A Review of Urban Science Methods for Characterizing Mobility Patterns

A Case Study of Mexico City

Vincent Fighiera, Jeremy A. Kho, Pierre Melikov and Marta C. González

Abstract Throughout the world, people struggle with safety, economic opportunity, and health. Seamless access to destinations they value, such as workplaces, schools, hospitals, and parks, influences their quality of life. One of the first steps to planning and improving accessibility is to estimate the number of trips being made across different parts of a city. However, the spatial distribution and availability of urban services, may come short in supplying the needs of their inhabitants. Relying on expensive and infrequently collected travel surveys for modeling trip distributions to their facilities slow down the decision making process. The growing abundance of data already collected, if analyzed with the right methods, can help us in planning and understanding cities. In this chapter, we examine the use of points of interests (POIs) registered on Google Places to approximate trip attraction in a city. First, 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 provides a high degree of accuracy when compared with the official Origin-Destination Matrices from the latest census in Mexico City. Moreover, we apply percolation theory to the POIs to produce meaningful insights into the functional dynamics of the city. In particular, we estimate the available number of facilities per capita, taking into account the different income groups. We illustrate novel applications of urban science, adapted to data readily available worldwide in the form of geo-mapped points of interest and population density. Related codes are available on-line.

V. Fighiera (vincent.fighiera@berkeley.edu) – J. A. Kho (jerkho@berkeley.edu) –
P. Melikov (pierre_melikov@berkeley.edu) – M. C. González (martag@berkeley.edu)
University of California, Berkeley
Berkeley, CA

https://github.com/VincentFig/urban_computing_mexico

1 Introduction

As more people continue to migrate from rural to urban settings, the challenges of improving the cities we live in increase in pace and complexity. The analysis and understanding of mobility, particularly within large metropolitan areas, is becoming one of the crucial issues of the coming years. Climate change, overpopulation, and the cost of living in highly urbanized areas are issues at the heart of public policies designed to better meet the population's transport needs while limiting its negative economic and environmental impact. Inevitably, an understanding of mobility specific to a metropolis is therefore necessary to establish efficient strategies.

Taking advantage of advancing methodologies and improved computational capacities of today, this chapter explores novel approaches to analyzing human mobility patterns. These include parsing an alternative source of geospatial data, applying a physics model to simulate travel behavior, and implementing unsupervised machine learning techniques to characterize different types of commuters through the set of modes they take on a regular basis.

Case Example

In the sample use case for this chapter, 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 amount of people and a high level of vibrancy, mobility in the region can be quite a challenge.

In 2017, a major household travel survey [1] was completed for the Metropolitan Zone of the Valley of Mexico. Conducted from January to March of 2017, the survey obtained information to facilitate better understanding of mobility of the inhabitants in the metropolitan region. This included 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 the study zone.

Need for Alternative Sources of Data

While conventional travel surveys provide a wealth of valuable information, they are also very expensive and time-intensive. For most major cities, these are conducted about once a decade; for smaller cities and towns, it is more seldom than that or none at all. Between the of publications 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, 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 proxy to the data obtained in conventional surveys. One example of this is the use of triangulated mobile phone data to form mobility networks with trip chains. [2]

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.

2 Data Collection of POIs

2.1 Problem Setup

In order to obtain POIs from Google Places, programming scripts are written to utilize the Application Programming Interface (API) that Google provides [3]. 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 tends 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 are framed and constrained by geometries defined by the Hexagonal Hierarchical Geospatial Indexing System (H3) of Uber Technologies, Inc (Uber) [4]. 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 1a shows a sample resolution applied on a district in Mexico City.

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

2.2 Parsing Algorithm

An initial resolution for the size of the hexagons is determined. The lower the initial resolution, the more efficient the script is likely to run, as excessive requests

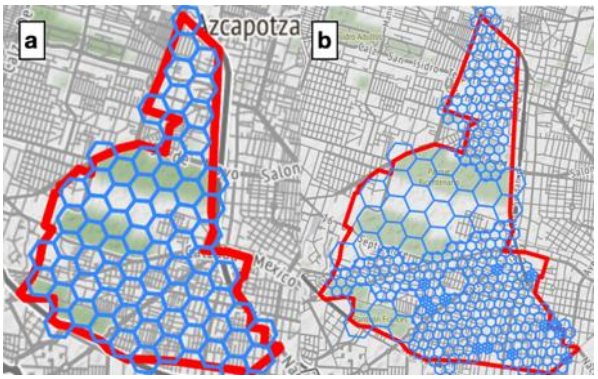

Fig. 1 a Initial state and resolution of parsing algorithm, **b** Final state after recursively increasing resolution in hexagons that reach the API request limit

are avoided in sparsely developed areas. On the other hand, low resolutions also increase the marginal areas near the borders of irregular shapes unaccounted by the algorithm. Before issuing any API request, the initial resolution is tuned and visualized to balance these tradeoffs.

For each hexagon, an API request is made at the centroid. If the request reaches the limit of POIs 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 Figure 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.

3 Exploratory Analysis

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 analyzing data from the travel survey that could generate insights on characteristics of the megacity.

3.1 Supplementary Data to Travel Surveys

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 'electronics store'. Clustering relevant tags together, POIs may be classified as either commercial or public service establishments. Combining this new dimension of data with the original travel survey, Figure 2a maps the relationship of the sociodemographic status of a district with the ratio of the number of public service establishments to the population. In this case, sociodemographic

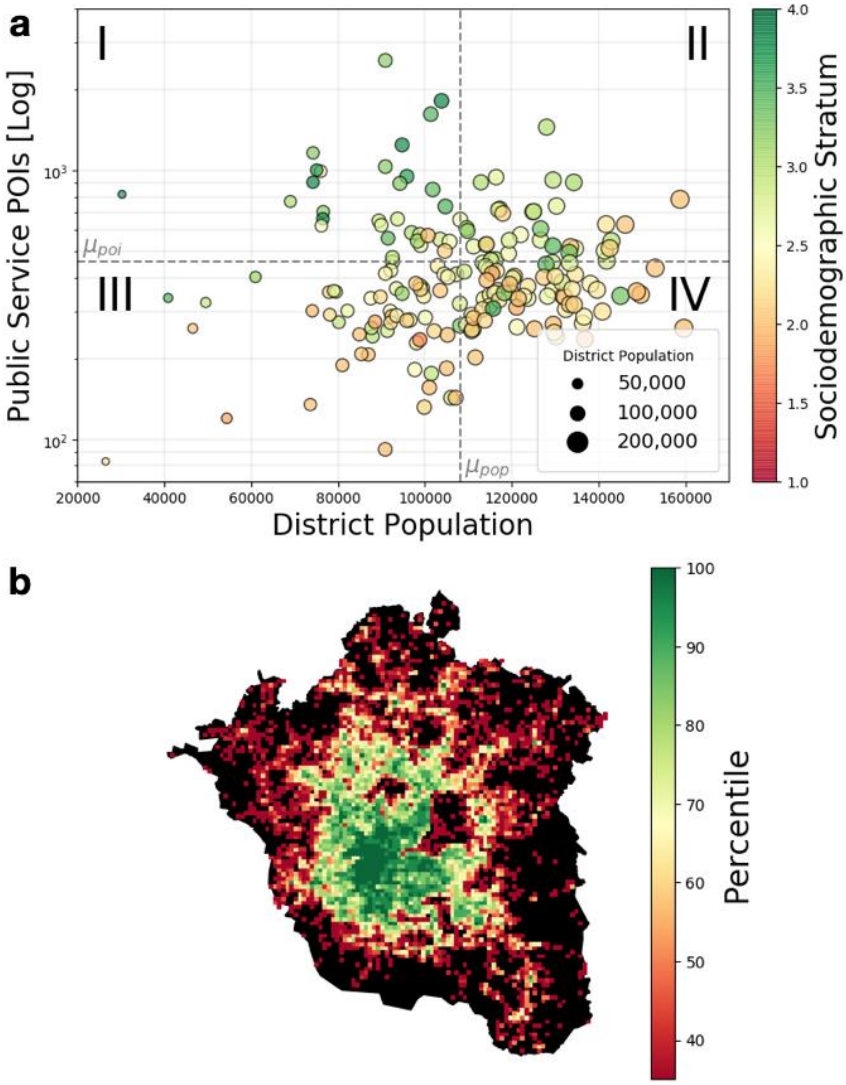

Fig. 2 a Relationship of the sociodemographic stratum of a district with the ratio of the number of public service establishments to the population, **b** Percentiles of the number of public service POIs for each 1 sq.km. block

strata are indices defined by the travel survey to characterize a respondent's social and economic condition.

In Quadrant I, where the number of public service establishments are above average and the population is below average, districts tend to enjoy the highest sociodemographic stratum. Quadrant II has districts of intermediate

sociodemographic status, while Quadrant III performs noticeably worse. 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 establish-
ments to serve more people in a small amount of space. Certainly, these are mere observations that may fuel hypotheses for further extensive research.

3.2 Geospatial Granularity of Data

Another advantage gained through the POIs is the spatial granularity of the collected data. Travel survey respondents are often organized by district of residence, whereas establishments on Google Places are pinpointed to near-exact coordinates. Since cities and districts are never truly homogeneous, this level of detail provides a more realistic perspective of city dynamics, highlighting functional interaction over arbitrary political boundaries.

In Figure 2b, the coordinates of public service establishments are truncated to two decimal places, binning them to grids that are approximately a kilometer to a side. Due to the orders of magnitude in difference between the urban core and more rural areas, the number of public service establishments are 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.

4 Extended Radiation Model for Human Mobility

4.1 Pre-processing: Trip Attraction and POIs

Counting the number of POIs per district is necessary for direct comparison with the 2017 travel survey data where the most relevant granularity is at the level of districts. Mapping these per district in Figures 3a and 3b, a direct comparison can be made with trip attraction reported in the 2017 travel survey.

While the comparison is not perfect, the distribution of points of interest makes a good approximation of 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

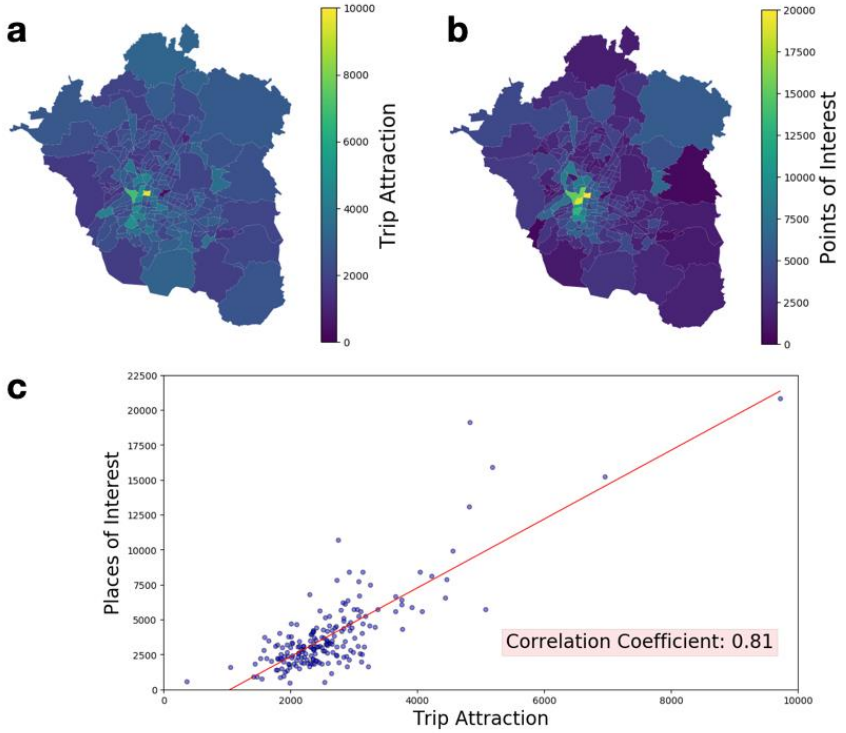

Fig. 3 a Values of trip attraction, **b** Number of points of interest, **c** Correlation plot of trip attraction and points of interest

interest in Figure 3c yields a similar observation, with the correlation coefficient of the two variables determined to be quite high at 0.81. This comparison will be of great relevance in Section 4.2, where the POIs are used to model mobility patterns in the city, in place of travel survey data.

4.2 The Model

Many models have been developed in order to predict population movement at different scales. In the context of the Greater Mexico City we want to investigate how accurate such models are and how well do they perform to reconstruct mobility patterns. Despite the widespread use of the gravitation model for population movement [5, 6, 7], we present an application of a derivation of the radiation model: the extended radiation model [8].

The radiation model [9, 10] is a stochastic process that differentiate itself by a parameter-free nature and enables, without previous mobility measurements, prediction of mobility in good agreement with mobility and transport patterns [10]. The radiation model only relies on population densities and offers a realistic approximation to the observed commuting patterns [10]. Using the natural partition of the city in district, 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/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/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

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{4.2.1}$$

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

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

$$\langle T_{ij} \rangle = \gamma T_i \frac{\left[(a_{ij} + m_j)^{\propto} - a_{ij}^{\propto}\right](n_i^{\propto} + 1)}{(a_{ij}^{\propto} + 1)\left[(a_{ij} + m_j)^{\propto} + 1\right]} \tag{4.2.2}$$

where $a_{ij} = n_i + s_{ij}$, γ is the percentage of the population commuting, and empirically we set $\propto = \left(\frac{l}{36\,[km]}\right)^{1.33}$, where l is the characteristic length of the system at hand. $\propto$ account for the fact that

The performance of the extended radiation model, meant to be applied without trip data for calibration, can be measured thanks to the Common Part of Commuters based on the Sørensen index

$$CPC(T, \tilde{T}) = \frac{2 \sum_{i=1}^{n} \sum_{j=1}^{n} \min(T_{ij}, \widetilde{T_{ij}})}{\sum_{i=1}^{n} \sum_{j=1}^{n} T_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{n} \widetilde{T_{ij}}} \tag{4.2.3}$$

It gives a quantitative measure of the goodness of the flow estimation, 0 meaning no agreement found and 1 perfect estimation.

4.3 The Experiment

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 trips coming to the district), trip production (number of trips leaving from the district), point of interest and characteristic length (given as the square root of the area of the district).

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

Four different set ups are then used to compare the performance of the model based on different approximations of the trip generation from the origin districts and the trip attraction of the destination districts: (1) we use trip attraction and trip production as a baseline, (2) we use POI as a proxy for trip attraction, (3) we use population as a proxy for trip production, and (4) we combine (2) and (3). The resulting CPC values are shown in Table 1.

Origin Destination	Trip Production Trip Attraction	Trip Production POI	Population Trip Attraction	Population POI
CPC	0.69	0.67	0.64	0.63

Table 1. Comparison of the goodness of fit depending on different input data in the model

Table 1 shows the goodness of fit of the extended radiation model is close to other recently proposed models [citation]. Moreover, we investigate the impact of different proxies for flow generation and attraction volumes as input in our model and find that the use of more easily acquired data sources such as population and POI density achieves the same level of accuracy. POIs seem really interesting because they may enable to get a very accurate model compared to the actual trip attraction of each districts which is very promising given their much easier and cheaper access. On the other hand, the use of population in place of trip attraction, aims at predicting future mobility patterns given the knowledge of γ the proportion of the total population of the system commuting, and assuming this ratio remains the same at the prediction time. Here, we extract γ on the 2017 survey and use it again on the same data. Consequently, we cannot validate the predictive power of the model but nonetheless, we distorted the population data of each district by multiplying it by γ but still observe encouraging results.

5 Analysis of Behaviors Related to Means of Transport

5.1 Introduction of the Problem and Context

This section is devoted to the analysis of mobility behavior within Mexico City. Using broad user surveys, the aim is to identify types of dominant behavior in the population: modes of transport used, geographic distribution, and socio-demographic characteristics..

To do this, we have a large database collected during a survey of Mexico City residents that includes more than half a million trips. For each trip identified, the means of transport used, the districts of departure and arrival, the time of departure and arrival, the purpose of the trip, the gender of the traveler, his age and socio-demographic category are indicated. Twenty different means of transport are thus identified among the 196 districts mentioned in this survey.

5.2 Method

Although we have many indicators other than means of transport for each trip, we will only base our groupings in terms of mobility, without associating the other metrics. The latter will then be involved in the analysis of these clusters formed. In doing so, we seek to distinguish the main transport behaviors, which will themselves be the association, in certain proportions, of several possible transport modes.

It is clear that all the means of transport mentioned in this database will not be significantly present in the main behaviors. We expect 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. We therefore do not need such a large number of variables - initially twenty - to describe our database. Given the number of variables and realizations of the database, the application of a method to reduce the size of the variables is not an absolute necessity here. However, we will apply the Principal Component Analysis (PCA) method to address the problem mentioned above. This will also allow us to reduce computation time and complexity when using clustering algorithms in a second phase of the study. Finally, projecting into a lower dimensional base provides us with a different perspective on the database and will inform our understanding of it [12, 13].

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 criterion retains here to set the size of the new projected database is that the total variance captured by the N first PCs has to be more than a threshold of

85% - a widely used value -, we therefore choose to keep only the first five PCs for the rest of the study [14].

To group trips around main behaviors, we use the K-means clustering algorithm. While the ideal number of clusters can be estimated by quantitative tools, such as the Elbow method for example, the best way remains to know the subject in question and the data available. In this case, it appears, from the results obtained subsequently and the preliminary studies carried out to determine this number, that six clusters is the most relevant. Each journey of the database can initially be represented as a vector composed of zeros and ones, depending on the transports used during it. We only consider its projection in the PCs database when applying the Means algorithm. K-means will work 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 allow patterns to be identified within the dataset. As a result, we obtain a list that reflects the membership of each trip in a particular cluster. The six average behaviors are then determined as the average, for all twenty initial variables, of all trips belonging to the same cluster.

5.3 Results

Figure 4 shows the six median behaviors that characterize mobility in Mexico City. 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 use several of them and do not relate exclusively to one of them. On the contrary, an analysis of the purposes of these trips later shows that residents tend to favor certain clusters depending on the purpose of their trip: going home, going to work, etc. In Figure 4, only the three most important components for each transport are shown. Each of these components is associated in ordinates with its proportion within the cluster. While the percentages associated on the x-axis with each of the clusters can be understood as the proportion of a way of moving among all the others.

Thus, for cluster 2, it appears that about 35% of trips within Mexico City are exclusively made by walking. Indeed, the proportion of the "walking" means of transport on the ordinate is equal to one, while that of the second most present means of transport in this cluster, Mexibus & Metrobus, has a proportion of 0.027. Thus, only about 2.7% of the trips attached to this cluster combined their walking on a section made in Mexibus or Metrobus. It can therefore be said that these trips are made almost exclusively by walking.

Figure 5 shows, for each of the six clusters, the proportion, per cluster, of each of the ten purposes of the trips considered by the survey: going home, going to work, going to school, going to do some shopping, going to a place to share a moment with friends, family or sports or relaxing, picking someone up, going to make a

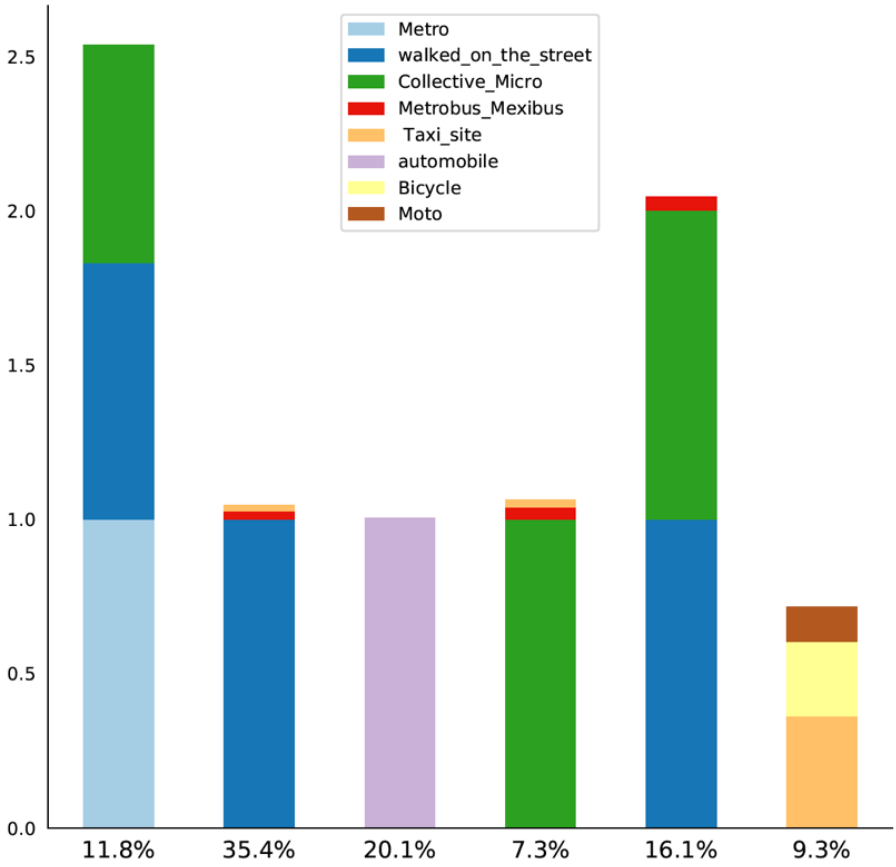

Fig. 4 Main mobility behaviors in Mexico City. We indicate the three most important main means of transport for each group. The percentage of trips contained by each group is indicated at the bottom.

procedure, going to see a doctor or receiving care, going to do a religious act or, for any other purpose not mentioned, the "other" category.

If we compare to the average of the clusters, we notice that when people move by walk (cluster 2) it is particularly to go to study, do shopping, pick up someone or for a religious act. For example, about 16% of the trips associated with the first

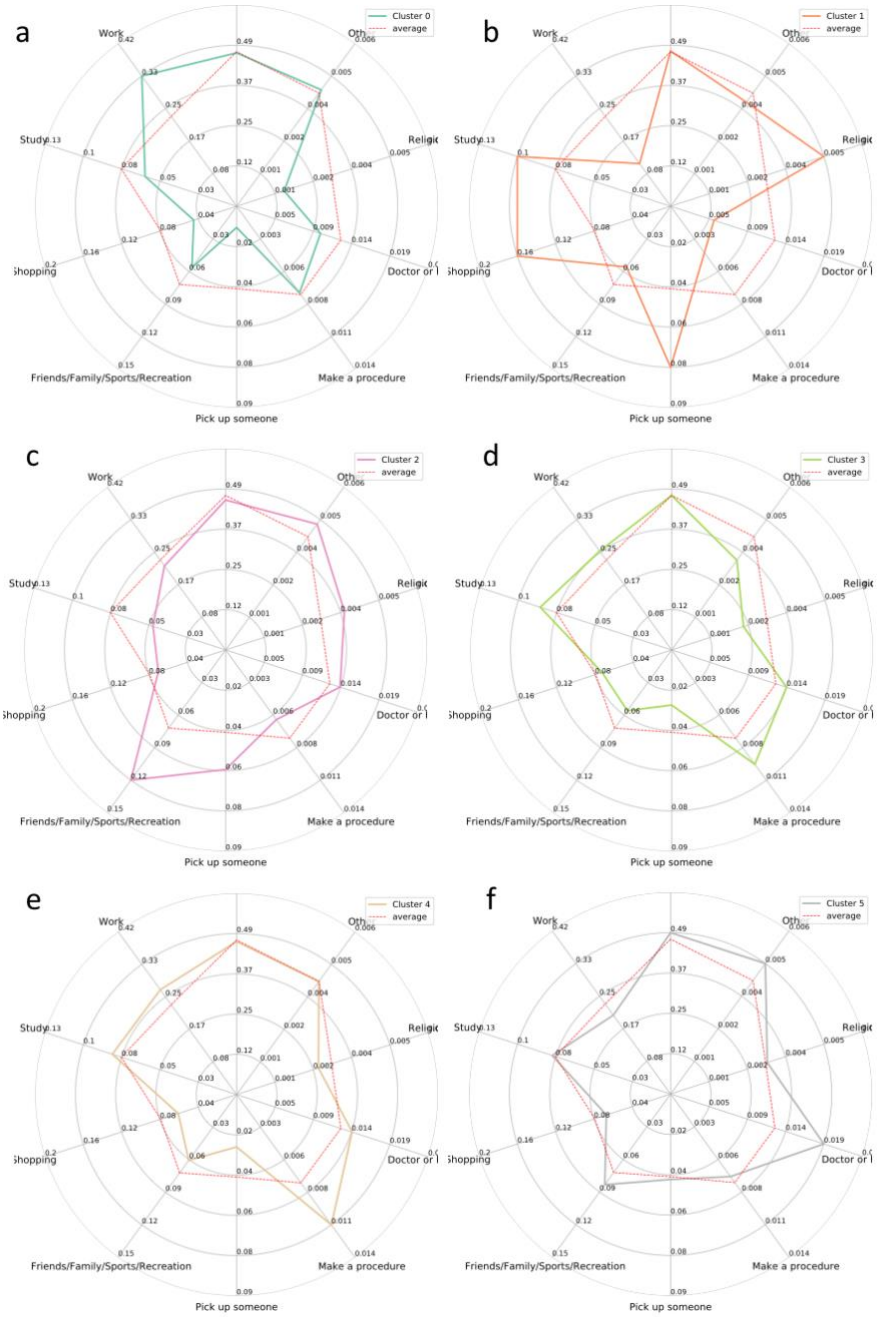

Fig. 5 Comparison of the mean of purpose trip variables within each group with respect to the median of all groups. Each letter is linked, in order, with the identified clusters of the Figure 4 (from 1 to 6 from left to right)

cluster are for shopping purposes, where the average number for all clusters is around 8% for this category, i.e. half as much. On the contrary, it seems that walking is only rarely used to get to work or to go to the doctor. Indeed, the percentages of trips within the first cluster are respectively 10% and 0.5% for an average within the clusters of 23% and 1.3%. But these figures do not reflect the fact that walking would not be used to get to work for instance. Since more than 35% of journeys are made on foot, a significant proportion of journeys to work are made on foot. But this indicates that walking, compared to other clusters, is particularly associated with some purposes. In addition, since the average travel time of this cluster is about 20 minutes where the average cluster is about twice as long, this cluster can therefore be associated with local trips. This suggests that workplaces or care centers are generally located further from family homes than shops, schools or religious places.

The same analysis can be carried out for the third cluster. This one, which represents 20% of the trips made in Mexico City, is exclusively composed of the car as a means of transport. This case, this cluster stands out for a relatively high proportion in the category "going to a place to share a moment with friends, family or to play sports or relax" compared to other clusters. By the same logic, we could underline on the contrary its withdrawal for the category "go to school". Finally, this trend suggests that study sites are probably better served by public transit, or at least more accessible than those in the category of "going to a place to share a moment with friends, family or to play sports or relax".

The fifth cluster includes the routes that exclusively combine walking and micro collective. It is also the one whose average travel time is the longest of all clusters, with an average of about one hour per trip. There are great similarities with the fourth cluster, which groups together the paths that are only made thanks to the micro collective. The proportions of the goals of the trips are similar, as are other metrics such as average age or socio-demographic status. These similarities seem logical given their similarity in terms of means of transport.

The use of walking, metro and micro collective during the same journey is characterized by the first cluster. Indeed, the metro obtains a proportion equal to 1, walking 0.83 and the micro collective 0.71. Not all the journeys in this cluster therefore systematically combine these three means of transport, but this average behaviour indicates that in the vast majority of cases these three means of transport are combined. This behaviour 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 phenomenon is probably due to the high concentration of metros and micro collectives in this part of the city, making travel much faster and more convenient by linking these modes of transport, particularly to get to work, as shown in Figure 5.

The last cluster is more complicated to interpret, because none of the twenty possible transports is fully shared by all the ways of moving that this cluster groups.

However, it should be noted that it is mainly concentrated in the agricultural regions that make up some of Mexico City's districts.

6 Conclusion

The well-defined analysis of complex human socio-technical systems has become the interest of interdisciplinary groups around the World, as urban planner seek to analytically amend current city and country infrastructure to better accommodate the continued expansion of the World's major cities and metropolises. The purpose of this study was broad, that is, the work of the researchers demonstrates initial steps that city and urban planners may take to better characterize their cities based on current data techniques drawn from various disciplines. Many of the methods used in this paper reduce the complexity of the dataset, while simultaneously extracting useful information. To this end, the exponential growth of measured and recorded data lends import to the understanding and implementation of these and other methods for use in city and urban planning.

Furthermore, techniques which extract values from the features of the dataset, and produce equations and models that determine variables of interest (e.g. travel time) can be used in the absence of data for both model calibration and prediction. As indicated in our final section, the researchers are highly interested in objective measurements for the accessibility of POIs by sociodemographic stratum, as we hope to utilize these methods as a tool to produce social equity and accessibility in the world's major cities.

Acknowledgments

We are grateful to Jorge Audiffred, Jose Luis Mateos, Emmanuel Landa, and Irving Morales of Data Lab MX for collaborating with us in collecting data and in gaining better insights to the Metropolitan Zone of the Valley of Mexico. We are also grateful to Fahad Alhasoun, who created and provided the core of the parsing algorithm to collect points of interest from the Google Places API.

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

[2] Jiang S, Fiore GA, Yang Y, Ferreira J. Jr, Frazzoli E, Gonzalez 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, 11 August 2013

[3] https://developers.google.com/places/web-service/search

[4] https://eng.uber.com/h3/

[5] Barthélemy, M. Spatial networks. Phys. Rep. 499, 1–101 (2010). 4. Erlander, S. & Stewart, N. F. The GravityModel in Transportation Analysis: Theory and Extensions (VSP, 1990).

[6] Jung, W. S., Wang, F. & Stanley, H. E. Gravity model in the Korean highway. EPL 81, 48005 (2008)

[7] Thiemann, C., Theis, F., Grady, D., Brune, R. & Brockmann, D. The structure of borders in a small world. PLoS ONE 5, e15422 (2010)
[8] Yang Y, Herrera C, Eagle N, & González M. C. (2014) Limits of predictability in commuting flows in the absence of data for calibration. Scientific Reports. doi :10.1038/srep05662
[9] Simini, F., Maritan, A.&Ne´da, Z. Human mobility in a continuum approach. PloS one 8, e60069 (2013).
[10] Simini F, González M. C., Maritan A., & Barabási A. L. (2012). A universal model for mobility and migration patterns. Nature. doi: 10.1038/nature10856
[11] Makse, H. A., Havlin, S., & Stanley, H. E. (1995). Modelling urban growth patterns. Nature. doi: 10.1038/377608a0
[12] Eagle N. and Pentland A. S. (2009). Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63:1057-1066
[13] Dorothy C. Ibes (2015) A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application. Landscape and Urban Planning 137 122-137
[14] Shlens J. A tutorial on Principal Component Analysis (December 10, 2005; Version 2)