EXTRACTING AND COMPARING PLACES USING GEO-SOCIAL MEDIA

F. O. Ostermann [a*], H. Huang [b], G. Andrienko [c], N. Andrienko [c], C. Capineri [d], K. Farkas [e], R. S. Purves [f]

[a] Department of Geo-Information Processing (ITC), University of Twente, Enschede, Netherlands – f.o.ostermann@utwente.nl
[b] Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria - haosheng.huang@tuwien.ac.at
[c] Fraunhofer Institute IAIS, Sankt Augustin, Germany / City University London, London, UK – {gennady, natalia}.andrienko@iais.fraunhofer.de
[d] Dipartimento Scienze Sociali Politiche e Cognitive, Università di Siena, Siena, Italy - cristina.capineri@unisi.it
[e] Department of Networked Systems and Services, Budapest University of Technology and Economics, Budapest, Hungary - farkask@hit.bme.hu
[f] Department of Geography, University of Zurich, Zurich, Switzerland – ross.purves@geo.uzh.ch

KEY WORDS: user-generated geographic content, volunteered geographic information, geo-social media, semantic similarity, geographic places

ABSTRACT:

Increasing availability of Geo-Social Media (e.g. Facebook, Foursquare and Flickr) has led to the accumulation of large volumes of social media data. These data, especially geotagged ones, contain information about perception of and experiences in various environments. Harnessing these data can be used to provide a better understanding of the semantics of places. We are interested in the similarities or differences between different Geo-Social Media in the description of places. This extended abstract presents the results of a first step towards a more in-depth study of semantic similarity of places. Particularly, we took places extracted through spatio-temporal clustering from one data source (Twitter) and examined whether their structure is reflected semantically in another data set (Flickr). Based on that, we analyse how the semantic similarity between places varies over space and scale, and how Tobler's first law of geography holds with regards to scale and places.

INTRODUCTION AND RELATED WORK

This study aims to combine a solid theoretical foundation with a localized case study, for which we combine the expertise of a research group with a diverse background. Our objective is to contribute to the understanding of the semantics of places by combining methods from diverse disciplines and datasets from several sources, in particular geo-social media.

The amount of user-generated geographic content (UGGC) and geo-social media (GSM) continues to increase. While its quality is heterogeneous and – depending on which tasks and use cases it is employed for – it can be noisy, it also represents a rich and multi-faceted view on the perception and semantics of geographic places contributed by a subset of the population. However, much current research is focused on improving the presumed lack of quality or management of issues related to (near) real-time processing (Steiger 2015). While these are pressing issues, we propose to explore the somewhat more stable geographic semantics of places as expressed by UGGC and GSM. The motivation for this approach is the assumption that an improved understanding of geographic place semantics can in turn improve interoperability between existing geospatial datasets, and the quality of geographic information retrieval for future streams of geographic information, both from authoritative as well as non-authoritative or citizen sources.

To do so, we draw on several approaches to describe places. First, Agnew (1987, 2011) distinguished three criteria for distinct places, i.e. specific location, locale, and a sense of place, which together define a place. The first criterion, a specific location, allows distinguishing a place from other locations in space, thereby answering the question of *where* something is or happens. Locale is defined by the *properties* of the space, i.e. its boundaries such as walls or streets that delineate a public park or an activity carried out such as shopping, travelling, celebrating, etc. (Teobaldi and Capineri 2014). Finally, sense of place is the people's personal and emotional *attachment* to a place. Our aim is to find specific locations through terms and expressions capturing notions of locale and sense of place, i.e. descriptions and annotations used by citizens for the places they frequent on a regular basis (see below).

Second, we also refer to Winter and Freksa (2012) for an additional motivation to explore the concept of place through contrast: In order to be a distinctive place, a location has to be sufficiently distinct from neighbouring locations: the distinctive traits of UGGC may be better grasped in the contents or annotations included in the data.

Lastly, our approach to measure this distinctiveness is the measurement of semantic similarity expressed through a user's mentioning of elements, activities, and qualities (Tversky and Hemenway 1983, Purves et al. 2011).

We are also interested in the opportunities offered by the similarities or differences between GSM platforms in the description of places, since many studies have focused only on a single source (Purves and Derungs 2015, Huang et al. 2013), and we can assume that many users are not equally active on all GSM platforms. For this study, we decided to combine data from two distinct GSM platforms, i.e. Twitter (micro-blogging) and Flickr (image sharing). Their distinct characteristics (short messages of max. 140 characters vs. photographs with rich

[*] Corresponding author

metadata), wide user base, and well-developed APIs make them feasible candidates.

The aim is to use the abundant Twitter data for an initial, strictly spatial clustering to identify the potential locations of places. However, while giving a first impression, Tweets as signals are very noisy, since they cover a wide range of topics, from personal to event-focused ones, and georeferenced Tweets are not necessarily related to the space or place from where they originate, especially at fine-grained scales (e.g. describing individual locations *within* cities) such as are of interest in this paper (Hahmann et al. 2014).

Hence, we then turned to the second source of data that has shown to provide rich information on the semantics of places: Tagged photographs which can be considered the footprint of place appropriation by the users. We consider Flickr data to be (potentially) richer than Twitter data because a user has more opportunities to provide information on both the image itself and also the circumstances under which it was taken. Flickr allows a user to enter long titles and descriptions for every image, and additionally tag them with free-form expressions. Further, depending on the uploaded image, detailed EXIF (exchangeable image file format) metadata can be retrieved, including such useful information like orientation. A semantic analysis of the richer Flickr data elicited information relating to locale and sense of place. Furthermore, the very nature of photography, and the motivation for describing images, means that we expect images to be more strongly related with their local surroundings at the sub-city scale than tweets, with common categories of tags including location names and artefacts or objects (Sigurbjörnsson & van Zwol 2008).

This kind of analysis enables to address urban centrality with a different approach - based on the concept of *espace vécu* (Frémont 1976) - from traditional methods based on service distribution and concentration: GSM content combines locations with attributes concerning experiences of different nature in such places, thus revealing convergences of city users in the same place who also produce information and – eventually - knowledge about it. They may also be considered as a proxy of city "consumption" or appropriation. Identifying distinct places is necessary because we start from the location, then move to the discovery of the locale by analysing their similarity or difference based on the information attached to them.

Our semantic analysis of locations involved the extraction of Flickr images from the same geographic areas and analysis of the tags used by the authors to describe them. Based on these tags, we then built an aggregated image term vector for each potential place (Tweet cluster), which can be considered as a representation of the semantics of this cluster (place). By carrying out cosine similarity analysis to these term vectors, we were then able to investigate the semantic similarity of places and the relationships between it and space and scale. Following these steps, we aim to answer the following research questions:

- Can we identify distinct places in one data source (Flickr) based on purely spatial clustering from another data source (Twitter)?
- How does semantic similarity between places vary over space and scale? How does Tobler's first law of geography hold with regards to scale and space?

The study area is the Greater London Area. We chose to focus on London because it is a very diverse study area and the availability of data is good (both social media as well as ancillary open data sets).

This study is a first step towards a more in-depth study of semantic similarity of places found in different data sources, as well as a more fine-grained analysis of the fabric of geographic semantics using activities, elements and qualities.

DATA, METHODS AND RESULTS

As outlined in the introduction, our approach follows five phases, which are described in more details in this section alongside the results.

1. Mine Twitter data for London for potential places
2. For each potential place, identify Flickr images that might "belong" to this place
3. Build a binary term vector of elements, qualities and activities
4. Calculate cosine similarity between all potential places
5. Analyse correlations between semantic similarity and space and scale

Starting with phase 1, we collected all geo-referenced Tweets in the Greater London Area between Nov 5, 2012 and October 3, 2013 (334 days). Since our aim was to learn more about place as expressed also in repeated regular behaviours, we decided to focus on Tweets classified as having been generated by residents only. There are several possible ways to separate residents from tourists and guests. One possibility is to use the location as reported in the user profile, but this information is known to be unreliable (Hecht et al. 2011). Other approaches are to filter by the overall duration (dates of first and last Tweets), or to filter by a number of distinct days. We experimented with both methods, selecting thresholds based on natural breaks in histograms. More information can be found in Andrienko and Andriekno (in press). In this study, to filter out tourists, we chose the criterion that any user that had tweeted only within a 30 day time window was eliminated from further analysis. While we could not validate these results formally due to a lack of ground truth, this filtering approach aligns with the research objectives, and resulted in 15,246,565 Tweets from 40,246 users.

Extracting places from a collection of geotagged Tweets can be considered as a clustering problem identifying locations where many Tweets were contributed (and thus, potentially, many people gathered). As proposed in Andrienko et al. (in press), the task of place extraction requires a point clustering algorithm that is insensitive to the density variation and allows limiting the spatial extents of the resulting clusters. We applied the spatially bounded point clustering algorithm, proposed by Andrienko et al. (in press), for this purpose. In short, the algorithm places points in circles with a user-specified maximum radius R_{max} of 300m, a choice grounded in previous empirical research (Ostermann et al. 2013). When a point is added to a circle, the circle centre is re-computed by averaging the x- and y-coordinates of all its points. When there is no suitable circle for a point, a new circle with the centre at this point is created. After processing all points, the circles containing fewer points than the user-chosen minimal number (i.e., 10 in this paper) are discarded, and spatial clusters are formed from the points of the remaining circles. The algorithm allows different point densities in different circles and does not allow the clusters to grow beyond the specified limit R_{max}. Note that the resulting clusters only consist of the points and do not include the enclosing circles; hence, the clusters may have smaller radii than R_{max} and may have arbitrary shapes.

Figure 1 shows the histogram of Twitter cluster radii extracted by the point clustering algorithm. As can be seen from the figure, most of the Twitter clusters have radii between 50 and 100 meters.

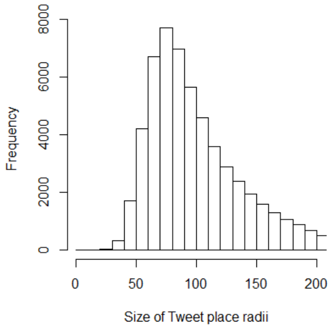

Figure 1: Histogram of the radii of Tweet place clusters

Figure 2 shows the histogram of the number of distinct visitors in each Twitter cluster. As can be seen from the figure, most of the Tweet clusters are small and have only between 10 and 50 distinct visitors.

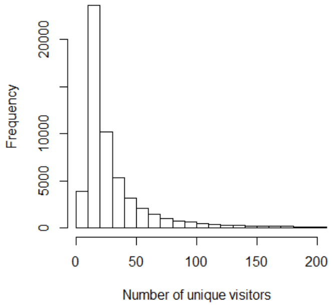

Figure 2: Histogram of the number of unique visitors per Tweet place cluster

The number of more than 55,000 potential places was too high for an exploratory analysis. We therefore decided to filter for clusters that have more than 100 distinct users, as we wished to focus on potential places frequented by many distinct users. In total, we identified 3501 clusters of this kind for further analysis.

In phase 2, after having extracted the potential places (i.e., Tweet clusters), we then needed to identify all Flickr images which belonged to each of these places. Our data set consisted of all geo-referenced Flickr images within a Greater London Area bounding box, retrieved in November 2014, with a total of more than five million images. Two approaches are feasible assigning them to a Tweet cluster: 1) create non-overlapping convex hull polygons representing the extent of the Twitter clusters and identify Flickr images belonging to each of these places; or 2) buffer the Twitter clusters using their radius, and identify Flickr images belonging to each of these buffers. The former will result in Flickr images belonging exclusively to one Twitter place, while the latter allows Flickr images to belong to more than one Twitter place because of overlapping areas. We chose the second approach, in order to avoid unnecessary bias in the results through restricted research design, and because of the characteristics of the Flickr data: A common problem with georeferenced images is that their recorded physical location does not exactly match the physical location of what is being depicted in the image, i.e. the geographic object photographed (since coordinates are typically of the photographer's location). In urban settings, this discrepancy between recorded geographic location and location of the geo-semantics is less of a problem than in rural settings because of the limited line of sight afforded by built-up areas. However, by assigning Flickr images exclusively to the Twitter place in which area it falls, we would be aggravating this problem.

Consequently, first we buffered the Twitter places using their radius in meters, followed by point-in-polygon analysis by intersecting the point location (geographic coordinates) of Flickr images with the Twitter place areas using PostGIS (inner join with multiple entries). A new table thus contains all Flickr images and the id(s) of the corresponding Twitter places. We then dropped Flickr images which were not assigned to one or more Twitter places from further analysis.

During phase 3, we built term vectors for the remaining Flickr images by looking up any activities, elements, or qualities (Purves et al. 2011) in the titles, descriptions, or tags of each Flickr image. This was accomplished through simple lexical matching. Next, we aggregated the term vectors of individual Flickr images to aggregated term vectors representing the Twitter places, i.e. added all term vectors of all images belonging to a Twitter place. There were 17 Twitter places with empty term vectors, which we excluded from further analysis.

Some users bulk uploaded many images of the same area, resulting in high values for some terms for some Twitter places. However, this does not mean that the semantics of that place are necessarily different from another place where fewer users uploaded fewer images. Therefore, in order to avoid the common problem of contributor bias in user-generated content, we normalized the Twitter place term vectors to a binary representation, i.e. replaced all values > 1 with 1.

Phase 4 was initiated with an exploratory analysis, for which we calculated the cosine similarity only for complete term vectors, i.e. combined activities, elements and qualities, between all pairs of Twitter places. We chose cosine similarity measurement because it is an established and well-understood technique for comparing text-based term vectors and computationally feasible. It measures the cosine of the angle between two vectors, thus if the vectors have the same orientation, the angle is 0°, and the cosine similarity is 1, while orthogonal vectors have cosine similarity of 0. It serves as an approximation of semantic similarity between places and has been used successfully in Geographic Information Retrieval (Vockner et al. 2013). Since we are interested in the relationships between space and scale with semantic similarity, we also calculated the Euclidean distance between all pairs, aware of the limitations that Euclidean distance has in an urban context.

In phase 5, we first identified the nearest neighbour for each Twitter place - this relationship can be asymmetrical - and compared the cosine similarity with the distance. The underlying hypothesis here is that based on Tobler's first law of geography, we could expect a strong negative correlation between distance to nearest neighbour and cosine similarity (the farther away, the less semantically similar). This relationship is shown in Figures 3 and 4, which show a histogram of cosine similarity values and a scatter plot of physical distance vs. cosine similarity for all nearest neighbour pairs.

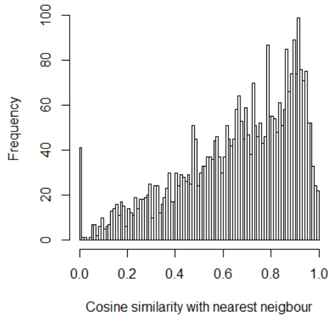

Figure 3: Histogram of the cosine similarities with the nearest neighbours

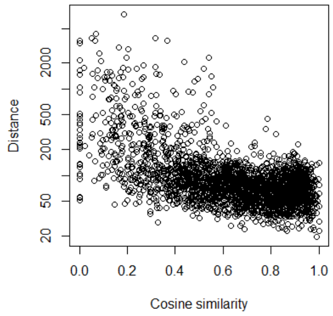

Figure 4: Scatter plot of the cosine similarity and distance of nearest neighbour pairs

As Figures 3 and 4 above show, many place pairs indeed display a very high similarity and are also nearby. This suggests that they might not be distinct places in the sense of Freksa and Winter (2012) but rather a sort of convergence of meanings assigned to places.

In order to test for correlation, we first tested for normality (Shapiro-Wilk) which is not given (distance: W = 0.2508, p-value < 0.000, cosine similarity: W = 0.9354, p-value < 0.000). Since they were not normally distributed, we used non-parametric Kendall's Tau correlation tests, resulting in weak to moderate negative correlation (Kendall's rank correlation tau z = -25.8158, p-value < 0.000, sample estimates tau -0.2921797). To some extent, this negative correlation result between distance and similarity is consistent with Tobler's first law of geography, and shows that near things are in general more related than distant things.

An exploratory visual analysis (Figure 5 on next page) suggests a geographically uneven distribution of cosine similarities with nearest neighbours, i.e. there are areas with higher or lower similarity.

Next, we computed the correlations between cosine similarity and distance for every Twitter place with all other Twitter places, shown in Figure 6:

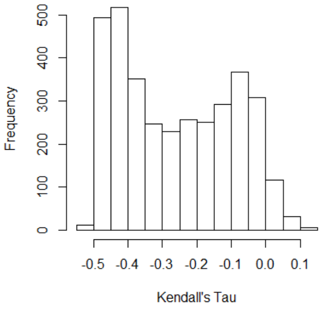

Figure 6: Correlation between distance and cosine similarity for all Tweet places

Figure 7 (next page) shows that the correlation between distance and cosine similarity is much stronger in the city centre than in the more outlying areas. A potential explanation is that Twitter places in the centre have shorter distances to all others, and that the correlation between distances and cosine similarity breaks down at longer distances (compare the plots above on nearest neighbour distance and cosine similarity).

In a next step, we compared the cosine similarity over several distance thresholds: we calculated each Twitter place's average similarity with other nearby places (e.g., within a distance threshold of 100 meters; the lower number of places shown for shorter distance bands is due to excluding places that do not have neighbours within that distance).

As Figures 8-10 (next page) show, the overall average similarity decreases with increasing distance, as expected from our preliminary statistical analysis. However, it is also clearly visible that the centre of the study area, downtown London, has higher average similarity scores than the periphery. Further, in the lowest distance band (Figure 8), there are clearly visible clusters of high average similarity scores. These suggest that these areas are internally more semantically similar than others are. This might potentially allow us to discover distinctive places (as proposed by Winter and Freksa (2012)), which will be our main aim for the future work.

DISCUSSION AND ONGOING WORK

Our analysis is only the first step towards a better understanding of the relationship between geography, semantics, and different data sources. We took places extracted through spatio-temporal clustering from one data source (Twitter) and examined whether their structure is reflected semantically in another, richer data set (Flickr).

One interpretation of the results is that semantic similarity with neighbouring places is stronger in London city centre, the hot spot of cultural activities, shopping malls and services. Despite the social and cultural melting pot of global cities like London, the GSM reveal a sort of shared routine of daily activities. This decreases towards the outskirts of the city where users return to more specific environments. The "espace vécu" offers more heterogeneous stimuli to the users in producing their geo-social content.

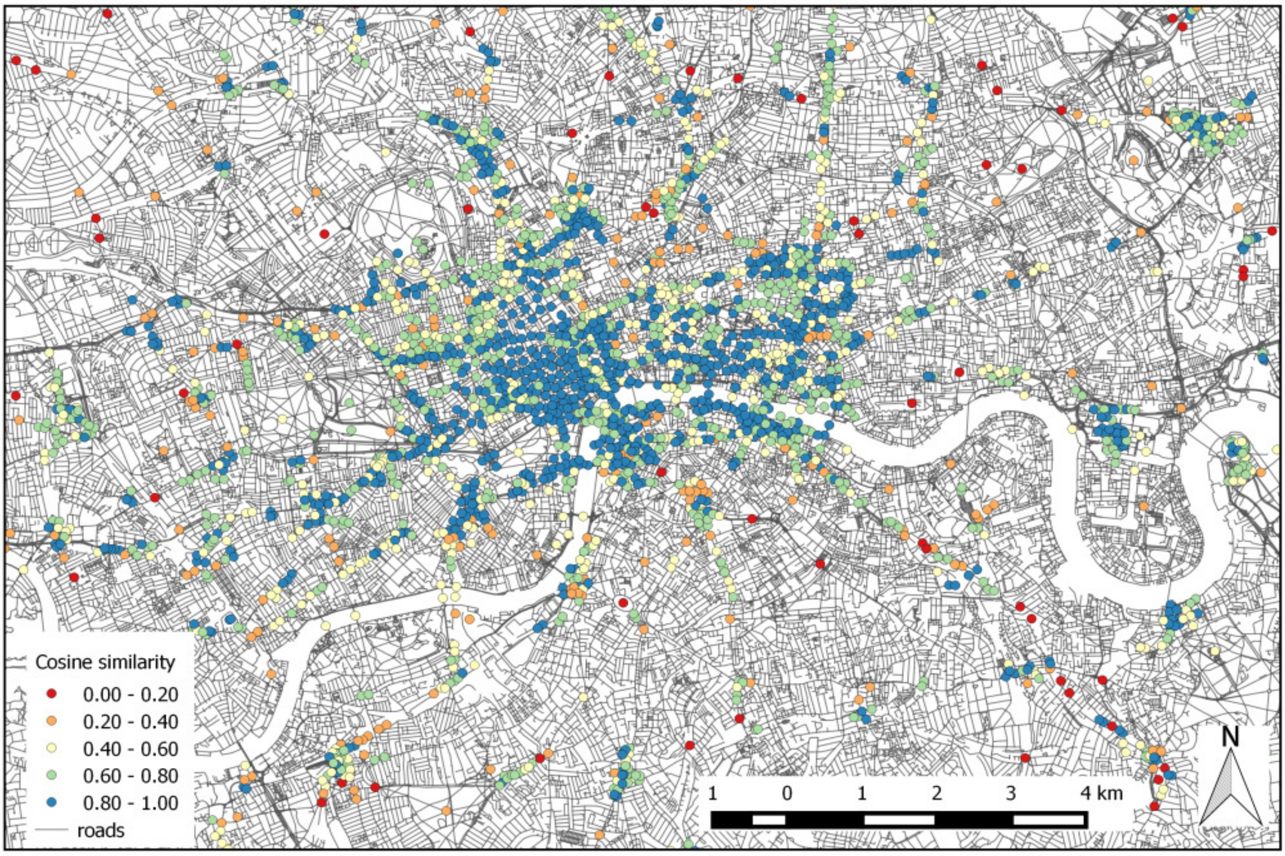

Figure 5: Cosine similarity with nearest neighbours

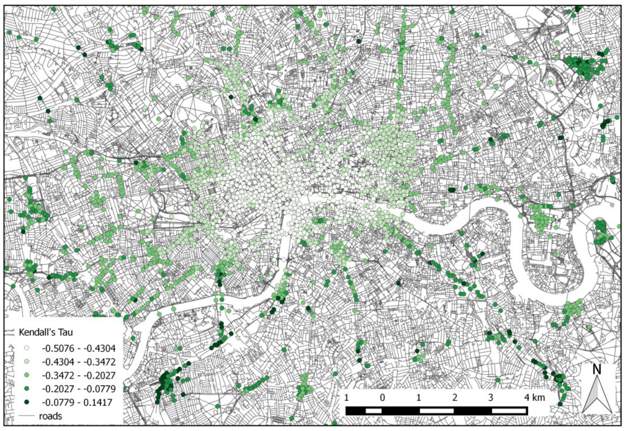

Figure 7: Kendall's Tau for all Twitter places

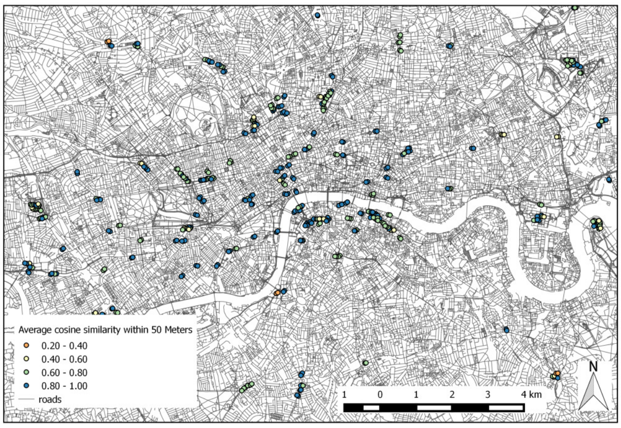

Figure 8: Average cosine similarity within 50 meters

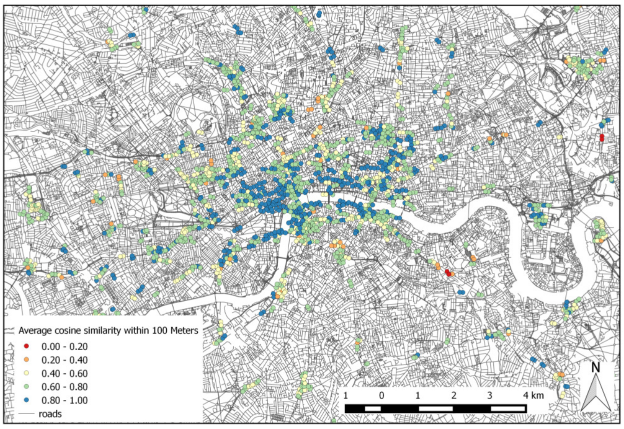

Figure 9: Average cosine similarity within 100 meters

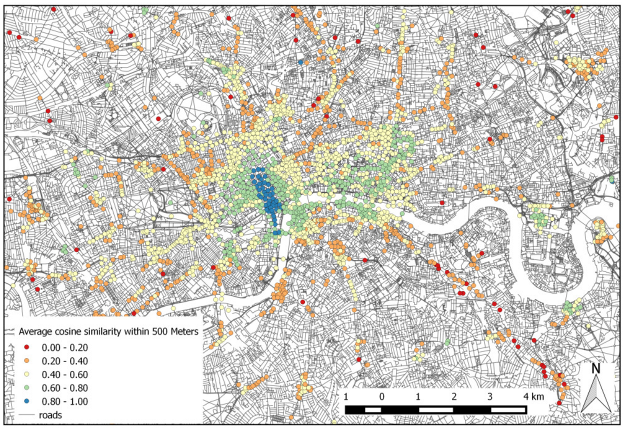

Figure 10: Average cosine similarity within 500 meters

Regarding our research questions, we can identify several coarse places when comparing the average cosine similarity for low distance bands (see Figures 8 and 9). The geographically uneven distribution of similarity suggests that distinct locales could be identified. Concerning our second research question, it seems that the negative correlation between distance and cosine similarity is the strongest for smaller distances, and flattens out over longer distances. This is consistent with Li et al. (2014), which show that Tobler's first law of geography is only consistently true within a specific distance range, and beyond that distance, it no longer holds. These results also support our assumption that distinct locales are discoverable through geographic semantics in user-generated geographic content.

This first exploratory analysis needs further support through in-depth geostatistical analysis. We suggest the following steps:

1. Measure correlation between similarity independently in the 3 dimensions of activities, elements and qualities.
2. Measure the impact of the temporal dimension by investigating time slices of Twitter and Flickr data.
3. Merge neighbouring places with cosine similarity greater than some given threshold value in an iterative clustering process until we reach a stable state.
4. Ground the resulting places (e.g. through POIs from OSM)
5. Select a sample of groups of places and conduct and in-depth qualitative analysis (which terms are similar, which aren't) and compare with the results from the grounding.
6. Analyse the implications of the results with respect to the theoretical framework in particular with reference to the sense of place and urban experiential patterns.

Together with these analyses, we expect to improve the understanding of the semantics of places as well as how geo-social media can contribute to that. We plan to continue with these and present the newest results at the ISSDQ.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of COST Action 1203 ENERGIC in bringing this work to fruition.

REFERENCES

Agnew, J., 1987. Place and Politics: *The Geographical Mediation of State and Society*. Boston and London: Allen and Unwin.

Agnew, J., 2011. Space and place. In: Agnew, J., Livingstone D. (eds.). *The SAGE handbook of geographical knowledge*, London, SAGE Publications Ltd., pp. 316-330.

Andrienko, N., Andrienko, G., Fuchs, G., Jankowski, P., In press. Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces. *Information Visualization*, accepted.

Frémont, A., 1999. La region, espace vécu. Flammarion.

Hahmann, S., Purves, R. S., Burghardt, D. 2014. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science, 9, 1-36

Hecht, B., Hong, L., Suh B., Chi E. 2011. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, 237–46. Vancouver, BC, Canada: ACM.

Huang, H., Gartner, G., Turdean, T., 2013. Social media data as a source for studying people's perception and knowledge of environments. *MÖGG Mitteilungen der Österreichischen Geographischen Gesellschaft (Communications of Austrian Geographical Society)*, 155, pp. 291-302.

Li, T., Sen, S., Hecht, B., 2014. Leveraging Advances in Natural Language Processing to Better Understand Tobler's First Law of Geography. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York: ACM Press.

Ostermann, F.O., Tomko, M., Purves, R.S. 2013. User Evaluation of Automatically Generated Keywords and Toponyms for Geo-Referenced Images. *Journal of the American Society for Information Science and Technology* 64 (3): 480–99..

Purves, R.S., Edwardes, A., Wood, J., 2011. Describing Place through User Generated Content. *First Monday,* Volume 16, Number 9 - 5 September 2011.

Purves, R.S., Derungs, C., 2015. From space to place: place-based explorations of texts. *International Journal of Humanities and Arts Computing*, 9(1), pp. 74–94.

Sigurbjörnsson, B., van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th International Conference on World Wide Web*:327–336.

Steiger, E., Albuquerque, J., Zipf, A., 2015. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, doi:10.1111/tgis.12132.

Teobaldi, M., Capineri, C., 2014. Experiential tourism and city attractivness in Tuscany. *Rivista Geografica Italiana*,121, pp.259-274

Tversky, B., Hemenway, K., 1983. Categories of environmental scenes. *Cognitive Psychology*, 15, pp. 121–149.

Vockner, B., Richter, A., Mittlböck, M.. 2013. From Geoportals to Geographic Knowledge Portals. ISPRS International Journal of Geo-Information 2 (2): 256–75.

Winter, S., Freksa, C., 2012. Approaching the Notion of Place by Contrast. *Journal of Spatial Information Science*, 5, pp. 31–50.