Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement

Dmytro Karamshuk [*]
IMT Lucca, Italy
dmytro.karamshuk
@imtlucca.it

Anastasios Noulas
Computer Laboratory
University of Cambridge, UK
an346@cam.ac.uk

Salvatore Scellato
Computer Laboratory
University of Cambridge, UK
ss824@cam.ac.uk

Vincenzo Nicosia [*]
School of Math. Sciences
Queen Mary University, UK
v.nicosia@qmul.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge, UK
cm542@cam.ac.uk

ABSTRACT

The problem of identifying the optimal location for a new retail store has been the focus of past research, especially in the field of land economy, due to its importance in the success of a business. Traditional approaches to the problem have factored in demographics, revenue and aggregated human flow statistics from nearby or remote areas. However, the acquisition of relevant data is usually expensive. With the growth of location-based social networks, fine grained data describing user mobility and popularity of places has recently become attainable.

In this paper we study the predictive power of various machine learning features on the popularity of retail stores in the city through the use of a dataset collected from Foursquare in New York. The features we mine are based on two general signals: geographic, where features are formulated according to the types and density of nearby places, and user mobility, which includes transitions between venues or the incoming flow of mobile users from distant areas. Our evaluation suggests that the best performing features are common across the three different commercial chains considered in the analysis, although variations may exist too, as explained by heterogeneities in the way retail facilities attract users. We also show that performance improves significantly when combining multiple features in supervised learning algorithms, suggesting that the retail success of a business may depend on multiple factors.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining

Keywords

optimal retail location, machine learning, location-based services

---

[*]This work was done while the author was based in the Computer Laboratory of the University of Cambridge.

1. INTRODUCTION

The geographical placement of a retail store or a new business has been of prime importance from the establishment of the first urban settlements in ancient times to today's modern trading and commercial ecosystems in cities. Open a new coffee shop in one street corner and it may thrive with hundreds of customers. Open it a few hundred meters down the road and it may close in a matter of months.

In this paper we take advantage of the new layers of information offered through check-in data in Foursquare to frame the problem of optimal retail store placement in the context of location-based social networks. That is, given a set of candidate areas in the city to open a store, our aim is to identify the most promising ones in terms of their prospect to attract a large number of check-ins (i.e, become popular). We formulate this problem as a data mining task, where, by extracting a set of features, we seek to exploit them to assess the retail quality of a geographic area. In more detail, our contributions are:

Spatial and mobility analysis of retail store popularity. We conduct an analysis of the popularity of Foursquare venues: these follow a power-law distribution, which indicates the existence of very heterogeneous check-in patterns across places. Focusing on the venues of three retail stores chains we explore how their popularity is shaped by spatial and human mobility factors. We find that 50% of user movements originate from nearby venues within 200 to 300 meters, and 90% of movements occur within 1km. This suggests a strong *local* bias in the attraction of customer crowds. Moreover, the geographic placement of retail stores with respect to different venue types is non-random: there is, in general, a higher likelihood of observing a store near a transportation hub or a touristic spot (museum, hostel, etc.), compared to a randomly picked Foursquare venue type. We subsequently extend this analysis by considering the movements between retail chains and other places in proximity, discovering that co-location does not necessarily imply higher probability of movement between two types of venues. This highlights how a deeper insight into human mobility patterns captured through location-based services in the city can improve the performance of static prediction frameworks on local business analytics.

Mining features in location-based services for the retail assessment of a geographic area. Driven by the findings reported above we mine Foursquare user check-ins so as to capture a variety of signals that may be informative about the retail quality of a geographic area. The data mining features we choose belong to two

well-defined classes: geographic and mobility. The *geographic features* encode spatial information about the properties of Foursquare venues in an area. This includes information about the existence of certain types of settlements (coffee shop, nightclub, etc.) in an area or its density. Moreover, we factor the *competitiveness* of an area by examining the influence of the presence of competitor business venues on the popularity of a retail facility. The set of *mobility features* we devise involves the measurement of the popularity of an area in terms of number of check-ins: for instance, we model transition probabilities from nearby venues to the target store or the attraction of crowds from remote locations.

Individual feature and supervised learning evaluation. Finally, we assess the performance of the features mined in the previous step in terms of their ability to predict high rankings for the most popular geographic spots individually, and combined in a supervised learning framework. Feature performance can vary across different retail chains as there are observed idiosyncrasies in the ways customer crowds are attracted at those. In principle, however, features accounting for the degree of competition in an area or those that model the spatial structure of an area, based on the existence of certain venue types, perform best in the ranking task. From the class of mobility features, the best predictors are those that rank areas based on their attraction of users from distant locations and those which consider the *transition quality* of an area by means of venue types. When the features are combined using supervised learning models we observe a clear improvement in performance with the optimal retail spot being constantly ranked in the top-5% of the prediction list in one over two cases. In addition, the supervised learning algorithms benefit marginally, yet steadily, when we employ mobility features in comparison to using those which encode solely the static geographic properties of an area.

We envisage that similar approaches can influence research in urban mining where a variety of applications may benefit, including the provision of better services for businesses and citizens or predictions of house price evolution, development indices of urban areas and location-based marketing. While Foursquare offers business accounts where shops can register themselves and obtain basic analytics such as the times of the week their business is becoming more or less popular, or the demographics of the users that check in to them [1], data mining approaches that exploit the rich spatio-temporal datasets sourced from these services can considerably boost their business model by offering new commercial opportunities to their users beyond their principal application scope so far that has been location-based activity recommendations [4].

2. DATASET ANALYSIS

In this section we provide essential details about the Foursquare check-in dataset and subsequently we elaborate on the analysis of venue and retail store popularity in the service. Particularly, we concentrate on the three most popular retail chains in the New York area and study their spatial and mobile interactions with other Foursquare venues and the mobility patterns of the users that visit them.

Data Collection

Foursquare was launched in 2009 and it has quickly become the most popular location-based service, with more than 35 million users as of January 2013 [2]. Per-user Foursquare check-in data is not directly accessible, however, users can opt to share their check-ins publicly on Twitter. We thus were able to crawl for publicly-available check-ins via Twitter's streaming API[1]. Note that we can

[1]`https://dev.twitter.com/docs/streaming-api`

Chain Name	Check-Ins	Places	$\langle c_p \rangle$
Starbucks	$210,174$	186	1129.97
Dunkin' Donuts	$26,955$	104	259.18
McDonald's	$15,014$	66	227.48

Table 1: Summary of Chain Statistics: Total number of check-ins as observed through Foursquare's Venue API[2] at the time of data collection, number of places and average number of check-ins per place ($\langle c_p \rangle$)

only access those check-ins that users explicitly chose to share on Twitter, although users have the possibility to set this option as default. In the present work we use a dataset of check-ins and venue information in city of New York and the surrounding area. New York is the city where the service was launched, and due to this fact the adoption of the service in the area is significantly larger than in any other place in the world. Numerically speaking, we consider the square region of 10×10 km around the geographical center of New York (Manhattan Area, Coordinates: 40° 45'50"N 73° 58'48"W), featuring $37,442$ geo-tagged venues, $46,855$ users and $620,932$ check-ins collected in a period of 5 months (May 27th to November 2nd 2010). We note that according to our estimates this sample accounts for approximately 25% of all check-ins collected by Foursquare in the aforementioned region and time frame.
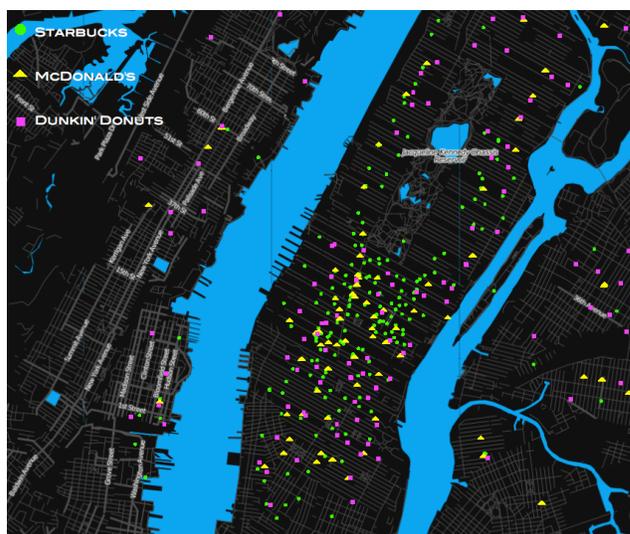

Figure 1: Spatial distribution of the three store chains in the New York area.

Analysis of Retail Store Popularity

Over the dataset just described we have analyzed the popularity of the different places present. By popularity of a place we mean the total number of check-ins we have observed in this venue. Figure 2 illustrates the Complementary Cumulative Distribution Function (CCDF) of check-ins per venue in the dataset, considering all places in New York. The functional form of the distribution resembles a power-law and the check-in frequency spans a large number of orders of magnitude.

From this point on we concentrate on the analysis of venues' popularity for individual store chains. We have observed consid-

[2]`https://developer.foursquare.com/docs/venues/venues`

erable discrepancies between the check-in patterns across different chains. In Table 1 we show basic statistics for the three chain stores we have elected based on their number of venues in New York. Starbucks is the chain with the highest number of venues, 186. Dunkin' Donuts and McDonald's follow with 104 and 66 stores, respectively. In Figure 1, we present a visualization of the spatial distribution of the stores for the three retail chains we analyze. The mean number of check-ins per place at Starbucks is equal to 1129.97, almost five times larger than the rest. Similarly, the CCDF of check-in volume for Starbucks restaurants, as shown in Figure 2, features a significantly longer tail than in the other two cases. Among other reasons, that could be attributed to the fact that visitors of a coffee shop stay there longer and thus are more likely to check-in. On the other hand, fast food chains may attract opportunistic visits just to pick up food. Addressing the problem thus for each chain separately can help us alleviate the biases that may be introduced by check-in variations across place types in Foursquare.
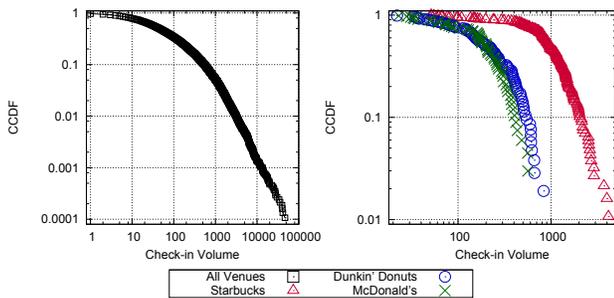

Figure 2: Complementary Cumulative Distribution Function (CCDF) of check-ins per place for all venues in the dataset (left) and for the three considered store chains (right).

Spatial Interactions between Business Activities

Previous work aiming to assess the retail quality of places based on a dataset with similar spatial characteristics was proposed by Jensen in [10]. The author studied the spatial interactions between different place categories in the city of Lyon, France. The proposed metric computes the frequency of co-location for different dyads of place types, for instance, coffee shops close to railway stations, restaurants close to shopping centers, and compares the results to the corresponding expectation when settlements are distributed uniformly at random in the city. If the measured frequency is higher than the expected value, it is proposed that the categories tend to attract each other. If it is lower the two place types repel each other. The resulting inter-type coefficients are then exploited to assess the retail quality of an area for a given place type.

Provided with the rich spectrum of place types (e.g., coffee shop, fast food restaurant, train station etc.) that Foursquare venues can be semantically annotated with, we compute the inter-category attraction coefficients for 248 different venue categories (we discuss the details of the computation in Section 3). The results of the Top 10 attractors for each chain with the corresponding inter-category coefficients are shown in Table 2. We observe that Starbucks and McDonald's restaurants are often placed near Train Stations with frequency 11.80 (for Starbucks) and 3.08 (for McDonald's) times higher than expected in the scenario when venues are distributed randomly in the city. This can be intuitively explained by the aspiration to benefit from the massive flows of people generated around transportation hubs. These crowds may seek food and refreshment opportunities as they travel. Another common attractor for the two

Starbucks		Dunkin' Donuts		McDonalds	
Train Station	11.80	Hostel	5.02	Flower Shop	5.87
Light Rail	8.60	Gas Station	3.05	Office Supplies	3.16
Stadium	7.25	Automotive Shop	2.66	Train Station	3.08
Airport	6.24	Flower Shop	2.36	Theater	2.84
Museum	5.10	Post Office	2.19	Light Rail	2.32
Convention Center	4.93	Flea Market	1.84	Gift Shop	2.26
Hostel	4.82	School	1.72	Subway Station	2.21
Corporate Office	4.57	Drug Store	1.70	Department Store	2.17
Hotel	4.13	Subway Station	1.67	Bank / Financial	1.92
Bank / Financial	4.09	Bike shop	1.64	Drug Store	1.89

Table 2: Top 10 most attractive categories for each chain as defined by Jensen's inter-category coefficients with the corresponding values of the coefficients

chains are Corporate Offices, as these are good sources of regular clients. Similarly, the two coffee shop chains, Starbucks and Dunkin' Donuts, are often placed around Museums and Hostels are a source of tourist crowds. Less frequently, but, still considerably more often than expected at random, Dunkin' Donuts can be found next to Gas Stations.

Exploiting Mobility Data for Retail Analysis

Motivated by Jensen's approach [10] discussed above we extend the analysis of the retail quality of urban areas by also considering the fine grained information on human movements available in Foursquare. We will see how location-based social networks provide a unique opportunity to assess a geographic area not only by considering static spatial information, but also the dynamics of the movements of mobile users. In more detail, we study users' transitions between places inferred from their consecutive check-ins in different venues. As we show below, the aggregated transition data can be effectively exploited to analyze the flows of the users towards the place of interest and generally in the surrounding neighborhood. In Figure 3 we plot the Cumulative Distribution Function (CDF) of incoming transition distances, i.e., distances that users travel from other venues in the dataset towards the three chains under analysis. As we observe in the plots, the vast majority of the incoming transitions, i.e., $80 - 90\%$, are done from within a radius of less than 1km and 50% of transitions from 200-300m. This result suggests that the customers of Starbucks coffee shops, Dunkin' Donuts and McDonald's restaurants usually come from local places. On the other hand the analysis of user transitions can also characterize interactions between Foursquare venues located at the different corners of the city: this information is not captured by the place co-location approach discussed in the previous paragraph.
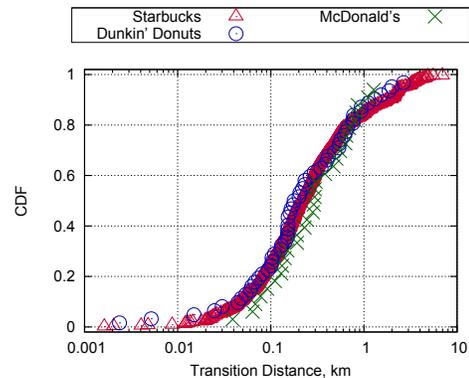

Figure 3: The Cumulative Distribution Function (CDF) of the transition distances towards the shops of the three chains under analysis.

Starbucks		Dunkin' Donuts		McDonalds	
Hostel	17.44	Convention Center	8.18	Parks & Outdoor	15.76
Flea Market / Fair	7.28	Laundry	5.51	Sculpture	6.10
Sculpture	6.80	Post Office	3.22	School	4.84
Drug Store	4.96	Mall	2.20	Light Rail	3.94
Train Station	4.79	Playground	1.49	Bus Station	3.40
Bank / Financial	4.32	Drug Store	1.43	Post Office	2.70
Post Office	4.01	Subway Station	1.30	Plaza / Square	2.20
Technopark	3.76	Bank / Financial	1.26	Bank / Financial	2.18
Admin. Building	3.64	School	1.18	Airport	2.09
Convention Center	3.39	Gas Station	1.09	Theater	1.87

Table 3: Categories with the highest transition ratio ρ towards the three chains under analysis.

We further measure the probability of transitions between different types of venues to determine the main donors of users towards the three chains under analysis (we discuss the details in Section 3). The transition ratio ρ is defined as the ratio between the transition probability of a given pair of venue categories over the random transition probability between any pair of categories. The Top-10 place categories with the highest transition ratio towards the three chains under analysis are presented in Table 3. Numerous similarities are observed with respect to the inter-category attractiveness presented in Table 2. The main attractors that we have identified by measuring the frequency of pairwise category co-location are often the main sources of customers. Thus, Hostels, Train Stations and Financial Centers, which had one of the highest attractiveness coefficients towards Starbucks restaurants, have also the highest transitional ratio towards them, i.e., 17.44 (for Hostels), 4.79 (for Train Stations) and 4.32 (for Financial Centers) times higher than in a random scenario. The same observation holds for the attractiveness towards Schools and Subway Stations, in the case of Dunkin' Donuts. However, a deviation between the two measurements is also possible. The fact that a place type tends to be co-located with another may not necessarily mean that movement occurs across the two types of places. Thus, considering information about user flows between venues can provide additional value to tackle the optimal retail store placement problem as will also become apparent during evaluation in Sections 4 and 5.

3. OPTIMAL RETAIL STORE PLACEMENT IN LOCATION-BASED SERVICES

In this section we formalize the problem of optimal retail store placement in the context of location-based social networks. Our goal is to identify the best area amongst a candidate set of potential areas for a new store to be opened. After formulating the problem, we define and discuss the features we have mined from the Foursquare dataset to predict the best geographic spots by ranking geographic areas according to the predicted retail quality.

Problem Formulation

Formally, by considering the existence of a candidate set of areas L in which a commercial enterprise is interested in placing their business, we wish to identify the *optimal* area $l \in L$, such that a newly open store in l will potentially attract the largest number of visits. An area l is encoded by its latitude and longitude coordinates and a radius r, as depicted in Figure 4. We have experimented with different values for the radius r and have selected it to be equal to 200 meters as this has yielded the best experimental results, i.e., the highest prediction performance across independent experiments with various values of r, and it is also in agreement with what the urban planning community considers as the optimal neighborhood size [16]. The ranking of places according to their popularity is then estimated using the features mined by incorporating the char-

acteristics of the area nearby (features are described next). For each feature we compute a score $\hat{\chi}_l$ for every candidate area l: the top ranked area in terms of that score will be the optimal area for the new store placement. Our main assumption in the formulation of this task is that the number of empirically observed check-ins by Foursquare users can be used as a proxy for the relative popularity of a place.
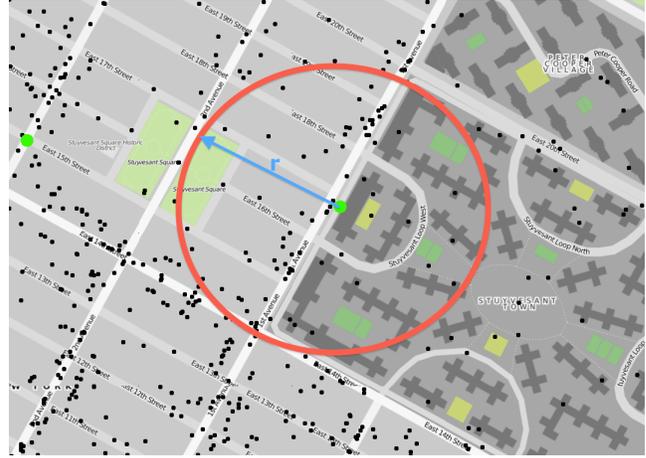

Figure 4: Area of radius $r = 200$ meters around a Starbucks coffee shop in New York. Black dots correspond to nearby Foursquare venues.

Prediction Features

We now introduce the features we have mined from the Foursquare dataset in the city of New York. Each feature returns a numeric score $\hat{\chi}_l$ that corresponds to a quality assessment of the area for the optimal retail store placement problem. We have categorized the extracted features into two broad categories: *place-geographic* features which integrate information about the types and spatial interaction between the places, and *user mobility* features which exploit knowledge about user movements and transitions between places.

Geographic Features

This class refers to features that describe the environment around the place under prediction as encoded through the spatial distribution of Foursquare venues. More specifically, we measure the density, heterogeneity and competitiveness of the surrounding area by analyzing the set $\{p \in P : dist(p, l) < r\}$ of places that lie in a disk of radius r around location l. The function $dist$ denotes the geographic distance between two places and P the set of venues in New York.

Density: By measuring the number of neighbors around the place we assess to what extent the popularity of a place depends on the density of other places in the same area. Formally:

$$\hat{\chi}_l(r) = |\{p \in P : dist(p, l) < r\}|$$

We note that given that the radius r is the same, the density of candidate geographic areas depends only on the number of places they include. We denote the number of neighbors of a place within a radius r with $N(l, r)$. Intuitively, a denser area could imply higher likelihood for an opportunistic visit to a retail facility.

Neighbors Entropy: To assess the influence of the spatial heterogeneity of the area on the popularity of a place, we apply the

entropy measure from information theory [7] to the frequency of place types in the area. We denote the number of place neighbors of type γ with $N_\gamma(l, r)$. The entropy defines how many bits are required to encode the corresponding vector of type counters $\{N_\gamma(l, r) : \gamma \in \Gamma\}$, where Γ is a set of all types, and is higher the more heterogeneous the area is:

$$\hat{\chi}_l(r) = -\sum_{\gamma \in \Gamma} \frac{N_\gamma(l, r)}{N(l, r)} \times \log \frac{N_\gamma(l, r)}{N(l, r)}$$

In general, an area with high entropy values is expected to be diverse in terms of types of places, whereas the least entropic areas would imply that the number of check-ins is biased towards a specific category, for instance *Home* if the area is residential.

Competitiveness: We devise a feature to factor in the competitiveness of the surrounding area. Given the type of the place under prediction γ_l (for example Coffee Shop for Starbucks), we measure the proportion of neighboring places of the same type γ_l with respect to the total number of nearby places. We then rank areas in reverse order, assuming that the least competitive area is the most promising one:

$$\hat{\chi}_l(r) = -\frac{N_{\gamma_l}(l, r)}{N(l, r)}$$

It is worth noting however that competition in the context of retail stores and marketing can have either a positive or a negative effect. One would expect that, for instance, placing a bar in an area populated by nightlife spots would be rewarded as there is already an ecosystem of related services and a crowd of people being attracted to that area. However, being surrounded by competitors may also mean that existing customers will be shared.

Quality by Jensen: To consider spatial interactions between different place categories, as we anticipated in Section 2, we exploit the metrics defined by Jensen et al. in [10]. To this end, we use the inter-category coefficients described in the previous section to weight the desirability of the places observed in the area around the object, i.e., the more the places that attract the object exist in the area, the better the quality of the location. More formally, we define the quality of location for venue of type γ_l as:

$$\hat{\chi}_l(r) = \sum_{\gamma_p \in \Gamma} \log(\kappa_{\gamma_p \to \gamma_l}) \times (N_{\gamma_p}(l, r) - \overline{N_{\gamma_p}(l, r)})$$

where $\overline{N_\gamma(l, r)}$ denotes how many venues of type γ_p are observed on average around the places of type γ_l, Γ is the set of place types, and $\kappa_{\gamma_l \to \gamma_p}$ are the inter-type attractiveness coefficients. To compute the latter, we analyze how frequently places of type γ_l are observed around γ_p on average, and normalize that value with the expectation for a random scenario. Formally we get:

$$\kappa_{\gamma_p \to \gamma_l} = \frac{N - N_{\gamma_p}}{N_{\gamma_p} \times N_{\gamma_l}} \sum_p \frac{N_{\gamma_l}(p, r)}{N(p, r) - N_{\gamma_p}(p, r)}$$

where N, N_{γ_l} and N_{γ_p} denote the total number of places considered in the analysis and the number of places for types γ_l and γ_p respectively. Similarly, the intra-categories coefficient $\kappa_{\gamma_l \to \gamma_l}$ are computed between the places of the same type, thus, assessing to which extent the places tend to group into spatial clusters, e.g., financial centres at Wall Street.

Mobility Features

In this section we show how information about the check-in patterns of Foursquare users can be exploited to assess the retail quality of an area. Our goal is to identify to what extent information crowdsourced from mobile users can improve geographic business analytics and what are the benefits with respect to information exploiting only static spatial information about venues such as the features presented above. We shall consider characteristics that measure the general popularity of the area and features that exploit transitions amongst venues.

Area Popularity: To assess the influence of the overall popularity of the area on the popularity of individual places we measure the total number of check-ins empirically observed among the neighboring places in the area:

$$\hat{\chi}_l(r) = |\{(m, t) \in C : dist(m, l) < r\}|$$

where tuple (m, t) denotes a check-in recorded in place $m \in P$ at time t, and C is a set of all check-ins in the dataset.

Transition Density: Assuming that increased mobility between places in the area can increase the number of random visitors towards the target place, we measure the density of transitions between the venues inside the area. Formally, by denoting as T the total set of consecutive check-in transitions between places and as a tuple, $(m, n) \in T$, the places $m \in P$ and $n \in P$ involved in two consecutive check-ins, we have:

$$\hat{\chi}_l(r) = |\{(m, n) \in T : dist(m, l) < r \wedge dist(n, l) < r\}|$$

Incoming Flow: We also define a feature to account for the incoming flow of external user traffic towards the area of the place in question. We consider transitions between places denoted by a tuple, $(m, n) \in T$, such that first place m is located outside and second place n inside the area under prediction. Formally:

$$\hat{\chi}_l(r) = |\{(m, n) \in T : dist(m, l) > r \wedge dist(n, l) < r\}|$$

One would expect that an area that is a good attractor of remote users would be a promising one for establishing a new retail facility.

Transition Quality: Another aspect of location attractiveness comes from the potential number of local customers that the place might attract from the area. We measure the probability of transitions between all other types of venues and venues of the same type as the venue itself. The resulting probabilities allow us to weight the nearby places as potential sources of customers to the place under prediction. More formally:

$$\hat{\chi}_l(r) = \sum_{\{p \in P : dist(p, l) < r\}} \sigma_{\gamma_p \to \gamma_l} \times C_p$$

where C_p is the number of check-ins at a place p and $\sigma_{\gamma_p \to \gamma_l}$ is a probability of transitions between two categories γ_p and γ_l, defined as the average percentage of all check-ins to place p of category γ_p that are followed by transitions to places of category γ_l:

$$\sigma_{\gamma_p \to \gamma_l} = \mathrm{E}\left[\frac{|\{(m, n) \in T : m = p \wedge \gamma_n = \gamma_l\}|}{C_p}\right]$$

$$\rho_{\gamma_p \to \gamma_l} = \sigma_{\gamma_p \to \gamma_l} \frac{N - N_{\gamma_p}}{N_{\gamma_p} \times N_{\gamma_l}}$$

Having defined the two classes of geographic and mobility features, covering a diverse set of signals that can be exploited in location-based services, we evaluate their performance in the next sections aiming to understand what sources of information form the best predictors of the popularity of a retail chain unit.

4. FEATURE EVALUATION

As discussed, a primary aim of this paper is to identify the most important information signals that can be mined from location-based services, in order to forecast the area where a retail store may attract the maximum number of check-ins. To this end, we evaluate the performance of the individual prediction features introduced in the previous section. We first describe the methodology and the metrics we employ; then we compare the performance of each individual feature across these metrics.
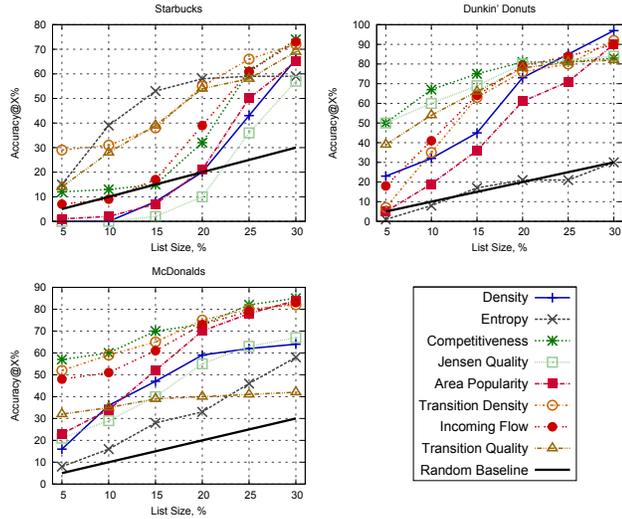

Figure 5: The Accuracy@X% scores of the individual feature predictions for Starbucks, Dunkin' Donuts and McDonalds.

Methodology and Metrics

Given a specific retailer brand like the three we are considering in this paper, we analyze a set of prediction areas L for a new store. For each feature and area $l \in L$ we compute χ_l as defined in the previous section and then rank the locations based on this. As a result we obtain a ranked list of locations $R = (l_1, l_2, ..., l_{|L|})$ and denote with $rank(l_i)$ the position of location l_i in R. We also compute the ranked list $\overline{R}$ of locations based on the actual popularity (number of check-ins) of the stores in those locations and denote with $\overline{rank}(l)$ the position of location l in $\overline{R}$. Given the two lists ranking locations in terms of predicted and ground truth values, we formally define the metrics we use to assess the quality of predictions achieved by the different features.

NDCG for top-K location ranking: Firstly, we aim to measure the extent to which the top-k locations in the list of actual popularity $\overline{R}$ are highly ranked in the predicted list R. To this end, we adopt the NDCG@k (Normalized Discounted Cumulative Gain) metric frequently used in the performance evaluation of information retrieval systems [9]. The metric assesses the cumulative gain achieved by placing the most relevant instances in the top-k of the prediction list as formally defined by the Discounted Cumulative Gain measure:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(l_i)} - 1}{\log_2 (i + 1)} \qquad (1)$$

where $rel(l_i)$ is the score relevance of an instance at position i in the predicted ranking R. The result is then normalized by the DCG

Feature	Starbucks	Dun.Don.	McDon.
Geographic Features			
Density	0.60	0.79	0.73
Entropy	0.65	0.60	0.69
Competitiveness	**0.70**	0.68	**0.78**
Jensen Quality	0.54	**0.81**	0.72
Mobility Features			
Area Popularity	0.54	0.77	0.78
Transition Density	0.62	0.79	0.78
Incoming Flow	0.60	0.75	**0.79**
Transition Quality	**0.66**	**0.81**	0.73
Random Baseline			
	0.48	0.51	0.53

Table 4: The Average NDCG@10 results of the individual feature predictions for each of the three chains.

of the ideal prediction (iDCG), when the instances are sorted by the relevance with the most relevant in the first position. The resulting scores, thus, lie in the range from $0 < $ NDCG@$k \leq 1$. As a relevance score for an instance l_i we use its relative position in the actual ranking $\overline{R}$, i.e., $rel(l_i) = \frac{|L| - \overline{rank}(l_i) + 1}{|L|}$. The $rel(l_i)$ score is equal to 1 when the area is ranked first in terms of check-ins and it linearly decreases to 0 as the rank goes down the list. As a baseline for comparison, we use the expected value of NDCG@k for a random ranker which is achieved by randomly permuting the instances in the testing set.

Assessing the best prediction: Considering now the application scenario where the best geographic area for a new business has to be discovered, for instance by a geo-analytics team, we would like to compare the different ranking strategies in terms of their ability to yield high quality top locations. To this end, we measure the fraction of times that the optimal location in the predicted list R is at the top-X% of the the actual popularity list $\overline{R}$ which represents our ground truth. We refer to this metric as Accuracy@X% and we note that we have used the % instead of absolute list size values (i.e., top-K) to allow for comparison across different chains.

Geographic Cross Validation: We use a random sub-sampling method [15] to select subsets of geographic areas for validation. In each experiment (repeated here 1000 times) we randomly sample 33% of the total areas associated with the stores of the brand we analyze, obtaining L candidate areas for prediction. The rest of the areas are subsequently used to form our training set on which features (or supervised learning algorithms as we see in Section 5) will be trained. We assume that the stores (and the corresponding check-ins) in the test set areas L do not yet exist and our goal is to forecast the popularity ranking of the stores, if they were to be opened there. We note that the same assumption is used to define the features in the training and test sets. The mean NDCG@k scores are measured by averaging across all testing sets.

Individual Feature Performance

NDCG Top-K location ranking: In Table 4, we present the results obtained for the NDCG@10 metric for all features across the three chains. In all cases we observe a significant improvement with respect to the random baseline, yet there are features which perform considerably better than others. In particular, with regards to the geographic class of features, the Jensen Quality does best with NDCG@10 = 0.81 for Dunkin' Donuts, whereas Competitiveness

is the top feature for Starbucks and McDonald's scoring 0.70 and 0.78 respectively. Interestingly, as suggested by the Competitiveness feature, the lack of competitor venues in an area can have a positive influence in attracting customers. A potential interpretation of this observation is that an urban area, for instance a neighborhood, is expected to provide a number of services to its local residents. A retail facility can benefit by being the main provider at an area whereas, in contrast, its customer share is expected to drop as more and more competitors are opening nearby.

When considering the class of mobility features, Transition Quality does best in the cases of Starbucks and Dunkin' Donuts with an NDCG@10 score of 0.66 and 0.81. This signifies that knowledge about the types of places that can initiate large customer flows to a target business can be useful to the placement of a new shop. In the case of McDonald's, however, the most effective feature is Incoming Flow which has achieved an NDCG@10 score of 0.81. A careful inspection of Figure 3 suggests that this may be due to the fact that McDonald's tend to attract customers from remote locations with higher probability than the other chains. In the case of McDonald's the probability of attracting a Foursquare user from a distance beyond 200 meters is almost 0.65, when the value drops to 0.50 and 0.45 for Starbucks and Dunkin' Donuts respectively.

Another observation which stands out is that in the case of Dunkin' Donuts, Jensen Quality and Transition Quality achieve the same performance (NDCG=0.81), better than the rest of the features. This is a case where geographic and mobility features agree in performance revealing that spatial structure alone is sufficiently correlated with the mobility of users in the area. A comparison between Tables 2 and 3 shows that the rankings of the different types of places across the three chains agree more for Dunkin' Donuts compared to Starbucks and McDonald's, where larger deviations may signify a large discrepancy between the geographic and mobility properties of an area.

Assessing the best prediction: We now evaluate the individual feature predictors in terms of the Accuracy@$X\%$ metric. As defined above, this metric considers the top predicted location $l \in L$ and measures how high it is ranked in the list of actual rankings $\overline{R}$. For small list fractions 5% the relative performance of the various features is qualitatively similar to the NDCG metric. While the ranking in the performance of different features persists across different list $X\%$ values, it is worth mentioning that Competitiveness tends to rise faster than other features reaching a score of Accuracy@15% higher than 70%, outperforming the winning (in terms of NDCG) Incoming Flow for McDonald's and Jensen Quality for Dunkin' Donuts. The results in terms of Accuracy@$X\%$ are much lower in the case of Starbucks. An explanation of the drop in performance (also observed for NDCG) for Starbucks may relate to their high density in the city: as shown in Table 1 (and hinted by observing Figure 1), the number of Starbucks stores is almost two times higher than McDonald's and Dunkin' Donuts in the same 100km^2 area around Manhattan. Thus, the potential geographic overlap of the areas covered by two Starbucks picked randomly is higher and, therefore, while the underlying features may be very similar their corresponding popularity may differ. To provide a numerical indication of this issue we note that the probability of encountering a store of the same chain within 100 meters was 35% for Starbucks, whereas only 11% and 7% for McDonald's and Dunkin' Donuts, respectively.

5. SUPERVISED LEARNING APPROACH

In this section we combine features in a supervised learning framework. Our aim is to exploit the union of individual features in or-

Algorithm	Starbucks	Dun.Don.	McDon.
Geographic Features			
Lin. Regression	0.73	0.80	0.78
M5 Dec. Trees	0.72	0.80	0.78
SVR	0.73	0.81	0.81
RankNet	0.72	0.81	0.79
All Features			
Lin. Regression	0.76	0.82	0.82
M5 Dec. Trees	0.77	0.82	0.82
SVR	0.77	0.83	0.84
RankNet	0.77	0.81	0.83
Random Baseline			
	0.48	0.51	0.53

Table 5: The best Average NDCG@10 results of the supervised learning algorithms applied to combinations of features grouped into two classes: combinations of geographic features only and combinations of both mobility and geographic features.

der to improve predictions, testing if the popularity of places in Location-based Social Networks can be better predicted by considering a composition of signals. We use different supervised models to learn how feature vectors $\mathbf{x}$ can be associated with the check-in scores y of the areas under prediction. Two different ranking methodologies are employed which we explain in the following paragraphs.

Supervised Regression for Ranking: The three algorithms we exploit here are Support Vector Regression [8], M5 decision trees [20] and Linear Regression with regularization. The latter case assumes that the output score y is a linear combination of the input features $\mathbf{x}$ with the weights $\mathbf{w}$ being calculated from the training data. The goal of the prediction algorithm is to minimize the error between actual and predicted outputs:

$$\min_{\mathbf{w}} \|\mathbf{x}^{\mathsf{T}}\mathbf{w} - y\|^2 + \gamma\|\mathbf{w}\|^2 \qquad (2)$$

where γ is the regularization parameter set here equal to 10^{-8}. We have used the corresponding implementations that are publicly available through the WEKA machine learning framework [21]. By training the supervised learning algorithms to obtain regression scores and, subsequently, rank the candidate geographic areas, we are essentially reducing the regression problem to a ranking one.

Supervised Learning to Rank: Additionally, we consider a pairwise learning-to-rank approach, RankNet [6], that learns the ordering relation between a pair of venues based on their features. Given a pair of venues A and B, characterized by features $\mathbf{x^A}$ and $\mathbf{x^B}$, RankNet identifies if venue A has to be ranked higher than B. The models assume a ranking function $H(\mathbf{x}) : \Re^{|\mathbf{x}|} \to \Re$ such that the rank order of instances is specified by the real values of H. Specifically, $H(\mathbf{x^A}) > H(\mathbf{x^B})$ assumes that venue A has to be ranked higher than venue B. As the ground truth in this model we employ a linear ranking $\overline{R}$ of test set venues according to the number of their check-ins. We finally note that in the RankNet algorithm the ranking function $H(x)$ is modeled as a neural network and we use a publicly available implementation of the algorithm from the RankLib library[3]. Overall, the testing and evaluation of the supervised algorithms is conducted using the exact methodology we have used for the evaluation of individual features, and all features have been normalized before training.

---

[3] `http://people.cs.umass.edu/~vdang/ranklib.html`

Results We now present the results of supervised learning using the algorithms described above. We measure the performance of the algorithms trained separately on different combinations of geographic features and then combining them with mobility features. We aim to assess the extent to which the fusion of the features yields better results than the prediction based on individual features and secondly, to what extent adding mobility features improves predictability based only on geographical information. The supervised models based on the grouping of geographic features offer a slight increase (0.02-0.03) in NDCG@10 performance in comparison to individual feature prediction in two out of three cases: McDonald's and Starbucks. However, adding mobility features to the supervised models considerably improves the prediction results (Table 5) across all chains. For example when using supervised learning in the case of Starbucks an NDCG@10 score of 0.77 is reached an improvement in the performance of Transition Quality, our best single feature prediction, by 10%. In the Dunkin' Donuts and McDonald's cases the best NDCG@10 in supervised learning reaches 0.83 and 0.84 in comparison to the best single feature performance that achieves 0.81 and 0.79, respectively. The improvement in performance when supervised learning is applied is more clearly indicated when taking into account the Accuracy@X% measurement. The top predicted geographic location for Starbucks, as shown in Figure 6, is positioned with an accuracy of 67% and 76% amongst the Top-10% and the Top-15% in the list $\overline{R}$ of the most popular places. When considering the Top-20% most popular places, accuracy goes over the mark of 80% across all considered chains, reaching a maximum of 93% in the case of Dunkin' Donuts. These results significantly outperform the best individual features predictions which are 58% for Starbucks, 75% and 81% for McDonald's and Dunkin' Donuts respectively.
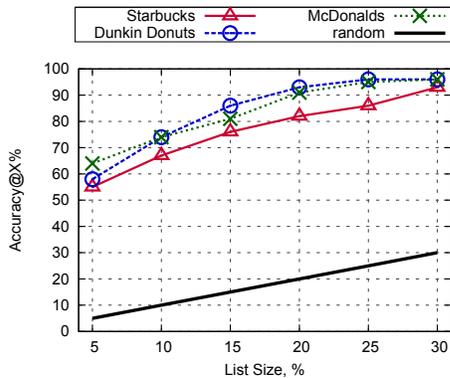

Figure 6: The Accuracy@X% results of the best performing combination of features for the three store chains.

Overall, the combination of geographic and mobility features in a supervised learning algorithm yields better results than the single feature predictions. Moreover, when considering mobility features in addition to the geographic features we have observed large improvements, highlighting that information about the check-in movements of Foursquare users adds value to the prediction. Across different prediction techniques, we have observed that the optimal performance is achieved for the predictions based on the Support Vector Regression algorithm.

6. IMPLICATIONS

The analysis and evaluation presented in the previous sections highlight how the new generation of location-based services can play a significant role in the commercial evolution of the web. The rich spatio-temporal datasets sourced from services like Foursquare can provide access to new layers of data bringing us one step towards the integration of local businesses in the economic model of social networking platforms. As we have seen in our experiments, knowledge on the semantics of geo-tagged venues can provide effective data representations to model the commercial value of urban areas allowing us to measure the influence of competitive retail facilities nearby or the presence of types of venues that can generate a large flow of customers for a target business.

The fact that different features may prove more or less effective across chains signifies that the problem of optimal retail store placement is not trivial. Different types of businesses can demonstrate significant variations in the ways they become attractors to their respective customer crowds. As empirically studied in Section 2 the spatial properties and the patterns of customer movement can be chain specific, in spite of the fact that common patterns of attraction are also observed. As we have shown, supervised learning classifiers can exploit a number of data mining features and seamlessly alleviate the biases due to heterogeneities in the way retail facilities attract users. Moreover, the combination of geographic and mobility features in the majority of cases has yielded superior performance suggesting that the dynamics of human movement matter in understanding the retail quality of an area.

Besides, is especially clear in the case of Starbucks, noise can impede the prediction task. One factor that has contributed to this is the fact that very proximate retail facilities of the same chain will have by definition very similar feature values, although, at the same time, their popularity may vary significantly. Other factors such as local architecture and planning idiosyncrasies of an area can effect the flow of customers towards a place. Especially, in extremely dense cities, like New York, two venues can have similar latitude and longitude coordinates but one of them may be placed at the corner of a high street and the other at the top of a skyscraper nearby. Along these lines, we are planning to extend our work to include more cities and chains. Informal experiments we have conducted suggest the existence of large heterogeneities in the spatial properties of different cities, but also a strong biases in the amount of user check-ins from area to area. Therefore, the development of techniques that deal with these issues is a challenging task with potentially high value for location-based services and urban data mining in general.

7. RELATED WORK

The retail store placement problem has, in recent decades, attracted researchers from a broad spectrum of disciplines. Land economy community research has concentrated on spatial interaction models, which are based on the assumptions that the intensity of interaction between two locations decreases with their distance and that the usability of a location increases with the intensity of use and the proximity of complementary arranged locations [3, 5, 12]. It was shown, however, that the applicability of these models is limited to agglomerations, such as big shopping centres, and their predictive accuracy decreases when smaller, specialized stores are considered. With respect to previous work in the general area, in this paper we examined how the problem can be framed in terms information signals mined from location-based services. As we have seen, the richness of information provided in these services could help us to study the retail quality of an area in a fine grained manner: various types of geographic, semantic and mobility information not only can complement traditional techniques, but also form the basis for a new generation of business analytics driven by services such as Foursquare.

The present paper has been largely inspired by Jensen's seminal work [10, 11] on the identification of the appropriate geographic positioning of retail stores. Jensen's approach uses a spatial network based formulation of the problem, where nodes are 55 different types of retail stores and weighted signed links are defined to model the attraction and repulsion of entities in the network. Subsequently, a retail quality index is devised which has been used to empirically assess the positioning of stores in the city of Lyon, France. Another approach based on the analysis of the spatial distribution of commercial activities was proposed by Porta et al. in [17, 18]. The authors investigate the relationship between street centrality and retail store density in the cities of Bologna and Barcelona respectively, verifying how the former acquires a significant role in the formation of urban structure and land usage. We extend the results of these works by adding to the analysis features mined from the human mobility traces and effectively show that the combination of the geographic and mobility features provides better insights on the quality of an area as a potential spot to open a new retail facility. To our knowledge this is also one of the first efforts to tackle the problem exploiting machine learning algorithms. These techniques may prove crucial in doing similar analyses in the future, as they have been devised to operate in settings where large amounts of dynamic data are available.

Finally, from a data mining perspective, we could classify our work in the area of *urban mining*, which studies the extraction of knowledge from spatial or geographic datasets and aims to improve services and intelligence in the city. In [19], Quercia et al. mine cellular data of user movements in the city of Boston to recommend social events. In [13] the authors analyze the movement of passengers of the London metro and provide insights into the financial spending of transport users, and in [14] the relationship between social deprivation on citizen mobility is investigated. In [22] the authors attempt to infer the functions of different regions in the city by analyzing spatial distribution of commercial activities and human mobility traces in the city of Beijing. The present paper is well aligned with this stream of work and extends the applicability of the urban mobility mining methods to the field of retail analytics.

8. CONCLUSION

In this paper we tackled the problem of optimal retail store placement in the context of location-based social networks. We collected human mobility data from the leading location-based service, Foursquare, and analyzed it to understand how the popularity of three retail store chains in New York is shaped, in terms of number of check-ins.

We developed and evaluated a diverse set of data mining features, modeling spatial and semantic information about places and patterns of user movements in the surrounding area. We evaluated each feature separately and found that, among those exploiting place semantics, the presence of user attractors (i.e., train station or airport) as well as retail stores of the same type to the target chain (i.e. coffee shop or restaurant) encoding the local commercial competition of an area, are the strongest indicators of popularity. However, additional improvement in the prediction performance may be achieved by assessing potential flows of users that a place may attract from other venues nearby or far away locations. We further combined different features in a set of supervised learning algorithms and showed that the popularity of places can be better explained by the fusion of geographic and mobility features. We plan to extend our work to the comparison of the impact of features in different cities and in new types of places and retail chains.

10. REFERENCES

[1] Everything you need to know about foursquare's new merchant tools. http://goo.gl/3DhfT.

[2] Foursquare at 35 million users. 2013. http://goo.gl/ThfnF.

[3] A. Athiyaman. Location decision making: the case of retail service development in a closed population. In *Academy of Marketing Studies*, volume 15, page 13, 2010.

[4] J. Bao, Y. Zheng, and M.F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *ACM SIGSPATIAL GIS*, 2012.

[5] O. Berman and D. Krass. The generalized maximal covering location problem. *Computers & Operations Research*, 29(6):563–581, 2002.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.

[7] T.M. Cover, J.A. Thomas, and J. Wiley. *Elements of information theory*, volume 6. Wiley Online Library, 1991.

[8] M.A. Hearst, ST Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[10] Pablo Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Phys. Rev. E*, 74:035101, Sep 2006.

[11] Pablo Jensen. Analyzing the localization of retail stores with complex systems tools. In *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII*, IDA '09, pages 10–20, Berlin, Heidelberg, 2009. Springer-Verlag.

[12] A. Kubis and M. Hartmann. Analysis of location of large-area shopping centres. a probabilistic gravity model for the halle-leipzig area. *Jahrbuch für Regionalwissenschaft*, 27(1):43–57, 2007.

[13] Neal Lathia and Licia Capra. Mining mobility data to minimise travellers' spending on public transport. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1181–1189, New York, NY, USA, 2011. ACM.

[14] Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: sensing community well-being from urban mobility. *Pervasive Computing*, pages 91–98, 2012.

[15] O.Z. Maimon and L. Rokach. *Data mining and knowledge discovery handbook*. Springer US, 2010.

[16] M. Mehaffy, S. Porta, Y. Rofè, and N. Salingaros. Urban nuclei and the geometry of streets: The 'emergent neighborhoods' model. *Urban Design International*, 15(1):22–46, 2010.

[17] S. Porta, V. Latora, F. Wang, S. Rueda, E. Strano, S. Scellato, A. Cardillo, E. Belli, F. Càrdenas, B. Cormenzana, et al. Street centrality and location of economic activities in barcelona. *Urban Studies*, 2011.

[18] S. Porta, V. Latora, F. Wang, E. Strano, A. Cardillo, S. Scellato, V. Iacoviello, and R. Messora. Street centrality and densities of retail and services in bologna, italy. *Environment and Planning B: Planning and Design*, 36(3):450–465, 2009.

[19] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Proceedings of ICDM '10*, pages 971 –976, dec. 2010.

[20] J.R. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian joint Conference on Artificial Intelligence*, pages 343–348. Singapore, 1992.

[21] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[22] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.