A. Yair Grinberger
Tessio Novack
Michael Schultz
Peter Mooney
Alexander Zipf
(Eds.)

Proceedings of the

"Geographical and Cultural Aspects of Geo-Information: Issues and Solutions"

AGILE 2019 Workshop

June 17th 2019

Limassol, Cyprus

Table of Contents

Proceedings of the “Geographical and Cultural Aspects of Geo-Information: Issues and Solutions” AGILE 2019 Workshop, June 17th 2019, Limassol, Cyprus

The Geographical and Cultural Aspects of Geo-Information: An Introduction

Tessio Novack[1], novack@uni-heidelberg.de

A. Yair Grinberger[1,2], yair.grinberger@mail.huji.ac.il

Michael Schultz[1], michael.schultz@uni-heidelberg.de

Alexander Zipf[1], zipf@uni-heidelberg.de

Peter Mooney[3], peter.mooney@mu.ie

[1] GIScience Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120, Heidelberg Germany

[2] Department of Geography, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, 9190501, Israel

[3] Department of Computer Science, Maynooth University, Maynooth, Co. Kildare, Ireland, W23 F2H6

1 Introduction

As part of the investigations made in the context of LandSense, a citizen-science project for land-use monitoring (Moorthy et al. 2019), a group of experienced land-use researchers were asked to associate widely used OpenStreetMap (OSM) tags to classes of the CORINE land classification system. The results showed that many tags were not associated to the same CORINE classes (Novack et al., 2018). A qualitative analysis of the results taking into consideration the heterogeneous cultural backgrounds of these researchers led to the conclusion that this disagreement in the association of OSM tags to land-use classes is due to the different instantiations, i.e. physical expressions, and cultural meanings of the geographic concepts represented by the classes and tags.

Such a result is just one manifestation of the seemingly inherent tension between the ambitions of Geographical Information Science (GIScience), i.e. providing answers to fundamental and generic questions about its subject matter, geo-information (Goodchild, 1992), and the contingencies of spatial reality and the data representing it on cultural and geographical contexts. The perhaps most noticeable embodiment of this tension were the intense debates between the proponents and antagonists of Geographical Information Systems (GIS) during the early 1990’s (Schuurman, 2006). Since then however, the discourse had changed and GIScientists have become more sensitive to the social and cultural nature of geo-information and geo-informatics, leading to the formation of research approaches committed to understanding the social bias and implications of GIS, such as GIS and Society and Critical GIS (Goodchild, 2015). Furthermore, in attempts to work across worlds of meaning towards data interoperability, geo-ontology and geo-semantics research assisted in forming new models for representing the world (Goodchild, 2010). And yet, as in the case discussed above, this fundamental issue of geo-cultural dependency has yet to be resolved.

Convinced of the importance of achieving progress on this issue, especially in a context where geo-datasets, geospatial applications, and GIScience methodological approaches strive to be universally effective and relevant, the ‘Geographical and Cultural Aspects of Geo-Information: Issues and Solutions’ workshop was organized. The aim of the workshop was to engage with relevant discussions, relating to issues such as the influence of geographic and cultural aspects on the production and usage of volunteered geographic information (VGI); potential local effects of the usage of global VGI datasets such as OSM; approaches for dealing with geographic and cultural aspects in different analysis contexts and application purposes; the discursive contention of generalization versus specificness in GIScience; and more generally – the relevance of different social and material geographies for GIScience.

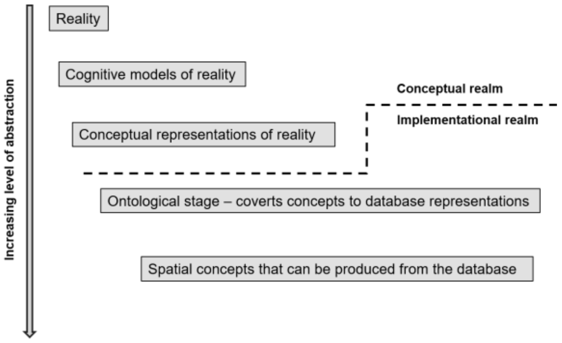
Accordingly, the workshop combined research papers with more general discussions on the progress of GIScience given the challenges that geo-cultural heterogeneity presents. One such discussion was the one which opened the workshop and presented a framework for theorizing about the transition from conceptualization to implementation, which is summarized in the next section.

2 The Ground for Discussion: A Framework for Theorizing on the Transition from Conceptualization to Implementation

In order to support a discussion on the above topics, a basic theoretical framework proposed by Brodeur et al. (2003) was presented (Figure 1). This framework establishes five conceptual levels of abstraction in the path from physical reality to the digital representation of geographic information. This graduation is divided into two main parts, namely, a *conceptual* and an *implementational* realm. The former is inherently human and springs from our cognitive models of reality. The latter is formal, i.e. it refers specifically to the representation of geographic concepts and dynamics as computational ontology.

In accordance to this framework, it can be argued that, within the conceptual realm, the interplay between physical and socio-cultural aspects dynamically produces and re-produces conceptual representations. If the ontology of GIS should mirror these representations, and if the dynamics and

Figure 1: Theoretical framework proposed by Brodeur et al. (2003) and used for grounding the workshop’s discussions.


Reality
Cognitive models of reality
Conceptual realm
Implementational realm
Conceptual representations of reality
Ontological stage – coverts concepts to database representations
Spatial concepts that can be produced from the database
Increasing level of abstraction

output of this interplay varies geographically, historically, and socially, then GIS ontologies must also be, if not specific to each place, time, social group, and use, flexible enough to enable the representation, systematization, and analysis of different geographic and socio-cultural aspects. In other words, dealing with geographic and cultural differences in GIS and geo-information requires not only theorizing on how conceptual representations are contingent upon local environments and cultural contexts, but also requires designing GIS ontologies (i.e. data models, taxonomies, visualization techniques, algorithms) that are specific or flexible enough for enabling the representation of geographic scenarios according to local cultural contexts as well as the deployment of locally relevant epistemologies.

2.1 Scale, Ontological Complexity, and Transferability

Besides the realization that specific and flexible GIS ontologies are necessary for representing, structuring, and analyzing complex social, cultural, and geographical differences, researchers and practitioners need also to care for an adequate alignment between the complexity of the ontology, the geographic scale and the intended degree of the methodological transferability. The aspect of scale also refers to the degree of conceptual generalizations of the categories of analysis, e.g. individuals, social groups, entire populations, etc. The argument being made here is that generalizations and specificness are both possible if this alignment is adequate. For example, the Global Urban Footprint aimed to map all urban areas of the world through the processing of remote sensing images is a pertinent agency producing useful results as the degree of generalization of the category of analysis, i.e. urban areas, is adequate to its global pretension. Another example is the Level 1 of the CORINE land classification system with its five general classes being reasonably applicable for a continental scale of analysis. More detailed land-use taxonomies, however, such as that from CORINE Level 2, might not find relevance and applicability in some specific areas. In her paper Schuurman (2006), the statement is reported that this classification does not match vegetation types from Ireland or the United Kingdom and that conservationists and ecologists in these areas do not share the epistemologies of those from, for example, Russia.

The incompatibility between scale, ontological complexity and intended methodological transferability results in or is caused by a disregard of local geographic and cultural aspects. More specifically, issues of over-simplification and misrepresentation arise when, for example, general taxonomies or taxonomies designed for a specific area are transferred and applied to areas for which they do not reflect local social and geographic idiosyncrasies. This misalignment between ontologies and places results in an imposition of power by the analyst (and the institution or social group he/she represents) on the local affected social groups. At times, this imposition of power is unconscious and the result of the analyst’s negligence. Examples of the unintended application of alien taxonomies/concepts are numerous in VGI research and practice. Is the widely adopted road categorization of OSM (originally conceived for England) pertinent for all urban areas worldwide? Are the feature tagging adopted in OSM remote mapping parties taking into consideration local material and semantic idiosyncrasies? These are questions that need to be critically considered by GIS/VGI researchers and practitioners. At other times, however, the imposition of an ontology is conscious and aimed to strengthen a certain discourse. For example, administration agencies might be interest in reporting an effective preservation of ‘forest’. Thus, the prevalence of one or a few species resulting from a reforestation program is “swept under the hood” (Robbins & Maddock, 2000).

2.2 The Spectrum of Formalizations

In terms of GIS ontology design, we might consider a spectrum of purposes and goals, at its extremities critical GIS scientists and geo-ontologists may be placed. The former group of scholars is interested in local specific contexts and its detailed representation with the minimum loss of meaning. GIS is seen as a tool for representing and empowering local communities and minority groups. For them, the main interest is often a positive real-world impact benefiting these groups. On the other hand, the interests of geo-ontologists are focused on generalization and operationalization, which require proper ways of systematizing, cataloguing and standardizing geographic information as well as analyses. As discussed above, as long as the aspects of scale and conceptual generalization, ontology complexity, and transferability are adequately aligned, the two approaches are equally relevant for GIScience research and practice. In this context, the thriving research field of ontology matching is a promising source of proposed approaches for achieving the interoperability between communicable (specific or general) ontologies. Geo-data conflation and the development of databases embedding context are research avenues that are contributing significantly for the interoperability of GIS ontologies, what extends epistemological possibilities.

2.3 Reflux – The Influence of the Implementational Realm in the Conceptual Realm

An important topic closely related to the discussions in the workshop is how digital representations of the geospace (as GIS, VGI, Webmaps, and WebGIS) are affecting ways in which we perceive, structure, and deploy geographic concepts. In a time where geo-spatial services are more and more part of our lives, human scientists have been discussing ways in which our conceptual representations are being influenced by existing computational ontologies. More specifically, critical GIScientists are calling attention to the fact that the implementational (i.e. formalization, ontological) realm is influencing and "dictating terms" in the conceptual realm. What happens when we rely on existing ontologies to make sense of the world instead of designing ontologies that mirror our differentiated ways of understanding and acting in the world? Are we collecting and structuring geographic information in terms of layers just because GIS are ontologically designed to display and store information this way? What about the influence of location-based services on our spatial behavior? Does the widespread use of these tools has the power of gradually decreasing geographic differences, since they are constantly used by ever larger groups of people? Although these relevant questions related to digitally mediated spatial behavior can rapidly move us towards other inquiries less related to the topic of the workshop, they are surely relevant considerations for GIScientists.

3 Outcome and Outlook

The papers included in the workshop and these proceedings touch upon different aspects of the process of transitioning from conceptualization to formalization. Grinberger et al. (2019), for example, study the extensive roles of institutions in the production of OSM, calling for a more explicit repositioning of institutional epistemologies in the conceptualization of VGI. Zhu et al. (2019) offer an approach relying on spatial signatures for understanding the relations between different sets of categories, i.e. those of streets types and places types. Finally, Ludwig & Zipf (2019) presented an exploratory approach for characterizing the differences between representations across regions, focusing on the case urban green spaces in OSM, as a means towards working with and across these differences.

The diverse dimensions of the relations between geo-cultural contexts and geo-information, and the diverse set of possibilities for approaching these were addressed in the workshop via a concluding discussion relating to the metaphor of "The Glass Bead Game". This game, introduced in Herman Hesse's fictional work of the same title, is a manipulation and creation of symbolic forms for finding links across all areas of human knowledge. This perhaps reflects to some extent the original ambitions of geo-ontology research (cf. Smith & Mark, 2001) – identifying fundamental categories which can be used as the building blocks for any GISystem. Yet, taking the topic of geo-information for disaster preperdness, management, and resilience as a useful case study and point of departure, the discussion had pointed to difficulties with this approach. In such situations, higher-level constructs, to the degree they actually exist, are translated into actions through culturally directed processes. Hence, utilizing the representation of one scenario to another is not straightforward and requires some knowledge regarding the rules of transfer. These rules are geo-culturally contingent and hence require explicitly integrating geography and cultural into geo-ontologies, a challenge which remains open for GIScience to explore even today.

References

Broudeur, J., Bedard, Y., Edwards, G., & Moulin, B. (2003) Revisiting the concept of geospatial data interoperability within the scope of the human communication processes. *Transactions in GIS* 7 (2): 243-265.

Goodchild, M. F. (1992) Geographical Information Science. *International Journal of Geographical Information Systems* 6 (1), 31-45.

Goodchild, M. F. (2010) Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science* 1: 3-20.

Goodchild, M. F. (2015) Two decades on: Critical GIScience since 1993. *The Canadian Geographer* 59 (1): 3-11.

Grinberger, A. Y., Schott, M., Raifer, M., Troilo, R., & Zipf, A. (2019) The institutional contexts of volunteered geographic information production: A quantitative exploration of OpenStreetMap data. In A. Y. Grinberger, T. Novack, M. Schultz, A. Zipf, & P. Mooney (eds.) *Proceedings of the GeoCultGIS AGILE 2019 Workshop* (pp. 4-9).

Ludwig, C., & Zipf, A. (2019) Exploring regional differences in the representation of urban green spaces in OpenStreetMap. In A. Y. Grinberger, T. Novack, M. Schultz, A. Zipf, & P. Mooney (eds.) *Proceedings of the GeoCultGIS AGILE 2019 Workshop* (pp. 10-13).

Moorthy, I. et al. (2019). The LandSense Engagement Platform: Connecting citizens with earth observation data for Land Use and Land Cover Monitoring. In: Living Planet Symposium 2019, 13-17 May 2019, Milan, Italy.

Novack, T., Voss, J., Schultz, M., & Zipf, A. (2018) Associating OpenStreetMap tags to CORINE land-cover classes using text and semantic similarity measures. In: VGI ALIVE Workshop at the AGILE 2018 Conference. Lund (Sweden).

Robbins, P., & Maddock, T. (2000) Interrogating land cover categories: Metaphor and method in remote sensing. *Cartography and Geographic Information Science* 27 (4): 295-309.

Schuurman, N. (2006) Formalization matters: Critical GIS and ontology research. *Annals of the Association of American Geographers* 96 (4): 726-739.

Smith, B., & Mark, D. M. (2001) Geographical categories: An ontological investigation. *International Journal of Geographical Information Science* 15 (7): 591-612.

Zhu, R., McKenzie, G., & Janowicz, K. (2019) Are streets indicative of place types? In A. Y. Grinberger, T. Novack, M. Schultz, A. Zipf, & P. Mooney (eds.) *Proceedings of the GeoCultGIS AGILE 2019 Workshop* (pp. 14-19).

The Institutional Contexts of Volunteered Geographic Information Production: A Quantitative Exploration of OpenStreetMap Data

A. Yair Grinberger[1]
yair.grinberger@uni-heidelberg.de

Moritz Schott[1]
M.Schott@stud.uni-heidelberg.de

Martin Raifer[1]
martin.raifer@uni-heidelbeg.de

Rafael Troilo[1]
rafael.troilo@uni-heidelberg.de

Alexander Zipf[1]
zipf@uni-heidelberg.de

[1] GIScience Research Group, Heidelberg University,
Im Neunheimer Feld 348, 69120, Heidelberg, Germany

Abstract

The original notion of volunteered geographical information (VGI) offers a vision of democratizing geographical information systems (GIS) via the contributions of non-expert individuals, replacing authoritative episetemologies with more open and local geographical representations. Recent studies have questioned this vision, with empirical and conceptual investigations pointing to the effects of data production procedures on the resulting representation. In practice, many organizations and social institutions hold important roles in the production of VGI, thus integrating institutional epistemologies into VGI. This paper explores the role of such institutions in the production of OpenStreetMap (OSM) data by identifying and analysing large-scale contribution events, such as data imports or organized mapping efforts. The paper deploys a global event-identification query on the historical OSM database. The results show that large-scale events are responsible for a significant portion of OSM activities, especially in relation to the creation of data. The procedure identifies several event hotspots, prevalent in either highly developed regions or developing ones. Characterizing the events according to the institutional context that drives them, the paper suggests a relation between socio-economic contexts and the integration of specific institutional perspective into local representations. Hence, the paper contributes to our understanding of VGI as a product of complex interactions of social and institutional perspectives and offers a method towards considering these in research and practice.

Keywords: OpenStreetMap, VGI, Context, Data Imports, Institutions, Remote Mapping.

1 Introduction

From their early days, online geographical information systems (GIS) were hailed as a means towards “democratizing GIS” (Butler, 2006), visioning systems based on individuals of varying skills and perceptions contributing VGI (Goodchild, 2007). Recent studies however point to conceptual and empirical issues that subvert this individual-based vision (Byrne & Pickard, 2016; Haklay, 2013, 2016; Sieber & Haklay, 2015; Stephens, 2013). According to some of these, it is impossible to understand VGI without considering contribution procedures and the technical and institutional framework that they rely upon (Fast & Rinner, 2014; Sieber & Haklay, 2015). This is especially true when large volumes of data are contributed over a short time period, termed here large-scale data production events. Such events require the cooperation of multiple individuals via some kind of organization. Given their volume and impact on data, a possible implication is significantly biasing representation towards the institutional contexts through which they emerge.

One example of this are bulk imports of ready-made datasets into OSM, events reflecting the work of certain (usually governmental) institutes and their employees. While increasing coverage, these events carry with them institutional conceptual and epistemological baggage that, when producing data not fitting well to the project’s structure, may lead to representation issues (Zielstra et al., 2013). Hence, imports can enforce institutional perspectives into OSM on the expanse of more local and individual epistemologies.

OSM, a collaborative mapping project that makes a prominent VGI example, also includes other event types. For example, local chapters organize ‘field mapping parties’ or ‘mapathons’ and organizations such as the Humanitarian OSM Team (HOT) mobilize different communities to make large-scale contributions from afar. Such institutions, while operating within the OSM framework, still hold their own epistemology and enforce it through guidelines and control structures (Palen et al., 2015). These epistemologies may still be different from the ones emerging via the individual-based process initially imagined in VGI.

Hence, the existence of large-scale contribution events in OSM, while adding much to the data, still subvert the initial VGI vision in general. This paper quantitatively explores this issue by studying the spatial distribution of large-scale events and relating these to institutional and social contexts. Below, we detail the data and procedure used for identifying events, the emerging results, and their implications.

2 Methodology

2.1 Event Identification

In this paper, we base our analysis on an assumption that a generic development of OSM data for a specific area would follow three stages, similar to the model described by Gröching et al. (2014): (a) initial interest from a small number of mappers, leading to low contribution numbers; (b) an increasing interest and awareness leading to a rise in the number of mappers and/or contributions; (c) saturation of the data leading to a decrease in the number of mappers and contributions. Over time, the number of contributions will create a normal-like distribution, meaning the cumulative function would take an S-shaped form (Figure 1). Large-scale events disrupt such developments, leading the process to continue as if it jumped forward in time (see cumulative curve w/ event in Figure 1).

Based on this conceptualization, the analysis here relies on fitting a logistic curve describing the development of the cumulative number of contributions C_t over time t (equation 1; α, β, ρ and μ are scaling coefficients) to observed data within a given region. Cases when the curve underestimates actual contribution volumes are indications of events, hence we use estimation errors to identify events. However, time series errors tend to be non-stationary showing a non-random temporal pattern in errors (see errors in Figure 1). We neutralize this by using time-lagged errors to identify events, i.e. error in time *t* minus error in time *t-1*, assuming a normal distribution of lagged errors. We define here events as periods with positive and significant errors at 95% confidence.

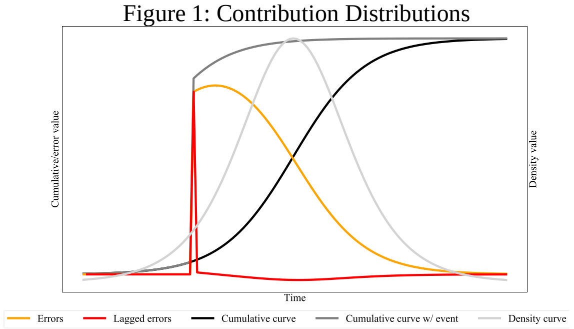
$$C_t = \frac{\alpha}{1+\rho * e^{-\beta(t-\mu)}} \quad (1)$$

2.2 Data extraction and processing

The above procedure requires producing time series data on cumulative contributions for a given spatial division and temporal resolution. For this, we have utilized the OSM History Database (OSHDB; Raifer et al., 2019) tool, which allows querying and aggregating OSM history data in a flexible way on a global scale using custom spatial divisions. The spatial division we used is based on the number of existing OSM entities – a quad-tree-like procedure starting from dividing the world into quadrants and continuing to divide each quadrant as long as the number of entities in one of its sub-quadrants is larger than 50,000. The resulting spatial system thus presents cells of varying sizes and number of entities[1]. The analysis did not consider cells with less than 20,000 entities (see Figures 2 and 4 for the resulting division). The temporal resolution we used is of one month, thus reducing the procedure’s sensitivity to smaller events, and the temporal extent included all data since the beginning of the OSM project and up to April 2019.

The query designed for this research extracted for each spatio-temporal unit (i.e. for each cell and month combination) the total number of contribution actions by breaking down each contribution made during a specific month into basic operations. The number of operations in a contribution of the ‘creation’ type was defined to be the number of added nodes plus the number of created tags. Edit actions considered the total number of changes, i.e. the number of new nodes/tags plus the number of deleted nodes/tags. Deletion contributions were treated as one operation, since such edits can usually be carried by one click of a mouse. These operations were then aggregated to compute the monthly total. This query related to tagged nodes and ways only, excluding relations as they are responsible for only a small fraction of the data yet greatly increase computational load.

Figure 1: Contribution Distributions



Accumulating the monthly total of contribution operations for each cell over time creates the basic time-series data for the analysis detailed above (the time cumulative curve). The query also produced additional information for each spatio-temporal unit for post-processing, such as the number of active users (Users), the relative change in the number of contributions from *t-1* to *t* (Change), the maximal share of contributions made by one user (Max. Actions), the number of edited entities (Entities), the average number of geometry and tag actions per entity (Geometry Actions, Tag Actions), and the share of each contribution type out of all contributions (Deletions, Creations, Tag Changes, Geometry Changes). Notice that the choice of temporal resolution holds an implication for these statistics, meaning they may include non-event activities.

3 Results

3.1 The weights of events within OSM data

Out of 10,136 cells, 494 (4.9%) produced errors during the curve fitting procedure. For the remaining 9,642 cells, the procedure identified 56,578 events (5.9 events per cell, maximum of 19 events in one cell). These events produced 808,117,670 contributions and 6,318,493,481 actions, i.e. 14,283 contributions and 111,677 actions per event (maximum of 2,064,875 contributions and 12,851,643 actions).

To understand the impact of events on OSM, these figures were compared with the total number of contributions and actions in the history of OSM (Table 1). The weight of events is significant, with more than 40% of actions and contributions originating from events. Events especially dominate data creations with more than half of the data ever created in OSM attributed to events. While these results surely include some overestimations relating to the temporal resolution of the analysis, the volume of these events and the lack of results for 4.9% of the cells due to error probably compensate for this. Even so, eliminating the lower decile of events from the analysis (i.e. treating these as false positives) still results in

Table 1: Events' weight in OSM data

Measure	Entire OSM History	Events	% in Events	Median % per Cell	Interquartile Range
Total actions	$1.3*10^{10}$	$6.3*10^{9}$	46.7%	45.7%	26.2%
Geometry actions	$9.5*10^{9}$	$4.2*10^{9}$	44.1%	43.4%	26.9%
Tag actions	$3.9*10^{9}$	$2.1*10^{9}$	53.4%	46.9%	33.8%
Total contributions	$1.9*10^{9}$	$8.1*10^{8}$	41.5%	39.5%	25.6%
Creation contributions	$9.5*10^{8}$	$5.0*10^{8}$	52.4%	50.1%	35.9%
Deletion contributions	$1.3*10^{8}$	$4.3*10^{7}$	33.0%	25.0%	35.9%
Tag change contributions	$4.7*10^{8}$	$1.7*10^{8}$	36.4%	20.6%	29.8%
Geometry change contributions	$4.0*10^{8}$	$9.7*10^{7}$	24.4%	22.7%	27.3%

events representing 41.0% of contributions and 45.9% of actions. Hence, events are a significant driver of OSM data.

Breaking down the share of events in contributions by cell (Figure 2), exposes an uneven distribution with hotspots of event impacts existing in areas such as western and eastern Africa, Indonesia and the Philippines, Nepal, U.S.A, Canada, and to a certain extent Japan, France, Poland, Norway, and Italy. This uneven distribution of institutionalized contributions and hotspots within very different regions suggests the impact of other contextual influences the pattern of events.

3.2 Types and distributions of events

As a means towards exploring such influences and the different characteristics of events (as mentioned in the introduction), we have used the k-means clustering procedure to group events. The variables used for this were the maximal share of actions by one user (Max. actions, percentage) and the share (in percentage) of each type of contribution type out of all contributions, as these represent how centralized this contribution was and on what kind of themes/operations it focused. The procedure clustered events into six groups. To determine the number of clusters, we have computed several cluster separation measures (Davies-Bouldin index, the silhouette coefficient, and the Calinski-Harabasz score) for a range of *k* values. While these produced the optimal values for k=4, this result was judged as too restrictive in terms of representing the diversity of events. The separation measures did not agree on which *k* makes the second-best choice (ranging from 6 to 8) and thus we based our decision on a visual analysis of clustering results.

Figure 3 shows for each cluster the average values of the clustering variables and other available data using parallel coordinates. These allow distinguishing and labelling clusters. Four clusters show high Max. Actions values, meaning one user made most of the contributions, i.e. pointing to a bulk data

Figure 2: Events' share in OSM contributions by cell

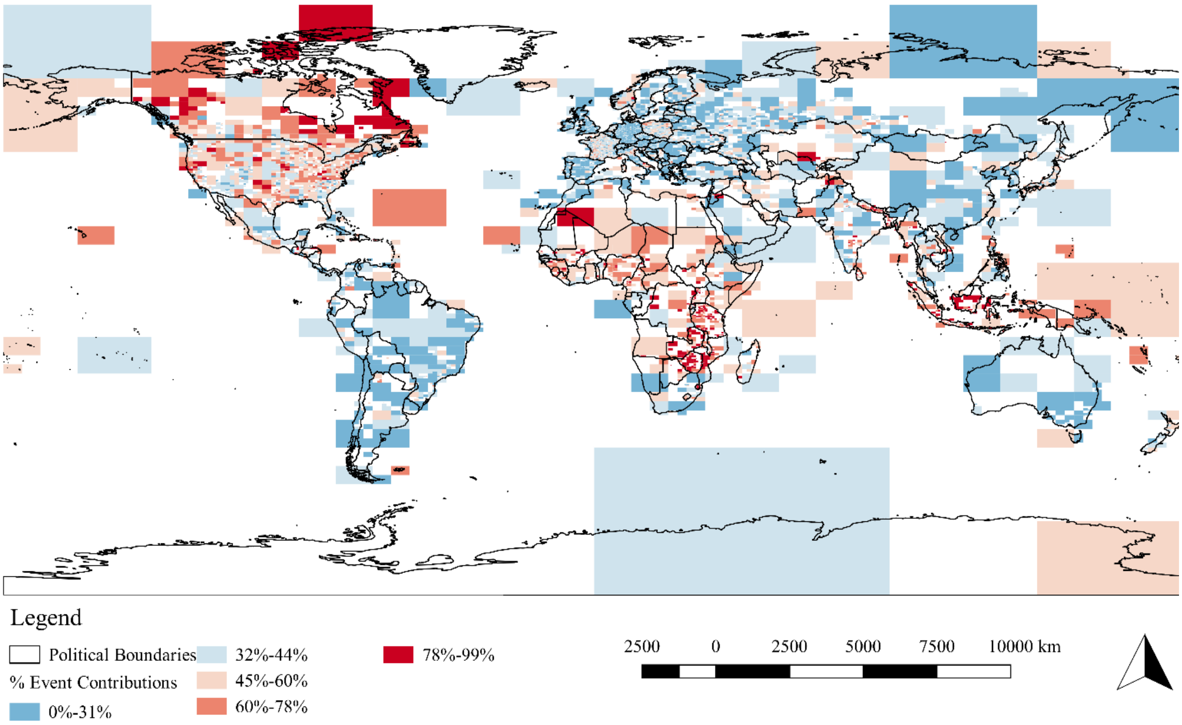
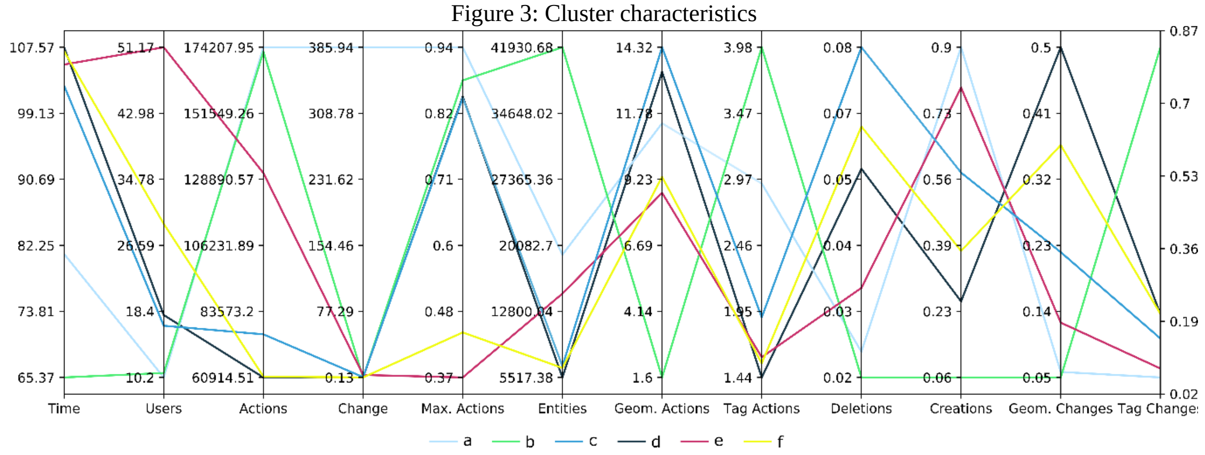


Figure 3: Cluster characteristics



import. Variables such as the share of contribution types and time (number of months since the first contribution to the area) differentiate between these imports (see Fig. 3):

(a) Early imports – the term early refers here both to chronology (*t* value) and to the event’s timing – these events take place relatively early and create a very large effect (average change value of 386%), pointing to an underdeveloped database. Not surprisingly, these events mostly add new data, with creations making 90% of all contributions on average.

(b) Tag imports – another type of early imports including mostly tag operations (more than 85% of contributions, almost 4 tag actions per entity). Despite having high contribution volumes on average, these events do not affect geometry much. Incidentally, these take place mostly in the U.S.A.

(c) Late imports – these are bulk imports taking place in a more mature data region, hence change values are low, creations shares are still high, but geometry and tag changes become more prevalent.

(d) Data updates – this may represent the most ‘mature’ import, where creations receive less weight and the primary activity is updating of geometries, as evident also in the average number of geometry actions per entity.

The two other types present a more distributed kind of large-scale contributions, with actions spread across more users:

(e) Remote mapping event – representing the kind of practices common within HOT tasks, such events include high creation volumes but less tagging activity, indicative of little local knowledge. The average number of users however is very high, thus producing large contribution volumes.

(f) Local mapping event – while similar to remote mapping events in many aspects, these events still show much more focused work and local knowledge, as evident in the relatively high shares of tagging and geometry update contributions and low average number of edited entities.

In the context of institutional epistemologies, event types a-d conceptually seem to represent the same phenomenon – an import of a governmental/external epistemology into OSM. These make the majority of events (70.8% of all events; Table 2) with early and late imports being the most common types. The last two, representing the 3rd and 4th most common types (Table 2), do show difference, as the first represents the epistemological stance of the institute mobilizing the global community, mostly HOT, while the other represents more local epistemologies.

Identifying the most common event type for each cell (Figure 4) and comparing with Figure 2 suggests a pattern. Visually, there seems to be a correlation between event hotspots and event types, mediated by the socio-economic status of the region: late imports dominant the more affluent countries (Japan, France, Poland, Norway, Canada, with the U.S.A. dominated by tag imports) while remote mapping events being more common in the more developing economies (e.g. Indonesia, Eastern and Western Africa). Interestingly, many areas presenting lower event impacts are ones where early imports are most common. These include highly developed economies (e.g. Germany, Spain, the U.K., the European part of Russia, and most major urban areas of Australia), along with some emerging economies (e.g. eastern parts of China and parts of India).

Comparing events discussed in previous studies to the results here validates our results, showing these events were identified and correctly classified for the most part (Table 3). The exceptions are the 2009 Gaza Strip event, caused by multiple local contributions aggregated into one contribution, and some cases of the May 2015 event in Nepal, perhaps pointing to the fieldwork of the Katmandu Living Labs organization and the volunteers it attracted.

Table 2: Events by type

Event type	Frequency	Percentage
Early imports	15,852	28.0%
Tag imports	3,218	5.7%
Late imports	13,901	24.6%
Data updates	7,090	12.5%
Remote mapping events	7,244	12.8%
Local mapping events	9,273	16.4%

Figure 4: Most common event type, by cell

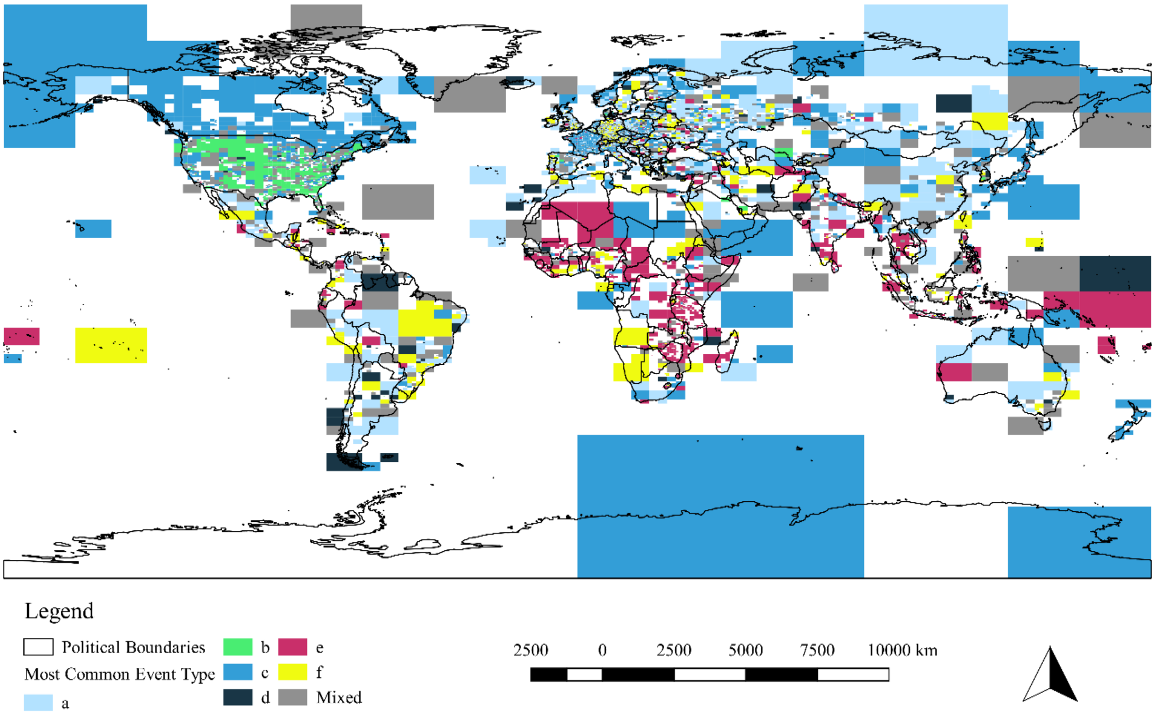


Table 3: Validation of events

Event location and time	Source	Details	Classification by the procedure
Gaza Strip, September 2009	Grinberger, 2018	Bulk import of the work of multiple local mappers	Early import
Gaza Strip, Summer 2014		HOT project	Remote mapping event
Tel Aviv, December 2012		Bulk import of official data	Early import
Tel Aviv, January 2013		Deletion of redundant data and tags after import	Tag import
Nepal, April and May 2015	Poiani et al., 2016	HOT project	Remote mapping wvent; the May 2015 portion of the event was classified as a local mapping event for several cells

4 Discussion and Conclusions

In this paper we have set out to evaluate the individual-driven vision of VGI by investigating large-scale contributions to OSM. The results here allow quantitatively assessing the relevance of this vision, showing that a significant share of the activity in OSM relies on some form of organized contribution, either that of an external data-collecting agency imported into OSM or of organizations operating within this project's framework. Hence, OSM data relies very much on, or is a product of, the work of institutional mediators that are not included in the original vision.

While such a pattern is not inherently problematic, it does hold the potential for introducing bias into representation in OSM. In the case of bulk imports, this may be caused when the workings of a small group of experts (those who created the data and those importing them) replace the democratic concept of crowdsourced contribution. Mapping events organized by local chapters or HOT, on the other hand, enforce epistemologies derived from these institutes' agendas via the organization and direction of data collection efforts. These epistemologies may be different than those emerging otherwise, e.g. when remote mapping events increase the involvement of non-local mappers in an area.

The results pertaining to the spatial patterns and types of events expose such potential impacts, also pointing to their complex relations to geo-social contexts. The negative correlation between the frequency of early import events and the weights of events in total data found for affluent and emerging economies[2] suggests that socio-economic context is both the driving force behind the 'problem' (institutional epistemologies dominating the data) and the 'solution' (an active local community reshaping the data). Imports require a minimal population of educated, skilled, and engaged mappers, the kind of mappers that also make more competent individual contributors. In less developed economies, such mappers are harder to come by, meaning that the impacts of remote mapping events, typical of such regions, tend to last. Hence, while such events rely more on the contributions of individual mappers, they seem to fossilize an institutional perspective which was originated outside of these areas and do not necessarily reflect local views, needs, and perspectives.

With these results and the ability to compare trends across regions, this paper contributes to our understanding of the social, geographical, and institutional contingency of OSM data and procedures. The question remains whether this phenomenon is endemic to OSM, or whether it is common within VGI. In principle, even projects such as citizen reports on vandalism or biodiversity have parallel institutional databases that could be imported, yet such occasions may still be rare. Even so, as OSM makes perhaps the most celebrated and widely utilized VGI project, this issue requires further attention, especially given the increasing impact of corporate mappers on the data (Anderson et al., 2019). Future steps of the analysis would include looking at individual events, measuring their specific impacts and studying the development of data after these. Doing so would allow producing a deeper understanding of the interplay between local communities, institutions, social contexts, and data, pointing towards possible steps and interventions to institutional practices in OSM.

Endnotes

[1] While not considering human perceptions or administrative borders, this spatial division still captuers in most cases regional differences, at least at the national scale (see figure 2).

[2] Using the following definition: affluent economies - western Europe, U.S.A, and Australia; emerging economies - China and India; least developed areas - Sub-Saharan Africa and parts of the south-east Asia and Oceania.

Acknowledgements

We thank the two anonymous reviewers for their useful comments. This research was supported by the Humboldt Research Fellowship for Postdoctoral Researchers.

References

Anderson, J., Sarkar, D. and Palen, L. (2019) Corporate mappers in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), 232.

Butler, D. (2006) Virtual globes: The web-wide world. *Nature*, 439(16), 776-778.

Byrne, D. and Pickard, A. J. (2016) Neogeography and the democratization of GIS: A metasynthesis of qualitative research. *Information, Communication & Scoeity*, 19(11), 1505-1522.

Fast, V. and Rinner, C. (2014) A systems perspective on volunteered geographic information. *ISPRS International Journal of Geo-Information*, 3(4), 1278-1292.

Goodchild, M. F. (2007) Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211-221.

Grinberger, A. Y. (2018) Identifying the effects of mobility domains on VGI: Towards an analytical approach. *Short Paper presented at VGI-ALIVE Workshop at AGILE 2018 Conference*, June 2018, Lund.

Gröching, S., Brunauer, R. and Rehrl K. (2014) Digging into the history of VGI data-sets: Results from a worldwide study on OpenStreetMap mapping activity. *Journal of Location Bassed Services*, 8(3), 198-210.

Haklay, M. (2013) Neogeography and the delusion of democratization. *Environment and Planning A: Economy and Space*, 45(1), 55-69.

Haklay, M. (2016) Why is participation inequality important? In: Capineri, C., Haklay, M., Huang, H., Antoniou, V. Kettunen, J., Ostermann, F. & Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*. London, Ubiquity Press, pp. 35-44.

Palen, L., Soden, R., Anderson, T. J. and Barrenechea, M. (2015) Success & scale in a data-producing organization: The socio-technical evolution of OpenStreetMap in response to humanitarian events. In: Mayer, T. & Do, E. Y.-L. (eds.) *Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems.* New York, The Association for Computing Machinery, 4113-4122.

Poiani, T. H., Rocha, R. d. S., Degrossi, L. C. and Albuquerque, J. P. d. (2016) Potential of collaborative mapping for disaster relief: A case study of OpenStreetMap in the Nepal earthquake 2015, *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Koloa, HI, pp. 188-197.

Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.-B. and Zipf. A. (2019) OSHDB: A framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software and Standards*, 4(3), 1-12.

Sieber, R. E. and Haklay, M. (2015) The epistemology(s) of volunteered geographic information: A critique. *Geo: Geography and Environment*, 2(2), 122-136.

Stephens, M. (2013) Gender and the GeoWeb: Divisions in the production of user-generated cartographic information. *GeoJournal*, 78(6), 981-996.

Zielstra, D., Hochmair, H., H. and Neis, P. (2013) Assessing the effect of data imports on the completeness of OpenStreetMap – A United States case study. *Transactions in GIS*, 17(3), 315-334.

Proceedings of the “Geographical and Cultural Aspects of Geo-Information: Issues and Solutions” AGILE 2019 Workshop, June 17th 2019, Limassol, Cyprus

Exploring regional differences in the representation of urban green spaces in OpenStreetMap

Christina Ludwig[1],
christina.ludwig@uni-heidelberg.de

Alexander Zipf[1],
zipf@uni-heidelberg.de

[1] GIScience Research Group, Heidelberg University,
Im Neuenheimer Feld 348, 69120, Heidelberg, Germany

Abstract

OpenStreetMap (OSM) has been gaining importance in land use mapping due to its free and global availability and high information content especially in urban areas. Since OSM data is created by volunteers and without strict mapping rules, the OSM tags used to mark geographic objects may vary across space. This is especially the case for urban green spaces which leads to different representations of them in OSM. A good understanding of these differences is necessary for the design of a universally applicable algorithm for urban vegetation mapping using OSM data. This study explores which OSM tags indicate urban green spaces, how strong this indication is and how much this varies across different regions. This is done using an exploratory data analysis based on statistical and graphical methods applied to four different cities. Results show that the representation of urban vegetation is influenced by socio-cultural context and the purpose of the map production. Furthermore, the inherent vagueness in the conceptualization of natural objects leads to different associations between OSM tags and vegetation presence across regions.

Keywords: OpenStreetMap, VGI, Sentinel-2, urban land use mapping, urban green space

1 Introduction

Urban green spaces such as parks, semi-natural areas or private gardens are an important factor in cities due to their positive influence on the micro climate, air quality and the wellbeing of citizens. Therefore, sustainable urban planning requires detailed information about the distribution of urban vegetated areas.

Most methods for (urban) land cover mapping rely on remote sensing imagery (Yan et al., 2015). But in recent times, the usage of OpenStreetMap (OSM) has been gaining importance as well (Dorn et al., 2015; Jokar Arsanjani et al., 2015; Schultz et al., 2017). In regard to urban green space mapping, Lopes et al. (2017) investigated whether OSM data is suitable for the derivation of different natural land cover types. They found that OSM offers valuable information, but is not suitable to distinguish between sparse and dense forests due to a lack of data in OSM.

The main advantages of OSM are its free availability and its global community of volunteers generating a rich source of geospatial information especially in urban areas. However, there are also some obstacles to its usage for land cover mapping. In OSM, objects are mapped using a tagging system based on key-value pairs e.g. a building may be mapped as a polygon with the tag *building=yes*. In principle, the users can freely create and choose the tags, but there are mapping guidelines set up by the OSM community to assure the homogeneity of the map. Still, the choice of the appropriate tag is not always unambiguous as Ali et al. (2014) has shown. In a later study, they proposed a methodology to assess the plausibility of OSM tags related to vegetated surfaces to assist mappers in choosing the right tag for a feature (Ali et al., 2016).

Still, this ambiguity and vagueness of certain tags introduces heterogeneity into the data which complicates the application of automatic classification algorithms across large regions using OSM data. Alleviating this problem requires a better understanding of the different ways urban green spaces are mapped in OSM and which aspects need to be considered when interpreting the data. In this regard, this study investigates the following research questions:

- Which OSM tags indicate the presence of urban vegetation?
- How strong is this indication?
- How does this change across regions?

These questions will be answered using an explorative data analysis based on statistical and graphical evaluation methods to quantify the association between certain OSM tags and vegetation presence. The Normalized Difference Vegetation Index (NDVI) derived from Sentinel-2 imagery is used as a reference for vegetation presence. In the following section, the study sites and the explorative data analysis are described. In section 3, selected findings are presented and subsequently discussed in section 4. A conclusion is given in section 5.

2 Data and Methods

2.1 Study sites

Four cities in different geographic regions were evaluated including Munich and Dresden in Germany, Dar es Salaam in Tanzania and Tel-Aviv in Israel. The size of the study sites was set to 7 by 7 km covering the city centre and in parts the suburban area. To exclude the effect of data quality on the representation of green spaces, only cities were chosen which show a high degree of completeness considering roads and buildings in OSM.

2.2 Data processing

To assess the relationship between OSM tags and vegetation presence an explorative data analysis was performed using OSM data and Sentinel-2 imagery. The OSM data was retrieved for April 21st 2019 using the OSM History Database and the OHSOME API (Raifer et al., 2019). All features were retrieved that contained one of the following keys: *leisure, landuse, natural, surface, waterway, wetland, water, building, amenity*. Features that are overlapping another larger feature were cut out (e.g. buildings and roads were erased from a residential area polygon).

The Normalized Difference Vegetation Index (NDVI) derived from Sentinel-2 imagery was used as a proxy to quantify vegetation presence. To get rid of the influence of clouds and seasonal variations in vegetation cover, a maximum NDVI composite was calculated from a time series of Sentinel-2 images spanning the year 2018. The NDVI was calculated based on the near infrared and red spectral bands at a spatial resolution of 10 by 10 meters.

Finally, for each OSM tag all NDVI values that lie within respective features were extracted. Pixels at the edges of OSM features do not provide reliable information, because they may cover multiple land cover types. Therefore, only pixels which are almost fully contained within a feature were extracted.

2.3 Analysis of OSM tags

The association between OSM tags and NDVI values was evaluated using statistical and graphical data exploration methods. For visual analysis, an interactive, web-based dashboard containing different graphs and maps was created using Python. The distributions of NDVI values for different OSM tags and cities were visualized and compared using histograms. Interactive maps were used to compare OSM features to very high resolution satellite imagery.

In order to get an overview of the strongest OSM indicators for urban vegetation, probability values for vegetation presence were derived from the NDVI distributions by calculating the quantiles described in Table 1. NDVI values larger than 0.6 usually indicate pixels that are fully covered by vegetation. By ranking the OSM tags by vegetation probability *p(vegetation)* the strongest indicators for urban greenness were revealed. The *p(mixed)* can be seen as a measure of uncertainty, since mixed pixels do not provide any useful information. *p(no vegetation)* indicates evidence for the absence of vegetation.

Table 1: Thresholds for the calculation of probabilities for vegetation presence of each OSM tag

Probabilities	Thresholds
p(vegetation)	$0.6 < \text{NDVI} \leq 1.0$
p(mixed)	$0.3 < \text{NDVI} \leq 0.6$
p(no vegetation)	$-1.0 < \text{NDVI} \leq 0.3$

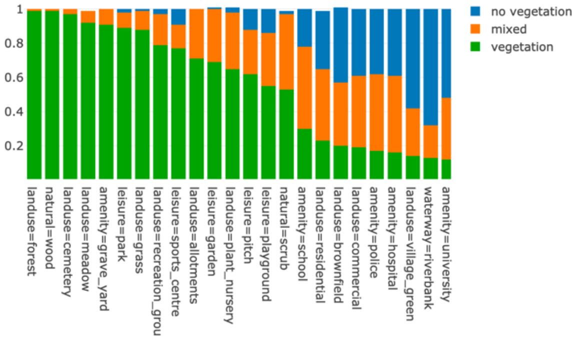
To automatically identify OSM tags whose association with vegetation presence differs between cities, two statistical distance measures were calculated to quantify the similarity of the NDVI distributions. The *Kolmogorow-Smirnov*-Test (KS-test) is a common test to assess whether two samples were created by the same process or not. The KS distance however does not always give a good estimation of the similarity of two distributions. Therefore, a second measure, the *Wasserstein* distance, was calculated in addition.

The OSM wiki and forum were consulted to get information about the evolution and meaning of certain OSM tags and the guidelines that describe their usage (Mocnik et al., 2017).

3 Results

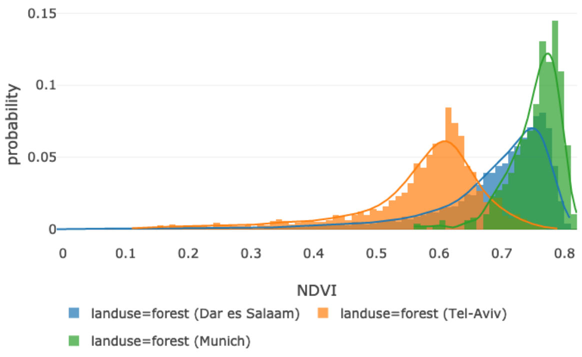
Across all cities, the tags *landuse=forest* and *natural=wood* are always amongst the strongest indicators for vegetation presence with *p(vegetation)* exceeding 0.98 in most cases (e.g. Figure 1) For Tel-Aviv the association is less pronounced (*p(vegetation)*=0.71), This is due to the fact that in this city small areas with scattered trees are often tagged using *landuse=forest,* while in other places this would not be classified as such (Figure 2). Instead, it is more common to map such patches using tags like *landuse=grass* or *leisure=park.* Scattered trees inside those areas would be mapped as nodes with the tag *natural=tree.*

Figure 1: OSM tags ranked by probability for vegetation presence for the study site in Munich.



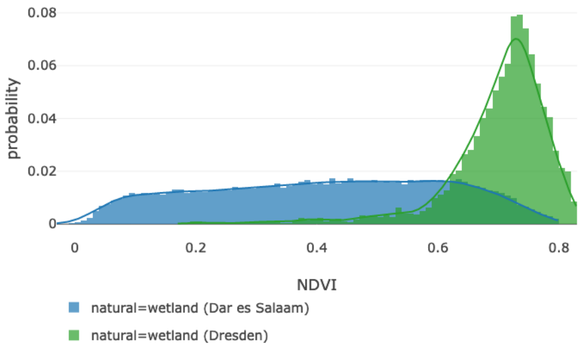
The extent to which individual trees are mapped also differs considerably between the cities. Tel-Aviv shows the lowest number of trees (n=68) in contrast to Dar es Salaam where more than 46 000 trees have been mapped. Even compared to the other cities this is an extraordinarily high number and can be explained by the fact that these trees were mostly mapped by volunteers during a Missing Maps campaign which aimed at mapping relevant objects for flood risk management.

Figure 2: Distribution of NDVI values for the OSM tags *landuse=forest* in Tel-Aviv, Munich and Dar es Salaam.



Comparing the NDVI distribution of the tag *natural=wetland* in the district of Dresden (Germany) and Dar es Salaam (Tanzania) shows large differences (Figure 3). While in Germany wetlands are densely vegetated areas mostly free of human influence, wetlands within the city of Dar es Salaam often contain informal settlements. So, although having the same OSM tag these areas are profoundly different land use types. The OSM wiki contains the *wetland=** tag, which is to be used to further characterize the type of wetland. However, this tag does not contain a value describing artificial, managed or inhabited wetlands. But even though there is no designated OSM tag to mark anthropologically influenced wetlands, the information about the human influence is still contained in OSM through the presence of features that indicated human influence such as *building=** or *highway=**.

Figure 3: Distribution of NDVI values for the OSM tag *natural=wetland* in Dar es Salaam and Dresden.



Sometimes OSM tags seem to be used differently even within the same region. The tag *landuse=village_green* usually denotes a central part of a village covered by grass. This is why it is usually quiet a good indicator for urban greenery. A statistical comparison between Dresden and Munich, however, indicates strongly differing distributions with high values for the KS statistic (0.61) and the Wasserstein distance (0.28). Further analyses show that this detected outlier is due to the "Theresienwiese", a large open space for municipal events, which is tagged as *landuse=village_green* despite being completely covered by asphalt.

Table 2: Probability for vegetation presence of *landuse=village_green*.

City	*p(vegetation)*	*p(non-vegetation)*
Dresden	0.43	0.01
Tel-Aviv	0.33	0.08
Munich	0.14	0.58

Among the best predictors for vegetation are sometimes also tags which do not explicitly describe the area itself, but rather what it is used for. However, this can vary strongly across cities. A good example for that are cemeteries. While the presence of the tag *landuse=cemetery* is a very good predictor for the presence of vegetation in Munich, it is very much the opposite in Tel-Aviv (Figure 4).

Figure 4: Distributions of NDVI values for the OSM tag *landuse=cemetery* for Tel-Aviv and Munich.

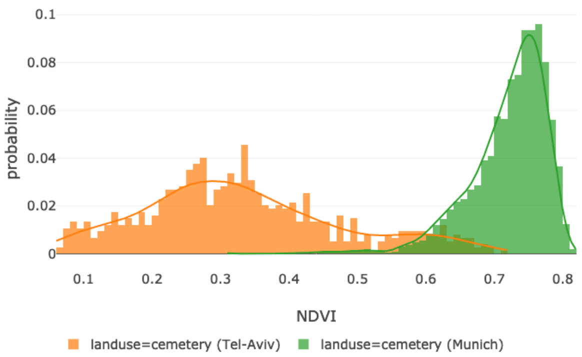
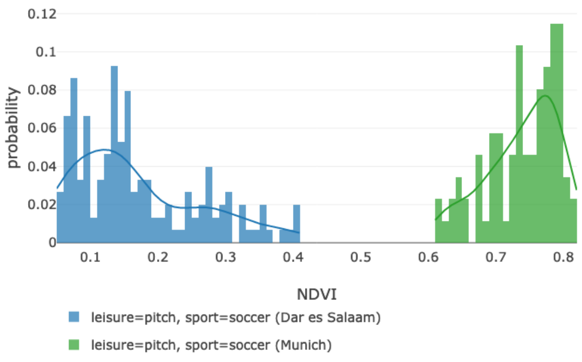


Figure 5: Distribution of NDVI values for soccer fields in Dar es Salaam and Munich.



The explorative data analysis also revealed the importance of secondary tags to increase the specificity of OSM tags for predicting certain land cover classes. Across all study sites, the tag *leisure=pitch* alone is not an unambiguous indicator for vegetation presence. This is due to the fact that some sports require a grass surface, while others require sand or bare soil. Sometimes this is indicated with an additional *surface=** tag. In the case of Munich and Dar es Salaam this tag is mostly not provided in combination with *leisure=pitch*, but the explorative analysis revealed that considering the additional tag *sport=soccer* can help specifying the land cover type as well (Figure 5). However, this is also very much dependent on the cultural context.

4 Discussion

The results show that there are commonalities but also some differences in how urban green spaces are represented in OSM. Natural objects such as forests or wetlands are generally vague concepts and therefore not easy to define unambiguously as shown by Bennett (2001). The consequences of this vagueness can be observed in OSM. Different conceptualizations of forests held by mappers from different socio-cultural contexts lead to different representations of forests in OSM. To which extent these differences can be explained by local socio-cultural or even bio-climatic conditions could not be answered in this study, since a larger number of study sites would have been needed to derive robust statistics.

In regard to wetlands, it became clear that the OSM wiki contains a western bias in the definition of certain geographic concepts. Wetlands are tagged using *natural=wetland* which implies that it is a land use type which is by default free of human settlements. While this is usually the case in western countries, wetlands in other parts of the world are often inhabited or under strong human impact. Currently, this is not explicitly represented in the OSM tagging system, but a strongly discussed proposal to introduce the key *landcover=** might help in reducing such kinds of implicit biases of OSM tags in the future. This case also shows that considering the geographic context of OSM features is crucial in drawing the right conclusions about the underlying land cover.

Another important factor influencing the representation of urban vegetation in OSM is the map production context. The purpose for which the data is produced and by whom plays an important role. In Dar es Salaam, OSM is used as an information source for flood risk management by local organizations. Hence, the overrepresentation of trees compared to other areas where OSM is mainly shaped by independent mappers.

The results also showed how much the association between certain cultural places and the presence of vegetation varies across regions (e.g. cemeteries or sport fields). Deriving information about vegetation presence indirectly from land use information can be a very strong indicator, but it is highly dependent on the cultural context.

5 Conclusion

This study explored the representation of urban green spaces in OSM and its variations across space. Using an explorative data analysis based on graphical and statistical methods the association between OSM tags and the presence of vegetation was investigated. The NDVI derived from Sentinel-2 imagery was used as a proxy for vegetation presence. The analysis was conducted for several cities in different geographic regions to evaluate how much this association varies across space.

The results showed that there are commonalities but also some differences in how OSM tags are used to mark urban vegetation. The vagueness of certain natural objects combined with the different socio-cultural backgrounds of mappers leads to differences in the representations of urban green spaces in OSM. In addition, the purpose of the map production influences the focus of the OSM data. Important information about the presence of vegetation can also be drawn indirectly from tags describing the land use. However, this strongly depends on the cultural context.

For future studies, it would be worth investigating the reasons behind the observed differences in the usage of certain OSM tags such as socio-cultural or bio-climatic context, data quality or the mapping process. These will help in developing locally adaptable algorithms for land use classification using OSM.

References

Ali, A., Sirilertworakul, N., Zipf, A., Mobasheri, A., 2016. Guided classification system for conceptual overlapping classes in OpenStreetMap. ISPRS Int. J. Geo-Inf. 5, 87.

Ali, A.L., Schmid, F., Al-Salman, R., Kauppinen, T., 2014. Ambiguity and plausibility: managing classification quality in volunteered geographic information, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 143–152.

Bennett, B., 2001. What is a forest? On the vagueness of certain geographic concepts. Topoi 20, 189–201.

Dorn, H., Törnros, T., Zipf, A., 2015. Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany. ISPRS Int. J. Geo-Inf. 4, 1657–1671.

Jokar Arsanjani, J., Mooney, P., Zipf, A., Schauss, A., 2015. Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets, in: Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M. (Eds.), OpenStreetMap in GIScience: Experiences, Research, and Applications, Lecture Notes in Geoinformation and Cartography. Springer International Publishing, Cham, pp. 37–58. https://doi.org/10.1007/978-3-319-14280-7_3

Lopes, P., Fonte, C., See, L., Bechtel, B., 2017. Using OpenStreetMap data to assist in the creation of LCZ maps, in: 2017 Joint Urban Remote Sensing Event (JURSE). IEEE, pp. 1–4.

Mocnik, F.-B., Zipf, A., Raifer, M., 2017. The OpenStreetMap folksonomy and its evolution. Geo-Spat. Inf. Sci. 20, 219–230.

Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.-B., Zipf, A., 2019. OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data. Open Geospatial Data Softw. Stand. 4, 3.

Schultz, M., Voss, J., Auer, M., Carter, S., Zipf, A., 2017. Open land cover from OpenStreetMap and remote sensing. Int. J. Appl. Earth Obs. Geoinformation 63, 206–213. https://doi.org/10.1016/j.jag.2017.07.014

Yan, W.Y., Shaker, A., El-Ashmawy, N., 2015. Urban land cover classification using airborne LiDAR data: A review. Remote Sens. Environ. 158, 295–310. https://doi.org/10/f6zhvf

Proceedings of the "Geographical and Cultural Aspects of Geo-Information: Issues and Solutions" AGILE 2019 Workshop, June 17th 2019, Limassol, Cyprus

Are Streets Indicative of Place Types?

Rui Zhu[1],
ruizhu@geog.ucsb.edu

Grant McKenzie[2],
grant.mckenzie@mcgill.ca

Krzysztof Janowicz[1],
jano@geog.ucsb.edu

[1] STKO Lab, Department of Geography, UC Santa Barbara,
Ellison Hall, Isla Vista, CA 93117, USA
[2] Platial Analysis Lab, Department of Geography, McGill University,
805 Rue Sherbrooke Ouest, Montréal, QC H3A 0B9, Canada

Abstract

Places, and here more specifically Points of Interest (POI), can be described by characteristics such as their location, names, capacity, atmosphere, accessibility, reviews, opening hours, prices of services or products they offer, and so forth. Most importantly, however, places can be categorized into place types, e.g., *museum* or *fire station*. These types are best understood as proxies for a wide range of latent characteristics that we do not typically model explicitly in an information system. For example, we would expect to hear sirens nearby fire stations, find parking restrictions nearby, etc. Nonetheless, many modern (geographic) information retrieval systems treat place types as labels, i.e., atomic tokens. The same can be said about names of places and their locations, e.g., addresses. With regards to place (type) embedding, for instance, such a view ignores the cultural structure of these types, names, and addresses, thereby loosing important information. In this work we will show that addresses, here street types, are more than just atomic tokens. They are indicative of the types of places we can expect to encounter. Both the proximity to and suffix of streets are investigated to model the interaction between place types and streets, which are

1 Introduction

In contrast to typical spatial analysis, place-based (or platial) analysis focuses on characteristics that go beyond metric information about locations or geometries (Couclelis, 1992; Goodchild and Li, 2011; Merschdorf and Blaschke, 2018). Work towards place-based GIS and analysis is currently attracting significant attention in the GIScience community (Gao et al., 2013; Merschdorf and Blaschke, 2018; Blaschke et al., 2018; Westerholt et al., 2018), with multiple techniques being developed to analyze places from the perspective of the place hierarchies they form and what they afford to citizens. One family of these approaches focuses on crowdsourced textual descriptions of places, e.g., Adams and McKenzie (2013); Steiger et al. (2015); Siragusa and Leone (2018). These approaches are prevalent nowadays because they are capable of capturing moods, opinions, and experiences towards a place as well as many other latent characteristics such as atmosphere. Many place-based operations use these characteristics to derive a notion of place similarity (Medin et al., 1993) as an analogue to distance in space.

Places, specifically Points of Interest (POIs) in this work, and their types can be studied from a behavioural perspective by considering the thematic, temporal, and spatial patterns in which humans tend to interact with places of specific types. These patterns jointly form *semantic signatures*, i.e., the set of thematic, temporal, and spatial bands that uniquely characterize place types (Janowicz et al., 2019). Intuitively, places of type *museum* may be clustered in a specific district while *fire station* has to maximize coverage. Similarly, we would expect minimal activity around museums at night and early in the morning, but a more uniform distribution of temporal activity patterns at fire stations. Finally, news or reviews about museums are more likely to be about art, exhibitions, tickets, and so on than about rescues, emergencies, fires, and floods. Zhu et al. (2016), for instance, specifically investigated the role of spatial signature in modelling the semantics of place types through applying spatial statistics that quantify the spatial structures and interactions of places of given types.

Our work follows the aforementioned argumentation and further delves into one specific aspect, namely the spatial interaction between place types and addresses, here the street types (suffixes) associated with a place type. Put differently, street suffixes such as Avenue or Boulevard are not just atomic tokens, they carry meaning and reflect the types of places we can expect to encounter at a location. For example, airports are frequently located by main avenues that are close to highways while bookstores would be found on quieter and smaller streets. This paper introduces the proximity to and suffix of the closest street as two forms of spatial signature that describe the spatial interaction between places (and their types) and streets.

2 Related Work

Semantic signatures have been discussed considerably in the literature (Adams and Janowicz, 2015; McKenzie et al., 2015; Zhu et al., 2016; Miller et al., 2019). From a spatial perspective, Zhu et al. (2016) introduced 41 spatial statistics to describe the spatial structure of places and their interactions with other geographic features such as population, climate zones, and street networks. Though a preliminary street interaction analysis was included in this work, street networks were examined in combination with a number of other approaches and not explicitly investigated themselves. In addition, these previous studies focused on aligning feature types across different gazetteers in which most of the features are natural resources such as mountains, rivers, and valleys. In contrast, this work focuses on places in urban areas, where the street networks play a larger role in place and place type identity.

Rather than characterizing the semantics of place types, street networks have also been investigated to model urban functional zones (Yuan et al., 2015), to measure the complexity of urban forms (Boeing, 2018), to predict the traffic interactions of streets (Liu et al., 2017), and so on. However, these techniques only model the interaction of street within a street network, without the association with places being taken into account.

3 Data

Two Point of Interest (POI) datasets were accessed in Maryland, USA, namely Google Places [1] and Foursquare Venues[2]. The data were accessed in January of 2018 using the respective companies' application programming interfaces (API). While both datasets offer similar spatial coverage, each employs a different *place type* schema. These different schemata reflect the underlying purpose for which these datasets were generated. Google Places puts an emphasis on navigation and local business search while Foursquare focuses on local venue recommendations, ratings, and reviews. Given this difference in purpose, Foursquare venues are classified at a finer thematic resolution than Google and include place types such as *Mexican restaurant* and *Japanese restaurant.* In contrast, Google provides only one *restaurant* place type. In total, 383,545 Google places were accessed and categorized into 99 different place types and 132,429 Foursquare venues were accessed and grouped into 403 place types. We selected the Maryland Road Centerlines dataset[3] for the street network, which contains about 4,816 street centerlines for all public roadways in Maryland.

4 Methods

With our goal of differentiating and characterizing place types, we explore two forms of interactions between places and streets, (a) *Proximity* to the closest streets and (b) The *suffix* of the closest street. The closest street of a place in this work is defined as the centerline that contains the point having the smallest geographic distance to the target place.

4.1 Proximity to the Closest Street

The geographic distance between a place and the closest street plays a significant role in identifying the *type* of the place. Such a theory comes from the observation that *nature features,* for instance, are often isolated and further from streets than *cafés* and *restaurants*, place types that must be close to streets in order to attract business. Put differently, the type of a place is implicitly embedded in its interaction with a street network given that the relationship between places and streets differs based on the properties and affordances of the place type. For example, people interact with restaurants on a daily basis as they provide necessary sustenance and social interactions, whereas natural features such as forests, lakes, and parks do not necessarily serve a human-centric purpose.

Considering this, we identify “distance to closest street” as one measure on which to differentiate place types. A set of statistics can be extracted from the distribution of this measure. For example, Equation 1 quantifies the mean distance between a place type and its closest streets, where d_j represents the distance of a place j to its closest street, and N is the total number of places associated with the target place type. Additional distance statistics such as minimum (*min*), maximum (*max*), and standard deviation (*std*) are computed as well to aid in describing the interaction between places and streets.

$$s^{proximity} = \frac{\sum_{j=1}^{N} d_j}{N} \qquad (1)$$

Three Google Places types are shown in Table 1 along with the “distance to closest street” values that distinguish them from one another. As expected, the place type *restaurant* reports a relatively small mean distance to the closest street, while *natural feature* shows a relatively larger distance. These values align with our aforementioned street interaction notion. With the inclusion of additional measures, i.e., *min* and *max*, we can further characterize place types such that *stadium* in Maryland has a much greater minimum but smaller maximum distance to their closest (major) streets when compared to *restaurants*, even though their means are relatively similar. Note that distances are computed based on centroids as places in Google Places and Foursquare Venues are represented as points and that our dataset contains only public streets. This effects the distance between large scale features and streets, particularly in more rural areas.

Table 1: Example statistics for proximity to closest street. Values are based on a sample of > 50 POI per place type

Place Types	Distance to Closest Street (in meters)			
	Min	Max	Mean	Std
restaurant	0.01	15084.88	503.29	785.35
natural feature	8.90	14881.89	1423.70	2172.93
stadium	15.20	1870.40	468.42	387.72

4.2 Closest Street Suffix

In addition to street proximity, place types can also be characterized through other properties such as *street width*. This rational lies on the notion that place types such as *café* or *bakery* are more likely to be close to local, narrower single lane streets as opposed to place types such as *car dealerships*. Fortunately, thanks to the historical and cultural conventions, many properties of a street are implicitly encoded in its suffix[4]. For instance, one expects to find a short and narrow street categorized by the suffix *lane* in a local neighborhood. In contrast, the *parkway* suffix implies a wide, multi-lane street. Based on this, we propose to utilize the distribution of closest street suffix to identify and characterize place types.

Using the Maryland Street Centerlines dataset, we find that streets are categorized into 14 suffix types including streets (*RD*), turnpikes *(PIKE*), avenues (*AVE*), boulevards (*BLVD*), streets (*ST*), parkways (*PKWY*), connectors (*CONNECTOR*), circles (*CIR*), lanes (*LA*), ramps (*RAMP*), drives (*DR*), express ways (*EXPWY*), and no names (*NO NAME*). For each place type, we build a suffix distribution based on each place’s closest

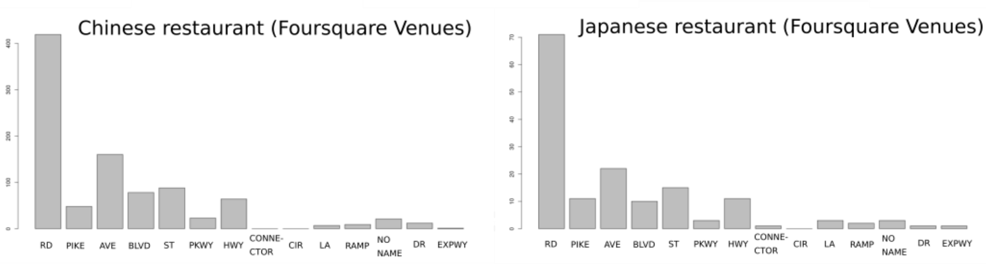
[1] https://cloud.google.com/maps-platform/places/

[2] https://developer.foursquare.com/

[3] http://data.imap.maryland.gov/datasets/

[4] https://pe.usps.com/text/pub28/28apc_002.htm

Figure 1: The distribution of street suffix for *Chinese restaurant* and *Japanese restaurant* from Foursquare Venues.



street and compare it with those produced from other place types. Figure 1 illustrates the distribution of *Chinese restaurant* and *Japanese restaurant* from Foursquare Venues. As expected, they share relatively similar patterns with the type *RD* occurring the most in both, with *ST* and *AVE* second and third, respectively. Moreover, we observe that these two types are barely located close to streets that belong to *CONNECTOR* or *CIR*.

In addition to characterizing *similar* place types, Figure 2 demonstrates how street suffix distribution is capable of distinguishing different place types. For example, the three types, *football stadium, department store*, and *bakery*, illustrate different patterns, despite the common domination of *RD* in their distributions. Specifically, *RAMP* has a prominent contribution in the pattern of *football stadium*, which we barely observe in other place types. Bakeries in general are located more close to *AVE* and *ST*, while department stores have a relatively equal likelihood of being near a *PIKE*, *AVE, BLVD, ST* or *HWY* .

In order to extract representative statistics from the distribution, Equation 2 is introduced, which measures the entropy of closest street suffix for each place type. In Equation 2, p_k represents the probability of observing the suffix k in a distribution of *M* different street suffixes (*M* equals 14 in this work). The larger the value, the more balanced (i.e., uncertain) the distribution. For example, *department store* shows a relatively larger entropy value (2.63) as compared to *aquarium* (1.78). This is due to the fact that department stores can be found near a wide range of street suffixes, while this is not the case for aquariums.

$$S^{suffix} = -\sum_{k=1}^{M} p_k \log p_k \qquad (2)$$

In summary, we propose five descriptive statistics to quantitatively describe the interaction between places and their closest streets. These five statistics are: the *mean, minimum, maximum* and *standard deviation of distance to closest streets,* and *the entropy of closest street suffix.*

5 Experiments and Discussions

Next we discuss exploratory experiments to verify the feasibility of the proposed street-based signatures on characterizing and differentiating place types. First, we used the street signatures to explore the relation of place types within one dataset (i.e., Google Places). Second, we use these measures to assess the similarity of place types across different datasets.

5.1 Experiments Within One Dataset

As a first step, we applied multidimensional scaling (MDS) to our place type dataset using the five-dimensional (i.e., *min, max, mean, std distance* to street and *entropy* of street suffix), street-based, spatial signatures computed from the interaction with closest streets. MDS transforms the relation of place types in high dimensional space into a lower one, by which we can visualize in a 2D map the perceived similarity between place types as reported by our new street-based spatial signatures. Using this method, the relationship between place types of

Figure 2: The distribution of street suffix for *football stadium, department store*, and *bakery* from Foursquare Venues.

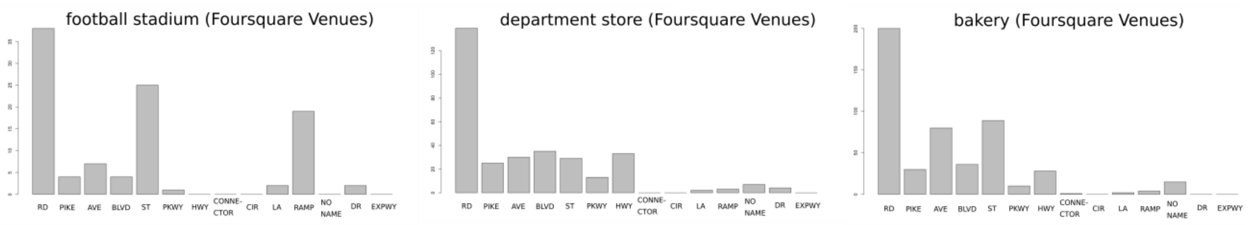
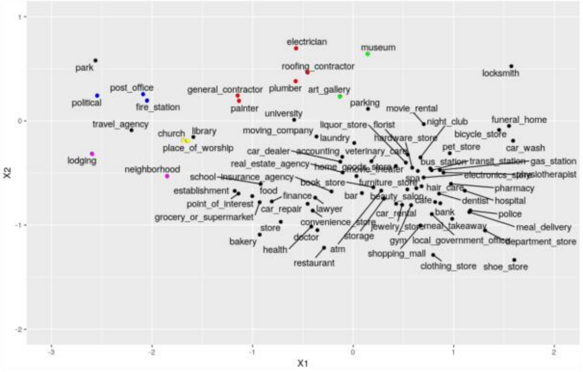


Figure 3: Multi-dimensional scaling map for place types of Google Places.



Google Places were visualized as a two-dimensional chart shown in Figure 3, with the scaling stress achieved at 6.46%. Note that the x1 and x2 axes of Figure 3 are transformed dimensions implying the greatest variation of the signatures without any practical interpretations.

From this initial experiment, we observe that the proposed signatures are capable of revealing similarities between place types. First, place types such as *electrician, roofing contractor*, *plumber, general contractor*, and *painter* form a noticeable group in this map (highlighted in red). Interestingly, they are all related to the construction trade. Second, *post office, political* and *fire station* cluster together providing public services (in blue). In addition, we observe that *museum* and *art gallery* are in close proximity in the figure (in green), both relevant to arts. Finally, the religion-related place types, *church* and *place of worship*, are near to each other (in yellow), indicating a high degree of similarity. Many other types of places exhibit similarity to one another, as can be seen in the figure.

In summary, statistics designed by leveraging the interaction with closest streets have the ability to uniquely characterize and cluster place types (in the Google Places dataset), similar to what most humans would intuitively perceive. Specifically, we demonstrate here that street-based signatures are capable of quantitatively characterizing place types with respect to religions, art, housing modeling and public services.

5.2 Experiments across Different Dataset

In addition to understanding place types within one dataset, this section concentrates on employing the proposed measures to compare place types across different datasets. We particularly investigated the distribution of closest street suffix with the goal of aligning place typing schemata between Google Places and Foursquare Venues. We applied Jensen-Shannon divergence (JSD) to compare the suffix distribution of place types between two datasets. Specifically, the pairwise JSD are computed and ranked, based on which of the top places are selected as candidate matches for a target place type.

Table 2 depicts examples of top matches from Foursquare Venues to Google Places. These examples show the merits of using the proposed signature in aligning place types. First of all, many place types are labeled as different tokens in different data sets, hence using traditional string matching (e.g., Levenshtien distance) would fail to align them. However, the interaction between place type and street suffix helps to address this issue. For instance, *amusement park* and *theme park* have different string names while their similar distributions of street suffix correctly align them, as shown in Table 2. On the other hand, even though two place types from different data sources share the same string names, they are by no means guaranteed to have the same semantics. Take the *hospital* from Google Places as an example, its top 5 matching candidates do not include the *hospital* from Foursquare Venues despite their exactly the same string names. On the contrary, *medical center* is ranked semantically closest to *hospital* in Google Places (with respect to the interaction with streets). As Figure 4 illustrates, hospitals in Foursquare Venues have a high probability of being located near a ST suffix, while both medical centers in Foursquare Venues and hospitals in Google Places are more likely to be found close to a RD suffix. However, it is still worth noting that street-based signatures do not work for all cases. As the third row of Table 2 illustrates, only applying proposed street-based signatures fails to align *post office* in Google Places to its correspondence in Foursquare Venues.

In summary, this section demonstrates that a "suffix-based" spatial signature is of use when aligning two different place type vocabularies. Further work, outside of this short paper, will investigate the limits of this approach.

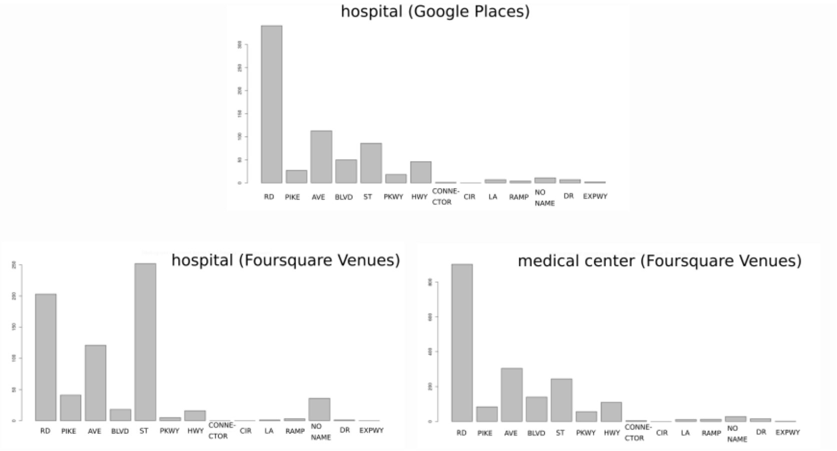
6 Conclusions and Future Work

This paper introduces a new aspect of spatial signature to quantify the semantics of place types based on the interaction with streets. Two types of statistics were proposed: the distance to the closest street, with the *mean, minimum, maximum* and *standard deviation* being selected as the specific statistics, and the distribution of the closest street suffix, with the *entropy* being extracted as the statistic. A series of experiments were conducted to illustrate the feasibility of proposed signatures in terms of understanding the semantics of place types both within one dataset and across different datasets. Thanks to the cultural implication behind both place types and street names, we discovered that the streets, specifically their geographic footprints and suffixes, are in fact indicative of place types. The interaction between places and streets is particularly beneficial

Table 2: Example of typing schema alignment from Foursquare Venues to Google Places. They are ranked by the Jensen-Shannon divergence on their street suffix distribution.

Place Type in Google Places	Top 5 Match in Foursquare Venues				
	1	2	3	4	5
amusement park	**theme park**	bike rental bike share	motel	lounge	market
hospital	**medical center**	salon barbershop	miscellaneous	drugstore pharmacy	laundry service
post office	fire station	city	bridge	flower shop	brewery

Figure 4: Street suffix distribution of hospital from Google places and hospital and medical center from Foursquare. Venues



to identify semantics that are relevant to public services, home improvement, art, health and so on.

However, our current work, as an initial exploration, has several limitations. First, the proposed street-based signatures were represented equally in the multidimensional scaling (MDS) map illustrated in Section 5.1, but such an assumption is not preferable in practice and assigning different weights to different signatures will be explored in future studies. Second, the MDS exploration only focused on a small subset of place types and the analysis was rather subjective and qualitative. Future studies will extend the work to the whole set of place types, and new approaches, such as clustering algorithms, will be introduced to quantitatively investigate the semantic relevance of place types using street-based signatures. Furthermore, we only showed several examples of using proposed signatures to align place types across different data sources, more sophisticated models and systematic evaluations will be investigated in future studies. In practice, the proposed signature has the potential to address practical challenges such as co-reference resolution, open geospatial data cleaning, and place disambiguation, which are the future directions of this work as well. Last but not least, we plan to apply the approach across different cities and countries as a new means to compare and understand the culture implication on places.

References

Adams, B. and Janowicz, K. (2015) Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science* 29 (4), 556-579.

Adams, B. and McKenzie, G., (2013) Inferring thematic places from spatially referenced natural language descriptions. In *Crowdsourcing Geographic Knowledge* (pp. 201-221). Springer, Dordrecht.

Blaschke, T., Merschdorf, H., Cabrera-Barona, P., Gao, S., Papadakis, E. and Kovacs-Györi, A. (2018) Place versus Space: From Points, Lines and Polygons in GIS to Place-Based Representations Reflecting Language and Culture. *ISPRS International Journal of Geo-Information*, *7*(11), p.452.

Boeing, G. (2018) Measuring the complexity of urban form and design. *URBAN DESIGN International*, *23*(4), pp.281-292.

Couclelis, H. (1992) Location, place, region, and space. In: Abler, R. F., Marcus, M. G., Olson, J. M. (Eds.), *Geography's Inner Worlds*. Rutgers University Press, New Jersey, pp. 215-233.

Gao, S., Janowicz, K., McKenzie, G., Li, L. (2013) Towards platial joins and buffers in place-based gis. In: *Comp@ Sigspatial*. pp. 42-49.

Goodchild, M., Li, L. (2011) Formalizing space and place. In*:* *CIST2011-Fonder les sciences du territoire*. pp. 177-183.

Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., Gao, S. (2019) Using semantic signatures for social sensing in urban points of interest with representation learning. *Computers, Environment and Urban Systems* 75, 146–16.

Liu, K., Gao, S., Qiu, P., Liu, X., Yan, B. and Lu, F., (2017) Road2vec: Measuring traffic interactions in urban road system from massive travel routes. *ISPRS International Journal of Geo-Information*, 6(11), p.321.

McKenzie, G., Janowicz, K., Gao, S., Gong, L. (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.

Medin, D. L., Goldstone, R. L., Gentner, D. (1993) Respects for similarity. *Psychological Review* 100 (2), 254–278.

Merschdorf, H., Blaschke, T. (2018) Revisiting the role of place in geographic information science. I*SPRS International Journal of Geo-Information* 7 (9), 364.

Miller, H.J., Jaegal, Y. and Raubal, M., (2019) Measuring the Geometric and Semantic Similarity of Space–Time Prisms Using Temporal Signatures. *Annals of the American Association of Geographers*, pp.1-24.

Steiger, E., Westerholt, R., Resch, B., Zipf, A. (2015) Twitter as an indica- tor for whereabouts of people? correlating twitter with uk census data. *Computers, Environment and Urban Systems* 54, 255–265.

Westerholt, R., Gr¨obe, M., Zipf, A., Burghardt, D. (2018) Towards the statistical analysis and visualization of places (short paper). In: *10th International Conference on Geographic Information Science*, GIScience 2018, August 28-31, 2018, Melbourne, Australia.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H. (2015) Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering* 27 (3), 712–725.

Zhu, R., Hu, Y., Janowicz, K., McKenzie, G., (2016), Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statis- tics. *Transactions in GIS* 20 (3), 333–355.