Measuring Spatial Subdivisions in Urban Mobility with Mobile Phone Data

Eduardo Graells-Garrido
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Universidad del Desarrollo
Santiago, Chile
eduardo.graells@bsc.es

Irene Meta
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Universitat Internacional de Catalunya (UIC)
Barcelona, Spain
irene.meta@bsc.es

Feliu Serra-Burriel
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
feliu.serra@bsc.es

Patricio Reyes
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
patricio.reyes@bsc.es

Fernando M. Cucchietti
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
fernando.cucchietti@bsc.es

ABSTRACT

Urban population grows constantly. By 2050 two thirds of the world population will reside in urban areas. This growth is faster and more complex than the ability of cities to measure and plan for their sustainability. To understand what makes a city inclusive for all, we define a methodology to identify and characterize spatial subdivisions: areas with over- and under-representation of specific population groups, named *hot* and *cold* spots respectively. Using aggregated mobile phone data, we apply this methodology to the city of Barcelona to assess the mobility of three groups of people: women, elders, and tourists. We find that, within the three groups, cold spots have a lower diversity of amenities and services than hot spots. Also, cold spots of women and tourists tend to have lower population income. These insights apply to the floating population of Barcelona, thus augmenting the scope of how inclusiveness can be analyzed in the city.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Urban Mobility, Mobile Phone Data, Spatial Analysis

ACM Reference Format:
Eduardo Graells-Garrido, Irene Meta, Feliu Serra-Burriel, Patricio Reyes, and Fernando M. Cucchietti. 2020. Measuring Spatial Subdivisions in Urban Mobility with Mobile Phone Data. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3366424.3384370

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3384370

1 INTRODUCTION

As of 2020, more than half of the population live in cities [25], and by 2050, two thirds of the population will reside in urban areas [28]. In this scenario of urban growth, the United Nations have declared a goal for sustainable development that aims to make cities "inclusive, safe, resilient, and sustainable" [1]. To reach these objectives, disciplines such as urbanism, architecture, ecology, sociology, and others provide frameworks to model the functioning of cities. Typically, the main data source for analysis is household and time-use surveys as well as travel diaries. Such instruments provide rich information that represents the general population of a city, which then informs urban design and policy making.

However, the goals of improving inclusiveness and safety are limited by the purpose of traditional methods, because typical data sources tend to under-represent specific sub-populations, including women [8] and elders [20]. For example, surveys fail to measure that women trips are shorter than those of men [6], and how these trips are chained to others, partially due to household and caretaking purposes [17]. Likewise, elders also move in shorter trips, but, in contrast to younger people, their trip purposes are focused mostly on feeling independent and interacting with others in social situations [30]. While it is known that traditional methods have these shortcomings, improving them to reach finer representation is expensive and impractical on a global scale. Even though specific methods can be designed for under-represented groups, such efforts may miss the global picture, which includes the relationship between mobility of different population groups. As a consequence, there is need of fine-grained city-scale data for the design of inclusive and safe urban spaces. Finally, data biases often occur due to underlying societal biases. Biased data means incorrect population statistics which can mislead city planning and design into amplifying the problems they aim to fix [26].

Recent technological advances and the availability of non-traditional data sets have allowed to study urban phenomena at spatio-temporal granularities that traditional methods cannot. Mobile phone data, for example, allows a cost-effective way to perform

studies about urban human mobility [7], as mobile operators already generate, store, and analyze the data for billing and marketing purposes. The aggregation of digital traces from mobile phone usage was used to uncover data gaps in mobility [11], a true seminal piece of work towards using this data source for inclusive cities. Inspired by this line of work, we extend the scope to understand mobility aspects of three groups of urban visitors: women, elders, and tourists. Under the assumption that all people access the city equally, our research questions are: *How to identify places with more (or less) presence of these groups than expected?* If these places can be pointed out, *what characterizes them?* Our proposed method is a pipeline that starts on the definition of visitor metrics related to these three groups; then, we perform spatial analysis on these metrics to identify whether there is spatial concentration of visitors (or the lack thereof). If so, we identify areas with over-representation of these groups, or *hot spots*, as well as the opposite, places with under-representation, or *cold spots*. We then proceed to use up-to-date information about income, amenities, and services in the city to characterize these areas based on their economic development and urban environment.

As a case study, we analyzed the city of Barcelona, the second largest city in Spain and one of the largest in Europe. The city is known for its urban planning tradition and was often ahead of its time compared to the Spanish general developments [16]. However, there are still challenges in improving the city for everyone from a sustainable perspective: overtourism [21], gender accessibility problems [15], and one fifth of its inhabitants are elderly people [9]. In an effort to augment its understanding of the city's urban dynamics, the local government (*Ajuntament de Barcelona*) acquired anonymized and aggregated mobile phone data from the mobile phone operator Vodafone. Access to this type of data for planning is a clear opportunity to compare and advance over other data sources focused only on census and residential populations, as it could help to understand the mobility patterns of residents and non-residents alike.

We find that the studied population subgroups behave differently. Cold spots for all groups are characterized by lower population income than hot spots, as well as less diversity of amenities and services. Hot spots for all groups are characterized for being less associated with public transport than the rest of the city. Urban infrastructure such as highways and the main streets of the city play a role when interpreting the locations of these areas of over- and under-representation, as cold spots tend to be outside of the area delimited by highways around the city whereas hot spots tend to be close to relevant primary streets. As such, our work contributes: (i) a methodology to identify and characterize hot and cold spots of the floating population in a city; (ii) a case study applying the methodology to Barcelona.

We conclude the paper with a discussion focused on the implications in public policy and the usage of non-traditional data to solve the complex problems that affect cities today.

2 RELATED WORK

Mobile phone data usually refers to the set of billing records from mobile phone networks, known as Data Detail Records (XDR) [7]. Other types of non-traditional data that has been used to understand mobility include micro-blogging platforms with geo-location [19] or inferred user attributes regarding mobility [29]; check-ins from location-based services [24]; and photos from photo-sharing services [5]. In comparison to these data sets, XDR allows a fine-grained analysis, not only in spatio-temporal aspects to, for instance, observe changes in mobility in short periods of time [12], but also in demographic ones, such as measuring the social diversity of visitors in shopping malls [4].

In terms of scope, our work is similar to recent efforts to uncover gender gaps in urban mobility [11]. The differences in our approach are two-fold. On the one hand, we use a data set that is aggregated from XDR in spatial and temporal aspects. As such, it does not include individualized information, allowing us to perform analysis at the area-level but not on the individual level. On the other hand, our methods rely primarily on established spatial analysis [3, 23], which brings a different perspective to the distance-based approach employed before. Then, our work contributes a different approach to an already identified problem, with extended coverage in terms of population groups, adding elders and tourists, and focusing on a different city, Barcelona.

3 CONTEXT AND DATA SETS

More than 1.6 million people reside in Barcelona. Its 100 Km^2 area is composed of 12 districts, split into 73 neighborhoods. Natural boundaries delimit the city: the Besos river limits the city at the north-west, and the Llobregat river does so at the south-west side. The Metropolitan Area is much wider, and it is impossible to distinguish the limit between Barcelona and the surrounding municipalities. The city extends on a mild slope from the sea (south-east) up to the edge of the Collserola mountain chain (north-west). The Collserola and the Montjuic (south) have limited the city expansion because of their relatively hard accessibility, and now are important areas of leisure and biodiversity within the city [16].

The social aspects of mobility that affect subpopulations of the city [9, 15] along with rising overtourism [21] and alarming pollution levels have urged urban planners to focus on sustainability [2]. In this context, we focus on one of the qualities of sustainability, inclusiveness.

City Data. The *Ajuntament* provides open access to socio-demographic attributes at the neighborhood level at a yearly frequency (some metrics are quarterly), including income and house pricing among other things. All these variables are scaled to the mean, which allows us to compare the different neighborhood areas in relative terms. We measure income through the mean family income (RF, or *Renda Familiar*),[1] which contains mean income at the neighborhood level (see Figure 1 (a) for its spatial distribution), normalized so that the whole city mean income equals 100.

Mobile Phone Data. The data obtained from the mobile phone operator consists of the number of visitors observed during the year 2018, at periods of four hours, grouped into 212 regions or cells (see Figure 1 (b)). The number of visitors is defined as the total number of mobile phones active inside each region during each period. *Active* means that the phone was initiated or received some activity (call, browse, text, etc) other than passive connections to the

[1] https://opendata-ajuntament.barcelona.cat/data/es/dataset/est-renda-familiar

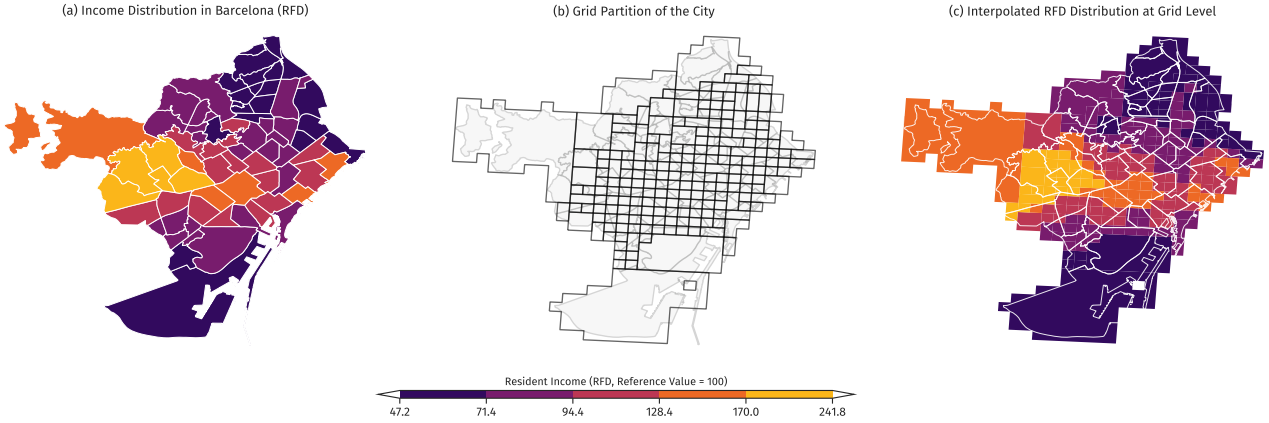

Figure 1: a) Map of neighborhoods with income data; b) Overlap of the grid areas with the neighborhoods; c) Spatial join of neighborhood income data with the grid.

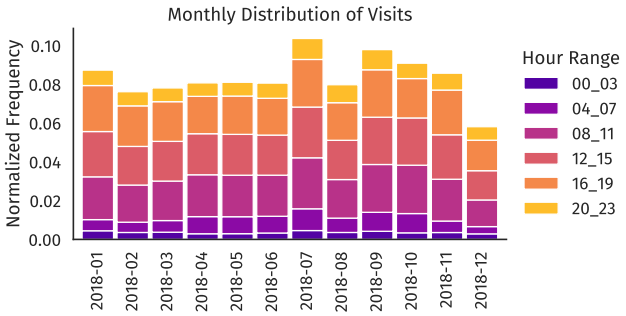

Figure 2: Normalized frequency of the sum of the monthly number of visitors captured in the dataset, provided by Vodafone to the Barcelona city hall. The colors stratify the sample into time-frames.

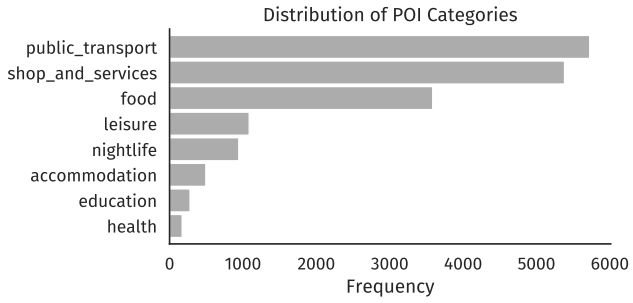

Figure 3: POI count per category. There are three main categories that represent more than 80% of all the POI.

network. This may introduce bias in the data, as people do not call while driving or at night, or they connect to their home or job WiFi falsely indicating less presence. In addition to the total number of visitors, the operator also provides the number of visitors according to specific demographic characteristics, including gender (binary, female and male), age cohorts, and tourists (national and foreign). The determination of these characteristics and its aggregation into a number of visitors was made directly by the mobile operator using activity criteria as well as billing and other information when available. In addition, cells with less than a given number of observations during each period were discarded from the data and cells that consistently exhibited few observations (below 500 visitors) were consolidated into grouped regions.

We estimated the total number of visits accounted per month and per hour range (see Figure 2, normalized to avoid revealing commercially sensitive data). The number of total visitors per month lies within the same order of magnitude, with fluctuations that could be explained by changes in the market share of the operator and seasonal factors such as tourism in July.

We estimated a mean income for each cell, defined as the weighted interpolation of the incomes in all areas that intersect with that cell (see Figure 1 (c)).

Open Street Map. To include aspects of the built environment and accessibility to amenities and services, we use data from OpenStreetMap (OSM).[2] OSM provides spatial and geographical data contributed freely and voluntarily by its members, and it has been identified as an accurate source of urban information [13].

From OSM we obtain Points of Interest (POIs) that allow us to understand part of the urban environment in our analysis. Points of interest (POIs) are geolocated attractors, such as shops, food spots, and tourist sights. We categorized most of POIs in the city as shown on Figure 3. These include (sorted by descending frequency): *public transport* (e.g., bus stops, metro stations), *shops and services* (e.g., convenience shops, government offices, professional services), *food* (e.g., cafés, restaurants), *leisure* (e.g., natural attractions, parks, stadiums), *nightlife* (e.g., bars), *accommodation* (e.g., hotels), *education* (e.g., universities, schools), and *health* (e.g., hospitals).

By analyzing the integration of these data sets, we aim to identify areas of the city with over- and under-representation of women,

[2]http://openstreetmap.org

elders, and tourists, and characterize these areas according to their economic development and urban environment.

4 METHODS

In this section we describe how we measure spatial subdivisions in urban mobility. First, we define the metrics we evaluate. These metrics are local, in the sense that they cover a specific point or area and do not consider the context or their surroundings. Second, we perform a spatial analysis using established techniques, allowing us to take into account the spatial context and evaluate both local and global patterns to identify significant areas of over- and under-representation. And third, we define how to characterize the significant areas with respect to economic development and urban environment.

4.1 Cell Level Metrics

In our context, the city is partitioned by a grid which comprises neighboring cells. Cells may have edges and/or vertices in common but they do not overlap. Here we define cell-level metrics regarding the presence of women, elders, and tourists in them. The three metrics we define are: the women ratio G, the elder ratio E, and the tourist ratio T.

Women Ratio G. This metric captures the ratio of female visitors in an area i during the whole period under study. We first define the women ratio G' as:

$$G'_i = \frac{\text{\# women visiting area i}}{\text{\# total visitors in area i}}.$$

Note that it is likely that the mobile operator has a non-representative sample of the population. For instance, the sample ratio at the city level may not be 1, even though the population ratio may be close to it. To counter this effect we define a standardized version of the women ratio as

$$G_i = \frac{G'_i - \bar{G}'}{s(G')},$$

where s is the sample standard deviation function and $\bar{G}'$ corresponds to the mean of G'_i for all the cells. In this way, if $G_i = 0$, the area i has a women ratio equivalent to the average of the city. Positive and negative values of G indicate how many standard deviations the ratio deviates from the sample mean. Notice that, to focus on the floating population, we consider visitors between 8am and midnight.

Elder Ratio E. In a similar way to G', the elder ratio before standardization is defined as:

$$E'_i = \frac{\text{\# elder visitors in area i}}{\text{\# total visitors in area i}}.$$

We choose the threshold age to be considered elder as 65 years old or more, as defined by the *Ajuntament*.[3] Analogous to G, the metric E is the standardized version of E'.

[3] https://ajuntament.barcelona.cat/personesgrans/es/canal/la-gent-gran-de-barcelona

Tourist Ratio T. Our last metric is similar to the previous ones, as it represents the proportion of tourists in an area:

$$T'_i = \frac{\text{\# tourists (both foreign and national) in area i}}{\text{\# total visitors in area i}}.$$

Analogous to G and E, the metric T is the standardized version of T'.

4.2 Spatial Patterns

Our aim is to find places where each population group of interest is over- or under-represented according to its floating population patterns, expressed in the metrics G, E and T. To do so, we evaluate whether values of these metrics tend to concentrate in geographical terms, *i.e.*, if nearby areas have similar values. The Moran's I coefficient of spatial autocorrelation [23] measures this concentration. It is defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

where N is the number of spatial units (in our case, grid cells) under analysis, x_i is one of $\{G_i, E_i, T_i\}$, w_{ij} encodes the spatial weight of cell j into cell i, and W is the sum of all spatial weights. Note that w_{ij} is a matrix where $w_{ii} = 0$. The value of w_{ij} is a normalized version of the following schema:

$$w'_{ij} = \begin{cases} 1 & \text{if area i and j are contiguous.} \\ 0 & \text{otherwise.} \end{cases}$$

Here, contiguity between cells is defined as sharing an edge or sharing a vertex. This is coherent when using grids composed of square cells, as it is possible to move from one square to another through a corner. Then, w_{ij} is normalized in the following way:

$$w_{ij} = \frac{w'_{ij}}{\sum_j w'_{ij}}.$$

With these definitions, $I = -1$ when the variable under analysis is perfectly dispersed in space, $I = 1$ when it is completely clustered, and $I = 0$ when values are randomly arranged.

Next, for each metric, if indeed there is spatial autocorrelation, we proceed to estimate Local Moran's I, a coefficient that allows us to identify groups of areas that have high (or low) values that are surrounded by other areas with high (or low) values [3]. It is defined as follows:

$$I_i = \frac{x_i - \bar{x}}{s(x_i)^2} \sum_{j=1, j\neq i}^{n} w_{ij}(x_j - \bar{x}),$$

where $s(x_i)$ is the standard deviation of values of x of contiguous areas to area i. Note that, in global and local I, significance is estimated through permutation tests. Areas with significant high values of local I are known as *hot spots*, and areas with significant low values are known as *cold spots*. The other areas present neutral or average behavior.

4.3 Characterizing Hot and Cold Spots

After identifying areas of interest, our purpose is to characterize each type of spot. With this aim, we analyze the income of each spot and the availability of services and activities through Points of Interest (POIs).

Income. We estimate a mean income for all spatial units. Under the null hypothesis that income is independent of spatial subdivisions, one would expect that population income in hot spots has the same distribution as population income in cold spots. We test this hypothesis by comparing the income in all hot spots with the income for all cold spots using the two-sample Kolmogorov-Smirnov (KS) test. This non-parametric test evaluates whether two underlying one-dimensional probability distributions that generated those samples differ. If the result of a test is significant, it means that there is evidence to reject the null hypothesis of same income for both types of area for a given visitor metric.

Association and Diversity of Points of Interest. Next, we estimate how each category of POIs is associated with each area. This problem is analogous to document categorization in Information Retrieval where there are frequent words in many (if not all) documents that do not necessarily characterize them. In our context, areas are analogous to documents, and POIs are analogous to words. For instance, bus stops may be available in all areas of the city, but some areas may have more bus stops than others, while having less POIs of other categories. In that case, these latter areas have a stronger association with bus stops than other areas. Given that most areas have many kinds of POIs, we need to use a weighting schema that controls for frequency and variability. While a common technique to do so is Term Frequency–Inverse Document Frequency (TF-IDF), we resort to a technique that does not overweight elements with low frequencies. This method is known as Log-Odds ratio with Uninformative Dirichlet Prior [22]. It defines the weight of a word through the following point estimate:

$$\widehat{\delta}_{kw}^{(i)} = \log\left[\frac{(y_{kw}^{(i)}+\alpha_{kw}^{(i)})}{(n_k^{(i)}+\alpha_{k0}^{(i)}-y_{kw}^{(i)}-\alpha_{kw}^{(i)})}\right] - \log\left[\frac{(y_{kw}+\alpha_{kw})}{(n_k+\alpha_{k0}-y_{kw}-\alpha_{kw})}\right],$$

where kw is the frequency of the POI type at cell i, and α is the prior distribution. Positive values of this metric indicate positive association, while negative values indicate disassociation. Thus, we would expect larger amounts of specific kinds of categories in specific regions and some other categories that are rare in other regions. Values close to 0 indicate independence between the POI type and the cell.

Ref. [11] found relationships between accessibility gaps with the diversity of places, and between economic development and the diversity of social connections in places [10]. To explore this potential relationship, for each cell i we estimate its POI entropy, defined as the Shannon Information Entropy H:

$$H_i = -\sum_c p_c \log p_c,$$

where c is a POI category, a p_c is the fraction of POIs from category c within cell i.

By following this methodology, it is possible to identify spatial subdivisions in urban mobility according to who visits each area of the city, particularly women (G), elders (E), and tourists (T). The spatial subdivisions are defined as those areas identified as hot/cold spots of visits from these groups, which then can be characterized according to their economic development and urban environment.

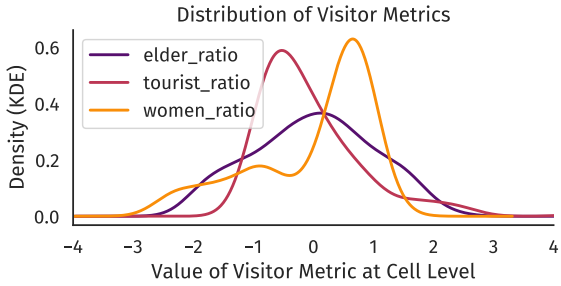

Figure 4: Probability density functions of cell-level metrics estimated with Kernel Density Estimation.

5 RESULTS

In this section we describe the results of applying our proposed methods to the data sets, with the aim of understanding spatial subdivisions in Barcelona, as seen from mobility data.

Cell-level Metrics. We estimated the women ratio G, elder ratio E, and tourist ratio T for all cells in the grid of Barcelona. Of all cells, 195 have the same size. However, a few of them have bigger sizes, because they were merged by the mobile phone operator to ensure privacy. We considered the number of regular cells as a scaling factor (i.e., the most common) that would fit in a merged cell. Thus, we divided the value of each cell according to its scaling factor.

The distributions of each observed metric (Figure 4) have different shapes. The elder ratio distribution is unimodal with fat tails on both sides. The tourist ratio distribution is positively skewed, having a negative mode. Conversely, women ratio distribution is negatively skewed with positive mode. It has a group on the negative values but the majority of the values are positive.

Global Spatial Autocorrelation. In spatial terms, the three metrics present spatial autocorrelation ($I_G = 0.25$, $I_E = 0.34$, $I_T = 0.39$, all significant with $p \leq 0.001$). Women and elders cover most of the densely populated areas of the city (see Figure 5 (a) and (b), respectively). However, the extent of elder concentration is smaller, thus having a greater autocorrelation than women. Tourists being the most concentrated group makes sense given the touristic attractiveness of the city, which tends to be concentrated in the historical districts, with few spots in other places such as the beaches, the highway that connects the airport to the city, and Barcelona's soccer team stadium (see Figure 5 (c)). Note, however, that all these concentrations are smaller than the concentration of income ($I_{RFD} = 0.83$, see Figure 1 (c)).

Local Spatial Autocorrelation. Next, we estimated the Local Moran's I coefficient of each area of the city for G, E, and T. Figure 6 shows the spatial location of all relevant areas, with hot spots of each metric in the top row, and cold spots of each metric in the bottom row. Color indicates the income level of each cell.

Both G and E hot spots are located mostly above the Diagonal Avenue (the avenue the goes from west to east), and they overlap in three different sectors. The first point corresponds to the Sants area (west). The second sector is north-west of the city, right above the Diagonal Avenue. It corresponds to a middle and high income

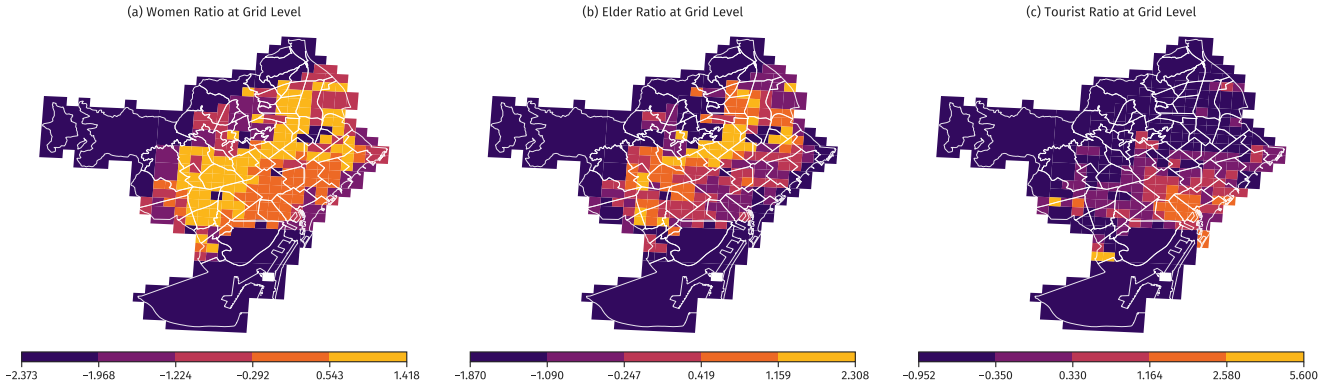

Figure 5: Maps of the spatial distribution for metrics G, E and T. Color of cell i corresponds to the value of G_i, E_i and T_i respectively. Notice that the scales are different for each one of the metrics.

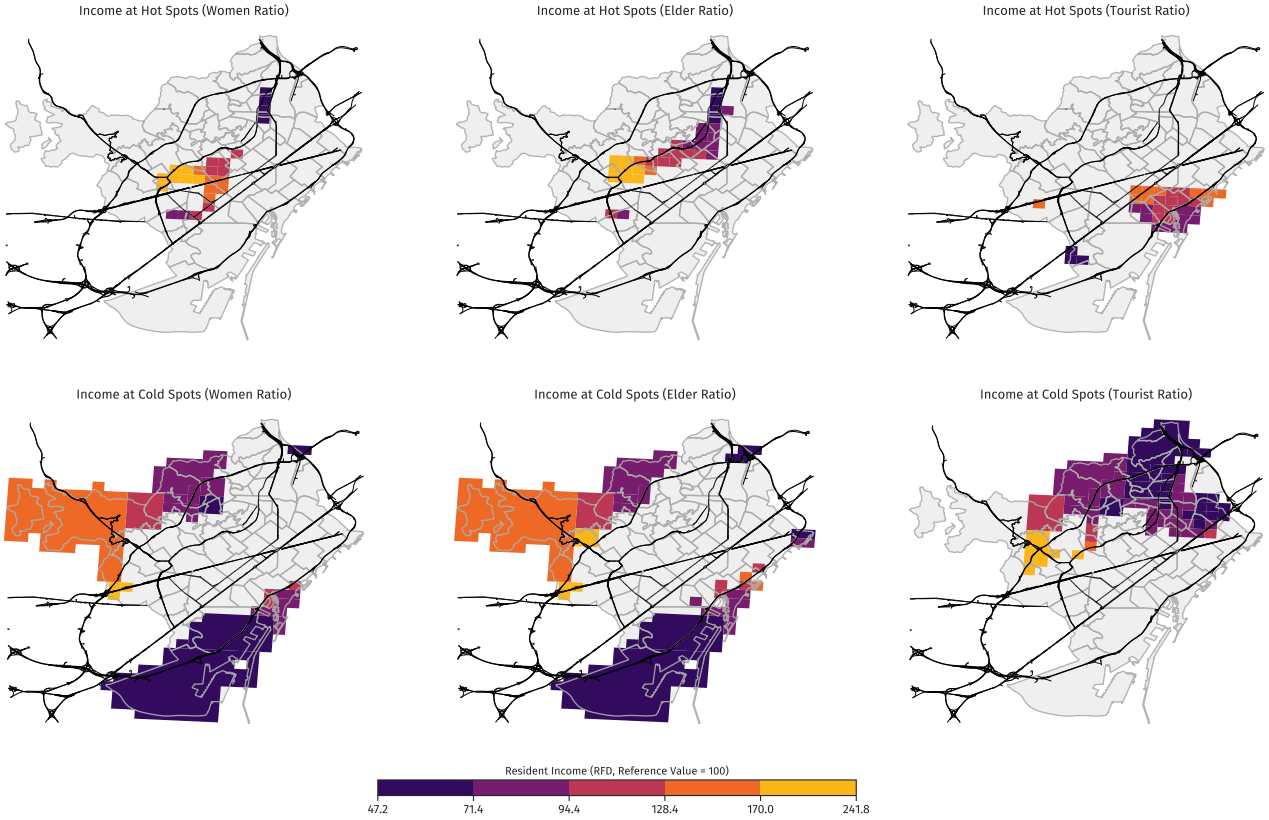

Figure 6: Hot/cold spots for each metric according to Local Moran's I metric. The first column is the gender ratio, the second column the elder ratio and the third column the tourist ratio. The top row contains hot spots, while the bottom row contains cold spots. The color represents the income of that area as in Figure 1.

area of the city, while the third sector, to the north of the city, corresponds to a low income area. It is hard to interpret this overlapping: we know the percentage of female population is larger with age [9], but also that women are the ones who dedicate more time to care-taking activities [9]. The E hot spot shows a unique cluster, mainly just below the middle beltway. Within this hot spot, we observe heterogeneity on the socio-economic status, having the south-western part a much larger income ratio. The city center does not show under or over-representation of women and elders, as both hot and cold spots are absent there. Conversely, the area in its

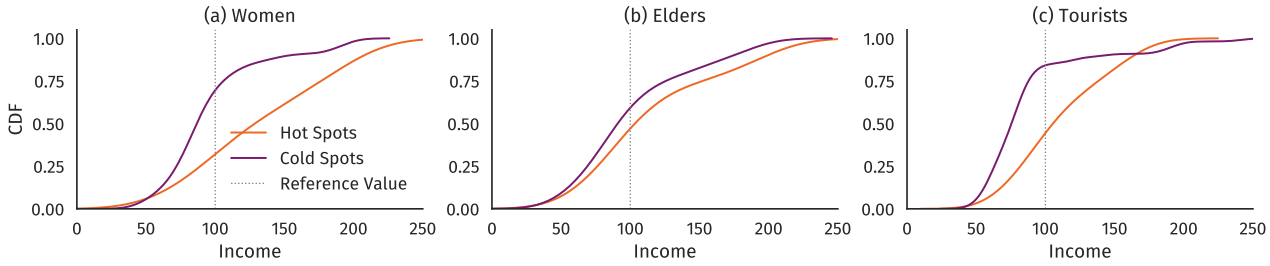

Figure 7: CDFs of the income on hot and cold spots for each of the metrics.

totality is signalled as a hot spot for tourists (T), encompassing the Old district and part of the adjacent Eixample, below the Diagonal Avenue. This observation seems reasonable given the density of historical sites and leisure spots in the area. Two more isolated areas show a hot spot of tourist activity. The one on the west is nearby the Barcelona Football Club Stadium. Particularly, it is not the cell that contains the stadium, although it contains one of the metro stations that is closer to it. At the same time, it contains the southern campus of Universitat Politecnica de Catalunya, which may also receive foreigners regularly. The other area is located on the south and contains the *Fira Gran Via Barcelona*, the biggest venue of the city for international congresses and expositions (including the Mobile World Congress). This is expected given that the mobile operator may use roaming connections to identify foreign tourists.

The cold spots are mostly spread around the periphery of the city and have different levels of income. There are three G cold spots. The smallest area is characterized by an infrastructural node that links motorways. The largest area covers mixed income but also low population density, as it is located in the periphery of the city, near the mountains. The third area, on the south, covers a leisure sector (the Montjuic hill), a working sector (the Port), and a densely populated neighborhood, La Barceloneta, which holds many touristic attractions, including restaurants and the beach. The E cold spots are similar to those of G. The western T cold spot corresponds to the richest areas, and to the toll access tunnels from the valley behind the Collserola chain mountains. In the north side, T cold spots comprehend the Sagrera high-speed train station and one of the main accesses to the city for road transport, with the infrastructural node between three motorways and two beltways. It is an area of low income and less leisure amenities than the rest of the city. Other areas of the city, such as the southern, are also characterized by similar income levels but they are not tourism cold spots.

Income Characterization. There is a variety of income levels in hot and cold spots. Women and tourist cold spots have different income distributions than their corresponding hot spots, according to a Kolmogorov-Smirnov (KS) test ($p_G = 0.004$, and $p_T < 0.001$, Bonferroni-corrected). Elders do not show that difference ($p = 0.798$, Bonferroni-corrected). To explore these differences visually, Figure 7 shows the cumulative distributions of income for hot/cold spots of each metric, estimated with Kernel Density Estimation (KDE). Hot spots tend to be shifted towards the larger income areas and cold spots appear to be on the low income areas.

Table 1: Kolmogorov-Smirnov tests (significance level $p < 0.05$, values show have been Bonferroni-corrected) for each of the metrics and significant POI categories on each studied subdivisions. Visitor metrics are women ratio (G), elder ratio (E), and tourist ratio (T).

Visitor Metric	POI Category	KS	# Cells (Hot)	# Cells (Cold)	p
G	Accommodation	0.703	26	22	< 0.001
G	Education	0.633	26	22	0.001
G	Food	0.587	26	22	0.006
G	Shops & Services	0.787	26	22	< 0.001
G	Nightlife	0.549	26	22	0.019
G	Leisure	0.580	26	22	0.007
G	Public Transport	0.657	26	22	0.001
E	Education	0.471	32	27	0.043
E	Shops & Services	0.678	32	27	< 0.001
E	Public Transport	0.524	32	27	0.007
T	Accommodation	0.500	27	55	0.003
T	Food	0.707	27	55	< 0.001
T	Leisure	0.502	27	55	0.003
T	Public Transport	0.593	27	55	< 0.001

POI Characterization. The distribution of POIs in the city exhibits different functional regions based on the activities and services available (see Figure 8). Categories such as accommodation, food, and nightlife are more concentrated than the others, while health, shops and services and education are more scattered, indicating that most of the city has access to a diversity of amenities and services.

To evaluate differences in POI (or *amenities*) association between hot and cold spots, we performed pairwise KS tests for each POI category and each metric (see Table 1). Then, we built swarm plots of each area type per metric, per POI category (see Figure 9). Every dot is a cell in a hot/cold spot of the associated variable, and its color represents its tendency (either hot or cold). Its y-position represents the corresponding POI association, while its x-position is only for legibility. Women (G) have the largest number of POI categories with significant differences between hot and cold spots association. Only the health category has the same distribution for hot and cold spots. Elders (E) present differences in the distribution of POI association for education, shops and services, and public transport. They present similar distributions to Tourists (T), where hot spots tend to be positively associated with amenities, except for the Public Transport category that presents some negative associations, similar to G and E. The hot spot association to amenities

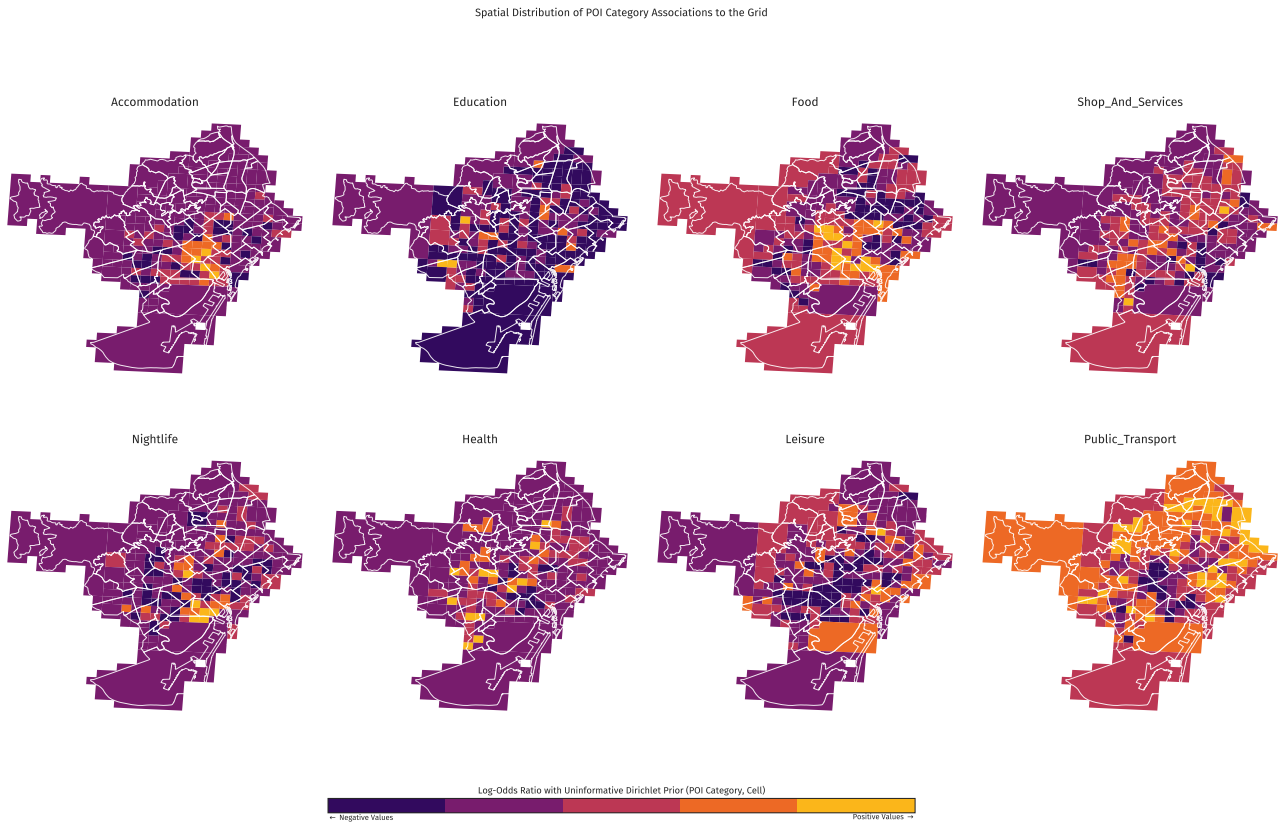

Figure 8: Log-odds Ratio with Uninformative Dirichlet Prior for each category of POI. Map colors according to the log-odds ratio value within each category.

may be related to gender or age based mobility behaviours, regarding trip chaining and trip purposes; however, we lack a clear understanding of the disassociation to public transport, which is arguably unexpected. Tourists (T) present differences between hot and cold spots association on accommodation, food, leisure, and public transport. The first three categories describe tourist attractors, as the hot spots are positively associated with these amenities. The public transport negative association to hot spots, similar to G and E, may be explained due to the historic district being comprised mostly by pedestrian streets. There are other associations that can be discussed, but we omit them due to space reasons.

Finally, regarding the diversity of POIs measured through entropy, cold spot areas are more associated with low diversity of POIs. The KS test was significant for the three pairwise comparisons ($p_G = 0.030$, $p_E = 0.005$, and $p_T < 0.001$, Bonferroni-corrected). The differences are illustrated through the KDE-based cumulative density functions on Figure 10.

In this section we explored how three population groups (women, elders, and tourists) were present in several areas of Barcelona in the year 2018. We observed that, indeed, there are areas of the city that tend to be visited by these groups (hot spots), as well as areas that tend to have an under-representation of them (cold spots), effectively creating subdivisions of over and under-representation in the city. Income and the availability and diversity of POIs play a role in characterizing these relevant areas. The most salient characterizations are two. On the one hand, cold spots of activity for women and tourist visitors are associated with less population income. On the other hand, hot spots of the three types of visitors are associated with less public transportation. Cold spots of all types are associated with a lesser diversity of POIs. We discuss further implications of these results in the next section.

6 DISCUSSION AND CONCLUSIONS

A city is experienced uniquely by each individual, although people with shared characteristics may experience it in similar ways. Urban disciplines have been studying these experiences for decades with the goal of improving quality of life in cities through urban planning and design. In this paper we have shown that aggregated mobile phone data allows us to identify relevant areas in terms of over- and under-representation of subpopulations such as women, elders, and tourists. Being a cost-effective source of data, our proposal brings knowledge of which places are relevant in terms of presence (or absence) of people from these groups as well as what characterizes these places in terms of the urban environment. Then, our methodology provides knowledge about under-represented groups in urban and policy design.

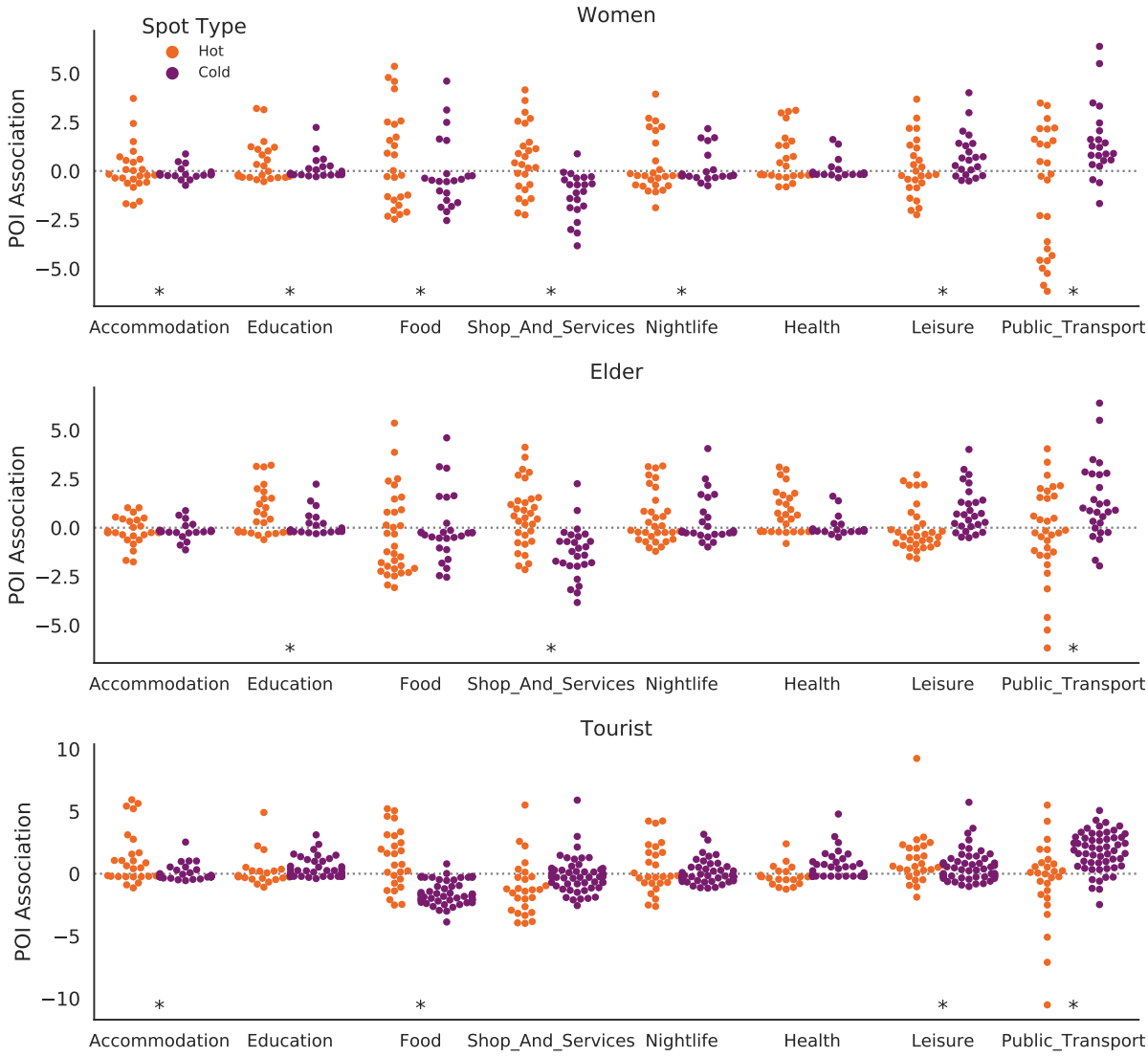

Figure 9: Swarm Plots of the POI association for each category of POI discerning hot and cold spots, for each kind of ratio. Plots marked with a star (*) indicate significant differences (according to K–S tests from Table 1) in POI association between Hot and Cold spots for the corresponding metric.

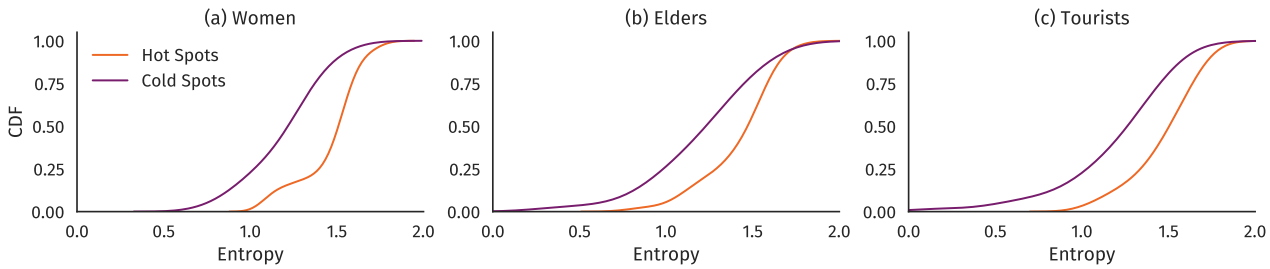

Figure 10: CDFs of the entropy for hot and cold spots on each of the ratios.

We have shown that the places visited by specific groups are related to income and the presence and diversity of amenities and services. By using mobile phone data, we were able to present these insights for the floating population of Barcelona, in contrast to and thus complementing traditional data sources that focus on the resident population only.

Our work has two main limitations. First, the analysis is bound to the market share of the mobile phone operator, which is likely to be biased toward specific socio-economic and demographic groups. Given that the data is aggregated and anonymized, we cannot control for this fact. This motivated the usage of standardized metrics to entice a clearer interpretation of our results. Second, there are intersections between the groups we analyzed, namely elderly female tourists. Hence, our analysis on income and POIs raises questions while providing preliminary answers which need further, deeper exploration, perhaps with more granular data.

In addition to improving the factors that limit the scope of this work, we devise three main lines of future work: the integration of additional area-level data sets, the definition of time-aware metrics, and multi-city analysis. Including data about crime or health would improve the characterization of hot/cold spots.

This aspect makes a time-aware analysis relevant, which would allow to measure the effect of urban interventions and seasonality according to our metrics. Finally, the issues studied here are not exclusive to one city only. In order to advance on the path to inclusive, safe, resilient and sustainable cities, quantitative methods are required to compare cities within and between them, as well as fine-grained data sets to which apply these methods to. This would allow us to distinguish between systematic subdivisions and those specific to a city.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857191 (IoTwins project). E. Graells-Garrido was partially funded by CONICYT Fondecyt de Iniciación project #11180913. We acknowledge the following libraries used in the analysis: *matplotlib* [14], *seaborn*, *PySAL* [27], *pandas* [18], and *geopandas*. Part of the map data used in this work is copyrighted by OSM contributors. Thanks to Leo Ferres for insightful discussion, and to Xavier Paradis for his help in proofreading. Finally, we thank the people from *Ajuntament de Barcelona* and Vodafone for providing access to the data and for useful discussions.

REFERENCES

[1] [n.d.]. About the Sustainable Development Goals. https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Accessed: 2020-01-20.
[2] [n.d.]. Urban Mobility Plan of Barcelona. https://www.barcelona.cat/mobilitat/es/actualidad-y-recursos/nuevo-plan-de-movilidad-urbana-2019-2024. Accessed: 2020-01-20.
[3] Luc Anselin. 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis* 27, 2 (1995), 93–115.
[4] Mariano G Beiró, Loreto Bravo, Diego Caro, Ciro Cattuto, Leo Ferres, and Eduardo Graells-Garrido. 2018. Shopping mall attraction and social mixing at a city scale. *EPJ Data Science* 7, 1 (2018), 28.
[5] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. 2016. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science* 5, 1 (2016), 30.
[6] Orna Blumen. 1994. Gender differences in the journey to work. *Urban Geography* 15, 3 (1994), 223–245.
[7] Francesco Calabrese, Laura Ferrari, and Vincent D Blondel. 2014. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Computing Surveys (CSUR)* 47, 2 (2014), 1–20.
[8] Sylvia Chant. 2013. Cities through a "gender lens": a golden "urban age" for women in the global South? *Environment and Urbanization* 25, 1 (2013), 9–29.
[9] Consorci Sanitari de Barcelona. 2019. La Salut a Barcelona 2018. (2019).
[10] Nathan Eagle, Michael Macy, and Rob Claxton. 2010. Network Diversity and Economic Development. *Science* 328, 5981 (2010), 1029–1031.
[11] Laetitia Gauvin, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. 2019. Gender gaps in urban mobility. *arXiv preprint arXiv:1906.09092* (2019).
[12] Eduardo Graells-Garrido, Leo Ferres, Diego Caro, and Loreto Bravo. 2017. The effect of Pokémon Go on the pulse of the city: a natural experiment. *EPJ Data Science* 6, 1 (2017), 23.
[13] Mordechai Haklay. 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design* 37, 4 (2010), 682–703.
[14] John D Hunter. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9, 3 (2007), 90.
[15] Regidoria de Feminismes i LGTBI de l'Ajuntament de Barcelona. 2016. Plan para la Justicia de Género 2016-2020. (2016). http://hdl.handle.net/11703/98743
[16] Tim Marshall. 2004. *Transforming Barcelona: The Renewal of a European Metropolis*. Routledge.
[17] Nancy McGuckin and Elaine Murakami. 1999. Examining Trip-Chaining Behavior: Comparison of Travel by Men and Women. *Transportation Research Record* 1693, 1 (1999), 79–85.
[18] Wes McKinney et al. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing* 14, 9 (2011).
[19] Graham McNeill, Jonathan Bright, and Scott A Hale. 2017. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science* 6, 1 (2017), 24.
[20] David H Metz. 2000. Mobility of older people and their quality of life. *Transport Policy* 7, 2 (2000), 149–152.
[21] Claudio Milano, Marina Novelli, and Joseph M Cheer. 2019. Overtourism and degrowth: a social movements perspective. *Journal of Sustainable Tourism* 27, 12 (2019), 1857–1875.
[22] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16, 4 (2008), 372–403.
[23] Patrick AP Moran. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10, 2 (1948), 243–251.
[24] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. 2012. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE* 7, 5 (2012), e37027.
[25] Population Division of the United Nations. Department of Economic and Social Affairs (UN DESA). 2018. The 2018 Revision of the World Urbanization Prospects. (2018).
[26] Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men.* Random House.
[27] Sergio J. Rey and Luc Anselin. 2010. PySAL: A Python Library of Spatial Analytical Methods. *Handbook of Applied Spatial Analysis* (2010), 175–193.
[28] Hannah Ritchie and Max Roser. 2018. Urbanization. *Our World in Data* (2018).
[29] Paula Vasquez-Henriquez, Eduardo Graells-Garrido, and Diego Caro. 2019. Characterizing Transport Perception using Social Media: Differences in Mode and Gender. In *Proceedings of the 10th ACM Conference on Web Science*. 295–299.
[30] Friederike Ziegler and Tim Schwanen. 2011. 'I like to go out to be energised by different people': an exploratory analysis of mobility and wellbeing in later life. *Ageing & Society* 31, 5 (2011), 758–781.