The Effect of Regional Variation and Resolution on Geosocial Thematic Signatures for Points of Interest

Grant McKenzie and Krzysztof Janowicz

Abstract Computational models of place are a key component of spatial information theory and play an increasing role in research ranging from spatial search to transportation studies. One method to arrive at such models is to extract knowledge from user-generated content e.g., from texts, tags, trajectories, pictures, and so forth. Over the last years, topic modeling techniques such as latent Dirichlet allocation (LDA) have been studied to reveal linguistic patterns that characterize places and their types. Intuitively, people are more likely to describe places such as Yosemite National Park in terms of *hiking*, *nature*, and *camping* than *cocktail* or *dancing*. The geo-indicativeness of non-georeferenced text does not only apply to place instances but also place *types*, e.g., state parks. While different parks will vary greatly with respect to their landscape and thus human descriptions, the distribution of topics common to all parks will differ significantly from other types of places, e.g., night clubs. This aggregation of topics to the type level creates thematic signatures that can be used for place categorization, data cleansing and conflation, semantic search, and so on. To make full use of these signatures, however, requires a better understanding of their intra-type variability as regional differences effect the predictive power of the signatures. Intuitively, the topic composition for place types such as *store* and *office* should be less effected by regional differences than the topic composition for types such as *monument* and *mountain*. In this work, we approach this regional variability hypothesis by attempting to prove that all place types are aspatial with respect to their thematic signatures. We reject this hypothesis by comparing the signature similarities of 316 place types between major cities in the U.S. We then select the most and least varying place types and compare them to thematic signatures from regions outside of the U.S. Finally, we explore the effects of LDA topic resolution on differences between and within place types.

G. McKenzie
Department of Geographical Sciences, University of Maryland, College Park, USA
e-mail: gmck@umd.edu

K. Janowicz
STKO Lab, Department of Geography, University of California, Santa Barbara, USA
e-mail: janowicz@ucsb.edu

Key words: semantic signature, point of interest, place type, geosocial, topic modeling

1 Introduction

Selecting discriminative features[1] is a key prerequisite for the classification of places into categories, which, in turn, contribute to a wide variety of tasks such as data deduplication, cleansing, recommender systems, geographic information retrieval, and so on (Bao et al., 2015). One source for extracting features is user-generated content, e.g., tags, check-ins, ratings, and many other forms of structured or semi-structured data. To give a concrete example, previous work has shown that the time of the day and day of the week users of geo-social networks check in to places is highly indicative of the place type (Ye et al., 2011; Gao et al., 2013; Shaw et al., 2013). Simplifying, the underlying assumption is that if one is provided with a sufficiently large sample of places of a given type as well as user check-ins to these places, the extracted temporal features will follow our everyday intuition that *restaurants* are visited during lunch and dinner times and that *universities* show a strong decline in visiting intensity in the evenings and during weekends. Such temporal features can be expected to effectively discriminate between restaurants and universities but will more likely fail in telling apart high schools from universities. Consequently, multiple kinds of features are combined to improve classification accuracy. Examples include features extracted from spatial second-order analysis of place patterns by type (Mülligann et al., 2011), features based on the sequence of check-ins (Cheng et al., 2013), features based on user-centric commonalities (Scellato et al., 2011), to name a few.

The examples discussed so far, make use of 2 of the 3 components of geographic information, namely space and time. The third component, specifically *theme*, can be employed to derive features as well. In fact, a multitude of work has investigated user-contributed tags. More recently, the quest for additional features has lead researcher to explore unstructured data, e.g., full text user reviews and tips, as well. Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a particularly popular technique that takes a *bag-of-words* approach to classifying documents based on the co-occurrence of words. An LDA approach produces distributions of topics that can be used to differentiate place types based on the terms and words used in describing these places. These *thematic signatures* offer an additional dimension through which place

[1] The term *feature* has varying meanings across different communities. Here, following common practice in machine learning, we will use it to refer to the measurable characteristics of points of interest as extracted from geosocial data. We will refer to geographic features as *places*.

types can be understood. In many cases, thematic signatures uncover nuanced differences between places that spatial and temporal signatures cannot.

Extracting such statistical features for places raises the important question of whether these features are stable across space. Previous work with *temporal signatures* has shown that some place of interest (POI) types vary regionally while others do not (McKenzie et al., 2015). This is an important finding as it means that features extracted from global datasets will perform well for certain types such as *drugstores* but will be less effective for other types such as *theme parks*. In this work, we extract thematic signatures from use-generated content contributed to POI across the United States and internationally to study the effect of regional variation on place types. We assume that regional effects will be even more prominent across linguistic patterns compared to temporal patterns. **The research questions and contributions addressed in this work are as follows.**

RQ1 With regards to thematic similarity, **is there regional variation between POI types?** This question will be answered by testing a null hypothesis which states that any variation in the thematic properties of POI types can be explained by sampling variation and noise. The *Chi Squared Goodness of Fit* test will be used to test this hypothesis.

RQ2 Are regional variations in topics (themes) equally prevalent across all POI types? In other words, **are some POI types more influenced by a change in region than others?** To respond to this, we compare POI types from the top three largest cities in the United States (New York City, Los Angeles and Chicago). Two methods, namely *Jensen-Shannon Distance* and *Cosine Similarity*, are used to compute dissimilarity and their results are compared for concordance via *Kendall's W*. Second, does regional (in)variance transfer across feature type hierarchies? When exploring the parent types of lower-level POI types, are certain top-level parent POI types more or less regionally prevalent?

RQ3 Having first explored POI type regional variability on a national level, we must then ask, **is the measured thematic variability in POI type consistent inter-nationally?** A subset of POI from Sydney, Australia and London, England are compared thematically with those from within the United States. Through this international comparison we will show that highly regional variant types within the United States continue to remain highly variant when compared to POI across national, cultural and physical borders. The opposite is also true for regionally invariant types.

RQ4 On the topic of resolution of thematic signatures, **does the number of selected LDA topics impact the similarity *within* and *between* place types?** By constructing a wide range of topic models, we show that the number of topics indeed influences place type similarity and that some place types are more susceptible to this influence than others.

2 Related Work

The complex relationship between language and place has been the subject of a considerable amount of research (Basso, 1996; Cresswell, 2014; Graham and Zook, 2013; Tuan, 1991; Kinsella et al., 2011; Stefanidis et al., 2013). While much of this work has focused on linguistic descriptions of place, a subset has discussed textual characteristics and the geo-indicativeness of terms and phrases. Along these lines, work by Hollenstein and Purves (2010) explored the ability of tags and textual content garnered from user-contributed geotagged Flickr photos to define local regions such as city centers. Cheng et al. (2010) build on the geo-indicativeness of language and terms to predict twitter users' rough locations based purely on the content of their tweets. Hecht and Gergle (2010) discuss the role of *localness* as it relates to contributing content to various online sources. The authors found that there are strong differences between the contributions of local users to different platforms such as Wikipedia or Flickr.

The recent rise in the use of topic models for describing and classifying documents has also played a role in the geospatial realm. Work by Adams and Janowicz (2012) has employed topic modeling to estimate the geo-indicativeness of non-georeferenced unstructured text. Additional work in this area (Adams et al., 2015) has shown that terms and phrases have probabilistic spatial extents. Combining a topic modeling approach of textual data with spatial clustering analysis and temporal check-in behavior has been successfully employed to reclassify existing POI types into unique and more semantically appropriate top-level types (McKenzie et al., 2015).

The process of extracting unstructured text in the form of tips and reviews in order to compare places and place types has also been pursued in other research areas. While not specific to regional differences, Tanasescu et al. (2013) explored the personality of venues through analysis of the keywords and textual review data contributed about a place. The authors then referenced the *five-factor model of personality* as proposed in the psychology literature to assign personality traits to places. Similarly, Hu and Ester (2013) extract data from online social posts and review sites with the purpose of location prediction and place recommendation.

3 Data and Thematic Signatures

The data used for this research was accessed from the geosocial networking application *Foursquare*. Through the application programming interface (API), 938,031 POI[2] were accessed from three cities across the United States, 437,358 from New York City, New York, 213,279 from Los Angeles, California

[2] Foursquare refers to Points of Interest (POI) as *venues*.

and 249,169 from Chicago, Illinois. These regions were selected based both on their geographic location (East Coast, West Coast, and Midwest, respectively) and the fact that they constitute the top three most populated urban areas in the United States as reported by the 2010 U.S. Census.

3.1 Place Types and Tips

Aside from geographic coordinates, attribute information attained from these POI included a *type*, which is assigned by the user contributing the POI, verified by other users, and confirmed by the application administrators. All POI types accessed via the API had been assigned types based on the Foursquare *Category Hierarchy*[3] consisting of 421 POI types. These range from types such as *Mexican Restaurant* to *Police Station* or *Mountain Top*. To ensure the validity of the thematic similarity method proposed in the remainder of this paper, only those *types* that consisted of 30 or more POI instances in each region were included. Given the specificity of some types (e.g., College Cricket Pitch) this reduced the number of POI types to 321.

Next, *tips* were accessed for each POI in the dataset. *Tips* consist of unstructured, textual comments and reviews of a POI contributed by users of Foursquare. To ensure a fair representation of the POI types, only those types to which 30 or more tips were contributed per region were included in our analysis. Otherwise some POI type may have been represented merely by a small number of terms. Cleaning the data in this way further reduced the number of POI types from 321 to 316.

3.2 Thematic Signatures

A topic modeling approach was taken to generate thematic signatures on which to measure the variability of POI types across regions. These signatures model the fact that people are likely to write about cocktails and loud music after visiting a nightclub, and contribute reviews about hiking routes, waterfalls, and camp grounds when visiting state parks. *Latent Dirichlet allocation (LDA)* (Blei et al., 2003) was employed through the use of the MALLET Toolkit (McCallum, 2002) to generate a range of topics on which the regional similarity of POI types could be assessed. *LDA* is an unsupervised, generative topic model that takes a bag-of-words approach to organize content. In this case, all of the unstructured textual data (tips) are grouped together by POI type and region (316 types $\times$ 3 regions). The co-occurrence of words across these documents is examined, exposing latent topics within

[3] `https://developer.foursquare.com/categorytree`

the data. A probability distribution of these topics is returned and can then be used to thematically define each type split by region. Since the topics remain the same across all types, the similarity of POI types and regions can be measured through calculating the (dis)similarity between probabilistic topic distributions.

4 Regional Variation

In this section, methods for exploring regional variation between different POI types are presented. The first step involves determining whether some types are regionally *invariant* (aspatial) while others are regionally *variant*. Once this has been determined, the degree to which each type is influenced by regional variations is studied as well as the sensitivity of the type to regional nuances. Last, regional type variability is abstracted to the top level of the POI type hierarchy with the goal of determining whether or not regional variability transcends hierarchy levels. The thematic signatures on which the analysis is based are constructed from 65 LDA topics. Further discussion on the resolution of thematic signatures is given in Section 7.

4.1 Significance of Regional Variations

The first task, as outline in **RQ1**, is to investigate the possibility that regional variations in thematic signatures are simply a sampling artifact and merely the result of random variation. To test this, the null hypothesis is defined stating that all POI are regionally invariant in term of their thematic signatures. Using the χ^2 *Goodness of Fit Test* (Bentler and Bonett, 1980), this hypothesis can be tested. The χ^2 Goodness of Fit Test involves comparing two distribution samples (thematic signatures) and determines the amount by which the two are statistically different. Equation 1 shows the comparison of two thematic signature distributions, P and Q, divided by the variance, σ^2, of the observation.

$$\chi^2(P, Q) = \sum \frac{(P - Q)^2}{\sigma^2} \tag{1}$$

Using thematic signatures (here topic distributions) of the same POI type from two different regions, variability of said type can be modeled via the p-value reported from the χ^2 test. Table 1 shows the percentage of the 316 POI types that are regionally variant split by regional pair and level of significance (0.1, 0.05, 0.01).

The highest percentage of regionally variant POI types exists between Los Angeles and Chicago at roughly 48% followed by New York and Los Angeles

	0.01	**0.05**	**0.1**
NY & LA	29.7%	32.9%	33.0%
NY & CHI	22.5%	25.0%	25.6%
LA & CHI	46.5%	47.5%	48.7%

Table 1: Percentage of POI types that are statistically different between regions as determined by the *Chi Squared Goodness of Fit Test*. The results for three p-values (0.01, 0.05, 0.1) are reported.

at around 33%. A discussion on possible explanations for these regional differences is presented in Section 5. These statistically variant regional results can be broken down further to look at the agreement between regions. Focusing specifically on the 0.05 χ^2 level (Table 2), we find that close to 50% of POI types are either regionally invariant or variant across all the three U.S. cities (11.1% invariant and 38.6% variant). The complimentary 50% is split between some combination of agreement between one or two of the regional pairs.

NY&LA	**NY&CHI**	**LA&CHI**	**Percentage**
1	1	1	11.1%
0	1	1	6.3%
1	0	1	9.5%
0	0	1	20.3%
1	1	0	4.4%
0	1	0	2.8%
1	0	0	6.6%
0	0	0	38.6%

Table 2: Agreement between regions on variant (0) and invariant (1) POI types, shown with percentage of overall POI types. Based on a χ^2 p-value of 0.05.

These results confirm our intuition and reject the null hypothesis stated in **RQ1**. There is regional variation in POI types and it is unlikely caused by random fluctuations, but rather by statistical differences in the thematic signatures. This finding points to a difference in words and language choices between regions which is in accordance with numerous existing studies (Tuan, 1991; Johnstone, 2004; Cheng et al., 2010; Graham and Zook, 2013). Interestingly though, not all POI types are shown to vary by region implying that the linguistic characteristics of certain POI types are less regionally specific. This is important as it implies that features extracted from global datasets will not perform well for certain types thereby affecting tasks such as classification. Given that we have shown that some types do vary by region, but others do

not, the next question asks *which* POI types are regionally (in)variant and by what amount?

4.2 Variability Between POI Types

RQ2 follows up on the previous section by examining the ways in which POI types differ between regions and the amount by which some types vary regionally while others remain invariant. Two different methods are used to assess the (dis)similarity between POI type using their thematic signatures.

4.2.1 Jensen-Shannon Divergence

The Jensen-Shannon Divergence (JSD) is a method for measuring the dissimilarity between two probability distributions *(P,Q)* (Lin, 1991). This method takes a one-to-one bin matching approach to comparing discrete datasets. The *distance metric* calculated here is computed by taking the square root of the divergence shown in Equation 2. The metric is finite, bounded between 0 and 1. *KLD* represents the *Kullback-Leibler Divergence* and is specified in Equation 3.

$$JSD(P \parallel Q) = \frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M) \tag{2}$$

$$KLD(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{3}$$

Table 3 shows the top five most dissimilar POI types as well as the top five most similar types split by regional pairs. Note that while there are difference between regions, a number of types are found across regions. *Monument / Landmark* is listed in the top five dissimilar POI types for all region pairs and *Nail Salon* is in all similar POI groups. Using Kendall's Coefficient of Concordance W (Kendall and Smith, 1939) to calculate the agreement between the *JSD* dissimilarity values across the three pairs of regions resulted in a value of 0.9 ($p < 0.01$) indicating that there is strong agreement in dissimilarity in terms of POI types between region pairs.

4.2.2 Cosine Similarity

Cosine similarity is a measure that reports the similarity between two vectors, or distributions in this case (Equation 4). Using the Euclidean dot product, the cosine of two vectors (P and Q) is computed. Provided non-negative values for P and Q, the resulting similarity value is bounded between 0 and 1.

NYC & LA	**NYC & CHI**	**CHI & LA**
Dissimilar POI Types		
Football Stadium (0.90)	Vineyard (0.90)	Football Stadium (1.00)
Rest Area (0.77)	Meeting Room (0.84)	Meeting Room (0.93)
Plaza (0.76)	Independent Theater (0.78)	Vineyard (0.90)
Monument/Landmark (0.63)	Assisted Living (0.75)	Mountain Top (0.89)
Mountain Top (0.62)	Monument/Landmark (0.73)	Monument/Landmark (0.78)
Similar POI Types		
Airport Lounge (0.00)	Nail Salon (0.02)	Mobile Phone Shop (0.00)
Nail Salon (0.05)	Pet Store (0.04)	Office Supply Store (0.03)
Barbershop (0.07)	Automotive Shop (0.06)	Nail Salon (0.05)
Dentist's Office (0.10)	Airport Lounge (0.06)	Electronics Store (0.07)
Automotive Shop (0.10)	Frozen Yogurt (0.08)	Pet Store (0.07)

Table 3: Top five and bottom five dissimilar POI types based on normalized Jensen-Shannon Divergence and split by region pairs.

$$CosSim(P,Q) = \frac{\sum_{i=1}^{n} P_i \times Q_i}{\sqrt{\sum_{i=1}^{n} (P_i)^2} \times \sqrt{\sum_{i=1}^{n} (Q_i)^2}} \tag{4}$$

Table 4 again shows the top five most dissimilar types as well as the top five most similar types split by regional pairs. Note that cosine similarity is a *similarity measure* and the values reported in the table are actually $1 - CosSim$ in order to mirror the dissimilarity values reported by the JSD. While there are clear differences between regions, a number of POI types are found between region pairs. In line with the types reported using JSD (Table 3), *outdoor* types such as *Scenic Lookout* and *Monument / Landmark* appear in the top dissimilar POI types. Comparatively, types related to *shopping* and *service* type activities such as *Optical Shop* or *Nail Salon* appear in the top similar types list. Applying Kendall's W again, the agreement between the cosine similarity values across the three pairs of regions resulted in a value of 0.5 ($p < 0.01$). While not as strong as the coefficient reported by *JSD*, this value still indicates agreement in similarity across types between region pairs.

4.3 Concordance Between Dissimilarity Measures

While JSD and CosSim are both highly popular in the LDA literature, they originate from different families of similarity measures and thus reflect different views on what *similarity* means. Hence, the two (dis)similarity measures produce individual results for inter-signature comparison. The value of these

NY & LA	NY & CHI	LA & CHI
Dissimilar POI Types		
Park (0.99)	Greek Restaurant (1.00)	Military Base (0.98)
Platform (0.98)	City Hall (0.97)	Platform (0.98)
Football Stadium (0.91)	Scenic Lookout (0.97)	Mid. East. Restaurant (0.95)
Gay Bar (0.91)	Monument/Landmark (0.92)	Conference Room (0.93)
Baseball Stadium(0.85)	Beach (0.90)	Water Park (0.87)
Similar POI Types		
Car Wash (0.00)	Synagogue (0.03)	Airport Lounge (0.01)
Dessert Shop (0.03)	Women's Store (0.05)	Motel (0.02)
Accessories Store (0.12)	Shoe Store (0.19)	Cemetery (0.20)
Trade School (0.15)	Credit Union (0.24)	Women's Store (0.27)
Optical Shop (0.18)	Optical Shop (0.26)	Nail Salon (0.28)

Table 4: Top five and bottom five dissimilar POI types based on normalized cosine similarity and split by region pairs. Note that to align with JSD, all values reported in this table are computed as *1 - CosSim*.

approaches is shown through their agreement. As was done between region pairs within each method, Kendall's coefficient of concordance W is used. Each of the three regions is compared to each other region using Jensen-Shannon Distance and Cosine Similarity. These methods produce single similarity (or dissimilarity for JSD) values for each region pair and for each POI type. *Kendall's W* is then used to calculate the concordance between these measures across all POI types.

Region Pair	Kendall's W	p-value
New York & Los Angeles	0.59	<0.02
New York & Chicago	0.56	<0.07
Los Angeles & Chicago	0.57	<0.05

Table 5: Kendall's coefficients of concordance W between Jensen-Shannon Distance and Cosine Similarity for pairs of regions.

A Kendall's W value of 1 indicates complete concordance where a value of 0 represents no concordance at all. As shown in Table 5, all W values are greater than random. This indicates a significant level of agreement between dissimilarity measures, thus reducing the possibility that the discovered similarities are simply artifacts of choosing one specific measure. Given these findings, we focus on JSD for the remaining analysis.

4.4 Hierarchy Homogeneity

Many Point of Interest type vocabularies are constructed as hierarchies[4] with subsumption relationships between subtypes and supertypes. The Foursquare POI type vocabulary follows this model by mapping every type to one of three distinct levels. The top-level into which all subtypes are assigned consists of 9 *supertypes*. For example, *Mexican Restaurant* is a subtype of *Food* and *Scenic Lookout* is a subtype of *Outdoors & Recreation*. These subsumption relationships are essential for many aspects of knowledge organization and play an important role in understanding POI type variability. Here we discuss how the regional variability of types traverses levels in this hierarchy.

As shown in Tables 3 and 4, the *most* and *least* regionally variant types are seen to be different. The most regionally variant POI types are primarily related to outdoor activities while the most regionally invariant types can be typically categorized as shops, stores or service-related places. Here we extracted the top 20% most regionally variant and 20% least regionally variant POI types (as determined by *JSD*) and grouped them by their supertypes. Figure 1 shows the distribution of these supertypes grouped by their level of regional variability. The most invariant types clearly consist of primarily *Shop & Service* types while the most regionally variant POI types have a higher *information entropy*, distributed more evenly across the parent supertypes with peaks in *Outdoors & Recreation* and *Travel & Transport*.

In summary and with respect to the second part of **RQ2**, POI hierarchies, such as Foursquare's are not completely homogeneous with regards to the regional (in)variability of POI types There are clear differences in the top-level types that are regionally variant and those that are not. This is important as it implies that there are larger sets of types that do not vary significantly across space and thus can be learned from global datasets, while POI related to travel and outdoor themes need to be approached from a more local perspective.

5 Exemplary Investigation of Differences in Thematic Signatures

In this section, a subset of the POI types are examined further with the aim of better explaining the regional variability, or lack thereof, within and between types.

To gain a better understanding of the regional variability results presented in Table 1, one must understand the regions of interest. Relative to the layout of the United States, the cities of New York, Los Angles and Chicago vary greatly in their geographic locations. The climate deviates significantly

[4] e.g., Schema.org or the place hierarchy used by the Ordnance Survey.

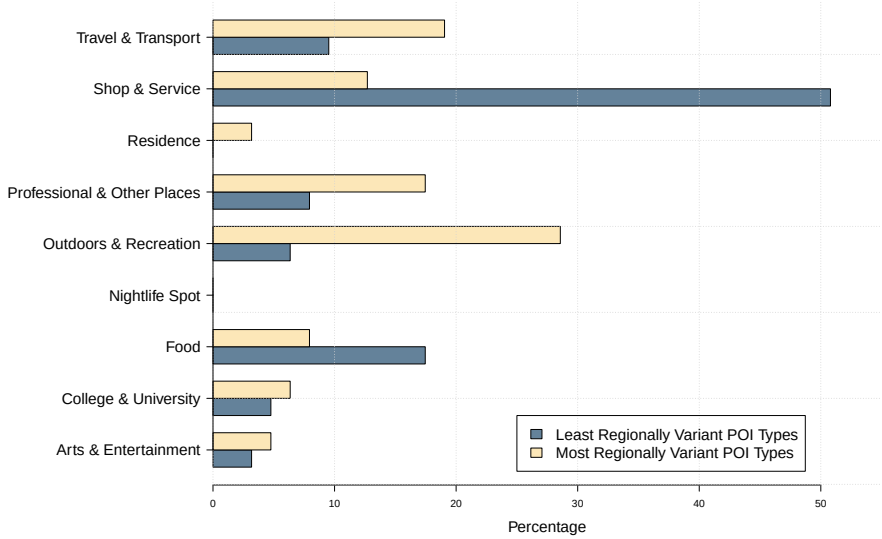

Fig. 1: Top level class for the top 20% (63) regionally dissimilar POI types as well as the top 20% (63) regionally similar POI types.

between these cities with Los Angeles in a Mediterranean climate and New York and Chicago consisting of humid/continental climates. These climates allow for very different types of activities, specifically those *outdoor* activities that take place at POI types shown to have high variability (see Figure 1). In addition to climate, there are important cultural differences between these regions. New York City, the "Gateway to America," is often viewed as a *melting-pot* of diverse cultures from around the world. Los Angeles, while still a multi-cultural city, is composed of a large Latino immigrant population with very different traditions and cultural backgrounds.

The POI type that consistently shows the highest level of thematic dissimilarity (regional variance) across all regions according to the *JSD* metric is *Monument / Landmark*. Figure 2 show the topic distribution for this type across the three U.S. cities using 65 LDA topics. Note that for visualization purposes the cube root of the distribution values are shown in order to make the very low topic values visible. Three of the most prominent topics are displayed as *word clouds* below the distribution. Not surprisingly, words such as *exhibit*, *visit* and *admission* appear to contribute significantly to thematically defining this POI type. In comparing the topic distributions by region, we see that there is considerable disagreement between the regions as to which topics contribute to the type. Topics 1 and 55 contribute strongly to this type in Chicago while Topics 13 and 34 are more important, and unique, for defining *Monuments* in Los Angeles.

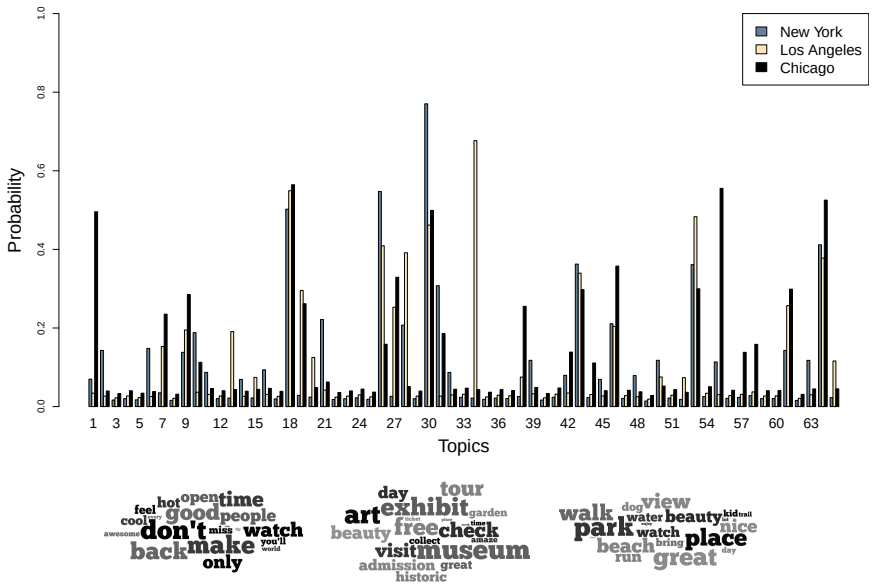

Fig. 2: The POI type *Monument / Landmark* depicted as a probabilistic distribution of topics, split by region. The top words contributing to the most prevalent topics are shown below the graph. Note for visibility, as most values are close to zero, the cube root of the data is shown.

A *Monument or Landmark* is, almost by definition, something that is unique to the region in which it exists. Hence, the terms and language used to describe a landmark such as *The Hollywood Sign* in Los Angeles, CA are quite different than those used to describe the *Cloud Gate* in Chicago, IL or the *Grand Central Terminal Clock* in New York City, NY. Not only does this type invite the use of regionally specific terms and linguistic characteristics from locals, but POI of type *Monument / Landmark* are also the focus of many tourists or visitors to a region. These non-locals often visit these places from locations outside of the United States and bring with them their own unique descriptive terms and phrases.

An alternative POI type example is one that is highly regionally invariant. The type *Nail Salon* fits this description and is shown to have one of the lowest regional variation values reported by both *JSD* and *CosSim*. As opposed to the type *Monument / Landmark*, Figure 3 visually depicts a higher topic probability agreement between regions. In most cases, the three U.S. regions agree on the prominent topics contributing to thematically defining *Nail Salons*. As shown in the word clouds below the distribution graph, topics 25, 43 and 61 use words such as *nail*, *hair*, *service* and *staff* to define this type.

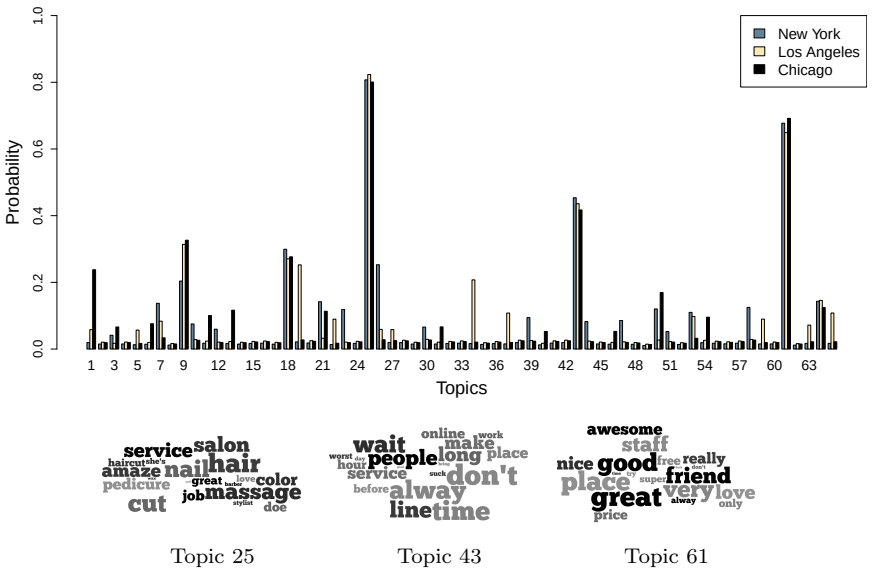

Fig. 3: The POI type *Nail Salon* depicted as a probabilistic distribution of topics, split by region. The top words contributing to the most prevalent topics are shown below the graph.

Given the agreement in topic probability across regions it is not surprising that this type is considered highly regionally invariant.

Nail Salons fit into what is often referred to as the *Service Industry*. Many service industry types appear in the regionally invariant group and most of them are based on activities that occur with regular frequency. The type *Nail Salon* is prototypical of this group as it involves an activity that occurs semi-regularly and traditionally involves a specific customers group, e.g., defined by gender and socio-economic status (Kang, 2010). In addition, the focus of the services conducted at this POI type are highly specific to a part of the human body. In contrast to *Monument / Landmark*, this implies that the terms and words used to describe the type are more focused on certain theme (of nails) and less influenced by the geographic region in which the POI exists.

6 International Regional Variability

Up to this point, the focus of this research has been on understanding the regional variability of textual descriptions of POI types *within* the United States. The next step is to examine regional variation as it pertains to places

outside of the United States. While the previous sections have shown that there are differences within a single country, **RQ3** questions the level of (dis)similar of POI types between countries. To answer this question, data from two cities outside of the United States were accessed. *Sydney, Australia* (SYD) and *London, England* (LDN) were selected based on the facts that (1) English is the primary language spoken and written in both regions, (2) technologically, both regions are quite similar to the United States wrt the use of location-based services, and (3) geographically, both regions are far away from the U.S. cities in this study.

The data accessed from Foursquare consists of 5,717 venues, 49,711 tips (SYD) and 8,179 venues, 132,193 tips (LDN). With the limitation that analysis requires a minimum of 20 venues and 20 tips per POI type, due to lighter usage of Foursquare outside the U.S., this reduced the number of types on which regional similarity could be calculated to 96 (SYD) and 172 (LDN). For this reason, these regions were not included in the original analysis, but instead are the focus of supplementary analysis. This subsequent work involved running a separate topic model with a reduced set of POI types in order to compare these new regions with those in the U.S.

6.1 POI Type Similarity Comparisons

Based on the JSD values reported for the three U.S. cities, 20 of the highest regionally variant types were selected as well as 20 of the highest regionally invariant types. POI types such as *Nail Salon*, *Pet Store* and *Doctor's Office* are examples of high regionally invariant types while *Monument / Landmark*, *Scenic Lookout* and *Skate Park* contributed to the group of high regionally variant types. Not all types that appeared in the top intra-U.S. JSD values were chosen as, for example, types such as *New American Restaurant* had insufficient data.

The Jensen-Shannon distance was calculated between POI of the same type across all five pairs of regions (New York, Los Angeles, Chicago, Sydney and London). The JSD values were then averaged across the 20 regionally variant and the 20 regionally invariant types independently. Figure 4 shows these *mean JSD* values for each group of types split by region pair. The JSD dissimilarity values between pairs of U.S. cities is notably less than the international comparisons, both for the high regionally variable types as well as the high regionally invariant. One should also note that those POI types that are influenced by region changes within the US are also highly regionally influenced when compared internationally, and vice versa.

In response to **RQ3**, this demonstrates that regional (in)variability with respect to POI types exists both within the U.S. and internationally. Furthermore, highly variable types within the U.S. are also highly variable when compared to regions outside of the U.S. and this trend holds for regionally

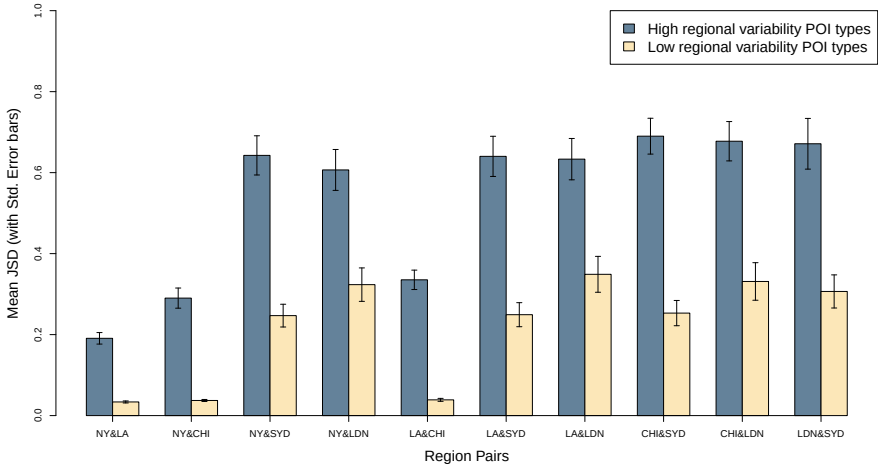

Fig. 4: The mean *JSD* of the top 20 regionally dissimilar POI types and top 20 regionally similar POI types along with standard error bar split by region pairs.

invariant types as well. These results have strong implications for the development of global POI thematic signatures suggesting that at least a subset of types can be sufficiently described by such global signatures even at an international level.

7 Thematic Resolution

One of the more difficult aspects with taking an unsupervised topic modeling approach to exploring thematic similarity is choosing and justifying the number of topics used in such analysis. Existing work in this area (Arun et al., 2010; Griffiths and Steyvers, 2004) has presented approaches for choosing the number of topics to use in LDA topic modeling. The finding in both of these cases is that determining the number of "correct" topics is often complex, hard to validate, and dependent on the underlying dataset.

In this section, we take an exploratory analysis perspective to understand the affect of changing the number of topics on the similarity of POI types. Specifically we are interested in examining the process through which a corpus of words contributed to a type, such as *Park*, begins to merge with, and possible become indistinguishable from all other types in a sample dataset. Furthermore, how do types differ from one another as the number of topics increase (**RQ4**)? To conduct this analysis, the text from all tips across the three U.S. regions were collected at the *instance* level and tagged by their POI type. A series of LDA topic models were constructed from this data ranging

in topic number from 10 to 500 in increments of 10. The resulting place type topic distributions were then analyzed in two different ways. First, Kullback-Leibler Divergence (KLD) similarity was calculated *within* a specific type (e.g., Park). Similarity was measured from each instance of the type to each other instance of the same type. Second, KLD similarity was measured from each instance of a given type (Park, in this example) to all other instances of *other* types (e.g., Donut Shop, University). Computing these two sets of KLD values for each type and for each topic number allowed us to examine how the number of topics influences *within* type similarity and *between* type similarity. Figure 5 shows a sample of five of these POI type KLD densities split by the number of topics.

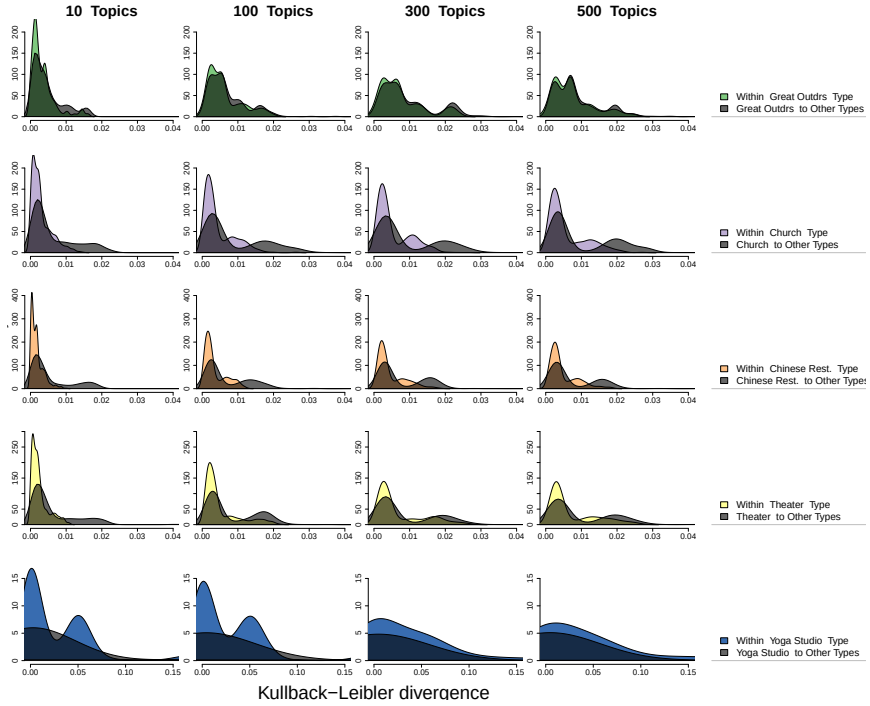

Fig. 5: Kullback-Leiber Divergence density graphs for five types shown as the number of LDA topics increase. The colored graphs show density graphs for KLD values for POI instances within the type. The gray density graphs show KLD values for POI instances within the listed type to all place instances not associated with the given POI type. Note that the Y-axis is different for each row in this figure.

These densities show a number of interesting phenomena. First, we find that regardless of the POI type, very low numbers of topics show a relatively high peak of similarity within the given type (peak around 0 on the X-axis).

As the number of topics increases, the dominant peak at 0 for *within* place types decreases, in all cases. While we see a small decrease in peakedness of the *between* POI types as the number of topics increases, it is less pronounced indicating that *between* POI type similarity, at least for these example cases, is less influenced by a change in the number of topics.

There is a notable difference when comparing POI types to one another. All types show some percentage of overlap between the *within* and *between* KLD densities but the amount of overlap and magnitude differs across POI types. *Great Outdoors* for example shows a high amount of overlap at all number of topics with extremely high overlap in the high number of topics (e.g., 300 and 500). *Yoga Studio* on the other hand shows a relatively low amount of overlap. These overlap values are shown in greater detail in Figure 6.

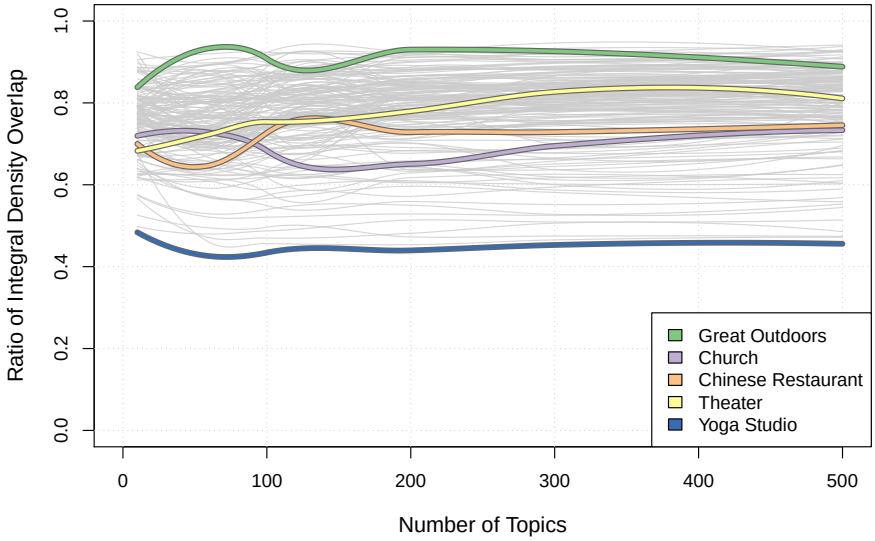

Fig. 6: Ratio of integral overlap between *within* type and *between* type densities shown varying by the number of topics. The gray lines show all POI types while the colored lines show a selection of five types.

This figure shows all of the POI type overlap ratios as gray lines with our select five types highlighted. It was constructed by calculating the amount of overlap in the *within* and *between* KLD density plots. The integral, or the area under the density curve, was computed for both the between and within KLD densities for all topics between 10 and 500 at 100 topic intervals. The within and between density integrals were merged for the total area and the *ratios* were calculated as the overlap divided by this merged value. As was shown in Figure 5, we see a range of density overlaps across our sample types. *Great Outdoors* shows a consistently high ratio of overlap across all

topics. This reflects the *catch-all* nature of this place type. In the Foursquare category hierarchy, *Great Outdoors* is a higher-level category. It is therefore not surprising that the terms used within the category would be very similar to terms used outside of the category. In contrast, we find the type *Yoga Studio* to have a low ratio of overlap. A *Yoga Studio* is a very unique type of place which focuses on a specific activity and textual content about yoga studios are likely contributed from a narrow demographic of people. It is therefore not surprising that the overlap of words related to this type with terms from other types is relatively small.

Figure 6 also depicts a larger amount of overlap variance with smaller topic numbers than with larger ones. There is notable change in all types between 10 and 200 topics with an increase in the topic number after 200 showing much more limited influence. With respect to **RQ4**, we find that the influence of the number of topics, i.e., the resolution, on thematic similarity between POI types is dependent on the POI type itself and our findings confirm there is no hard and fast rule for determining the correct number of topics for place type similarity analysis.

8 Conclusions and Future Work

The terms used to describe places of interest play an important role in differentiating them from one another, thereby forming a prerequisite for classification, retrieval, and recommender tasks. Intuitively, reviews, social media postings, news, and so forth, about state parks are more likely to use terms such as trail, hike, and landscape, than descriptions of other places, say universities. This raises the important question how regional such type-based key characteristics are, i.e., whether we can learn a single embedding for POI types from global datsets or not. By constructing *thematic signatures* through topic models build on unstructured, user-generated text, we show that some types of places vary regionally, e.g., Monuments/Landmarks, while others do not, e.g., Nail Salons. Not only does this hold true for regions within the US, but also for many other English speaking regions. These findings are important as they speak to linguistic and cultural differences and similarities between people and the ways in which they interact with their environment. Furthermore, this work shows that regional variability does traverse levels in a place type hierarchy, e.g. generally, outdoor place types are more regionally variant, and that the resolution of the topics does impact the differentiation of place types.

Future work in this area will focus on exploring additional methods for constructing thematic signatures to better understand the robustness of the findings. Latent Dirichlet allocation has been shown to be a reliable model for studying similarities between places, but further research in this area will benefit from the incorporation of additional language-based similarity

models. A method that combines these thematic signatures with previously constructed temporal signatures is currently underway with the goal of producing a robust set of *semantic signatures* for use in place-based activity behavior research.

References

Adams, B. and K. Janowicz (2012). On the geo-indicativeness of non-georeferenced text. In *ICWSM*, pp. 375–378.

Adams, B., G. McKenzie, and M. Gahegan (2015). Frankenplace: Interactive thematic mapping for ad hoc exploratory search. In *24th International World Wide Web Conference*. IW3C2.

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining*, pp. 391–402. Springer.

Bao, J., Y. Zheng, D. Wilkie, and M. Mokbel (2015). Recommendations in location-based social networks: a survey. *GeoInformatica 19*(3), 525–565.

Basso, K. H. (1996). *Wisdom sits in places: Landscape and language among the Western Apache*. UNM Press.

Bentler, P. M. and D. G. Bonett (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin 88*(3), 588.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *The Journal of machine Learning research 3*, 993–1022.

Cheng, C., H. Yang, M. R. Lyu, and I. King (2013). Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*, Volume 13, pp. 2605–2611.

Cheng, Z., J. Caverlee, and K. Lee (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768. ACM.

Cresswell, T. (2014). *Place: An Introduction*. John Wiley & Sons.

Gao, H., J. Tang, X. Hu, and H. Liu (2013). Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 93–100. ACM.

Graham, M. and M. Zook (2013). Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. *Environment and Planning A 45*(1), 77–99.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(suppl 1), 5228–5235.

Hecht, B. J. and D. Gergle (2010). On the localness of user-generated content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 229–232. ACM.

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

Hu, B. and M. Ester (2013). Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 25–32. ACM.

Johnstone, B. (2004). Place, globalization, and linguistic variation. *Sociolinguistic Variation: Critical Reflections*, 65–83.

Kang, M. (2010). *The managed hand: Race, gender, and the body in beauty service work*. University of California Press.

Kendall, M. G. and B. B. Smith (1939). The problem of m rankings. *The annals of mathematical statistics 10*(3), 275–287.

Kinsella, S., V. Murdock, and N. O'Hare (2011). I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 61–68. ACM.

Lin, J. (1991). Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on 37*(1), 145–151.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

McKenzie, G., K. Janowicz, S. Gao, and L. Gong (2015). 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.

McKenzie, G., K. Janowicz, S. Gao, J.-A. Yang, and Y. Hu (2015). POI Pulse: A multi-granular, semantic signatures-based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization 50*, 71–85.

Mülligann, C., K. Janowicz, M. Ye, and W.-C. Lee (2011). Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *Spatial information theory*, pp. 350–370. Springer Berlin Heidelberg.

Scellato, S., A. Noulas, and C. Mascolo (2011). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1046–1054. ACM.

Shaw, B., J. Shea, S. Sinha, and A. Hogue (2013). Learning to rank for spatiotemporal search. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 717–726. ACM.

Stefanidis, A., A. Crooks, and J. Radzikowski (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal 78*(2), 319–338.

Tanasescu, V., C. B. Jones, G. Colombo, M. J. Chorley, S. M. Allen, and R. M. Whitaker (2013). The personality of venues: Places and the five-factors ('big five') model of personality. In *Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference on*, pp. 76–81. IEEE.

Tuan, Y.-F. (1991). Language and the making of place: A narrative-descriptive approach. *Annals of the Association of American geographers 81*(4), 684–696.

Ye, M., K. Janowicz, C. Mülligann, and W.-C. Lee (2011). What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 102–111. ACM.