scientific reports

OPEN

Activity of vehicles in the bus rapid transit system Metrobús in Mexico City

Jaspe U. Martínez-González[1,2] & Alejandro P. Riascos[1,2✉]

In this paper, we analyze a massive dataset with registers of the movement of vehicles in the bus rapid transit system Metrobús in Mexico City from February 2020 to April 2021. With these records and a division of the system into 214 geographical regions (segments), we characterize the vehicles' activity through the statistical analysis of speeds in each zone. We use the Kullback–Leibler distance to compare the movement of vehicles in each segment and its evolution. The results for the dynamics in different zones are represented as a network where nodes define segments of the system Metrobús and edges describe similarity in the activity of vehicles. Community detection algorithms in this network allow the identification of patterns considering different levels of similarity in the distribution of speeds providing a framework for unsupervised classification of the movement of vehicles. The methods developed in this research are general and can be implemented to describe the activity of different transportation systems with detailed records of the movement of users or vehicles.

The study and understanding of human mobility in cities is an important and challenging problem since more than half of the world population lives in urban areas[1]. Nowadays human mobility can be explored in detail thanks to the digital traces people leave on mobile/digital platforms[2,3]. The identification of patterns in human mobility[4–9] is necessary in topics like urban planning, dealing with traffic congestion[10], the influence of the spatial distribution of a city[1,11–14], the encounters or contacts that emerge[15,16], among many others[1,2,17]. In these problems, the science of networks with well-established tools and methods to characterize and model complex systems[18–20], provide a valuable framework to study transportation modes and their interactions[21–23].

As one type of transit mode, bus rapid transit (BRT) systems have gained popularity worldwide for providing fast and easy access for citizens to fulfill their transportation needs[24] and have been adopted widely over the world[24–27]. The merit of the BRT system lies in its ability to provide a high-quality public transit service with limited infrastructure and at relatively low capital and operating cost[25]. The benefits of a typical BRT system consist of dedicated lanes and proper vehicles and stations; such a layout guarantees a significant advantage in terms of operability[26]. In addition, BRT systems stand to significantly decrease personal vehicle mode share[25] and might pull together connecting parts of the city in ways which other systems do not, especially at the level of service and spatial coherence[28]. In many BRT systems, vehicles have a preinstalled global positioning system (GPS) device which helps in collecting the travel time-related data, this information gives a global picture of the system in real time and can be used for improving the overall performance and schedule adherence of the vehicle. In recent works, the availability of trajectory data collected from operational vehicles in transportation systems has made possible the statistical analysis of travel time of vehicles in roadway segments[29–31], the development of mathematical tools for the estimation of travel times and temporal changes in public transport[32,33], the implementation of techniques to detect patterns in vehicle trajectories[34–36] and public traffic congestion estimation by using artificial neural networks[37–41]. However, approaches to systematically analyze information and identify activity patterns in BRT systems are limited; specifically, very little past research in BRT systems focused on the statistical analysis of the speed of vehicles in specific zones of the system.

In contrast, community detection in networks[42–47] has been proved as an important tool to detect patterns in different complex systems. For example, in the identification of correlations in financial markets[48,49], the study of physiological networks[50], the classification of patents based on their semantic content in technology[51] and pharmaceutics[52], describing the network of stations in bike-sharing systems[53,54] or the geographical structure of the Twitter communication network at the global scale[55], just to mention a few examples of the applicability of this method for pattern detection, unsupervised-classification, and data mining.

[1]Instituto de Física, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510 Mexico City, Mexico. [2]These authors contributed equally: Jaspe U. Martínez-González and Alejandro P. Riascos. ✉email: aperezr@fisica.unam.mx
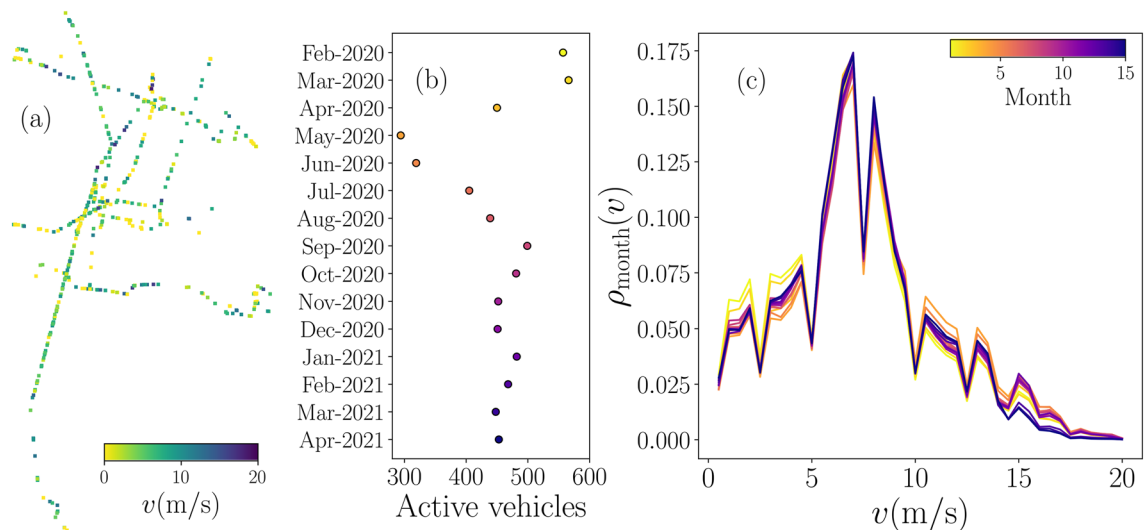
Figure 1. Global characteristics of the BRT system Metrobús in Mexico City. (**a**) Activity of vehicles at 13:00 h observed in registers on March 2nd, 2020. Each active vehicle in the system is represented by a point with speeds v encoded in the color bar. (**b**) Monthly active vehicles from February 2020 to April 2021. (**c**) Probability density $\rho_{\text{month}}(v)$ of the speed v of active vehicles for each month in the color bar [the same colors are shown in the results in panel (**b**)]. We consider bin counts with increments $\Delta v = 0.25$ m/s. All figures were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org).

In this research, we analyze the activity of vehicles in the BRT system Metrobús in Mexico City. The database encompasses 383 days with registers of each active vehicle in the system with GPS geographical coordinates and speeds updated every 30 seconds. For this study, we divide the system into segments. In the first part, we applied statistical methods to characterize the movement of vehicles in each segment by comparing daily activity with the total data. Using the Kullback–Leibler distance between probability densities of speeds we identify zones with regular operation. In the second part, we compare the movement of vehicles in the system using a similarity network. In this structure, each segment is represented as a node, and links are added when the probability densities of speeds in two segments are similar. The exploration of different levels of similarity in terms of a parameter H define networks for which community detection algorithms allow an unsupervised classification of the segments based on the speeds of vehicles. The methods introduced are general and provide a framework for the study of different transportation systems in cities when massive databases with geolocalized activity are available. This approach will help to a better understanding of variations in the speeds over space and time by means of statistical analyses and complement other techniques such as time reliability-based performance indicators introduced to study the travel time variations of vehicles in specific routes.

Results

Global characteristics of the system.
The dataset analyzed in this research is part of an effort of Mexico City to have open databases for transportation systems, human mobility and other topics of interest[56]. In particular, for the BRT system Metrobús it is possible to have access to real-time registers with geographical coordinates (longitude, latitude) for positions and speeds of all the active vehicles in the system[57]. We analyze data for 383 days collected from February 2020 to April 2021 (see "Methods" section for a detailed description of the data). From these registers of activity we have the speed of each vehicle in m/s. In this way, we have access to a global picture of the system's activity at specific moments.

In Fig. 1 we illustrate a global analysis of the dataset. In Fig. 1a we show the positions and speeds of the vehicles in March 2nd, 2020 at 13:00 h, the system has 516 active vehicles at this moment. Each vehicle is depicted with a point and the color encodes its speed (see Supplemental Material with a video of the complete data collected for this day). This representation gives a general overview of the data, such as regions with the highest activity, and zones where the speed of the units is higher than the average. On the other hand, considering that each vehicle has a unique ID, we can count the total number of active vehicles in the collected data on a determined time scale. In Fig. 1b, we show the number of active units at the scale of months considering the 15 months covered in this research. In this analysis, we assume that a vehicle is active in the system if at least has one register in the respective month. The results show that the number of vehicles changes significantly, especially due to the modifications in the system introduced in response to the different stages of the COVID-19 pandemic in Mexico City. In this respect, the number of vehicles in February and March of 2020 represents the common operation pre-COVID-19 in Mexico. This number reduced, particularly, from April to August 2020. In the last 8 months of this study, the number of active vehicles increased from the low in May 2020 but not to the same levels observed in the first two months. To complement this part, in Fig. 1c, we analyze the information of the speeds of the vehicles in the whole system in each month. We explore the probability density $\rho_{\text{month}}(v)$ of non-null speeds v for the records of each month in the dataset. In strong contrast with the results for the number of
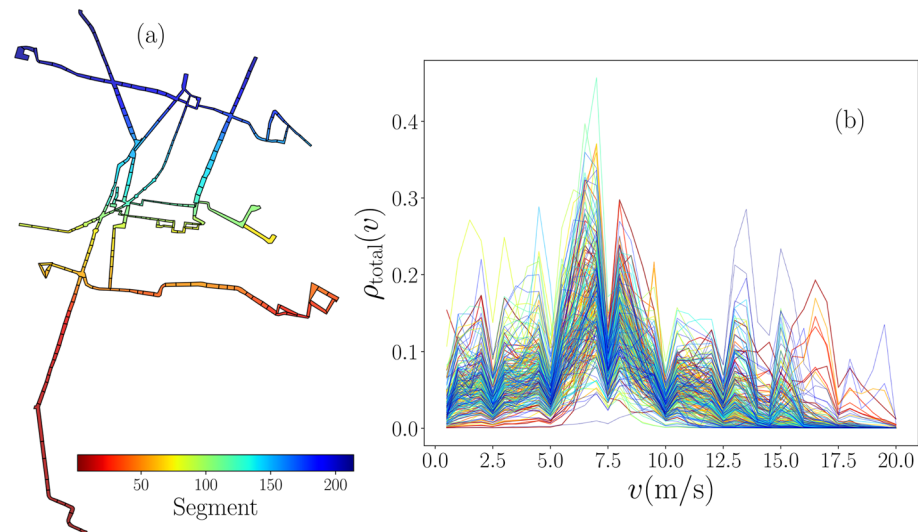
Figure 2. Activity of vehicles in segments of the system Metrobús in Mexico City. (**a**) Map with polygons associated to each of the $\mathcal{N} = 214$ segments in the system. (**b**) Probability density $\rho_{total}(v)$ of the speed v of vehicles in each segment. The values $\rho_{total}(v)$ are obtained from bin counts in the interval $0 < v \leq 20$ m/s with increments $\Delta v = 0.25$ m/s. Segments are sorted considering their geographical coordinates with a number codified in the color bar that applies for polygons in (**a**) and the respective $\rho_{total}(v)$ in (**b**). All figures were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org). The map in panel (**a**) was created using the geopandas (0.10.2) package (https://geopandas.org).

vehicles, we see that the distributions of speeds of the vehicles present similar characteristics with small changes at lower speeds $v \leq 5$ m/s and higher speeds $v \geq 10$ m/s, minimal variations are observed for 5 m/s $< v < 10$ m/s.

Vehicle activity in segments. In addition to the vehicles, the infrastructure of the system Metrobús includes 195 stations where users access this service and are distributed in 7 lines with 225 km exclusive roadways dedicated to buses. We use all the information available for the movement of vehicles to study the operation of the system in different zones of Mexico City. To this end, we divide the system into $\mathcal{N} = 214$ segments defined by polygons that include stations and the lanes that connect them. A simple segment is described by an elongated rectangle defining a specific geographical region that includes the system's roads and stations at two of its ends. In this study, our partition of the systems considers 205 simple segments. In addition, 9 segments are general polygons, located in zones where different lines converge. In Fig. 2a, we present all the segments of the system. In this representation, polygons are sorted according to the geographical coordinates of their geometrical centers starting from the south-west and considering the latitudes (from south to north) as the first variable and the longitudes (from west to east) as the second variable. An index $i = 1, 2, \ldots, \mathcal{N}$ codified in the color bar denotes the segment number; we maintain the same index in all the following analyses.

Once we define the information of polygons representing the segments, we proceed to analyze the activity of vehicles in each one of them. Due to the high volume of data, we use the geographical coordinates of the vehicles to divide the complete dataset into registers associated with each segment. In this manner, we have all the speeds in each segment during the 383 days in our study. The statistical analysis of the speeds v considering the movement of vehicles (i.e. only registers with $v > 0$ m/s corresponding to 164 867 137 speed values, a 76.7% of the total database) are presented in Fig. 2b, results show the probability density $\rho_{total}(v)$ calculated with the relative frequencies of v in regular bin counts with $\Delta v = 0.25$ m/s. In this representation, we maintain the same colors that codified the segments in Fig. 2a. We define a maximum speed $v_{max} = 20$ m/s $= 72$ km/h, only a 0.0114% of the total database contains registers with $v > v_{max}$. In addition, our analysis shows that in a high number of segments, most frequent values of speeds are observed in the interval 5 m/s $\leq v \leq 10$ m/s.

Also, the values v can be divided into sets considering a defined time window; for example, registers in a particular hour, day, weekday, month, among others. Once we establish a particular partition of the speeds, we have probability densities that we can compare with the total density $\rho_{total}(v)$ in each segment. On a temporal scale of days, we obtain 383 densities $\rho_{day}(v)$ per segment with the information of the movement of vehicles in each day considered in our study. Then, using the Kullback–Leibler distance[58]

$$\mathscr{D}_{KL} \equiv \int_0^{v_{max}} \rho_{day}(v) \log \left[\frac{\rho_{day}(v)}{\rho_{total}(v)} \right] dv, \tag{1}$$

we calculate the "*distance*" between each $\rho_{day}(v)$ and $\rho_{total}(v)$ (see the "Methods" section for a discussion about $\mathscr{D}_{KL}$). In this way, we have 383 values of $\mathscr{D}_{KL}$ in each segment comparing daily registers with the respective $\rho_{total}(v)$ in Fig. 2b.
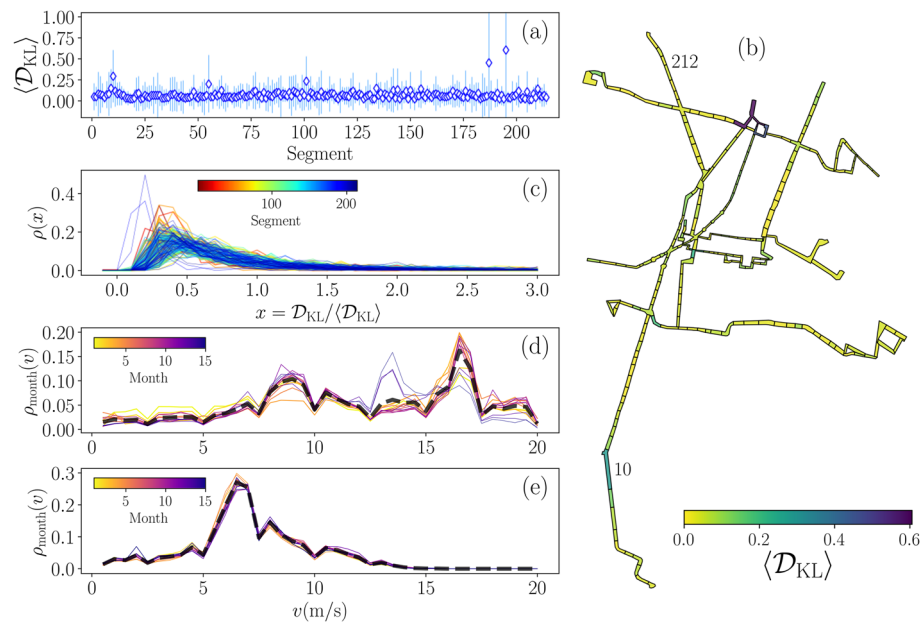
Figure 3. Statistical analysis of the distances $\mathscr{D}_{KL}$ for speed densities. (**a**) Average Kullback–Leibler distances $\langle\mathscr{D}_{KL}\rangle$ in each segment, error bars represent the standard deviation of the values. (**b**) Average values $\langle\mathscr{D}_{KL}\rangle$ represented in the segments map. (**c**) Probability densities of the values $\mathscr{D}_{KL}/\langle\mathscr{D}_{KL}\rangle$ in each segment codified in the color bar. Probability densities $\rho_{month}(v)$ of the speed v at the scale of months for: (**d**) segment 10 with $\langle\mathscr{D}_{KL}\rangle = 0.296$ and (**e**) segment 212 with $\langle\mathscr{D}_{KL}\rangle = 0.050$, dashed lines represent $\rho_{total}(v)$ for the segment. All figures were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org). The map in panel (**b**) was created using the geopandas (0.10.2) package (https://geopandas.org).

In Fig. 3 we present the statistical analysis of the distances $\mathscr{D}_{KL}$ found. Figure 3a shows the average values $\langle\mathscr{D}_{KL}\rangle$, error bars are obtained with the standard deviation of the values in each segment $\sigma_{KL} = \sqrt{\langle\mathscr{D}_{KL}^2\rangle - \langle\mathscr{D}_{KL}\rangle^2}$. We observe that the average values $\langle\mathscr{D}_{KL}\rangle$ lie in the interval $0.0204 \leq \langle\mathscr{D}_{KL}\rangle \leq 0.6058$ and 209 segments present distances that can be considered as small with $\langle\mathscr{D}_{KL}\rangle \leq 0.155$. In contrast, in Fig. 3a we identify 5 segments $i = 10, 55, 101, 187, 195$ with average distances $\langle\mathscr{D}_{KL}\rangle > 0.2$. A detailed analysis of the Kullback–Leibler distances for these polygons reveal that in some days the distribution $\rho_{day}(v)$ differs of the respective $\rho_{total}(v)$, this may be due to modifications in the routes of vehicles. In Fig. 3b we plot the map of the system Metrobús, the colors represent the value $\mathscr{D}_{KL}/\langle\mathscr{D}_{KL}\rangle$ of each segment. This map allows the identification of zones with regularity in the movement of vehicles (small values of $\rho(\mathscr{D}_{KLS})$) and particular segments where average distances $\langle\mathscr{D}_{KL}\rangle > 0.2$ show that the daily distributions differ with the total activity captured in $\rho_{total}(v)$. In Fig. 3c we depict the statistical analysis of the values $x = \mathscr{D}_{KL}/\langle\mathscr{D}_{KL}\rangle$ in each segment, the respective probability densities $\rho(x)$ are skew with a high fraction of distances below the average, (i.e. with $x < 1$).

In addition to the results for the $\mathscr{D}_{KL}$ distances, we show in Fig. 3d,e the monthly distribution of speeds $\rho_{month}(v)$ for two particular segments. Our analysis is similar to the presented for the total system in Fig. 1c, but now for the vehicles in segment 10 with $\langle\mathscr{D}_{KL}\rangle = 0.296$ and segment 212 with $\langle\mathscr{D}_{KL}\rangle = 0.050$ (the numbers of these particular segments are included in the map in Fig. 3b). With different colors, we represent the 15 months considered in our study. In the case of the segment 10 with higher distance $\langle\mathscr{D}_{KL}\rangle$, the values of $\rho_{month}(v)$ suffered deviations with respect with the total data $\rho_{total}(v)$ represented with a dashed line. On the other hand, in the segment 212 with $\langle\mathscr{D}_{KL}\rangle$ closer to zero, the $\rho_{month}(v)$ remain approximately the same as $\rho_{total}(v)$ for the total data in this segment.

The results in Fig. 3d,e are two particular examples to show how $\langle\mathscr{D}_{KL}\rangle$ allows identifying variations in the activity of vehicles in a particular segment, similar analyses can be implemented at different scales of time to characterize changes or the regularity in the vehicular activity in a particular region of the system. In Fig. 3, it can be seen that except for the 5 segments discussed above, the activity of vehicles (described by $\rho_{total}(v)$ in each segment) maintains some regularity. This result is important considering the variations in the number of active vehicles in Fig. 1b. A particular case is found in segment 10 that includes stations close to the National Autonomous University of Mexico. Since many of the activities in this university were developed virtually, the number of users in these stations was reduced significantly. As a consequence of the low demand, the movement of vehicles in this segment was restructured as we can see in Fig. 3d.

Network of similarity between segments. Now we compare all the speed probability densities associated with the activity of vehicles in each segment. We consider the complete probability density for all segments $i = 1, 2, \ldots, 214$, with the records $0 < v \leq 20$m/s in our database, denoted as $\rho_{total}^{(i)}(v)$ and presented in Fig. 2b.
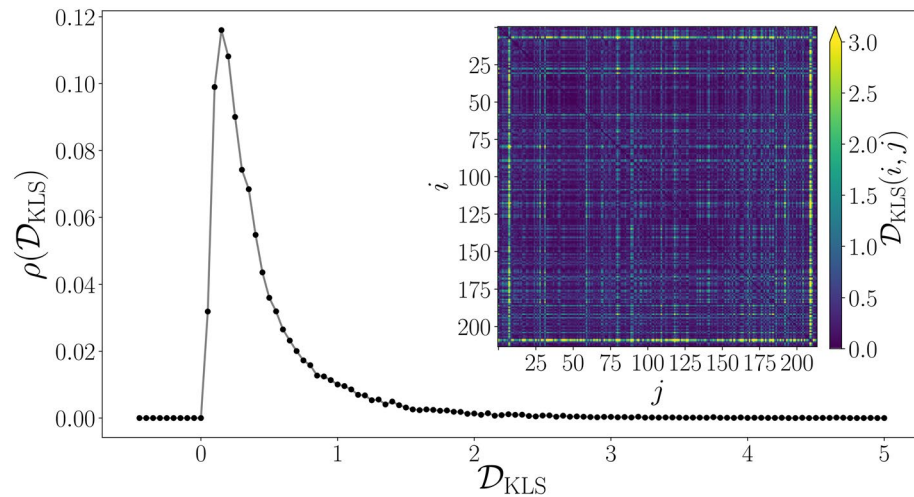
Figure 4. Statistical analysis of symmetric Kullback–Leibler distances between segments. The values $\mathscr{D}_{\text{KLS}}$ are obtained from Eq. (2) for all the segment pairs $i, j = 1, \ldots, 214$. We present the probability density $\rho(\mathscr{D}_{\text{KLS}})$ obtained with bin counts in intervals with $\Delta \mathscr{D}_{\text{KLS}} = 0.05$, the inset shows the matrix representation with entries $\mathscr{D}_{\text{KLS}}(i, j)$ codified in the color bar. All figures were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org).

In this case, it is convenient using a symmetric distance $\mathscr{D}_{\text{KLS}}(i, j)$ between two probability densities of the segments i and j obtained with the average of the Kullback–Leibler distance, hence we define

$$\mathscr{D}_{\text{KLS}}(i, j) = \frac{\mathscr{D}_{\text{KL}}(i, j) + \mathscr{D}_{\text{KL}}(j, i)}{2} \tag{2}$$

with

$$\mathscr{D}_{\text{KL}}(i, j) \equiv \int_0^{\nu_{\max}} \rho_{\text{total}}^{(i)}(\nu) \log \left[\frac{\rho_{\text{total}}^{(i)}(\nu)}{\rho_{\text{total}}^{(j)}(\nu)} \right] d\nu. \tag{3}$$

With the definition in Eq. (2), we obtain a $\mathscr{N} \times \mathscr{N}$ symmetric matrix with the information of similarity between the vehicle activity in the segments, the value $\mathscr{D}_{\text{KLS}} = 0$ is obtained for two equal speed distributions and values $\mathscr{D}_{\text{KLS}}$ large show cases where the segments have a completely different activity. Our findings are depicted in Fig. 4, where we present as an inset the matrix with elements $\mathscr{D}_{\text{KLS}}(i, j)$ for $i, j = 1, \ldots, 214$. We also statistically analyze all the entries of the matrix of distances, obtaining the probability density $\rho(\mathscr{D}_{\text{KLS}})$. Both representations show that a high fraction of the distances $\mathscr{D}_{\text{KLS}}$ between segments have values in the interval $0 \leq \mathscr{D}_{\text{KLS}} \leq 0.5$, revealing different degrees of similarity in the activity in segments.

Looking for a better understanding of the similarities in all the distributions presented in Fig. 2b, we use the values of the distances in Fig. 4 to define a similarity network. In this representation, each node is associated with a segment of the BRT system, the size of the network is $N = 214$ and an edge connecting two different nodes i and j is established if $\mathscr{D}_{\text{KLS}}(i, j) \leq H$ where H is a given threshold limit to decide if two segments have similar activity. The result is an undirected network with an adjacency matrix $\mathbf{A}$ with elements $A_{ij} = 1$ if $0 < \mathscr{D}_{\text{KLS}}(i, j) \leq H$ and $A_{ij} = 0$ for $\mathscr{D}_{\text{KLS}}(i, j) > H$. By definition, the adjacency matrix considers the diagonal entries $A_{ii} = 0$ to avoid loops or connections of a node to itself. In the general case, it is hard to have intuition about the values of H to define a similarity network and its choice depends on the particular structure of the dataset explored and the metric used for the distance or relation between two nodes. In this respect, it is convenient to perform the statistical analysis of all the entries $\mathscr{D}_{\text{KLS}}(i, j)$ presented in Fig. 4. In this representation of $\rho(\mathscr{D}_{\text{KLS}})$, the area under the curve $\int_0^H \rho(z) dz$ gives the fraction of edges included in the network with respect to a fully connected graph. The higher the threshold H, the more edges are included in the network. In particular, in our analysis for the BRT system Metrobús, we see that the interval $0 < H \leq 1$ could produce networks with useful information. For $H \gg 1$, a high fraction of the edges are included in the similarity network losing any particular structure at the level of groups of zones. For the readers interested in this part, we refer to the recent work of Rincón et al.[52], where a network of patents is explored using similar methods with a metric implemented to compare keywords in texts.

In this way, for each value of H we have a network of segments and we can apply standard methods for its analysis. In Fig. 5 we centered our study on the different networks generated and in the largest connected component (LCC), defined by the largest set of connected nodes within the network. In Fig. 5a we present the number of nodes in the LCC as a function of the similarity threshold H in the interval $0.01 \leq H \leq 1$. In this representation of the results, we observe the effect of H defining different scales in the similarity between segments. For small values $H < 0.02$ the similarity network is formed by small size disconnected clusters showing that a reduced
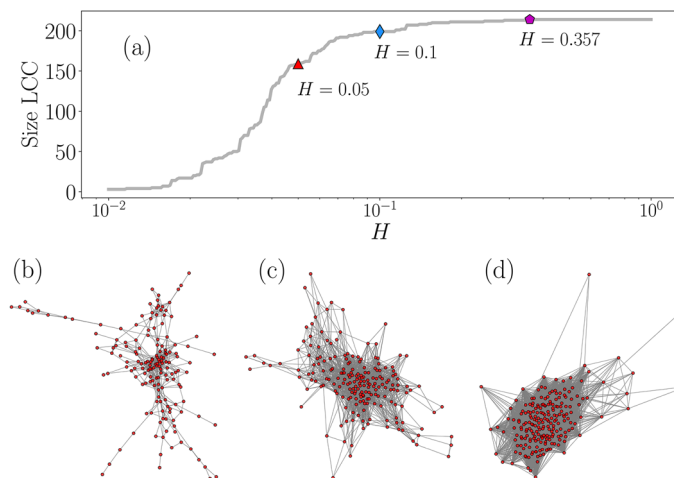
Figure 5. Similarity networks in the activity of vehicles in the BRT system Metrobús. (**a**) Size of the largest connected component (LCC) as a function of the threshold value H for the similarity between segments. Three particular LCCs obtained for $H = 0.05$, $H = 0.1$ and $H = 0.357$ are presented in panels (**b**)–(**d**), respectively. $H = 0.357$ is the lowest value of H that produces a connected network including the 214 segments. All figures were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org). Networks in panels (**b**)–(**d**) were created using the networkx (2.6.3) package (https://networkx.org).

number of probability densities are almost identical. This behavior changes in the interval $0.02 \leq H \leq 0.1$ where the size of the LCC increases monotonically with H. For $H > 0.1$ a high fraction of the network is connected and in $H \geq 0.357$ the LCC includes all the $N = 214$ nodes. Here $H = 0.357$ is the lowest value of H that produces a connected network including the 214 segments.

In this manner, all the information contained in the distance matrix can be analyzed considering different degrees of similarity of probability densities establishing connections between nodes as the structures in Fig. 5b–d obtained for the values $H = 0.05$, $H = 0.1$ and $H = 0.357$. In the case with $H = 0.05$, each edge requires high similarity between two segments and the LCC contains $N = 159$ nodes, with an average degree $\langle k \rangle = 7.7$, a global clustering coefficient $\langle C \rangle = 0.410$ indicating that the structure has a low fraction of triangles; also, the average number of edges in the shortest path connecting two nodes in the network is $\langle l \rangle = 4.42$ (see "Methods" section for a formal definition of $\langle k \rangle, \langle C \rangle, \langle l \rangle$ for networks with N nodes). In contrast, for $H = 0.1$ the similarity network includes more edges defining a LCC with $N = 199$ nodes, $\langle k \rangle = 29.0$, $\langle C \rangle = 0.598$ and $\langle l \rangle = 2.64$ revealing a more connected structure. For $H = 0.357$ the LCC contains all the $N = 214$ nodes in a network with $\langle k \rangle = 127$, $\langle C \rangle = 0.822$ and $\langle l \rangle = 1.48$ allowing a coarse-graining description of the segments in the system Metrobús. In networks with $H \gg 0.357$, increasing H we lose information of the similarity between segments and the network gradually approaches to a fully connected graph.

Community structure and identification of patterns. In networks, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of nodes and low connectivity between these groups. This feature in networks is called community structure. Communities (also clusters or modules), are sets of vertices that probably share common properties and/or play similar roles within the network[45]. Community detection endorses the identification of local connectivity patterns and guides the understanding of interactions in a complex structure. In this work, the communities represent groups of segments with similar activity of vehicles, which arose from considering all the information contained in the similarity network for different thresholds H, something not immediately visible if comparing the probability densities of speeds by pairs of segments. We apply modularity-based clustering algorithms[42,43] to analyze the community structure of the LCC of similarity networks generated with $H = 0.05$ and $H = 0.357$ and depicted in Fig. 5b,d, the results derived from the community structure are presented in Figs. 6, 7 and Table 1.

In Fig. 6 we explore the Metrobús system using the similarity network with threshold limit $H = 0.357$. This network has two communities: $\mathscr{C}_1$ with 108 nodes and $\mathscr{C}_2$ formed by 106 nodes; the community structure is represented in Fig. 6a, other quantities that characterize the subnetworks defining each cluster (average degree $\langle k \rangle$, global clustering $\langle C \rangle$ and average lengths of the shortest path connecting different nodes $\langle l \rangle$) are presented in Table 1(a). In Fig. 6b we show the communities in a map, in which segments in communities $\mathscr{C}_1$ and $\mathscr{C}_2$ are distributed in the whole system. We also present the total density of speeds $\rho_{\text{total}}^{(i)}(v)$ for the segments $i \in \mathscr{C}_1$ and $i \in \mathscr{C}_2$ (see panels in Fig. 6c,d), dashed lines are generated with the registers of v in all the segments representing the total activity in each community. The speed distributions found in the communities $\mathscr{C}_1$ and $\mathscr{C}_2$ present particular features. One of the differences observed is that in $\mathscr{C}_1$ the speeds $v \geq 10\,\text{m/s}$ appear in a 30.46% of the non-null records; whereas, only a 9.34% in $\mathscr{C}_2$ fulfill this condition. In this way, the classification of segments through community detection in similarity networks with $H = 0.357$ establishes a coarse-grained classification with two categories: $\mathscr{C}_1$ for high-speed segments and $\mathscr{C}_2$ for low-speed zones. The average speeds $\bar{v}$ in each community reported in Table 1(a) also confirm this characteristic observed in both categories, we also include the
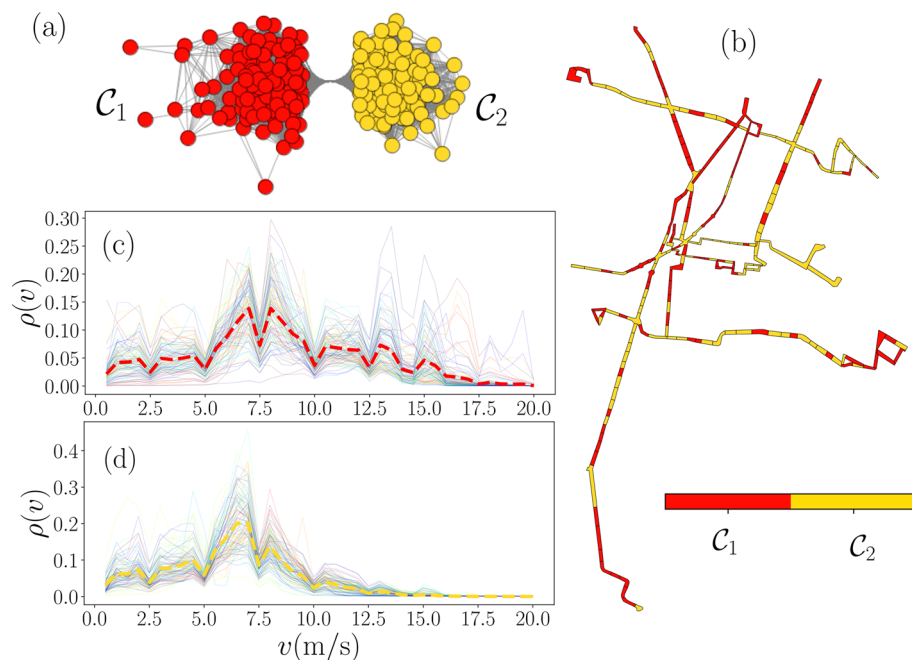
Figure 6. Community detection and activity of vehicles for similarity networks. (**a**) Similarity network with threshold value $H = 0.357$ with two communities $\mathscr{C}_1$ and $\mathscr{C}_2$. (**b**) Segments map representing the communities. Probability densities $\rho(v)$ for the segments in $\mathscr{C}_1$ (**c**) and $\mathscr{C}_2$ (**d**). In panels (**c**) and (**d**), thin lines represent the $\rho_{\text{total}}^{(i)}(v)$ for each segment as in Fig. 2b and thick dashed lines depict the probability density found for all the values of v in $i \in \mathscr{C}_1$ and $i \in \mathscr{C}_2$. The network in panel (**a**) was generated using Mathematica 12.3.1 (https://www.wolfram.com/mathematica/). (**b**)–(**d**) were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org). The map in panel (**b**) was created using the geopandas (0.10.2) package (https://geopandas.org).

Community	Segments	$\langle k \rangle$	$\langle C \rangle$	$\langle l \rangle$	$\bar{v}$(m/s)	σ_v(m/s)	$\% \, v \geq 10$ m/s
(a) $H = 0.357$							
$\mathscr{C}_1$	108	77.11	0.877	1.32	8.07	4.03	30.46
$\mathscr{C}_2$	106	89.1	0.921	1.15	5.99	2.97	9.34
(b) $H = 0.05$							
$\mathscr{C}_1$	55	11.7	0.625	2.33	6.41	3.17	13.28
$\mathscr{C}_2$	41	3.85	0.287	3.36	8.05	3.81	30.69
$\mathscr{C}_3$	32	4.75	0.494	2.83	5.71	2.61	5.31
$\mathscr{C}_4$	12	2.67	0.281	2.41	6.93	3.09	14.2
$\mathscr{C}_5$	10	3	0.250	1.91	4.73	2.69	3.16
$\mathscr{C}_6$	4	1.5	0	1.67	8.68	3.57	37.81

Table 1. Classification of segments using community detection. Analysis of similarity networks for (a) $H = 0.357$ and (b) $H = 0.05$. The number of segments in each community is reported with the average degree $\langle k \rangle$, global clustering $\langle C \rangle$, and average lengths $\langle l \rangle$ of the shortest path connecting the nodes for each subnetwork. For the non-null speeds in the segments in each community $\bar{v}$ denotes the average speed and σ_v the standard deviation, we also report the fraction (as percentage) of speeds with $v \geq 10$ m/s.

standard deviations σ_v of the speed values. In the results in Fig. 6c,d, for $\rho_{\text{total}}^{(i)}(v)$ (represented with thin lines for each segment i), some probability densities deviate from the result obtained with the records for the whole community, something that is more marked in the $\mathscr{C}_1$ community. This is because having the threshold $H = 0.357$, some differences are allowed in the similarity network.

In Fig. 7, we present the results for the analysis using $H = 0.05$, a criterion that requires greater similarity to form a link in the network of segments. In this case, community detection algorithms applied to the LCC allow defining 6 communities $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_6$ that contain at least four segments as we illustrate in Fig. 7a. In the map in Fig. 7b, we see that these categories produce a more varied map, although as H is small, several segments (represented in white) cannot be grouped into a community, being outside the LCC (we also omitted a
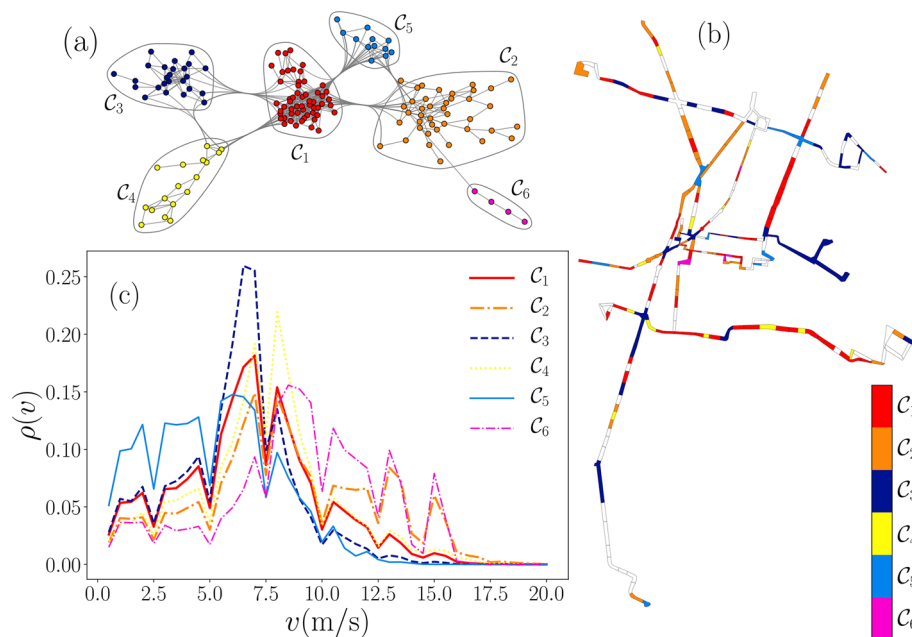
Figure 7. Segment classification through similarity networks with $H = 0.05$. (**a**) Largest connected component for a similarity network with communities $\mathscr{C}_l$ ($l = 1, 2, \ldots, 6$) represented with different colors. (**b**) Segments map with the representation of communities. (**b**) Probability densities $\rho(v)$ for all the values in the segments in each community. The network in panel (**a**) was generated using Mathematica 12.3.1 (https://www.wolfram.com/mathematica/). Figures (**b**)–(**c**) were created using python 3.8 and the matplotlib (3.4.3) package (https://matplotlib.org). The map in panel (**b**) was created using the geopandas (0.10.2) package (https://geopandas.org).

community defined by three segments in the LCC). In Fig. 7c we present the probability density for v with the records of non-null speeds in the segments of each community. In this case, the unsupervised classification of the segments produces more varied results that are reported in Table 1(b). For example, in the analysis of the proportion in which the velocities $v \geq 10$ m/s appear, $\mathscr{C}_2$, $\mathscr{C}_6$ define groups of segments with high speeds where more than 30% of the data meet $v \geq 10$ m/s. Communities $\mathscr{C}_1$, $\mathscr{C}_4$ have a fraction of around 14% for these speeds, and $\mathscr{C}_3$, $\mathscr{C}_5$ define low-speed zones with less than 10%. This type of classification is also evident for the average velocities $\bar{v}$ in each community, being the highest $\bar{v} = 8.68$ m/s in community $\mathscr{C}_6$ and the lowest $\bar{v} = 4.73$ m/s obtained for $\mathscr{C}_5$. The measures that describe the communities as networks also give us important information; for example, $\mathscr{C}_1$ and $\mathscr{C}_2$ are the subnetworks that have more nodes; however, considering the links, $\mathscr{C}_1$ is much more connected, a fact that is evidenced in the highest average degree and clustering. Finally, the most valuable information is the probability density $\rho(v)$ in Fig. 7c for the data in each community. The obtained distributions have particular characteristics that describe the vehicular movement in each group of segments. All this information and the map in Fig. 7b, help us to understand the global activity of the Metrobús system and its operation since the results obtained with the combination of methods implemented in this research allow us to detect emerging patterns when comparing the activity of the entire system.

Discussion

From the study of the data with the activity of vehicles in the BRT system Metrobús for 383 days and a partition of the regions where these vehicles move defining 214 segments, it is found that this system operates with relative regularity in each zone. In particular, the distributions of speeds in the entire system and in each of the segments are preserved, presenting small variations depending on the day, with some exceptions that also can be detected using the statistical methods implemented for this study. In this way, the speed distribution of each segment is a good reference for the specific behavior of the vehicles in each of the geographical zones defining the segments. The variations in each segment at different temporal scales can be effectively studied using Kullback–Leibler distances.

In addition, the analysis of the Kullback–Leibler distance between speed distributions of all pairs of segments allows the representation of the entire system as a network. In this structure, community detection algorithms allow identifying groups of segments with similar vehicular activity, the number of communities found varies according to parameter H required to define the similarity network. In a case with $H = 0.357$ two categories are established, one with segments in which speeds with $v \geq 10$ m/s are more frequent and another in which these speeds appear less frequently. The analysis of a network with a greater similarity between segments with $H = 0.05$ gives a classification with more specific characteristics in the speed distributions, in this case the records with $v \geq 10$ m/s appear in different proportions in each community. Our findings show that the Metrobús BRT system presents certain regularity in its operation, in the sense that the distribution of speeds of vehicles in 209 segments of the system suffered only small variations even with the reduction of active vehicles implemented

due to the COVID-19 pandemic in Mexico City. It seems plausible to associate this regularity with the exclusive lanes in the system and the rules that operators of the vehicles must follow.

The statistical methods and the network science approach implemented in this research can be used for the multi-scale study of different transportation systems. In systems like taxis, buses, car-sharing services, in which large amounts of data are available with registers of the movement or quantities associated with vehicles or agents along with their geographic coordinates, this approach can lead to the unsupervised detection of regions with similar activity of vehicles. Other studies can incorporate the statistical analysis of different quantities of interest; for example, schedule adherence of the vehicles, carbon emissions, or user's accessibility to stations. A profound understanding of the vehicle activity and similarities detected in groups of segments can help researchers and transit specialists to draw up strategies tailored to improving operational aspects of the system.

Methods

Dataset description.
With the implementation of location-enabled devices on public transportation, a large amount of bus trajectory data is being generated. Since April 2019 all the information of the movement of all active vehicles in the BRT system Metrobús is available under request to the public[57]. The information provided contains the timestamp, vehicles ID, and registers in real-time of the GPS coordinates (longitude, latitude) of each vehicle in the system, their speeds in meters per second (m/s), and qualitative descriptions of the state of each vehicle or the levels of congestion for example: *on time*, *stopped*, among others. Each vehicle updates this information every 30 seconds. The Metrobús system operates from 4:30 to 00:00 h on weekdays (Monday to Friday) and starts operation at 5:00 h on weekends (Saturday, Sunday) and holidays, the description of stations and routes is available to the public in the webpage of Metrobús[59]. By using a code written in Python, we request the data automatically (waiting 30 seconds between requests), an initial treatment of the retrieved records is performed to save the data. We maintained the download of data from February 16th, 2020 to April 8th, 2021. In total, we have data for 383 days with 215 025 258 registers of position and speed of vehicles. On some particular days, the data was not available for download due to maintenance or problems connecting with the server.

Kullback–Leibler distance.
The Kullback–Leibler distance is a standard method to calculate the difference between two probability distributions $P(z)$ and $Q(z)$ describing a stochastic variable z[58,60]. This tool is widely used for database comparison. For continuous distributions, this distance is given by[58]

$$\mathscr{D}_{KL}(P||Q) = \int P(z) \log \left[\frac{P(z)}{Q(z)} \right] dz. \tag{4}$$

Here Q acts as a reference distribution. Also, it is important to emphasize that $\mathscr{D}_{KL}(P||Q)$ is not a distance in the sense of a metric since the distance between P and Q is not necessarily the same as between Q and P. Also, from the definition in Eq. (4), it is clear that $\mathscr{D}_{KL}(P||Q) > 0$ and is null when $P = Q$.

Networks.
Symmetric networks with N nodes are described by an adjacency $N \times N$ adjacency matrix $\mathbf{A}$ with entries 1 if two different nodes are connected and 0 otherwise. An important quantity in the study of networks is the degree of node i given by $k_i = \sum_{l=1}^{N} A_{il}$, that gives the number of connections to that node. In terms of this quantity we define the average degree as

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i. \tag{5}$$

Another measure to characterize the topology of networks is the clustering coefficient[18]. This coefficient C_i of the node i, quantifies the fraction of connected neighbors $\triangle_i$ of the node i with respect to the maximum number of these connections given by $k_i(k_i - 1)/2$. In terms of the adjacency matrix we have for $k_i \geq 2$

$$C_i = \frac{(\mathbf{A}^3)_{ii}}{k_i(k_i - 1)}, \tag{6}$$

otherwise $C_i = 0$. Here $(\mathbf{A}^3)_{ii} = (\mathbf{AAA})_{ii} = \triangle_i/2$. The average clustering coefficient is given by

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{7}$$

From the information in the adjacency matrix, different algorithms allow the calculation of the shortest path connecting the nodes i and j, the length l_{ij} with the number of edges in this shortest path is a measure of the distance between two nodes[18]. This information allows defining an average distance $\langle l \rangle$ given by

$$\langle l \rangle = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} l_{ij}. \tag{8}$$

In this way, for a particular connected undirected network with N nodes, we can calculate the adjacency matrix $\mathbf{A}$ and obtain the global quantities $\langle k \rangle$, $\langle C \rangle$ and $\langle l \rangle$ that describe this structure.

References

1. Batty, M. *The New Science of Cities* (MIT Press, 2013).
2. Barthélemy, M. *The Structure and Dynamics of Cities: Urban Data Analysis and Theoretical Modeling* (Cambridge University Press, 2016).
3. Barbosa, H. *et al.* Human mobility: Models and applications. *Phys. Rep.* **734**, 1–74. https://doi.org/10.1016/j.physrep.2018.01.001 (2018).
4. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782. https://doi.org/10.1038/nature06958 (2008).
5. Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823. https://doi.org/10.1038/nphys1760 (2010).
6. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: Universal patterns in human urban mobility. *PLoS One* **7**, e37027. https://doi.org/10.1371/journal.pone.0037027 (2012).
7. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100. https://doi.org/10.1038/nature10856 (2012).
8. Loaiza-Monsalve, D. & Riascos, A. P. Human mobility in bike-sharing systems: Structure of local and non-local dynamics. *PLoS One* **14**, e0213106. https://doi.org/10.1371/journal.pone.0213106 (2019).
9. Riascos, A. P. & Mateos, J. L. Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. *Sci. Rep.* **10**, 4022. https://doi.org/10.1038/s41598-020-60875-w (2020).
10. Pérez-Méndez, D., Gershenson, C., Lárraga, M. E. & Mateos, J. L. Modeling adaptive reversible lanes: A cellular automata approach. *PLoS One* **16**, e0244326. https://doi.org/10.1371/journal.pone.0244326 (2021).
11. Portugali, J. *Complexity, Cognition and the City* (Springer, 2011).
12. Louail, T. *et al.* From mobile phone data to the spatial structure of cities. *Sci. Rep.* **4**, 5276. https://doi.org/10.1038/srep05276 (2014).
13. Lee, M. & Holme, P. Relating land use and human intra-city mobility. *PLoS One* **10**, e0140152. https://doi.org/10.1371/journal.pone.0140152 (2015).
14. Riascos, A. P. Universal scaling of the distribution of land in urban areas. *Phys. Rev. E* **96**, 032302. https://doi.org/10.1103/PhysRevE.96.032302 (2017).
15. Riascos, A. P. & Mateos, J. L. Emergence of encounter networks due to human mobility. *PLoS One* **12**, e0184532. https://doi.org/10.1371/journal.pone.0184532 (2017).
16. Riascos, A. P. & Sanders, D. P. Mean encounter times for multiple random walkers on networks. *Phys. Rev. E* **103**, 042312. https://doi.org/10.1103/PhysRevE.103.042312 (2021).
17. Ortúzar, J. & Willumsen, L. G. *Modelling Transport* 4th edn. (Wiley, 2011).
18. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, 2010).
19. Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks* (Cambridge University Press, 2008).
20. Barabási, A.-L. *Network Science* (Cambridge University Press, 2016).
21. Gallotti, R. & Barthélemy, M. Anatomy and efficiency of urban multimodal mobility. *Sci. Rep.* **4**, 6911. https://doi.org/10.1038/srep06911 (2014).
22. Aleta, A., Meloni, S. & Moreno, Y. A multilayer perspective for the analysis of urban transportation systems. *Sci. Rep.* **7**, 44359. https://doi.org/10.1038/srep44359 (2017).
23. Bassolas, A., Gallotti, R., Lamanna, F., Lenormand, M. & Ramasco, J. J. Scaling in the recovery of urban transportation systems from massive events. *Sci. Rep.* **10**, 2746. https://doi.org/10.1038/s41598-020-59576-1 (2020).
24. Ko, J., Kim, D. & Etezady, A. Determinants of bus rapid transit ridership: System-level analysis. *J. Urban Plan. Dev.* **145**, 04019004. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000506 (2019).
25. Wirasinghe, S. C. *et al.* Bus rapid transit—a review. *Int. J. Urban Sci.* **17**, 1–31. https://doi.org/10.1080/12265934.2013.777514 (2013).
26. Trubia, S., Severino, A., Curto, S., Arena, F. & Pau, G. On BRT spread around the world: Analysis of some particular cities. *Infrastructures* **5**, 1–13. https://doi.org/10.3390/infrastructures5100088 (2020).
27. Global BRT database. https://www.brtdata.org
28. Venter, C., Jennings, G., Hidalgo, D. & Valderrama Pineda, A. F. The equity impacts of bus rapid transit: A review of the evidence and implications for sustainable transport. *Int. J. Sustain. Transp.* **12**, 140–152. https://doi.org/10.1080/15568318.2017.1340528 (2018).
29. Barahimi, A. H., Eydi, A. & Aghaie, A. Multi-modal urban transit network design considering reliability: Multi-objective bi-level optimization. *Reliab. Eng. Syst. Saf.* **216**, 107922. https://doi.org/10.1016/j.ress.2021.107922 (2021).
30. Ostrowski, K. & Budzynski, M. Measures of functional reliability of two-lane highways. *Energies* **14**, 4577. https://doi.org/10.3390/en14154577 (2021).
31. Chen, Z. & Fan, W. Data analytics approach for travel time reliability pattern analysis and prediction. *J. Mod. Transp.* **27**, 250–265. https://doi.org/10.1007/s40534-019-00195-6 (2019).
32. Ji, K. & Ma, J. A modified network-wide road capacity reliability analysis model for improving transportation sustainability. *Algorithms* **14**, 7. https://doi.org/10.3390/a14010007 (2021).
33. Nie, Y. & Wu, X. Shortest path problem considering on-time arrival probability. *Transp. Res. Ser. B Methodol.* **43**, 597–613. https://doi.org/10.1016/j.trb.2009.01.008 (2009).
34. Abraham, S. & Sojan Lal, P. Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations. *Transp. Res. Part C Emerg. Technol.* **23**, 109–123. https://doi.org/10.1016/j.trc.2011.12.008 (2012).
35. Kim, J. & Mahmassani, H. S. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transp. Res. Procedia* **9**, 164–184. https://doi.org/10.1016/j.trpro.2015.07.010 (2015).
36. Zheng, L. *et al.* Spatial-temporal travel pattern mining using massive taxi trajectory data. *Physica A Stat. Mech. Appl.* **501**, 24–41. https://doi.org/10.1016/j.physa.2018.02.064 (2018).
37. Lin, Y., Yang, X., Zou, N. & Jia, L. Real-time bus arrival time prediction: Case study for Jinan, China. *J. Transp. Eng.* **139**, 1133–1140. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000589 (2013).
38. Fan, W. & Gurmu, Z. Dynamic travel time prediction models for buses using only GPS data. *Int. J. Transp. Sci. Technol.* **4**, 353–366. https://doi.org/10.1016/S2046-0430(16)30168-X (2015).
39. Liu, L. & Chen, R.-C. A novel passenger flow prediction model using deep learning methods. *Transp. Res. Part C Emerg. Technol.* **84**, 74–91. https://doi.org/10.1016/j.trc.2017.08.001 (2017).
40. Dabiri, S. & Heaslip, K. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transp. Res. Part C Emerg. Technol.* **86**, 360–371. https://doi.org/10.1016/j.trc.2017.11.021 (2018).
41. Gu, Y., Wang, Y. & Dong, S. Public traffic congestion estimation using an artificial neural network. *ISPRS Int. J. Geo-Inf.* **9**, 152. https://doi.org/10.3390/ijgi9030152 (2020).

42. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113. https://doi.org/10.1103/PhysRevE.69.026113 (2004).
43. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8577–8582. https://doi.org/10.1073/pnas.0601602103 (2006).
44. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008 (2008).
45. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002 (2010).
46. Schaub, M. T., Delvenne, J.-C., Rosvall, M. & Lambiotte, R. The many facets of community detection in complex networks. *Appl. Netw. Sci.* **2**, 4. https://doi.org/10.1007/s41109-017-0023-6 (2017).
47. Dey, A. K., Tian, Y. & Gel, Y. R. Community detection in complex networks: From statistical foundations to data science applications. *Wiley Interdiscip. Rev. Comput. Stat.* e1566. https://doi.org/10.1002/wics.1566 (2021).
48. Münnix, M. C. *et al.* Identifying states of a financial market. *Sci. Rep.* **2**, 644. https://doi.org/10.1038/srep00644 (2012).
49. Pharasi, H. K. *et al.* Identifying long-term precursors of financial market crashes using correlation patterns. *N. J. Phys.* **20**, 103041. https://doi.org/10.1088/1367-2630/aae7e0 (2018).
50. Barajas-Martínez, A. *et al.* Metabolic physiological networks: The impact of age. *Front. Physiol.* **11**, 587994. https://doi.org/10.3389/fphys.2020.587994 (2020).
51. Bergeaud, A., Potiron, Y. & Raimbault, J. Classifying patents based on their semantic content. *PLoS One* **12**, e0176310. https://doi.org/10.1371/journal.pone.0176310 (2017).
52. Rincón-López, J., Almanza-Arjona, Y. C., Riascos, A. P. & Rojas-Aguirre, Y. When cyclodextrins met data science: Unveiling their pharmaceutical applications through network science and text-mining. *Pharmaceutics* **13**, 1297. https://doi.org/10.3390/pharmaceutics13081297 (2021).
53. Zaltz Austwick, M., O'Brien, O., Strano, E. & Viana, M. The structure of spatial networks and communities in bicycle sharing systems. *PLoS One* **8**, e74685. https://doi.org/10.1371/journal.pone.0074685 (2013).
54. Zhou, X. Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago. *PLoS One* **10**, e0137922. https://doi.org/10.1371/journal.pone.0137922 (2015).
55. Hedayatifar, L., Morales, A. J. & Bar-Yam, Y. Geographical fragmentation of the global network of twitter communications. *Chaos* **30**, 073133. https://doi.org/10.1063/1.5143256 (2020).
56. CDMX Open Data. https://datos.cdmx.gob.mx/
57. Metrobús Open Data. https://metrobus.cdmx.gob.mx/portal-ciudadano/datos-abiertos
58. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86. https://doi.org/10.1214/aoms/1177729694 (1951).
59. Route Map Metrobús Mexico City. https://www.metrobus.cdmx.gob.mx/mapas-rutas
60. Kullback, S. *Information Theory and Statistics* (Wiley, 1959).

Author contributions

J.U.M.G. and A.P.R. designed and performed the research. A.P.R. wrote the manuscript, both authors concur with the submission and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-04037-6.

Correspondence and requests for materials should be addressed to A.P.R.

Reprints and permissions information is available at www.nature.com/reprints.

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