Chapter 11
Characterizing Urban Mobility Patterns: A Case Study of Mexico City

Pierre Melikov, Jeremy A. Kho, Vincent Fighiera, Fahad Alhasoun, Jorge Audiffred, José L. Mateos, and Marta C. González

Abstract Seamless access to destinations of value such as workplaces, schools, parks or hospitals, influences the quality of life of people all over the world. The first step to planning and improving proximity to services is to estimate the number of trips being made from different parts of a city. A challenge has been representative data available for that purpose. Relying on expensive and infrequently collected travel surveys for modeling trip distributions to facilities has slowed down the decision-making process. The growing abundance of data already collected, if analyzed with the right methods, can help us with planning and understanding cities. In this chapter, we examine human mobility patterns extracted from data passively collected. We present results on the use of points of interest (POIs) registered on Google Places to approximate trip attraction in a city. We compare the result of trip distribution models that utilize only POIs with those utilizing conventional data sets, based on surveys. We show that an extended radiation model provides very good estimates

P. Melikov · J. A. Kho · V. Fighiera · M. C. González (✉)
University of California, Santa Barbara, USA
e-mail: martag@berkeley.edu

P. Melikov
e-mail: pierre_melikov@berkeley.edu

J. A. Kho
e-mail: jerkho@berkeley.edu

V. Fighiera
e-mail: vincent.fighiera@berkeley.edu
URL: https://github.com/VincentFig/urban_computing_mexico

F. Alhasoun
Massachusetts Institute of Technology, Cambridge, USA
e-mail: fha@mit.edu

J. Audiffred
Data Lab MX, Mexico City, Mexico
e-mail: ja@digitalstate.mx

J. L. Mateos
Universidad Nacional Autónoma de México, Mexico City, Mexico
e-mail: mateos@fisica.unam.mx

W. Shi et al. (eds.), *Urban Informatics*, The Urban Book Series,
https://doi.org/10.1007/978-981-15-8983-6_11

when compared with the official origin–destination matrices from the latest census in Mexico City.

Keywords Trip distribution models · Transit use · Clustering methods · Mobility science

11.1 Introduction

As more people continue to migrate from rural to urban settings, the challenges of improving cities increase in pace and complexity. Planning for daily mobility within metropolitan areas is one important topic of the coming years. The estimates of the total daily trips specific to a metropolis are the first step to establish efficient strategies that inform the transportation-planning process. However, the lack of reliable and accessible data sources of individual mobility greatly slows down the planning progress. Data on human mobility have thus far been collected through individual surveys with small and potentially biased sample sizes because they require active participation and often rely on self-reporting (Cottrill et al. 2013). While conventional travel surveys provide a wealth of valuable information, they are very expensive and time-intensive. For most major cities, these surveys are conducted about once a decade; for smaller cities and towns, it is less frequent than that or not at all. Between the publication of these surveys, a lot can happen that could change the dynamic of the city: new attractions, redevelopment of entire city blocks, changing economic trends, the impact of a natural calamity, or just the gradual shift of a city's characteristics. These changes would not be captured until the next travel survey is issued, which could be anywhere from the following year to a decade. With the abundance of information and connectivity today, other sources of easily accessible data could prove to be useful as a proxy for the data obtained in conventional surveys. One example of this is the use of triangulated mobile phone data to form mobility networks and extract individual trip chains (Jiang et al. 2013). Another such potential is points of interest (POIs) registered on Google Places, a feature of the mapping service developed by Google LLC (Google), which are extensive, updated frequently, and relatively accessible for most people. Google Places lists various types of establishments, such as restaurants, schools, offices, and hospitals, allowing it to serve as a good indicator of trip attraction. For an overview of mining POI data for urban land-use classification and disaggregation, see the work of Jiang et al. (2015).

As a complement to the development of statistical methods to carefully treat travel diaries (Ben-Akiva and Lerman 1985; Hall 1999; de Dios Ortúzar and Willumsen 2011), alternative, cheaper, and larger data sources are necessary to push our understanding of human mobility efforts further. The evolution of technology over the past decade has given rise to ubiquitous mobile computing, a revolution that allows billions of individuals to access people, information, and services through information technologies such as their cellular or mobile phones. Using today's large-scale computing infrastructure and data gathered from sensing technologies, one can

combine methods from computer science with urban planning, transportation, and environmental science, to tackle specific problems with fined-tuned methodologies in a data-centric computing framework.

Urban-science methods for characterizing human mobility should take into account the complexity of these dynamics. However, despite being a complex system, recent results have indicated some patterns or general features that can clarify these dynamics. These features are called universals in analogy with phenomena in the physical sciences. First, there is a set of models to analyze aggregated human mobility in cities or large-scale migrations. On the one hand, we have gravity-like models, and on the other radiation models (Simini et al. 2012). In 2008, González et al. (2008) used data from mobile phones to show that the step-length distribution can be described by a truncated power law. To understand the mechanism that gives rise to this distribution, the authors used the radius of gyration: a quantity that characterizes the radius enclosing the most visited locations of an individual over months of observation. Simulations suggest that the step-length distribution of the entire population is produced by the convolution of Lévy flight processes, each with a different characteristic jump size within the individual radius of gyration of each person. The observed power law is the result of the heterogeneity in the radius of gyration of the population. While the great majority of users have a radius of a few kilometers, there is a minority of users that cover thousands. Similar to the income and other variables following a power law, following the Pareto principle 80% of the distance covered comes from 20% of the subjects.

Another interesting pattern of human mobility is the interplay between randomness and predictability. There is a high rate of return to previously visited locations such as home or work. The nature of these returns follows a probability inversely proportional to the rank of the location, following then a Zipf law. Subsequent work by Song et al. (2010a, b) using data from mobile phones, revealed two important characteristics of human behavior. First, the number of distinct visited locations increases as a power of time with exponent less than 1, indicating a very slow rate of explorations. Second, the probability that an individual returns to a previously visited place scales with the inverse of the rank of that location, a phenomenon labeled as a preferential return. With a perspective from information theory, Song et al. (2010a, b) used different kinds of entropy measures to analyze the limits of predictability of human mobility.

Another approach to study human mobility is by mobility motifs, introduced by Schneider et al. (2013) as an abstract (semantic) way to define periodic trajectories in the daily movements of individuals. A daily mobility motif is a directed network (digraph) where unlabeled nodes represent locations and the edges are trips from one location to another. Counting motifs in data from mobile phones and traditional travel surveys, they amazingly found that despite over 1 million unique ways to travel between 6 or fewer locations, just 17 motifs are used by 90% of the population. For an overview of these works, see the papers by Jiang et al. (2013) and Toole et al. (2015), and the recent review of human mobility by Barbosa et al. (2018).

In this chapter, we focus on statistical methods of the type described above in the analysis and modeling of human mobility both in the aggregate and individually. We

take advantage of novel data sources passively collected, to enrich the information on human mobility patterns. Namely we parse an alternative source of geospatial data, apply trip distribution models to estimate aggregated trips, and implement unsupervised machine learning to characterize different types of commuters by their mode of transportation and travel time.

As a sample case, we focus on Mexico City, one of the largest cities in the world with over 21 million people in the greater metropolitan area. It is also one of the most important cultural and historical centers in the Americas. With such a large number of people and a high level of vibrancy, mobility in the region can be quite a challenge. In 2017, a major household travel survey (Encuesta Origen-Destino en Hogares de la Zona Metropolitana del Valle de Mexico 2017) was completed for the Metropolitan Zone of the Valley of Mexico. Conducted from January–March 2017, the survey obtained information to facilitate a better understanding of the mobility of the inhabitants in the metropolitan region. This includes data on trip generation, trip attraction, mode choice, trip purpose, trip duration, socio-demographics, and more, which is representative of 34.56 million daily trips occurring in our study zone.

11.2 Data Collection of POIs

In order to obtain POIs (Jiang et al. 2015) from Google Places, programming scripts were written to utilize the application programming interface (API) that Google provides (Documentation of Google Maps API no date). However, Google sets limits on the number of POIs a single request can return and on the number of API requests an account is allowed to make in order to differentiate commercial and non-commercial applications. While the conduct of this undertaking is non-commercial, the data to be collected tend to exceed Google's limitations. Hence, an efficient algorithm needs to be implemented to collect the most information from a minimal number of API requests.

To achieve this, API requests were framed and constrained by geometries defined by the Hexagonal Hierarchical Geospatial Indexing System (H3) of Uber Technologies, Inc (Uber Engineering 2018). Uber's H3 system is an application of the concept of fractals. Maps are divided into large hexagonal tiles, with each tile further divided into seven smaller hexagons. With 16 supported resolutions, the system is flexible to most use cases. Figure 11.1a shows a sample resolution applied to a district in Mexico City.

Hexagons serve as good approximations of circles while minimizing the overlap between cells. This is useful as the Google Places API requires a radius parameter within which the search for POIs will be made.

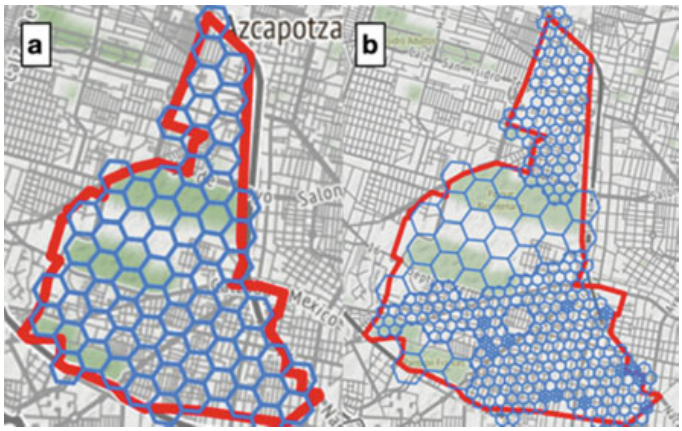

Fig. 11.1 Hierarchical sampling method to extract POIs. **a** Initial state and resolution of parsing algorithm, **b** Final state after recursively increasing ressolution in hexagons that reach the API request limit

11.2.1 Parsing Algorithm

An initial resolution for the size of the hexagons was determined. The coarser the initial resolution, the more efficiently the script is likely to run, as excessive requests are avoided in sparsely developed areas. On the other hand, coarse resolutions also increase the marginal areas near the borders of irregular shapes that are unaccounted by the algorithm. Before issuing any API request, the initial resolution was tuned and visualized to balance these tradeoffs.

For each hexagon, an API request was made at the centroid. If the request reaches the limit of POIs that it can return, the algorithm subdivides that hexagon into smaller hexagons. This process is repeated until each request is met without reaching the limit. In Fig. 11.1b, some areas, such as parks and nature reserves, do not need numerous API requests. Downtown city blocks and dense neighborhoods, on the other hand, are recursively splintered.

11.3 Spatial Distribution of POIs

In the use case for this chapter, the parsing algorithm returned a total of over 733,000 POIs from Google Places across the Metropolitan Zone of the Valley of Mexico. These points of interest provide new dimensions to analyze data from the travel survey that could generate insights on the characteristics of the megacity.

For instance, the API requests return tags for each POI, indicating the nature of the establishment. This may include broad categories, such as store, or more specific labels, such as electronic store. Clustering relevant tags together, POIs may be classified as either commercial or public-service establishments. Combining these data with the travel survey, Fig. 11.2a maps the relationship of the sociodemographic

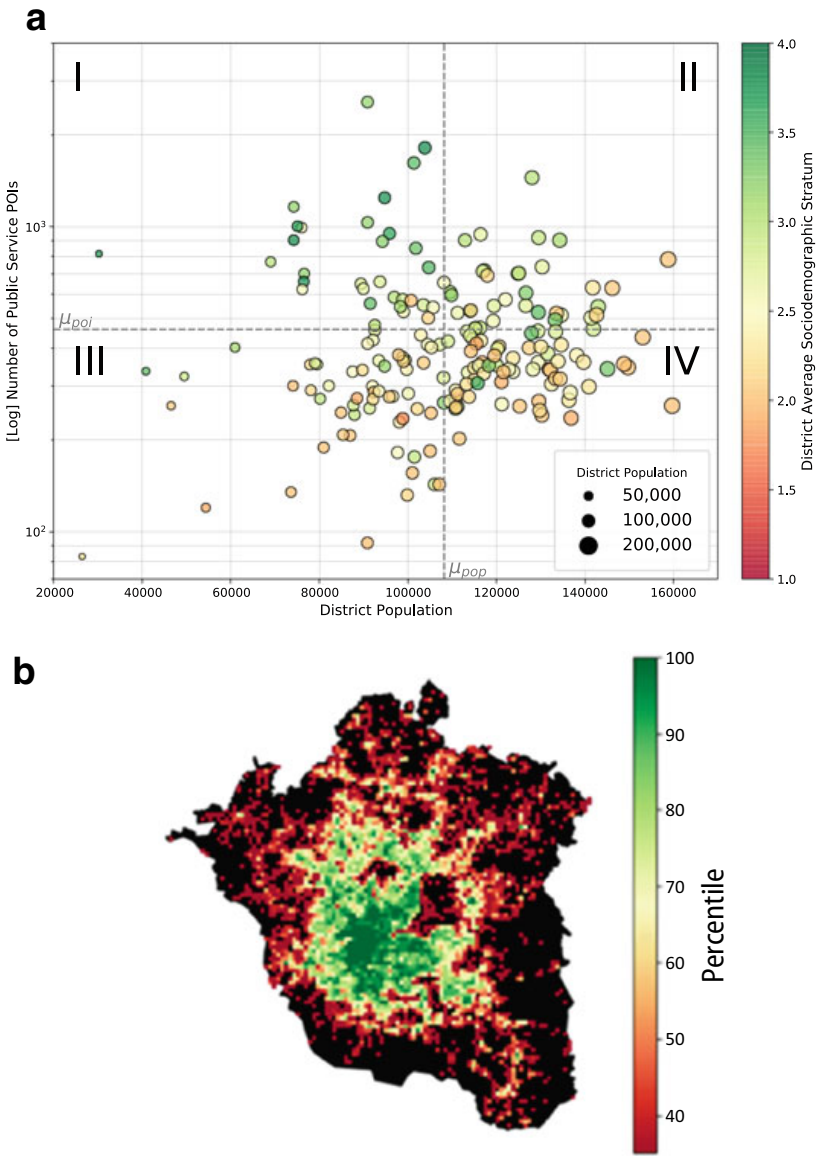

Fig. 11.2 Spatial distribution of population and services. **a** Relationship of the sociodemographic stratum of a district with the ratio of the number of public service establishments to the population, **b** Percentiles of the number of public service POIs for every 1 km^2 block

status of a district with the ratio of the number of public service establishments to the population.

In this case, sociodemographic strata are indices defined by the travel survey to characterize a respondent's social and economic conditions, with numbers from 1 to 4 denoting increasing economic well-being. In Quadrant I, the number of public-service establishments is above average and the population is below average: such districts tend to enjoy the highest sociodemographic stratum. Quadrant II has districts of intermediate sociodemographic status, still benefiting from an above-average number of POIs. Quadrant III has both less than the average population and number of

facilities and a lower socio-economic stratum. Interestingly, Quadrant IV has districts on opposite ends of the sociodemographic spectrum, possibly due to the diversity of inner cities and the efficiencies of density that allow fewer establishments to serve more people in a small amount of space. These enrich the spatial information of the surveys and deserve further research.

Another advantage gained through the POIs is the spatial granularity of the collected data. Travel survey respondents are often organized by the district of residence, whereas establishments on Google Places are pinpointed to street address coordinates. Since cities and districts are not homogeneous, this level of detail provides a more realistic perspective on city dynamics, highlighting functional interaction over arbitrary political boundaries.

In Fig. 11.2b, the coordinates of public-service establishments are truncated to two decimal places, binning them to grids that are approximately a kilometer per side. Due to the orders of magnitude in the difference between the urban core and more rural areas, the number of public-service establishments is abstracted to intervals of 5 percentile points. As it is, mapping these establishments may have a strong dependency on population density. Nevertheless, a hidden structure to the city is revealed, with a strong urban core, some urban corridors expanding outwards from the city center, and regional centers further away from the center. Significantly, there are large regions on the outskirts of the study area where public services are sparse. Further insights may be gained when supplemented by population distribution data at a similar level of granularity.

11.3.1 Extended Radiation Model for Human Mobility

Counting the number of POIs per district is necessary for direct comparison with the 2017 travel survey data, which have the smallest granularity only at the level of districts. Mapping these per district in Fig. 11.3a, b, a direct comparison can be made with trip attraction reported in the 2017 travel survey.

While the correspondence is not perfect, the distribution of points of interest makes a good approximation to the distribution of trip attraction obtained from the travel survey. Most notably, the difference between the city center and the rest of the region is similarly stark.

Plotting the relationship between trip attraction and points of interest in Fig. 11.3c yields a quantitative plot, with the correlation coefficient of the two variables determined to be quite high at 0.81. This comparison will be of great relevance later, where the POIs are used to model mobility patterns in the city, in place of travel-survey data.

Many models have been developed in order to predict population movement at different scales. In the context of Greater Mexico City, we want to investigate how accurate such models are and how well they perform to reconstruct mobility patterns. The models of trip distribution can be divided into gravity-model types (Barthélemy 2010; Erlander and Stewart 1990; Jung et al. 2008; Lenormand et al. 2016), or

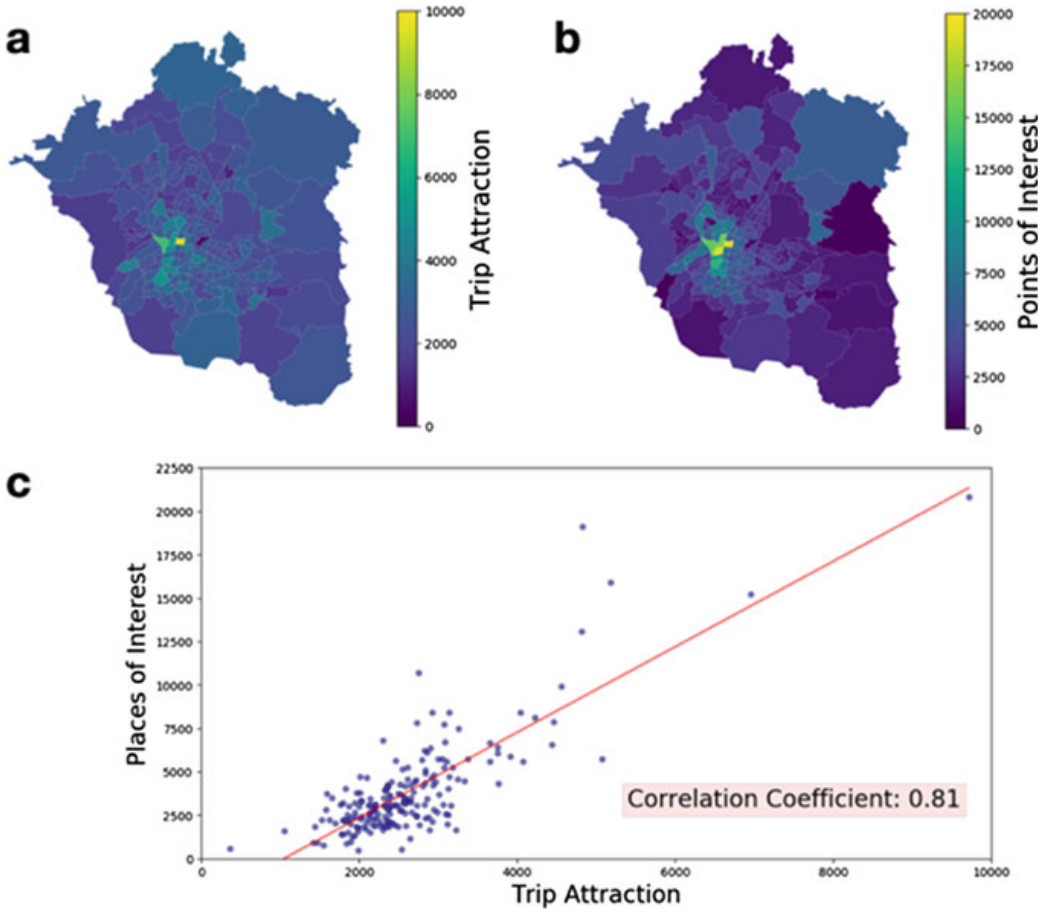

Fig. 11.3 Trip attraction versus POIs. **a** Values of trip attraction, **b** The number of points of interest, **c** Correlation plot of trip attraction and points of interest

intervening-opportunity types (Lenormand et al. 2016). In this chapter, we present an application of the latter, named the extended radiation model (Yang et al. 2014), to estimate trip distributions in Mexico City.

The radiation model (Simini et al. 2012, 2013) is based on a stochastic process that is parameter-free and enables, without previous mobility measurements, estimates of trip distributions in good agreement with mobility and transport patterns (Simini et al. 2013). The original radiation model only relies on population densities to estimate commuting patterns between US counties (Simini et al. 2013).

Here, we use the natural partition of the city in districts. The model states that a trip occurs based on the number of opportunities that can be found in each district if the two following steps are met: (1) an individual seeks opportunities from all districts, including his or her home district (the number of opportunities in each county is proportional to the resident population); (2) the individual goes to the closest district that offers more opportunities than his or her home district. To analytically predict the commuting fluxes with the radiation model, we consider locations i and j with population m_i and n_j, respectively, at distance r_{ij} from each other. We denote with s_{ij} the total population in the circle of radius r_{ij} centered at i (excluding the source and

destination population). The average flux T_{ij} from i to j is:

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_i + s_{ij})} \tag{11.1}$$

where $T_i = \sum_{i \neq j} T_{ij}$ is the total number of commuters that start their journey from location i, or the trip production of location i.

The extended radiation model aims at predicting flows without first calibrating the data. Thus, it introduces a scaling parameter $\propto$ by combining the derivation of the original radiation model with survival analysis and gives:

$$\langle T_{ij} \rangle = \gamma T_i \frac{[(a_{ij} + m_j)^{\propto} - a_{ij}^{\propto}](n_i^{\propto} = 1)}{(a_{ij}^{\propto} + 1)[(a_{ij} + m_j)^{\propto} + 1]} \tag{11.2}$$

where $a_{ij} = n_i + s_{ij}$, γ, is the percentage of trips between all places found between the origin and destination, and empirically set $\propto = (\frac{1}{36[\text{km}]})^{1.33}$, where $\boldsymbol{i}$ is the characteristic length of the study area, and $\propto$ accounts for the fact that the trip distributions depend on the area of study.

The extended radiation model was meant to be used when we lack trip data for calibration. When there are actual trip data as in this case, one can evaluate them with the common part of commuters based on the Sørensen index (Lenormand et al. 2016):

$$\text{CPC}(T, \tilde{T}) = \frac{2 \sum_{i=1}^{n} \sum_{j=1}^{n} \min(T_{ij}, \tilde{T}_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{n} T_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{T}_{ij}} \tag{11.3}$$

It gives a quantitative measure of the goodness of the flow estimation, 0 meaning no agreement found and 1 perfect estimation. CPC compares the model estimates T_{ij} versus the empirical observations $\tilde{T}_{ij}$, between all origin–destination pairs.

11.3.2 Results

From the survey data, we extracted the different variables to run the extended radiation model. First, we extracted the 194 districts that compose Greater Mexico City with their respective population, trip attraction (number of daily trips coming to the district), trip production (number of daily trips leaving from the district), points of interest, and characteristic length, given as the square root of the area of the district.

Then, we set i as the mean of the characteristic length of each district. We also constructed the distance matrix that gives for every row i and column j the distance between the centroids of the districts i and j. Finally, γ was set to the total number of trips as a proportion of the total population.

Table 11.1 Comparison of the goodness of fit depending on different input data in the model

Origin	Trip production	Trip production	Population	Population
Destination	Trip attraction	POI	Trip attraction	POI
CPC	0.69	0.67	0.64	0.63

Four different setups were then used to compare the performance of the model based on different approximations of the trip production from the origin districts and the trip attraction of the destination districts: (1) we used trip attraction and trip production as a baseline, (2) we used the number of POIs as a proxy for trip attraction, (3) we used population as a proxy for trip production, and (4) we combined (2) and (3). The resulting CPC values are shown in Table 11.1.

Table 11.1 shows that the CPC of the estimates of the extended radiation model was close to other recently proposed models (Lenormand et al. 2016). Moreover, we investigated the impact of different proxies for flow generation and attraction volumes as input in our model and found that the use of more easily acquired data sources such as population and POI density achieves nearly the same level of accuracy. POIs seem particularly interesting because they enable good estimates without travel surveys, but with data of much cheaper access. On the other hand, the use of population in place of trip production aims at predicting future mobility patterns given the knowledge of γ, the proportion of the total population of the system commuting, and assuming changes in this ratio. Here, we extracted γ from the 2017 survey and used it for the models. Consequently, we cannot validate the predictive power of the model; but nonetheless, when distorting the population data of each district by multiplying it by γ, we still observe encouraging results.

11.4 Analyzing Human Mobility by Mode of Transportation

This section is devoted to the analysis of individual travelers within Mexico City. One advantage of a broad user survey is to identify types of dominant behavior in the population, with respect to the modes of transportation used, their geographic distribution, and socio-demographic characteristics.

We analyzed the large database collected by the Mexico City survey, containing information on individual residents; it details information on more than half a million trips. For each trip identified, we have the mode of transportation, the districts of departure and arrival, the time of departure and arrival, the purpose of the trip, the gender of the traveler, and his or her age and socio-demographic stratum. As many as twenty different modes of transportation can be identified among the 196 districts of the survey.

We wanted to reduce the complexity of this information by grouping the trips based on transportation mode, without associating the other metrics. The latter would then

be involved in the analysis of clusters formed. In doing so, we sought to distinguish the main mobility behaviors, which would, in turn, combine various proportions of the possible transport modes and trip purposes.

By simple inspection, it is clear that all the means of transport mentioned in the database were not significantly present in the main groups of behaviors. We expected to see certain modes of transport, such as cars or walking, as the majority in certain behaviors and others, such as the category "Other means of transport," very poorly represented or even absent. It is, therefore, not necessary for such a large number of variables, initially twenty, to describe the individual trip database. We applied principal component analysis (PCA) to determine the main variables. This allowed us to reduce computation time and complexity when using a clustering algorithm. Projecting into a lower dimensional base informs our understanding (Eagle and Pentland 2009; Ibes 2015).

The PCA method aims to capture as much of the total variance of the data as possible with a reduced number of variables, called principal components (PC). Since the objective was to set the size of the new projected database such that the first N PCs had to account for 85% of the total variance, we, therefore, chose to keep only the first five PCs for the rest of the study (Shlens 2005).

To group trips around main behaviors, we used k-means clustering (Jiang et al. 2012). Each journey of the database was initially represented as a vector composed of zeros and ones, depending on the mode of transportation used. We only considered its projection in the PCs database when applying the k-means algorithm. K-means works iteratively to ultimately minimize the sum of the distances between each projected journey and the centroids of the clusters determined by the algorithm, and thus allows patterns to be identified within the dataset. As a result, we obtained a list that reflected the membership of each trip in a particular cluster. We also calculated the proportions of the modes of transport for each cluster to determine their average behavior (Jiang et al. 2012). While the ideal number of clusters can be estimated via various metrics, such as the elbow method, the best number of clusters depends on the interpretability of the data available. In this case, we decided to keep six clusters.

11.4.1 Detected Mobility Groups

Figure 11.4a at the top shows the six clusters that characterize daily mobility in Mexico City and their percentages. They represent the main ways of moving around the city. Since the database reports journeys, several of which may have been made by the same person, and residents can have several trips. The analysis groups journeys and not individuals. Note that these journeys also have the purposes of these trips such as: going home, going to work, errands, shopping, etc. Their average percentage is shown at the bottom of Fig. 11.4a. In the top of Fig. 11.4a, only the three most reported modes of transportation in each cluster are shown. Each of these components is associated in the y-axis with its fraction within the cluster. The % in the x-axis

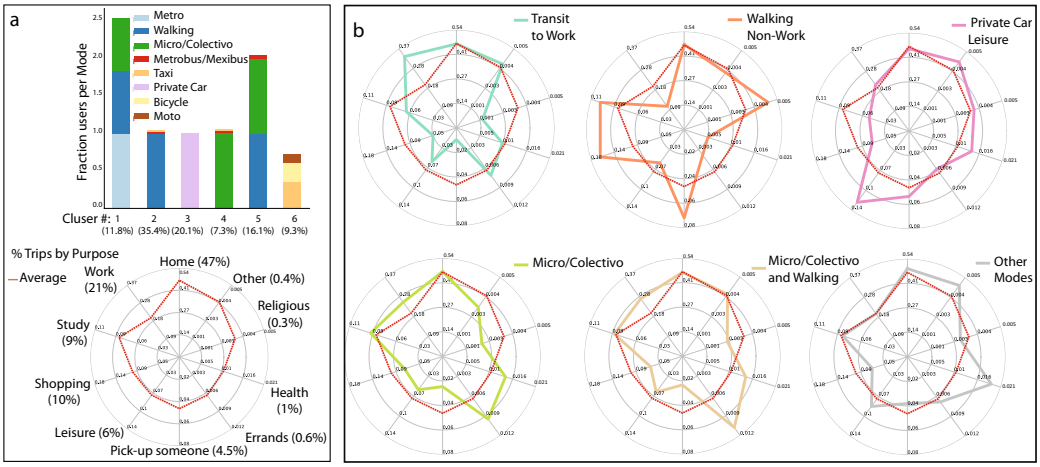

Fig. 11.4 Mobility groups in Mexico City. **a** The fraction of users per mode in each behavioral group or cluster. The lower part shows the legend displaying the percentage of trip purposes averaging the entire population. **b** Comparison of the percentage of trip purposes by cluster in contrast to the mean. The clusters are from 1 to 6 from left to right, starting at the top. We see that certain purposes are more present in each group. Cluster 1 uses combined transit modes with a higher percentage of work travel, Cluster 2 groups shopping, school, and social activities (picking someone up) by walking. Cluster 3 groups leisure trips via private car. Clusters 5 and 6 group errands done by Micro/Colectivo or combining Micro/Colectivo and Walking

shows the fraction of the total journeys in each cluster. We can see that the majority of journeys in Clusters 1 and 5 combines three or two modes respectively.

Cluster 2 contains 35% of all the trips in the Mexico City survey. The fraction of walking on the ordinate is equal to one, while that of the second most present mode of transportation in this cluster, Mexibus & Metrobus, has a fraction of 0.027. Thus, only about 2.7% of the trips attached to this cluster combined their walking with Mexibus or Metrobus. It can therefore be said that these trips are made almost exclusively by walking.

Figure 11.4b shows, for each of the six clusters, the proportion, per cluster, of each of the ten purposes of the trips considered in the survey: going to home, going to work, going to school, shopping, leisure, errands, picking someone up, religion, health purposes, or all other purposes.

We compared the average percentage of trip purposes with the average within each cluster. Cluster 1 represents 11.8% of all the trips and has 33% of them with work as its purpose, larger than the average of 21% among all trips. We see that when people walk (Cluster 2), the shopping purpose is twice the average. While about 16% of the trips associated with the second cluster are for shopping purposes, the average number for all trips is around 10% for this category. On the contrary, it seems that walking is not commonly used for commuting or going to the doctor.

In addition, since the average travel time of this cluster is about 20 min while the average travel time for the total population is about twice as long, this cluster can therefore be associated with local trips. This suggests that workplaces or healthcare

centers are generally located further from family homes than shops, schools, or religious places.

Cluster 3 groups 20% of the daily trips made in Mexico City; it is exclusively composed of private cars as a mode of transportation. This case has leisure in higher proportion compared to other clusters. This can be a consequence of the lack of transit to cover distant journeys, or being inconvenient for such purpose.

Cluster 5 contains 16% of the trips and includes the routes that exclusively combine walking and micro/colectivo, while Cluster 4 with 7% of the trips does not include walking. These two clusters are similar in purpose to the average and their average travel time is the longest, about one hour per trip.

The use of walking, metro and micro/colectivo during the same journey is also observed in the first cluster. Indeed, metro obtains a proportion equal to 1, walking 0.83 and micro/colectivo 0.71. Not all the journeys in this cluster, therefore, systematically combine these three means of transport, but on average in the great majority of cases these three means of transport are combined. This group is over-represented in the heart of the capital's historic district, where more than 55% of the trips undertaken are associated with this cluster. On the other hand, it becomes absent as soon as one moves away from this geographical area. This is due to the high concentration of metro and micro/colectivo in this part of the city, making travel much faster and more convenient by linking these modes of transport, particularly to get to work.

Cluster 6 is not possible to interpret, because it does not represent any particular mode. However, it should be noted that it is mainly concentrated in the agricultural regions that make up some districts.

Koelbl and Helbing analyzed data from the UK National Travel Surveys during nearly three decades, in the years 1972–98, observing that the average journey times for different modes of transport are inversely proportional to the energy consumption rates measured for the respective human physical activities. In Figure 11.5a, we show the distribution of the travel times per mode divided by their mean, inspired by the

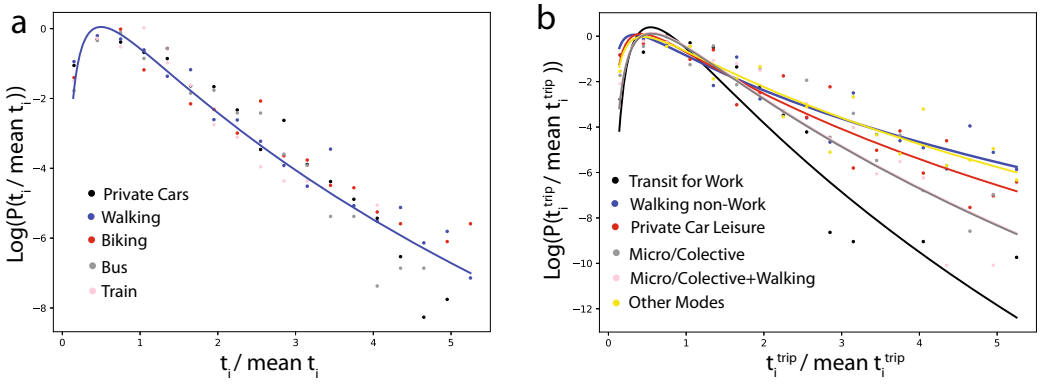

Fig. 11.5 Comparison of travel times by mode and by cluster group. **a** Lognormal fit for the scaled time-averaged travel-time distributions for different modes of transport on a logarithmic scale as reported by Schneider et al. (2013) based on UK surveys. **b** Lognormal fit for the scaled time-averaged travel-time distributions for the clusters found in the Mexico City travel survey

Table 11.2 Comparison of the fitted parameters for the clusters

Cluster	μ	var	mean trip time
Transit to work	−0.03	0.21	89
Walking non-work	−0.28	0.63	20
Private car + leisure	−0.19	0.69	40
Micro/collective	−0.07	0.41	49
Micro/collective + walking	−0.11	0.41	58
Other modes	−0.30	0.47	30
Results Kölbl et al. (2003)	−0.14	0.51	N/A

results reported by Kölbl and Helbing (2003). The authors presented five transport modes, and they all collapse well in one lognormal distribution with parameters reported in Table 11.2. To further investigate our clusters, we made the same analysis of the travel time of the individual trips divided by the mean travel time. We observed a lognormal with different parameters for each cluster; only Cluster 5 has closer parameters to the ones reported by Kölbl and Helbing (2003). Given the challenges of mobility in Mexico City, we observed larger variance among the members of each cluster, except for the trips of Cluster 1, which groups a higher fraction of the journeys to work. The differences between the results reported in the UK and Mexico City could be related to more strained transit service and longer commuting journeys in a vast metropolis. The universal scaling which is shown in different modes by Kölbl and Helbing (2003) could still serve as a guide to target improvements in the transit system. Note that the variance of private-car travel times is less than half that for transit. If the travel times were more similar, transit could be more attractive for those that can afford traveling by private car.

11.5 Conclusions

Data-informed analysis of complex socio-technical systems has become the interest of interdisciplinary groups around the world. These techniques can inform urban planning with an analytical angle in the complex task of amending current cities and their infrastructures. This increases its relevance to better accommodate the continued expansion of major cities and metropolises around the world. The purpose of this study was to summarize statistical methods to analyze human mobility in the urban context. We combined alternative data sources and methods in the topic that has mostly used travel diaries and econometric methods. The common aim of the data analysis presented is to reduce the complexity of the dataset at hand, while simultaneously extracting useful information. To this end, the recent growth of passively collected data lends important opportunities to the understanding and the implementation of these and other methods. In particular, we analyzed and modeled human mobility in Greater Mexico City, one of the largest cities in the world with over

21 million people. We explored a data set of a recent major travel survey conducted in 2017, using clustering methods, and compared the trip distributions with the one inferred from an extended radiation model that uses population and points of interest.

Future extensions should include the sociodemographic stratum, and possible interventions to plan for social equity and accessibility.

Acknowledgements We are grateful to Emmanuel Landa and Irving Morales of DataLabMX for collaborating with us in collecting data and in gaining better insights into the Metropolitan Zone of the Valley of Mexico. These contents of this chapter were initiated as a class project based on the content covered in CYPLAN 257: Data Science for Human Mobility and Sociotechnical Systems. The codes and data used in this chapter are available at https://github.com/VincentFig/urban_computing_mexico.

References

Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: models and applications. Phys Rep 734:1–74

Barthélemy M (2010) Spatial networks. Phys Rep 499:1–101

Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge

Cottrill CDA, Pereira FCA, Zhao FA, Dias IF, Lim HB, Ben-Akiva ME, Zegras PC (2013) Future mobility survey. Transport Res Record 2354:59–67

de Dios OJ, Willumsen LG (2011) Modelling transport. Wiley, Chichester

Documentation of Google Maps API. https://developers.google.com/places/web-service/search

Eagle N, Pentland AS (2009) Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63:1057–1066

Encuesta Origen-Destino en Hogares de la Zona Metropolitana del Valle de Mexico (2017) Instituto Nacional de Estadistica y Geografia, Mexico. https://en.www.inegi.org.mx/programas/eod/2017/. Accessed 11 Oct 2018

Erlander S, Stewart NF (1990) The gravity model in transportation analysis: theory and extensions, vol 3. Vsp

González MC, Hidalgo C, Barbási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782

Hall RW (ed) (1999) Handbook of transportation science. In: International series in operations research and management science, vol 23. Springer, Boston

Ibes DC (2015) A multi-dimensional classification and equity analysis of an urban park system: a novel methodology and case study application. Landscape Urban Plann 137:122–137

Jiang S, Ferreira J, González MC (2012) Clustering daily patterns of human activities in the city. Data Min Knowl Disc 25:478–510

Jiang S, Fiore GA, Yang Y, Ferreira Jr J, Frazzoli E, González MC (2013) A review of urban computing for mobile phone traces: current methods, challenges, and opportunities. Paper presented at the 2nd ACM SIGKDD international workshop on urban computing, Chicago, Illinois, August 2013.

Jiang S, Alves A, Rodrigues F, Ferreira J Jr, Pereira FC (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. Comput Environ Urban Syst 53:36–46

Jung WS, Wang F, Stanley HE (2008) Gravity model in the Korean highway. Europhys Lett 81(4):48005

Kölbl R, Helbing D (2003) Energy laws in human travel behaviour. New J Phys 5(1):48

Lenormand M, Bassolas A, Ramasco JJ (2016) Systematic comparison of trip distribution laws and models. J Transp Geogr 51:158–169
Schneider C, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. J R Soc Interface 10(84):20130246
Shlens JA (2005) Tutorial on principal component analysis (December 10, 2005; Version 2)
Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. Nature 484(7392):96–100. https://doi.org/10.1038/nature10856
Simini F, Maritan A, Neda Z (2013) Human mobility in a continuum approach. PLoS ONE 8:e60069
Song C, Koren T, Wang P, Barabási AL (2010a) Modeling the scaling properties of human mobility. Nature Phys 6:818–823
Song C, Qu Z, Blumm N, Barabási AL (2010b) Limits of predictability in human mobility. Science 327(5968):1018–1021
Toole JL, de Montjoye YA, González MC, Pentland AS (2015) Modeling and understanding intrinsic characteristics of human mobility. In: Goncalves B and Perra N (eds) Social phenomena: from data analysis to models. Springer
Uber Engineering (2018). https://eng.uber.com/h3/
Yang Y, Herrera C, Eagle N, González MC (2014) Limits of predictability in commuting flows in the absence of data for calibration. Sci Rep. https://doi.org/10.1038/srep05662

Pierre Melikov holds a Master of Science in Systems Engineering of the Department of Civil and Environmental Engineering of the University of California, Santa Barbara, and also a Master's degree from CentraleSupélec.

Jeremy A. Kho is a Master of Science graduate in Civil Systems Engineering at the University of California, Santa Barbara, and a Bachelor of Science graduate in Civil Engineering at the University of the Philippines, Diliman. He is the Data Science Lead in GrowSari, a leading enterprise e-commerce startup in the Philippines.

Vincent Fighiera is a master's degree student in Urban Computing at UC Berkeley and earned an engineering master's degree from the École Nationale Supérieure d'Arts et Métiers, Paris, France. His research mainly deals with a game theory approach on the impact of app use on traffic patterns with the Institute of Transportation Studies at UC Berkeley.

Fahad Alhasoun is a Ph.D. candidate in the Computational Science and Engineering program at Massachusetts Institute of Technology, his Ph.D. work focuses on applications of machine learning in urban computing using street view imagery. His research interest is in machine learning and applications across domains.

Jorge Audiffred is a Mexican entrepreneur who applies Data Science to design informed strategies for, among other things, the improvement of mobility in urban areas. He is the Founder of Data Lab Mx, based in Mexico City.

José L. Mateos is Research Professor at the Institute of Physics and Research Director of the Center of Complexity Sciences at the National Autonomous University of Mexico UNAM. His areas of expertise include Statistical Physics, Non-linear Dynamics, Network Science and Urban Mobility.

Marta C. González is an Associate Professor in Engineering and City & Regional Planning at UC Berkeley, she leads the Human Mobility and Networks Laboratory. Her group applies Complex Network and Complex Systems Sciences to understanding and planning for the interactions of humans with the built and natural environments.