Juxtaposing Thematic Regions Derived from Spatial and Platial User-Generated Content

Grant McKenzie[1] and Benjamin Adams[2]

1 **Department of Geographical Sciences, University of Maryland, College Park, MD, USA**
gmck@umd.edu

2 **Department of Geography, University of Canterbury, Christchurch, New Zealand**
benjamin.adams@canterbury.ac.nz

Abstract

Typical approaches to defining regions, districts or neighborhoods within a city often focus on place instances of a similar type that are grouped together. For example, most cities have at least one bar district defined as such by the clustering of bars within a few city blocks. In reality, it is not the presence of spatial locations labeled as bars that contribute to a bar region, but rather the popularity of the bars themselves. Following the principle that places, and by extension, place-type regions exist via the people that have given space meaning, we explore user-contributed content as a way of extracting this meaning. Kernel density estimation models of place-based social check-ins are compared to spatially tagged social posts with the goal of identifying thematic regions within the city of Los Angeles, CA. Dynamic human activity patterns, represented as temporal signatures, are included in this analysis to demonstrate how regions change over time.

1998 ACM Subject Classification H.1.1 Systems and Information Theory

Keywords and phrases place type, thematic region, temporal signature, topic modeling, user-generated content

Digital Object Identifier 10.4230/LIPIcs.COSIT.2017.20

1 Introduction

Colloquially, inhabitants often refer to vernacular places within a city by their *thematic place type*, e.g., *the bar district* or *the shopping area* of the city [18]. Though each of us has a vague understanding of where these regions are (and are not) in the city, and they can be fundamental units of infrastructure for understanding urban dynamics, how we choose to delimit the boundary of these regions remains a topic of discussion for many in the spatial information science community [22, 15, 10]. This research has focused on this task from both topological and cognitive perspectives, identifying the various ways that humans choose to partition their environment. Differentiating commercial centers from residential areas, identifying a city's "downtown," and separating tourist areas from non-tourist areas have all been the subject of empirical and theoretical studies [9].

New types of data have emerged over the past few years that offer the opportunity to re-examine the concept of region delineation through a different lens. User-generated content (UGC) in the form of volunteered geographic information (VGI) and geosocial media have given inhabitants and visitors to a city a range of platforms on which to share their observations [8, 7]. While the majority of previous work in this area has focused on theoretical, simulated or small-sample surveys to gain insight into how individuals and groups

13th International Conference on Spatial Information Theory (COSIT 2017).
Editors: Eliseo Clementini, Maureen Donnelly, May Yuan, Christian Kray, Paolo Fogliaroni, and Andrea Ballatore;
Article No. 20; pp. 20:1–20:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

think about the urban environment, these new sources of data offer a rich set of *geotagged*, heterogeneous, in situ observations. The form that these *geotags* take and their relationship to the contributed content facilitate a revised debate on the both space and place, and the role they play in identifying thematic regions.

In his early work on the concept of place, Yi-Fu Tuan focused on better understanding and defining the concept and its relation to geographic space [31]. In his writings, he describes how places are spaces instilled with meaning given to them by the people that experience them. This notion is reflected in the observations that are made by people as they move throughout the city. The spatial location of a set of geographic coordinates obtained from the GPS of a mobile device becomes a place when someone tweets about their first kiss at that location. Similarly, the location of a geotagged photograph is a place given meaning by the photographer and subjects of the photograph, and simply preserved in time through a camera. The question is then, how cohesive are these places and is there enough similarity between them to construct regions of common themes? For example, do people tend to talk about activities related to bars (e.g., cocktails, dancing) within distinct spatial areas? What is more, do the locations where people talk about bars align with actual brick-and-mortar bars? In essence, there are two ways of defining thematic regions, one focusing on the linguistic content of spatially tagged observations and another based on the clustering of place instances of a thematic type. With respect to the latter, a current approach might find that a *bar district* of a city is defined as such based on the density of bars in that part of the city. Following the logic that *people define places*, however, means that while a brick-and-mortar venue that sells liquor may be labeled as a *bar*, it is arguably not one until it has a patron. Continuing this thread, an area that has a high density of popular bars contributes more to a consensus of a *bar district* than a cluster of establishments labeled as bars that have no customers.

The times of day that people conduct activities is of particular importance when discussing regions. Kevin Lynch, in his writings on *the image of the city*, describes how one's understanding of the city changes based on time of day, season, etc. [17, p.86] Using our bar place type example, if a bar district were to exist, it would clearly be most "bar-like" at 11pm on a Friday night. Does that bar district still exist at 9am on a Tuesday morning though? We argue here that thematic regions within a city are dynamic and as the activities conducted by people change, so do the regions in which those activities are conducted. The section of the city that facilitated entertainment and alcohol related activities on Friday night ceases to socially afford these activities on Tuesday morning, instead functioning as a space for office or workplace related activities.

The effect of time on thematic regions unveiled through spatially tagged content compared to those built from place-instances remains a point of discussion. So does the influence of environmental characteristics. Some thematic regions are related to physiographic features, e.g., beaches, while other are socially constructed based on a common activity theme, e.g., bars. The regions that emerge from grouping similar spatial or platial[1] thematic instances vary depending on these characteristics. For many people, especially college students, drinking activities tend to dominate the topic of conversation meaning that observations and content related to bars present themselves under spatiotemporal conditions where no bar or drinking activity currently exist, e.g., "Really looking forward to going to the bar with friends tonight." By comparison, observations about physiographic or environmental features tend to be less influenced by time and more restricted to the spatial extent of the physiographic feature.

In this work we examine the role that place, space and time play in defining thematic

[1] We use the neologism *platial* here in reference to place, similar to how spatial refers to space.

regions within a city. Specifically, we investigate how these regions can be identified through the following tasks:

- We use a kernel density estimation model to construct thematic regions from user-generated place instances. Using attribute information associated with these place instances as a proxy for popularity, we demonstrate that regional boundaries change when comparing places that are frequented by visitors with those that are places in name only.
- Using place type-specific temporal signatures generated from millions of human activity patterns, we demonstrate that regions can be represented dynamically. Depending on the time of day, and day of the week, a region may grow or shrink in size.
- We use a topic modeling approach to extract place type linguistic patterns from spatially tagged social media. Through these topics we show how thematic regions can be exposed from natural language content. We compare and contrast thematic regions based on place type and show that there is a substantial difference between content associated with physiographic features and content related to human constructed features.
- Last, we discuss the implication of both the spatial and platial approach to defining thematic urban regions. We show how the constraints and limitations of the various content platforms have an impact on the resulting region definitions.

2 Related Work

In recent years, the explosion in new forms of user generated volunteered geographic data has rekindled a research interest in using quantitative analysis to explore the notion of place [33]. In particular, this new data is seen as an opportunity to tap more closely into the phenomenological idea of place as tied to individual human experiences of their environments [26]. Due to increased "citizen" sensing of the environment via social media and mobile device usage, people are increasingly generating data about their environment through their activities, and human sensors are able to directly collect sensory information about the environment and contextualize and communicate it in language that other humans can understand [11]. Thus, this data gives us unique insight to place identity and sentiment as well as relationships between individuals, groups, and the physical environment [27]. By and large this research has explored the use of a variety of spatial analysis techniques as well as other data science methods, such as text mining, to operationalize place in geographic information systems.

A sub-area of this work revolves around creating spatial representations for regions, which are vaguely or non-canonically defined. Examples of this work include generating spatial footprints of vaguely defined *platial* regions such as tourist areas and city centers [14, 6], and for constructing spatial footprints for digital gazetteers [16]. Much of this work is not specifically at the *place-type* level nor has it compared place-based regions with regions generated from spatial footprints.

Because many human sensor observations are stored as natural language text, a variety of research projects have used text mining and natural language processing (NLP) tools and methods to generate structured place knowledge. Textual and narrative descriptions arguably provide a unique perspective on human interpretations and conceptualizations of place, because of the opportunity to infer information about what people "think" about places. In practice, different NLP methods provide this to different degrees. Existing work has focused on extracting place semantics from user-generated spatial content and narratives [13, 28]. In addition, computational sentiment and emotion analysis has been used to infer regularities in the emotional content of place descriptions [4].

One approach to better understand the thematic contents of place narratives is topic

modeling, a family of probabilistic machine learning methods commonly used to infer thematic structure in a corpus of text documents. The simplest topic model is Latent Dirichlet Allocation (LDA), which models the generation of a document set as the result of a random process [5]. First, a document is assigned a mixture of topics, then each word is randomly picked from those topics proportionately based on the importance of that word for the topic. Given this model as a starting point, the inferencing algorithm identifies the topics that would have most likely generated the existing corpus. Thus, it is an unsupervised method in that it derives the topics from the document set without any additional information or pre-defined structure. LDA is a bag-of-words model where word-order, parts of speech, and other grammatical structures are ignored. Despite this simplification, LDA is widely used as a way to quantitatively characterize the topics that are in individual documents and across an entire corpus. LDA has been used to map regions and times that are described using those themes [2]. It has also been used to discover thematic signatures of place types from text for the purpose of enriching place-based linked data [1]. In our previous work we developed a thematic search engine that uses geotagged natural language content from Wikipedia and travel blogs to cartographically present the results of topic-specific searches [3]. Additionally, topic modeling was used to extract vague cognitive regions such as *Southern California* from user-generated spatial content [10].

3 Data

A sample of 213,279 user-generated place instances were accessed via the Foursquare application programming interface[2] for the greater Los Angeles area. These place instances are categorized into one of 421 user-contributed place types (e.g., Bar, Park, Police Station) curated in a hierarchical vocabulary.[2] Of these, we selected 20 of the most dominant and unique place types restricting our place instance set to 37,302. Similarly, a random sample of 684,776, 699,113, and 642,059 geo-tagged (geographic coordinates) social media posts were collected from *Twitter*, *Instagram* and *Yik Yak* respectively over a three month time span starting January 2015. Twitter is a microposting service which restricts posts to 140 characters.[3] Yik Yak is a mobile application allowing users to post anonymous content to other users within a 5 mile radius of their location. The photo sharing platform, Instagram, geotags photographs and captions by default.[4] Note that only the text-based captions, not the photographs, were used in this analysis. These platforms range in the demographics of their user-base but most users are between the ages of 19 and 29 [23].

3.1 Temporal Signatures

Hourly temporal signatures for each of the 421 place types were constructed from check-in data collected in Los Angeles over a three month time period and aggregated to hours in a single week (see [20] for details). These normalized temporal signatures represent the *default* activity behavior at a given place type in Los Angeles. Figure 1 shows a sample of 3 place type temporal signatures. A higher value represents an increase in likelihood of finding someone at a place of that type at that time. *Mexican Restaurant* displays peaks at lunch and dinner time throughout the week while *Bar* activity is shown increasing late at

[2] `https://developer.foursquare.com/`
[3] The removal of this character limit occurred in 2016.
[4] Data was collected for this project prior to Instagram changing their location-tag settings.

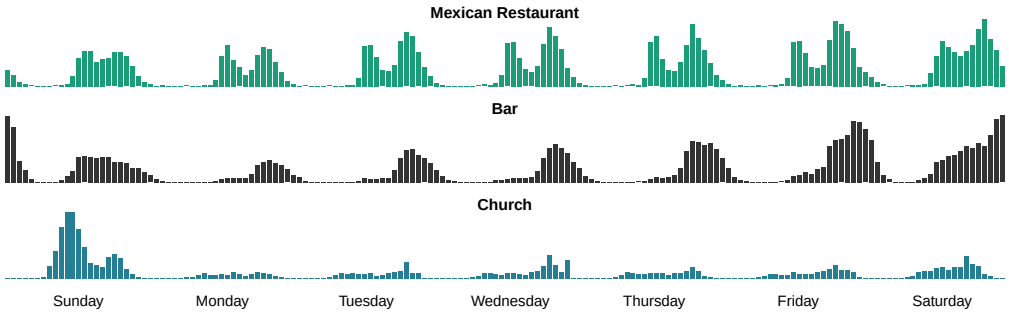


Figure 1 Hourly temporal signatures constructed from geosocial media check-ins for three place types in Los Angeles, CA.

(a) Mexican Restaurant **(b)** Bar **(c)** Park

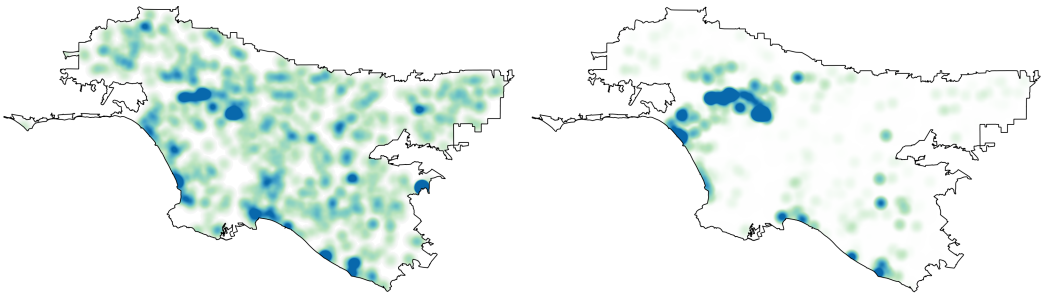
Figure 2 Three place type specific topic word clouds extracted from Foursquare tips.

night throughout the week. *Church* presents an expected peak on Sunday morning and a smaller one on Sunday afternoon with negligible activity the remainder of the week. The purpose of visually depicting these temporal signatures is to show that default activity behavior towards places does vary significantly between place types. The involvement of these temporal signatures in identifying regions will become apparent in Section 4.

3.2 Linguistic Signatures

A hexagonal grid was generated over the greater Los Angeles area with grid cells at 0.01 degrees wide in latitude (roughly 1.1 km). All geotagged tweets, Instagram captions and Yik Yak posts were intersected with the hexagonal grid and each post was assigned to a grid cell. The textual content of these posts were cleaned by removing all non-alphanumeric characters as well as removing all stop words and words less than three characters.

Topic modeling was used to extract common themes across the spatial data sources. Previous work has used topic modeling to derive thematic signatures for places from unlabeled text [2, 1]. However, in this case we had a dataset of labeled tips from Foursquare, so we could use a form of supervised topic modeling called Labeled LDA (L-LDA) to train for topics that match the 20 place types selected [25]. Similar to LDA, L-LDA models a topic as a probability distribution over words, indicating the likelihood that someone writing on that topic will use particular words. It differs from LDA in that a one-to-one relationship is maintained between the user-supplied place type labels and the topics that are learned. Figure 2 represents the topics learned from the Foursquare tips for three place types.



(a) Non-weighted KDE **(b)** KDE weighted by unique visitor check-ins

Figure 3 Two kernel density estimate representations of the *Bar* place type in Los Angeles, CA.

4 Thematic regions identified through place instances

In many gazetteers, place instances are stored as point representations. This also holds true for many geosocial media place dictionaries such as Facebook and Foursquare. One approach to constructing regions from point data is to use kernel density estimation (KDE). This has been successful for constructing spatiotemporal regions from spatial locations of photographs [12, 30] and georeferenced text [3, 24]. Here, we split the place instances in the greater Los Angeles area by their place type and construct kernel density estimations based on these points. The KDE bandwidth used in each of these was calculated using the method proposed by Sheather and Jones [29].

One approach to identifying regions for a specific place type, e.g., *Bar*, using kernel density estimation is to weight each instance of a bar equally. The assumption here being that all bars are equal in their *bar-ness* and that the presence of a bar in the city, regardless of location, size or popularity should contribute equally to the identification of one or multiple regions. While this is a reasonable approach, it does make the arguably erroneous assumption that all bars contribute equally to what one might consider a *bar region*. We argue here, that a popular bar, for example, should contribute more to defining a *bar region* than a venue that has had little to no visitors in the past year. Ascertaining the actual popularity of a venue is a monumental task however. Fortunately, new sources of user-contributed place-based data now exist that work as proxies to actual venue popularity measures. Geosocial media content such as *check-ins* offers additional information concerning both place types and place instances that were previously only accessible through cost and time-prohibitive surveys or simulated data. Interaction behavior with the Foursquare representation of a place instance is accessible via a number of attributes including *unique visitor check-ins*, *total number of check-ins* and *total number of likes*. A check-in in this case refers to the act of an individual using the Foursquare application on their mobile device to indicate that they are at the physical place represented in their application as a Foursquare *venue*. A *Like*, on the other hand, does not imply that the user is or was at the actual physical place. To account for place instances being added to the Foursquare gazetteer at different times, attribute values used in this work were restricted to the last three years.

Not surprisingly there is a high correlation ($Pearson > 0.83$, $p < 0.01$) between the three Foursquare attributes which is reflected in the regions exposed by attribute weighted kernel density estimation models. Given these high correlation values, we chose to focus on *unique visitor check-ins* in a KDE weighted model. Applying this attribute as a weight in the KDE ensures that place instances that have had a higher number of unique visitors,

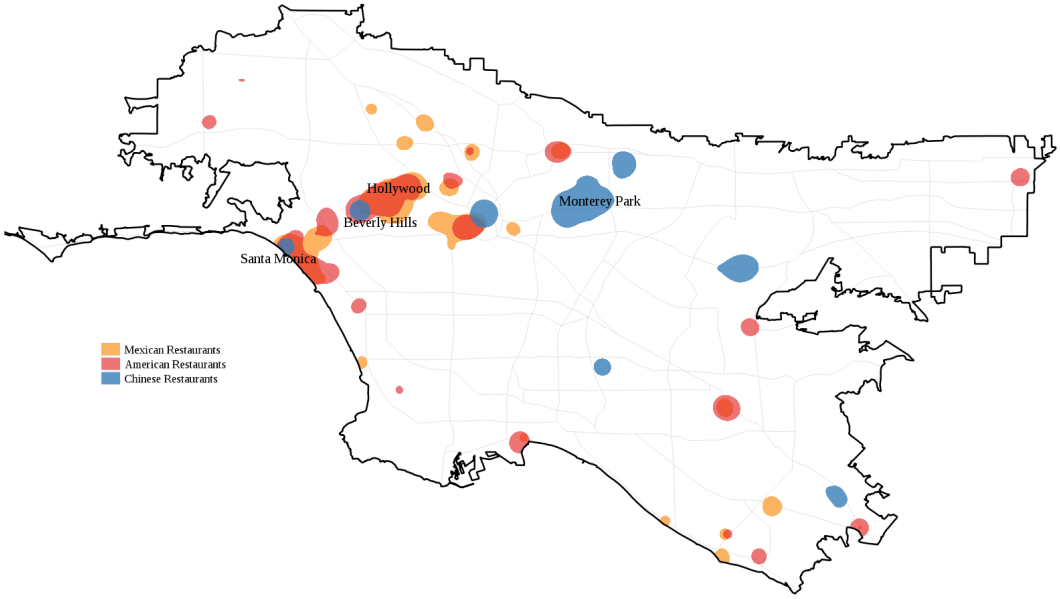


Figure 4 Three restaurant types in Los Angeles, CA.

also have a larger influence on the KDE and the discovery of thematic regions. Figure 3 shows a cartographic representation of two KDE models for *bars* using no weight (Figure 3a) and *unique visitors check-ins* (Figure 3b). The difference in these two maps highlights the influence that the popularity of a set of bars has on defining bar-type-regions. Note that over our sample set of 37302 place instances, 18% (6717) had no visitor check-ins. In our subset of *bars*, 45% had less than 10 unique visitors while 14% listed more than 3000.

4.1 Specifying region boundaries

Generating and mapping a kernel density estimation model for a place type produces a cartographic representation of a region with vague or fuzzy boundaries. As shown in Figure 3, opaque blue highlights the most *bar*-like areas while semi-transparent green indicates areas that are less *bar*-like. While this representation of regions via fuzzy boundaries is often appropriate for discussion purposes, as it reflects our cognitive perception of thematic regions, specifying a threshold on which to state that a region is either a *bar region* or not is of value in some cases [15]. For example, certain urban planning laws in the United States require that commercial land-use be specified by a hard boundary (typically streets) and restricting these boundaries or limiting place-types to a certain neighborhood or set of city blocks is often necessary for zoning purposes.

To construct these hard boundaries, we removed all raster pixel values below two standard deviations above the mean (for the given KDE raster) and assigned all other pixels a value of 1. Three types of restaurants in Los Angeles were analyzed in this way and are presented in Figure 4. There are clear similarities and differences in the regions produced through this analysis. From a qualitative perspective, the most notable similarity is that all three types of restaurants have a regional presence in the city of Santa Monica and surrounding area. Both Mexican and American restaurants are popular along the coast to Venice Beach and Marian Del Ray neighborhoods as well as inland around West Los Angeles. Similarly, Hollywood is a hot spot for both American and Mexican cuisine while Chinese Restaurants

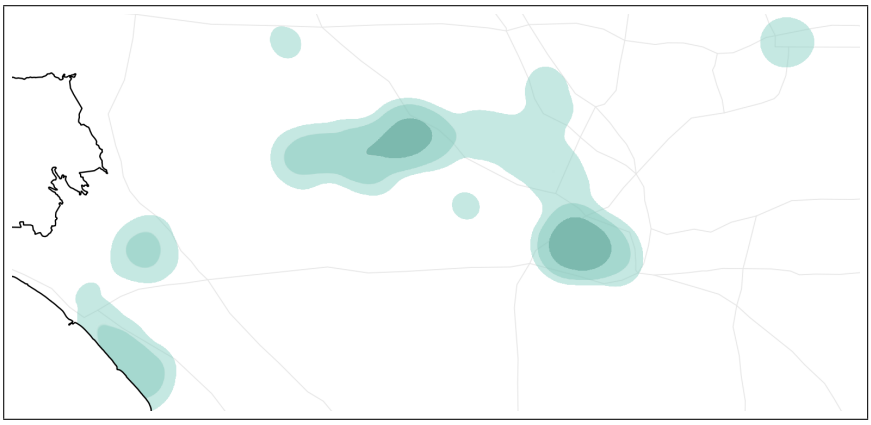


Figure 5 The *Bar* thematic place type region changing by time of day. Darkest to lightest: Friday 3pm, 7pm, 11pm.

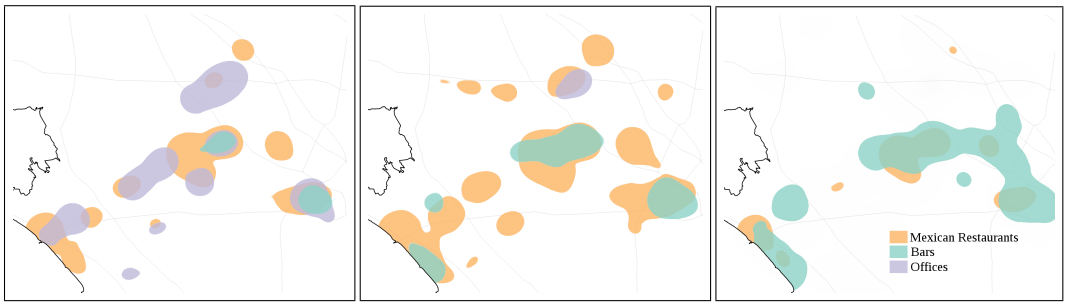
are also popular in the city of Beverly Hills. There are notable differences between these restaurant regions as well. The city of Monterey Park overlaps with the largest thematic region of Chinese Restaurants in this figure. According to the 2010 U.S. Census, Monterey Park contains a population that identifies as 66.9% Asian decent with a large concentrations of Chinese Americans [32].

4.2 Temporal dynamics of thematic regions

In our previous work on place types, we used social check-ins to generate temporal signatures of human activity behavior based on time of day [19]. This work confirms the notion that certain place types are more popular at certain times of day and days of the week. For example, people are more likely to patron restaurants during midday and evenings and employees have a high probability of being in office buildings between 9am and 5pm on weekdays. These temporal signatures offer unique insight into what place type activities happen when. Combining these temporal signatures with our thematic place type regions allows us to model the temporal dynamics of a city like Los Angeles. This work continues existing efforts in examining the *pulse* of a city, focusing on thematic regions instead of point representations [21].

Regions for three place type, namely *Mexican restaurants*, *bars* and *offices* were constructed using the unique visitor check-in weighted KDE method described in Section 4. As previously mentioned, the threshold value two standard deviations above the mean was recorded for each of these place types. The original place location data containing normalized values for unique visitor check-ins was then multiplied by the normalized temporal signature for the respective place type at three different times. This produced three new values on which to weight three new kernel density estimates (three for each place type). In this example, the times were Friday at 3pm, 7pm and 11pm. The threshold value from the original non-temporally weighted KDE was applied to each of the three new KDE maps which resulted in larger or smaller regions depending on the temporal probability value. Figure 5 shows the bar region for one section of Los Angeles as three temporal snapshots overlain on top of each other. The regions are represented temporally from darkest to lightest in this example with the smallest *bar region* around 3pm and the largest *bar region* (highest temporal probability) at 11pm.

The effect of time is different on each thematic region as shown by three examples in Figure 6. The *Office* region, shown in purple, decreases in area from 3pm to 11pm on Friday



(a) 3pm Friday **(b)** 7pm Friday **(c)** 11pm Friday

Figure 6 Merging three place-type regions with their default temporal signatures at three times on a typical Friday.

(in fact it is non existent at 11pm) while the regions representing *Mexican restaurants*, in orange, peak in size at 7pm. *Bars*, as shown in Figure 5, grow significantly from 3pm to 11pm. These visual representations reflect the idea that regions of a city are transitional [17]. While the buildings and spaces that contain the place instances exist atemporally, the places themselves and the regions that they contribute to are temporally dynamic.

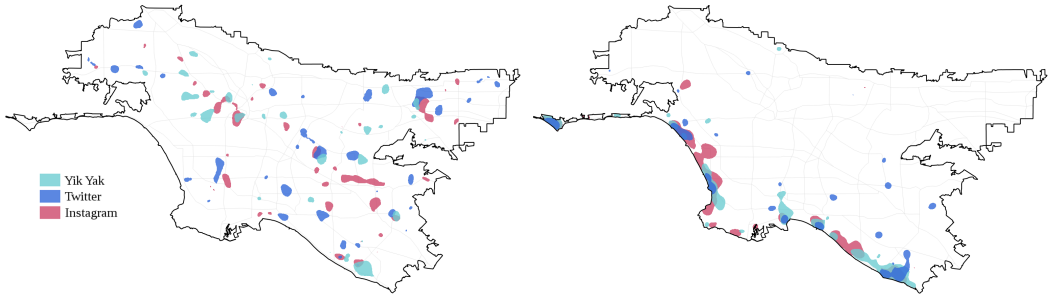
5 Thematic regions identified through spatially tagged content

The L-LDA topics extracted from geosocial spatial data were used to identify spatial regions in Los Angeles. Following a similar approach to the regions built from place-based data, kernel density estimate models were plotted from the 0.01 degree hexagonal grid using the topic values as the weight. As mentioned in Section 3.2, these topics were extracted from Foursquare tips, trained by the appropriate place type, and used to label our *spatial* data: *Twitter* tweets, *Yik Yak* posts and *Instagram* photo captions. For example, text related to *bars* in Foursquare tips were used to identify and label tweets with similar textual content. However, we found that in most cases, the thematic regions defined by the spatial datasets did not align with the place instance-identified regions. Moreover, in many cases, there was no common agreement between the different social media platforms themselves. Further investigation found that the main difference impacting agreement between datasets was the broader category of place type. More specifically, whether the thematic place type was tied to a feature in the natural or human-built environment.

5.1 Physiographic vs. human defined place types

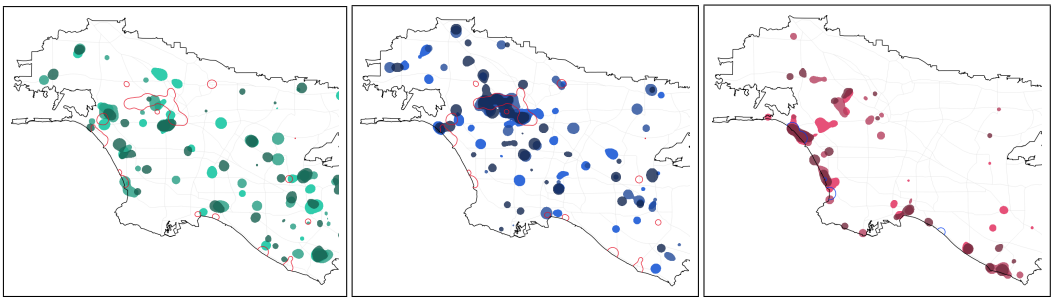
There is very little agreement between the three spatial datasets for human-built regions such as *bars* (Figure 7a), *Mexican restaurants* or *office buildings* discussed in the previous section. In fact a pixel-based Jaccard similarity coefficient (threshold at $mean + 2SD$) for these three datasets was consistently below 0.05 for all pairs of datasets.[5] By comparison, place types that mapped to physiographic features in the natural environment such as *Beach* (Figure 7b) showed much higher agreement between the three datasets with an average Jaccard coefficient of 0.28. These differences likely reflect how people refer to these different

[5] Jaccard is bounded between 0 and 1, with the latter indicating the datasets are identical.



(a) Bar **(b)** Beach

Figure 7 Regions for *Bar* and *Beach* identified from three different spatially tagged social media platforms.



(a) Bar regions via Yik Yak **(b)** Bar regions via Twitter **(c)** Beach regions via Instagram

Figure 8 Bar topics for two datasets at 3pm, 7pm, 11pm on Friday.

groups of place types as well as the demographics of the application users. This will be discussed further in Section 6.

5.2 Temporal dynamics of spatial content

Following the temporal example described in Section 4.2 of Friday at 3pm, 7pm and 11pm, we extracted spatially tagged content from each of the three social media platforms for those times periods. Though this reduced the sample size from which to generate topic signatures, Friday evening is a popular time for all social media applications meaning there was adequate data on which to run the analysis. Using the *bar* example again, we compared the regions identified for the three times of day across the three data sources. The results were inconsistent. In some cases, e.g., Yik Yak (Figure 8a), small regions were identified in different parts of Los Angeles, at different times of the day. There is little overlap between time periods and in general and the overlap that does exist does not align with the place-instance based thematic regions, outlined in red. On the other hand, Twitter (Figure 8b), again shows small, inconsistent regions outside of the city center, but there is significant overlap with the place instance-based regions over all time periods on a Friday. Notably, the number of regions do not increase nor is there a change in the size of the regions as the evening progresses.

By comparison, the *beach* regions identified via Instagram photograph captions are primarily clustered around the Los Angeles coastline and reflect a similar pattern to the one shown in Figure 7b. The overlap across hour layers is high and though there are regions identified inland, they are primarily clustered around the downtown city core where there is

significant social media activity. Potential explanations for this are discussed in the next section.

6 Discussion & Conclusions

The results presented in the previous two sections deserve further discussion, specifically on the difference between regions identified through spatial data and those identified through place instances. The biases and limitations associated with these data are also discussed.

6.1 Spatial tags vs. Place instances

There are important differences between regions identified through spatial and platial sources. Spatially tagged social content reflect observations of individuals at certain locations and times. The content of an observation, however, need not reflect the affordances or activities associated with the space from which the observation was made. The bar regions we identified, for example, tend to be dispersed across all of Los Angeles at many times of day. The reason for this can be understood from an example Yik Yak post at 3pm on Thursday in North-East Los Angeles: *"Can't wait to hit up the bars tonight, it has been a long week."* What we find is that this disconnect between what place instances identify as regions and what spatially tagged posts identify as regions, varies by place type. Human constructed places tend to have a larger disconnect between the spatially and platially identified regions while physiographic regions show greater alignment.

Also of importance is the difference in the intention of the data source. For example *beaches* contributed to Foursquare tend to be officially designated public beaches. Spatially-tagged social content, however, rarely explicitly identifies a beach. The content reflects observations and language related to beaches in general. The contributor of the latter data may not actually care that she is standing on an officially designated public beach. In all likelihood her definition of a beach simply consists of a sandy area adjacent to a body of water. In this case it is not unexpected that the regions identified from the place-based data may differ from those identified from spatially tagged observations, even for physiographic features.

6.2 User-generated Content

User-generated content, of which social media is one type comes with its own set of biases. Like all data, it is influenced by the views of its creators. In the case of these geotagged data, the contributors are predominantly young adults. This demographic has biases towards certain place types and the amount and nature of the content reflects these biases. For example, young adults arguably have a much more complex relationship with *bars* than they do *beaches*. While the social capital involved with posting about beach activities is high, it pales in comparison to activities related to drinking alcohol.

The structure of the Foursquare place type vocabulary also impacts how this work identifies regions. The place type *Bar* shares a number of similarities related to entertainment and alcohol with other place types such as *nightclubs, lounges, Karaoke venue*, etc. It is likely that the words and topics extracted for bars via L-LDA are quite similar to those of these other place categories yet the regions identified may be slightly different. The language used in social content may align with these similar place types as well. Similarly, the terms identified as being most *bar*-like may potentially be used to describe social interactions with friends in a dorm room or *tailgating* outside of a stadium. Though our topic modeling

approach assigned many alcohol and entertainment terms to the *bar* place type, the noise and ambiguousness common to social media posts could have lead to some mis-labeling. These are some of the known issues of working with natural language classification.

6.3 Conclusions

Understanding how thematic regions are identified within a city has been a topic of discussion in the spatial science community for many years. This work makes use of two unique types of geographic content, namely spatially tagged social media posts and thematically labeled place instances. Novel aspects of these data offer insight into how people interact with a city, allowing us to identify thematic regions through the use of weighted analysis models. The heart of this research, however, lies in a discussion of space and place. Does having access to user-contributed geographic content enhance our understanding of the relationship between space and place? Does the inclusion of new and alternative datasets change our existing cognitive and theoretical approaches to how regions are defined? These are questions that we have just scratches the surface of in this work and will continue to examine in future research.

References

1. Benjamin Adams and Krzysztof Janowicz. Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science*, 29(4):556–579, 2015.
2. Benjamin Adams and Grant McKenzie. Inferring thematic places from spatially referenced natural language descriptions. In Daniel Sui, Sarah Elwood, and Michael Goodchild, editors, *Crowdsourcing Geographic Knowledge*, pages 201–221. Springer, 2013.
3. Benjamin Adams, Grant McKenzie, and Mark Gahegan. Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th International Conference on World Wide Web*, pages 12–22. ACM, 2015.
4. Andrea Ballatore and Benjamin Adams. Extracting place emotions from travel blogs. In *Proceedings of AGILE*, volume 2015, pages 1–5, 2015.
5. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
6. Jiaoli Chen and Shih-Lung Shaw. Representing the spatial extent of places based on flickr photos with a representativeness-weighted kernel density estimation. In *International Conference on Geographic Information Science*, pages 130–144. Springer, 2016.
7. Jeremy W. Crampton, Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and geographic information science*, 40(2):130–139, 2013.
8. Andrew Crooks, Dieter Pfoser, Andrew Jenkins, Arie Croitoru, Anthony Stefanidis, Duncan Smith, Sophia Karagiorgou, Alexandros Efentakis, and George Lamprianidis. Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, 29(5):720–741, 2015.
9. Clare Davies, Ian Holt, Jenny Green, Jenny Harding, and Lucy Diamond. User needs and implications for modelling vague named places. *Spatial Cognition & Computation*, 9(3):174–194, 2009.
10. Song Gao, Krzysztof Janowicz, Daniel R. Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 2017, Ahead of Print.

11 Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

12 Christian Grothe and Jochen Schaab. Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation*, 9(3):195–211, 2009.

13 Heidelinde Hobel, Paolo Fogliaroni, and Andrew U Frank. Deriving the geographic footprint of cognitive regions. In *Geospatial Data in a Changing World*, pages 67–84. Springer, 2016.

14 Livia Hollenstein and Ross Purves. Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, (1):21–48, 2010.

15 Christopher B. Jones, Ross S. Purves, Paul D. Clough, and Hideo Joho. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1065, 2008.

16 Carsten Keßler, Patrick Maué, Jan Torben Heuer, and Thomas Bartoschek. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In *International Conference on GeoSpatial Sematics*, pages 83–102. Springer, 2009.

17 Kevin Lynch. *The image of the city*, volume 11. MIT press, 1960.

18 G. Mattson. Bar districts as subcultural amenities. *City, Culture, Society*, 6(1):1–8, 2015.

19 Grant McKenzie and Krzysztof Janowicz. Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Computers, Environment and Urban Systems*, 54:1–13, 2015.

20 Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015.

21 Grant McKenzie, Krzysztof Janowicz, Song Gao, Jiue-An Yang, and Yingjie Hu. POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2):71–85, 2015.

22 Daniel R. Montello, Michael F. Goodchild, Jonathon Gottsegen, and Peter Fohl. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3):185–204, 2003.

23 Pew Research Center. Demographics of key social networking platforms, 2015. Accessed: 2017-02-27. URL: `http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/`.

24 Ross S. Purves and Curdin Derungs. From space to place: Place-based explorations of text. *International Journal of Humanities and Arts Computing*, 9(1):74–94, 2015.

25 Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP'09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

26 Edward Relph. *Place and placelessness*, volume 67. Pion London, 1976.

27 Stéphane Roche. Geographic information science II Less space, more places in smart cities. *Progress in Human Geography*, 40(4):565–573, 2016.

28 Simon Scheider and Ross Purves. Semantic place localization from narratives. In *COMP@ SIGSPATIAL*, pages 16–19, 2013.

29 Simon J. Sheather and Michael C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690, 1991.

30 Yeran Sun, Hongchao Fan, Marco Helbich, and Alexander Zipf. Analyzing human activities through volunteered geographic information:Using Flickr to analyze spatial and temporal pattern of tourist accommodation. In *Progress in location-based services*, pages 57–69. 2013.

31 Yi-Fu Tuan. *Space and place: The perspective of experience*. Univ. of Minnesota, 1977.

32 United States Census. United States Census Quick Facts: Monterey Park, CA, 2010. Accessed: 2017-02-27. URL: `https://www.census.gov/quickfacts/`.

33 Stephan Winter and Christian Freksa. Approaching the notion of place by contrast. *Journal of Spatial Information Science*, 2012(5):31–50, 2012.