OPENSTREETMAP ELEMENT VECTORISATION - A TOOL FOR HIGH RESOLUTION DATA INSIGHTS AND ITS USABILITY IN THE LAND-USE AND LAND-COVER DOMAIN

M. Schott[1,*], S. Lautenbach[1,2], L. Größchen[1], A. Zipf[1,2]

[1]GIScience, Institute of Geography, Heidelberg University, Germany
[2]HeiGIT at Heidelberg University, Germany
(moritz.schott, sven.lautenbach, leonie.groesschen, zipf)@uni-heidelberg.de

Commission IV, WG IV/4

KEY WORDS: Data Analysis, OpenStreetMap, Software, Data Quality, Land-use and Land-cover, Intrinsic Data Quality.

ABSTRACT:

OpenStreetMap offers manifold possibilities for spatial analysis and location based services. However, fitness for purpose is a commonly discussed issue. As external datasets of high quality are frequently missing, many assessments rely on intrinsic methods. Existing tools for intrinsic data analysis tend to focus on specific topics and/or regions. We present a tool that provides access to currently 32 attributes or indicators at the level of single OpenStreetMap objects. These indicators cover aspects concerning the element itself, surrounding objects and the editors of the object. The usability of the tool was proven on the use case of land-use and land-cover polygons. We applied the tool to 1000 randomly sampled polygons. A tendency that OpenStreetMap objects in more densely populated areas were smaller was detected. Age and size of the objects differed across the continents with older and smaller objects in Europe and North America. A k-means cluster analysis was used to identify groups in the data. This detected a cluster highly influenced by North American lakes that originate from imports. The tool offers amble opportunities for future research, supports the OpenStreetMap community by making informed planning decisions for future activities and enables data consumers to make informed decisions on data usage. While the development was made with land-use and land-cover information in mind, the tool can be seamlessly applied to any polygonal OpenStreetMap data and also supports linear and point data.

1. INTRODUCTION

OpenStreetMap (OSM) has evolved to one of the most used geographic databases. It is a major knowledge source for researchers, professionals and the general public to answer geographically related questions. As a free and open community project, the OSM database can not only be edited but also used by any person with very limited restrictions like internet access or usage citation. This open nature of the project enabled the establishment of a vibrant community that curates and maintains the projects' data and infrastructure, but also a growing ecosystem of tools that use or analyse the data (OpenStreetMap Contributors, 2022b,a). Prominent examples for these tools are the routing platform OpenRouteService[1] or the analyses platform ohsome[2], which is based on the OSM History Database (OSHDB) (Raifer et al., 2019).

The large variety of interests for using OSM data is mirrored in the variety of stakeholders and tools surrounding the project. For mappers and the community interests often revolves around mapping tools, map change notifications, activity reports or leader-boards (Neis, 2022). Professionals on the other hand are more interested in easy data access and usability, while researchers are additionally interested into studying the OSM community and analysing the OSM data. One of the most prominent analysis topics is data quality that has been covered in theory (see for example Barron et al., 2014; Senaratne et al., 2017) as well as in many practical studies (e.g. Jokar Arsanjani et al., 2015; Brückner et al., 2021).

Senaratne et al. (2017) characterises analyses into extrinsic metrics, where OSM is compared to another dataset, and intrinsic indicators, where metrics are calculated from the data itself. While the quality gold-standard is mostly defined for extrinsic metrics through an external dataset of higher or known quality and standards, intrinsic indicators are sometimes not so easily interpretable in the context of quality and are an ongoing topic of research. However, external datasets of high quality are frequently not available. In recent years the interest has also more and more shifted towards doing large scale or even global analyses (e.g. Herfort et al. (2021)) which further limits the number of potential high quality alternatives as reference.

This diversity of tools can be a challenge for data users who will find themselves in a universe of highly specialised or complex tools using different programming languages, platforms, interfaces, output formats etc. While there have been efforts to provide users with higher level data insight and analyses platforms, these still mostly concentrate on or are limited to certain topics or regions. To our knowledge no tool exists to analyse and combine topic independent aspects of the data at the highest possible resolution: single OSM elements. This work sets out to bridge this gap by integrating multiple aspects of the OSM ecosystem into one workflow that allows the quantitative assessment of selected OSM elements. While the proposed tool is yet another contribution to the ecosystem, it sets out to bundle a set of analyses, that would otherwise have to be run or implemented by users seperately and individually. The new software and its functionality is described in section 2.1 before applying it to the example use-case of Land-use and Land-cover (LULC) information in OSM in section 2.2. The use-case findings are described and interpreted in sections 3 and 4 before section 5

* Corresponding author
[1] https://openrouteservice.org
[2] https://ohsome.org

concludes with a summary and outlook.

1.1 Related Work

The presented tool combines the findings and suggestions of more than 15 scientific studies, community initiatives or community projects. The full list of references sourced by the tool is available in the respective documentation linked below. These works have analysed various aspects of the OSM data such as properties of the OSM object itself, properties of its surrounding OSM objects and properties of the OSM editors who edited the object. Many but not all of these studies focused on the aspect of data quality.

The general assessment of OSM geometries and their properties is e.g. described by Mooney et al. (2010). Measuring the mean distance between vertices of OSM polygons, they found that this metric may vary across different land cover types. Smaller distances are related to higher detail which can have implications for the quality of a feature, i.e. a forest polygon with low detail might not be representative of the actual boundaries on the ground (Mooney et al., 2010).

Although individual OSM objects can be viewed individually, they are always embedded in a larger context of surrounding OSM objects, communities of contributors and other classification systems, such as biomes or socio-economic factors. Comparing contributions and communities for selected cities, Neis et al. (2013) found a positive correlation between contributor density and Gross National Product per capita and showed that community sizes vary between Europe and other regions. In 2021, Schott et al. (2021) described 'digital' and 'physical locations' in which an OSM object is located. These 'locations' consist of OSM-specific measures such as density and diversity of elements, but also include aspects of economic status, culture and population density to describe the surrounding of an object.

Furthermore, the OSM database does not only offer the possibility to exploit elements, but also to gather information about the mappers who edited the objects. In previous studies, users were categorised by their experience (Neis et al., 2013), the distance of the edited object to other edits by the same user (Neis and Zipf, 2012) and the tag diversity which can show how specialised a user is (Schott et al., 2021).

2. MATERIALS AND METHODS

2.1 Software

The presented tool named OSM Element Vectorisation Tool (OEVT)[3] defines an extensible workflow for OSM data analyses. The goal of the tool is to transform raw OSM data objects into interpretable or machine readable attributes, sometimes also called vectorisation, embedding or feature construction. Many of these attributes where in the past linked to data quality or presumed to be indicating a certain quality. The term 'indicator' is therefore used synonymous to 'attribute' or data 'metric'. Raw OSM data itself represents the same two information as any other geographic data: 1) what is 2) where. The Volunteered Geographic Information (VGI) aspect of OSM though adds a huge amount of additional information, that is hidden behind this strict technical view on the data. Sieber and Haklay (2015) called this "the epistemology(s) of VGI". In contrast to 'official' geodata, OSM and its communities transparently communicate the existence of these epistemologies that express themselves e.g. though the absence or variability of standardised data collection and mapping methods which in turn add inherent information to the data. For example any OSM object has a story to tell about *who* created it *why*, *when* and *how*. The presented tool makes a step in this direction to facilitate and formalise the analyses and explicit communication of this information.

Out of the endless number of possible data aspects, a set of 32 indicators are currently available for the user to choose. These cover aspects concerning the element itself (e.g. object area, geometric complexity and object age) but also the surrounding data (e.g. the mapping saturation and community activeness) and the editors (e.g. their experience, remoteness or editing software used). These indicators are a collection of metrics that have been analysed in dedicated scientific studies as well as by the OSM community. The collection was augmented by indicators the authors deemed interesting or useful. A full list of the indicators, their calculation, description, data type, used data sources and the inspirational source is contained in the documentation of the repository linked above.

The tool is centred around a Python package. Figure 1 provides a visual representation of the tool. The package bundles all data collection from the different used sources, data transformation, some data analyses and the backend handling. The tool uses a PostGIS enhanced PostgreSQL database as backend data storage and computation framework. It draws on other sources where necessary such as POST-requests and Java. Further data processing is done using the R scripting language. The database enables an efficient server-side computation on the potentially large and interwoven data. Additionally, the backend enables the tool to be fail-safe between computations. It can recover from many common issues like failed connections because the data and all intermediate results are persisted and the tool will automatically detect incomplete analyses when it is restarted. The workflow is also resilient towards missing data. The main Python module to run the full workflow as well as parts thereof can be applied by the user in their own code via the provided Python-API. This though requires a certain backend setup that must then be provided by the user. Additionally, a command line interface exists that makes the tool available for any user, independent of any programming experience. Alternatively a Docker Compose[4] script will set up a suitable backend and run the tool. The Docker setup also provides a minimal reproducible example that enables interested users to test the tool within a few minutes.

Apart from the static connection parameters to the backend and the necessary external data sources, the workflow only requires a location and timestamp as input to analyse a certain set of elements. Benchmarks have shown that the tool is capable of processing around 1k elements per hour making it a suitable tool for larger analyses of custom regions or element sets. However, producing meaningful benchmarks for the tool is nearly impossible as the computational time highly depends on the analysed objects. While some analyses durations are linked to the amount of data processed, others are more influenced by the data composition or have an O(1) time complexity. Apart from the database, the tool can optionally write the output to a Geopackage and thematic CSV-files.

[3] source code available at `https://gitlab.gistools.geog.uni-heidelberg.de/giscience/ideal-vgi/osm-element-vectorisation`

[4] `https://docs.docker.com/compose/`

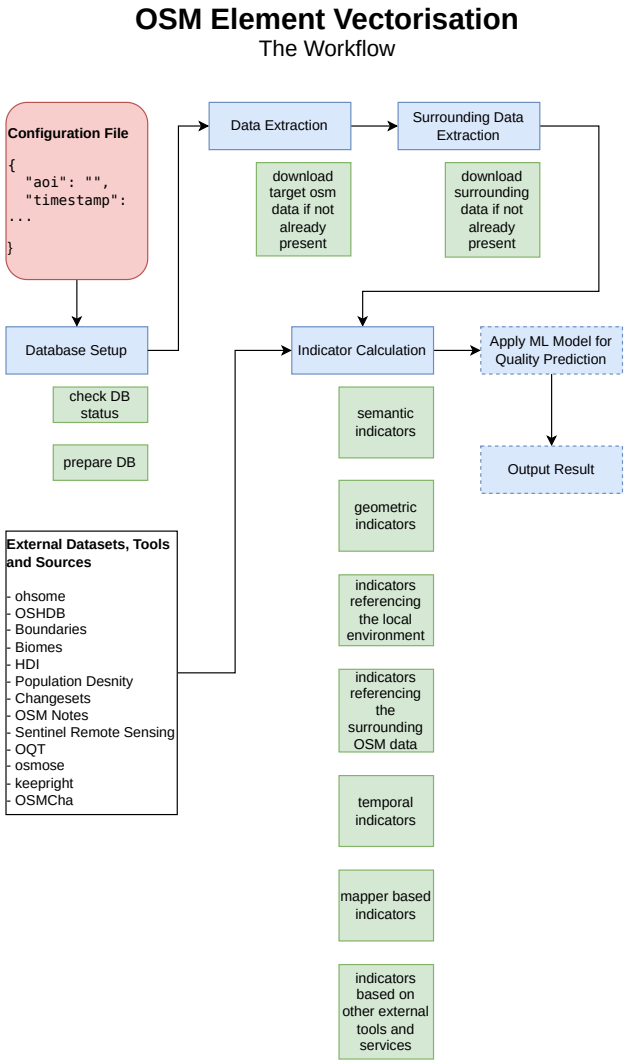

Figure 1. A visual representation of the workflow stages.

Development of the tool was done with LULC information in mind and many calculated metrics can be related to OSM data quality. The software though refrains from any interpretation or intended usage of the results. In fact the tool can be applied to any geometry type like points and lines as well as any OSM information topic where most indicators produce valuable and meaningful insights. Users are therefore encouraged to use the output for their own ideas. Machine learning for example is one application that is currently investigated, by linking the calculated attributes to a certain set of labels for model training. The following section will provide three concrete usage examples to demonstrate the usability of the tool. Namely a certain hypotheses is tested and two explorative data analyses are conducted.

2.2 Application

To prove the potential of the tool, it was applied in a use-case study of LULC elements. LULC information is an interesting aspect of OSM. On the one hand this information is central as it provides the base layer and background for most rendered maps based on the data. It also gets more and more attention from the remote sensing community where it is used e.g. as a training label source for machine learning for image classification problems (e.g. Schultz et al., 2017; Fonte et al., 2019; Schott et al., 2022; Vargas-Munoz et al., 2021). Yet, this information has a challenging stand within the OSM ecosystem as it is more difficult to map compared to other map topics. Natural ambiguities and an ever growing tagging scheme consisting of sometimes overlapping or ambiguous definitions are some challenges mappers face.

1000 out of the then globally existing 62.9 million LULC elements were randomly sampled on 2022-01-01. Only polygonal objects with at least on of the LULC defining tags[5] were considered. These elements' IDs were then fed to the OEVT to extract the data and calculate the described metrics (see Appendix). The sample therefore represents a random selection. Yet, there is a second dimension of randomness that could be used, when sampling polygonal geographic objects: the random location. Sampling at random location would potentially avoid a bias towards more densely mapped areas such as Canada or Germany. However, it would introduce another bias as large objects, such as those representing the Sahara dessert, have a much higher chance of being selected than smaller objects. An analysis of all OSM LULC polygons in 2021 showed this bias (c.f. figure 2).

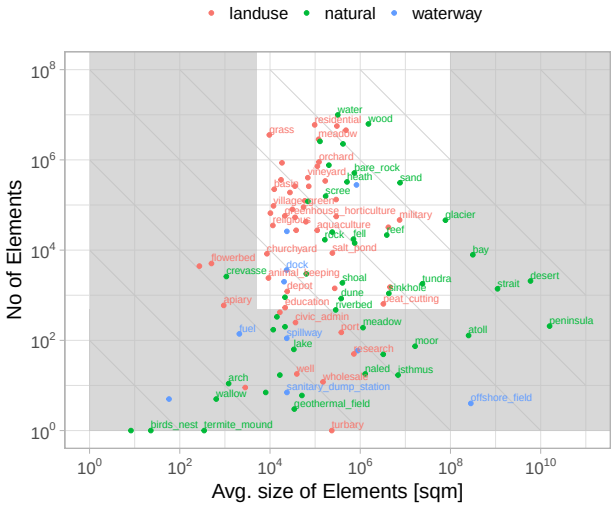

Figure 2. OSM LULC tags by their frequency and size, based on all OSM LULC polygons in 2021. Colors represent the tag-key. The gray horseshoe defines an area where tags can be classified as too small (too detailed), too large (too general) or too rare (not in use). Diagonal grid lines define tags of equal importance meaning equal global areal coverage.

Grasping the amount of new data and insights generated by the tool can be challenging, especially when it comes to indicator interactions. We therefore chose a) a hypotheses driven statistical approach to selected data aspects and b) an automated clustering method to analyse and structure the data and demonstrate the usability of the new toolchain.

Three hypotheses describe our assumptions of the triangular relation between the size of OSM objects, their age and their location in terms of population density (c.f. figure 3). We first hypothesise, that a general mapping order exists where the OSM

[5] The ohsome-filter is a textual composite that clearly defines a set of elements. The used filter is defined in `https://github.com/GIScience/ohsome-quality-analyst/blob/c04f965cff819f529188dfb3061be96fdb48f948/workers/ohsome_quality_analyst/ohsome/layer_definitions.yaml#L310`.

community first concentrates on or arises from urban areas before moving to rural areas. This order has been found in snapshot assessments where at a given point in time, rural areas were more complete and of higher quality than rural areas (Haklay, 2010; Zielstra and Zipf, 2010; Girres and Touya, 2010). Yet, it is unclear if these findings still hold true today on a global scale and how strong this relation is. This assumption is tested by hypotheses 1 (H1): *There is a positive correlation between the object age and the population density*.

Hypotheses 2 (H2) targets the size of elements in relation to their location in terms of population density. Namely, we assume a 'natural' interaction between these factors where areas with higher population density are more fragmented and therefore exhibit smaller elements while areas with low population density such as forest are often larger objects. *There is a negative correlation between the object size and the population density*.

Hypotheses 3 (H3) targets the size of elements in relation to their age. Here, two opposing assumptions exist: Large geographical entities may be mapped first and regions may be first coarsely drafted before adding details. This would lead to old objects being of larger size. Yet, hypotheses 1 and 2 contradict this tendency: according to H1 and H2 younger objects would be in areas with less population density and therefore tend to be larger. We therefore hypothesise that *there is no significant correlation between the objects' age and their size*.

All three hypotheses were tested separately using Kendall's τ as a non-parametric correlation metric (Hollander and Wolfe, 1973) and the method introduced by Benjamini and Hochberg (1995) to adjust the p-value for multiple comparisons.

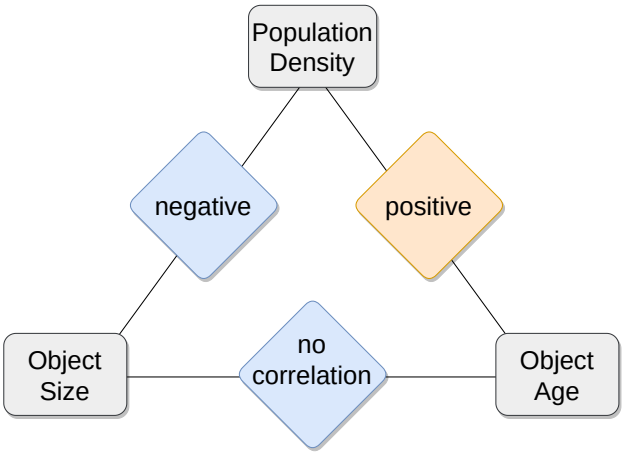

Figure 3. Hypothesised triangular relationship between the object size, the object age and the population density. Interactions that were confirmed by statistical testing are marked in blue.

In a second experiment these three data aspects of object size, object age and population density were tested for regional tendencies meaning their spatial distribution over the seven continents. First, South America, Oceania and Antarctica were grouped into an 'other'-group to assure a minimum of 30 elements per continent(-group). A general Kruskal–Wallis rank sum test (Hollander and Wolfe, 1973) detected significant ($p < 0.02$) differences between the continents for all three data aspects. These effects were then further analysed in pairwise Wilcoxon rank sum tests (Hollander and Wolfe, 1973) using the same p-value correction method as above.

Additionally a k-means clustering (Hartigan and Wong, 1979) with five cluster centers was performed to identify multi-attribute clusters. K-means is prone to extreme values. Therefore, any values extending beyond the lower or upper bound defined by equation 1:

$$b_l = Q_1 - 1.5 \times IQR; b_u = Q_3 + 1.5 \times IQR \quad (1)$$

where b_l, b_u = lower/upper bound
Q_1, Q_3 = first/third quartile (25/75%)
IQR = inter quartile range ($Q_3 - Q_1$),

were rounded to the respective bound (McGill et al., 1978). In addition all numeric variables were standardised to have a mean of 0 and a standard deviation of 0.5 (following Gelman, 2008) and all non-numeric variables were converted to dummy variables. In addition the data was stripped off any geographic information as we were hypothesising that the different clusters might be linked to different geographic regions. Namely, the continent, biome and Human Development Index (HDI) information was removed. The best fit from 1000 runs with random starting points was used for the cluster assignment of the OSM objects.

3. RESULTS

The main hypotheses regarding the mapping order (H1) could not be confirmed. In fact, the estimated correlation was slightly negative meaning that for the used sample, objects in urban areas were younger than in rural area. The negative correlation between the object size and the population density (H2) could be confirmed with a p-value$<$0.01 though the τ was only -0.096 meaning that while a correlation was found, the effect was small (figure 4). H3 was equally confirmed, but only after p-value correction (p-value=0.14). Because the conducted analyses was of explorative nature, we were willing to accept significant results up to a p-value of 0.1.

The random dataset was dominated by European objects (61%) followed by North America (18%), Asia (9%), Africa (7%), South America (4%), Oceania (2%) and the Antarctic (0.1%). North America and Europe exposed significantly smaller object sizes compared to the 'other'-group (South America, Oceania, Antarctica). More significant differences were observed for the age comparison where Africa contained the youngest objects followed by Asia having significantly younger objects than Europe and North America. North America on the other hand contained the oldest objects in comparison to all other continents (figure 5). The population density in the region the objects were located at also showed significant differences between the continents. Asian elements were located in more densely populated areas compared to Africa and Europe, that had medium population densities, while North America and the 'other' group showed the lowest population densities.

3.1 k-means

The clusters detected by the k-means algorithm were similar in size with between 131 and 255 elements. Cluster 1 was composed of many large, complex objects that were often changed,

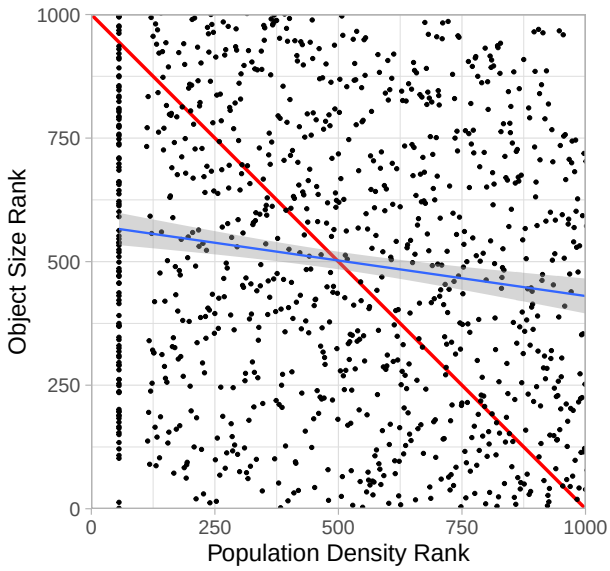

Figure 4. Interaction between the population density and the object size. The diagonal red line represents our idealised second hypotheses, the blue line is a linear regression estimate.

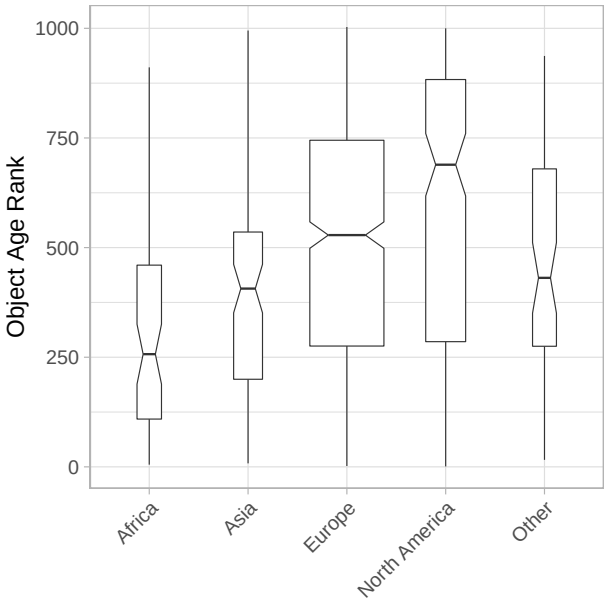

Figure 5. Object age in relation to the continents the OSM LULC objects were located. Africa contained significantly younger objects while North America contained significantly older objects.

especially by inexperienced mappers. A relatively high share, in comparison to the whole random sample, of this cluster were forest or built-up areas and elements on the European continent. Contributions to objects in this cluster were often flagged as 'bad data' by OSMCha[6].

Cluster 2 contained a remarkable share of objects that were located in low HDI areas as well as a relatively high share of elements in Africa or Asia. Many objects were relatively young and contributed by experienced but non-local mappers. Over 80% of these polygons were flagged as imported data by OSMCha.

Cluster 3 was dominated by lakes in North America, specifically Canada. The data source for these lakes was often marked as an import from official sources. While these imports already dated back relatively long, the data has seldom been updated since then, even though a relatively high amount of individual mappers was active in these areas.

Cluster 4 was more diverse than the previous clusters, meaning it was more similar to the general random sample distributions. It had a slight tendency towards recently added elements, elements in Europe and forest polygons.

Cluster 5 had similar attributes to cluster 2. Yet, elements in this cluster were often added by inexperienced mappers and mappers who exclusively used the iD editor[7]. This cluster also had a higher share of agricultural objects like farmland.

4. DISCUSSION

The presented OEVT, although still under active development, enables users to generate multi-faceted data insights. Their usefulness was shown in three example applications. The new insights gained from these applications already increased our knowledge of OSM LULC information at the global scale.

The hypotheses driven testing showed that the 'natural' relation between the object size and the population density exists, even if this effect is surprisingly small. The analysed sample may though be too generic to gain detailed insights in this domain. On a global scale, many influencing factors may overlap or intervene with each other, hindering the extraction of single detailed effects. For the example at hand we can assume that there are multiple regional communities or active mappers with individual mapping styles. The mapping detail in urban or rural regions will therefore be linked to these and other factors as well, not only the population density. The population density itself may not be generalisable on a global scale. The same level of fragmentation, meaning object size distribution, may be reached at different population density values, depending e.g. on the continent.

A similar interpretation has to be drawn when looking at the two other tests. No global mapping order could be proven where mappers first concentrate on urban areas before moving to rural areas. Yet, this does not imply that this mapping order does not exist in certain sub-regions. In addition, the age of an object is a fragile metric that highly depends on the mapping style of local mappers. Mappers frequently decide to delete and redraw elements instead of changing the original object, especially if the object was only a coarse approximation. This though 'resets' the object age, meaning that urban areas may have a high share of young objects because they are still actively mapped and maintained, even though they started their map-appearance relatively early. A similar conclusion has to be drawn from the non-correlation between the object size and their age meaning that objects of different size are drawn at any point in time. Yet, regional specialities may exist and revisiting this data aspect in the future may lead to surprising insights. As a conclusion of the hypotheses driven testing, we can manly record that OSM

[6] https://osmcha.org/
[7] https://github.com/openstreetmap/iD

is diverse and multi-faceted and can hardly be described on a global scale, making our tool a valuable contribution facilitating these local and multi-faceted insights in the future.

Some of these regional trends were exposed by the second statistical analyses which characterised OSM as a predominantly European and North American project (c.f. Budhathoki and Haythornthwaite, 2013) that only in later stages spread to other continents. These findings also supported the notion of OSM being subject to a digital divide where European and American users have had easy access to mapping equipment which may be more difficult for African, Asian or South American users. Additionally, one can argue that Europe's LULC data may be of the same age as North American, yet has been more often updated and thereby redrawn, resetting the object age as described above. The object size difference on the other hand is not surprising, considering the findings and assumptions of H2, seeing that the 'other'-group is made up of continents with large areas of low population density like Australia, the Amazon rain forest, Antarctica and the Argentinian Pampas. The small object size in Europe and North America can though also be caused by the high level of detail already reached in these regions. The population density suggests that there may be a mapping order for OSM LULC objects present in Asia, where large areas with low population density exist (e.g. in Mongolia) but only high population density areas are mapped. Surprisingly, North America was very well mapped with respect to LULC in low population density areas, which explains the third k-means cluster. The imported North American lakes contributed a large number of objects to the database, while being located in very remote areas.

The aforementioned cluster 3 was the most striking finding by the algorithm. Although the cluster does not exclusively contain North American lakes, it describes a certain object archetype that is rarely found outside this domain e.g. with some forest or wetland elements. This element group makes up a considerable share of the global data and must therefore be taken into account when analysing or describing the global dataset. Additionally this data type, sourced from official data through imports, may also be relevant for other information domains such as the road network or building footprints, where imports are a legitimate data source. This supports the ambivalent findings by Juhász and Hochmair (2018), who stated that imports can lead to a more active community but may also hinder community development. Cluster 3 may cause the latter because these object are located in remote areas with little local population and therefore little local community potential.

The other clusters cannot be categorised or interpreted that easily. Cluster 2 also contained references to imported data. Yet, this flag was awarded by an automated mechanism through OSMCha that considers all changes made within a changeset[8]. Presumably, this mechanism mostly relies on the changeset size, i.e. the amount of data uploaded via one changeset. Therefore cluster 2 could also be caused by power-mappers, who create a large amount of data within short time frames. Objects from these changesets could therefore be misinterpreted as imports. This assumption is supported by the high experience of contributors to objects in cluster 2, which in turn also leads to a high diversity in object types touched by these users. Power mappers are frequently active in mapping campaigns organised e.g. by the Humanitarian OpenStreetMap Team (HOT) or UN Mappers. Combining this assumption with the low HDI, the low completeness, the large share of African and Asian elements, the youth of the objects as well as the relatively high remoteness of contributors and the data source coming from non-Bing aerial images further supports this assumptions. Bing can be seen as the standard aerial image source for OSM while non-Bing images are often used e.g. in distaster response contexts, where more recent possibly commercial aerial images are made available. Humanitarian mapping teams are often active in in the described world regions and predominantly rely on remote mapping campaigns. Only UN Mappers though are active contributors of LULC data while HOT concentrates more on building and infrastructure data.

Care has to be taken to not assume indicator interactions without proving them. For example cluster 2 contains a larger amount of young as well as a larger amount of African elements. Yet, it is unclear if the cluster contains young African elements or if in fact the African elements contribute to the older share of elements in the age distribution.

Cluster 5 had similar attributes to cluster 2. Striking differences were observed for the used mapping software, the low experience of mappers as well as the higher share of agricultural objects. UN Mappers could also play a role here assuming this platform attracts a large amount of newcomers like HOT. The remoteness of users though is average. This could either be caused by an inherent problem of the remoteness calculation, where the center of activity of mappers is considered as their home region. On-time contributors or beginners may have their center of activity at the location of their only map contribution, e.g. in Africa, even though they reside in Europe and made a remote contribution. Alternatively, they could actually be local mappers that joined the project individually or were activated by the aforementioned platforms. Again, it has to be emphasized that cluster 5 does not exclusively contain African and Asian elements. The object type created by inexperienced UN Mappers in Africa or Asia may though describe another object archetype that can also be found on all other continents and in all other LULC classes.

Cluster 1 did not show any strong relation to imports or organised mapping activities. It rather agglomerates larger objects that share similar attributes that may be caused by their size. Especially coarseness, complexity and frequent changes were common for large objects. It takes a lot of effort to draw large objects in high detail and often is not required e.g. for forests that have long straight edges. At the same time these objects often contain holes or other geometries that strongly deviate from simple triangles, circles or squares explaining the higher complexity. Frequent changes can also be caused by size as these objects overlap large areas and therefore interact with many changes or contributions in their surrounding. Additionally, this cluster had a relatively high share of built-up areas, mostly residential areas that inherently have a high community potential and therefore a high 'change potential'. This notion is supported by the high localness of mappers.

Finally cluster 4 could be seen as a sort of 'remainder' cluster, due to its high diversity or better said its high amount of relatively average indicator values. To fully understand this and all other clusters though, more detailed and multidimensional analyses are needed.

[8] A changeset in OSM is a collection of map edits similar to a commit in git.

5. CONCLUSION AND OUTLOOK

The presented work establishes a new software for OSM data transformation. This FOSS workflow that draws on different datasources, scientifc studies, community efforts and new methods helps to analyse OSM data in an automated fashion and converts the qualitative OSM raw data into quantitative indicators and metrics. Meanwhile it does not judge or interpret any of its results making its application a free choice for the user. Exemplary studies have shown the applicability and usefulness of the tool for specific hypotheses or explorative questions as well as for automated procedures such as clustering. An additional use-case is already contained within the tool that provides a machine learning based quality estimation. This feature though is still in an experimental phase. The future work on the tool will include more tests on its applicability on other OSM topics such as infrastructure, buildings or Points Of Interest as well as an increase in the number of calculated indicators. The cluster analyses for example provided hints with respect to organised mapping activities that can often be identified though hashtags (#) in changeset comments. Such an information could also be gathered by future releases of the tool. In addition, work has already started on further usage facilitation. While the current status allows the execution of the tool by any user via Docker, a hosted web interface where all necessary auxiliary data is already present would also enable less tech-savvy users to test the tool.

More concrete insights were gained by the exemplary analyses that gave a good overview of the current state of LULC information in OSM. Confirming the widespread notion of a 'diverse dataset' for this domain, yet, finding regional data communities as well as global archetypes hints towards a much larger amount of data aspects that still remain to be unveiled. The strong role that organised mapping activities are suspected to play in LULC data generation, as well as the strong regional and topical limitation of the 'imported' cluster 3 are only two insights that need further and more detailed investigation in the future. Especially, a more detailed look into objects that do not fit the general cluster narrative, e.g. for the North American lakes will be an inspiring topic that we are looking forward to drill into. Additionally, the gained insights should be linked to other topics such as data quality e.g. by identifying elements that need the communities attention.

Nevertheless the presented work already enables the OSM community to make a more informed planning of future activities like organised mapping or data curation efforts and enables data consumers to make informed decisions on data usage. It also facilitates the reproducibility of the cited studies which date back up to 12 years and are certainly worth being revisited. We hope that this new tool will fuel these and many other data analyses and enable more data analysts to join us on our path for a more thorough understanding of the OSM data base.

ACKNOWLEDGEMENTS

This work is a result of the IDEAL-VGI project funded by Deutsche Forschungsgesellschaft (DFG) under grant number 424966858. Sven Lautenbach acknowledges support by the Klaus-Tschira Stiftung, Germany.

REFERENCES

Barron, C., Neis, P., Zipf, A., 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 18(6), 877–895. https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12073.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.

Brückner, J., Schott, M., Zipf, A., Lautenbach, S., 2021. Assessing shop completeness in OpenStreetMap for two federal states in Germany. *AGILE: GIScience Series*, 2, 20.

Budhathoki, N. R., Haythornthwaite, C., 2013. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist*, 57(5), 548–575. https://doi.org/10.1177/0002764212469364.

Fonte, C. C., Lopes, P., See, L., Bechtel, B., 2019. Using OpenStreetMap (OSM) to enhance the classification of local climate zones in the framework of WUDAPT. *Urban Climate*, 28, 100456.

Gelman, A., 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865–2873.

Girres, J.-F., Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, 14(4), 435–459.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design*, 37(4), 682–703.

Hartigan, J. A., Wong, M. A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108. http://www.jstor.org/stable/2346830.

Herfort, B., Lautenbach, S., de Albuquerque, J. P., Anderson, J., Zipf, A., 2021. The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific Reports*, 11(1), 1–15.

Hollander, M., Wolfe, D. A., 1973. *Nonparametric statistical methods*. Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley, New York.

Jokar Arsanjani, J., Mooney, P., Zipf, A., Schauss, A., 2015. Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. J. Jokar Arsanjani, A. Zipf, P. Mooney, M. Helbich (eds), *OpenStreetMap in GIScience: Experiences, Research, and Applications*, Springer International Publishing, Cham, 37–58.

Juhász, L., Hochmair, H. H., 2018. OSM Data Import as an Outreach Tool to Trigger Community Growth? A Case Study in Miami. *ISPRS International Journal of Geo-Information*, 7(3). http://www.mdpi.com/2220-9964/7/3/113.

McGill, R., Tukey, J. W., Larsen, W. A., 1978. Variations of Box Plots. *The American Statistician*, 32(1), 12–16. http://www.jstor.org/stable/2683468.

Mooney, P., Corcoran, P., Winstanley, A., 2010. A study of data representation of natural features in openstreetmap. *Proceedings of GIScience*, 150, 150–156.

Neis, P., 2022. Overview of resultmaps. `https://resultmaps.neis-one.org/` (25 Mai 2022).

Neis, P., Zielstra, D., Zipf, A., 2013. Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5(2), 282–300.

Neis, P., Zipf, A., 2012. Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2), 146–165.

OpenStreetMap Contributors, 2022a. Category:osm processing. `https://wiki.openstreetmap.org/wiki/Category:OSM_processing` (25 Mai 2022).

OpenStreetMap Contributors, 2022b. List of osm-based services. `https://wiki.openstreetmap.org/wiki/List_of_OSM-based_services` (25 Mai 2022).

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, Software and Standards*, 4(1), 1–12.

Schott, M., Grinberger, A. Y., Lautenbach, S., Zipf, A., 2021. The Impact of Community Happenings in OpenStreetMap—Establishing a Framework for Online Community Member Activity Analyses. *ISPRS International Journal of Geo-Information*, 10(3), 164.

Schott, M., Zell, A., Lautenbach, S., Zipf, A., Demir, B., 2022. LULC multi-tags based on OSM, Version 0.1. https://gitlab.gistools.geog.uni-heidelberg.de/giscience/ideal-vgi/osm-multitag.

Schultz, M., Voss, J., Auer, M., Carter, S., Zipf, A., 2017. Open land cover from OpenStreetMap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 63, 206-213.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., Haklay, M. M., 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1), 139–167. https://doi.org/10.1080/13658816.2016.1189556.

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

Vargas-Munoz, J. E., Srivastava, S., Tuia, D., Falcão, A. X., 2021. OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1), 184–199.

Zielstra, D., Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. *13th AGILE international conference on geographic information science*, 2010, 1–15.

APPENDIX

The raw data as well as a fully documented R script with additional images for the example analyses is available at `https://gitlab.gistools.geog.uni-heidelberg.de/giscience/ideal-vgi/foss4g2022-analyses`.