AGILE Association of Geographic Information Laboratories in Europe

Uncovering spatiotemporal biases in place-based social sensing

Grant McKenzie[1], Krzysztof Janowicz[2], and Carsten Keßler[3]

[1] Platial Analysis Lab, McGill University, Montréal, Canada
[2] STKO Lab, University of California, Santa Barbara, USA
[3] Aalborg University, Copenhagen, Denmark
grant.mckenzie@mcgill.ca

Abstract. Places can be characterized by the ways that people interact with them, such as the times of day certain place types are frequented, or how place combinations contribute to urban structure. Intuitively, schools are most visited during work day mornings and afternoons, and are more likely to be near a recreation center than a nightclub. These temporal and spatial signatures are so specific that they can often be used to categorize a particular place solely by its interaction patterns. Today, numerous commercial datasets and services are used to access required information about places, social interaction, news, and so forth. As these datasets contain information about millions of the same places and the related services support tens of millions of users, one would expect that analysis performed on these datasets, e.g., to extract data signatures, would yield the same or similar results. Interestingly, this is not always the case. This has potentially far reaching consequences for researchers that use these datasets. In this work, we examine temporal and spatial signatures to explore the question of how the data acquiring cultures and interfaces employed by data providers such as Google and Foursquare, influence the final results. We approach this topic in terms of biases exhibited during service usage and data collection.

Keywords: place type, point of interest, bias, alignment

1 Introduction

As the field of Geographic Information Science grows to address the heterogeneity of data being produced today (e.g., mobile sensor data, digital social footprints, etc.), we are becoming increasingly concerned with the question of how humans conceptualize and categorize their environment. Affordance theory [1] describes how these categories form from the interaction of agents with their environment. For urban spaces, for instance, places can be categorized by the activities they afford into types such as cafés, offices, or hospitals. Each of these place types is characterized by a temporal activity footprint, we refer to as a *signature*, that arises from the fact that humans visit cafés in the morning, offices during weekday business hours, and hospitals throughout the day/week

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020

with peaks on the weekends, holidays, and during the winter season. In fact, these signatures are type-specific to a degree where they can be used to tell apart and categorize places based on the times they are frequented [2, 3]. Today, most of these signatures are generated through activity surveys or social sensing, i.e., from user-generated content. This, however, begs the question of how factors such as *perceived* social capital and privacy concerns impact the creation of truthful signatures given that humans are more likely to check-in at a trendy restaurant than a dermatologist's office. Furthermore, how do the interface limitations of the social media applications (e.g., users do not decide when they are *checked out* after *checking in*) impact these signatures, or the demographics of the application users? Are some of the identified patterns merely a function of how many place types a certain system supports? A lot of existing activity-based research has relied on these temporal patterns as truthful reflections of real-world human behavior while a small, but growing, amount of evidence indicates that there is little consistency between the different platforms[4, 5]. Similar to work on the data quality of Volunteered Geographic Information (VGI) [6], previous work has discussed the general biases that affect data collection [7]. Little empirical research, however, has quantified the biases inherent to check-in activities and signatures as such. This is a difficult undertaking as it requires ground truth data on which to compare user-generated temporal signatures.

We propose to make use of another, recently accessible dataset, namely *Popular Times*, temporal place profiles released by Google.[4] In contrast to geosocial check-in-oriented platforms such as Foursquare's Swarm,[5] users of mobile devices are passively identified as being at a place without actively deciding to check-in. Their *platial* location is inferred based on location information ascertained through *Google's Location Services*, a feature built in to many mobile devices on the market today. In order to use a mobile application such as Google Maps, Google Location Services must be enabled, both for Android and Apple iPhones. This service intermittently collects location information on millions of users who have enabled this service, forming the basis of their *popular times* feature. Given the size of their market share, these temporal signatures represent a broader demographic of the population than a geosocial media company such as Foursquare. The passive vs. active data collection approaches feeding these temporal signatures also speak to the different inherent biases of the platforms. They also have numerous ethical implications.

In theory, Google's passively fed temporal signatures should eliminate biases related to social capital and demography that are likely present in the Foursquare signatures. To test this theory, we compare the place type-level signatures mined from Foursquare and Google and discuss the arising differences framed through a number of different *biases*. To accomplish this, we first align the place type taxonomies from both data providers using a place instance co-occurrence matching method. This allows us to compare the temporal signatures from both data providers, further examining the variation between aligned place types.

[4] https://support.google.com/business/answer/6263531?hl=en
[5] https://www.swarmapp.com/

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

Lastly, we shift our focus away from the temporal dimension to explore the biases inherent in the contribution of places to these different data providers. Existing work has demonstrated that the spatial distribution of places plays an important role in differentiating place types [8]. Bars, for example, tend to cluster together whereas post offices are dispersed at regular intervals. The nuance of the category assigned to a place is important though as the clustering pattern of bars in one dataset may be more similar to the clustering patterns of pubs (not bars) in another. Continuing our focus on biases present in geosocial media data, we investigate the differences in spatial point pattern signatures with an eye on how they are contributed and the differences in place type taxonomies.

2 Related Work

User-generated geographic content, volunteered geographic information, and geosocial media data have formed the basis for a considerable amount of place-focused research in recent years. Stemming from a strong foundation in gazetteer research [9, 10], much of this focus has been on matching and conflating points of interest datasets [11, 12]. This is often done with the goal of gaining a better understanding of human activity and travel behavior through a combination of different datasets from different providers [13]. While significant efforts have previously targeted place instance matching, there is a genuine need to align different POI datasets at a place type level. There are commonalities that can be identified in places of the same type, such as the types of activities that they afford [14] and the demographics of visitors [15]. Quantitatively, these activity affordances are reflected in temporal visiting behavior and the spatial distribution of places. Temporal activity patterns have been identified and used in a range of work including everything from differentiating places based on temporal visiting behavior [3] to enhancing reverse geocoding services [16]. The spatial distribution of places and geographic features have also been used to differentiate place types [17] and identify similar spatial patterns in feature types across datasets [8]. These two types of signatures built from data aggregated at the place type level are often used as the foundation on which to examine changes in human activity behavior. The difficulty is that very little is truly understood about the biases inherent in these signatures.

At a broader scale, a rich literature has explored the biases associated with user-generated content and social media data. Biases related to the users contributing data to OpenStreetMap have been identified [18] as having contribution biases towards specific geographic regions [19]. Rost et al. [20] specifically studied check-ins on the Foursquare platform arguing that the platform is not really a "location-based service," but rather functions as a method for communication and sharing location information between friends. Furthermore, Tang et al. [21] identify two forms of location sharing in users of geosocial media applications, namely social-driven sharing and purpose-driven sharing. Works such as these highlight the need to further investigate the biases associated with these geospatial and place-based datasets.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

3 Data

We accessed information related to points of interest (POI) within the geographic boundary of the state of Maryland and the District of Columbia in the United States using the public application programming interfaces (API) provided for Google Places[6] and Foursquare.[7] The same exact same geographic boundaries were used both cases. In total we accessed 185,666 Google POI and 229,307 Foursquare POI. From these data, the following attributes were accessed: Geographic coordinates, name, and place type. Foursquare POI are classified with a single place type from the Foursquare taxonomy, while Google POI are classified with one or more place types from the Google Places taxonomy. For this research, the first (and finest resolution) place type was used when multiple place types were present. The Foursquare data contains 677 unique place types. A full list of the Foursquare Venue (POI) types is available at https://developer.foursquare.com/docs/resources/categories. The Google Places data contains 105 unique place types. The Google places taxonomy is available at https://developers.google.com/places/supported_types. For simplicity we will refer to the set of Google and Foursquare POI as POI_{Gi} and POI_{Fi}, respectively. POI_{Gt} and POI_{Ft} will reference the respective sets of place type taxonomies for each provider. Lowercase subscripts reference individual *i*nstances or *t*ypes within the datasets such that $POI_{gi} \in POI_{Gi}$ and $POI_{gt} \in POI_{Gt}$.

3.1 Temporal Signatures

In addition to the previously mentioned POI attributes, temporal data were accessed for the two sets of POI. *Popular Times*[8] were accessed for POI in POI_{Gi} resulting in a popularity value for every hour of the day over the course of a typical week. While popular times were requested from all POI_{Gi}, only 18,016 (9.7%) returned this attribute. These popular times were then aggregated by place type and an average set of popular times was calculated for each place type in POI_{Gt}. The Foursquare POI do not include temporal visiting behavior collected passively, but were instead generated through active POI-based geosocial check-ins. Check-ins to POI_{Fi} were accessed every hour over four months and split by Foursquare place type. These were then averaged as hours of a typical week, producing a set of POI_{Ft} temporal signatures. In previous work, it has been shown that such temporal signatures and their bands are type-indicative to a degree where places can be categorized into their proper types based on the times they are visited [16, 2].

[6] https://developers.google.com/places/web-service/details
[7] https://developer.foursquare.com/docs/api/venues/details
[8] Google uses "aggregated and anonymized data from users who have opted in to Google Location History" to compute *popular time* values. https://support.google.com/business/answer/6263531

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

3.2 Spatial Signatures

A wide array of metrics exist for the quantification of point processes, and, hence, for the creation of type-specific *spatial signatures*. *Ripley's K* [22] is a popular descriptive statistic for detecting deviation of a place type from spatial homogeneity. The K function is defined in Equation 1 where d_{ij} is the Euclidean distance between consecutive points $(i,\ j)$ in a set of n points, h is the scan distance, and A the area. I is the indicator function, returning 1 if true, 0 if false.

$$K(h) = (\frac{n}{A})^{-1} \sum_{i \neq j} \frac{I(d_{ij} < h)}{n} \tag{1}$$

Here we use a variance stabilized version (Ripley's L) defined as $(K(d)/\pi)^{1/2}$ as a simple means to establish signatures as it is well suited for comparisons since it controls for variance within each of the patterns. We calculated Ripley's L for all POI_{Gt} and POI_{Ft} resulting in characteristic curves for each place type in both datasets. For a detailed overview of spatial signatures and applicable methods; see [8].

3.3 Data and Software Availability

All relevant analysis scripts supporting this publication are available at https://github.com/ptal-io/TemporalBiases. The `R` and `PHP` scripts are split by analyzes, namely *Place Matching*, *Temporal Comparison*, and *Spatial Comparison* and released under BSD license. In addition, the temporal data access scripts used in this project are available at https://github.com/apollojain/popular_times. Research data used in this project is not publicly available due to the providers' terms of use, which prohibit re-distribution or re-publication of their data.[9] As all of these data were collected through the free-tiers of the public-facing APIs (URLs provided in Section 3), the analysis can be reproduced by accessing the Foursquare and Google data at the same temporal and spatial resolution reported in this work.

4 Place Type Alignment

First we align the place type taxonomies from Google and Foursquare by matching place instances between both datasets. Through this we can observe place classifications applied from both data providers. This alignment stage is very important for our work as we want to study differences in the data, e.g., whether people want others to know that they visited a place, while keeping the places themselves invariant. Figure 1 shows a single real-world place named *Donut Connection* identified by POI instances from both platforms. Each of these instances includes a place type assigned from both POI_{Gt} and POI_{Ft}. *Donut Shop*$_F$ was assigned to the Foursquare instance while *Café*$_G$ was assigned to the Google instance.

[9] https://foursquare.com/legal/terms

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

Fig. 1: *Donut Connection*, a single POI in Maryland, geographically identified by two place instances, one from Google (G) and one from Foursquare (F). Reference imagery by Digital Globe.

Place instance matching was done as follows. Each POI in the POI_{Gi} dataset was queried against all POI_{Fi} within 100m. This query distance was determined based on previous findings that the average distance between the same POI in two different datasets (Foursquare and Yelp) is 62.8 meters [23]. We then calculated *Levenshtein* distance between the name of each POI_{gi} and the name of each potential POI_{Fi} matched within the 100m radius. The resulting value represents the minimum number of character changes that must take place for one sequence to be changed to match the other. Any POI_{fi} name resulting in a Levenshtein distance greater than 0 (not a perfect match) was removed. If multiple POI_{fi} remained, the POI_{fi} closest in proximity to the POI_{gi} was identified as the match. While this is a simple approach for determining place instance matches, it is overly conservative by design – only allowing exact place name matches within 100 meters of each other. Given the number of POI available in these two datasets, we elected to be overly cautious and err on the side of false negatives rather than false positives. Through this approach, we matched 20,657 place instances, or 11% of POI_{Gi} to POI_{Fi}.

Following the matching process, we construct a co-occurrence matrix by counting the number of times each POI_{gt} co-occurred with a POI_{ft} at the same place instance. This matrix provides insight into how varied the two taxonomies are when applied to real-world points of interest. For example, the type $Café_G$ was assigned to 327 place instances which co-occurred with 35 different POI_{Ft}. The top 14 of these (those with co-occurrence counts more than 1) are shown in Figure 2. While some of these types are less intuitive, an argument can be made for each of them; Bar_F could refer to Cafés that serve wine, for example.

5 Usage Biases

Provided this basic place type alignment, we next investigate the nuanced differences between the place type temporal signatures with an eye towards factors

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

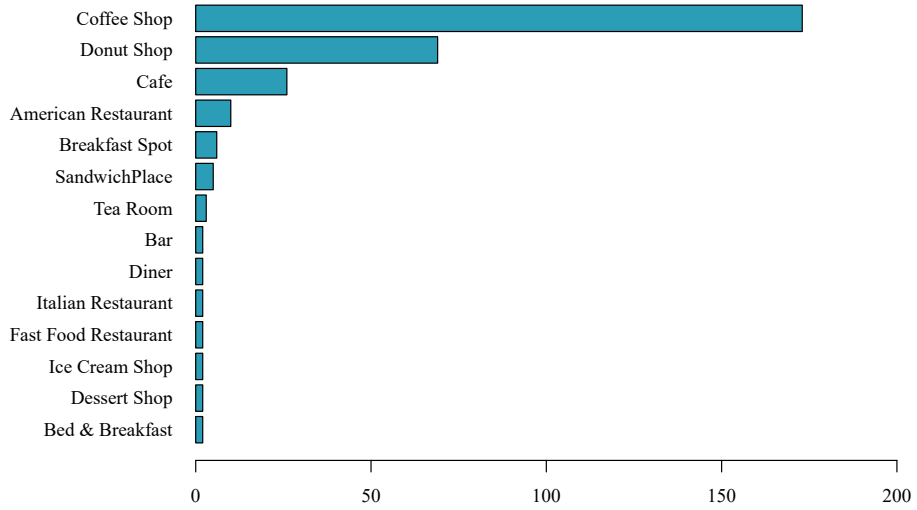


Fig. 2: Co-occurrence counts of Foursquare place types to the $Café_G$ place type. A total of 173 (53%) $Cafés_G$ are labeled as $Coffee\ Shop_F$ in Foursquare. Aside from the Foursquare place types shown here, an additional 21 place types in Foursquare also aligned with $Café_G$, one instance each.

that contribute to this difference. These discrepancies are examined from three perspectives, **(1)** social saliency bias, **(2)** user demographic bias, **(3)** interface and interaction bias, and **(4)** activity affordance bias.

First, we quantify the differences between temporal signatures of place types. *Cosine similarity* is used to measure the similarity between two vectors of equal dimensionality, or temporal signatures in our case. This produces a value bounded between 0 and 1 that can be used to compare place types based on activity times. We calculate cosine similarity between the temporal signatures for all POI_{Gt} and those for the aligned POI_{Ft}. The alignment is based on the place instance co-occurrence approach introduced in Section 4 and the POI_{ft} with the largest number of co-occurrences with a POI_{gt} is taken as the aligned place type. For example the similarity value of $Café_G \rightarrow Coffee\ Shop_F$ is 0.945, a value indicating a high degree of similarity between the two temporal signatures. Compare this to $Stadium_G \rightarrow Stadium_F$ an alignment that results in a temporal similarity value of 0.560. The ten most similar and ten least similar place types are reported in Table 1. Further examination of the place types in these lists identifies commonalities that are discussed in greater detail in the following sections.

5.1 Social Saliency

The influence of POI salience has a long history in navigation and wayfinding [24, 25]. The *social* salience of a place is often driven by the social capital that one

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

Table 1: The most and least similar place types between Foursquare and Google as determined by cosine similarity of the temporal signatures.

Most Similar Place Types		*Least Similar Place Types*	
Place Type	**CosSim**	**Place Type**	**CosSim**
Department Store	0.980	Lawyer	0.481
Park	0.967	Stadium	0.560
Liquor Store	0.957	Funeral Home	0.633
Shoe Store	0.955	School	0.644
Movie Theater	0.950	Hardware Store	0.666
Gym	0.949	Train Station	0.667
Jewelry Store	0.949	Synagogue	0.675
Bar	0.957	Insurance Agency	0.695
Clothing Store	0.947	Fire Station	0.698
Café/Coffee Shop	0.945	Travel Agency	0.699

gains not just from visiting a place, but making others aware of this fact [26]. To that end, users of geosocial media applications such as Foursquare choose to share their place-based *check-ins* with friends or the public, often with the goal of gaining social capital from an interaction with a specific place type. For example, being at a trending bar on a Friday night is more likely to increase a student's social capital (or perceived social capital) than visiting the dentist. The place type *Bar* in this case has a higher social saliency than a *Dentist's Office*. While most would agree with this assessment of these two place types, the relative social saliency of many other place types is less intuitive.

We theorize based on the data that the more socially salient the place type, the more similar the Google and Foursquare temporal signatures will be. For example, the temporal signature for Bar_G will reflect the times that visitors' mobile devices are physically detected at a bar. Foursquare users, on the other hand, will want their friends to know that they are at the bar and so will elect to share their platial location leading to an agreement between the information that is shared passively through Google's location services and the information shared actively by the Foursquare user. While the Google temporal signatures are likely to also record employees (less likely to assign social saliency to their place of employment), the overwhelming majority of visits are from customers and thus will increase activity during the expected popular times for a typical bar. In comparing the top most similar place types between providers to the bottom (Table 1), one could easily argue that those in the most similar set are more socially salient than those in the least similar set. In other words, Foursquare users presume they will gain more social capital through sharing their presence at a place type from the set on the left than on the right.

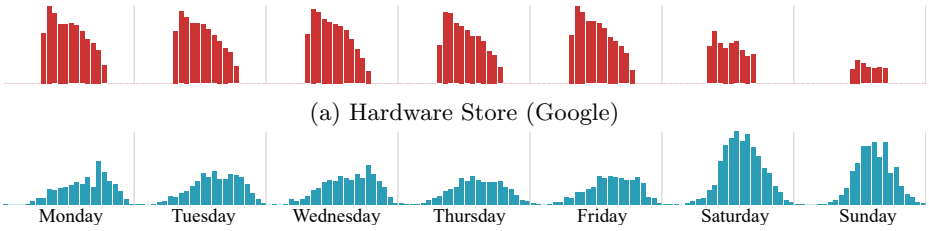
5.2 User Demographics

Target demographics for geosocial media platforms are notoriously difficult to ascertain but the most recent numbers [27] indicate that most Foursquare users

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

are between the ages of 25-34, have attended a college or university, and make between $28k-$58k per year (accounting for inflation). Knowing this, it is reasonable to assume that visiting a hardware store on a weekday offers little social capital to the typical Foursquare user. Not only are hardware stores not particularly socially salient (during the working week), they also highlight how different Foursquare's users are from Google's sample of the population.

The Google and Foursquare temporal signatures for *Hardware Store* are shown in Figure 3. The temporal signature representing Google visiting behavior (Figure 3a) largely reflects the typical visitor to a hardware store, namely trades workers procuring materials for their jobs. Throughout the work week we see a peak in the early morning trailing off by roughly 5pm with far fewer visits on the weekends. By comparison (Figure 3b), these same hardware stores in the Foursquare data show a very different temporal pattern. Weekends are much more popular for check-ins than weekdays and there is an increase in activity in the afternoon, not the morning. What we can learn from this is that a Foursquare user is unlikely to be a trades person or constructor worker given the difference in temporal signatures. Instead, these check-ins reflect casual visitors that may want to share their experience of picking up plants or starting a DIY project in their spare time. The very early morning and late evening Foursquare check-ins are likely due to the existence of 24hr hardware stores as well as erroneous check-ins and some likely mis-categorized places.



(a) Hardware Store (Google)

(b) Hardware Store (Foursquare)

Fig. 3: Temporal signatures for *Hardware Store* in (a) Google and (b) Foursquare.

Based on the variation between these two signatures, we argue that in general there is less social capital to be gained from visiting a hardware store during the week but a hardware store presents slightly more saliency over the weekend. Furthermore, this example clearly demonstrates a difference in the user base of these two platforms. While Foursquare's Swarm application boasts over 50 million monthly active users,[10] it is unlikely that tradespeople, construction workers, and those that frequent a hardware store during the week, are the application's target demographic. It is much more likely that Foursquare users are the types of people to visit hardware stores on the weekend for home improvement projects.

[10] https://foursquare.com/about

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

Computing the Earth Mover's Distance (EMD) between days across the two datasets yields Sunday as the most dissimilar day (normalized EMD=0.200). The EMD of days across the week *within* the Foursquare temporal signature returns Saturday as the most dissimilar day (normalized EMD=0.248). Put differently, the effect of demographics (and the activities places afford them) is largest on Sunday, while within the sample that includes over-proportionally many causal users, Saturday is the most prominent day. From an affordance point of view, visiting a hardware store may satisfy job routine needs for many, and leisure needs for others. In terms of Allen's interval algebra (and the working week), the resulting signatures for both affordances interact in the sense that both start at the same time (when the store opens) but the work-related activities end earlier.

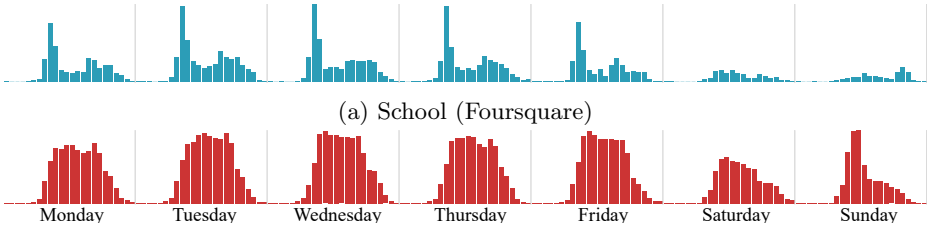
5.3 Interface and Interaction

The previous two examples highlight biases related to the users of platforms. Another aspect to consider is the interface of the application through which the data are contributed. While the exact resolution at which Google collects data from a user's mobile device varies, it is reasonable to assume that location information is taken at regular intervals. This implies that your location is attributed to a place for the duration of your time there. For example, walking into an office building and leaving eight hours later would result in Google attributing eight hours of your time to that office building. In contrast, Foursquare's Swarm application uses an event-based check-in model. A user checks in to a place once and Swarm stores their presence at that location for up to two hours or until their next check-in elsewhere.[11] There is no *check-out*, meaning that the duration of a visit is not recorded. This leads to an event-based effect where users typically check-in when they first arrive at a place and are automatically checked out 2 hours into their visit regardless of how long they choose to stay at the location. The impact of this is evident in the daily bimodal temporal signatures for the place type $School_F$ (Figure 4a). The dominant peaks shown in this Figure are at 8am on weekdays with a smaller increase in popularity between 3pm and 6pm. With knowledge of standard school hours in North America, we can identify these peaks as student drop-off and pick-up times, directly before and after school operating hours. By comparison, Figure 4b depicts the highest amount of activity during school hours on weekdays and decreased activity on the weekends. This reflects the continuous location data sampling method used by Google's location services to populate their temporal signatures and is likely constructed from data contributed from students, teachers, and school employees' mobile devices. While the patterns are very different between the two data providers, *within* the datasets, the dynamics are similar. Jensen-Shannon divergence (JSD) is used to assess the dissimilarity between (a) weekdays in the Foursquare dataset and (b) weekdays in the Google dataset. The results indicate that while the magnitude

[11] https://foursquare.com/dev/docs/venues/herenow

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

is different between datasets, Friday is the most dissimilar day of the week (compared to all other weekdays) with a JSD value of 3.98×10^{-3} and 6.36×10^{-4}. On a side-note, while we can't be certain, we speculate that the Sunday peak in Figure 4b is due to the Church (e.g., Sunday School) place types co-occurring or being labeled as schools in the Google dataset.



(a) School (Foursquare)

(b) School (Google)

Fig. 4: Temporal signatures for *School* in (a) Foursquare and (b) Google.

This example demonstrates that the interface and interaction mode through which a user shares place information has a substantial impact on how that information is reported. Big data research often assumes that sample size makes up for inherent biases but as can be clearly seen, this is not the case. Foursquare check-ins really just show aggregate arrival times whereas Google data offers duration. This begs the question, if Swarm changed their interface to include *check-out* functionality, would there be a significant change in their reported temporal patterns?

5.4 Affordances

The reality of classifying place instances into place types is that a degree of *type relaxation* is necessary in order to use one label to categorize multiple places. Places, by definition, are locations that have been given meaning by the people that visit or inhabit these places [28]. The *meaning* instilled on these locations is often reflected in the activities that people choose to do at these locations, or, put another way, the activities that a place *affords* [14] to them as an interaction of their own needs and capabilities and the (physical and social) properties of the environment. Most POI were designed with a small set of activities in mind that they can afford. Most restaurants, for example, afford eating, drinking, and socializing, but the degree to which each of these activities contributes to the place type varies. A bar, by comparison, also affords drinking, socializing, and eating (typically to a lesser degree), clearly overlapping with restaurant and many other place types. The affordances of these two example place types are almost identical, yet the adjustment in importance of these activities (i.e., predominantly drinking for a bar vs. eating for a restaurant) is what we use to

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

differentiate one from the other. Though both of these place types afford a range of activities, they pale in comparison to many other place types.

Let us examine this idea of affordance bias by exploring the place type *Stadium*. Most stadiums were designed as a place to hold events. These events range from sporting events such as football games or boxing matches, to music concerts or trade shows. The variety of activities that are afforded by a stadium is large, occurring at different times of the day, day of the week, or season of the year. In this way it is hard to define *Stadium* in terms of place type activities as each individual stadium is different from the next, more so than one bar is different from another. In exploring these place types from a temporal perspective, it then follows that aggregate temporal signature built from attendance to stadium events would likely include a large degree of variance depending on the types of events, activities, and the demographics of the people that attend these events. While Google's temporal signatures reflect a less biased sample of the population, Foursquare's temporal signatures produced for $Stadiums_F$ are, to some degree, dependent on the saliency of the event, and demography of the attendees. For example, the temporal signature for $Stadium_F$ would not likely see a significant impact from an Opera event held at a stadium (low saliency and outside target demographic), but would be more impacted by a performance from a new and upcoming DJ (high saliency and target demographic). It is for this reason that we see a substantial difference in the cosine similarity (Table 1) between the two data providers for this place type. Further statistical comparison of the temporal signatures for $Stadium_F$ to $Stadium_G$ results in an EMD value of 0.223, an order of magnitude larger than the EMD of either Bar_F to $Restaurant_G$ (0.056) or Bar_G to $Restaurant_F$ (0.064), demonstrating that the range of activities possible at a stadium contribute to greater temporal dissimilarity than bars and restaurants.

6 Contribution Biases

In much the same way that place types demonstrate unique temporal activity signatures, there has been a series of recent publications demonstrating that place types can be uniquely identified based on differences in spatial distribution of place instances [17, 8, 29]. Here we examine the use of spatial point pattern analysis to assess data collection and contribution biases between providers. We use Ripley's L as an example measure,[12] report on how the two datasets differ in their spatial coverage, and identify some of the reasons why this is the case. Specifically we examine the differences with respect to contribution biases. These are further refined as **(1)** the resolution bias of the taxonomies, and **(2)** bias in the data curation process.

Figure 5 shows Ripley's L functions for two place types, namely *Bar* and *Airport*, in each of the datasets. What is striking in this Figure, is just how different the $L(d)$ functions are for the same place type between data providers. $Airport_F$ demonstrates a high level of clustering at a very short distance whereas

[12] We chose this measure simply as one possible approach to quantifying the differences in spatial patterns. This could instead be Average nearest neighbor, Moran's I, etc.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

$Airport_G$ is less pronounced, gradually increasing as clustering distance increases. To a lesser degree, a similar discrepancy can be seen between Bar_G and Bar_F.

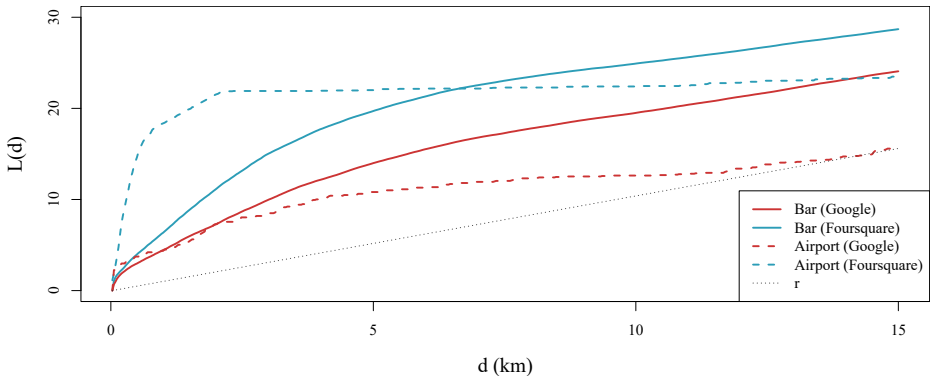


Fig. 5: Ripley's L functions plotted for *Bar* and *Airport* in Foursquare and Google.

6.1 Taxonomy Resolution

These two example place types highlight the substantial differences in the spatial clustering patterns between the two POI platforms. This can be partially attributed to the differences in taxonomy resolution, by which we mean how fine grained the used classification schema are. Given that there are 677 unique POI_{Ft} in our dataset and 105 POI_{Gt}, distinctions that can be made using Foursquare's taxonomy, cannot be made, and thus, observed, using Google's schema. The place type *Bar*, for example, is a single type in the Google taxonomy whereas it is explicitly associated with 11 different subtypes in Foursquare (e.g., *Sports* Bar_F, *Wine* Bar_F, *Gay* Bar_F), not to mention implicit types such as or $Brewery_F$ or $Winery_F$. Users have the option of labeling newly contributed POI with any of these including the broader type Bar_F. This difference in taxonomy resolution means that even though an alignment can be determined through place instance co-occurrence, the actual spatial distribution of POI in each dataset may vary greatly.

The increase in resolution also leads to many POI_{Ft} sharing many of their instances with other types. This means that it is not as simple as combining the point locations for all 11 Foursquare bar subtypes and generating one spatial signature. For example, *Aprés Ski* Bar_F, while intuitively a type of Bar, is actually considered part of the *Ski* $Area_F$ place type and presents a Ripley's L spatial signature more similar to *Ski* $Lodge_F$ than Bar_F or any associated type. Interestingly, our place instance co-occurrence method matched six place instances labeled as *Aprés Ski* $Bars_F$ matching them to place instances labeled as $Restaurant_G$.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

6.2 Place Curation

The methods employed for applying place type labels to place instances is considerably different depending on the provider. Foursquare relies on contributions from individual users through either of their two applications, Swarm or Foursquare. While adding a new place instance, users are asked to assign a place type from the pre-existing Foursquare taxonomy. While the company claims to corroborate much of these additions, they rely on verification and validation from their broader user base.[13] As is the case with many user-contributed data platforms [30], the accuracy and validity of place type labels varies substantially. Google's process on the other hand, is highly curated, involving multiple stakeholders (e.g., users, business owners, internal algorithms) and a robust verification process.[14]

This difference is clearly visible in the drastically different $L(r)$ functions for *Airports*. The clustering pattern for $Airport_F$ is not what one would intuitively expect, showing a sharp increase in POI at a very small distance with very little increase after 2km. Purely from an economical perspective, this clustering makes little sense as market segmentation should dictate that airports be spaced farther apart. Instead, one might reasonably expect a more gradual clustering based on distance, similar to $Airport_G$. Through further investigation, we find that many of the POI tagged as $Airport_F$ are actually terminals, food courts, or parking structures within individual airports. Contributors (those adding new POI_{fi}) to Foursquare have, arguably erroneously, applied the broader type *Airport* to entities within and associated with airports. This reflects the user-contributed nature of Foursquare data and the lack of consistency, verification, and validation on the part of the data curators. One possible future direction for our work is to identify these types of issues and mislabels through a more detailed approach involving spatial signature matching.

7 Conclusions

User-contributed data and geosocial media applications have opened up new avenues to study human behavior by promising easy access to vast amounts of data pertaining to the activities and movement of individuals in the environment. Many of these activities occur at *places* represented as points of interest by leading commercial data providers such as Google and Foursquare. These places are classified into *place types*, human constructed categories of places that afford similar activities. These activities are reflected in popular times of day or days of the week aggregated to produce place type temporal signatures. Similarly, the spatial distribution of POI contributed from individuals and labeled with place types permit the construction of spatial signatures reflecting the fact that bars are likely to be next to other bars, while police stations are not clustered as they have to serve a minimum area. The question then is, how biased are these

[13] https://support.foursquare.com/hc/en-us/articles/201066260
[14] https://www.google.com/business/?gmbsrc=ww-ww-et-gs-z-gmb-v-z-h~bhc-core-u

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

temporal and spatial signatures and how do these biases present themselves? This is not only an interesting question because it helps inform researchers on which dataset to use for a specific research design, e.g., active versus passive check-ins, but also because one would otherwise only expect minimal differences between two datasets that claim global coverage and tens of millions of users.

We address these questions by examining the differences and similarities between temporal and spatial signatures attributed to Foursquare and Google place types. We explore these data through the lens of six different forms of biases and present examples of how these biases manifest themselves in differences between the datasets. It is worth noting that the goal of our study is not to identify the most *accurate* dataset in terms of factual locations people visit, as the passive (often non-voluntary) check-ins would be superior. There is a clear difference between how people behave and how they think they (should) behave and studying this difference requires both datasets.

To showcase one such question that may be asked in the future: why do we see such a clear drop in school check-ins on Fridays in Foursquare but not Google? It looks as though passive check-ins still capture the presence of students, but the active pattern differs greatly. Interestingly, the same can be observed for different types such as *University* and even in entirely different check-in datasets such as the now defunct `Whrrl` platform (that also used active check-ins). Without having both types of sources available, one would simply assume that students tend to start their weekend early, when the reality is far more complex.

Finally, and to end this work with an open question, given that there are clear differences in some temporal signatures between active and passive check-ins and some of these differences can be explained by people preferring not to check in at certain place types, what are the type-specific privacy needs of citizens and should they not be respected?

References

1. James J Gibson. *The perception of the visual world*. Houghton Mifflin, 1950.
2. Krzysztof Janowicz, Grant McKenzie, Yingjie Hu, Rui Zhu, and Song Gao. Using semantic signatures for social sensing in urban environments. In *Mobility Patterns, Big Data and Transport Analytics*, pages 31–54. Elsevier, 2019.
3. Mao Ye, Krzysztof Janowicz, Christoph Mülligann, and Wang-Chien Lee. 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*, pages 102–111. ACM, 2011.
4. Carsten Keßler and Grant McKenzie. Consistency Across Geosocial Media Platforms. In *Proceedings of the 15th International Conference on Location Based Services, Vienna, Austria, 11—13 November 2019*, 2019.
5. Andrea Ballatore and Stefano De Sabbata. Los angeles as a digital place: The geographiesof user-generated content. *Transactions in GIS*, 2019.
6. Christopher Barron, Pascal Neis, and Alexander Zipf. A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS*, 18(6):877–895, 2014.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 |

7. Maryon F King and Gordon C Bruner. Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing*, 17(2):79–103, 2000.
8. Rui Zhu, Yingjie Hu, Krzysztof Janowicz, and Grant McKenzie. Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*, 20(3):333–355, 2016.
9. Linda L Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In *International Conference on Theory and Practice of Digital Libraries*, pages 280–290. Springer, 2000.
10. Martin Doerr. Semantic problems of thesaurus mapping. *Journal of Digital information*, 1(8):2001–03, 2001.
11. Junchul Kim, Maria Vasardani, and Stephan Winter. Similarity matching for integrating spatial information extracted from place descriptions. *International Journal of Geographical Information Science*, 31(1):56–80, 2017.
12. Lin Li, Xiaoyu Xing, Hui Xia, and Xiaoying Huang. Entropy-weighted instance matching between different sourcing points of interest. *Entropy*, 18(2):45, 2016.
13. Taha H Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S Waller. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 75:197–211, 2017.
14. Troy Jordan, Martin Raubal, Bryce Gartrell, and M Egenhofer. An affordance-based model of place in gis. In *Symposium on Spatial Data Handling, SDH*, volume 98, pages 98–109, 1998.
15. Andrea Ballatore and Stefano De Sabbata. Charting the geographies of crowdsourced information in greater london. In *The Annual International Conference on Geographic Information Science*, pages 149–168. Springer, 2018.
16. 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.
17. Christoph Mülligann, Krzysztof Janowicz, Mao Ye, and Wang-Chien Lee. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *Conference on Spatial Information Theory*, pages 350–370. Springer, 2011.
18. Giovanni Quattrone, Licia Capra, and Pasquale De Meo. There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1021–1032. ACM, 2015.
19. Jacob Thebault-Spieker, Brent Hecht, and Loren Terveen. Geographic biases are'born, not made': Exploring contributors' spatiotemporal behavior in openstreetmap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 71–82. ACM, 2018.
20. Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. Representation and communication: challenges in interpreting large social media datasets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 357–362. ACM, 2013.
21. Karen P Tang, Jialiu Lin, Jason I Hong, Daniel P Siewiorek, and Norman Sadeh. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 85–94. ACM, 2010.
22. Brian D Ripley. The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266, 1976.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.

23. Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41(2):125–137, 2014.
24. Alexander Klippel and Stephan Winter. Structural salience of landmarks for route directions. In *Conference on Spatial Information Theory*, pages 347–362. Springer, 2005.
25. David Caduff and Sabine Timpf. On the assessment of landmark salience for human navigation. *Cognitive processing*, 9(4):249–267, 2008.
26. Pin Luarn, Jen-Chieh Yang, and Yu-Ping Chiu. Why people check in to social network sites. *International Journal of Electronic Commerce*, 19(4):21–46, 2015.
27. Brian Chappell. Foursquare demographic data, 2010. https://www.ignitesocialmedia.com/social-media-stats/.
28. Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.
29. Elise Acheson, Stefano De Sabbata, and Ross S Purves. A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320, 2017.
30. Peter Mooney and Padraig Corcoran. The annotation process in openstreetmap. *Transactions in GIS*, 16(4):561–579, 2012.

AGILE: GIScience Series, 1, 2020.
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-14-2020 | © Authors 2020. CC BY 4.0 License.