Proceedings of the Ninth International AAAI Conference on Web and Social Media

Taxonomy-Based Discovery and Annotation of Functional Areas in the City

Carmen Vaca
Escuela Superior
Politecnica del Litoral
cvaca@fiec.espol.edu.ec

Daniele Quercia
University of Cambridge
dquercia@acm.org

Francesco Bonchi
Yahoo Labs
bonchi@yahoo-inc.com

Piero Fraternali
Politecnico di Milano
piero.fraternali@polimi.it

Abstract

Mapping the functional use of city areas (e.g., mapping clusters of hotels or of electronic shops) enables a variety of applications (e.g., innovative way-finding tools). To do that mapping, researchers have recently processed geo-referenced data with spatial clustering algorithms. These algorithms usually perform two consecutive steps: they cluster nearby points on the map, and then assign labels (e.g., 'electronics') to the resulting clusters. When applied in the city context, these algorithms do not fully work, not least because they consider the two steps of clustering and labeling as separate. Since there is no reason to keep those two steps separate, we propose a framework that clusters points based not only on their density but also on their semantic relatedness. We evaluate this framework upon Foursquare data in the cities of Barcelona, Milan, and London. We find that it is more effective than the baseline method of DBSCAN in discovering functional areas. We complement that quantitative evaluation with a user study involving 111 participants in the three cities. Finally, to illustrate the generalizability of our framework, we process temporal data with it and successfully discover seasonal uses of the city.

1 Introduction

Cities consist of functional areas, areas dedicated to specific functions (food, entertainment, residence). A quick understanding of a complex city might be provided by new ways of discovering functional areas. These might benefit a variety of stakeholders: tourists who look for historical sites; locals who are after niche shopping; walkers who prefer specific types of urban smells (Quercia, Schifanella, and Aiello 2015) and sounds (Quercia, Schifanella, and Aiello 2016); and retail analysts who have to recommend where new brick-and-mortar shops are best placed (Karamshuk et al. 2013).

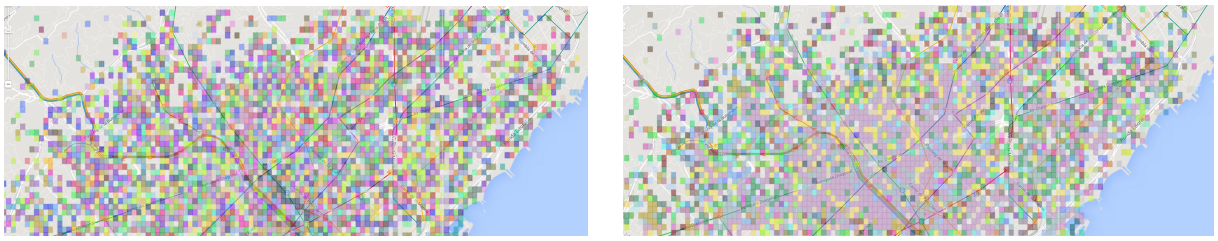
Automatically discovering functional areas is still "a very challenging problem in traditional urban planning" (Yuan, Zheng, and Xie 2012). To partly fix that, we propose a framework for discovering functional areas and, in so doing, make three main contributions:

- We frame the area discovery problem in terms of maximization of a simple objective function to be integrated into any clustering algorithm (Section 3). This function aims at finding and labeling the largest possible areas with the most specific labels (e.g., the label 'clothing stores' is preferable to 'shops').
- We evaluate the framework with a hierarchical clustering algorithm upon Foursquare data in the cities of Barcelona, Milan, and London (Section 4). We find that it is more effective than DBSCAN in discovering functional areas in those three cities.
- We complement our quantitative evaluation with a qualitative one (Section 5). We ask 111 participants to suggest areas where to carry out specific tasks (e.g., where to best place a new tech startup, where to go shopping). We then compare the areas they suggested with those automatically returned by our framework.

We conclude by discussing desirable properties of the framework, including its flexibility in working upon any type of data, including temporal one (Section 6).

2 Related Work

The simplest way of finding functional areas is to use a spatial clustering technique. One of the most common techniques is the Density-based Spatial Clustering of Applications with Noise (DBScan) (Ester et al. 1996). It finds a number of clusters starting from the estimated density distribution of points. A few years ago, for example, it was run on Foursquare data in the three cities of New York, London, and Paris (Bawa-Cavia 2011). To test the hypothesis that a modern city functions as a 'social archipelago' (i.e., "a fragmented set of islands characterized by high-density



(a) (b)

Figure 1: Cells in Barcelona, each of which is labeled using: (a) the category with the highest *TF-IDF* score; (b) the most frequent venue category.

social activity"), the author modeled Foursquare venues as geo-located points and found that Paris is less spatially fragmented than London; by contrast, New York is the most fragmented, twice as much as Paris.

Most of the latest research effort has gone into finding functional areas in the city. Researchers have done so in three main ways. The first way has relied on grouping together semantically-annotated points of interests. Cao *et al.* identified popular signatures (e.g., frequency distribution of different types of buildings) to find urban patches that frequently occur in different parts of the city. For example, the signature of residential areas might well be the high presence of single houses and garages (Cao et al. 2013). Noulas *et al.* exploited semantic annotations of Foursquare venues for grouping geographic areas in New York and London (Noulas et al. 2011), and saw how those changed from day to night.

The second way of finding functional areas has relied on human mobility. Mobility is derived from mobile phone traces (Quercia et al. 2010) or Foursquare check-ins (Long, Jin, and Joshi 2012; Zhang, Amy X. and Noulas, Anastasios and Scellato, Salvatore and Mascolo, Cecilia 2013). The premise of those approaches is that people's movements signal the potential and intrinsic relations among locations. In this vein, by tracking where Foursquare users check-in, Cranshaw *et al.* were able to move beyond the politically-defined boundaries of neighborhoods and discovered areas that effectively reflected the character and life of a few US cities (Cranshaw et al. 2012).

The third (and latest) way of finding functional areas is to combine semantic annotations with human mobility. Yuan *et al.* inferred the functions of each area using a topic-based inference model: areas are modeled as documents, functions as topics, categories as metadata, and human mobility patterns as words (Yuan, Zheng, and Xie 2012). They found that this way of discovering functional areas is far more effective than if one were to apply *TF-IDF* or *LDA* on their datasets. Indeed, if one were to gather all the foursquare venues in the city of Barcelona, divide the city map into 100x100m walkable cells, and color each cell with either the most frequent venue category or the category with the highest *TF-IDF* score, then the resulting map would be either too homogeneous (most of the cells in Figure 1(a) are labeled as 'food') or too fragmented (cells in Figure 1(b) are labeled with a variety of categories).

All the previous approaches find areas and, only after that, manually describe their functions. Understanding what an area really means requires, however, considerable human effort. Here we propose an approach in which the two steps of clustering and labeling incrementally unfold *at the same time*.

3 Proposal

3.1 Problem Statement

Imagine the entire city map divided into n equally sized grid cells: $A = \{a_1, a_2, \cdots, a_n\}$ (cell size is an input parameter). Within each cell, there are points that are labeled. As such, cell a contains different fractions of points having label l ($cov(l, a)$). Our goal is to find clusters of cells such that the points inside the clusters are *geographically closer* and more *semantically related* than what points outside those clusters are.

More formally, the map can be represented as a graph $G = (A, E)$ with vertex set A and edge weights $\{w_{i,j}\}_{a_i,a_j \in A}$. Vertices are our map cells, and edge weights reflect how related two adjacent cells are. For our purposes, we define $w_{a,b} = \frac{1}{k_a}$, if a and b are adjacent, and 0 otherwise. The value k_i is the number of neighbors of a. We use the 'rook case' notion of adjacency, in which the four surrounding cells (above, below, left, right) are considered adjacent[1]. There are also a taxonomy arranged in a tree $T = (V, E)$ whose leaves are in the set $L \subseteq V$, and an initial labeling function $\ell : A \to L$ that assigns a leaf of the tree to each cell. These assignments should be considered initial ones that the algorithm might well then change. Our goal is to find a labeling $\ell^* : A \to V$ of the cells such that:

- The labeling ℓ^* generalizes the initial labeling; that is, the new label $\ell^*(a)$ of each cell $a \in A$ should be an ancestor of $\ell(a)$ in T;
- The labeling ℓ^* gives adjacent cells the same or very similar labels;
- The labeling ℓ^* prefer specific labels over general ones; that is, the closer the labels to the leaves of T, the better.

[1]One could also assume that the graph G is undirected. In that case, the weights of two adjacent cells (a, b) are the same ($w_{a,b} = w_{b,a}$) and measure to which extent a and b are semantically related. By contrast, the weights of two non-adjacent cells are zero.

There is a natural tradeoff between those objectives. Consider the extreme case in which all the cells are labeled with the root of the tree: on the one hand, we would have perfect homogeneity of labeling; on the other hand, however, we would have over-generalized. At the other extreme, setting $\ell^* = \ell$ would result in no generalization, but adjacent cells would have overly diverse labels.

We formalize those competing objectives next. Let $adv : V \times A \rightarrow \mathbb{R}$ denote a function representing the advantage of assigning a label to a cell. Given a user-defined parameter $\lambda \in [0, 1]$, we want to find the labeling $\ell^* : A \rightarrow V$ that maximizes the sum $\sum_{a \in A}$ of:

$$\left(\lambda \sum_{b \in A} w_{a,b} \, \mathbb{1}[\ell^*(a) = \ell^*(b)] + (1 - \lambda) adv(\ell^*(a), a) \right) \quad (1)$$

The first term measures to which extent the proposed labeling $\ell^*(a)$ for a's neighborhood is homogeneous (the indicator function $\mathbb{1}$ is one, if the two proposed labels are the same; zero otherwise). The second term quantifies the advantage of cell a's proposed labeling and is defined as:

$$adv(l, a) = cov(l, a) - cov(l, A)$$

The higher label l's coverage for the points inside cell a ($cov(l, a)$) and the lower its coverage for the points across the whole map A ($cov(l, A)$), the higher label l's advantage. The *cell coverage* $cov(l, a_i)$ is the fraction of points with label l or with labels that are under l in T. More formally, we are given $\ell : L \times A \rightarrow [0, 1]$ such that $\forall a \in A$, $\sum_{l \in L} \ell(l, a) = 1$. For a node $l \in V$ of the taxonomy, let $L(l) \subseteq L$ denote the set of leaves contained in the subtree rooted at l. Label l's coverage for cell a then measures the presence of l inside a:

$$cov(l, a) = \sum_{l \in L(v)} \ell(l, a)$$

In a similar way, the entire *map coverage* is:

$$cov(l, A) = \frac{log_{10}\left(10 + \sum_{i=1}^{n} cov\left(l, a_i\right)\right)}{log_{10}(10 + n)}$$

This adds all the values of $cov(l, a)$ across the map and divides that sum by the number of cells. Both numerator and denominator are log-transformed to account for the skewness of the labels: it is likely that a few labels are present in many cells, while most labels are present in only a few cells. If this were not to be the case in a specific domain, then one could simply divide the overall coverage by the number of cells: $\frac{\sum_{i=1}^{n} cov(l, a_i)}{n}$.

3.2 Our proposal

To find the areas and labeling for the map that maximize the objective function, we would need to test all possible labels assigned to all possible areas and select the configuration for which the function is maximum. Since that would be computationally prohibitive, we need to find an efficient way of finding a satisfactory partitioning and labeling of areas. To this end, we modify the hierarchical clustering algorithm.

Algorithm 1 Hierarchical clustering - pseudocode

```
1:  procedure HAC(T, A, λ, contr())
2:      for each cell a_i in A do
3:          Assign a_i to a newly created cluster C_k
4:          Assign label l to C_k: contr(C_k, ℓ*(C_k)) is max
5:      end for
6:      for each pair of adjacent clusters C_k, C_h do
7:          checkEnqueuePair(C_k, C_h, λ, T)
8:      end for
9:      while (priorityqueue is not empty) do
10:         Get next tuple (M_kh, l_ij, contr(M_kh, ℓ*(M_kh)))
11:         mergePair(C_k, C_h, l_ij, contr(M_kh, ℓ*(M_kh))
12:         Replace references to C_k and C_h with M_kh
13:         Update contr() for each neighbor of M_kh
14:     end while
15: end procedure
```

This allows us to merge candidate cluster pairs in incremental ways: by incremental, we mean that each potential merge is independently evaluated and it takes place only if the objective function increases as a result.

We start with the initial labeling ℓ in which each cell is assigned the most popular label inside it. Next, we apply the hierarchical clustering that works as follows (Algorithm 1). Each cell in the map is initially assigned to a new cluster (*line 3*), resulting in as many clusters as cells. For each cluster C_k, since we can select any of the *candidate labels*: the labels present in it and their ancestors in T, we need to compute the contribution to the objective function for each of those labels and select the one that results in the maximum (*line 4*):

$$contr(C_k, \ell^*(C_h))$$

To start merging those clusters, we augment the original algorithm (not tailored to spatial applications) with a simple spatial notion: only *adjacent* clusters can be merged [2]. By testing which clusters are adjacent and which are not, we have a set of *cluster pairs* that could be potentially merged (*line 6*).

In the $checkEnqueuePair$ procedure (*line 7*), we test whether we are better off with the merge of C_k and C_h or not. For each candidate pair, we compute the first cluster's contribution to the objective function, and the second cluster's contribution. The two contributions are computed considering the two clusters' current labels. The contribution of the first cluster C_k is computed with the previous formula over all C_k's cells, and the contribution of the second cluster is computed summing over all cells in C_h. Having those two individual contributions, we are now able to decide whether to merge the two clusters or not. We merge them only if that merging operation contributes to the objective function equally or more than the sum of the two individual contributions; otherwise, the two clusters are best left separate. The

[2] One desirable by-product of this restriction is that the algorithm's search space is greatly reduced.

contribution of the newly merged cluster M_{kh} is computed with the previous formula: the only difference is that the sum is done over all the cells in *both* clusters. That contribution $contr(M_{kh}, \ell^*(M_{kh}))$ changes depending on the label assigned to the newly merged cluster. Since we can assign any of the *candidate labels* (i.e. the intersection of C_k and C_h's *candidate labels*), we need to compute the contribution for each of those labels and select the ones that result in a non-negative merging benefit:

$$\big(contr(M_{kh}, \ell^*(M_{kh})) - \\ (contr(C_k, \ell^*(C_k)) + contr(C_h, \ell^*(C_j))\big) \geq 0$$

In selecting C_k and C_h, we generate a priority queue in which cluster pairs are ordered by their merging benefits.

After putting all cluster pairs with non-negative merging benefits in the queue, we visit the queue by performing ordered merging operations starting with those with highest benefits (*line 10*). At each merging operation (*line 11*), the queue is updated (*line 12*): after combining, say, C_k and C_h, we refresh the queue by replacing all references to either C_k or C_h with M_{kh} and updating the contributions to the objective function of M_{kh}'s neighbors. The merging operations end when the queue is empty.

4 Evaluation

The goal of our algorithm is to cluster points in a map in a way that the points in the same cluster are geographically close and semantically related. To ascertain whether our proposal meets this goal, we ought to answer two main questions:

(*Area Distinctiveness*) To which extent is our proposal able to group points that are related to each other in the same areas?

(*Labeling Accuracy*) Does the label assigned to each area well describe the area's points?

Baseline. To answer those questions in a comparative fashion, we need to resort to a baseline algorithm. DBScan is widely used for spatially clustering points but it does not return any label for its clusters. We thus augment it by labeling DBScan's clusters with the most popular label in each cluster as that is what practitioners tend to do. The algorithm works by iteratively aggregating spatial points into clusters based on a threshold distance ϵ and a minimum cluster size $cmin$. We try $\epsilon = 400m$ and $cmin = \{3, 4, 5\}$, as they are commonly used values (Bawa-Cavia 2011). Points that cannot be assigned to any cluster are marked as 'noise'.

Foursquare dataset. We use the Foursquare REST Public API and crawl 22K, 60K and 37K venues located within the bounding box of the three cities of Barcelona, London and Milan. We then consider only the venues with at least 10 check-ins, leaving us with 14K, 30K, 18K venues. Each venue comes with its unique identifier, name, latitude, longitude and category. Those three cities have been chosen for their very different characteristics. From a geo-demographic perspective, with such a choice of cities, we explore different population sizes (3,2M for Barcelona, 8,3M for London, and 1,3M for Milan) and population densities (5,060 inhabitants per square kilometer in Barcelona, 4,542 in London, and 7,536 in Milan). From a Foursquare (or, more generally, social media) perspective, both Barcelona and Milan are less 'mature' than London, having a smaller user base and smaller number of check-ins. For our experiments, we crawl the entire tree of categories used by Foursquare to categorize venues and divide the bounding box of each city into walkable cells, each of which is initially labeled with its most frequent category and is roughly 100x100 meters in size. Previous research has established that 200m tends to be the threshold of walkable distance in urban areas (O'Sullivan and Morrall 1996; C.L. and TfL 2006) (in dense parts, this is equivalent to 2.5-minute walk), making our choice of 100m sufficiently conservative.

4.1 Distinctiveness

To measure the quality of the clustering, we consider the correlation-clustering problem. The values of the correlation clustering CC are experimentally affected by the number of clusters less than other clustering measures (e.g., normalized cut):

$$CC = \sum_{\substack{p_i \in C_k \\ p_j \in C_k}} (1 - dist(p_i, p_j)) + \sum_{\substack{p_i \in C_k \\ p_j \notin C_k}} dist(p_i, p_j)$$

Since points in the same area should ideally be closer to each other than points in different areas, this measure reflects the quality of the clustering as it increases with the points' closeness $(1 - dist(p_i, p_j))$, if the two points are in the same area; and with the distance $dist(p_i, p_j)$, if the two points are in different areas. To define the notion of semantic distance in a taxonomy, we resort to Jiang and Conrath (Jiang and Conrath 1997)'s definition:

$$dist(p_i, p_j) = 2\, log(Pr(lca(\ell^*(p_i), \ell^*(p_j)))) - \\ (log(Pr(\ell^*(p_i))) + log(Pr(\ell^*(p_j))))$$

where $Pr(\ell^*(p_i))$ is the occurrence probability of the label assigned to p_i in the city map (i.e., number of points labeled with $\ell^*(p_i)$ over the total number of points), and $lca(\ell^*(p_i), \ell^*(p_j))$ is the lowest common ancestor of the two labels assigned to p_i and p_j.

We compare the semantic correlation clustering values in the three cities of Barcelona, London, and Milan (Figure 2). In all cities, we observe the same two main results. First, our proposal performs consistently better than the best DBScan results (obtained with $cmin = 3$). The values of the correlation clustering for our framework are all above 0.90. Second, as λ increases, the semantic correlation clustering stays flat for DBScan (as it does not depend on λ) and slightly decreases for our proposal. In Barcelona, that decrease become noticeable only for $\lambda > 0.6$: yet, after that value, the correlation clustering is still above the best DBScan's values. This suggests that enforcing homogeneity with high values of λ does not impact the distinctiveness of the resulting functional areas in both Milan and London but, to a limited extent, impacts that in Barcelona. By inspecting the data, we

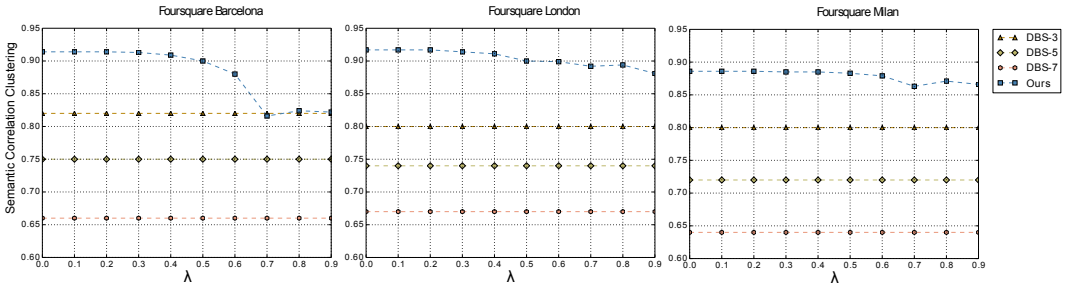


Figure 2: Semantic correlation clustering values for DBSscan and for our proposal as a function of λ in: (a) Barcelona; (b) London; and (c) Milan. The parameters of DBScan are $minpts = \{3, 4, 5\}$ and $\epsilon = 100$ meters.

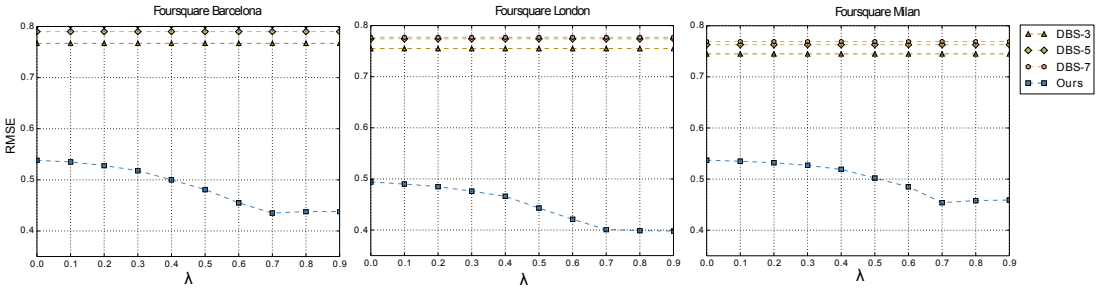


Figure 3: Labeling error (RMSE) for DBScan and for our proposal as a function of λ in: (a) Barcelona; (b) London; and (c) Milan.

know that the values of the semantic correlation clustering decrease after $\lambda > 0.6$ because labels in Barcelona tend to be more diverse than those in the two other cities.

4.2 Labeling Accuracy

After testing the distinctiveness of the functional areas, we need to test whether the areas are properly labeled. To assess whether a label assigned to each area well describe the points inside it, we use the Root Mean Squared Error (*RMSE*) and refer to it as labeling error:

$$RMSE(A, \ell^*(A)) = \sqrt{\sum_{C_k \in A} \frac{1}{|C|} \sum_{p_i \in C_k} \frac{1}{|C_k|} dist(\ell^*(p_i), \ell^*(C_k))^2}$$

In words, over all clusters (areas) C_k in the map, and over all points in each area, we measure the *semantic* distance between the label $\ell^*(p_i)$ assigned to the point and the label $\ell^*(C_k)$ assigned to the corresponding area. The higher the map's *RMSE*, the lower the accuracy of the labels (i.e., the higher the mismatch between the labels of the individual points and those of the corresponding areas).

We compare the labeling error values in the three cities of Barcelona, London, and Milan (Figure 3). By choosing the most popular label in each DBScan's cluster, the error is above 0.75 in the three cities. For our proposal, the error is always below 0.5. As λ increases, the error decreases (as high values for lambda enforce labeling homogeneity), and for $\lambda > 0.7$, the error is minimum.

By combining the two sets of results presented so far, we conclude that values of λ in the range $[0.35, 0.55]$ strike the right balance between area distinctiveness and labeling accuracy.

5 User Study

To complement our quantitative results and evaluate whether the found functional areas and their labels actually align with what residents perceive about their city, we conduct a mixed method user study.

5.1 Experimental setup and execution

We recruit 111 study participants in the three cities through university mailing lists and advertising on Facebook. Each participant is presented with a map of the city containing 5 (Milan) or 6 (Barcelona and London) functional areas that are highlighted and sequentially numbered (e.g., the bottom panel of Figure 4). Those functional areas have been generated setting $\lambda = 0.4$. Milan has 5 areas instead of 6 because it has no specific area where to buy electronics as opposed to Barcelona and London.

We start our study by asking each of our participants to read a consent form and optionally provide age, gender, years living in the city, and email address. We recruit 40 (Barcelona), 40 (London), 31 (Milan) participants. Among them, the percentage of female-male is 75%-25% for London, 53%-47% for Barcelona, and 58%-42% for Milan. The most common age band is that of 30-35 (London) and 24-29 (Barcelona and Milan), and all our participants have lived in their city for more than four years, allowing us to go beyond a typical student demographic.

We then provide brief instructions about the study and present six different tasks, one at a time. The tasks are chosen based on the area labels returned by our framework, and

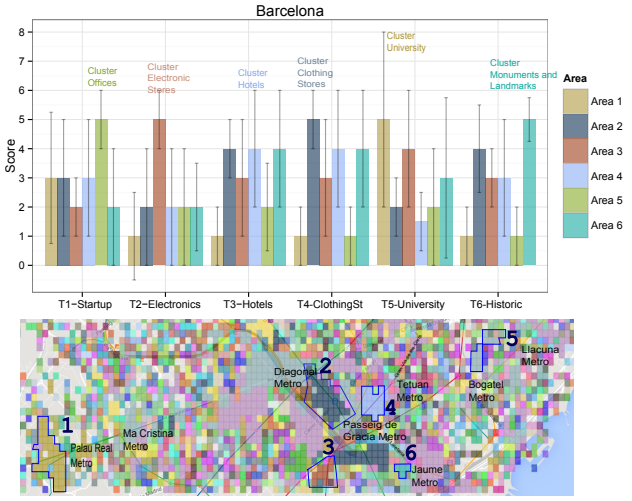


Figure 4: Barcelona areas our framework identified to be best for each hypothetical task (bottom panel), and the extent to which our respondents suggested the very same areas (top panel).

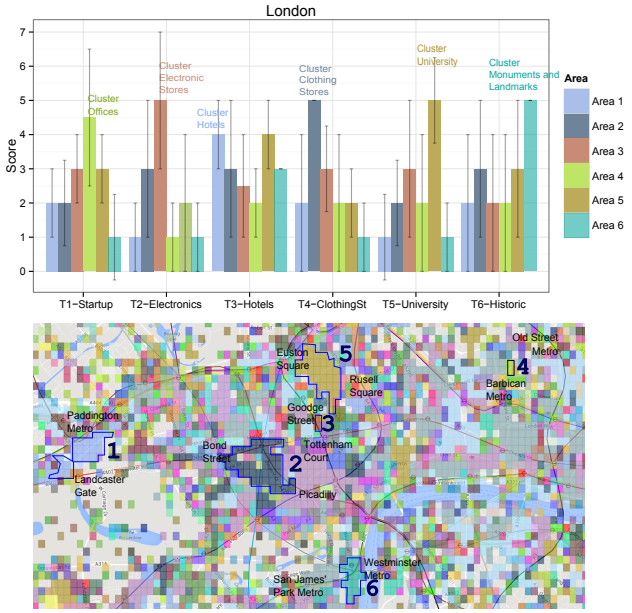


Figure 5: London areas our framework identified to be best for each hypothetical task (bottom panel), and the extent to which our respondents suggested the very same areas (top panel).

each task corresponds to one label. Of course, the area labels are unknown to the participants. A participant has to imagine the following six tasks and rate the extent to which each of the areas is suitable for each of the tasks. The ratings are expressed on a Likert scale (i.e., strongly disagree, disagree, neither agree nor disagree, agree and strongly agree).

Task-Office. A young entrepreneur is looking for a location where to base his new tech startup. You would recommend ...

Task-Electronics. A newly arrived student needs to buy electronics. You would recommend ...

Task-Hotels. A guy working in the hotel service industry has to visit as many hotels as possible in a short time. You would recommend ...

Task-Clothing. A friend of yours wants to visit as many clothing stores as possible. You would take her to ...

Task-University. A newly arrived foreign student wishes to experience university life. You would take her to ...

Task-Monuments. A friend of yours is visiting the city for the first time and wishes to see historic places and monuments. You would take her to ...

After providing answers for each task, the participant is asked to motivate his/her answers in free-text form and eventually indicate the name of the area that (s)he would have recommended for that task (it could but does not have to be one of the six areas).

5.2 Quantitative and Qualitative results

In Barcelona, the functions proposed by our framework (unknown to our respondents) match those suggested by our respondents for the office, electronics, clothing, and historic tasks (top panel of Figure 4): the area labeled with a given category (e.g., electronics) is associated with the corresponding category's task (e.g., 'task-electronics') with a 'strongly agree' assessment (i.e., with a median score of 5). These results are also reflected in our respondents' comments. For example, the majority of our respondents correctly consider area 5 to be best for the 'startup-task' and identify it to be the "22@ area", which is known, as one respondent puts it, "as the innovation & entrepreneurship area in Barcelona". By contrast, for the two remaining tasks (i.e., hotels and university), the variability of the answers is high. For the 'hotels-task', our framework has identified area 4 to be best. Some respondents agree with that, while others add that areas 2 and 6, being located in the central part of town, have many hotels too. For the 'university-task', our respondents suggested area 3 because most students live in that area, while there are just university buildings in area 1, which was the area identified by our framework.

As for London, the functions proposed by our framework match those suggested by our respondents for the office, electronics, clothing, university and historic tasks (top panel of Figure 5): for those tasks, the median scores are all 5. Once again, for the 'hotels-task', our respondents identified multiple areas (mainly central ones) to be best (not only the one identified by our framework).

Finally, in Milan, the functions proposed by our framework match those suggested by our respondents for the clothing and university tasks (top panel of Figure 5), and, to a certain extent, for the hospital task (the answer variability is lowest for the area identified by our framework). Again, for the 'hotels-task', a central area in addition to the one proposed by our framework is often mentioned. For the

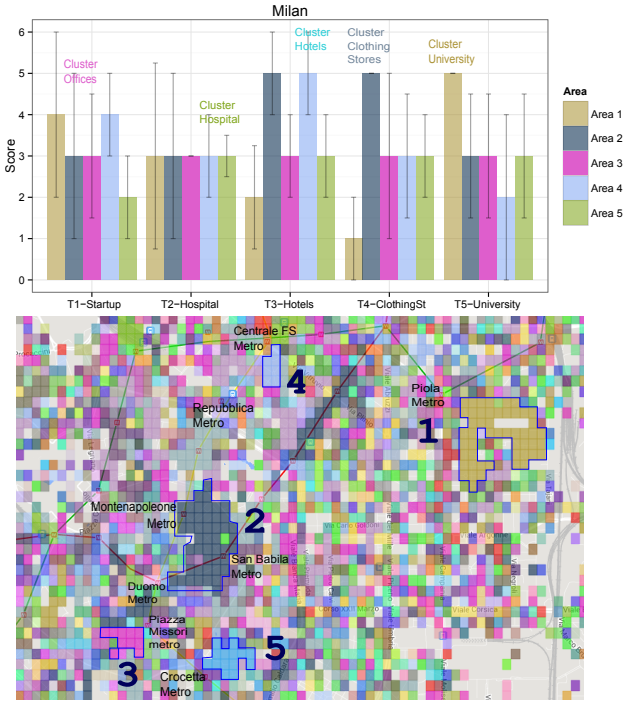


Figure 6: Milan areas our framework identified to be best for each hypothetical task (bottom panel), and the extent to which our respondents suggested the very same areas (top panel).

remaining 'startup-task', our respondents offer a variety of answers: some suggest the area labeled by our framework, while others suggest the area near the main technical university (Politecnico). They motivate this latter answer by saying that, if placed near Politecnico, the startup could benefit from technology transfer and could avoid the problems of more central areas of the city, which "are not specialized in technology and are simply crowded and expensive".

Taken together, the user study's quantitative and qualitative results both suggest that our framework is able to identify areas in the three cities and effectively label them with their functions.

6 Discussion

We now dwell on some of our framework's desirable properties and discuss some open questions.

6.1 Flexible framework

Previous approaches for discovering functional areas have modeled contextual factors such as time. For example, Yuan *et al.* not only derived Beijing's functional areas but also their evolution from 2010 to 2011 (Yuan, Zheng, and Xie 2012). Next, we briefly show that considering a temporal taxonomy allows for studying how the city fabric changes during different seasons.

To this end, we gather a random sample of more than 1M geo-referenced pictures within the bounding box of Barcelona from the Flickr public API, 400K of which are

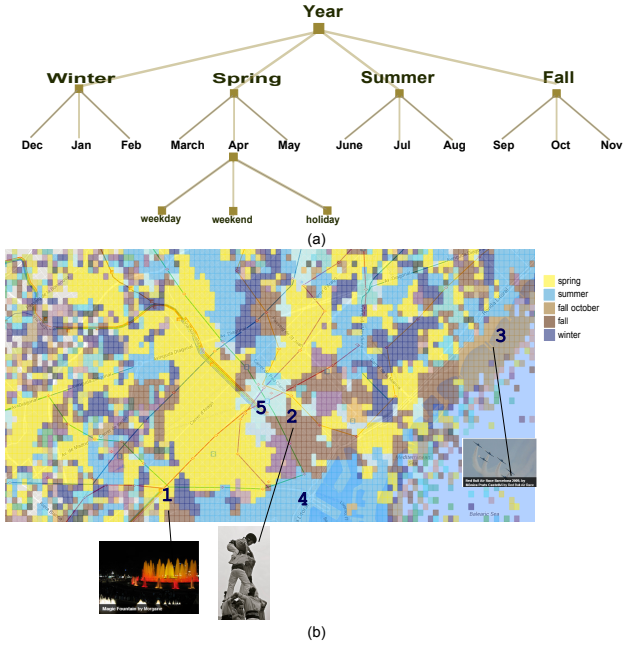


Figure 7: Temporal analysis of Flickr pictures in Barcelona. The areas reported in (b) are determined by our framework using the temporal taxonomy in (a) with $\lambda = 0.4$.

generated by distinct users in distinct locations. For each picture, we gather its unique identifier, latitude, longitude, the owner's identifier, the date of creation and the date of upload. We then consider the temporal taxonomy in Figure 7(a) and accordingly label each picture with one of the taxonomy categories depending on the picture's date of creation. We then run our framework with, again, $\lambda = 0.4$ and obtain the areas in Figure 7(b), which we number from 1 to 5:

Area 1. The framework categorizes this area with the 'fall' label. The area mainly covers a road called 'Avinguda de la Reina Maria Cristina', which is close to Plaza d'Espana. By looking at the actual pictures, we find that most of them depict the Magic Fountain (pictured at the bottom of Figure 7) during 'La Merce' festival. This celebrates the 'La Virgen de La Merce', patron of Barcelona, towards the end of September, and it is one of the most important festivities for the residents of Barcelona.

Area 2. This area is marked with 'fall' and is related with the Merce festival too. It mainly includes the road of 'Via Laietana', in which, during the festival, the world-renowned 'castellets' (towers of individuals on top of each other) are built. They are world-renowned since they receive coverage from major international news outlets.

Area 3. This area is marked with the very specific label of 'fall-october'. This label comes as a surprise as the area includes beaches popular during the summer. By looking at the pictures, we find out that they are about the 'Festa al cel', which we now learn is the greatest air showcase in the country. The label fall-october is not the most frequent among the pictures in that area but understandably is the

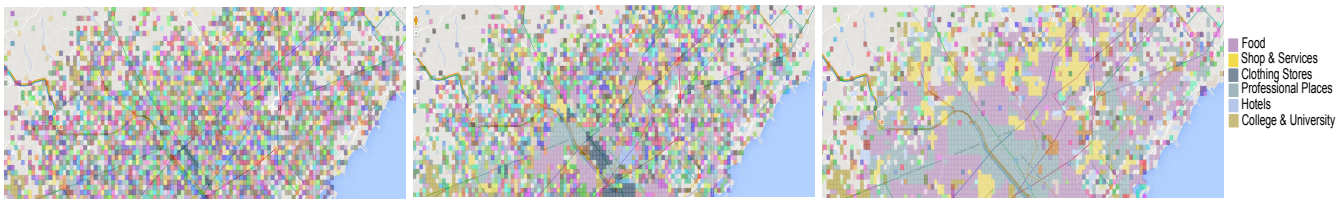


Figure 8: Areas in Barcelona identified by our framework from Foursquare venues with: (a) $\lambda = 0.1$; (b) $\lambda = 0.4$; and (c) $\lambda = 0.9$.

most discriminative compared to nearby areas, which happen to be beaches.

Areas 4 and 5. They contain the most famous beaches and central squares (e.g., Placa de Catalunya) popular among summer tourists. As one expects, the areas are categorized as 'summer'.

Those results suggest that, with a temporal taxonomy, one is able to identify key events in specific seasons. More importantly, those results speak to our framework's flexibility: different definitions of taxonomy result into very distinct notions of functional areas. Take a demographic taxonomy, which segments users into different socio-demographic classes according to age, gender, and profession. With it, our framework could potentially discover areas that serve similar or distinct functions for a variety of lifestyles (Yuan et al. 2013). Another example is a weather taxonomy. By using such a taxonomy, our framework could discover areas that are visited preferentially when, say, it is sunny or raining.

In practice, our framework's flexibility enables key applications. For example, a number of mobile personalized services are trying to figure out how to make geofencing a reality. The term geofencing refers to the use of geofences in combination with mobile services. The idea is that notifications are sent to mobile users whenever they cross a geofence (a geographic boundary). Many common geofencing scenarios are based on a simple radius around a point of interest, like a shop. One of the most important challenges is that geospatial calculations are complex and, since they require the use of GPS chips, they tend to drain the battery in a few hours. Our proposal partly fixes that as it offers a scalable solution: one could imagine to download the shape files of different functional areas on the phone and cheaply support background mobile applications that send tourist information or personalized shopping offers. To see how, imagine a user saving books and electronics on her electronic wishlist while at work. When traveling back home, her phone could generate alerts with the list of functional areas in which she could stop by and acquire some of the items on the list.

6.2 Usable framework

The proposed framework aims at being not only flexible but also usable. We believe it is so for two main reasons: i) it requires to fine tune a single parameter; and ii) it supports current visualization paradigms.

Single parameter. To discover functional areas, only one parameter has to be set. This parameter is intuitive and goes from 0 to 1: 0 corresponds to the finest-grained division of functional areas, while 1 corresponds to the most homogeneous division. The value of this parameter depends on what the algorithm's user is after. The difference between retail analysts' needs and tourists' is a case in point. Retail analysts might wish to emphasize areas that do not have a single function but have compound functions, so they would set a low value for the parameter (Figure 8(a) reports the Barcelona map for $\lambda = 0.1$). By contrast, tourists might be after a quick and digestible snapshot of the city fabric and, a such, would set a high value for it (Figure 8(c) reports the Barcelona map for $\lambda = 0.9$).

Supporting current visualization paradigms. The labels with which we annotate areas are not only human interpretable but also part of a taxonomy. That makes it possible to show such annotations simply as a list of categories next to a map, which is what people nowadays are used to and, as such, avoids pushing them outside their comfort zone. Having the map of a city and, next to it, the list of categories in our taxonomy, a user could click on a category and see the areas annotated with that category highlighted on the map. This was not readily possible with the most popular approaches of discovering functional areas (i.e., topic-based inference models, segmentation techniques). That is because those approaches express area annotations as category distributions, which are not easily translatable in a drop-down menu or, for that matter, in any visualization paradigm individuals are used to.

7 Conclusion

We have proposed a taxonomy-driven framework for discovering functional areas and have extensively tested it in three cities. By changing the type of taxonomy under study, we have shown that our framework offers flexibility on the types of areas that could be potentially discovered: for example, we have shown that it discovers not only functional areas but also seasonal ones. Based on those positive results, our framework promises to partly overcome a wide range of challenges, including: recent industry efforts in the area of mobile and personalization like geofencing; spatio-temporal studies of how the city is effectively used by social media users; and urban socio-cultural investigations such as that of how, given a language taxonomy, ethnic groups geographically sort themselves. To

collectively meet those challenges, in the near future, we will make the framework's code publicly available at researchswinger.org/tbscan.

References

Bawa-Cavia, A. 2011. Sensing The Urban: Using location-based social network data in urban analysis. In *Pervasive Urban Applications (PURBA)*.

Cao, Z.; Wang, S.; Forestier, G.; Puissant, A.; and Eick, C. F. 2013. Analyzing the Composition of Cities Using Spatial Clustering. In *Proceedings of the* 2^{nd} *ACM SIGKDD International Workshop on Urban Computing (UrbComp)*.

C.L., P., and TfL. 2006. Legible London wayfinding study. In *TfL*.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. M. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proceedings of the 6th AAAI Conference on Weblogs and Social Media (ICWSM)*.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, 226–231.

Jiang, J., and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the Conference on Research in Computational Linguistics*.

Karamshuk, D.; Noulas, A.; Scellato, S.; Nicosia, V.; and Mascolo, C. 2013. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In *Proceedings of the* 19^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Long, X.; Jin, L.; and Joshi, J. 2012. Exploring Trajectory-driven Local Geographic Topics in Foursquare. In *Proceedings of the ACM Conference on Ubiquitous Computing (UbicComp)*.

Noulas, A.; Scellato, S.; Mascolo, C.; and Pontil, M. 2011. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proceedings of 3rd Workshop Social Mobile Web (SMW'11). Colocated with ICWSM 2011*.

O'Sullivan, S., and Morrall, J. 1996. *Walking Distances to and from Light Rail Transit Stations*. Transportation Research Board.

Quercia, D.; Lathia, N.; Calabrese, F.; Di Lorenzo, G.; and Crowcroft, J. 2010. Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the* 10^{th} *IEEE International Conference on Data Mining (ICDM)*.

Quercia, D.; Schifanella, R.; and Aiello, L. M. 2015. Smelly Maps: The Digital Life of Urban Smellscapes. In *Proceedings of the* 9^{th} *AAAI Conference on Web and Social Media (ICWSM)*.

Quercia, D.; Schifanella, R.; and Aiello, L. M. 2016. Chatty Maps: The Digital Life of Urban Soundscapes. In *To Appear*.

Yuan, N. J.; Zhang, F.; Lian, D.; Zheng, K.; Yu, S.; and Xie, X. 2013. We Know How You Live: Exploring the Spectrum of Urban Lifestyles. In *Proceedings of the* 1^{st} *ACM Conference on Online Social Networks (COSN)*.

Yuan, J.; Zheng, Y.; and Xie, X. 2012. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *Proceedings of the* 18^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Zhang, Amy X. and Noulas, Anastasios and Scellato, Salvatore and Mascolo, Cecilia. 2013. Hoodsquare: Modeling and Recommending Neighborhoods in Location-based Social Networks. In *Proceedings of IEEE Social Computing Conference (SocialCom)*.