If I build it, will they come? Predicting new venue visitation patterns through mobility data

Krittika D'Silva
University of Cambridge
krittika.dsilva@cl.cam.ac.uk

Anastasios Noulas
New York University
noulas@nyu.edu

Mirco Musolesi
University College London
m.musolesi@ucl.ac.uk

Cecilia Mascolo
University of Cambridge
cecilia.mascolo@cl.cam.ac.uk

Max Sklar
Foursquare Labs
max@foursquare.com

ABSTRACT

Estimating revenue and business demand of a newly opened venue is paramount as these early stages often involve critical decisions such as first rounds of staffing and resource allocation. Traditionally, this estimation has been performed through coarse measures such as observing numbers in local venues. The advent of crowdsourced data from devices and services has opened the door to better predictions of temporal visitation patterns for locations and venues. In this paper, using mobility data from the location-based service Foursquare, we treat venue categories as proxies for urban activities and analyze how they become popular over time. The main contribution of this work is a prediction framework able to use characteristic temporal signatures of places together with k-nearest neighbor metrics capturing similarities among urban regions to forecast weekly popularity dynamics of a new venue establishment. Our evaluation shows that temporally similar areas of a city can be valuable predictors, decreasing error by 41%. Our findings have the potential to impact the design of location-based technologies and decisions made by new business owners.

CCS CONCEPTS

• **Database applications** → *Data mining*; • **Social and Behavioral Science** → *Miscellaneous*;

KEYWORDS

Human mobility prediction, urban traffic, spatio-temporal patterns, urban computing

1 INTRODUCTION

Cities are complex systems that constantly change over time. The way in which neighborhoods become popular over time has been a fundamental area of study in traditional urban studies literature as it is critical to city governance [2]. The rise of mobile technologies and collective sensing in the last decade has contributed to the generation of large datasets that describe activity dynamics in cities and has created new opportunities for research in the area [3, 4, 9, 12]. Nevertheless, little work has focused on predicting important properties of a new business. In this paper, we provide an analytical framework that captures the popularity dynamics of urban neighborhoods. We begin with a temporal characterization of urban activities across regions, showing how the popularity dynamics of venue categories give rise to the temporal patterns of the urban areas that contain them. We highlight how urban activities and population levels in a neighborhood are inherently interconnected temporal processes and then exploit these temporal patterns across areas to predict the popularity dynamics of newly established venues. Our work enables a fine-grained dynamic estimation of activity for new venues and provides analytics which can help plan the provision of services to customers.

SIGSPATIAL'17, November 7-10 2017, Los Angeles Area, CA, USA
 978-1-4503-5490-5/17/11…$15.00
DOI: 10.1145/3139958.3140035

2 RELATED WORK

The rise of geo-tagging applications, such as Foursquare, Twitter, and Flickr, combined with the accessibility to their corresponding APIs has led to more granular representations of urban activities across time and space and time [7]. These findings have also helped inform location-based technologies. For example, Foursquare exploited weekly temporal visitation patterns of venues to power its local search engine [13] and Google Maps incorporated features to inform users of popular times [8]. Additionally, data from cellular networks have helped to understand the collective dynamics of urban activities. Ratti et al. in [12] presented one of the first works that demonstrated how urban landscapes transform in real time as populations move around the city. Beyond dynamic visualizations, Calabrese et al. in [4] provided interpretations of observed mobility patterns in terms of the underlying urban activities which drive population volumes, such as transport and residential land uses. In [3] Becker et al. used cellular data to characterize mobility trends across different metropolitan areas, while Jiang et al. in [9] proposed using cellular data as an alternative to travel surveys for more accurate spatio-temporal representations of mobility flows.

3 NOTATION AND DEFINITIONS

3.1 Dataset

We use a longitudinal dataset describing urban mobility and activity patterns in Greater London from Foursquare. For each venue, our data set contains the geographic coordinates, specific and general category, creation date, total number of check-ins, and unique number of visitors. The specific and general categories fall within Foursquare's API of hierarchical categories which can be found by

querying the Foursquare API [1]. The dataset also contains "transitions" which are defined as a pair of check-ins to two different venues within the span of three hours. A transition is identified by a start time, end time, source venue, and destination venue. Our dataset includes 18,018 venues and 4,000,040 transitions for Greater London from December 2010 to December 2013.

3.2 Formalization

In this section, we introduce a formalization of our model. Electoral wards are the main building blocks of administrative geography in the United Kingdom. Greater London consists of 649 electoral ward and these spatial units uniquely identify London boroughs [6]. We use wards $w \in \mathbf{W}$ as a means of subdividing Greater London. We also consider venues $v \in \mathbf{V}$. A venue has a precise geographic location in a ward. A venue v is represented with a tuple $v =< loc, g_v, s_v >$ where loc is the geographic location of the venue, g_v is its general category and s_v is its specific category.

We define a *time interval* t as the interval $[t\Delta, (t+1)\Delta]$ of duration Δ. For example, the time interval $t = 0$ indicates the interval $[0, \Delta]$, the time interval $t = 1$ indicates the interval $[\Delta, 2\Delta]$ and so on.

Definition 1: Temporal Profile of a Ward. Similarly, we define the temporal profile of a ward w in an interval $[0, T]$ as the following sequence (i.e., time series):

$$C^w[0,T] = \{c_t^w\} \quad \text{with} \quad t = 0, 1, ..T-1 \tag{1}$$

where c_t^w is the *total* number of check-ins in the ward w during the time interval t.

Definition 2: Temporal Profile of a Venue. We define the *temporal profile of a venue* v in an interval $[0, T]$ as the following sequence (i.e., time series):

$$C^v[0,T] = \{c_t^v\} \quad \text{with} \quad t = 0, 1, ..T-1 \tag{2}$$

where c_t^v is the *total* number of check-ins to venue v during the time interval t.

Definition 3: Aggregate Temporal Profile of Venues of a Generic (Specific) Category in a Ward. We then define $\mathbf{V}_{g,w}$ as the set of the venues of *generic* category g in a ward w. Similarly, we define $\mathbf{V}_{s,w}$ as the set of the venues of *specific* category g in a ward w. Therefore, the *aggregate temporal profile* of venues of *generic* category g in a ward w in a time interval $[0, T]$ is defined as the following sequence (i.e., time series):

$$C^{V_{g,w}}[0,T] = \{c_t^v\} \quad \text{with} \quad t = 0, 1, ..T-1 \quad \text{and} \quad v \in V_{g,w} \tag{3}$$

where c_t^v is the *total* number of check-ins in the ward v during the time interval t, but as it can be seen in the formula, we also set the condition that v is a venue of general category g in the ward w under consideration. The temporal profile of venues of a specific category in a ward can be defined in a similar way.

4 TEMPORAL PATTERNS OF USER ACTIVITY

4.1 Regional temporal activity patterns

Figure 1 presents the characteristic temporal profile of two categories: Nightlife Spots and Gyms or Fitness Center. Each profile is a direct function of a users's propensity to visit at a given hour of the day and day of the week. The profiles of different venue categories in a ward establish the overall profile of that venue.

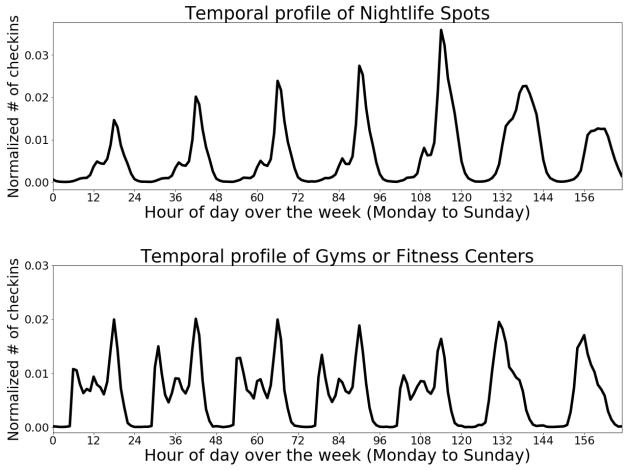

Figure 1: Normalized temporal profile of different categories of venues.

To illustrate this, let us consider two wards of interest: St. Pancras & Somers Town, which contains a major transportation hub and offices, and Camden Town with Primrose Hill, which contains a variety of venues and tourist attractions. Figure 2 shows the average number of check-ins in each ward for each hour of the day over the course of one week, aggregating across a number of weeks. This signal creates a characteristic temporal profile which acts as a temporal signature for the ward. The overall signal, shown in black, is different for these two wards. The number of check-ins at Camden Town steadily increases over the course of the day while the number check-ins at St. Pancras has two large peaks, one in the morning and another in the evening. Examining the three main categories (Food, Travel & Transport, Nightlife Spots) that characterize these two wards can help to better understand this observation. Camden Town has significant contributions from Nightlife venues which gradually increases over the course of a day. Conversely, St. Pancras is dominated by Travel & Transport, causing the overall temporal profile of the ward to peak at rush hour. These trends suggest that Camden Town is likely a more youth dominated area while St. Pancras is a hub for commuters or travelers, as they actually are [1].

4.2 Utilizing similarities in visitation patterns

Similar observations can be generalized to the rest of the wards in London. Different regions feature different degrees of similarity, an insight which we exploit in Section 5 to predict the characteristic temporal curves of new venues. We quantify the similarity between two temporal profiles using the Jensen-Shannon divergence (JSD) [10][2]. The JSD between two wards w_i and w_j is calculated as

[1] https://developer.foursquare.com/categorytree

[2] We use the JSD instead of Kullback-Leibler divergence since the former is a symmetric similarity measure between two functions, whereas the latter is not.

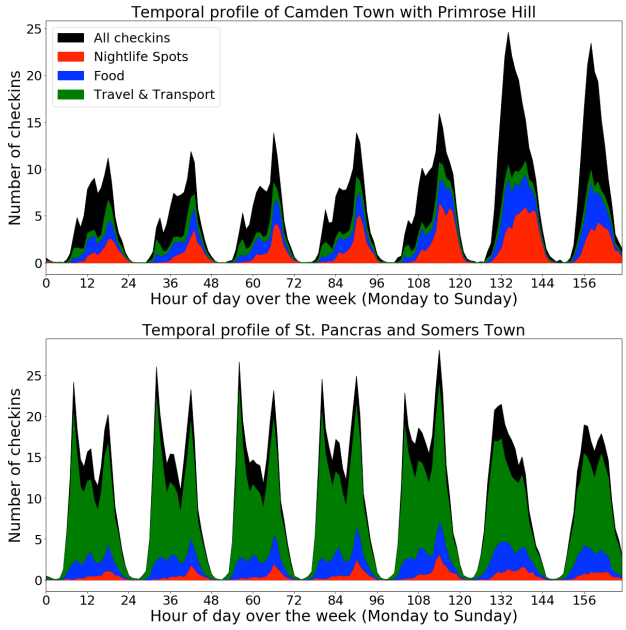

Figure 2: Daily temporal profiles and category breakdown of St. Pancras & Somers Town and Camden Town with Primrose Hill, two contrasting wards in London.

follows:

$$JSD(C^{w_i}, C^{w_j}) = H\left(\frac{C^{w_i} + C^{w_j}}{2}\right) - \frac{H(C^{w_i}) + H(C^{w_j})}{2} \quad (4)$$

where H is the Shannon entropy. JSD provides a metric that quantifies how two profiles, which can be seen as distributions over time, are similar. A low value of the JSD between the temporal profile of two wards represents a high similarity.

4.3 Temporal visitation patterns of new venues

We now focus on the temporal characteristics of new venues. These venues represent an interesting case study as upon their launch, unlike existing venues or geographic areas, there is no historic information on their expected popularity patterns over time.

Identification of new venues. The Foursquare dataset includes a list of all public venues in the city of London and for each includes the *creation time* which referes to the date the venue was crowdsourced by Foursquare users. Prior research on Foursquare data has shown that venues added after June 2011 were highly likely (probability above 0.8) to actually be new venues opening in an area rather than existing venues being added to the system for the first time [5]. We look at all new venues that were added to Foursquare after June 2011 with a minimum of 100 check-ins. This results in a list of 305 venues that is used for the following analysis.

Defining a venue's stable temporal profile. For each of the new venues in our set, we first examine the total number of check-ins at each time step for each week after the venue opened (ie. the weekly temporal profile). To avoid sparsity issues working at this level of granularity, we create a cumulative temporal profile per week, summing the total number of checkins at each time step with each consecutive week. The trend over the course of the week represents the characteristic curve of the venue and indicates the weekly demand trend. We normalize this curve each week by dividing by the sum of all checkins for the venue, up to the time of observation. With each consecutive week, we expect this curve to stabilize. We measure the stability of this temporal profile over time by calculating the variance of the temporal curve at time t relative to time $t-1$. Our data suggests that the temporal profile of a new venue becomes stationary when the value of the variance relative to the prior week is $\sigma^2 < 2.6e-05$. On average, this occurs 5 weeks after a venue has opened. Note that we build the profile of a venue considering a week's temporal span. This captures the most essential temporal patterns of activity at a venue, which includes diurnal variations, but also differences between weekends and weekdays.

5 PREDICTING A TEMPORAL SIGNATURE

5.1 Discovering similarities in dynamics

For our model, we begin with the basis that two venues of the same category in two different wards are likely to have similar temporal patterns if the overall temporal patterns of their wards are similar. For a given new venue v_i, our methodology to predict its temporal profile is as follows. For clarity, we will describe an example in which we assume v_i is an Italian restaurant called "The Meaning of Life" in ward 23.

(1) Determine the general category, specific category, and ward of that venue. For our example, the general category is "Food", the specific category is "Italian restaurant", and the ward is 23.
(2) Determine the temporal profile of the ward for the general category of interest. In this example, we would determine the overall temporal profile of "Food" venues in ward 23. Formally, we determine $C^{V_{g,w}}[0,T]$ where $T = 168$.
(3) Determine the N most similar wards. For all other wards in the city, compare their general category's temporal profile to that of our ward of interest and determine the N most similar wards where similarity is defined as $JSD(C^v, C^w)$ where $v \neq w$. This is referred to as the set of *temporally similar wards*. For our example, this would entail finding the N wards whose Food temporal profile is most similar to that of ward 23.
(4) Calculate the specific temporal profile for each ward in the set of *temporally similar wards*. For our example, we would calculate the temporal profile of Italian restaurants for each of the N similar wards.
(5) Create a representative curve. These N temporal curves serve as the basis of our prediction of the profile of our new venue v_i. To create a representative curve from those N profiles, we use each of the profiles as inputs to a Gaussian Process (GP) because of its ability to recognize latent periodic trends. The output of the GP becomes our prediction.

5.2 Gaussian Processes model

Our algorithm finds temporal profiles that are likely to be similar to the venue of interest. We harness Gaussian Processes (GPs) to build a regression model to capture the periodic trends in those profiles. GP regression is a Bayesian non-parametric which models a

distribution over an infinite set of random variables and is described by its prior mean and covariance functions. For this work the prior mean was set to zero [11]; the product of two Radial Basis function kernels were used as the base kernel functions. These describe the two types of periodicity in our data, over the course of a week as well as over the course of a day. Given the periodicity over the course of a day and a week, we posit that Gaussian Processes are able to recognize latent periodic trends in the data and more accurately create a prediction for a temporal profile. The inputs to our Gaussian Process are the temporal profiles of the similar wards. We then have the GP predict a temporal pattern for an interval of $[0, T]$ where $T = 168$ for the week's hourly profile. We then compare this prediction to stable temporal profile of the venue of interest.

Algorithm	NRMSE
Temporally similar wards, general category	1.614
Temporally similar wards, specific category	**1.575**
Random wards	2.692
Same ward, all categories	2.1941
Same ward, specific category	1.760
Same ward, general category	1.884
All wards, all categories	1.937
All wards, specific category	2.028
All wards, general category	2.190

Table 1: Error analysis of different similarity algorithms.

5.3 Evaluation

In this section, we introduced the idea of temporally similar wards suggesting that areas in a city that share temporal trends can be used to provide insight into the temporal profile of a new venue.

Baselines. We compare our results with a number of baseline approaches. For each, the temporal profiles produced are used as features for a GP whose outputs are the prediction of the characteristic temporal profile of the new venue. Table 1 lists the algorithms. We analyze wards with the most temporally similar profile of venues with the same general category as well as those with the same specific category. Additionally, we examine the overall temporal profile of random wards. Looking within the same ward as the new venue, we look at the temporal profile of all venues from all categories, as well as those with the same general category, and those with the same specific category. We also look at all wards in the city, examining venues from all categories, the same general category, and the same specific category as the new venue of interest.

Metrics. To analyze the accuracy of our prediction, we calculate the normalized root mean squared error (NRMSE) between the predicted temporal profile and the stable profile for each venue. We first look at the value of NRMSE as we vary the number of neighbors, N. Our results showed $N = 10$ to be the best indicator of temporal similarity of neighbors. This value was chosen for the subsequent analysis presented in this paper.

Results. We calculate the NRMSE for the output of each algorithm compared to the actual stable curve of each new venue. Table 1 presents a summary of these results. Temporally similar wards using the specific category of the venue proves to be the best predictor of the temporal profile of a new venue. This result suggests that the use of temporally similar wards as a predictor of the dynamics of a new venue could be a more robust predictor than the history of that same venue itself, insight that is especially valuable when working with sparse datasets.

6 DISCUSSION & CONCLUSION

We have investigated the prediction of the temporal dynamics of newly established venues using the check-in data of millions of Foursquare users. We have also introduced the concept of *temporally similar areas* in a city: areas that share patterns in the movement of people to venues within those areas. On the neighborhood level, we have seen that areas that are far from each other can be synchronized with regards to their temporal activities. Moreover, the temporal frequencies of such activities tend to be stationary over a certain period of time due to regularities in human mobility patterns. We exploited this information to predict the temporal popularity profiles of newly established venues, essentially transferring information from the level of an urban region to that of a specific venue. This form of analytics can provide new insights to new business owners who can plan supplies and staffing in their facilities during the cold start period of a new opening. Beyond retail venues, the idea can be expanded to other types of places, such as parks or outdoor spaces. Predicting how urban spaces are used over time can improve planning, including the design of schedules for their maintenance or police them.

ACKNOWLEDGMENTS

This work was supported through the Gates Cambridge Trust and partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

REFERENCES

[1] Greater London Authority. LSOA Atlas, 2014. https://data.london.gov.uk/dataset/lsoa-atlas.
[2] Michael Batty. *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-based Models, and Fractals.* The MIT press, 2007.
[3] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.
[4] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.
[5] Matthew L. Daggitt, Anastasios Noulas, Blake Shaw, and Cecilia Mascolo. Tracking urban activity growth globally with big location data. *Royal Society Open Science*, 3(4), 2016.
[6] Office for National Statistics. Number of Electoral Wards/Divisions in the United Kingdom, 2011. https://www.ons.gov.uk/.
[7] Urbano França, Hiroki Sayama, Colin McSwiggen, Roozbeh Daneshvar, and Yaneer Bar-Yam. Visualizing the "heartbeat" of a city with tweets. *Complexity*, 21(6):280–287, 2016.
[8] Google. Popular times and visit duration, 2017. https://support.google.com/business/answer/6263531.
[9] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. The TimeGeo modeling framework for urban motility without travel surveys. *Proceedings of the National Academy of Sciences*, page 201524261, 2016.
[10] Jianhua Lin. PDivergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 1991.
[11] Carl Edward Rasmussen and Christopher Williams. *Gaussian processes for machine learning.* MIT Press, 2006.
[12] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.
[13] Blake Shaw, Jon Shea, Siddhartha Sinha, and Andrew Hogue. Learning to rank for spatiotemporal search. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*, pages 717–726. ACM, 2013.