AGILE: GIScience Series, 2, 20, 2021. https://doi.org/10.5194/agile-giss-2-20-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

Assessing shop completeness in OpenStreetMap for two federal states in Germany

Josephine Brückner[a](corresponding author),Moritz Schott[a], Alexander Zipf[a,b] and Sven Lautenbach[a,b]

ww400@uni-heidelberg.de, moritz.schott@uni-heidelberg.de, zipf@uni-heidelberg.de, sven.lautenbach@heigit.org

[a]Institute of Geography, GIScience, Heidelberg University, 69120 Heidelberg, Germany
[b]HeiGIT at Heidelberg University, 69120 Heidelberg, Germany

Correspondence: Josephine Brückner (ww400@uni-heidelberg.de)

Abstract. The completeness of the number of OpenStreetMap (OSM) retail stores was estimated for two federal states of Germany at district level. An intrinsic measurement was applied that fits saturation models on the cumulative curve of the number of OSM retail stores over time. Even though the mean completeness of retail stores was estimated high in both states, the values within the states varied between 42 % and 100 %. The question therefore arises in which areas retail stores are well represented in OSM and whether economically weaker regions are possibly also digitally disadvantaged on the map. We investigated the influence of the urban-rural gradient as well as the influence of socioeconomic factors (gross domestic product, the unemployment rate, the proportion of academics) on the estimated completeness by means of a generalized linear model. Our results indicate that average big cities with low unemployment rate are better mapped with respect to retail stores.

Keywords. VGI, OpenStreetMap commpleteness, intrinsic quality analysis, limited growth functions

1 Introduction

The ubiquity of smartphones has lead to a continuous availability of geodata. In day to day life, especially at less familiar locations, shops and restaurants are some of the most frequently searched points of interest (POI). Having an up to date and complete collection of these POIs is of great interest for the potential customers. Shop owners equally have great interest in being well represented in these POI collections for advertisement and visibility reasons. While big players such as Apple, Google and Microsoft dominate the market, the Volunteered Geographic Information project OpenStreetMap (OSM) provides an established non-commercial crowd sourced alternative. OSM contains an enormous amount of various geodata, that are continuous edited by the great number of more than 7.5 million volunteers (state of March 2021, OpenStreetMap contributors (2021)). The open nature of the OSM project additionally provides the potential to cast niches for specialized services such as for shops without packaging (https://cartovrac.fr), cigarette vending machines (http://www.ubahnverleih.de/osm/zigaretten) or farm shops (https://farmshops.eu). According to https://taginfo.openstreetmap.org/keys/shop 4.5 million objects with the key shop=* existed in OSM at the end of March 2021. Especially in Germany where the active OSM community has created an extremely rich data set that is well known and used, OSM would be suitable as a flexible platform for online-shop-searching and navigation apps. The advantage of OSM over commercial providers of geodata is the free availability under the Open Data Commons Open Database License that allows an unrestricted use in many commercial or non-commercial applications. A persistent question for the usability of OSM is the quality of its data base. OSM roads are often mapped first and are now considered almost complete in many regions (Zielstra and Zipf, 2010). Due to the high growth rates of OSM in recent years, a high completeness can also be assumed for other objects such as buildings or stores.

Even though extensive research has been carried out on the OSM data quality, no single study exist - to the best of our knowledge - on the completeness of OSM retail stores. Such an investigation is not only necessary to clarify for which areas OSM data is suitable for shop finding platforms, but also to clarify the usability of OSM for researchers to analyze spatio-temporal patterns of the stationary retail sector. In addition, knowledge on influencing factors on shop-completeness can be used to predict, which parts of the physical world are digitally mapped or digital lacking and to counteract a digital disadvantage of retailers in particular regions, that may even already be economically weaker. Moreover, the socio-economic system of retailing is an interesting study field in OSM, because it can be assumed, that - in contrast to roads - local knowledge is necessary to locate, tag and add specific information (e.g. opening hours) to these locations.

For these reasons, this study examined the completeness of retail stores, a main quality criterion for online-searching-platforms, using the case study of two economically different states in Germany, Baden-Württemberg and Saxony. We further investigated how the urban-rural gradient and socio-economic factors (gross domestic product, unemployment rate and proportion of academics) were associated with completeness of OSM retail stores.

2 Methods and data

OSM quality analyses can be categorised into extrinsic or intrinsic approaches. Extrinsic approaches compare OSM with an external data set of presumably higher quality (see for example Zielstra and Zipf (2010); Arsanjani et al. (2015); Hecht et al. (2013); Törnros et al. (2015); Fan et al. (2014); Neis et al. (2012). A major drawback of extrinsic approaches is the necessity for a compatible external data set, which may not always be available. For example for shops, official statistics, if available at all, may only be in reference to a certain level of administration and a specific definition of 'retail', that cannot be directly transferred to the definition of OSM. Therefore, we assess the fitness for purpose of an intrinsic completeness estimation using only OSM data itself (see Ballatore and Zipf (2015); Degrossi et al. (2017); Barron et al. (2014); Barrington-Leigh and Millard-Ball (2017) for some examples on intrinsic OSM data quality assessment).

The underlying idea of the intrinsic completeness analysis is that the added number of OSM objects of a specific feature class per time period decreases as the number of mapped objects converges against the (unkown) true number of objects. The cumulative number of OSM objects would then saturate. Given a sufficient mapping activity it is possible to estimate the saturation level using a suitable function in the context of a non-linear regression approach. Baden-Württemberg and Saxony were particularly suitable for this intrinsic investigation, as no bulk data imports of retail stores into the OSM database have been recorded so far.

Table 1. Regional data and economic information of Baden-Württemberg and Saxony: Number of administrative districts, the GDP in 1000 euros per employed person (for 2016), unemployment rate as percentage of unemployed in the civilian labor force (for 2017), proportion of academics as employees to social security contributions at the place of residence with an academic qualification per 100 inhabitants of working age, total area and population density (2019) of the respective federal state.

	Baden-Württemberg	Saxony
Rural districts	9	8
Urban districts	27	2
Independent cities	8	3
GDP [1000 € per employed person]	73.6	57.2
Unemployment rate [%]	3.5	6.8
Academics per 100 inhabitants	9.3	10.2
Area [km^2]	35673	18449
Inhabitants per km^2	311	221

2.1 Experimental setup

Retail stores are defined as places, where goods or services are sold to the final consumer (Bankim and Vaja (2015)). Our analysis was restricted to stationary retail stores, that were tagged as "shop" or "amenity" and that were listed in the OSM Wiki (OSM Wiki (26.02.2021)). We included all key-value combinations that could not be clearly excluded from retail trade (for a detailed list of the used tags see the linked source code in section 2.2). These included combinations of retail trade and direct marketing - such as farm shops - or services - such as car repair shops.

The research area is characterized by contrasts, both economically and in the degree of urbanisation. Baden-Württemberg, located in South Germany, has been one of the economically strongest federal states of Germany. Saxony in East Germany has been economically weaker. Altogether, both states contain 57 administrative districts of rural as well as urban character (table 1).

We fitted various limited growth curves to the OSM history for each administrative district and estimated the completeness level via their saturation parameter. The curves used originate from two families. On one hand, the family of sigmoid curves seems adequate for a three-phase mapping process as also described

by Barrington-Leigh and Millard-Ball (2017). On the other hand, curves of the non-logistic growth curves family tend to represent a mapping process without the initial phase of slow growth. In this analysis, we used the following functions: the three and four parameter logistic function (equation 1 and 2), that are assigned to the sigmoid curve family as well as the rectangular hyperbola (equation 3) and the asymptotic function (4) of the non-logistic growth curves family .

$$y = \frac{Asym}{1 + e^{\frac{t_{mid} - t}{scale}}} \tag{1}$$

$$y = Asym_{low} + \frac{Asym - Asym_{low}}{1 + e^{\frac{t_{mid} - t}{scale}}} \tag{2}$$

$$y = \frac{Asym * t}{t_{1/2} + t} \tag{3}$$

$$y = Asym + (y_0 - Asym) * e^{-e^{lrc} * t} \tag{4}$$

where:

$Asmp$	= a numeric parameter representing the saturation to which the curve converges
$Asmp_{low}$	= lower Asymptote
t	= time at which half the saturation level is attained
t_{mid}	= mid point of the logistic curve
$scale$	= the steepness of the logistic curve
$t_{1/2}$	= time, 50 % saturation
y_0	= parameter, that specifies the value of y (here count of OSM contributions) at the begin of the period
rc	= 'rate constant', parameter that determines the spread of the curve with time
lrc	= log of the 'rate constant'

The reliability of estimation the number of retail stores in a district depends on the development of OSM contributions in several ways. First and foremost, the data history was checked for a decline in growth at all, that is a fundamental criterion to estimate a saturation level as a proxy for the number of retail stores.

Fitted models were filtered for unrealistic fits where the asymptote was estimated to be lower than the current number of OSM retail stores. To account also for the uncertainty of the models, we accepted fits whose asymptote was at most 2% lower than the actual latest amount. We chose the best fitting functional form of all accepted curves for each administrative district based on two criteria: i) the relative residual standard error and ii) the relative deviation of the slope between the historic development and the fitted curve during the last two years of the analysis period. If both criteria contradicted each other, the selection was made based on visual assessment. The completeness level was estimated as the quotient of the current number of retail stores and the asymptote of the estimated saturation curve.

Finally, we investigated the influence of factors on the completeness level based on a generalized linear model (GLM) with a negative binomial distribution and a log-link. We used the estimated asymptote (the estimated number of retail stores) as the response variable and used the logarithm of the number of shops as an offset. This standard procedure allowed to model the relation between estimated and observed counts without mixing up distributional assumptions. Since completeness was inversely proportional to the asymptote, negative coefficients of the regression indicated a higher completeness level, whereas positive coefficients meant a lower completeness level.

We examined the influence of the urban-rural gradient by the district type of the administrative units defined by Bundesinstitut für Bau-, Stadt- und Raumforschung (2019). Type 1 are independent cities with at least 100,000 inhabitants. Type 2 are urban districts with a medium population density of at least 150 inhabitants/km^2. Type 3 are rural districts with a low population density less than 150 inhabitants/km^2. We tested in addition the effects of three socio-economic factors as predictors: the gross domestic product (GDP) in 1000 euros per person in employment (for 2016), the unemployment rate as the percentage of unemployed in the civilian labor force (for 2017) and the proportion of academics as employees to social security contributions at the place of residence with an academic qualification per 100 inhabitants of working age (for 2017). We hypothesized that the completeness level would increase along the rural-urban gradient and with higher GDP, lower unemployment rate and higher share of academic employees. These hypotheses are based on Neis et al. (2013) who found a positive link between urban and OSM activity as well as GDP and OSM activity.

2.2 Software and data availability

Monthly counts of shops were extracted from OSM using the ohsome API at https://api.ohsome.org to query the OpenStreetMap History Database (OSHDB) (Raifer et al., 2019) for the time frame 01.01.2008 until 01.01.2020.

The data for the administrative units were taken from German federal agency for geodesy and cartography (Bundesamt für Kartographie und Geodäsie, http://gdz.bkg.bund.de) (reference 01.01.2020).

The data for the influencing factors GDP, unemployment rate, proportion of academics as well as the classification into district types originate from the federal institute for Research on Building, Urban affairs and Spatial Development (Bundesamt für Bau-, Stadt- und Raumforschung, INKAR, https://www.inkar.de/).

Further statistical information (population density, area) were queried via the regional database of Germany from the statistical offices of the federation and the federate states (Statistische Ämter des Bundes und der Länder, Regionaldatenbank Deutschland https://www.regionalstatistik.de/genesis/online/, data licence Germany – attribution – Version 2.0 www.govdata.de/dl-de/by-2-0).

The analysis was performed in R (R Core Team, 2021), using the packages sf (Pebesma, 2018), RCurl (Temple Lang, 2021), geojsonio (Chamberlain and Teucher, 2021), tidyverse (Wickham et al., 2019), ggplot2 (Wickham, 2016) and ggpubr (Kassambara, 2020).

All source code, preprocessed data and results can be found at https://github.com/GIScience/shop-completeness.

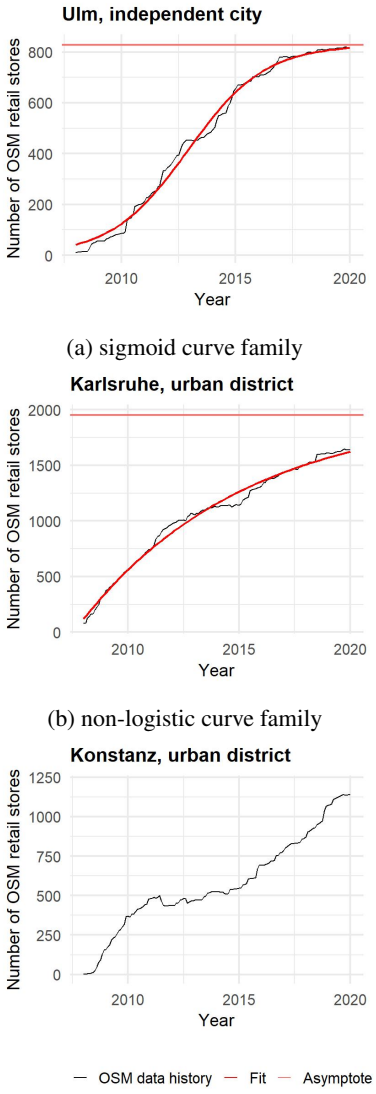
3 Results

Well fitting saturation models were generated for 44 of the 57 districts in both federal states. For five districts the data history showed a steady high increase in the number of OSM retail stores and did not indicate any slow down in growth rate while eight regions produced low quality saturation models (figure 1c) due to complex temporal pattern. These issues occurred independent of influencing factors such as population density due to non continuous mapping activities and the respective regions had to be ignored for the GLM. The non-linear fit of a sigmoid curve produced the best results for 28 districts (figure 1a), while a non-logistic curve showed the best fit for 16 districts (figure 1b). Ten of the thirteen districts for which no saturation level could be estimated were visually categorized as relatively far from saturation (table 2).

The mean completeness level of Baden-Württemberg was approximately 88 %, slightly higher than the mean value of Saxony of about 82 %. The completeness ranged from 42 % to almost 100 %. Even though the results showed heterogeneity in the completeness, the majority of 38 districts achieved at least 80 %.

Completeness was significantly higher in the independent cities than in the urban and rural districts (table 3). The completeness level of the data decreased significantly with a higher unemployment rate. In total, the GLM explained 18 % of the deviance in the data.

Districts, for which no suitable saturation model could be estimated, were represented in all district types.



(a) sigmoid curve family

(b) non-logistic curve family

(c) no suitable fit

Figure 1. Non-linear regression of varied curves to data history of OSM retail stores

Table 2. Number of districts for which a sigmoid fit, a non-logistic fit, and none of the fits were considered reasonable for the completeness estimation as well as the percent completeness as mean for the respective federal state.

	Baden-Württemberg	Saxony
Sigmoid curve family	21	6
Non-logistic curve family	14	3
No suitable fit	9	4
Avg. Completeness [%]	88	82

However, districts, that were clearly not saturated, were mostly of type rural or urban district and had a GDP, a proportion of academic as well as an unemployment rate slightly below the average of each fed-

eral state. Fitting a binomial GLM with a log-link (a logistic regression) did not reveal any significant relationship between the four predictors and successful fitting of a saturation curve. However, if compared visually for each district category unemployment was higher on average for those districts where the saturation level could not be estimated reliably.

Table 3. Coefficient estimates, standard errors and p-values of the GLM for the 44 districts with a reasonable fit of the asymptote. Coefficients and standard errors are provided at the link scale. The response was the asymptote - for a given observed count (included as an offset in the model) completeness goes up if the asymptote is lower. Negative coefficients therefore indicate a positive effect on completeness and vice versa. Rural districts were used as the reference level - the coefficient therefore represents the intercept. The θ parameter of the binomial distribution was estimated as 54.5 with a standard error of 12.3.

Factor	Estimate	std. error	p-value
Rural districts	0.026	0.076	0.733
Urban districts	0.009	0.050	0.859
Independent cities	-0.140	0.066	0.033
Unemployment rate [%]	0.037	0.014	0.01

4 Discussion

In comparison to previous studies e.g. on the completeness of OSM buildings in Germany ((Törnros et al. (2015))), the estimated completeness of retail stores in OSM was relatively high. It is in general more problematic to estimate the saturation level for incomplete districts than for complete districts. With this in mind, the mean completeness values tended to be overestimated since districts with lower saturation are not considered. Saturation may also occur due to a decrease in mapping activity resulting in a false intrinsic estimate of completeness. In our analysis, a sufficient number of active users was present in all districts which provides reasonable support for the assumption that saturation did not occur to a lack of user activity. Events, such as bulk data imports or mapping parties affect the form of the data history and require fitting functions of respective forms. In our analysis, the data history of only 8 of the 57 districts showed one of this deviating forms, due to which no suitable fit function was found. However, it might be suitable to included additional function types such as multiple sigmoid forms as well as step forms in other regions and for other OSM feature class, whose data history reflects such events.

The higher data completeness found for districts with a low unemployment rate was consistent with our hypothesis. The higher completeness level of independent cities - the district type with the highest population density - was similar to those reported by studies on the completeness of other OSM feature classes (Zielstra and Zipf, 2010; Mashhadi et al., 2015; Wang et al., 2020). However, the completeness for the category urban districts could not be distinguished from the completeness of rural districts. So the hypothesis, that completeness increases with the rural-urban gradient was not supported by our data.

The relatively small sample size makes the results sensitive to outliers. Two rural districts could be identified as influential by means of the usual regression diagnostics: “Nordsachsen” with the lowest of all estimated completeness levels (42%) with a high leverage and a high cook’s distance and "Görlitz" with a high leverage. If both districts would be omitted simultaneously, regression coefficients estimates would remain the same with slightly higher standard errors due to the reduced sample size. If only "Görlitz" would be ommited, the regression coefficients would be of similar magnitude and sign but if "Nordsachsen" would be omitted all coefficient estimates would render insignificant. Since we had to exclude districts with low completeness since we could not reliably estimate the saturation level, and those districts show a tendency for higher unemployment rate and to belong to rural or urban district types our results might be to conservative. To prove and clarify the effects of the factors on the completeness, further studies including a larger amount of data are necessary.

The major challenge of the intrinsic completeness estimation is the selection of the best fit among multiple options. Additionally, using different models to estimate the completeness makes comparison of results for different districts more challenging. In our analysis, curves of the non-logistic curve family tended to estimate a higher asymptote - and therefore a lower completeness level - than curves of the sigmoid family.

However, the diversity of OSM contribution histories seems to not allow a "one fits all" approach. We have started to extend our research in this direction to overcome this limitation.

5 Conclusion and outlook

The presented approach allowed a reliable completeness estimation and comparison of OSM data between regions with individual contribution histories. This study was applied to the use for case of retail stores but the approach may be transferred to e.g. roads or land use data by substituting the store count with road network length or land use area.

The estimated completeness level of more than 86 % on average indicated the high potential of OSM to

be used as a database for platforms offering online-shop-searching in densely mapped countries such as Germany. For a real world application, further quality elements like positional accuracy and moreover the completeness of the various attributes, such as opening hours, contact information as well as accessibility's, would also need to be investigated. Future research should further study how the completeness differ in the various types of retail stores, such as supermarkets or clothing stores, to identify lacks regarding store types.

The results of the GLM suggest that especially big cities with low unemployment rates can be expected to be of higher completeness of retail stores and therefor presumably fit for purpose. Furthermore, we expect of OSM to catch up in disadvantaged areas soon. This is due to the high estimated completeness level of retail stores compared to previous studies of other feature classes, that demonstrate the continued growth of OSM and the overall high activity of the OSM community.

References

Arsanjani, J. J., Mooney, P., Zipf, A., and Schauss, A.: Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets, in: OpenStreetMap in GIScience, pp. 37–58, Springer, 2015.

Ballatore, A. and Zipf, A.: A conceptual quality framework for Volunteered Geographic Information, in: International Conference on Spatial Information Theory, pp. 89–107, Springer, 2015.

Bankim and Vaja, R.: Retail Management, IJRAR- International Journal of Research and Analytical Reviews, 2, 22–28, 2015.

Barrington-Leigh, C. and Millard-Ball, A.: The world's user-generated road map is more than 80% complete, PLOS ONE, 12, e0180 698, 2017.

Barron, C., Neis, P., and Zipf, A.: A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis, Transactions in GIS, 18, 877–895, https://doi.org/10.1111/tgis.12073, 2014.

Bundesinstitut für Bau-, Stadt- und Raumforschung: INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung. Erläuterungen zu den Raumbezügen., https://www.inkar.de/documents/Erlaeuterungen%20Raumbezuege19.pdf, 2019.

Chamberlain, S. and Teucher, A.: geojsonio: Convert Data from and to 'GeoJSON' or 'TopoJSON', https://CRAN.R-project.org/package=geojsonio, r package version 0.9.4, 2021.

Degrossi, L. C., Albuquerque, J. P. d., Rocha, R. d. S., and Zipf, A.: A Framework of Quality Assessment Methods for Crowdsourced Geographic Information : a Systematic Literature Review, in: Proceedings of the 14th ISCRAM Conference, pp. 21–24, issue: May, 2017.

Fan, H., Zipf, A., Fu, Q., and Neis, P.: Quality assessment for building footprints data on OpenStreetMap, International Journal of Geographical Information Science, 28, 700–719, 2014.

Hecht, R., Kunze, C., and Hahmann, S.: Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time, ISPRS International Journal of Geo-Information, 2, 1066–1091, https://doi.org/10.3390/ijgi2041066, 2013.

Kassambara, A.: ggpubr: 'ggplot2' Based Publication Ready Plots, https://CRAN.R-project.org/package=ggpubr, r package version 0.4.0, 2020.

Mashhadi, A., Quattrone, G., and Capra, L.: The Impact of Society on Volunteered Geographic Information: The Case of OpenStreetMap, in: OpenStreetMap in GIScience, edited by Arsanjani, J. J., Zipf, A., Mooney, P., and Helbich, M., Lecture Notes in Geoinformation and Cartography, pp. 125–141, Springer, Cham, Heidelberg, New York, Dordrecht, London, 2015.

Neis, P., Zielstra, D., and Zipf, A.: The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011, Future Internet, 4, 1–21, 2012.

Neis, P., Zielstra, D., and Zipf, A.: Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions, Future Internet, 5, 282–300, https://doi.org/10.3390/fi5020282, 2013.

OSM Wiki: Key:amenity, https://wiki.openstreetmap.org/wiki/DE:Key:amenity, 26.02.2021.

Pebesma, E.: Simple Features for R: Standardized Support for Spatial Vector Data, The R Journal, 10, 439–446, https://doi.org/10.32614/RJ-2018-009, 2018.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2021.

Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.-B., and Zipf, A.: OSHDB: a framework for OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data, Open Geospatial Data, Software and Standards, 4, 1–12, https://doi.org/10.1186/s40965-019-0061-3, 2019.

Temple Lang, D.: RCurl: General Network (HTTP/FTP/...) Client Interface for R, https://CRAN.R-project.org/package=RCurl, r package version 1.98-1.3, 2021.

Törnros, T., Dorn, H., Hahmann, S., and Zipf, A.: Uncertainties of Completeness Measures in Openstreetmap - a Case Study for Buildings in a Medium-Sized German City, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, II-3/W5, 353–357, https://doi.org/10.5194/isprsannals-II-3-W5-353-2015, 2015.

Wang, S., Zhou, Q., and Tian, Y.: Understanding Completeness and Diversity Patterns of OSM-Based Land-Use and Land-Cover Dataset in China, ISPRS International Journal of Geo-Information, 9, 531, https://doi.org/10.3390/ijgi9090531, 2020.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, https://ggplot2.tidyverse.org, 2016.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani,

H.: Welcome to the tidyverse, Journal of Open Source Software, 4, 1686, https://doi.org/10.21105/joss.01686, 2019.

Zielstra, D. and Zipf, A.: Quantitative studies on the data quality of OpenStreetMap in Germany, in: Proceedings of GIScience, vol. 2010, 2010.