The WWW (and an H) of Mobile Application Usage in the City

The What, Where, When, and How

Eduardo Graells-Garrido
Data Science Institute
Universidad del Desarrollo
Santiago, Chile

Diego Caro
Data Science Institute
Universidad del Desarrollo
Santiago, Chile

Omar Miranda
Dept. of Computer Science
Universidad de Chile
Santiago, Chile

Rossano Schifanella
University of Turin
Turin, Italy

Oscar F. Peredo
Telefonica R&D
Santiago, Chile

ABSTRACT

People fulfill their informational needs through smartphones, however, little is known regarding how the urban fabric and the activities that take place in it affect the usage of mobile applications. In this regard, starting from an anonymized dataset of *Deep Packet Inspection (DPI)* data from the largest telecommunications operator in Chile, we focus on the following questions: *What* are the most popular applications used in the city? *Where* are they spatially clustered? *When* does an application is more frequently used? And *How* does the urban context and the mobility patterns relate to application usage? As a result, we observed that specific applications present high spatial clustering, while the most popular services are geographically dispersed throughout the entire city. Clusters appear in places of high floating population; however, hotspots vary in space depending on the application. Interestingly, we found that commuting plays an important role, both in terms of rush hours and transportation infrastructure. We present a discussion on these results, focusing on how the physical space and the daily commuting routine affect the pattern of data consumption and represent an important aspect in mobile users behavioral studies.

CCS CONCEPTS

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

KEYWORDS

Deep Packet Inspection, Spatial Analysis, Urban Informatics

ACM Reference Format:
Eduardo Graells-Garrido, Diego Caro, Omar Miranda, Rossano Schifanella, and Oscar F. Peredo. 2018. The WWW (and an H) of Mobile Application Usage in the City: The What, Where, When, and How. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3184558.3191561

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191561

1 INTRODUCTION

In the last decade, the adoption of user-generated content and interaction logs has been prominent when trying to understand the relationship between users and places. The availability of geolocated digital traces has enabled areas such as urban informatics and computational geography. While the aforementioned areas rely on geolocated data, they are generally not focused on a critical aspect of humans: they are dynamic multi-tasking agents whose needs constantly evolve during a day. This means, for example, that mobility patterns and informational needs are coupled with what they are doing and where. The city is experienced in sequence, and these sequences are unique for each person [18]. As such, there is the need to understand whether there is a relationship between activities performed and the urban context surrounding a user, *i.e.*, which applications a user is running, and the physical context where these activities take place in a city, such as the metro while commuting, the workplace, a recreational area, or the private home.

There is work done in this area that involves different approaches: mobile phones with custom logging applications; in-situ studies, and server logs analyses. However, they tend not to focus on the influence that the daily urban life has on the digital sphere. In this paper, we present a descriptive analysis of a large scale longitudinal dataset of mobile applications usage in Santiago, Chile, from the largest telecommunications operator in the country. The digital traces, anonymized, filtered and aggregated at cellular tower level to protect customer privacy, allow us to study spatial and temporal patterns according to different facets: the (i) *what*, or which applications are most accessed; (ii) *where*, or the places that are more associated with an app; (iii) *when*, or the times of the day where an app is more prevalent; and (iv) *how*, which we interpret as the physical context in which a mobile service is mainly used. In particular, we focus our analysis on the effect of commuting and the mode of transportation. In fact, commuting changes the surrounding and attentional context of users and thus plays an important role in modeling users behavior.

Our contributions can be summarized as follows:

- We map IP addresses to mobile applications identifying the most popular services accessed via a mobile device in Santiago.
- We apply spatial analysis tools to find application-specific traffic *hot spots* by estimating spatial auto-correlation metrics.

- We perform a regression analysis with generalized linear models at different time intervals to model usage patterns. As input for the regression, we label cellular towers according to contextual urban features, *e.g.*, point of interests. We show that some popular applications are dispersed across the city (more than expected by a random null model), while other applications present spatial clustering, which we analyze from a global and local perspective. Clusters appear in places of high floating population, but not in the same places for all applications.
- We show in what extent commuting plays an important role in application usage, both in terms of time (rush hours) and transportation infrastructure (metro stations, highways, bus corridors).

Finally, we present a discussion on these results, which shed light on what is possible to infer from anonymized and aggregated digital traces. This includes surfacing the spatial and temporal habits of passive users, *i.e.*, those that use applications but do not publish content; and how the relationship of usage trends with commuting empowers users in one of the most widespread daily activities.

2 METHODS AND DATASETS

In this section we present the methods and the datasets we adopt for the descriptive analysis.

2.1 What: Application Labeling

The primary input of this analysis is the log of accesses to specific IP addresses per mobile phone tower. Consequently, the first step of the methodology is to map each IP address to an application label. The application labels include specific applications (*e.g.*, WhatsApp, Facebook, or Twitter) as well as general categories (*e.g.*, Games, or Email Client). In the rest of the paper, we refer to both as applications, apps, or categories.

Given an IP address, we perform a WHOIS[1] query, and we decide on a label depending on the output of the call. Three cases are possible:

(1) The owner of the IP address is identified as a specific service. In this case, we label the corresponding address with the name of the owner. Well known apps were discovered in this case (*e.g.*, Facebook, Twitter, Spotify), along with local websites, newspapers, radio and television services.
(2) The owner of the IP address is a company that makes more than one app (*e.g.*, Apple, Microsoft, or game companies). In this case, we use a reverse DNS lookup service[2] that returns the current and previous hostnames that points (or pointed) to the IP address. Here we hypothesize that the hostname reveals the name of the application. For instance, an IP address associated with the hostname in the form xxx.yyy.com is related to the mobile application *xxx* (or, in some scenarios, *yyy*). With this approach we were able to identify applications such as Skype, WhatsApp, Apple Maps, and Google Maps.
(3) The IP address is hosted in a cloud service. In this case, we query directories of IP addresses generally accessed by Android applications. Such directories has been compiled to study privacy and geographical data breaches. One example is the PDTLoc[3] repository created by Eskandari *et al.* [10]. In this case, applications related with dating and car transportation were identified due to their dependency on external mapping services.

We filter out the IPs that fall in more than one case.

2.2 Where: Spatial Autocorrelation

We seek to understand whether the urban context, *i.e.*, the built environment and the functional areas of a city, influences application usage. This involves testing whether application usage tends to cluster in the spatial arrangement of towers. For instance, one would expect that photo-sharing applications are mostly used in proximity to towers within a touristic area. Since the boundaries of such areas are not known (as they may not be officially defined), one way to find photographed areas of the city, rather than specific points of interests, is through the identification of *Hot Spots*. A spatial cluster, or *Hot Spot*, is a region where the application usage is concentrated, with more accesses per tower than its surroundings. As example, if a tower i has a massive amount of accesses, but its neighbor towers do not exhibit similar behavior, then tower i does not lie within a *Hot Spot*. One way of measuring this is through global and local spatial auto-correlation.

In spatial analysis, there are several spatial auto-correlation metrics that capture this behavior. Here we use the Getis-Ord family of metrics [23], G (global) and G_i^* (local), which have been used in contexts such as traffic incidents [27] and crime-detection with social media [19]. The global G metric is defined as:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d)x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, j \neq i,$$

where $w_{ij}(d)$ is a quadratic distance decay function between towers i and j (by definition, this weight is symmetric); x_j is the total number of accesses to the application under study at tower j. This metric can be standardized, and thus, a value of $z(G)$> 1.96 (which, in a normal distribution, represents the 97.5% percentile point) indicates that app traffic tends to cluster spatially, in comparison with a null model of random distribution; conversely, $z(G) < 1.96$ (2.5% percentile point) indicates that app traffic tends to be more dispersed than expected in a null model, with 95% of confidence. We expect that some applications do not have significant spatial clustering, due to their general usage patterns or their popularity.

For applications that have a significant $z(G)$, we will estimate the local metric G_i^*, defined as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j}.$$

The formulation of G_i^* measures the concentration of app usage at tower i and its surroundings. As its parent metric, G_i^* can also be standardized, which allows to identify clusters of towers that have significantly more accesses by each application (with $z > 1.96$ with a confidence of 95%). Since the z score may be negative, this also

[1] https://whois.icann.org
[2] https://www.robtex.com/
[3] http://titan.disi.unitn.it/pdtloc/apps.php

includes clusters where app access is below than expected (*i.e.*, a *Cold Spot*). In this work, we focus only on *Hot Spots*.

2.3 When and How: Temporal Behavior

In this section, we focus on (1) the temporal dynamics of the application usage signal and (2) its relation with urban indicators. For instance, two spatial clusters may be in different areas of the city, yet they may share characteristics, such as the socio-economic indicators, or the availability of types of points of interests. To evaluate this, we performed a series of generalized linear model (GLM) [22] regressions over the set of IP accesses. Particularly, this model uses a Negative Binomial regression [13], which is used for count data (*e.g.*, the number of accesses) in over-dispersed scenarios (*e.g.*, the variance is larger than the mean). In prior work, we used this method to measure the effect of Pokémon Go in a city [11], which we adapt here for application accesses:

The model is specified as follows:

$$\log(E[X_a(b,t)]) = \log \alpha + \beta_0 + \sum \beta_i X_i$$

where $E[X_a(b,t)]$ is the expected number of connections to application a using a tower b at time-window t. The covariates X_i are defined as follows:

- `Indoor`, `Metro` (binary): whether tower b is installed indoor (*e.g.*, inside hospitals, malls, institutions) or within an underground metro station. `Indoor` and `Metro` towers guarantee that the user is within the perimeter of an enclosed space.
- `BusRT`, `HighWay`, `Pedestrian` (binary): whether tower b is located up to 500 meters from a transportation infrastructure: respectively, bus corridors, highways, and pedestrian streets.
- `MainStreet`, `SecondaryStreet` (binary): whether tower b is located within 500 meters of main streets (the main directives of a city according to OpenStreetMap), or secondary streets (streets usually connecting main streets).
- `IncomeDecile` (ordinal): the decile of mean resident income of the area in which b is installed.
- `GreenAreas` (binary): whether tower b is located up to 500 meters from parks or public squares, or within them.
- `Weekend` (binary): whether t's day is Saturday or Sunday.

Note that the covariates use dummy encoding in case of categorical variables. Moreover, the model exposure α represents the number of users connected to each tower, which allows to control for the fluctuations of population throughout the day.

By evaluating this model at every time-window available, it is possible to study the variation of the application usage during a day (β_0), as well as the effect of the covariates. To interpret the results, we visualize how the Odds-Ratio (OR) of a covariate ($\exp \beta_i$) evolves during the day. The OR is interpreted as the fraction of change in the observed variable. An OR of 1 implies no significant effect; an OR $>$ 1 implies a positive change (*e.g.*, OR = 1.5 means 50% increase), given all other factors remain equal.

2.4 Datasets

Our work analyzes the city of Santiago, Chile, a city with almost 8 million inhabitants and an urban surface of 867.75 square kilometers. It is composed of 35 independent administrative units called municipalities. Figure 1 displays the municipalities within the urban area of Santiago, according to their average yearly income.

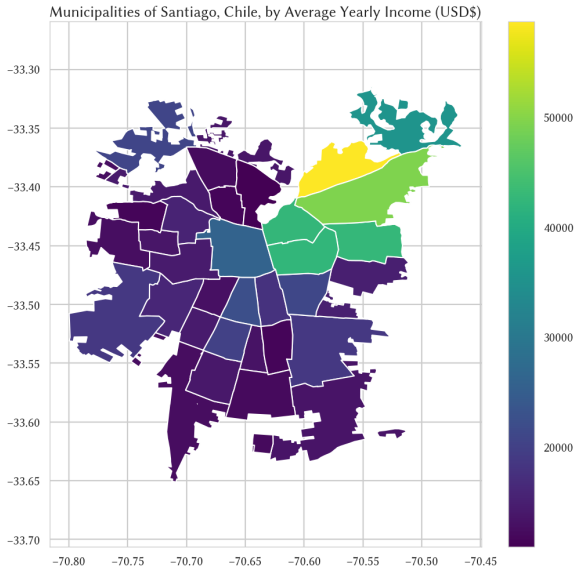

Figure 1: Choropleth map of the municipalities within Santiago, Chile. The color of each municipality represents the average yearly income in US dollars.

To describe how the urban context is related with application usage, we work with an aggregated Deep Packet Inspection (DPI) dataset from Telefonica Chile, the biggest telco in the country, with a market share of 33% in 2016. DPI is a family of probe methods for network data, that, given a network package, identifies several of its properties, such as requested IP address and port. In particular, our dataset comes from the aggregation of DPI in the HUAWEI SmartCare SEQ Analyst and NetProbe tools [14], which is aimed at improving quality of service in mobile phone networks. Specifically, the NetProbe tool performs aggregation of multiple TCP flows, multiple HTTP transactions, email transactions or streaming buffers within a single record, identifying the initial and final timestamp of the application usage.

The dataset contains the number of connections to the top 5,000 IP addresses accessed by devices with a Telefonica SIM card with a data plan, between the July 27th, 2016, to the August 10th, 2016. Each row of the dataset contains the following information: `tower id`, `date`, `time window`, `IP:port`, `number of accesses`. The 5,000 limit represents around 80% of the most accessed IPs. The number of connections is sampled in time windows of 10 minutes at each tower in the urban area of the city. These thresholds were determined to impede analysis of tower and IP pairs which would enable user identification (*e.g.*, access to private servers).

To give geographical context to each observation, we use the Telefonica tower network (*c.f.* Fig. 2 Top Left). It is worth noting that tower distribution is not uniformly distributed within the city. In fact, areas of the city with greater flows of people tend to have greater density to ensure a high quality of service.

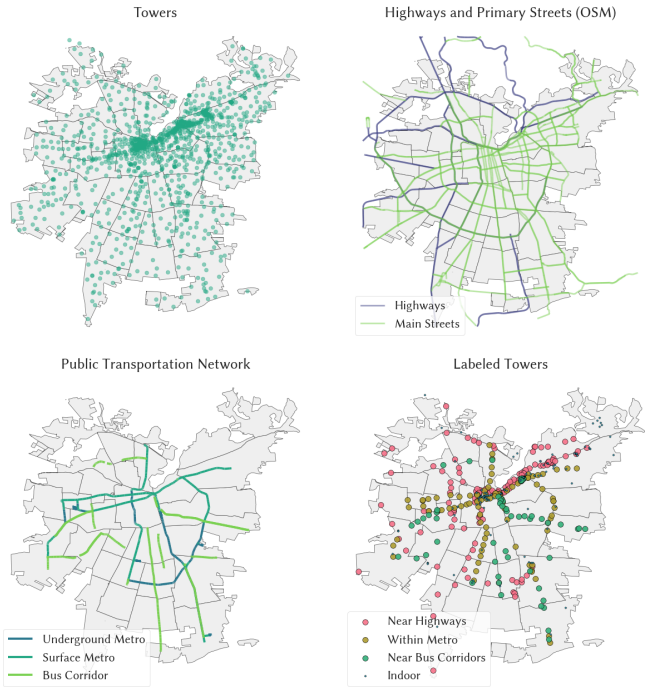

Figure 2: Schematic maps of Santiago. Top Left: spatial distribution of Telefonica towers. Top Right: OpenStreetMap network of highways and main streets. Bottom Left: OSM network of metro (surface and underground), and bus corridors. Bottom Right: towers labeled with specific features derived from distance to the OSM features.

To model the urban context, we downloaded the mobility network and the points of interest data from OpenStreetMap (OSM), dated in August 2016.[4] We use OSM to label towers according to their proximity to different elements of the mobility infrastructure. Fig. 2 gives a visual summary of the data: in the Top Right, highways and main streets of the city; in the Bottom Left, part of the public transportation network; and in the Bottom Right, labeled towers according to urban features.

To have an approximation of the number of application accesses per user, we used a dataset from a previous study [11], which contains a number of connected users to each tower. We consider it a lower bound on the number of people at each tower because it does not cover the entire market share of the telco. Note that this dataset covers the exact same period of analysis as the DPI dataset.

3 RESULTS

In this section we describe the results of applying the defined methods to the dataset under study.

3.1 Application Labeling

After applying the labeling procedure, we were able to assign an application/category to 1133 IP addresses. Table 1 shows the results. In addition to mobile apps, we categorized browser usage to specific content as a category. In the case of news, since we are interested in understanding news consumption instead of profiling access to specific outlets, we assigned them to the broader News category. Regarding the unlabeled IPs, the majority of them belonged to national ISP and DNS servers, cloud service providers, and analytic/ads services.

[4]Source: http://download.geofabrik.de/south-america/chile.html

Table 1: Found applications and categories of applications in the dataset. Note that News and Wikipedia refer to Web browser traffic.

Category	Apps	#IPs
Music	iTunes, Spotify, Soundcloud	33
Video	Youtube, Netflix, CrunchyRoll	38
Games	Zynga, Blizzard, King, Pokémon Go, GameLoft	31
Taxi	Uber, Cabify, EasyTaxi	6
Messaging	WhatsApp, Telegram, Line	735
Maps	Apple Maps, Google Maps	7
News	International (New York Times, The Guardian, *etc.*) and national news outlets (El Mercurio, COPESA, *etc.*)	43
Mail	Gmail, Yahoo-Mail, Outlook, Corporate Mail	99
Dating	Tinder, Grindr	5
Facebook	-	84
Twitter	-	41
Instagram	-	9
Wikipedia	-	2

The average daily traffic (and its standard deviation) of each identified category/application is shown on Figure 3. One can see that different patterns emerge from the distributions of their daily traffic. For instance, music applications seem to be strongly related with commuting, due to the location of their traffic peaks. Another pattern that can be seen is the steadily increase in traffic during the day for some categories, meaning that people play games, flirt (Dating) and watch videos mostly at the evening, when back at home from work (or study). Messaging and mail present a bell-shaped distribution, which account for their usage within work/study environments and when far away from family and friends. As such, these results support our hypothesis that application usage heavily differs within a day. We have left to see if such use differs also in the spatial and temporal contexts related to the urban built environment.

3.2 Global Autocorrelation and Hot Spots

We performed the analysis on the following periods of the day: business days between 6AM and 9AM (*e.g.*, morning commute), 2:30PM and 5:30PM (*e.g.*, afternoon activities, where we assume a majority of people tends to stay at one place), and 9PM to 12AM (*e.g.*, night at home or recreational places); weekends between 4PM and 9PM (*e.g.*, weekend activities, where we assume people tend to go outside of their homes). In these periods, we considered the cumulative traffic volume per tower/app pairs. Since we do not have all tower traffic, due to our focus on specific apps, and the

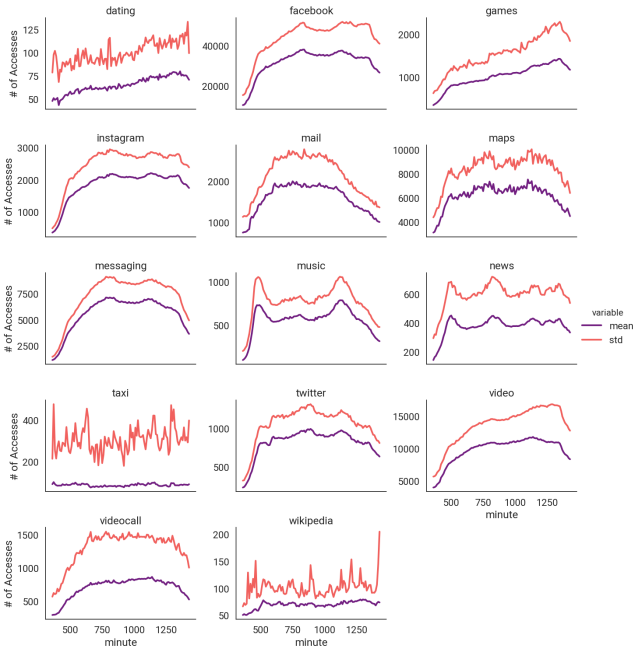

Figure 3: Time series of mean accesses (and their standard deviation) to the IP addresses in each category, according to the time of the day they were performed. The deviations showcase the over-dispersion of each time series.

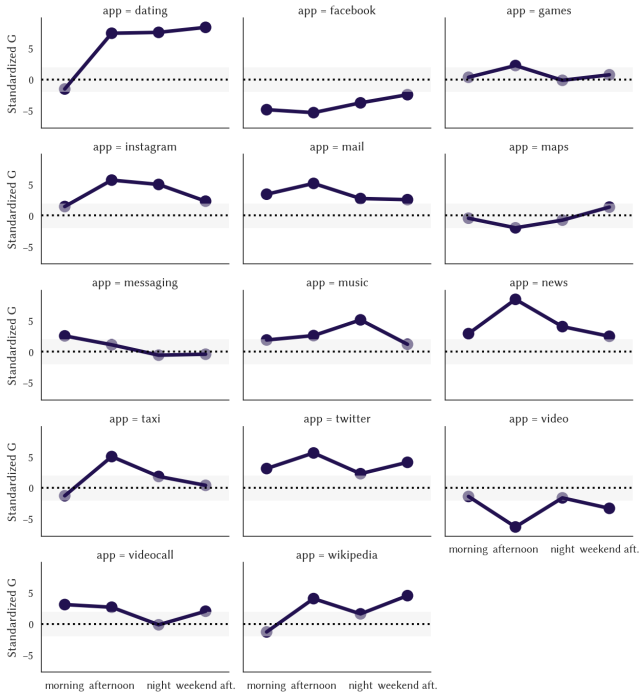

Figure 4: Point plots of the std. global Getis-Ord spatial autocorrelation G for each application, at business days (morning, afternoon and night) and weekend afternoon. The grey area represents the values of std. G that are not significant with respect to a null model. Positive values imply spatial clustering; negative values imply spatial dispersion.

limit of 5,000 IPs, we estimate the relative portion of traffic between the apps under study per tower. As distance band for the spatial autocorrelation, we defined a threshold of 1 Kilometer with quadratic decay. Then, we estimated G and G_i^* for all towers, except those in underground metro-stations, as they may be near other towers, but metro passengers are in a different context than people outside (note that we considered those towers in the temporal analysis).

Figure 4 shows the global correlation metric G using point-plots. One can see that there are dispersed apps in the city ($z(G) < -1.96$). This dispersed behavior occurs when high values repel other high values, and tend to be near low values, while clustered behavior occurs when high values cluster near other high values; or low values cluster near other low values. Note that Facebook and Video are always dispersed, while Maps are dispersed in only one time-window (business day afternoons). Null behavior, *i.e.*, apps without autocorrelation ($|z(g)| < 1.96$), also appear on the Figure. No autocorrelation implies neighbors values are random. Some applications present this behavior except in a specific time window: Messaging ($z(G) > 1.96$ in business day mornings), Games and Taxi (both $z(G) > 1.96$ in business day afternoons). The remaining apps present a majority (or all) of spatial autocorrelated time-windows.

We focused our local autocorrelation analysis on all the positive and significant time windows $z(G)$. Figure 5 shows the detected *Hot Spots* for all applications. It's worth noting that several apps present spatial clusters in similar locations, due to the high density of floating population in those areas (refer to Figure 2 for a mapping of the phone towers). As expected, some widespread applications present *Hot Spots* geographically distributed across the entire city: e.g., Mail, News, Twitter, VideoCall, Instagram, and Music. The category Games follows the same behavior with a less dense spatial pattern. On the contrary, other categories are densely clustered in specific areas, where their use is more prevalent. These include Messaging and Dating (areas with a high concentration of business and commerce), or Taxi (business districts in high income areas, and places with loaded public transportation). Wikipedia does not seem to show a direct interpretable pattern. An interesting result that comes from the inspection of the *Hot Spots* distribution, is that, while spatial clusters tend to appear in areas of high floating population, they are not the same for all applications. For instance, News cover a wider area than the rest.

In addition, Figure 5 shows how the *Hot Spots* vary during the day, with some being significant for specific time-windows only. For instance, Messaging apps register hot spots mainly in the mornings, on the contrary, Dating cover several time-windows, as illustrated in the circular-like set of towers in downtown and near an area with the biggest shopping center in the city.

3.3 Regression Results

We evaluated the regression model within every 10-minute time interval between 6AM and 12AM. As exposure parameter we computed the number of people connected to each tower, as previously

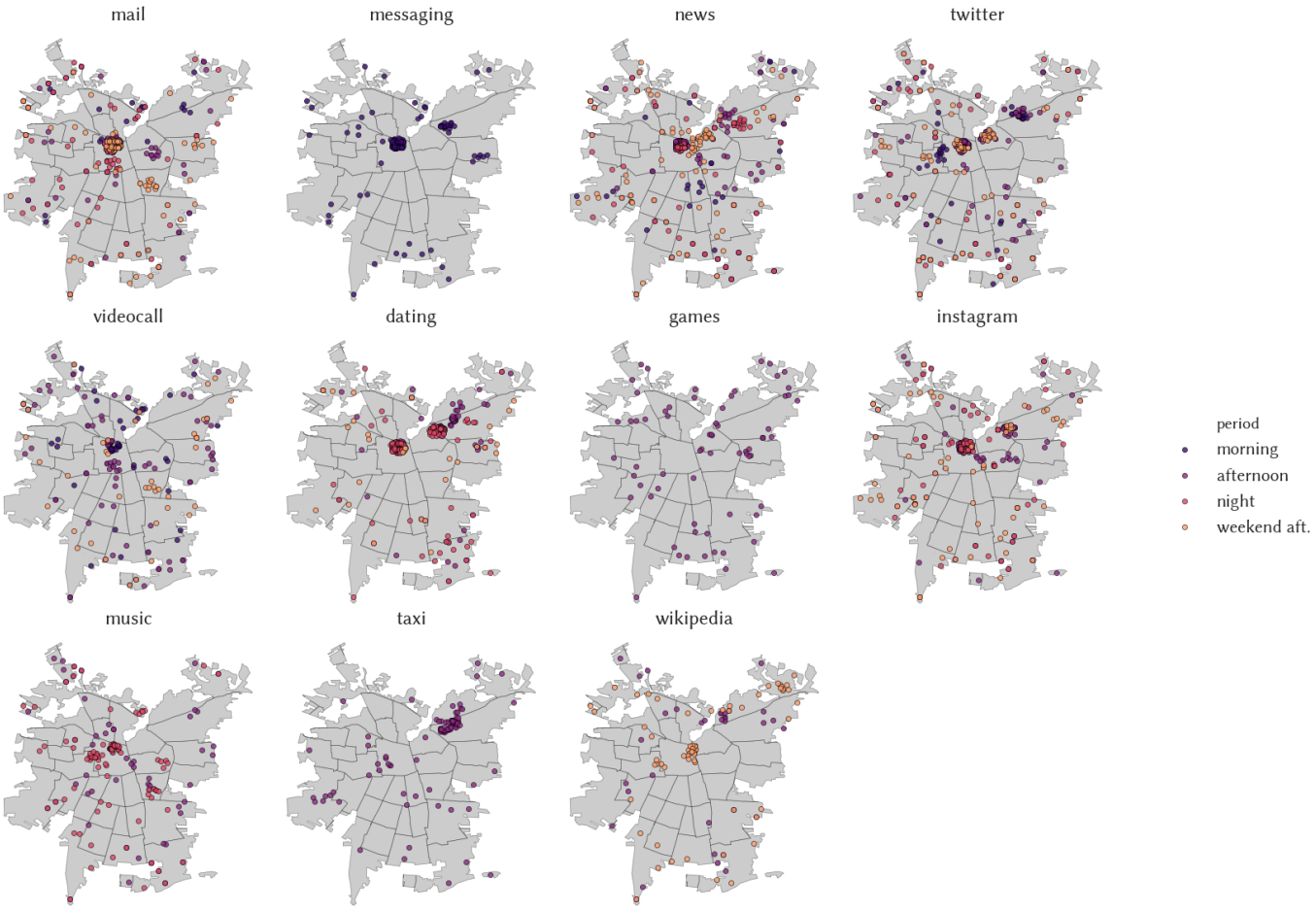

Figure 5: Dot maps of Santiago. Each map shows a set of application *Hot Spots* according to the local Getis-Ord G_i^* metric. Each spatial cluster is colored according to its respective time-window.

done in [11]. This allowed to modulate the application traffic in each tower according to the lower bound of people connected to it.

Figure 6 shows a summary of the β coefficient for each application category. To simplify the exploration of the results, we applied a hierarchical clustering. The three clusters found are: (I) Music, Mail, Video, Games; (II) Instagram, Twitter, Wikipedia, VideoCall, Maps, Dating, Taxi; and (III) Messaging, Facebook, News. The time-series in Figure 6 are colored according to the app clusters. However, a manual inspection of the charts did not reveal visually salient differences among them.

Conversely, the factor clusters analysis shows a clear semantic separation within them. The two clusters are: (A) *Is Weekend?*, *Near Secondary Streets*, *Near Primary Streets*, the *Intercept*, *Near Pedestrian Streets*; and (B) *Near Bus Corridor*, *Near Green Areas*, *Indoor*, *Income Decile*, *Near Highways*, *Within Metro*. On the one hand, cluster A refers to urban infrastructure from a city point of view (*e.g.*, "this is a main street"). On the other hand, Cluster B refers to kind of places and modes of transportation. Due to space constraints, here we summarize the main findings from these results, by focusing on some factors of each cluster.

In Cluster A: overall, the intercept captures the general patterns of application access (*c.f.* Fig. 3), including Music, which follows a commuting-related distribution with peaks around labor hours. Weekend seems to have a similar behavior in all apps – with the exception of Dating, which follows the opposite trend, and Taxi, which presents pulsating, yet significant, positive effects, mostly during the morning and afternoon. The street factors (Main Street, Secondary Street, Pedestrian) present mixed effects: some apps have positive effects on them while null or negative effects on the others, reflecting how people use streets (*e.g.*, Messaging has almost 10% more traffic in main streets, possibly because main streets have more people within them). Pedestrian streets have an overall null or negative relation with app traffic, except for Dating applications (*e.g.*, pedestrian streets contain good meeting points).

In Cluster B, the Metro factor shows distributions related to commuting hours. However, this behavior is not always symmetric. For instance, Music apps have 400% more traffic at the morning rush hour, while at the afternoon rush hour it is 300%. Highways usually have a higher association with applications, except with two that present larger amounts of text: Wikipedia and News (with

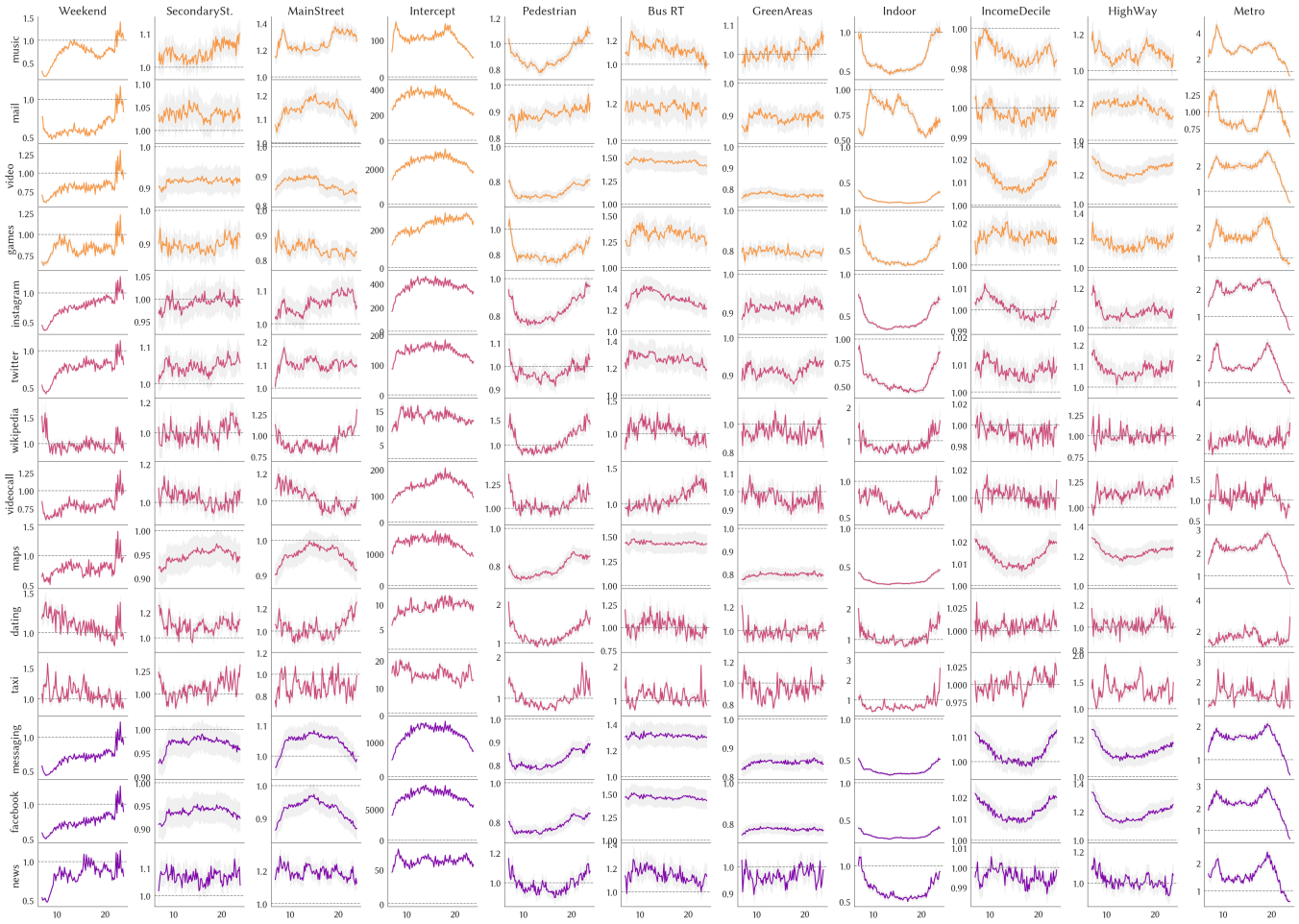

Figure 6: Matrix of time-series of the Odds-Ratio (effect size) of each factor under analysis, per application. The grey area around the time-series represents the 95% confidence interval of each factor, which should not intercept 1 to be significant. The factors and applications were ordered in the matrix using hierarchical clustering (colors represent the application clusters).

those two categories the effect is null). We expected to see positive effects of Green Areas in apps like Instagram, but the opposite was found. In terms of Indoor contexts, the effect is usually negative, with exception at specific times of the day for Mail. Bus Corridors tend to have positive effects for all apps, except some null cases like Taxi, Dating, VideoCall and Wikipedia. One would have expected to behave in similar ways to the Metro factor, but bus riders may have a different profile than Metro's. Finally, the effects of area Income are small (at most 2% of traffic increase per increase in income decile), meaning that it does not show a powerful interaction though the urban context, probably due to the effect of floating population (which comes from different city areas).

4 DISCUSSION

We performed a descriptive analysis of spatial and temporal patterns of application usage in the city. Some of these results are expected, due to our restriction to the most popular IP addresses, but others can be counter-intuitive, such as the small effect of area income or the lack of interesting effects of green area proximity. Even though application context is not a new topic, to the extent of our knowledge it has not been done at this scale, with the entire traffic of a major telco. for the most popular apps in a big city.

One interesting aspect is that many applications are scattered around the city (while still maintaining a degree of local clustering), yet they are perceived as applications with a biased user base. An example of such case is Twitter, where few users participate [3]. Following our results, Twitter may be biased in terms of who generates the content, but passive users may be less biased, as they are present in many areas of the city.

We observed that *Hot Spot* locations change during the day, and some factors vary their influence in app traffic. For application providers, this implies that information needs within the same application may vary depending on the context. An important context in which we put emphasis was commuting. Commuting is one of the most recurrent daily activities, and at the same time, it is the one least enjoyed [15]. However, not all commuters suffer, some

even say that "getting there is half the fun" [21]. The notion of *equipped time* [30] implies that, regardless of mode of transportation, commuters have several choices to make use of their time while traveling. While this has been studied in the past, the possibilities enabled by smartphones are yet to be explored–something that our study shows through the factors related with commuting, and their mostly positive associations with app traffic. Given that commuting times are on the rise [17], this is a relevant topic to study, with potential impact on user's lives.

5 RELATED WORK

The most common datasets from mobile phone networks are Extended and Call Detail Records (X/CDR), which have been studied extensively (see a survey [4]). Indeed, one of our auxiliary datasets is comprised by XDR data [11]. But, to the extent of our knowledge, there is no other work that analyzes Deep Packet Inspection data of applications with respect to to the city context. Having said this, our descriptive analysis is not new in terms of intent. The most similar study to ours is [5], however, the scale and focus is different–the study relied on users installing an application in their phones. Another alternative is to perform in-situ studies, which provide qualitative insights, but the scale is even smaller [7]. As such, even though the idea has been explored, it has been done with a different focus (users) and scale (thousands of users), while we have analyzed the application traffic of the entire city, using a dataset from the biggest telco. in the country.

Another source is mobile sensor data and mobile application logs, which have been used to understand application usage, particularly, on prediction of the next app users are going to open [2, 9]. Our focus is different, in the sense that we are looking at aggregated behavior and its relationship with the urban context. In fact, in our setup it is impossible to individualize users.

The DPI dataset we used is focused on quality of service. These measures are expensive to obtain in a device-driven context, but there have been endeavors on that area, such as [6]. That line of research focus on the mobile phone network status rather than the information needs of users.

The city context has been explored before, particularly in terms of commuting [20], land use and functional areas [12, 28], urbanism [8], and how the city is felt [1, 24] or remembered [29]. In this work, we have proposed another way to characterize the city through the applications its inhabitants use, which provides another perspective on what shapes (or is shaped) by the city.

6 CONCLUSIONS

We presented a descriptive analysis of how mobile phone application traffic is related with the city their users live in. We focused on the following aspects: *What* applications were used, among the most popular; *Where* and *When* those applications generated traffic; and *How* was the urban context surrounding users. The input was a dataset of Deep Packet Inspection from the biggest telco. in Chile, which, jointly with data from OpenStreetMap, allowed us to perform the analysis. We found that several applications present identifiable *Hot Spots* scattered around the city, and that the urban context, in terms of infrastructure, allows to understand the relationship between daily activities and application usage.

One limitation of our work is our focus on the most popular apps. There may be other apps, with a more fragmented user-base (*e.g.*, sport applications), that may still influence/be-influenced by the urban context, and that may show relevant differences in factors such as income. However, to solve this limitation, a strategy to aggregate data with the telco. should be defined, as a way to analyze more data without giving up privacy concerns. We hypothesize that if more apps were considered, the results could change incrementally, *i.e.*, the found patterns should remain. The extent of those patterns is left for future work.

Finally, an important aspect is privacy. We showed that, even though application traffic may be encrypted and aggregated, DPI data is able to reveal usage patterns on the city, even though these patterns present limits.

Acknowledgements. We thank Pablo García Briosso of Telefonica Research for granting access to the data, and Javier Bustos of NIC Chile for useful discussion regarding IP labeling. We acknowledge the following libraries used in the analysis: `PySAL` [25], `statsmodels` [26], `geopandas`,[5] and Project Jupyter [16]. Part of the data used in the schematic maps and our analysis is attributed to OpenStreetMap contributors.

REFERENCES

[1] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. 2016. Chatty maps: constructing sound maps of urban areas from social media data. *Open Science* 3, 3 (2016), 150690.
[2] Ricardo Baeza-Yates, Di Jiang, Fabrizio Silvestri, and Beverly Harrison. 2015. Predicting the next app that you are going to use. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 285–294.
[3] Ricardo Baeza-Yates and Diego Saez-Trumper. 2015. Wisdom of the Crowd or Wisdom of a Few?: An Analysis of Users' Content Generation. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 69–74.
[4] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4, 1 (2015), 1.
[5] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. 2011. Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*. ACM, 47–56.
[6] Javier Bustos-Jiménez, Gabriel Del Canto, Sebastián Pereira, Felipe Lalanne, José Piquer, Gabriel Hourton, Alfredo Cádiz, and Victor Ramiro. 2013. How AdkintunMobile measured the world. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 1457–1462.
[7] Juan Pablo Carrascal and Karen Church. 2015. An in-situ study of mobile app & mobile search interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2739–2748.
[8] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. 2016. The death and life of great Italian cities: a mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 413–423.
[9] Trinh Minh Tri Do and Daniel Gatica-Perez. 2014. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing* 12 (2014), 79–91.
[10] Mojtaba Eskandari, Bruno Kessler, Maqsood Ahmad, Anderson Santana de Oliveira, and Bruno Crispo. 2017. Analyzing Remote Server Locations for Personal Data Transfers in Mobile Apps. *Proceedings on Privacy Enhancing Technologies* 2017, 1 (2017), 118–131.
[11] 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.
[12] Eduardo Graells-Garrido, Oscar Peredo, and José García. 2016. Sensing urban patterns with antenna mappings: the case of Santiago, Chile. *Sensors* 16, 7 (2016), 1098.
[13] William Greene. 2008. Functional forms for the negative binomial model for count data. *Economics Letters* 99, 3 (2008), 585–590.

[5] http://geopandas.org

[14] HUAWEI. [n. d.]. Smartcare. http://www.webcitation.org/query?url=http3A%2F%2Fwww1.huawei.com%2Fenapp%2F9%2Fhw-u_256445.htm&date=2018-02-05. ([n. d.]). Accessed: 2018-02-05.
[15] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780.
[16] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. 2016. Jupyter Notebooks-a publishing format for reproducible computational workflows.. In *ELPUB*. 87–90.
[17] David Levinson and Yao Wu. 2005. The rational locator reexamined: Are travel times still stable? *Transportation* 32, 2 (2005), 187–202.
[18] Kevin Lynch. 1960. *The image of the city*. Vol. 11. MIT press.
[19] Nick Malleson and Martin A Andresen. 2015. The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science* 42, 2 (2015), 112–121.
[20] Graham McNeill, Jonathan Bright, and Scott A Hale. 2017. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science* 6, 1 (2017), 24.
[21] Patricia L Mokhtarian and Ilan Salomon. 2001. How derived is the demand for travel? Some conceptual and measurement considerations. *Transportation research part A: Policy and practice* 35, 8 (2001), 695–719.
[22] John A Nelder and R Jacob Baker. 1972. Generalized linear models. *Encyclopedia of statistical sciences* (1972).
[23] J Keith Ord and Arthur Getis. 1995. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis* 27, 4 (1995), 286–306.
[24] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. 2015. Smelly Maps: The Digital Life of Urban Smellscapes. In *Ninth International AAAI Conference on Web and Social Media*.
[25] Sergio J Rey and Luc Anselin. 2010. PySAL: A Python library of spatial analytical methods. *Handbook of applied spatial analysis* (2010), 175–193.
[26] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, Vol. 57. 61.
[27] Praprut Songchitruksa and Xiaosi Zeng. 2010. Getis-Ord spatial statistics to identify hot spots by using incident management data. *Transportation Research Record: Journal of the Transportation Research Board* 2165 (2010), 42–51.
[28] Carmen Karina Vaca, Daniele Quercia, Francesco Bonchi, and Piero Fraternali. 2015. Taxonomy-Based Discovery and Annotation of Functional Areas in the City. In *Ninth International AAAI Conference on Web and Social Media*.
[29] Shoko Wakamiya, Hiroshi Kawasaki, Yukiko Kawai, Adam Jatowt, Eiji Aramaki, and Toyokazu Akiyama. 2016. Lets not stare at smartphones while walking: memorable route recommendation by detecting effective landmarks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1136–1146.
[30] Laura Watts and John Urry. 2008. Moving methods, travelling times. *Environment and Planning D: Society and Space* 26, 5 (2008), 860–874.