Measuring the impacts of economic well being in commuting networks — A case study of Bogota, Colombia

Manuel A. Florez
Department of Earth, Atmospheric and Planetary Sciences, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: +1(617) 3088716; Email: mflorez@mit.edu

Shan Jiang
Department of Civil & Environmental Engineering, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: +1(857) 654-5066; Email: shanjang@mit.edu

Ruiqi Li
School of Systems Science, Beijing Normal University
No. 19 Xinjiekouwai St., Haidian District, Beijing 100875, China
Department of Civil & Environmental Engineering, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: +1(857) 891-3655; Email: liruiqi@mit.edu

Carlos H. Mojica
Head of Strategic Planning, Service Planning and Technical Affairs
Transmilenio S.A., Bogota, Colombia
Email: carlos.mojica@transmilenio.gov.co

Ramiro A. Rios
Inter-American Investment Corporation
1300 New York Avenue, N.W. Washington, D.C. 20577, USA
Tel: +1(202) 623-2399; Email: rarios@iadb.org

Marta C. González *
Department of Civil & Environmental Engineering, MIT
77 Mass. Ave., Cambridge, MA 02139
Tel: +1(857) 928-4546; Email: martag@mit.edu

* Corresponding author

Word count: 5133 words + 8 figures × 250 words + 1 table × 250 words = 7383
Submitted: November 16, 2016

ABSTRACT

Big data such as call detail records (CDRs) from mobile phones are novel resources for travel demand models. An important open question is how to use them to extract practical information in relation to urban mobility, socioeconomic development, and well-being. Can we study individual mobility characteristics by income group through the lens of Big Data? In this paper, we present a data analysis framework that uses urban mobility extracted from CDRs, to study various characteristics of the commuting network of Bogota, Colombia, relating them to income groups by their residential location. We show that the diversity of commuting trips, defined in terms of entropy of the trips, increases with the income of the population. Further, we show that vehicle travel times during commuting hours from lower income groups clearly suffer longer congested travel times. Our results detail a method to use passively generated mobile phone data as a low cost alternative for transportation policies that can benefit from economic well-being measures for population with different income levels.

INTRODUCTION

Rapid urbanization has become a common theme across the world, and its influence is profound in Latin America, where over 80% of the population lives in cities. Massive urban migration has imposed enormous burdens to the existing infrastructure, leading to increased congestion, longer travel time and delays, and severe environmental degradation. Spatial segregation is also a salient issue (1). Oftentimes, unintended effects of government policies have further accentuated this problem. For example, stratification became widely spread in major Colombian cities, and it was in fact sectioned as law in 1994 (nueva ley de sevicios publicos). Of particular interest is the case of Bogota, the capital and largest city of Colombia, where careful socioeconomic stratification—the basis of a cross subsidization scheme of public utilities for low income residents (2, 3)— has been implemented since the Eighties. New insights into the impact of this kind of policy on individual mobility and socioeconomic well-being can be obtained through the lens of Big Data.

Latin American cities, as those in much of the developing world, exhibit sharp differences in income levels among various neighborhoods and are characterized by economic and urban spatial segregation, closely related to job opportunities, economic mobility, and travel behavior (1, 2). There has been a lot of qualitative work on understanding spatial segregation (3, 4, 5, 6, 7), however, there are still ample opportunities to better quantify it using new data resources. In particular, it is necessary to come with effective alternatives at low cost to measure impacts of the spatial organization of cities on the well-being of their inhabitants. Understanding the interplay between economic segregation and mobility as extracted from communication technologies is important for more efficient means of transportation and urban planning (8, 9).

To that end, we propose a framework of analysis to use big data through the lens of urban transportation (10) to quantify the impact of economic segregation on individual mobility. We utilize anonymized call detail records (CDRs) with an observation period of 6 months (in 2013 to 2014) to estimate the origin-destination (OD) matrices for a typical day in the city of Bogota, Colombia. We validate the ODs against the estimates from local official household travel survey (2011). We then create sampled networks with around $20,000$ travelers for each of the 4 income brackets, over which we estimate the duration of home-based-work (HBW) trips by income group (socioeconomic stratum). We show differences in their mobility diversity and congested travel times. We demonstrate that the impact of segregation on urban mobility can be extracted from CDR data, which can be used as a good alternative to time consuming and expensive travel surveys.

LITERATURE REVIEW

Spatial segregation in its most extreme form originated in the ghetto, a clustering in space of identifiable ethnic groups, often times instituted by political authorities. The causes and impacts of modern spatial segregation on society are a matter of debate (3, 4, 5, 6, 7) . Vandell et al. (11) argue that in the urban space, segregation is the consequence of at least three factors: administrative policy, market forces, and individual choices. While many studies have shown the socioeconomic consequences of segregation, few have focused on using information technologies to measure the impacts of segregation on urban mobility. The available information has been limited (12), restricting the sizes of representative samples by income groups, particularly for residents in poverty.

In order to improve the mobility of a city, transportation and urban planners need to quantify the interplay between travel demand and existing infrastructures. This is typically done via models

of travel demand, which estimate daily trips of individuals aggregated to origin-destination (OD) matrices (13) by mode, purpose, and time of the day. The seed information to create these trip matrices need to capture travel preferences of various population groups across the day. This requires careful sampling of preferences through representative travel diaries. While these surveys are rich in detail per individual, they are expensive and quickly become outdated. They only contain one or two days for a thin sample (usually 1 percent households in a metropolitan area) to model trips over the years in a city with millions of residents (12, 14). These limitations are very problematic as local municipal budgets are increasingly constrained and people's mobility patterns are dynamic. As a result, cheaper and more up-to-date data sources and analytical methods are called to increase the efficacy of urban travel demand models to empower urban and transportation planning in the age of information and telecommunications.

Recent research has explored the opportunity to utilize cell phone data, known as *call detail records* (CDRs) for modeling travel behavior. CDRs are metadata on phone usage (such as phone calls, data or text messaging) collected by a cellular carrier for billing and operational purposes, and contain geospatial whereabouts of users during their phone usage. As cell phones have become ubiquitous, a surge of research (15) have demonstrated that various insights on human behavior can be extracted from the massive, passively collected CDR datasets. Techniques have been used to identify daily mobility motifs, which have simplified the extraction of daily trip chains of individuals without using surveys (16). Mobile phone meta data have also changed the process of modeling the spreading of infectious diseases such as dengue and malaria (17, 18, 19).

The main advantage of individual phone records is that they contain hidden valuable information on the most visited locations by each user across time. Mobile market share typically covers a great fraction of the population during various months and each individual tends to visit few repeated locations in their journeys (20). While the geospatial tags are not accurate in space and time to generate complete journeys in a day, depending on the data collection technologies, treated with right methods, it is possible to obtain average transportation demand matrices by purpose and by hour-of-the-day that represent the travel demand of an entire city (10, 13, 21).

Of particular interest in this domain is the estimation and validation of OD matrices. Early works presented transient OD matrices using CDRs (22, 23) which mapped road usage to the neighborhoods that originated the travels. But their validation with existing models that also included mode, routes and travel times remained a challenge. In the last two years, the techniques to generate ODs from phone data as well as their validation have been further developed. Using only CDRs and census population data, (21) and (10) demonstrate techniques that can estimate ODs by purpose (e.g. commuting trips, or home-based-other) and by time-of-day (AM, PM etc.). The resulting ODs in Boston and Rio de Janeiro matched the ODs from existing models that required survey data. The work in (13) describes a system architecture for an end-to-end software solution that transforms and integrates mobile phone data into estimates of travel demand and infrastructure performance, and applies it in five cities comparing favorably with existing models based on surveys. CDRs and population data are consistently transformed into OD matrices by purpose and time of day, and routes through road networks are constructed using open and crowd-sourced data repositories. The analytics on the system's output is fast and portable.

In the following sections, we first describe the data used in this study. We then employ similar methods tested for other cities (10, 13, 21) to transform CDRs into ODs and validate the modeling results with local household travel survey data for Bogota, Colombia (12). Next, we present a network analysis method to enable the understanding of the interplay between economic

segregation and individual mobility. The main challenge is how to obtain accurate transportation information by aggregating trips without loosing the representation by income from origins of the trips. While the technique is applied from end to end in Bogota, the results are portable to any other region with census and CDR data available.

DATA AND METHODOLOGY

Data

To estimate travel demand for population by income stratum in Bogota, Colombia, we obtained (i) an anonymized CDR dataset from a telecommunication operator in Colombia, including information for 1.5 million users for a period of 6 months across 2013 and 2014, (ii) population at the census block level with their income stratum (ranging from 1 to 6, representing income level from low to high). To validate our estimate results, we also obtained: (iii) a latest set of household travel survey data (in 2011) for transportation planning purpose, officially authorized by the city's department of transportation and conducted by an international transportation consulting firm, and (iv) a set of car travel times for the estimated OD matrices at the tower-level in a weekday morning peak hour (7:00 am -8:00 am), queried from the API of an online mapping service. We discuss the details of these data sets in the following subsections.

Call Detail Records

The CDR dataset was gathered at the cellular tower-level, with 659 towers distributed across the Bogota metropolitan area (with an area of 477 square kilometers, and a population density of 16,143 persons per sq. km.). The study area encompassing the Capital District with its 20 localities (localidades) and the neighboring municipalities of Soacha, Mosquera, Funza, Madrid, Chía, Cajicá, Cota, La Calera, Tenjo, Tabio, Sibaté, Zipaquirá and Facatativá, for 912 transportation analysis zones (ZATs). Each record in the CDR dataset contains an anonymous user ID, the geographical location in the form of the latitude and longitude of the cellular tower, and the time at the instance of the phone activity. We analyze the OD trips at both celluar tower level for network analysis purpose and at the locality and ZATs level for OD validation purposes. Figure 1 A-C. shows the basic statistics on the CDR dataset.

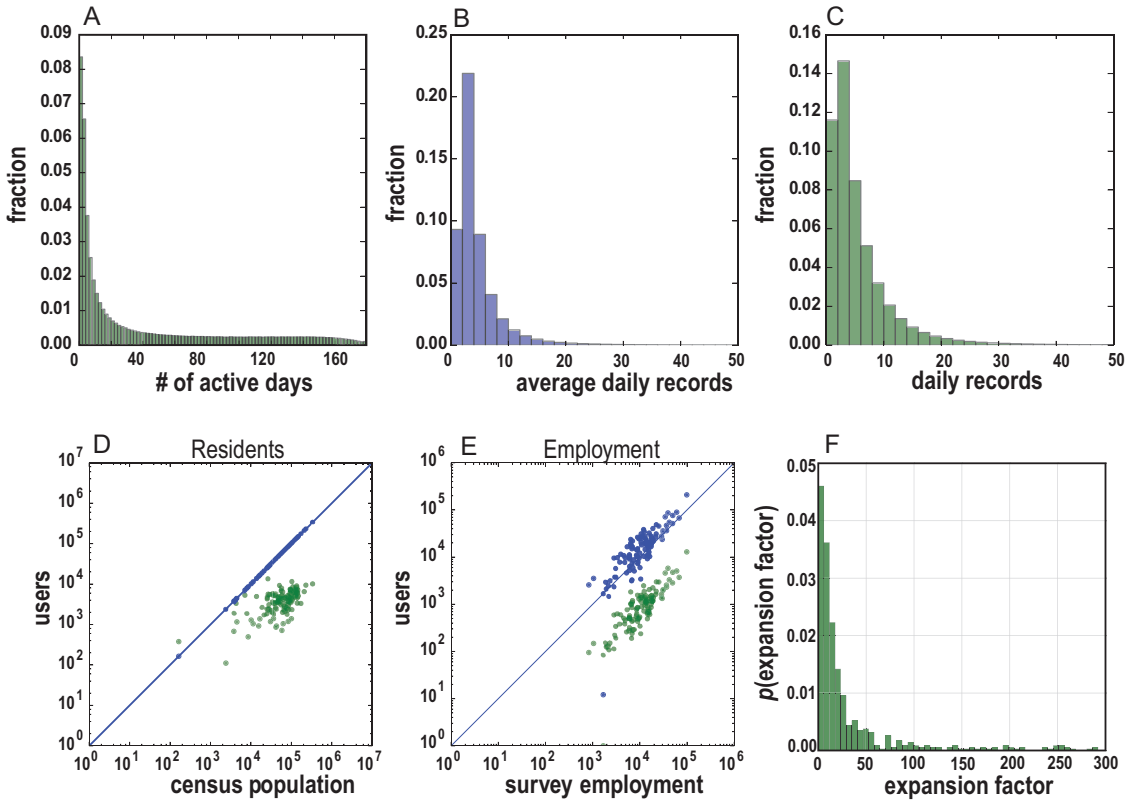

FIGURE 1 A-C. Distribution of number of active days of the Bogota phone data. B. Distribution of the individual users' average daily records. C. Distribution of daily phone records for all phone users over the 6-month observation period. D. Correlation between the WorldPop residential population and the mobile phone estimated residents before (green) and after (blue) expansion adjustment. E. Correlation between the employment population estimated from local household travel survey data and that estimated using the mobile phone data. F. Distribution of the expansion factors which are used to expand mobile phone users to total population.

Population and Income

To understand the economic status of local communities, we obtained a data-set at the city block level. The total population in Bogota is over 7 million. The census population data were also used to help expand the CDR sample users to the population for the Bogota metropolitan area, in order to generate representative urban travel demand estimates. To do so, in Fig. 1 D, we show the before (in green) and after (in blue) comparison of CDR users and census population at the ZAT level for Bogota. The purpose is to calculate expansion factors at the geographic level of analysis (e.g., shown here at the ZAT level). Fig. 1 F shows the distribution of expansion factors over Bogota.

We use the official socioeconomic stratification as a proxy for the spatial distribution of income within the city of Bogota. The department of city planning (DCP) is legally responsible for assigning a socioeconomic stratum to each city block (2). A scale from 1 through 6 is used, where 1 corresponds to the lowest socioeconomic level and 6 to the highest. The socioeconomic level of each city block is determined by a DCP official who relies on direct observation of the block and its surroundings and must take into account the following factors in his assessment (24): physical

characterizes of buildings, condition of local roads, presence and quality of sidewalks, ease of access to major roads and public transport, quality of urban space surrounding the block and overall urban context of the neighborhood. Socioeconomic stratification is the basis of a differentiated pricing scheme for public utilities (2). Residents of blocks classified in the three highest socioeconomic levels pay proportionally higher rates that are used to subsidize residents in the three lowest levels. Thus, careful assignment and regular updates to the socioeconomic stratification of the city are guaranteed by the current regulatory framework. Furthermore, in countries such as Colombia, where the informal economy plays a significant role, income data is grossly under-reported in both the census and local surveys (25). Using socioeconomic stratification as a proxy for income is a reasonable alternative that does not suffer from such biases.

There are about 45 thousand blocks classified by income (socioeconomic stratification) in the city but only 659 cell-towers. We need to assign and income rank to each cell-tower coverage area, This requires careful aggregation of the block-level income categories to the cell-tower scale. We map each city block to its corresponding cell-tower coverage area by calculating the centroid of the block and determining the converge area it falls within. Then, for each cell tower coverage area we calculate the population-weighted income rank:

$$\mu_n = \sum_{i=1}^{l_n} w_i s_i, \tag{1}$$

where l_n is the number of blocks that fall within coverage area n, w_i is the population weight of block i, defined as the fraction of the total population of block i, s_i is the discrete socioeconomic stratum (from 1 to 6) assigned to block i by the city planning officials.

Survey of Mobility

The 2011 Bogota Survey of Mobility (12, 26) was a local household travel survey, including 45 thousand individual samples to represent the over 7 million Bogota residents. The survey was conducted to collect trip information for residents during one sample day, including their trip purpose, and departure and arrival time and zonal information, and their social demographic information. Although there are more than 900 ZAT zones in the Bogota metropolitan area, due to the limited sample size, the survey only covered residents' travel in 767 ZATs. Using the estimates from the survey data, we validate our estimated employment population by inferring cellphone users' work location (discussed later), combined with our calculated expansion factors at the ZAT level (shown in Fig. 1 F). Fig.1 E shows high correlation between the employment estimates from the cellphone data model and those from the model informed by the survey data.

Travel Times in an AM Peak Hour

To asses the impact of spatial segregation on travel times, we estimate the trip duration for OD pairs that have a high number of home-based work trips. Bogota is a very congested city; estimates of trip duration have to take into account the increase in travel times during rush hours. To do so, we use the estimates of congested travel times provided by an online mapping service API (e.g., the Google traffic API, source: https://developers.google.com/maps/documentation/directions).

Generating OD Flows from CDR Data

Combined with population data and the CDR data, we generate ODs by trip purpose and by time-of-day based on three key steps: stay detection, activity labeling, and trip generation, as discussed

in (21).

Stay Detection

In order to discover users' activity location, we first filter out noise resulting from tower-to-tower call balancing performed by the mobile service provider, creating the appearance of false movements. We then employ a method (27) to distinguish users' stationary stay locations (when/where users engage in an activity) from their moving pass-by locations (when/where users are en-route to activities). As the data is tower-based, one can only know the closest tower to the user's actual location, so the estimate of a user's position is known up to the Voronoi cell for that tower. Due to the discrete nature of this data, the aforementioned call sequence simplification is carried out by joining sequences of calls made from a sets of towers within a certain distance threshold, followed by joining the sequence of calls made from the same tower. To address issues of temporal resolution, we only keep stays if the user is known to be in that location for at least 10 minutes.

Activity Labeling

To successfully extract purposes for every trip, we classify activities as *home*, *work* and *other*. Human mobility patterns as captured from mobile phone data that exhibit regularity and frequent returns to previously visited sites. For every user, her most visited location on weekdays from 7pm-8am and on weekends is classified as her *home*. Users with too little activity from their home locations are filtered out of the analysis. This is followed by assigning the user's *work* place, which is defined to be the non-home location that the user visits second most during the complement of the home time period on weekdays. Similarly, users with too few calls from their assigned workplace are excluded. Stays made from other locations are all classified under *other*. Once each stay is labeled with an activity purpose, then the resulting trips obtained from stay locations can be assigned with purpose pairs, such as home-based-work (HBW), home-based-other (HBO), or non-home-based (NHB). The ODs obtained in this way are then classified in terms of their purpose pairs. These methods are not necessarily definite solutions for perfectly estimating users' home and work locations, but they are straightforward and may lead in some cases to incorrect labeling of home and work locations. However, with increased spatial and temporal granularities of data and the inclusion of refined GIS information, more sophisticated algorithms can be developed.

Trip Estimation

After the call data have been assigned one of the three activity tags (*home*, *work* or *other*), the next step is to go through the time-ordered stay sequence for every user. Two consecutive calls on weekdays constitute a raw trip if they are not from the same location and are in the same effective day, which spans 3am of the previous day to the 3am of the next. Our method assumes that users typically travel from their home location at the beginning of an effective day and travel back home at the end. Therefore if a user's last call of the day is not from the home location, a raw trip is added to home. Similarly, if a user's first call of the day is made from a location other than home, a trip is added to ensure user's travel from her home. As CDR data is passive and generated when users choose to interact with their phones, one cannot assume that users start their trip at the exact time they make the call. We introduce a departure time estimation to account for the passiveness of CDR data as explained in (21). Fig.2 A-D shows the activity start time and duration by activity type of home, work, other, and all types. Fig. 2 E-H compares the trip departure time distribution by trip purpose of HBW, HBO, NHB and all types between the estimates of survey and the CDR-

data based model. The results show that in general the CDR based model presents similar temporal patterns to the survey, although there are some trade-off between the trips by trip purpose.

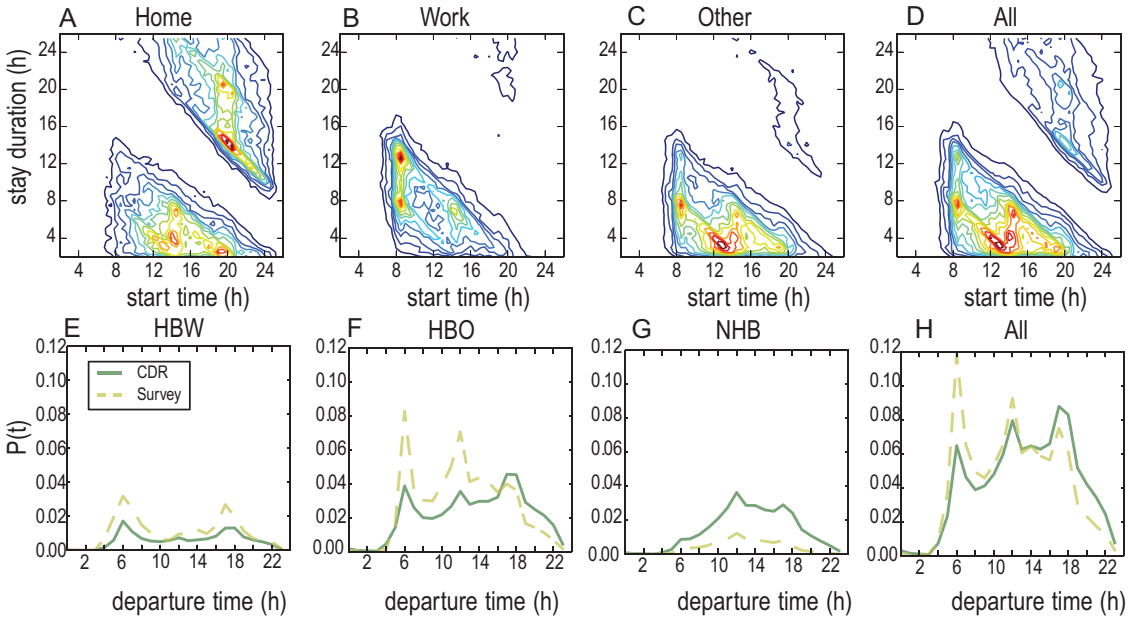

FIGURE 2 A-D. Temporal distribution (start time and stay duration) for inferred activities, including A. Home, B. Work, C. Other, and D. All types of activities. E-H. Comparison of trip start time by trip purpose, estimated with the local household travel survey data (in light green) and with the mobile phone data (in green), for different trip purposes including: E. Home-based-work (HBW), F. Home-based-other (HBO), G. None-home-based (NHB), and H. All trip purposes.

OD Generation

With the estimated trip departure time, and expanding the individual trips, we aggregate trips to construct OD matrices between zones by counting the number of expanded trips between each pair of zones on a typical day. We generated ODs for cell-tower based Voronoi-polygons, as well as for ZATs. We validate the ODs estimated from CDR data against those derived from the mobility survey at the ZAT level, since the survey only records trip origins and destinations at the ZAT level. We aggregated OD trips from ZAT to localities in Bogota. Fig. 3 shows the correlation of the ODs estimated from CDR and from survey data by time of day at the locality level. In general, the correlations of the CDR-based model and the Survey-based model are very high at the inter-locality level. The estimates of these two models on the total number of trips per day are reasonably close (as shown in Table 1). Given the good correlations and the validity of the method for other cities, we use the CDR-based travel demand model to further analyze the mobility characteristics by income group, because it is informed by many more users than the survey. This allows us to sample tens of thousands commuters per income bracket.

Network Analysis

The generated ODs naturally lend themselves to a network structure representation (10, 16, 28). The weight of the directed link connecting the origin node to the destination node is given by the

Bogota	HBW	NHB+HBO	AM	MD	PM	Total
CDR Trips (millions)	1.53	9.07	1.98	4.10	2.46	10.60
Survey Trips (millions)	2.82	9.80	1.99	5.09	2.48	12.62

TABLE 1 Number of trips on a typical day, split by trip purpose and time of day. AM, MD, and PM refer to 7am to 10am, 10am to 3pm, and 3pm to 7pm.

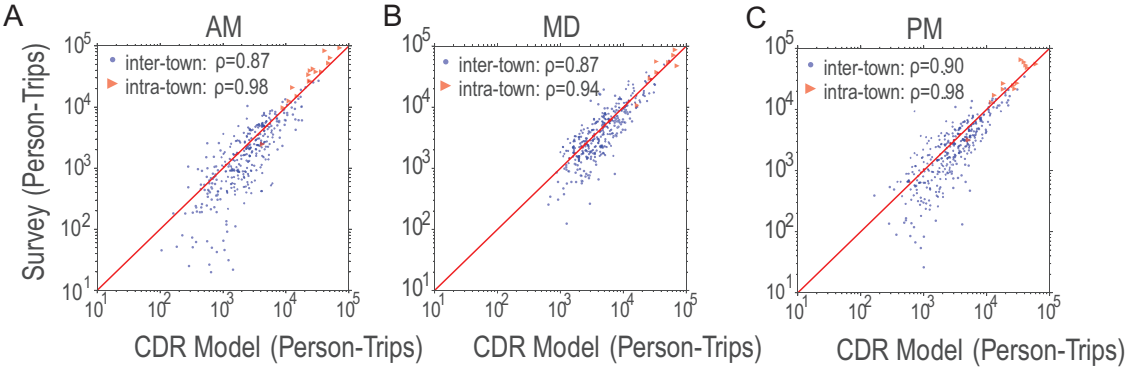

FIGURE 3 OD validation with Household Travel Survey at the Locality Level. by time of day, including A. morning (AM), B. midday (MD), C. evening (PM).

number trips estimated for that particular pair. Since we label trips by purpose, we select only OD pairs that represent commutes to work, as pairs where the number of HBW trips is significant. This selection enables us to quantify the relation between economic well being and the structure of the commuting network. The resulting network has 633 nodes and 27,939 links. Figure 5 shows the commuting network for Bogota using a geotagged layout, the distribution of degree or number of connections per node (Fig. 5B) and the trips per link (Fig. 5C). We see that while the degree distribution of the origins has an exponential decay, the distribution of the number of trips is broader with few OD pairs having between 100-1000 commuting trips. To assess the impact of income on the different network metrics we divide the nodes into four groups: low income, with weighted income stratum ranging 1.0 to 2.5 contained in 141 nodes, middle income, in the range 2.5-3.5 with 230 nodes, upper middle income, ranging in 3.5-4.5 with 134 nodes and higher income, from 4.5-6.0, and with 128 nodes.

In a seminal work Eagle et al. (29) analyzed landline calls and a nationwide mobile phone dataset in the UK. They proposed metrics of topological and spatial network diversity that displayed a strong correlation with the socioeconomic outcomes of the regional communities that each node represented. In this work we estimate mobility diversity (30) in similar way, first calculating the Shannon entropy of each node: $H_i = -\sum_{j=1}^{k} p_{ij} log(p_i j)$, where k is the number of destinations with origin in i, or the degree of node i. $p_i j$ is the relative proportion of trips between i and j: $p_{ij} = \frac{T_i j}{\sum_{j=1}^{k} T_{ij}}$. The spatial diversity of node i is then defined as:

$$D_i = \frac{H_i}{\log(k_i)}, \tag{2}$$

which is the ratio of the entropy observed in the trips divided by $H^{rand} = \log(k_i)$, the Shannon entropy when all trips from i are weighted equally among all the destinations, meaning $p_{ij} = 1/k_i$.

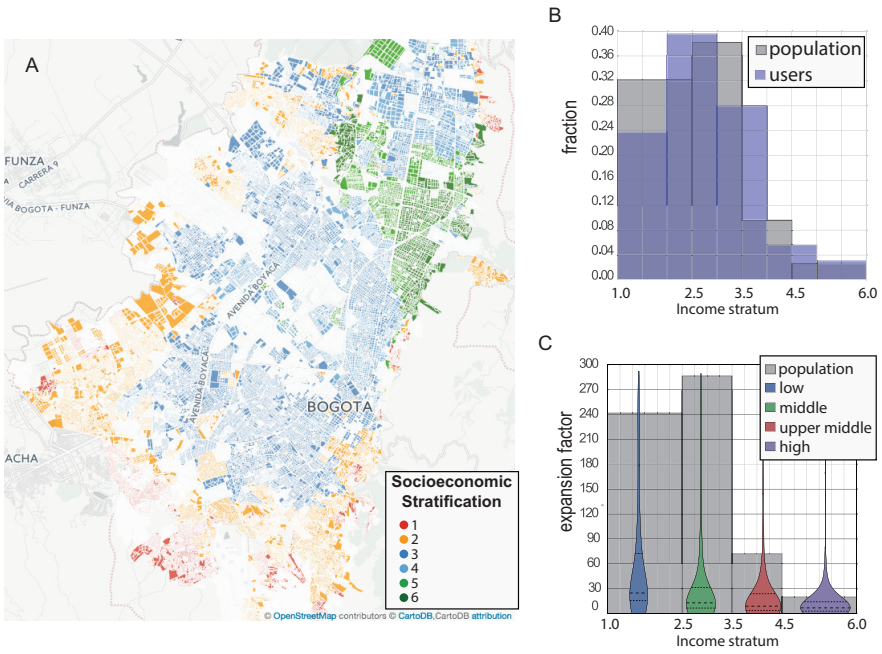

FIGURE 4 A. Map of population distribution by stratum from 1 to 6 going from lowest (1) to highest income (6). B. Population distribution of 4 income brackets (in gray) and fraction of users in each of 6 income strata. Income brackets are defined based on the average stratum on a given trip source. These are low income [1, 2.5], middle income [2.5-3.5], upper middle income [3.5-4.5] and higher income [4.5-6.0]. C. Distribution of factors to expand mobile phone users to population in the different origins separated in four income brackets– in each income bracket we have 1 user each 20 people, showing that the used mobile phone data cover all income levels with little bias.

Lower diversity implies that the commuting trips of origin i are more attracted to some destinations than others, among the k_i observed destinations. Entropy close to 1 implies that the trips are uniformly balanced. Also, the trip duration and the trip distance per origin are weighted over each node as follows: $< x >_i = \sum_{j=1}^{k} p_{ij} x_{ij}$, where x_{ij} is the quantity of interest as measured for OD pair i,j.

RESULTS

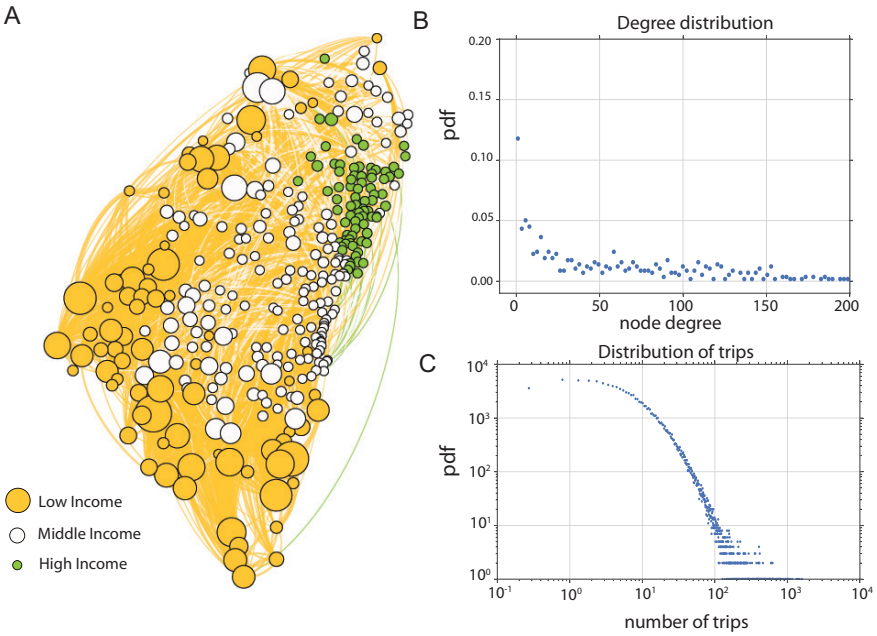

FIGURE 5 Network of commuting trips. A. 633 **origin nodes colored by income group (low, middle, high) and sizes representing the total number of commuting trips. B. Degree distribution represents the number of destinations of each node. C. Distribution of trips between the links.**

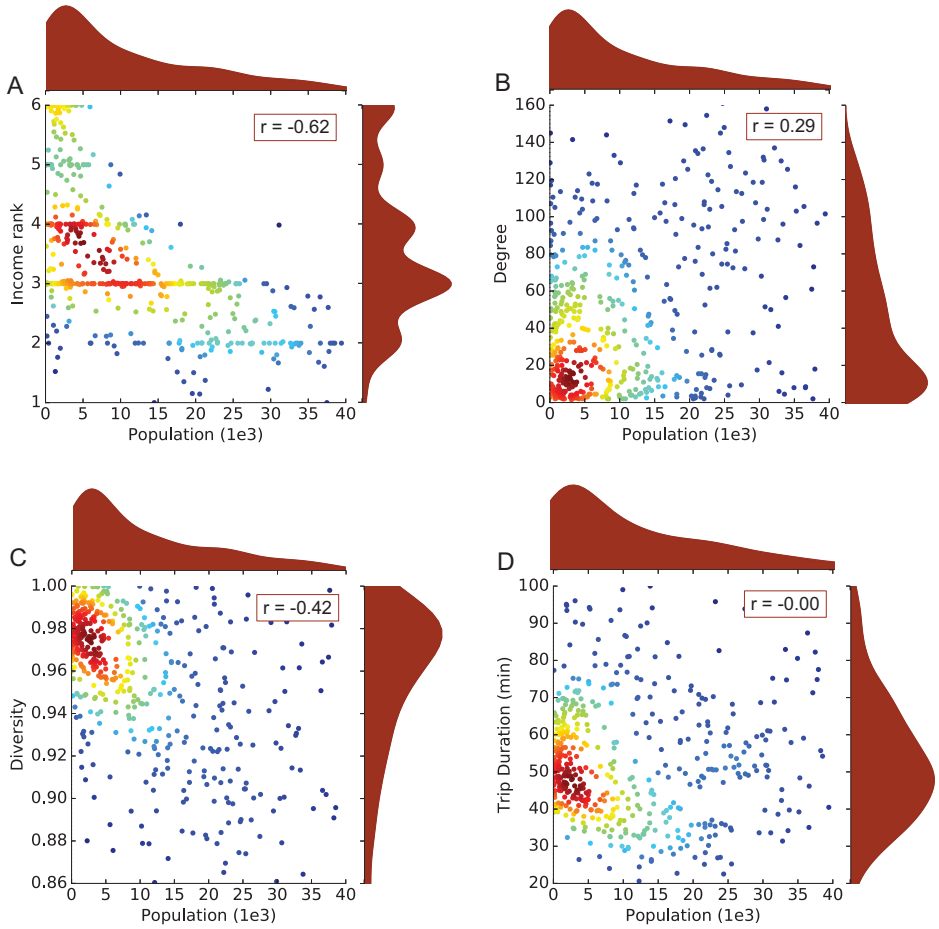

FIGURE 6 Income and various network metrics vs. population of the origins. A. Average income B. Degree C. Mobility diversity D. Commuting travel times. Most trip sources have around 5,000 people, those with larger population have lower income and lower mobility diversity. There is not correlation between the average commuting travel time and the population, and surprisingly, very low correlation between the degree and the population of origin.

We start by exploring the properties of the network in relation with the population of the origins (Figure 6). Most origins have similar population around 5,000 people. More populated nodes have lower income (Figure 6A) and slightly higher degree (Figure 6B). Origins with larger population have lower diversity (Figure 6C), indicating that they are preferentially attracted to particular destinations. There is not clear relation between the population of the origins and their average travel times in commuting (Figure 6D).

Next, we compare diversity and the entropy by income brackets (Figures 7 A-D). We observe that the higher the income the smaller the tendency of the origins to have preferential destinations for commuting (Figures 7 A-B), resulting in higher mobility diversity. Indeed, there is significant positive correlation between diversity and income rank (Pearson's r is 0.47). Entropy on the other hand does not display any significant relation with income (Figures 7C-D). In order to obtain estimates independent of the number of travelers per origin, we sample directly the commuting trips of $20,000$ mobile phone users for each of the 4 income brackets. They are selected with home locations following the population distribution of the city. Similar to the total network, the sampled network has 633 nodes and $25,073$ links. The observed effects of entropy and diversity by income group remain for both the sampled network and the entire network, as shown respectively in Figs 7 A vs. B and C vs.D.

Finally, we use the sampled network to study the distance to work, trips duration and time spent in congestion by income group (Figure 8). The sampling accounts for the wide differences in population among origins (Figure 6A), leading to similar statistics of the quantities of interest by income group and thus to more meaningful results. For each OD pair, congested travel times and road distances are queried form an online mapping service as discussed in the methods section. The distribution of values for the lowest income bracket tend to be broader. It is striking to see, however, that consistently, people in the lowest income group travel longer distances to work, spend more time in their commutes and are more affected by congestion. In fact, low income residents spend almost twice as much time in congestion than the high income sample (Figure 8C).

CONCLUSIONS

We use mobile phone data and query on-line maps to study the commuting network in Bogota, focusing on the relation between socioeconomic characteristics of the origins and their mobility characteristics. We find that mobility diversity has a significant positive correlation with income rank. Previous studies have confirmed this observation using information and communication technologies (ICTs) in the context of social contacts (29) and trips at a country scale (30). Eagle et al. (29) suggested that the diversity of social contacts is relevant because it showed very high degree of correlation (greater than 0.72) with socioeconomic indicators of well being. More recently, Pappalardo et al. (30) showed that social diversity may not be as important in the predictability of the Gini coefficient of a Municipality, and that the mobility diversity adds the largest predicting power to their regression models. The debate is only starting and many questions remain open, specially for both mobility and social networks within the urban scale, where less studies of this type have been done.

To our knowledge, this is the first study showing that mobility diversity is correlated with a socioeconomic indicator of well-being at the lowest possible degree of granularity, the mobile phone tower level. This opens up the possibility of using network based structural metrics from ICTs in the context of urban and transportation planning in relation to economic well being. In

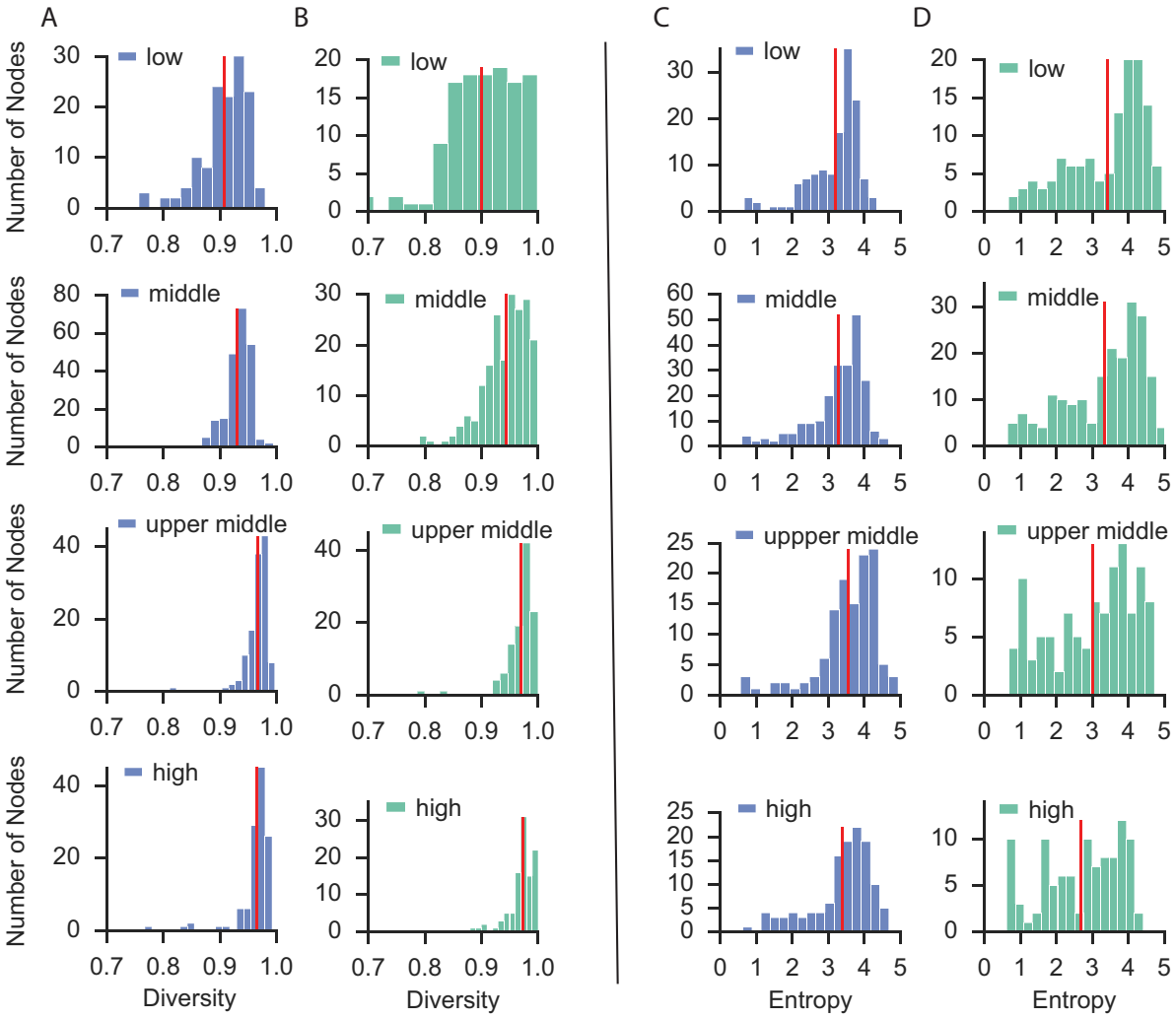

FIGURE 7 Distribution of mobility diversity and entropy by income group. A. Diversity in sampled network using same number of travelers per income group B. Diversity estimated in the complete commuting network of the CDR-based model C. Entropy in sampled network with same number of travelers per income group D. Entropy in the complete commuting network of the CDR-based model

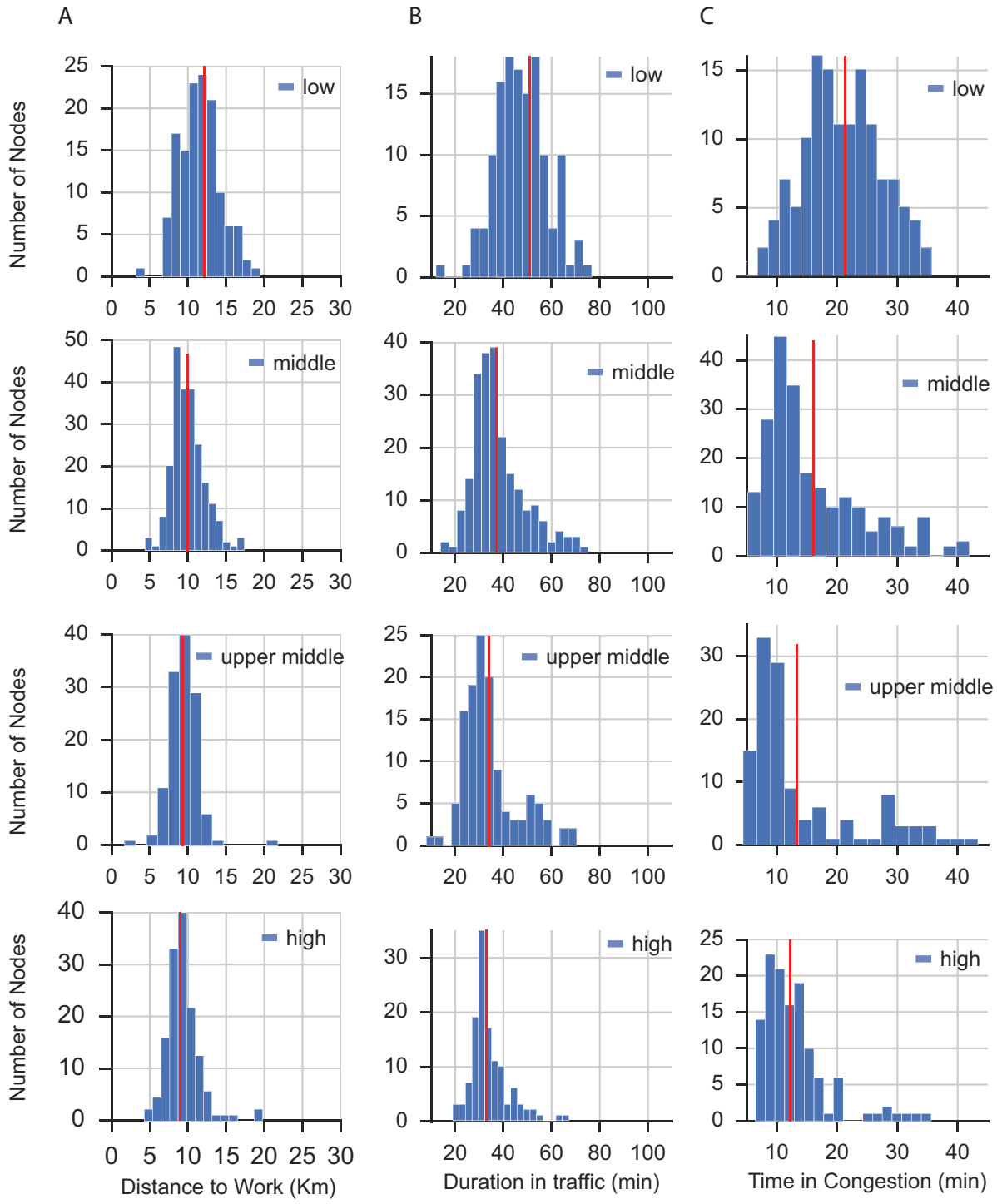

FIGURE 8 Travel by income group for commuters in a sampled network containing $20,000$ **travelers from each income bracket. A. Distance B. Duration of trips in congested travel time (calculated with the Google Maps api) C. Time spent in congestion.**

particular, the clear differences in the time spent in congestion by income group, suggest that travel information based on mobile phones can be used to measure the effects of transportation interventions to alleviate congestion for the most vulnerable sectors. Interesting avenues for future research related to this work, include the study of the characteristics of the social network in relation to the economic well-being at the urban scale. Also it is interesting to measure metrics of spatial exposure among different groups and how these affect the economic complexity of various cities. In the ideal scenario, we can learn types of urban organizations that promote better outcomes for their inhabitants. In the new data-rich reality of cities, deeper insights into their social and spatial connections will help make the places we live more sustainable, efficient and equitable.

ACKNOWLEDGMENTS

We thank Bradley Sturt, Nuria Oliver, Alvaro Ramirez Suarez, Gonzalo Durban Diez and Ricardo Hausmann for enlightening discussions during the design and motivation of this work. Data on the model based on the travel diary of Bogota was kindly provided by Laura Lotero. The research reported herein was funded in part by the Interamerican Development Bank, the MIT-Portugal program, the Samuel Tak Lee Real Estate Entrepreneurship Laboratory at MIT, DOT via the program New England UTC 25 and the Center for Complex Engineering Systems (CCES) at KACST.

REFERENCES

[1] Lora, E. et al., *The quality of life in Latin American cities: Markets and perception*. World Bank Publications, 2010.

[2] Uribe-Mallarino, C., Estratificación social en Bogotá: de la política pública a la dinámica de la segregación social. *universitas humanística*, Vol. 65, No. 65, 2008.

[3] Alzate, M. C., *La estratificación socioeconómica para el cobro de los servicios públicos domiciliarios en Colombia:¿ Solidaridad o focalización?* CEPAL, 2006.

[4] Benabou, R., Inequality and growth. In *NBER Macroeconomics Annual 1996, Volume 11*, MIT Press, 1996, pp. 11–92.

[5] Stiglitz, J. E., *The price of inequality: How today's divided society endangers our future*, 2012.

[6] Roback, J., Wages, rents, and the quality of life. *The Journal of Political Economy*, 1982, pp. 1257–1278.

[7] Blomquist, G. C., Measuring quality of life. *A companion to urban economics*, 2006, pp. 483–501.

[8] Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang, The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C*, 2016.

[9] Çolak, S., A. Lima, and M. C. González, Understanding congested travel in urban areas. *Nature Communications*, Vol. 7, 2016.

[10] Çolak, S., L. P. Alexander, B. G. Alvim, S. R. Mehndiretta, and M. C. González, Analyzing cell phone location data for urban travel: current methods, limitations and opportunities. In *Transportation Research Board 94th Annual Meeting*, 2015, 15-5279.

[11] Green, R. K., K. D. Vandell, et al., *Increasing Homeownership Opportunities through Modification of the IRS Rules Affecting Owner-Occupied Housing*. University of Wisconsin Center for Urban Land Economic Research, 1995.

[12] Lotero, L., A. Cardillo, R. Hurtado, and J. Gómez-Gardeñes, Several multiplexes in the same city: The role of socioeconomic differences in urban mobility. In *Interconnected Networks*, Springer, 2016, pp. 149–164.

[13] Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 2015.

[14] Stopher, P. R. and S. P. Greaves, Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, Vol. 41, No. 5, 2007, pp. 367 – 381, bridging Research and Practice: A Synthesis of Best Practices in Travel Demand Modeling.

[15] Blondel, V. D., A. Decuyper, and G. Krings, A survey of results on mobile phone datasets analysis. *EPJ Data Science*, Vol. 4, No. 1, 2015, p. 1.

[16] Schneider, C. M., V. Belik, T. Couronné, Z. Smoreda, and M. C. González, Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, Vol. 10, No. 84, 2013, p. 20130246.

[17] Wesolowski, A., T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee, Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, Vol. 112, No. 38, 2015, pp. 11887–11892.

[18] Wesolowski, A., N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, Quantifying the impact of human mobility on malaria. *Science*, Vol. 338, No. 6104, 2012, pp. 267–270.

[19] Tizzoni, M., P. Bajardi, A. Decuyper, G. K. K. King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza, On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology*, Vol. 10, No. 7, 2014, p. e1003716.

[20] Lu, X., L. Bengtsson, and P. Holme, Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, Vol. 109, No. 29, 2012, pp. 11576–11581.

[21] Alexander, L., S. Jiang, M. Murga, and M. C. González, Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 2015.

[22] Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González, Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63–74.

[23] Wang, P., T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, Understanding road usage patterns in urban areas. *Scientific reports*, Vol. 2, 2012.

[24] of Statistics (Colombia), N. A. D., *METODOLOGÃ■A DE ESTRATIFICACIÃŞN URBANA TIPO 1*, 2004.

[25] Jeffrey C. Moore, L. L. S. and J. Edward J. Welniak, Income Measurement Error in Surveys: A Review. *Journal of Official Statistics*, Vol. 16, No. 4, 2000, pp. 331–âĂŞ361.

[26] Secretaria Distrital de Movilidad, ., Informe de indicadores Encuesta de Movilidad de Bogotá, 2011.

[27] Hariharan, R. and K. Toyama, Project Lachesis: Parsing and Modeling Location Histories. In *Geographic Information Science* (M. Egenhofer, C. Freksa, and H. Miller, eds.), Springer Berlin Heidelberg, Vol. 3234 of *Lecture Notes in Computer Science*, 2004, pp. 106–124.

[28] Saberi, M., H. S. Mahmassani, D. Brockmann, and A. Hosseini, A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. *Transportation*, 2016, pp. 1–20.

[29] Eagle, N., M. Macy, and R. Claxton, Network Diversity and Economic Development. *Science*, Vol. 328, No. 5981, 2010, pp. 1029–1031.

[30] Pappalardo, L., M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti, An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2016, pp. 1–18.