like pattern, with long bundles connecting to the central area. This may indicate that peripheral regions are not well served by health services. Trips for studying reasons (Fig. 30) have the largest distances between the northeast and the western regions of the SPMA. Their pattern is somewhere in-between the work and health trips. Interestingly, education trips show several 'loops' in the center of the SPMA. Finally, shopping trips (Fig. 31) show the least dense, and overall also shortest, patterns, apart from a few outliers like the red (important) bundle connecting the center to the northeast. This tells that, unlike health, education, and work, shopping facilities (which are actually provided by private companies) are better distributed over the SPMA. This outlines that bundled visualizations are useful not only when they show the *presence* of certain data, *e.g.* trails linking far-apart regions; the *absence* of patterns is also insightful, as in the case of the lack of long shopping trips.

6 Conclusions and future work

In this work, we explored the usage of trail bundling for creating visualizations of various aspects of the urban mobility data from the São Paulo Metropolitan Area. Our analyses on the characteristics of the 2017 OD survey shows that bundling can be used to identify and compare different mobility patterns implied by different subsets of the data and subsets of the available attributes. By suitably combining filtering (to reduce the amount of data

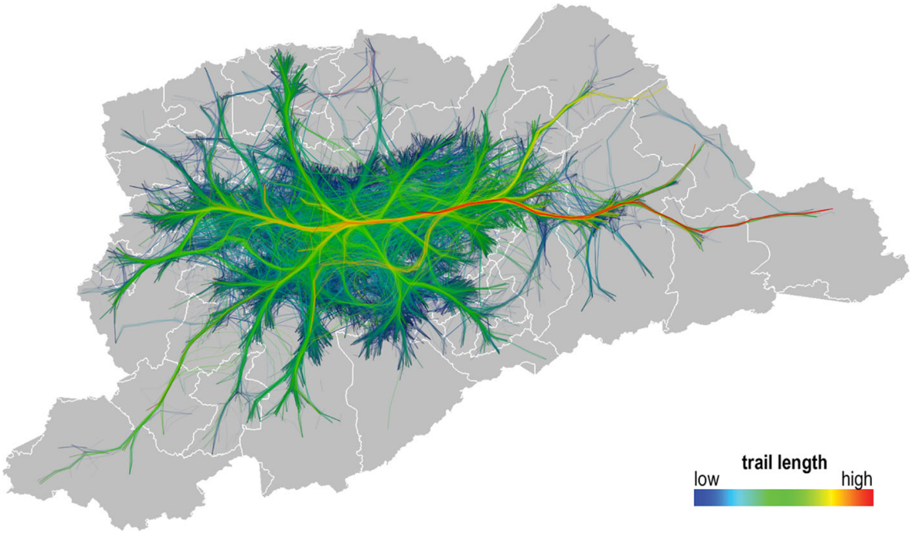


Fig. 28 Distance of trips for work reasons

and/or attributes to be explored) with bundling (to simplify the created visualizations and reduce visual clutter) and with the available visual channels (opacity, color, direction), we highlight different patterns in the OD17 dataset which would not have been easily obtainable by classical data mining and data analysis tools. In contrast to earlier work [14], this paper presents visual explorations of additional attribute combinations – density per social strata, mobility of young students per social strata, directions at peak hours, density by transportation mode, and trip distance per trip reasons. Together with earlier results [10, 14], our results strengthen the claim that trail bundling is an useful and usable tool for the visual analysis of large OD trail-sets.

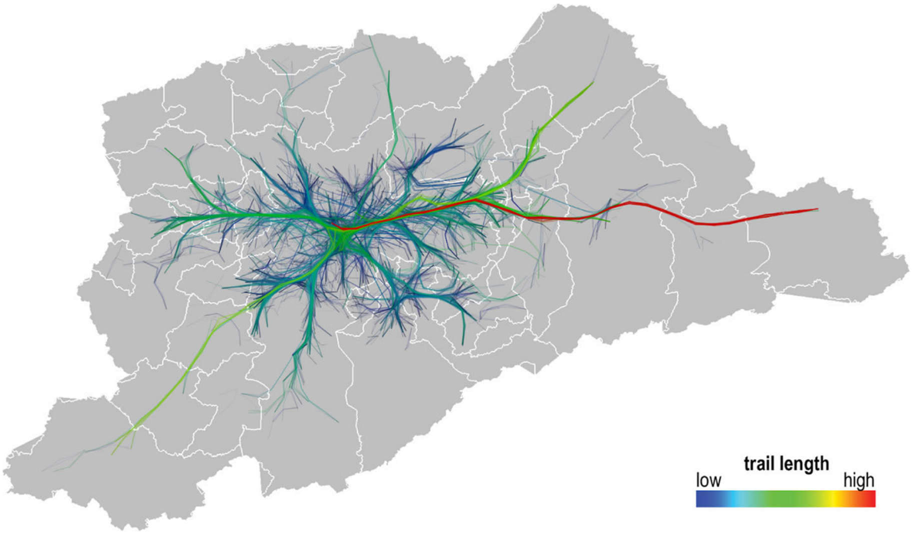


Fig. 29 Distance of trips for health-related reasons

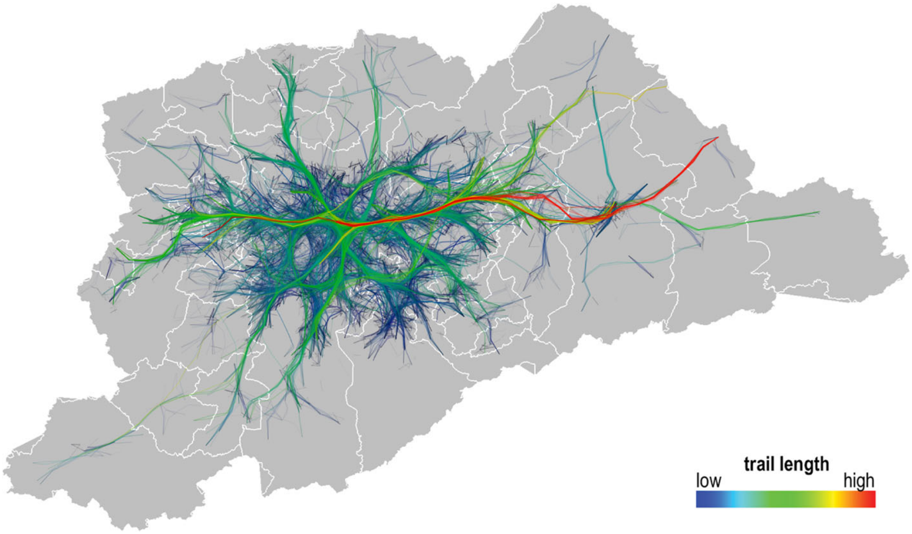


Fig. 30 Distance of trips for education reasons

The bundled layout for the trajectory dataset highlights its centrality structure over the represented area. Moreover, this structure matches the subway and rail lines infrastructure of São Paulo. Albeit this was not a surprise, the correlation suggests that our parameters were well tuned for the visualization in the metropolitan scale. Our methodology to reduce the dataset complexity from 42 million trips to less than a million, and also our customization of a general-purpose bundling framework (CUBu) to bundle specific subsets of data and/or attributes were key points that made this analysis possible.

As future work, we intend to explore improvements in the usage and usability of bundled visualizations. From a visualization perspective, improvements in the map to display region divisions on top of the trajectories, as proposed in Klein et al. [16], can help to better identify the connections between the regions. A different approach to convey the

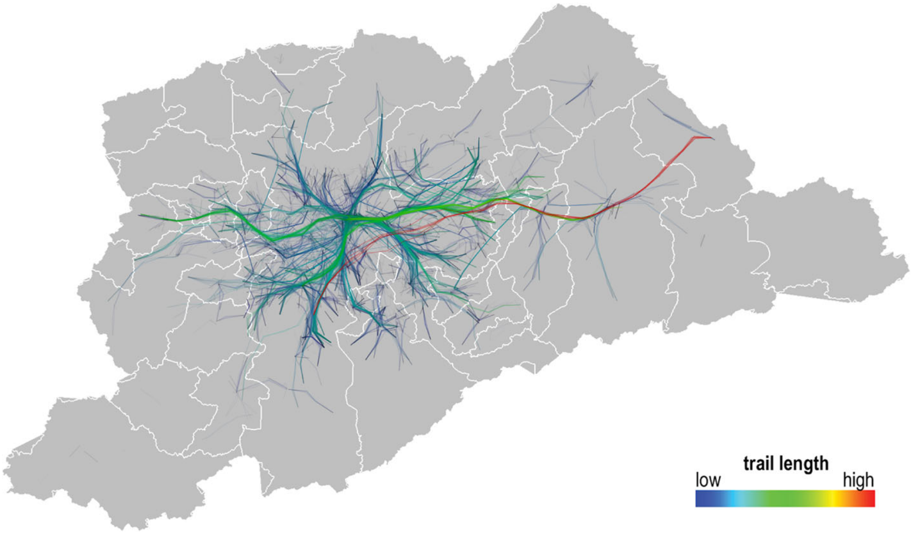


Fig. 31 Distance of trips for shopping reasons

density information of bundles could be to scale edge lines thickness proportionally to the expansion factor of each record instead of using colors, similarly to Lhuillier et al. [25]. This design is worth exploring, as it would eliminate the need to replicate edges and would therefore significantly reduce the size of the dataset. These are important steps to enable the use of bundling for real-time analysis over the Internet.

From an application perspective, there are many other possibilities for urban mobility analysis using data from the OD survey itself and along with other datasets, such as those from private mobility companies, IoT devices, and bike-sharing systems. Last but not least, we also intend to perform a user study to assess the usefulness of bundled visualizations with feedback from actual traffic managers.

Abbreviations

BECC: Brazilian economical classification criterion; CUBu: CUDA bundling; CUDA: Compute unified device architecture; IBB: Image based bundling; KDEEB: Kernel density estimation edge bundling; OD17: 2017 Origin–destination survey; RAEB: Road aware edge bundling; SPMA: São Paulo Metropolitan Area

Acknowledgments

We thank CNPq, CAPES, and FAPESP for the financial support granted to the work.

Authors' contributions

TGM has conducted the research with the guidance of the other authors. HAS has led the writing of this paper. NL, EFZS, AT, FK, and HAS have contributed with ideas, discussion, and effective revision of all sections of the paper. All authors read and approved the final manuscript.

Funding

This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPESP proc. 14/50937-1, and FAPESP proc. 15/24485-9.

Availability of data and materials

The datasets used and/or analyzed during this study area available from the corresponding author on request. The CUBu framework is available at https://webspace.science.uu.nl/~telea001/InfoVis/CUBu. The complete 2017 OD survey dataset is publicly available at http://www.metro.sp.gov.br/pesquisa-od.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Computer Science, University of São Paulo, São Paulo, Brazil. [2] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands.

Received: 29 October 2020 Accepted: 20 July 2021
Published online: 01 September 2021

References

1. Ahlgren B, Hidell M, Ngai EC-H. Internet of things for smart cities: Interoperability and open data. IEEE Internet Comput. 2016;20(6):52–6.
2. Municipality of Seattle USA. City of Seattle Open Data Portal. 2020. https://data.seattle.gov. Accessed 13 July 2020.
3. Vale RCC. The welfare costs of traffic congestion in São Paulo Metropolitan Area. Ribeirão Preto: School of Economics, Business Administration and Accounting – University of São Paulo; 2018. https://teses.usp.br/teses/disponiveis/96/96131/tde-01082018-091126/en.php.
4. Magrini M, Moroni D, Palazzese G, Pieri G, Leone G, Salvetti O. Computer Vision on Embedded Sensors for Traffic Flow Monitoring. In: Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE; 2015. p. 161–6.
5. Bongiorno C, Santucci D, Kon F, Santi P, Ratti C. Comparing bicycling and pedestrian mobility: Patterns of non-motorized human mobility in greater boston. J Transp Geogr. 2019;80:102501.
6. Netek R, Pour T, Slezakova R. Implementation of heat maps in geographical information systems – exploratory study on traffic accident data. Open Geosci. 2018;10(1):367–84.
7. Chen W, Guo F, Wang FY. A Survey of Traffic Data Visualization. IEEE Trans Intell Transp Syst. 2015;16(6):2970–84.
8. Zhou H, Xu P, Yuan X, Qu H. Edge bundling in information visualization. Tsinghua Sci Technol. 2013;18(2):145–56.
9. Lhuillier A, Hurter C, Telea A. State of the Art in Edge and Trail Bundling Techniques. Comput Graph Forum. 2017;36(3):619–45.
10. Van Der Zwan M, Codreanu V, Telea A. CUBu: Universal Real-Time Bundling for Large Graphs. IEEE Trans Vis Comput Graph. 2016;22(12):2550–63.

11. Willems N, Scheepens R, Van De Wetering H, Van Wijk JJ. Visualization of vessel traffic. Situat Aware Syst Syst. 2009;3:73–87.
12. Hurter C, Ersoy O, Fabrikant SI, Klein TR, Telea AC. Bundled Visualization of Dynamic Graph and Trail Data. IEEE Trans Vis Comput Graph. 2014;20(8):1141–57.
13. Blascheck T, Kurzhals K, Raschke M, Burch M, Weiskopf D, Ertl T. Visualization of eye tracking data: A taxonomy and survey. Comput Graph Forum. 2017;36(8):260–84.
14. Martins TG, Lago N, de Souza HA, Santana EFZ, Telea A, Kon F. Visualizing the structure of urban mobility with bundling: A case study of the city of São Paulo. In: Proceedings of the 4th Workshop of Urban Computing (CoUrb 2020). Porto Alegre: SBC; 2020. p. 178–91. https://sol.sbc.org.br/index.php/courb/article/view/12362.
15. Hurter C. Image-Based Visualization: Interactive Multidimensional Data Exploration. San Rafael: Morgan Claypool; 2015.
16. Klein T, van der Zwan M, Telea A. Dynamic multiscale visualization of flight data. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol 1. Piscataway: IEEE; 2014. p. 104–14.
17. Graser A, Schmidt J, Roth F, Brändle N. Untangling origin-destination flows in geographic information systems. Inf Vis. 2017;18(1):153–72.
18. Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell. 2002;24(5):603–19.
19. Hurter C, Ersoy O, Telea A. Graph bundling by kernel density estimation. Comput Graph Forum. 2012;31(3pt1): 865–74.
20. Zeng W, Shen Q, Jiang Y, Telea A. Route-aware edge bundling for visualizing origin-destination trails in urban traffic. Comput Graph Forum. 2019;38(3):581–93.
21. Cui W, Zhou H, Qu H, Wong PC, Li X. Geometry-based edge clustering for graph visualization. IEEE Trans Vis Comput Graph. 2008;14(6):.
22. Ersoy O, Hurter C, Paulovich FV, Cantareiro G, Telea A. Skeleton-based edge bundling for graph visualization. IEEE Trans Vis Comput Graph. 2011;17(12):.
23. von Landesberger T, Kuijper A, Schreck T, Kohlhammer J, van Wijk J, Fekete J-D, Fellner D. Visual analysis of large graphs: State-of-the-art and future research challenges. Comp Graph Forum. 2011;30(6):1719–49.
24. Peysakhovich V, Hurter C, Telea A. Attribute-driven edge bundling for general graphs with applications in trail analysis. In: Proceedings of the 2015 IEEE Pacific Visualization Symposium (PacificVis). Piscataway: IEEE; 2015.
25. Lhuillier A, Hurter C, Telea A. FFTEB: Edge bundling of huge graphs by the fast fourier transform. In: Proceedings of the 2017 IEEE Pacific Visualization Symposium (PacificVis). Piscataway: IEEE; 2017. p. 190–9.
26. Telea A, Ersoy O. Image-based edge bundles: Simplified visualization of large graphs. Comput Graph Forum. 2010;29(3):843–52.
27. Guo D, Zhu X, Jin H, Gao P, Andris C. Discovering spatial patterns in origin-destination mobility data. Trans GIS. 2012;16(3):411–29.
28. Moreira GC, Ceccato VA. Gendered mobility and violence in the São Paulo metro, brazil. Urban Stud. 2021;58(1): 203–22.
29. Slovic AD, Tomasiello DB, Giannotti M, de Fatima Andrade M, Nardocci AC. The long road to achieving equity: Job accessibility restrictions and overlapping inequalities in the city of São Paulo. J Transp Geogr. 2019;78:181–93.
30. Moreno-Monroy AI, Lovelace R, Ramos FR. Public transport and school location impacts on educational inequalities: Insights from São Paulo. J Transp Geogr. 2018;67:110–8.
31. Brazilian Institute of Geography and Statistics (IBGE). Estimativas da população residente no Brasil e unidades da federação com data de referência em 1° de julho de 2020. 2020. http://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2020/estimativa_dou_2020.pdf. Accessed 28 July 2021.
32. Becceneri LB, Alves H. P. d. F., Vazquez DA. Estratificação sócio-ocupacional e segregação espacial na metrópole de São Paulo nos anos 2000. Rev Bras Estud Urbanos Regionais. 2019;21:137–54. https://doi.org/10.22296/2317-1529.2019v21n1p137.
33. Jacobi P. Public and private responses to social exclusion among youth in São Paulo. Ann Am Acad Polit Soc Sci. 2006;606(1):216–30.
34. Kohara LT. Relação entre as condições da moradia e o desempenho escolar: estudo com crianças residentes em cortiços. Ph.D. thesis. 2009. https://teses.usp.br/teses/disponiveis/16/16137/tde-10052010-155909/.
35. Rolnik R, Klintowitz D. (Im)Mobility in the city of São Paulo. Estud Avançados (Adv Stud). 2011;25:89–108.
36. São Paulo Metropolitan Company (Metrô SP). Pesquisa Origem Destino. 2017. https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino.
37. Holten D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Trans Vis Comput Graph. 2006;12(5):741–8.
38. Slocum TA, McMaster RB, Kessler FC, Howard HH. Thematic Cartography and Geovisualization. Upper Saddle River: Pearson Prentice Hall; 2009.
39. Brazilian Association of Research Companies (ABEP). Critério de Classificação Econômica Brasil. 2003. https://www.abep.org/criterioBr/01_cceb_2018.pdf. Accessed 28 July 2021.
40. Municipality of São Paulo. Centros Culturais, Espaços Culturais e Casas de Cultura. 2018. https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/urbanismo/infocidade/htmls/6_centros_culturais_espacos_culturais_e_ca_2017_199.html. Accessed 03 Apr 2021.
41. Municipality of São Paulo. Empregos Formais no Setor do Comércio, segundo Escolaridade. 2018. https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/chamadas/32_trabalho_2018_1583331818.htm. Accessed 03 Apr 2021.
42. de Oliveira PR, Belluzzo W, Pazello ET. The public–private test score gap in Brazil. Econ Educ Rev. 2013;35:120–33.
43. Curi AZ, Menezes Filho NA. Mensalidade escolar, background familiar e os resultados do exame nacional do ensino médio (ENEM). Pesqui Planej Econ. 2013;2(42):223–54.
44. Kamakura W, Mazzon JA. Critérios de estratificação e comparação de classificadores socioeconômicos no Brasil. Rev Adm Empresas. 2016;56(1):55–70. https://doi.org/10.1590/S0034-759020160106.

45. Carruthers R, Dick M, Saurkar A. Affordability of public transport in developing countries. Transp Pap. 2005;TP-3:1–27.
46. Lucas K, Mattioli G, Verlinghieri E, Guzman A. Transport poverty and its adverse social consequences. In: Proceedings of the Institution of Civil Engineers – Transport, vol. 169, issue 6. London: ICE; 2016. p. 353–65.

Publisher's Note

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