DATABASE-SUPPORTED CHANGE ANALYSIS AND QUALITY EVALUATION OF OPENSTREETMAP DATA

A. Martini [1], P.V. Kuper [2]*, M. Breunig [2]

[1] Disy Informationssysteme GmbH, Karlsruhe, Germany – alexander.martini@disy.net
[2] Geodetic Institute, Karlsruhe Institute of Technology, Germany - (kuper, martin.breunig)@kit.edu

Commission IV, WG IV/7

KEY WORDS: OpenStreetMap, Volunteered Geographic Information, spatial database, temporal OSM data, data quality

ABSTRACT:

A significant advantage of OpenStreetMap data is its up-to-dateness. However, for rural and city planning, it is also of importance to access historical data and to compare the changes between new and old versions of the same area. This paper first introduces into a differentiated classification of changes on OpenStreetMap data sets. Then a methodology for an automated database-supported analysis of changes is presented. Beyond the information already provided from the OpenStreetMap server, we present a more detailed analysis with derived information. Based on this approach it is possible to identify objects with attributive or geometric changes and to find out how they exactly differ from their previous versions. The analysis shows in which regions mappers were active during a certain time interval. Furthermore, a time based approach based on various parameters to determine the quality of the data is presented. It provides a guideline of data quality and works without any reference data. Therefore, an indication about the development of OpenStreetMap in terms of completeness and correctness of the data in different regions is given. Finally, a conclusion and an outlook on open research questions are presented.

1. INTRODUCTION

Due to the concept of Volunteered Geographic Information (Goodchild, 2007), OpenStreetMap (OSM) has become a worldwide and open accessible geo dataset (Haklay and Weber, 2008). This data collection is characterized by its high dynamics and makes it very up-to-date in many aspects (Mooney and Minghini, 2017). At the same time, the focus lies on the current state of the "image" of our world. By additions and corrections of voluntary users, the geodatabase is maintained by the OSM community on a minute-by-minute basis and thus kept more up-to-date than a lot of proprietary map material.

However, also historic data are of high importance e.g. for city and rural planning scenarios. To the knowledge of the authors, there is no well-structured historicization approach of OSM data so that OpenStreetMap data can be maintained in a suitable data structure and responses to temporal queries can be delivered. Due to the large amounts of data used in OSM, the efficient storage of changes should be realized in a suitable data model with temporal references. Based on a historicization of OpenStreetMap data, changes can be classified more accurately than this is enabled today.

Furthermore, an automatic quality assurance process does not exist in OSM and takes place only with the help of corrections by the users. The OSM Project renounces on a fixed structure of the data. For unification there is only a recommended procedure describing how to collect the data correctly. Thus the question of data quality (Zielstra and Zipf, 2010; Koukoletsos et al., 2011; Fan et al., 2014; Antoniou and Skopeliti, 2015; Mohammadi and Malek, 2015; Senaratne et al., 2017; Fonte et al., 2017) is inevitable. The principle of data collection entails high risks and raises doubts about the quality of the data. In this paper we focus on a temporal approach to assess the quality of OSM data sets during given time intervals.

This paper is organized as follows: In section 2 related work on the analysis of OSM data is introduced. Section 3 figures out the methodology on data change analysis and data quality used in this work. Implementation issues are also discussed in this section. Section 4 presents experimental results using OSM data sets around the city of Karlsruhe, Germany. Finally, Section 5 gives a conclusion of the approach and discusses open research questions.

2. RELATED WORK

There exist multiple approaches for the assessment of the quality of VGI data which can be described by quality measures and quality indicators (Senaratne et al., 2017). Basically, a distinction is made between two different approaches: measures based on a comparison of VGI data with external data sources and intrinsic indicators (Antoniou and Skopeliti, 2015; Mohammadi, and Malek, 2015). Different measures for the quality of VGI data were specified and used if authoritative data is available: e.g. completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy, and usability (Fan et al., 2014; Antoniou and Skopeliti, 2015). Based on the well-known example of OSM data it was proven that the quality of VGI data can be compatible and better than authoritative data in multiple measures (Brovelli et al., 2017; Brovelli and Zamboni, 2018).

As VGI is getting more and more popular, some VGI datasets are more detailed and accurate than corresponding authoritative datasets. Furthermore, authoritative datasets are not always available. Therefore, internal indicators and intrinsic methods

* Corresponding author

are needed to measure the quality of such VGI datasets. In this context the heterogeneity of the data was identified as one major issue (Koukoletsos et al., 2011; Vandecasteele and Devillers, 2015; Mohammadi and Malek, 2015).

Multiple approaches were developed for an automated inspection of data completeness and positional accuracy based on intrinsic methods (Haklay et al., 2010; Koukoletsos et al., 2011; Fan et al., 2014; Mohammadi and Malek, 2015). Goodchild and Li (2012) divided such approaches in three categories: crowd-sourcing, social and geographic to assure VGI quality.

There are multiple studies on the completeness of street networks in OSM (Haklay, 2010; Ludwig et al., 2011). While most studies were based on the comparison of the data at a certain time, Corcoran et al. (2013) studied the growth of street networks in Ireland over a time period from 01-11-2007 to 01-10-2011. It was discovered that the major changes were found in densification and exploration. Zielstra and Zipf (2010) did a similar study for certain areas in Germany. Another study examines the degree of completeness of OSM data on buildings in Germany over time based on a comparison with authoritative data (Hecht et al., 2013). Brovelli et al. (2016) examined the completeness of buildings in OSM in the area of Milan in Italy.

3. METHODOLOGY

3.1 Handling of temporal data in OSM

It is desirable to have an OSM database that is always complete and up-to-date. But to map every real world change directly in the database is not a realistic scenario. Usually, mappers tend to limit themselves onto updating objects in certain regions or, for example, only improve certain object properties. In most cases, such changes are thematically similar and are eventually added to the database as a so-called changeset. The timestamp of such a changeset upload documents the moment of the change within the database. By definition, this is the *transaction time* and noted as the start time "valid_from" of the object. However, the validity period indicates the period in which the object in the real world correspond to the mapped state. Changes to objects in the real world are usually recorded by users in a delayed manner. Furthermore, it is not possible to change past periods. But it is always possible to alter the current state of an object. Consequently, there is no way to get a correct validity period of OSM data objects.

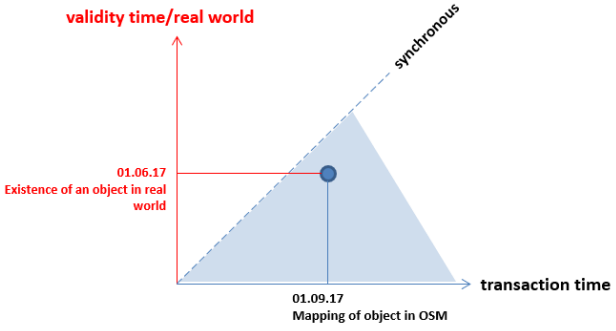

Figure 1. Orthogonality of validity and transaction time

Therefore, the validity period or the mapping of the real world cannot be captured and is therefore not available as information in the database. Only the transaction time, as the time at which the changes are recorded by mappers, is known. As a result, in comparison to Figure 1, all acquisitions in the database take place in time after changes in the real world happened, and thus lie below the axis of synchronicity.

3.2 Classifying changes in OSM datasets

In this work, the change classification is limited to geometric and attributive comparisons of OSM object versions. Figure 2 describes the procedure for classifying changes for OSM nodes.

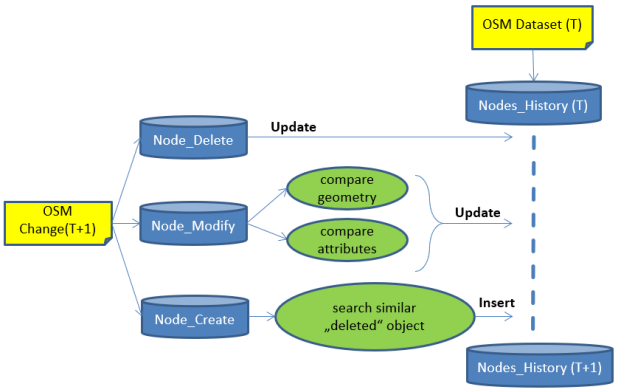

Figure 2. Classifying changes of nodes

The data base required in our approach of change classification is an OSM data record, for example, a planet file of an event to be analysed, and one or more subsequent change files. First, the planet file has to be imported into the database. Due to the schematic structure a change file is already structured in *Delete*, *Modify* and *Create* and must be further classified in the database after an import.

3.3 Database-supported change analysis

In our database approach objects that were deleted within a change file are not deleted accordingly, but only marked, since this information should be maintained. Here, the import mechanism and also our prototype system differs fundamentally from conventional OSM tools. Such tools usually only focus on the actuality of the data and therefore delete “unnecessary entries” completely from such a database.

Changed objects can now be subjected to a closer look. First, a database-side comparison by means of a spatial query is executed. Here it is checked whether the old object version matches the new one concerning its geometry. If this is not the case, a geometric change can be classified and the distance to the previous point is determined. This is followed by an attributive comparison. Special functions allow us to compare the tags that are stored as key-value pairs in a list of an object. Here it can be classified whether tags have been added, deleted or changed. After the classification, the results are noted in a classification column in the corresponding database table. During an import, the validity status of the corresponding old object is terminated, a change is classified and the new object is transferred to the database. New nodes can be loaded directly into the database. However, if a similar object already exists within a buffer region, the process of change analysis is triggered. The result is a database containing current, deleted and multiple versions of changed objects.

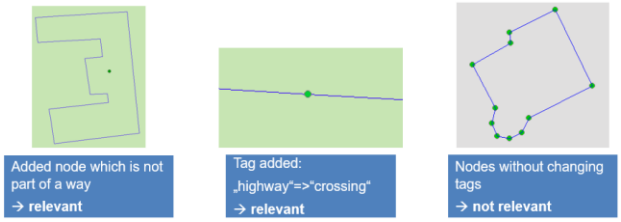

Figure 3. Determination of relevant and irrelevant changes

Due to the large number of changes, the extent to which a change is a relevant or irrelevant change is considered. Examples are given in Figure 3. If the number of changes within a particular region shall be determined, the nature of the change should be further investigated.

3.4 Evaluation of data quality

Our concept of data quality evaluation provides a summarized qualitative statement about collected OpenStreetMap data. The approach is based on intrinsic investigations and therefore does not require higher-level map material, but only the full history file of the region to be examined. As a consequence, however, the results of this evaluation can only be understood as a rough guideline. Methods are developed to obtain a rough overview of the data quality and a suitable ranking method for quality assessment based on various parameters. The ranking method to estimate the quality has to be adapted for individual use cases. In this paper the initial approach is to classify the parameters to well distributed rankings.

User contributions
Regions with a few users' participation appear to be poorly controlled. Furthermore, the expectation of good quality posts is higher for users with local knowledge than for outsiders. By analysing their worldwide activities, users with local knowledge could potentially be identified.

Temporal development of way length in a road network
The development of the way length alone, filtered after a certain tag, can provide information about the completeness of the data. For example, the increase in the length of motorways in well-recorded regions is stagnating over time. Although it may be that new roads are built, they do not significantly affect the total length of the road network.

Collection of data
The source of the data is an important factor of data quality. E.g. a data donation from an official department is much more geometrically accurate than a recorded point from a smartphone.

Ranking procedure for the determination of data quality
We introduce a procedure, which enables a quality statement of the OpenStreetMap data due to intrinsic investigations. Various parameters are shown for this purpose. Depending on the application, they can provide qualitative statements about the data individually or in a combined ranking result within a defined region. The concept can be flexibly extended to any parameter. Figure 4 shows an example of a procedure for determining the data quality.

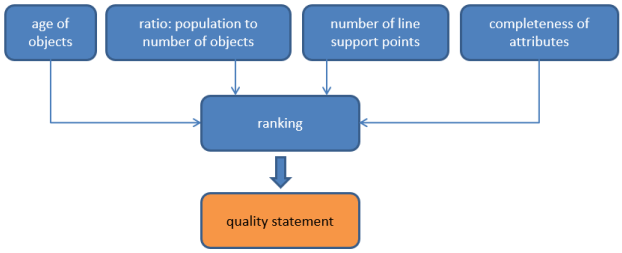

Figure 4. Derivation of an overall quality statement

As a result, a tendency emerges in which areas of a region a good or poor data quality can be expected. To realize the visualization with a good performance, the region is divided into grid cells. Each grid cell therefore represents the objects within their extents. The size of the cells is determined in advance.

Age of objects: The younger the objects, the better the data
First, a calculation is made of the average age of the objects within each grid cell of the region to be examined. This does not take into account the fact that older objects could already be detected correctly. If more than 15% of the objects within a grid cell are younger than 6 months, it is assumed that recent controls of this environment did not require corrections of the older objects. In order to take this aspect into account, the quality rating in these areas is improved by a ranking value.

Number of objects: The more inhabitants, the more objects
If many inhabitants are living within the area of a grid cell, it can be assumed that it has a good infrastructure and therefore many objects in OSM are to be expected. A rural area is usually characterized by large, uniform areas. Meadows, forests, few houses and roads result in a smaller number of objects to be detected. In comparison, an urban area, due to the dense house construction, includes a variety of objects.

Number of support points for lines: The more points, the more accurate
As a rule, the accuracy of an object improves by adding more points. In this way, the reality is depicted more accurately, especially in the case of curved objects. Even polygons of buildings can be depicted in greater detail. To determine the quality of the data, it is considered here on how many meters a line support point is expected on average.

Using this method detailed geometries shall be distinguished from less detailed ones. The problem here is that correctly detected objects, such as large building complexes or straight roads, may be classified as poor quality. A differentiated study between urban and rural regions is conceivable.

Number of attributes: The more attributes an object has, the more complete it is represented
It can be stated that a higher number of properties, or tags, suggests a detailed description of the object. With this method, only the set of attributes can be determined. A more meaningful examination of the quality is conceivable considering the recommended attributes from the OSM community depending on the object type.

Combination of parameters
Finally, all intermediate results of the ranking can be combined into a total ranking. The result is a statement about how good the data quality is within a grid cell.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-535-2019 | © Authors 2019. CC BY 4.0 License.

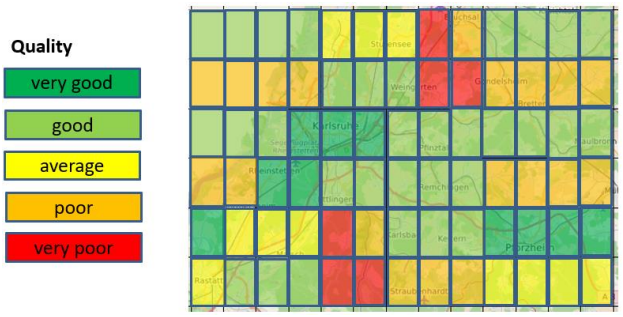

Figure 5. Map where each grid cells shows the total quality

The sum of the individual rankings is calculated. Thus an overall assessment of each grid cell can take place, see Figure 5. If the individual rankings of an area always achieve the best results, the overall ranking will be very good. If certain parameters are of greater importance, a corresponding weighting of the individual rankings can also be carried out in this step.

3.5 Implementation

Our realization of the spatio-temporal management of OSM data is based on a PostgreSQL database (PostgreSQL Development Team, 2019) with the spatial extension PostGIS (PostGIS Development Team, 2019). The change and quality analysis takes place on the database side. The visualization tool used is Cadenza (Disy Informationssysteme GmbH, 2019). This application is particularly suitable for map displays and diagrams and thus offers extensive statistical analysis functions. The basic workflow is shown Figure 6.

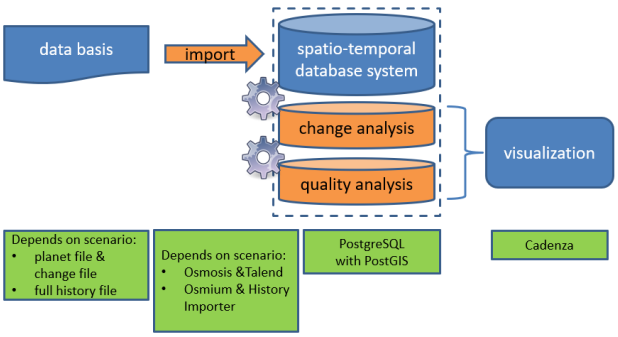

Figure 6. Workflow of change and quality analysis

Scenario "Change analysis"
For the analysis and classification of the changes within a certain period, the combination of a OSM planet file with temporally successive change files has proven to be a suitable solution.

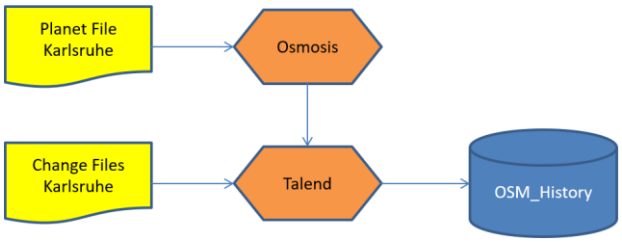

Figure 7. Import workflow: change analysis

A change file contains new or changed objects. To perform a comparison and thus a detailed change analysis, the previous version of the object has to be held available. This can be found in the corresponding Planet File to PostGIS. Osmosis is a command-line utility developed by the OSM community, which helps to import Planet Files. However, the change files still have to be parsed with a self-developed mechanism. Talend, a data integration tool, is used for an automation. Figure 7 shows the tools that are used for the analysis process.

Scenario "Quality evaluation"
For quality analysis, the full history file of the corresponding region should be used as data basis, see Figure 8. This includes all versions of the objects and thus offers the opportunity to analyse the temporal evolution. Even a small region takes up a large amount of data. Using History Importer, a tool based on the functions of Osmium, this data format can be loaded into a database and thus provides an optimal data basis for intrinsic investigations.

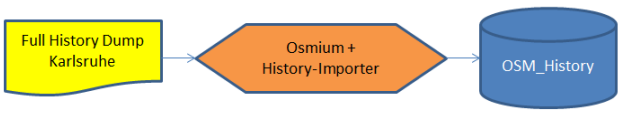

Figure 8. Import workflow: quality evaluation

The objects of both scenarios are ultimately managed in a temporal database. The main differences between the two scenarios "change analysis" and "quality evaluation" are the database and the associated different data formats that are used.

4. EXPERIMENTAL RESULTS AND DISCUSSION

After the processing process within the database, the objects of a change file are classified in such a way that it is possible to visualize the type of the changes that have taken place. When the object information of the respective change is called, additional information is provided, for example, the tags that were added or deleted. According to the legend, the nature of the change can be determined by the colour and shape. In addition to created, deleted or edited objects, objects without changes are also displayed. These arise when the temporal boundary of the Planet Files does not coincide exactly with that of the temporally following change files. The background map is the OpenStreetMap to provide a better overview. The change analysis identifies areas where mappers were active. Therefore, e.g., the update of maps can perform faster because only the rendering of these areas is necessary. Classified changes of nodes and ways are shown in Figure 9 and Figure 10.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-535-2019 | © Authors 2019. CC BY 4.0 License.

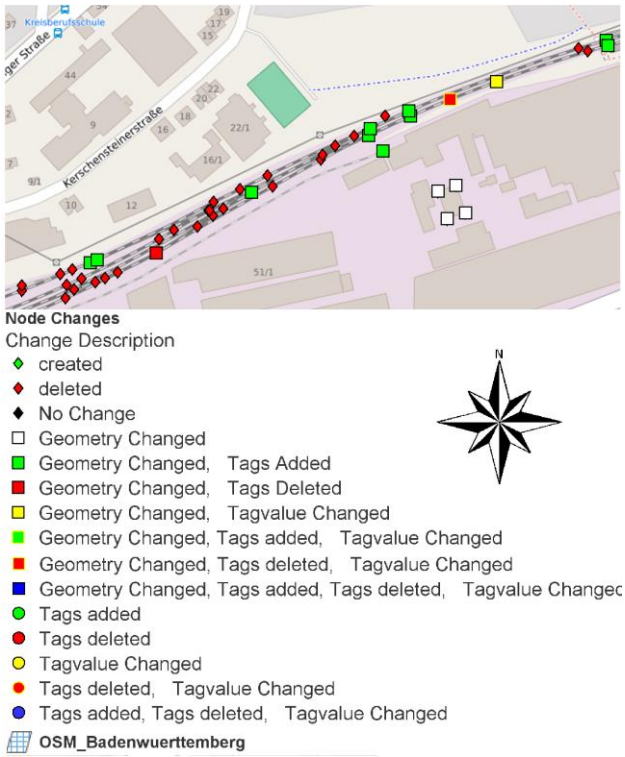

Figure 9. Classified changes of OSM nodes

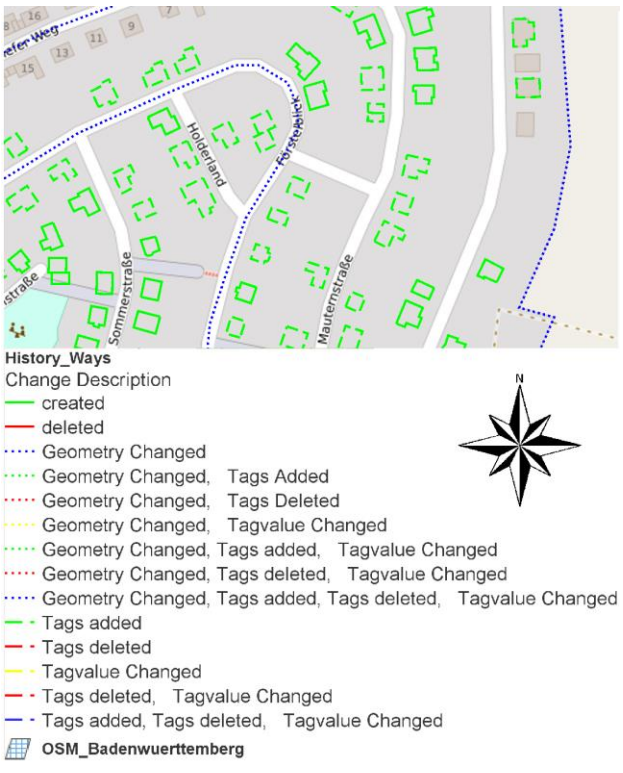

Figure 10. Classified changes of OSM ways

Fig. 11 compares the amount of OSM way data with the number of active users. Contributions are cumulative, starting with most posts from a user. It turns out that only 10% of the users are responsible for 90% of the contributions within the region of Karlsruhe. That means 100 active users are responsible for around 400.000 way objects.

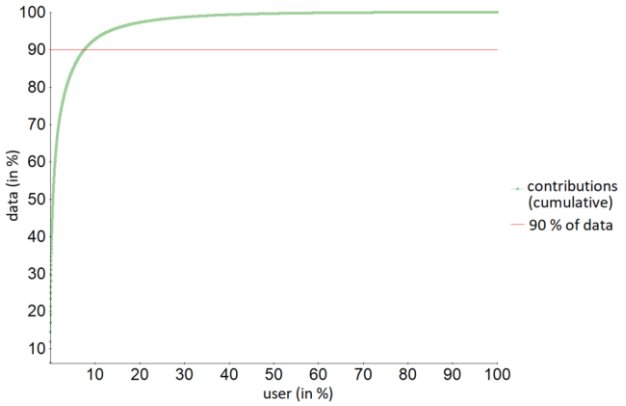

Figure 11. Number of users in comparison to OSM data

When looking at the newly created ways per year in Fig. 12, it is striking that at the beginning of the project only a few areas of the region were recorded. Mainly city centres, connecting roads, parts of the railway network and large rivers were recorded for the first time. Since 2010, a peak in newly acquired objects has been reached, the number is declining. In this case, it can be assumed that the region is largely completely covered and therefore only a few new objects have to be uploaded to the database in order to preserve the image of the real world.

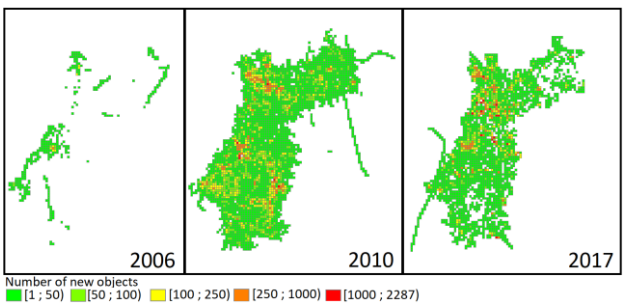

Figure 12. Number of new objects over time

By comparison, in 2006, based on the small amount of data at this time, there are initially no entries of changed objects, see Fig. 13. However, the number is increasing rapidly and remained at a high level since 2010. It is worth noting that a large number of changes in 2017 are in the area of metropolitan areas, such as Karlsruhe or Mannheim. By contrast, in 2010 the changes were more spacious and distributed in rural areas. The consistently high number of changes speaks for a good timeliness of the data.

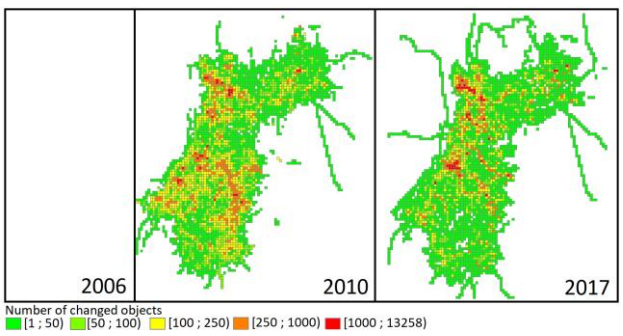

Figure 13. Number of changed objects over time

5. CONCLUSION AND OUTLOOK

In this work a concept for the historicization of OpenStreetMap data has been introduced. This serves as the basis for an automated change analysis and allows the visualization of change of OSM data for various time intervals. In addition, a method for an automated data quality evaluation has been developed using only the OSM Full History file, i.e. no reference data set is needed. The ranking methods include parameters such as the age of the objects, the relationship between the number of inhabitants to the number of objects, the number of support points of a line, and the completeness of the attributes of the objects. A prototypical implementation of our approach based on the geodatabase PostGIS and the visualisation and analysis tool Cadenza provides initial results. It is shown that the weighted sum of the single rankings is well suited to assess the data quality of OSM data for a-priori specified grid cells.

OpenStreetMap focuses on a high actuality of the data. Changes to objects are made by users in the database but are not marked or classified further in any way. With this work, an automated change analysis has been developed, which allows a classification of the differences between two points in time or within a period of time. This classification includes geometric and attributive changes. Nodes and ways of any period can be analysed and visualized. In addition to newly added, changed and deleted objects, also the composition of object changes is determined. The modification of object properties (add, change or delete) are noted in the database. In the case of geometric changes, the distance to the previous position of a node or the change in length of a route is specified. The change analysis can be used to visualize the areas and type of changes that were made within a region. In addition, the nature of the change history of a user can provide useful information about the user's thoroughness.

The principle of free data collection entails high risks and raises doubts about the quality. The content of the OSM project generated by users is not checked for quality or validity. As a consequence, there are potentially high quality differences for different areas. Therefore, a concept was developed, which provides a consolidated qualitative statement. An overview of the data quality as well as a ranking method for the determination of quality was developed with the help of intrinsic approaches. Various parameters for the analysis of the data quality were developed and finally summarized into an overall ranking. Based on the temporal management of the data, the evolution of the data quality over time since the beginning of a project until today can be considered. An annual representation of the parameters makes it possible to illustrate a trend in data quality. The time analysis was realized on the basis of various parameters. The annual number of newly created or changed ways indicate user activity and completeness within a region. As previously mentioned, another measure of completeness is the consideration of the ratio of number of objects to number of inhabitants, which usually will improve over the years. The quality of object properties can be determined by the number of attributes and can be tracked annually. The entire concept is generic and therefore additional parameters can be added. Due to the intrinsic approach, only relative quality values can be determined, which may only be used as an indicator. An absolute statement about the data quality is still only possible with superior reference data.

Future research will include the analysis of the temporal change of OSM relations and topology. Finally, the attributive analysis could be extended to compare the object properties with those that are recommended by the OSM community.

ACKNOWLEDGEMENTS

We thank Disy Informationssysteme GmbH, Karlsruhe, Germany for their support and excellent cooperation.

REFERENCES

Antoniou, V., Skopeliti, A., 2015. Measures and of VGI quality: An overview. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*., 2, 345–351. doi:10.5194/isprsannals-II-3-W5-345-2015.

Brovelli, M.A., Minghini, M., Molinari, M., Mooney, P., 2017. Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets. *Transactions in GIS*, 21(2), 191–206.

Brovelli, M.A., Minghini, M., Molinari, M.E., Zamboni, G., 2016. Positional Accuracy Assessment of the Openstreetmap Buildings Layer through Automatic Homologous Pairs Detection: the Method and a Case Study. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, XLI-B2, 615-620. doi:10.5194/isprsarchives-XLI-B2-615-2016.

Brovelli, M.A., Zamboni, G., 2018. A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints, *ISPRS Int. J. Geo-Inf*., 7, 289. doi:10.3390/ijgi7080289.

Corcoran, P., Mooney, P., Bertolotto, M., 2013. Analysing the growth of OpenStreetMap networks. *Spatial Statistics*, 3, 21–32. doi.org/10.1016/j.spasta.2013.01.002.

Disy Informationssysteme GmbH Cadenza, 2019. disy.net/de/produkte/cadenza (15 January 2019)

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.*, 28, 700–719. doi.org/10.1080/13658816.2013.867495.

Fonte, C.C., Antoniou, V., Bastin, L., Bayas, L., See, L., Vatseva, R., 2017. Assessing VGI data quality. Mapping and the Citizen Sensor, Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.M., Fonte, C.C., Antoniou, V., Eds. 137–164. doi.org/10.5334/bbf.g.

Goodchild, M.F., 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal*, Volume 69, Issue 4, 211–221. doi.org/10.1007/s10708-007-9111-y.

Goodchild, M.F., Li, L., 2012, Assuring the quality of volunteered geographic information, *Spatial Statistics*, 1, 110-120. doi.org/10.1016/j.spasta.2012.03.002.

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. doi.org/10.1068/b35097.

Haklay, M., Basiouka, S., Antoniou, V., Ather, A., 2010. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. *Cartographic Journal*, 47(4), 315–322. doi.org/10.1179/000870410X12911304958827.

Haklay, M., Weber, P., 2008. OpenStreetMap: User-Generated Street Maps. IEEE Pervasive Computing, 7(4), 12-18. doi.org/10.1109/MPRV.2008.80.

Hecht, R., Kunze, C., Hahmann, S., 2013. Measuring completeness of building footprints in OpenStreetMap over space and time. *ISPRS Int. J. Geo-Inf.*, 2(4), 1066–1091. doi.org/10.3390/ijgi2041066.

Koukoletsos, T., Haklay, M., Ellul, C., 2011. An automated method to assess Data Completeness and Positional Accuracy of OpenStreetMap. Proceedings of the 11th International Conference on GeoComputation, 236–241.

Ludwig, I., Voss, A., Krause-Traudes, M., 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. Geertman S., Reinhardt W., Toppen F. (eds) Advancing Geoinformation Science for a Changing World. Lecture Notes in Geoinformation and Cartography, 65-84. doi.org/10.1007/978-3-642-19789-5_4.

Mohammadi, N., Malek, M., 2015. Artificial intelligence-based solution to estimate the spatial accuracy of volunteered geographic data, *J. Spat. Sci.*, 60(1), 119–135. doi.org/10.1080/14498596.2014.927337.

Mooney, P., Minghini, M., 2017. A Review of OpenStreetMap Data. Foody, G, See, L, Fritz, S, Mooney, P, Olteanu-Raimond, A-M, Fonte, C and Antoniou, V. (eds.) Mapping and the Citizen Sensor, 37–59. doi.org/10.5334/bbf.c. License: CC-BY 4.0.

PostGIS Development Team, 2019. postgis.net (15 January 2019)

PostgresSQL Development Team, 2019. postgresql.org (15 January 2019)

Senaratne, H., Mobasheri, A., Loai Ali, A., Capineri, C., Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.*, 31(1), 139–167. doi.org/10.1080/13658816.2016.1189556.

Vandecasteele, A., Devillers, R., 2015. Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap, OpenStreetMap in GIScience, Lecture Notes in Geoinformation and Cartography, 59-80. doi.org/10.1007/978-3-319-14280-7_4

Zielstra, D., Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. Proceedings of the 13th AGILE International Conference on Geographic Information Science, 15p.