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)	$0.6 < NDVI \leq 1.0$
p(mixed)	$0.3 < NDVI \leq 0.6$
p(no vegetation)	$-1.0 < NDVI \leq 0.3$

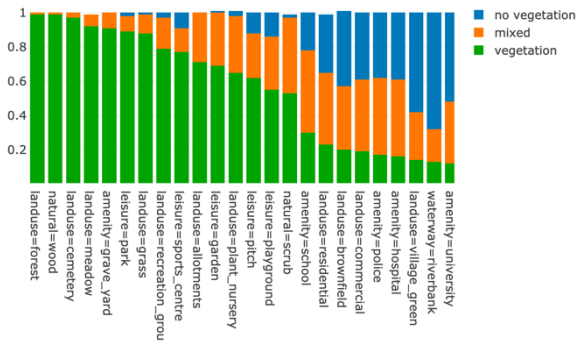
To automatically identify OSM tags whose association with vegetation presence differs between cities, two statistical distance measures were calculated to quantify the similarity of the NDVI distributions. The *Kolmogorow-Smirnov*-Test (KS-test) is a common test to assess whether two samples were created 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

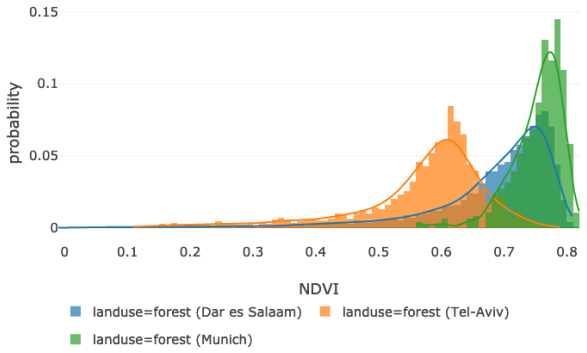
Across all cities, the tags *landuse=forest* and *natural=wood* are always amongst the strongest indicators for vegetation presence with *p(vegetation)* exceeding 0.98 in most cases (e.g. Figure 1) For Tel-Aviv the association is less pronounced (*p(vegetation)*=0.71), This is due to the fact that in 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.



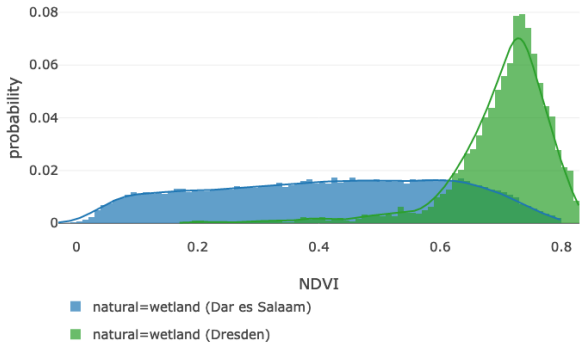
The extent to which individual trees are mapped also differs considerably between the cities. Tel-Aviv shows the lowest number of trees (n=68) in contrast to Dar es Salaam where more than 46 000 trees have been mapped. Even compared to the other cities this is an extraordinarily high number and can be 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.

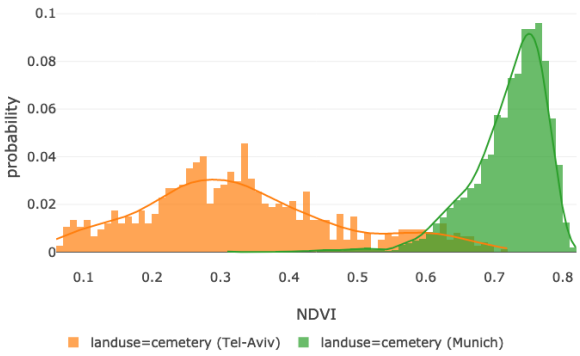
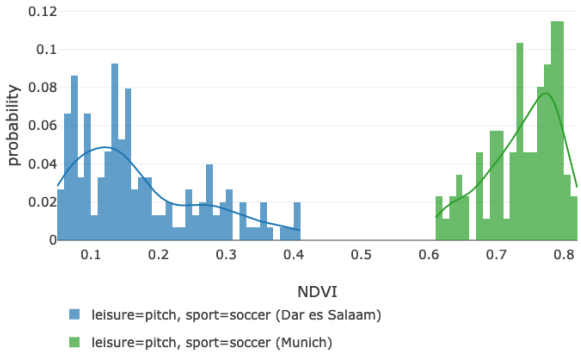


Figure 5: Distribution of NDVI values for soccer fields in Dar es Salaam and Munich.



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