Transportation Research Part C 115 (2020) 102640

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

# A data science framework for planning the growth of bicycle infrastructures☆

Luis E. Olmos[a,*,1], Maria Sol Tadeo[b,1], Dimitris Vlachogiannis[b], Fahad Alhasoun[c], Xavier Espinet Alegre[d], Catalina Ochoa[d], Felipe Targa[d], Marta C. González[a,b,e]

[a] Department of City and Regional Planning, University of California, Berkeley, CA 94720, United States
[b] Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, United States
[c] Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[d] Transport Global Practice, The World Bank, Washington, D.C 20433, United States
[e] Energy Analysis & Environmental Impacts Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

ARTICLE INFO

*Keywords:*
Bike infrastructure planning
Cycling facilities
Infrastructure planning
Mobile phone data
GPS traces
Network science
Percolation theory

ABSTRACT

Cities around the world are turning to non-motorized transport alternatives to help solve congestion and pollution issues. This paradigm shift demands on new infrastructure that serves and boosts local cycling rates. This creates the need for novel data sources, tools, and methods that allow us to identify and prioritize locations where to intervene via properly planned cycling infrastructure. Here, we define potential demand as the total trips of the population that could be supported by bicycle paths. To that end, we use information from a phone-based travel demand and the trip distance distribution from bike apps. Next, we use percolation theory to prioritize paths with high potential demand that benefit overall connectivity if a bike path would be added. We use Bogotá as a case study to demonstrate our methods. The result is a data science framework that informs interventions and improvements to an urban cycling infrastructure.

## 1. Introduction

Nowadays, walking and cycling trips are essential in cities that are compromised to shift to a more sustainable mobility. In this context, providing a connected and well designed infrastructure plays a key role in promoting cycling (Hull and O'Holleran, 2014; Heinen et al., 2015; Buehler and Dill, 2015). While much guidance has been done on the physical design (ITDP-México, 2011; Altrutz et al., 2010), little work has been focused on how to decide, prioritize, and choose places for cycling infrastructure investments. Efforts in that direction (Lovelace et al., 2017; Larsen et al., 2013; Zhang et al., 2014) are based primarily on survey data (online or household travel surveys) to develop a systematic, quantitative, and scalable methodologies for bike planning purposes. The Propensity to Cycle Tool (PCT) (Lovelace et al., 2017), is an online, interactive planning support system that was initially developed to explore and map cycling potential across England, based on total origin–destination data, the authors model and visualize fastest bike path routes at national scale, leaving the analysis and growth of the entire bicycle infrastructure network as a follow-up study. Larsen et al. (2013) develop a GIS-based, grid-cell model for bike facility prioritization and location, taking into account existing bicycle trips and short car trips, and integrating accidents data.

☆ This article belongs to the Virtual Special Issue on "Traffic flow modeling".

* Corresponding author.
*E-mail address:* leolmoss@berkeley.edu (L.E. Olmos).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.trc.2020.102640
Received 1 May 2019; Received in revised form 5 January 2020; Accepted 28 March 2020

In recent years, transportation and city planning studies have taken advantage of data from personal-tracking devices and network science to better understand urban dynamics and inform planning. It has been well established that origin–destination matrices can be extracted from CDRs of mobile phones (Toole et al., 2015; Çolak et al., 2016; Florez et al., 2017) and GPS-equipped vehicles can act as sensors of traffic conditions (Geroliminis and Daganzo, 2008; Li et al., 2015), allowing to uncover patterns of individual and city level mobility (Alexander et al., 2015; González et al., 2008; Olmos et al., 2018; Hamedmoghadam et al., 2019). Using GPS data for Beijing, China, along with a quantitative framework based on percolation theory, two recent works (Li et al., 2015; Zeng et al., 2019) have studied the traffic organization at a network level. They have determined the critical speed index at which the global city traffic disintegrates into small and isolated clusters of local flows, which are connected by bottleneck links. Ganin et al. (2017), also inspired by percolation theory, characterize the resilience of the road network of 40 US cities. With their analysis they were able to understand how does a network recovers after a disruption. Yildirimoglu and Kim (2018) used network science to understand people's activities dynamics across Brisbane Metropolitan Area in Australia. They aggregated flows of different agents (i.e. private cars, buses and passengers) to make them comparable and used community detection to distinguish areas of activities along the city. Finally, in a study closely related to our approach, Bao et al. (2017) propose a data-driven approach to develop bike lane construction plans based on large-scale bike trajectory data. They introduce a flexible objective function to tune the benefit between coverage of the number of existing bike users and the length of their trajectories.

We propose a data analysis framework to prioritize bike paths[2] at urban scale by integrating trajectories from biking applications with previously validated origin–destination matrices (ODs) extracted from mobile phone data (Florez et al., 2017; Toole et al., 2015; Çolak et al., 2016). First, we define the potential bike demand as a distance-based measure of the trips that could be done by bicycle. Using a rejection–sampling algorithm (von Neumann, 1951; Casella and Robert, 2004), we filter the trips extracted from phone data by a distance distribution. The chosen distance distribution emulates the obtained from a biking app. Having established this, the premise of the work is to design a bike path network that better serves our target distribution of trips. In doing so, we would select trips of drivers that could get out of the car (and not pedestrians), while also demanding global network connectivity with the minimum addition of links. To that end, after map matching the potential demand with the existing infrastructure, we obtain a weighted network of potential bike flow. Thus, we use a percolation analysis, similar to the developed in Li et al. (2015), that allows us to identify the minimal connected bike-path network that (1) covers the entire city and (2) is composed of the links with the highest bike flows. By subtracting the existing bike paths from the best connected component we obtain the links that the percolation proposes for having new bike infrastructure.

Performing this study in Bogotá, Colombia, is of great interest since the city has been making large efforts to increase the bike mode share. They have built one of the most extensive bike–path networks in the world. Today, the bike paths cover more than 500 km and cycling has a six percent share of daily trips (approx. 635,000 bike trips per day (Secretaría Distrital de Movilidad, 2015)). In the existing bike paths (called *ciclorrutas*) many segments are still not connected to the main network, and the infrastructure maintenance, renovation and management is slow and inefficient. These factors diminish the bike's use and slow down the improvement of the city's mobility. Nowadays, Bogotá's city administration extracts the bike travel demand from bike traffic counts, mobility surveys, and opinion polls, but this information is limited and gets rapidly outdated. A toolkit that uses novel data sources and on-line resources will serve well the assessment to prioritize infrastructure interventions. While performed for this particular city, the framework can be replicated in any other cities using similar data sets. It requires geo–referenced census population, socioeconomic information, bicycle trajectories, empirical estimates, overall OD demand, bike facilities, and road network infrastructure. It produces potential bike trips and proposes bike paths in addition to the existing infrastructure.

The rest of the article is organized as follows. In Section 1, we describe datasets and inputs used in this study. We then explain the methods used for processing GPS data and identifying potential bike trips, in Section 2. In Section 3, we first present comparisons of these results to estimates from surveys and counts from piezoelectric sensors. Next, we use modularity maximization to detect communities in the resulting bike trip network. We obtain a map of Bogotá composed of four main groups which, surprisingly, have a correspondence with the spatial distribution of the socioeconomic groups. Section 4 introduces a percolation-like framework that allows us to evaluate the global connectivity of the network. Applying the connectivity criterion, we propose new cycling infrastructure and perform spatial analysis to quantify how these interventions will distribute over the detected communities. Finally, we implement an algorithm which collects street-level imagery of road sections to intervene. With this tool, city decision-makers can remotely evaluate the viability of the proposed interventions visually.

### 1.1. Input and data description

To estimate the potential demand for bike mode in the city of Bogotá, this study leverages multiple sources of mobility data. The datasets and tools used to develop and demonstrate the capabilities of our methods are described as follows:

1. *Origin–destination (ODs) matrices:* From a previous work (Florez et al., 2017), we use OD matrices that model overall daily travel demand for the Bogotá metropolitan area. They cover a population over 9.5 million individuals and 659 towers distributed across an area of 477 km$^2$. They were extracted by coupling a large mobile phone data set and the population at the census block level. This trip's network has 612 nodes (towers) and 103,487 weighted links, representing more than 9.8 millions of daily trips. In the same work, the phone based OD matrices were validated with the 2011 mobility survey of Bogotá (Secretaría Distrital de

[2] We will refer to any type of cycleway as bike path, infrastructure or facility.

Movilidad, 2011). This dataset, which we refer as CDRs all trips, is the starting point to identify potential bike trips based on trip distance distribution.

2. *Smartphone-based bicycle GPS data:* The second dataset uses anonymized data collected through Biko, a smartphone application that encourages the use of the bicycle as a transportation mode (Biko, 2019). Riders win Biko points for each km they ride and can change them for goods in different local shops. This app was specially popular among cyclists in Bogotá during 2017. The dataset provided by Biko are GPS trajectories from April, 2017 to March, 2018, in which the position of a user is registered every few seconds. Each data point is identified by a user and an activity ID; we have 4,526 activity IDs corresponding to 2,507 users.
3. *Census population dataset:* We obtain a census population dataset at the city block level with their socio–economic status (SES), there are about 45,000 blocks classified by income in the city. This dataset is used for generating representative urban travel demand estimates, as well as, for analyzing the impact of possible new infrastructure and intervention according to the socio–-economic features of various demographics. The socio–economic data in Bogotá has a high spatial resolution, every real state property is classified in a SES, from 1 to 6, representing resident's income level from low to high, being 6 the highest. Other factors, such as quality of urban spaces and access to goods and services complement this classification, for details see Uribe-Mallarino (2008). In terms of socio–economic groups, 51.3% of the population is in the low–income bracket (SES 1 and 2), 44.4% in the medium–income bracket (SES 3 and 4) and, 4.4% in the higher–income bracket (SES 4 and 5) Secretaría Distrital de Planeación (2017).
4. *Mobility surveys:* For comparative purposes only, we use both the 2011 and the 2015 mobility surveys of Bogotá (Secretaría Distrital de Movilidad, 2011; Secretaría Distrital de Movilidad, 2015) and contrast their results to the output of our data analysis framework. The surveys provide detailed data of socio-demographic characteristics, modes of transport, age, gender, and types of daily activity. The samples comprises 15,500 (in 2011) and 28,025 (in 2015) households distributed across Bogotá and their surrounding municipalities. The metropolitan region is divided into 112 urban zonal planning units (UPZ), which are territorial units for planning urban development at the local level and for defining land–use and urban functions. Using the scaled–up bicycle trips from the surveys data and our potential bike trips, a comparison for trip productions and attractions was conducted at UPZ level.
5. *Data from automated bicycle counters:* As an effort to characterize non–motorized transportation (NMT), the Secretary of Mobility of Bogotá, or Secretaria Distrital de Movilidad de Bogotá (SDM), installed 12 bicycle counters (piezo–electric strips) along bike paths that run parallel to three major corridors of the city. These sensors collected data in each direction over 5 min periods during 588 days between 2016 and 2017. We contrast their 2017 data to the output of our potential demand.
6. *Road networks, and on-line map platforms:* In this study, we leverage on multiple capabilities of the OpenStreetMaps (OSM) project, an open source community dedicated to create and distribute free geographic data collected and served by thousands of volunteers across the world. For Bogotá, OSM has a very detailed road vehicular network along with a reasonable up–to–date data of the biking facilities. On it, one can distinguish different kinds of routes: physically separated cycle tracks (bike paths), on-street painted lanes, bicycle–only or multi-use roads. A core tool in our framework is their Open Source Routing Machine (OSRM) project(see Luxen and Vetter, 2011), an open–source efficient routing software, and also configurable and extensible. In particular, OSRM supports *profiles* that represent routing behavior and constraints for different transport modes such as car, bike and foot. In addition, it is also possible to create custom variations of these *profiles* like fastest/safest/greenest cycling profile. We use the back–end version of the OSRM for: (1) map–matching the Biko data to the street network, and (2) computing cycling routes for every OD pairs of the potential demand, using the underlying OSM road network infrastructure.
7. *Existing and projected bike–paths network:* For the case of Bogotá, the OSM geodata of the current state of the bike paths network is fairly updated, we thus will use them as the existing bike infrastructure. Additionally, the SDM provided us the shapefiles of the projected infrastructure updated to 2019.

## 2. Methods

### 2.1. Processing GPS bicycle trajectories

The Biko app identifies a single trip by an activity ID. However, since the beginning and end of an activity must be informed by the user, some users turn the app on, and forget to stop it when their trips ends. As a result, there are very long duration trips (e.g. 10 h) where the user is not riding all the time but making stops of different duration at different points.

To overcome this issue we parse the trajectory using the method proposed by Hariharan and Toyama (2004) adapted for our purpose, the code is detailed in Ref. Jiang et al. (2013). This method analyzes raw GPS data and classifies the records as whether the person is moving or staying in a place. They consider a stay as "spending some time in some place" and define it by two scale parameters, the roaming distance and stay duration. The roaming distance represents the maximum distance that an object can stray from a point location to count as a stay and a stay duration is the minimum duration an object must stay within the roaming distance of a point to qualify as staying at that location (Hariharan and Toyama, 2004). We use the method with a roaming distance of 0.1 km and a stay duration of 5 min and were able to classify each user record as whether the person is on the move or on a stay. Applying this method, divides each activity ID or individual trip of the Biko app into sub-trips in which the person is actually moving. The number of trips increased from 3,916 to 4,526 after the parsing.

The GPS data needs to be furthered treated since there are some points in the trajectories that present high speeds not compatible with bicycle trips, also known as hyperspacing. To that end, the speed between two consecutive points was calculated and any point with a speed higher than 30 km/h was removed. Fig. 1(b–c) show the final histograms for speed and travel time. Consistent with

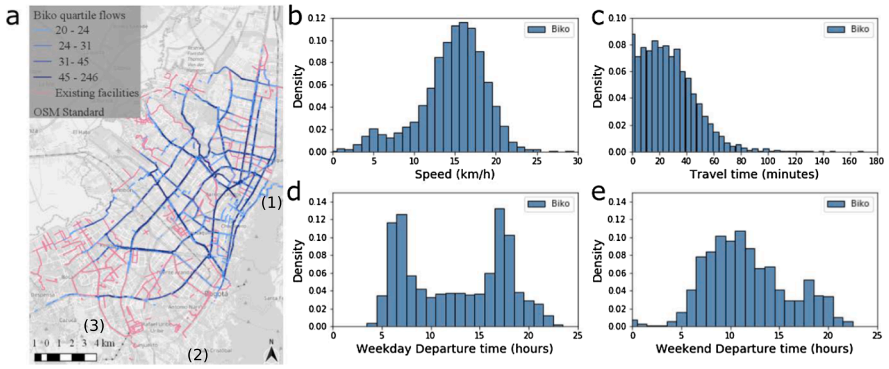


**Fig. 1. Main features of the post-processed Biko data.** (a) Biko flow network after a map matching procedure. Biko users (blue lines) mainly utilize the existing bike-paths (pink lines under the blue ones). Zones marked with numbers correspond to highest-altitude neighborhoods in Bogotá. With the exception of these three zones, the city is mostly flat. Map created using QGIS 3.4 software (QGIS Development Team, 2019). (b) Distribution of speeds in km/h. (c) Distribution of travel time in minutes. (d–e) Departure times for all trips during week and weekend days, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cycling reports, we found an average speed of $v = 15(4)$km/h for bikes on urban environments. To infer trip purposes, the beginning of each trip was classified as week or weekend and two separate histograms were generated with trip departure by time of the day (see Fig. 1(d–e)). Interestingly, the weekday histogram shows that the majority of the trips are conform to commuters behavior.

Once the bicycle GPS data is processed, we map matched it in the street network. The tool employed is the OSRM that takes as input GPS traces and the delta time between consecutive points and map matches it to the OSM street network. This step translates all the data into the same reference system (the OSM road network). In the raw data, if two users ride along the same street this cannot be noticed since most probably, they will not have the same coordinates for each point. The map match will associate each GPS point to a node from the OSM network and the result will be a sequence of nodes indicating the trajectory of each user. This can be also mapped into edges from the network. Edges of all users are added to create a weighted network in which the weight of the edge is the volume of users that passed by that link.

Fig. 1(a) shows the resulting flow network after the map matching procedure. Two features need to be mentioned. First, major corridors with bike paths attract the majority of Biko routes. Second, we do not see Biko users in most of the south bike paths of the city, a zone that corresponds mainly to low–income neighborhoods. This can be interpreted as socioeconomic bias in individuals using Biko. By coupling the home location of Biko users (origins of their earliest trip) with the census population dataset, we observe that 18.3% of users are low–income (SES 1 and 2), 54.7% middle–income (SES 3 and 4) and, 24.3% higher–income. There is then a clear mismatch with the empirical socio–economic distribution in Bogotá, where only the 4.4% belongs to the higher–income class and the 51.3% to the lower one, see Section 1.1. Since the zones not covered by the Biko data correspond to low–income areas, we conjecture that the low access to smartphones and mobile internet services hampers the adoption of the Biko app.

### 2.2. Identifying potential bike trips

A growing body of evidence suggests that high–quality infrastructure encourage more people to cycle. The case of Bogotá supports the evidence, with a clear plan of improving and extending the bike paths network, their number of trips has increased from $\approx$ 421, 000 in 2011 to $\approx$ 635, 000 in 2015. To meet the needs of current cyclists and attract future ones, it is important to determine and intervene the best location for new bike facilities.

Using the OD matrices derived from mobile phone data, we estimate the potential cycling trips as a function of the straight-line (Euclidean) travel distances between origin and destinations, $d$. We do not use routed distance here, for simplicity and making it agnostic to the existing infrastructure. For cities with landscape imposed detours, this step needs to be reviewed.

We first explore potential bicycle trips as those in which $d$ is shorter than a certain cut-off $d_c$. We then assign these trips to the road network in OSM using the routing service from OSRM with the *bicycle* profile, which considers the shortest time path. Figs. 2(d) and 2(e) show the resulting daily flow network for the case scenario of $d_c = 4$ km and $d_c = 10$ km. In the first case, Bogotá's workplace centers are not attracting bike trips (empty zones), while in the second case some main corridors appear as broken up strips (compare to the Fig. 1(a)). In both cases, the distribution of trips looks concentrated in the south and north-west of Bogotá, a feature not observed in Biko results. When comparing the Euclidean distance distribution in the two datasets (Fig. 2(b)), we can observe that a large fraction of the OD trips are so short that they are easily accessible by walking. In order to plan for medium distance trips, we sample the trips from the CDRs such that $d$ distributes similarly to those observed in the Biko app.

To doing so, we implement a rejection-sampling algorithm (von Neumann, 1951; Casella and Robert, 2004) that works as follows. First, we find a reasonable distribution fitting to each dataset. For Biko trips, Fig. 2(a) shows that $d$ follows a gamma distribution, $f(d) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} d^{\alpha-1} e^{-\frac{d}{\beta}}$, with mean 6.25 km and s.d. 4.03 km (i.e. $\alpha = 2.217$ and $\beta = 2.843$). In the case of the OD trips, Fig. 2(b), a reasonable fitting was done to a lognormal distribution, $g(d) = \frac{1}{\sqrt{2\pi}\sigma d} e^{-(\ln(d)-\mu)^2/2\sigma^2}$, with mean 6.10 km and s.d. 10.22 km (i.e.

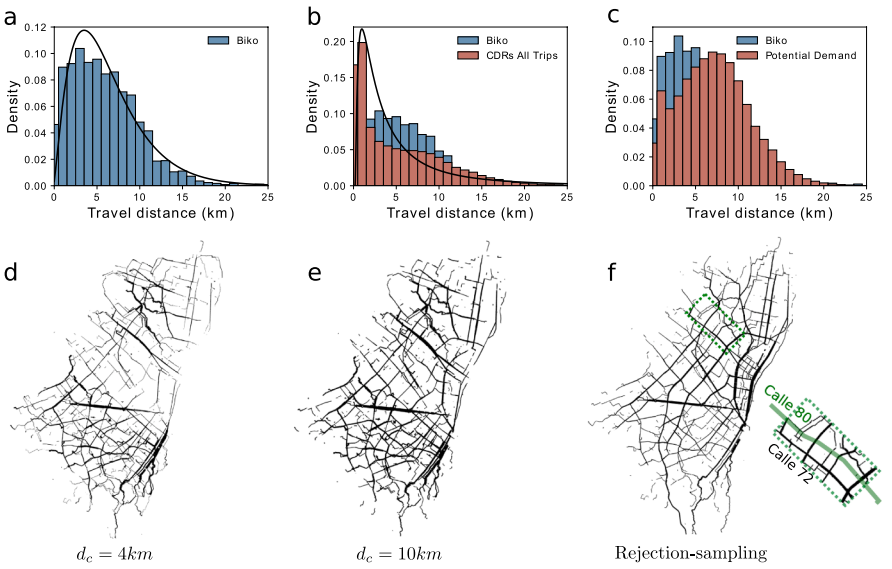


**Fig. 2. Distance criterion to define the bike potential demand.** (a–b) Distributions of Euclidean distances, $d$, for Biko and CDR-all trips. These were fitted by a gamma and a lognormal distribution, respectively. (c) Resulting distribution of $d$ after applying the rejection–sampling algorithm (Casella and Robert, 2004; von Neumann, 1951). (d–e) Daily flow networks of potential trips obtained with the cut-offs criteria, $d_c = 4$ km and $d_c = 10$ km, respectively. (f) Flow networks of the potential demand obtained with the rejection–sampling algorithm. (Lower inset) Zoom of the green box of dashed lines. Here, we show how the shortest–path assignment routes a large number of trips on Calle 72, street without bike facilities, leaving unused the bike paths on Calle 80 (green line). Biko users do the opposite, see Fig. 1(a). Maps created using Gephi software (Bastian et al., 2009). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$\mu = 1.132$ and $\sigma = 1.163$). Next, we find a scaling factor $k$, so that $kg(d)$ envelops $f(d)$ entirely. It will just be the maximum ratio between the two distributions, $k = \max(f(d)/g(d))$, for all $d$. Finally, iterating over all OD pairs, with probability $\frac{f(d_i)}{kg(d_i)}$ we keep the $i$-th trip, otherwise we reject it. The hidden idea of this simple method is to assure that the probability of drawing the $i$-th trip is exactly $f(d_i)$. The resulting potential demand consists then of ≈4.1 millions of bikeable trips, corresponding to the 42% of the total daily trips. Fig. 2(c) presents the distribution of $d$ for the accepted trips after applying the rejection method, to which we will refer as the potential demand. The observed mismatch with the Biko's distribution is a result of the simple fitting to a log–normal distribution. A more accurate distribution fit would improve this result; however, for seek of generality, we prefer to present the results with the simple fit.

Interestingly, after assigning the potential demand, the daily flow network is much better distributed in space, where the major corridors of the city serve the bulk of the bike trips, displaying a hierarchical structure and thus, much more satisfactory as potential bike trips from the empirical ODs. When comparing to the Biko flow network (Fig. 1), there is a clear consistency except for a small detail in the north-west region of Bogotá. Biko users prefer to use the Calle 80 corridor (see Fig. 2f) and inset) with bike paths, rather than Calle 72 (the next parallel corridor to the south), which is currently lacking bike facilities but shortens the routes according to OSRM software. Thus, the existence of bike infrastructure creates a deviation between our shortest-path assignment and the actual path taken by bikers. To reduce the impact of this issue, in the routing, we set the bike path speeds to 17 km/h in the OSRM bicycle profile (15 km/h is the default value) such that bike paths have a slight advantage to be selected over the vehicular road segments. From here, all the results have been obtained using this configuration.

Due to a large number of short trips, a simple cut-off criterion does not estimate properly potential demand. This is satisfactorily achieved when we use the Biko distance distribution to identify potential bike trips. This supports the premise that a proper design of bike path networks should prioritize medium-distance trips. Otherwise, instead of getting drivers out of their cars, bike paths will mainly attract pedestrians, reducing the impact of cycling in the city's mobility.

Note that our potential demand does not consider the elevation differences, a crucial factor in cycling. The reason is that the urban area of Bogotá is mostly flat, except for three residential areas located at the foothills of the north-eastern and southern hills of Bogotá, see marked zones in Fig. 1(a). However, if needed, integrating this factor in our methodology is relatively easy. Having the elevation profile of the city as external data, OSRM allows to load it in memory as node/ways features and then, one can penalize or ban them using a maximal slope cut-off, for details see Luxen and Vetter (2011).

It is important to emphasize that our potential demand is in contrast to the latent demand, which is based on behavioral models that would require demographic information from individual travelers, not available in the phone data records. Our potential demand acts as an upper bound of possible bicycle trips. We compare their magnitude with the estimates of bicycle trips based on surveys and path counters.

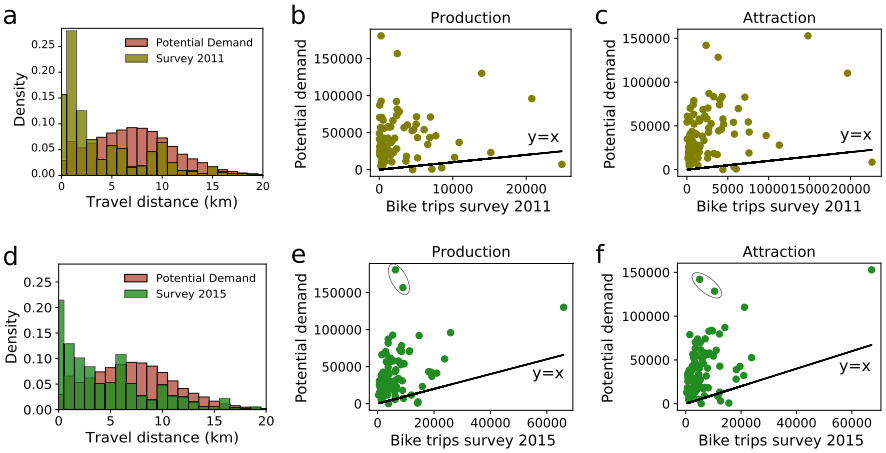


**Fig. 3. Comparison between bike potential demand and the scaled-up ODs based in 2011/2015 mobility surveys.** (a) and (d) Distribution of Euclidean distances, *d*, for bicycle trips. In the 2011–2015 period, the number of bike trips have been duplicated, and the average value of *d* have increased from 4.4 to 5.1 km. (b–c) and (e–f) Comparison of trip production and attraction at the UPZ level. As expected, there is not a good correlation, but there is a clear improvement from 2011 to 2015. These comparisons allow to pinpoint zones where the actual adoption is much lower than their potential demand and explore causes for this. That is the case of the marked outliers, which correspond to two main workplace centers in the city.

## 3. Results

### 3.1. Trip production and attraction

While we do not expect much a match between our potential demand and the empirical bike demand, our potential demand is not a latent demand, we look for a quantitative reference of the actual bike demand estimated by surveys and the potential scenario where 42% of all trips in the city are done by bike. Such comparison provides insights in the relative adoption of bicycle trips by origin and destination as of today vs. their actual potential based on the total travel demand.

The two last mobility surveys for Bogotá (Secretaría Distrital de Movilidad, 2011; Secretaría Distrital de Movilidad, 2015) show a large number of short bike trips. However, when comparing each other, those trips tend to reduce in that period. Instead, there is an increase of medium distance trips (5–10 km), and thus a shift on the average value of *d*, it went from 4.4 km in 2011 to 5.1 km in 2015 (see Fig. 3(a) and Fig. 3(d)). We then compare spatial distribution of origins and destinations for bike trips at a UPZ level. From both the potential and surveys data, we extract the daily trips originated in a zone (production) and the daily trips destinated for a zone (attraction). As expected, we find a small correlation between the observed in surveys and the potential demand, Fig. 3(b-c) and Fig. 3(e–f), with a small improvement in 2015. These comparisons are useful to highlight zones with large potential demand and low adoption or relative good adoption of bikes. For instance, the marked outliers in Fig. 3(e–f) correspond to two main workplace centers in Bogotá, namely: Chicó Lago and La Sabana. They are top producers and attractors of potential bike trips, but they have low adoption rates.

### 3.2. Piezoelectric sensor data vs Biko and potential demand.

A great limitation in the study and planning for bicycle demand is the absence of large data sources on empirical observations. In the case of Bogotá, the SDM has installed 12 permanent piezoelectric sensors along three bike corridors, see Fig. 4(a). We obtain the collected data by these sensors for 288 days in 2017, we aggregate them in weekdays (from 4 am to 10 pm), and then compute the average counted bicycles on a typical day for each location. We compare these daily bicycle traffic to both what observed in the Biko app and the volumes obtained from our potential demand. Before comparing, we should note that Biko reveals a local idiosyncracy. When a trip can be routed both by a bike path and a vehicular street, Bogotan cyclists sometimes may use them interchangeably (see Fig. 4(b)). Among the several reasons for this behavior, we should mention that local cyclists do that sometimes for avoiding assaults for stealing of their bicycles (Secretaría Distrital de Seguridad, 2018). Thus, for the comparison, Biko daily flows correspond to the sum of the flow in bike paths and the parallel vehicular street. For the potential demand, the $v = 17$ km/h configuration makes OSRM favors bike paths over vehicular streets, while still using the later (Fig. 4(c)).

Overall, we observe a reasonable correlation (Fig. 4(d–e)). In the case of Biko data, the dispersion of the points can be due to the socio–demographic biases that this application has. In the case of our potential demand, we find a remarkable correlation with the empirical observations. However, there are some places where the differences are significant. This is a consequence of our distance-based assumptions, recall we are discarding most of the short–distance trips and, in the empirical observation, those trips are a large share of the total. Surprisingly, potential flows and sensor counts are in the same magnitude, despite the significant difference in the number of trips between potential and actual demands. This is a consequence of the various route choices used in the empirical vs. the

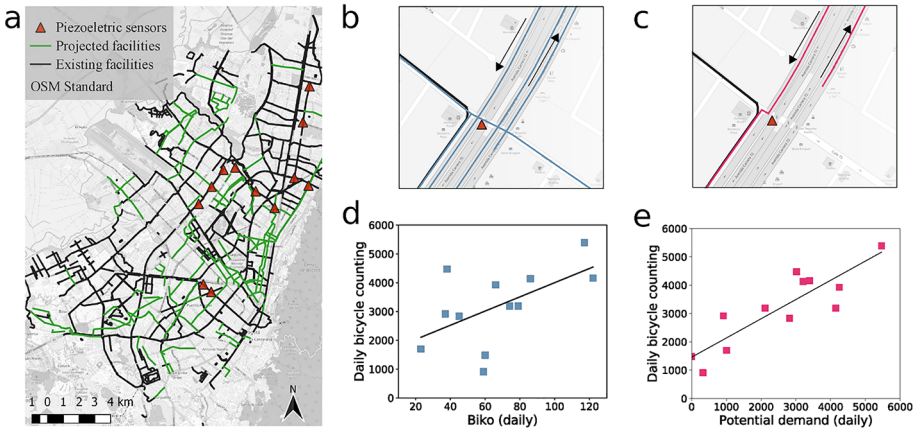


**Fig. 4. Piezolectric records vs Biko and potential demand.** (a) Distribution of the piezoeletric sensors (red triangles). Black (green) lines correspond to existing (projected) bike paths. (b) Biko reveals a very local behavior, in corridors with both bike paths (black line) and vehicular streets (dark grey strips), Bogotan cyclists use them interchangeably. Thus, in that case, Biko values correspond to the sum of bike paths and street flows. Arrows indicate the direction of the vehicular flow. (c) In the case of potential demand, the configuration of v = 17 km/h eliminates most of these situations. (d) Reasonable correlation between piezoelectric and Biko data. (e) The correlation is remarkably better in the potential demand case. Maps were created using QGIS 3.4. software by (QGIS Development Team, 2019). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

planned scenario. In the empirical demand, these are fairly constrained to the existing bike routes (as shown in Fig. 2(f)), which show high use captured by piezoelectric while in the potential demand, the use of alternative routes is possible.

### 3.3. Spatial clustering of the bike potential demand

In many complex networks, nodes cluster and form local sub-networks where their nodes are highly interconnected among themselves, called network communities. Detecting communities in mobility networks is an important problem, since it allows networks to be divided into coherent groups with high trip exchange, detecting those in an automatic manner, going beyond the static administrative subdivisions (Blondel et al., 2010). The usual measurement to quantify the quality of a network division is the notion of *modularity*, introduced by Newman (Newman, 2006). The modularity of a set of sub-network divisions or clusters reflects to what extent the density of links within groups is significantly higher than what may be expected in a random division with the same cluster sizes. Here, we examine the existence of communities in the potential demand by means of the so-called Louvain algorithm (Blondel et al., 2008), an algorithm that efficiently maximizes the resulting modularity of the resulting network communities.

Fig. 5 illustrates the network communities obtained based on the number of trips between nodes (towers) in the potential demand. Four spatially balanced communities appear naturally, they are composed of between 75 and 255 towers. As depicted, communities are organized around the Bogotá's CBD, the point with overlap of the four communities, revealing the monocentric structure of the trips. When comparing to the spatial distribution of socio–economic strata (SES), surprisingly, there is good correspondence between the detected communities and the residential socio-economic segregation in Bogotá. The lower-income groups cluster themselves in the southernmost region of the city while the higher-income households locate in the northern area.

## 4. Intervention strategies via a percolation approach

Thus far, our methodology allows determining the potential demand at a local level. Accordingly, we could simply prioritize the road segments to intervene according to their flow, $w$. However, we would be neglecting the global connectivity, a fundamental attribute of efficient transportation networks. From a transportation planning perspective, this condition allows touring the entire city without leaving the network of bike paths, reducing the risk of accidents and preventing the appearance of borders that boost even more the segregation. But more importantly, global connectivity ensures complete access from any pair of points. To accomplish this condition, we introduce a quantitative framework based on percolation theory, which is very similar to the one developed by Li et al. (2015).

To give a glimpse of the core concept behind percolation theory, consider the following example. Assume a 2D square lattice where each link $i$ is assigned a random weight $p_i \in (0, 1]$. The lattice is initially empty and as the threshold $p$ is changing form 1 to 0, you add links having weights larger than $p$ ($p_i > p$), one by one. At a certain point in this process, after adding a link with weight $p_i$, a path of connected links from the top to the bottom of the lattice emerges, the so-called percolating cluster. Averaging over many configurations, the percolation theory questions what is the likelihood to find a percolating cluster when all links with $p_i \geqslant p$ are added. It is clear that in the case of a very high $p$ (too few edges are added), it is unlikely to find a cluster of links that connects the top and the bottom of the lattice. Conversely, in the case of a very low $p$, such a cluster is likely to exist. Interestingly, the transition from one case to the other is not smooth but very drastic. Above a certain critical threshold, $p_c$, a percolating cluster is very unlikely to

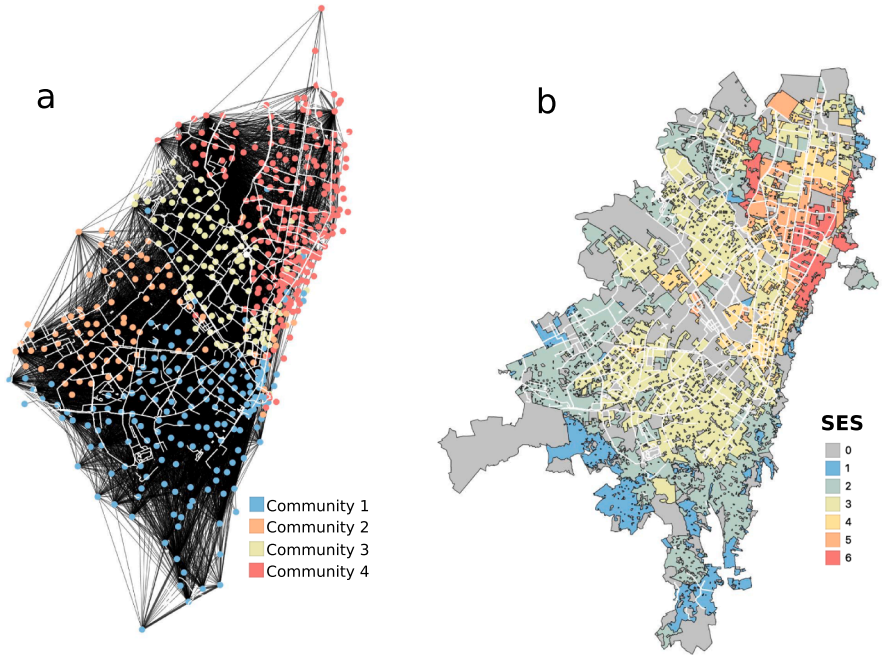


**Fig. 5. Spatial clusters of bicycles potential demand.** (a) Spatial clusters of network communities areas extracted based on the number of trips between nodes (antennas) in the potential demand. Four main communities can be distinguished by color. (b) Spatial distribution of socio–economic strata (SES), from 1 to 6 going from lowest (1, south) to highest income (6, north-east). Surprisingly, there is a clear correspondence between our communities and those given by the residential socio–economic segregation. Maps were created using QGIS 3.4 software by (QGIS Development Team, 2019).

exist, whereas below the critical threshold it is very likely. At criticality, the network is said to percolate, i.e. the likelihood of a connection between top and bottom side jumps from almost 0 to almost 1. This concept can be transferred to our problem. If we add a bike-path link according to if its flow is higher/lesser than a threshold q, we can try finding the minimum $q_c$ for which a connected network of links spans across the city. Noticing that in our case the weights are demand driven and not randomly located.

Starting from the flow network of the potential demand, we first set artificially the flows of the existing links to a $w_i = 5000$, a value higher than the maximum predicted flow. The aim of this step is to ensure that the existing network is present throughout the process. Then, for a given threshold q, we classify a link $e_i$ as "susceptible to intervene" when $w_i > q$ and not for $w_i < q$. In this way, given a q value, the bike path network to build can be defined from the potential flow network. For high values of $q$, the network will be fragmented in small clusters. In the opposite case, for very low values of $q$ the network will resemble the whole street network of the city. As said before, the goal is to find the maximal value of $q_c$ at which the connected network spans across the city, i.e. there is a giant connnected component. To find $q_c$, we can vary the value of q and track the cluster aggregation process, see Fig. 6(a–e) (from (e) in counterclockwise). For example, for $q = 3200$, as shown in Fig. 6(e), there are five middle-size clusters, which evidence the existing network is not well–connected. As the value of q decreases, the 2nd-largest cluster (dark green) grows faster than the 1st one (pink), in such a way that around $q = 2750\frac{trips}{day}$ there is a big jump in its size (green line in Fig. 6(f)). Continuing with the process, at $q = 1800\frac{trips}{day}$ the size of the second-largest cluster becomes maximal and merges with the 1st-largest cluster. This marks the point at which the so far proposed links connect most of the existing network. However, the network is still not globally connected, we need to go to $q_c = 520\frac{trips}{day}$ to see how the southern branch connects and the giant component emerges (see Fig. 6(a)).

This can be better understood in Fig. 6(f). As q increases, the size of the giant component (G) decreases and the network start to fragment. The critical threshold ($q_c$) separates the fragmented phase from the connected phase of the traffic network. Thus, the network in Fig. 6(a) corresponds to the minimal network with a reasonable use that spans across the city.

### 4.1. Proposed bike paths

From the percolation analysis, the giant component represents the target bike–paths network to have in the city. For this information to be useful for practical planning projects, we need to distinguish three types of links: (1) existing links, (2) links already projected by the SDM and (3) links without bike–path but that already have one across the street. Subtracting these links from the giant component, we will obtain the proposed links for intervention. While the existing links are already identified, for the other type of links, we need to implement a set of geo–processing steps. The routine can be summarized as follows: First, we calculate a buffer of 50 meters around each node of the giant component network. Then, we do a spatial join and identify whether or not a node was further than 50 meters from a bike facility. Thus, a link is classified as a proposed link if both nodes that conform it are further than

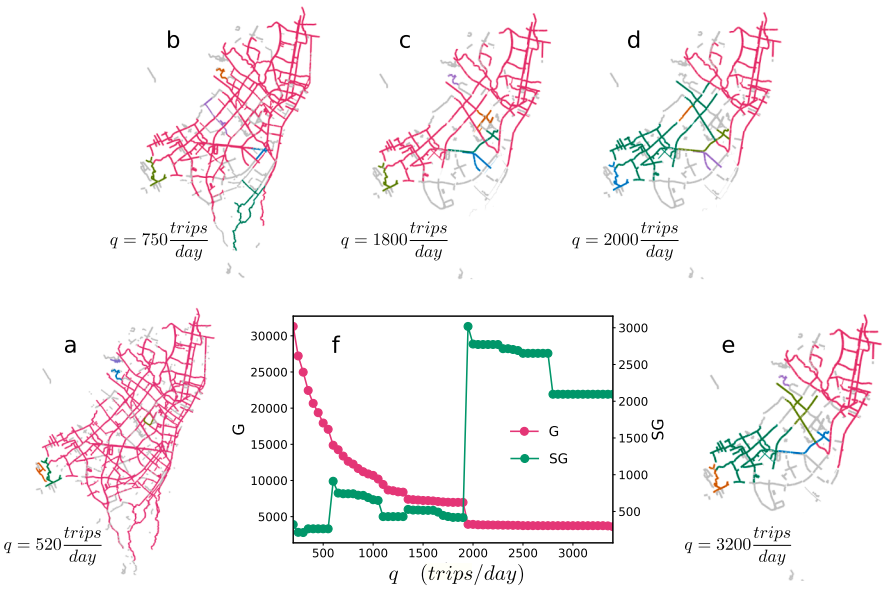


**Fig. 6. Percolation of the potential flow network for Bogotá.** (a–e) Proposed bike path network for five different q values corresponding to different states of connectivity. For clarity, only the largest clusters are plotted, e.g. pink (largest cluster), green (second-largest cluster), blue (third-largest cluster) and so on. (f) Size of the largest cluster (G) and the second-largest cluster (SG) of bike path networks as a function of $q$. Critical value, $q_c$, is determined as the q value when G connects globally for the first time, i.e. around $q = 520\frac{trips}{day}$. Thus, the network in (a) corresponds to the optimal bike path network indicated by the giant component using link flows of potential demand as threshold. Maps created using Gephi software (Bastian et al., 2009). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

50 meters from a bike facility. The output of this procedure is a shapefile of the proposed bike network to built in the city, an useful element for transportation and city planners. In Fig. 7(a) and Fig. 7(b), we can differentiate the existing, projected and proposed layers.

### 4.2. Characteristics per intervention Zone

In this section, we analyze the level of service of bike paths of two intervention scenarios for the communities depicted in

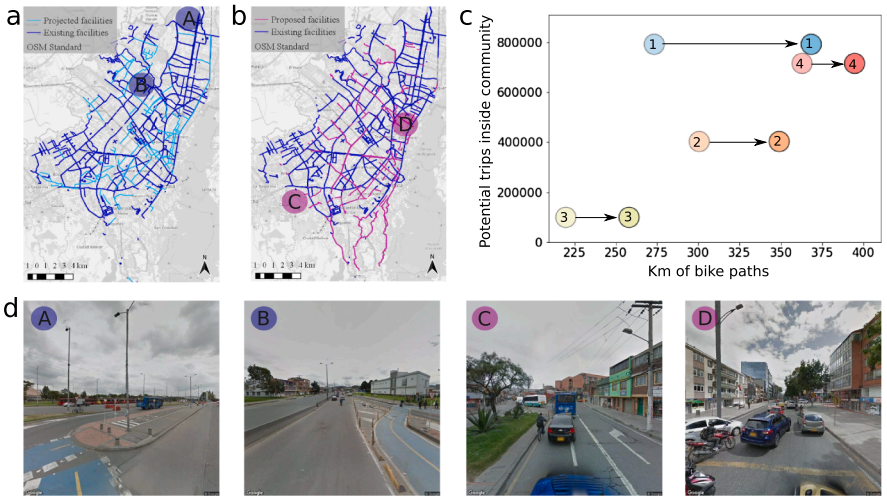


**Fig. 7. Proposed new facilities.** (a) Existing and projected bike paths updated to 2019. (b) Proposed facilities are the remaining network after subtracting existing and projected links from the connected network obtained in the percolation analysis, pink lines here. Map created using QGIS 3.4 software by (QGIS Development Team, 2019). (c) Distribution of infrastructure for the different communities. Potential trips inside a community against (1) the kilometers of existing and projected infrastructure (light-color circles) and (2) the same but for the scenario in which the proposed bike–paths have already been built (full-color circles). (d) Sample images scrapped from Google Street View for proposed facilities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5(a). To doing so, we first translate the communities to areas using the Voronoi diagram algorithm provided by QGIS 3.4 software (QGIS Development Team, 2019). Voronoi diagram is commonly used to define the catchment polygon for each point from a set of points in a two dimension space. All the points from the space inside a polygon are closer to the point that was used to generate it than to any other point in the set. In this process we observe that not all points from a community are adjacent between them. This happens because communities are defined as those points for which the connections to the points inside the community are stronger than to the points outside the community. So, there is no rule that requires points to be adjacent. On the other hand, most of the points in Fig. 5(a) seem to be adjacent. The reason is that networks related to human behaviors tend to be spatially clustered as well. The city center of Bogotá is were the communities are mixed which makes sense because of the monocentric structure of the city that attracts all the citizens to its centre for several reasons.

Once the area of each community is well defined, we estimate the number of potential trips inside each community, i.e those trips with origin and destination in the same community. Similarly, we compute the length of bike–paths inside communities for two scenarios (1) the existing and projected network and (2) the target escenario where the giant component is completely built. In Fig. 7(c), we compare these scenarios. As expected, the southern region of Bogotá receives most of the proposed facilities, evidencing that zone has been overlooked in the past intervention plans. Additionally, the target bike–paths network serves the communities accordingly to the volume of potential trips inside them.

### *4.3. Outlook: Intervention assessment via street imagery*

In an effort to make a feasibility analysis of the proposed infrastructure faster for the eye of experts, we scrape images from Google Street View. We create an image database from all the links and use it as needed. Our outlook, is to automatize this process in an interactive software application for better decisions and assessment, that can open avenues for training artificial intelligence in computer vision for a better classification of the paths and integration of other data attributes such as accidents and crime.

For this method, we need to indicate the Google maps API where we want to take the images. Since each link from the proposed network has two georeferenced ends each, those are our selected coordinates to extract images. Other variants like links' midpoints are also an option. Besides, other parameters like vertical angle from the camera can be set to get several images for various purposes. In Fig. 7(d), marked with labels (A) and (B), we present pictures of four pair of coordinates from the existing facilities and proposed. The remaining pictures in the same figure correspond to locations where we propose links for interventions. This is a small sample of a powerful tool that can be used to speed up planning processes in the information age.

## 5. Conclusions

In a data-rich reality, this work aims to open avenues for methods that support the planning of bicycle paths at urban scale.

We propose a data science framework that identifies streets that are candidates for adding new bike infrastructure, taking into account potential bike flow and preserving global connectivity. To that end, we identify potential bicycle trips by coupling mobile phone data and GPS traces from a smartphone application for bikers. Different from the latent demand concept derived from survey data, our potential demand are the trips that could be done by bicycle, we filter the trips extracted from phone data by a distance distribution.

Using percolation theory, we identified links from the street network that form a giant component and thus a well-connected bicycle infrastructure using the potential demand flows per link as a threshold. We obtained satisfactory results, first when comparing the estimated potential demand with survey and sensors data of bicycle demand. Second, we obtained reasonable bike paths extensions, when comparing the proposed bike infrastructure of our method with the existing and planned infrastructure of the secretary of mobility from Bogotá (SDM), our case study.

As mentioned in the outlook section, there are several avenues to enrich this framework. For example, by adding more information like unsafe areas in terms of crime, accidents and poor air quality. This would allow to prioritize interventions that take into account security measures that improve livability and overall well-being.

Finally, this is an open–source toolkit that has benefited from other open–source applications. The ultimate goal is to connect the data science community in projects that help cities with their bike facilities planning processes. It is a fully portable framework based on data from communication technologies and on-line maps. Far from being finalized, this may represent the beginning of a participative online community of bicycle advocates that could grow these ideas further helping in the ultimate goal of making our streets safer for those who ride.

## Credit authorship contribution statement

**Luis E. Olmos:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Maria Sol Tadeo:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Dimitris Vlachogiannis:** Software. **Fahad Alhasoun:** Software. **Xavier Espinet:** Data curation, Funding acquisition. **Catalina Ochoa:** Data curation, Funding acquisition. **Felipe Targa:** Data curation, Funding acquisition. **Marta C. González:** Conceptualization, Methodology, Supervision, Writing - review & editing, Investigation.

## Acknowledgements

The authors thank Prof. Daniel Rodriguez for having us initiated in the opportunities of this fascinating area of work and for his helpful comments and advise. This work was initiated as class project in CYPLAN 257, we are thankful to Robert O'Connor, Kanaad Deodhar, and Pierre-Louis, members of the initial project. This work was funded by the World Bank sponsored research project #63867 to UC Berkeley: Data Science Framework to Support Non-Motorized Transport. The authors also thank the Mobility and Logistics (MOLO) Multi-Donor Trust Fund, funded by the Austrian Ministry of Finance, German Ministry for Development Cooperation (BMZ), and Swiss Ministry of Economic Affairs (SECO). For contractual and privacy reasons, we cannot make the raw data available. We are pleased to make available the data of the OD matrices, software to replicate the methods, and the appropriate documentation. This information may be accessed at the GitHub repository https://github.com/humnetlab/BikesBogota/. This repository is sufficient to reproduce the results of this paper.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.trc.2020.102640.

## References

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C 58, 240–250.
Altrutz, D., Baer, R., Gwiazdy, P., Gaase, M., Hartkopf, G., Lerner, M. (Eds.), 2010. German guidelines for cycling infrastructure design (Empfehlungen für Radverkehrsanlagen, ERA 2010). FGSV Verlag GmbH.
Bao, J., He, T., Ruan, S., Li, Y., Zheng, Y., 2017. Planning bike lanes based on sharing-bikes' trajectories. In: KDD'17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, NY, USA. pp. 1377–1386.
Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: An open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.
Biko, 2019. Biking app. https://bikoapp.com/.
Blondel, V., Krings, G., Thomas, I., 2010. Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. Brussels Stud. [online] 42, 806.
Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. P10008.
Buehler, R., Dill, J., 2015. Bikeway networks: A review of effects on cycling. Transport Rev. 36, 9–27.
Casella, G., Robert, C.P., 2004. Monte Carlo Statistical Methods. Springer-Verlag.
Çolak, S., Lima, A., González, M.C., 2016. Understanding congested travel in urban areas. Nature Commun. 7, 10793.
Florez, M.A., Jiang, S., Li, R., Mojica, C.H., Rios, R.A., González, M.C., 2017. Measuring the impacts of economic well being in commuting networks - A case study of Bogota, Colombia. In: The 96th Annual Meeting of Transportation Research Board (TRB), Washington, DC. pp. 1–18. Paper Number 17-05211.
Ganin, A.A., Kitsak, M., Marchese, D., Keisler, J.C., Seager, T., Linkov, I., 2017. Resilience and efficiency in transportation networks. Sci. Adv. 3.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transp. Res. Part B 42, 759–770.
González, M., Hidalgo, C., Barabási, A., 2008. Understanding individual human mobility patterns. Nature 453, 779–782.
Hamedmoghadam, H., Ramezani, M., Saberi, M., 2019. Revealing latent characteristics of mobility networks with coarse-graining. Sci. Rep. 9, 7545.
Hariharan, R., Toyama, K., 2004. Project Lachesis: Parsing and modeling location histories. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (Eds.), Geographic Information Science. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 106–124.
Heinen, E., Panter, J., Mackett, R., Ogilvie, D., 2015. Changes in mode of travel to work: A natural experimental study of new transport infrastructure. Int. J. Behav. Nutrition Phys. Activity 12.
Hull, A., O'Holleran, C., 2014. Bicycle infrastructure: Can good design encourage cycling? Urban, Plann. Transport Res. 2, 369–406.
ITDP-México, I-CE., 2011. Manual integral de movilidad ciclista para ciudades mexicanas. Ciclociudades (Vol. V). http://ciclociudades.mx/manual-ciclociudades/.
Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J.J., Frazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing, ACM, New York, NY, USA. pp. 1–9.
Larsen, J., Patterson, Z., El-Geneidy, A., 2013. Build it. But where? The use of geographic information systems in identifying locations for new cycling infrastructure. Int. J. Sustainable Transport. 7, 299–317.
Li, D., Fu, B., Wang, Y., Lu, G., Berezin, Y., Stanley, H.E., Havlin, S., 2015. Percolation transition in dynamical traffic network with evolving critical bottlenecks. Proc. Nat. Acad. Sci. 112, 669–672.
Lovelace, R., Goodman, A., Aldred, R., Berkoff, N., Abbas, A., Woodcock, J., 2017. The Propensity to Cycle Tool: An open source online system for sustainable transport planning. J. Transport Land Use 10, 505–528.
Luxen, D., Vetter, C., 2011. Real-time routing with openstreetmap data. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, New York, NY, USA. pp. 513–516.
von Neumann, J., 1951. Various techniques used in connection with random digits. Monte Carlo methods. National Bureau Stand. 12, 36–38.
Newman, M., 2006. Modularity and community structure in networks. Proc. Nat. Acad. Sci. 103, 8577–8582.
Olmos, L., Çolak, S., Shafiei, S., Saberi, M., González, M.C., 2018. Macroscopic dynamics and the collapse of urban traffic. Proc. Nat. Acad. Sci. 115, 12654–12661.
QGIS Development Team, 2019. QGIS Geographic Information System. http://qgis.org.
Secretaría Distrital de Movilidad, 2011. Encuesta de movilidad (Mobility survey). Bogotá, Colombia. https://www.movilidadbogota.gov.co/.
Secretaría Distrital de Movilidad, 2015. Encuesta de movilidad (Mobility survey). Bogotá, Colombia. https://www.movilidadbogota.gov.co/web/encuesta_de_movilidad.
Secretaría Distrital de Planeación, 2017. Encuesta Multipropósito. http://www.sdp.gov.co/.
Secretaría Distrital de Seguridad, 2018. Convivencia y Justicia, Bogotá, Colombia. https://scj.gov.co/.
Toole, J.L., Çolak, S., Sturt, B., Alexander, P., L., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C 58, 162–167.
Uribe-Mallarino, C., 2008. Estratificación social en Bogotá: de la política pública a la dinámica de la segregación social. Universitas Humanística 65, 129–171.
Yildirimoglu, M., Kim, J., 2018. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. Transp. Res. Part C 89, 254–267.
Zeng, G., Li, D., Guo, S., Gao, L., Gao, Z., Stanley, H.E., Havlin, S., 2019. Switch between critical percolation modes in city traffic dynamics. Proc. Nat. Acad. Sci. 116, 23–28.
Zhang, D., Magalhaes, D.J.A.V., Wang, X.C., 2014. Prioritizing bicycle paths in Belo Horizonte City, Brazil: Analysis based on user preferences and willingness considering individual heterogeneity. Transp. Res. Part A 67, 268–278.