Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 653–658
DOI: 10.15439/2025F6262
ISSN 2300-5963 ACSIS, Vol. 43

Coverage of OpenStreetMap data at firm level. Assessing the suitability for social science representative firm samples and found

Christian Gerhards
0000-0002-8180-3248
Federal Institute for Vocational Education and Training
Friedrich-Ebert Allee 114-116,
53113 Bonn, Germany
Email: gerhards@bibb.de

***Abstract*— Establishment surveys are essential for gaining well-founded insights into company structures and developments, whereby representativeness ensures that the results can be reliably transferred to the population and enable bias-free conclusions. Company surveys in Germany are often based on the company file of the Federal Employment Agency (BA), which, however, only allows limited links to other data sources. The free online service OpenStreetMap (OSM) could serve as an alternative, publicly accessible data source. This article analyses the extent to which OSM is suitable for drawing a representative sample of companies. It is hypothesised that OSM covers a large proportion of businesses, especially in metropolitan areas, and that customer-oriented sectors are better covered than industrial or B2B businesses. To verify this, establishments are identified in OSM and compared with the BA establishment file, differentiated by region and sector. The results show a high but not complete coverage with regional differences in favour of metropolitan areas. There are limitations in terms of data quality and contactability of the establishments.**

***Index Terms*— OpenStreetMap (OSM), Company level data, Sectoral Coverage, Survey Methodology.**

I. Introduction: New approaches in Business Research: Computational Approaches with OpenStreetMap and Supplementary Digital Data Sources

DATA on companies is of central importance for social science research, as it provides insights into economic structures, labour markets, regional development and social dynamics. An establishment refers to a social workplace. Many analyses restrict the definition to establishments with at least one employee subject to social insurance contributions. This therefore excludes solo self-employed persons. Establishments in this sense are to be distinguished from companies, which may consist of several establishments.

Information on location, size, sector and company structures makes a significant contribution to understanding employment processes, innovation, education and social inequality. Both found data from sources such as OpenStreetMap (OSM) and data from systematic company surveys offer different, complementary perspectives: While found data can provide up-to-date and widely available information, targeted surveys enable a deeper analysis of specific company characteristics and contexts. The combination of both approaches significantly expands the potential for well-founded social science analyses and evidence-based policy advice.

The availability of large, freely accessible digital data sets such as OSM opens up new opportunities for interdisciplinary research in the social sciences, economics and humanities. As a first step, the data from OSM itself can be used to analyse companies in Germany, for example. There is also the potential to link additional data from company websites, job advertisements and company valuations. Finally, the company addresses available in OSM can be used to conduct new company surveys that are as representative as possible.

This article is positioned in the context of computational approaches that utilise digital methods for theoretical, methodological and applied questions. It looks at the extent to which OSM is a suitable data source for the representative mapping of businesses in Germany. On the one hand, this applies as a source for a complete survey (use of the OSM data itself) and as an applied approach as a basis for drawing representative establishment samples. In particular, the article examines the potential of OSM for the applied generation of representative firm samples and critically reflects on the quality, reliability and representativeness of such digital resources.

The article thus addresses the central challenges of linking heterogeneous data sources such as crowdsourcing data with regard to the methodological integration of quantitative and qualitative information (e.g. OSM data with big data analyses of job advertisements or company evaluations).

It will also discuss how computer-aided methods such as web mining can contribute to systematically accessing digital resources and critically reflecting on their use for social science, economic and humanities issues.

II. State Of Research: Quality of OpenStreetMap

Found data refers to data that was not collected specifically for research purposes but was originally generated for other purposes, while the opposite is primary data collected specifically as part of scientific studies. OpenStreetMap (OSM) can be regarded as "found data" in the context of firm data collection, as the information about firms is not systematically

obtained through structured surveys, but is collected through diverse, often uncoordinated contributions from different sources.

The mapping of businesses in OpenStreetMap (OSM) is based on a variety of data sources, which together determine the diversity and depth of the information available. A significant proportion of the entries come from volunteers who use their local knowledge to record points of interest (POIs) such as shops, service providers and offices directly on site and add them to the OSM database. These personal on-site surveys are complemented by the mapping of freely available aerial and satellite imagery, utilising high-resolution imagery from services such as Bing Maps to determine the exact location and structure of businesses. In addition, mass imports of open government and company data are carried out, which are integrated into the database in compliance with the OSM import rules. Such imports can include, for example, data from local authorities or other public bodies ([1]). Last but not least, organised mapping initiatives, for example by companies or non-governmental organisations, contribute to the expansion and updating of operational data in OSM through targeted collection campaigns.

Each of these sources has specific strengths and potential biases that must be taken into account when using the data for scientific analyses. The reliability of OSM in capturing and mapping firms in Germany has been the subject of numerous scientific studies (e.g. [2], [3]). These studies show that the data quality of OSM strongly depends on the geographical location: In urban areas, the completeness and accuracy of the data is generally higher than in rural regions. For example, a study by Zielstra and Zipf ([4]) found that in German cities, the coverage of OSM is comparable to proprietary geodata, while in rural areas there may be differences.

The completeness of the businesses recorded in OSM also varies. While larger retail chains and well-known establishments are often well documented, smaller or less publicised businesses are often missing. One of the reasons for this is that the collection of such data is heavily dependent on the local community and there is no systematic survey. A study by Neis et al. ([5]) emphasises the importance of active user participation for the data quality of OSM.

Various tools and methods have been developed to assess the data quality of OSM. For example, the Heidelberg Institute for Geoinformation Technology (HeiGIT) launched the "ohsome" platform in 2018, which enables a structured analysis of OSM data quality ([6]). This platform takes into account quality characteristics such as completeness, thematic accuracy and logical consistency in accordance with ISO 19157 standards.

III. Theoretical Background And Hypotheses

OpenStreetMap (OSM) is a collaborative project in which voluntary contributors collect and record geodata and make it available to the general public. These contributors - also known as mappers - are individuals or organisations who are active with different interests: from improving local map content to scientific purposes and commercial applications. While some specifically want to map their own environment in detail, others participate for altruistic or public interest motives. At the same time, there are also regions where few or no contributors are active, which can lead to uneven data coverage.

OSM is classic Volunteered Geographic Information (VGI): Data collection depends on the voluntary commitment of the population, but also on institutional imports and technical constraints. Goodchild's concept of "Citizens as Sensors" ([7]) describes precisely this area of tension between supply (data and contribution potential) and demand (benefits for user groups).

Agency districts of the Federal Employment Agency are regional areas of responsibility of the employment agencies that divide the German labour market into administratively delimited units. They are relevant for company analyses in Germany as they provide a geographical structure that can be used to systematically record company data, evaluate it in a regionally differentiated manner and link it to labour market indicators. The theoretical mechanisms originate from VGI research (participation and digital divide theories) and geodata governance. They provide hypotheses to systematically explain why some agency districts of the Federal Employment Agency in Germany (AAB) are significantly better covered with company POI in OSM than others. A combination of individual factors (digital competence, community size) and structural contexts (urbanity, open government data, economic profile) is likely to explain most of the variance.

The quality and completeness of company data in OSM varies considerably between different AAB. These differences can be explained by socio-demographic and economic structural characteristics of the regions. The following section presents a theory-based derivation of five hypotheses that address these differences. Empirical studies show that data density correlates strongly with the number of active mappers and that social and economic differences ("digital divide") limit regional participation. On this basis, the following hypotheses can be derived for the AAB:

Hypothesis 1: The more people live in an employment agency district, the better the coverage of businesses in OpenStreetMap.

The number of OSM contributions correlates positively with the population of a region. In regions with a higher population, there are more potential contributors, which leads to better data coverage. Herfort [3] show that the completeness of OSM data is higher in densely populated areas.

Hypothesis 2: The more businesses there are in the primary sector in an agency district, the poorer the coverage.

The primary sector includes businesses involved in the extraction of raw materials, the secondary sector in processing and the tertiary sector in services. Businesses in the primary sector, such as agriculture and forestry, are often located in rural areas, which tend to be mapped in less detail in OSM. This may be due to the fact that businesses with little customer contact are less likely to be the focus of mapping initiatives.

Hypothesis 3: The more businesses there are in the secondary sector in an agency district, the poorer the coverage.

Although the secondary sector is often present in urban areas, certain industrial establishments may be less well mapped in OSM due to access restrictions or low visibility to the public.

Hypothesis 4: The higher the population density in the agency district, the better the coverage.

High population density increases the probability that a location is known to mappers. This can facilitate mapping in OSM.

Hypothesis 5: The lower the average age in the agency district, the better the coverage.

Younger population groups tend to be more tech-savvy and more willing to participate in digital platforms such as OSM. A higher level of OSM contributor activity can therefore be expected in regions with a lower average age, which leads to better data coverage.

IV. Data Basis, Operationalisation and Model

A. Data basis

Empirically, a comparison has now been made of how many companies are found per AAB in the official statistics on the one hand and in OSM on the other. The Federal Employment Agency (BA) determines the number of companies with employees subject to social security contributions on the basis of employers' social security notifications. This information forms the BA's so-called company file. The notifications are required by law and are used to compile government employment statistics. It is therefore a complete survey with a high degree of reliability.

To analyse the data from OSM, it was downloaded as a file as of 2025-05-25T20:21:00Z (https://download.geofabrik.de/europe/germany.html). The query was programmed using Python by integrating libraries for processing OSM data (osmium), tabular data (pandas) and geographical data (geopandas).

In OSM, all companies with employees subject to social security contributions were also researched as far as possible. After a previous random search, all so-called nodes from OSM in Germany are selected for which one of the following tags is filled with content: 'shop', 'amenity' (maintenance, education, …), 'office', 'craft', 'industrial'. In OSM, a "node" refers to a single point, defined by its geographical coordinates (latitude and longitude).

The companies found with geo-coordinates were then assigned to the AAB. A so-called shape file from the Federal Employment Agency was used for this purpose. The BA file "Deutschland_Agenturbezirke.shp" contains the geographical boundaries and associated information of the 150 AAB in Germany. The file can be downloaded from the BA statistics website in SHP format. (https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Regionale-Gliederungen/BA-Gebietsstruktur-Nav.html) It was then compared which companies are located in which AAB. The number of establishments per AAB was determined and stored in a file..

B. Operationalisation

- To empirically examine the question of the extent to which OSM is suitable for drawing a representative sample of businesses, the registration rate of businesses in OSM is defined as the dependent variable. On the one hand, this rate results from the nodes or points of interest (POI) registered in OSM for businesses (see above). This number is set in relation to the number of official business census of the Federal Employment Agency (BA) statistics. In OpenStreetMap (OSM), points of interest (POIs) are specific nodes to which specific attributes are assigned in order to characterise locations of particular importance.
- Various predictors are used to explain this coverage rate. These include regional structural characteristics from the BA's regional data, including population figures and economic indicators. For the analyses, characteristics of the individual AAB were taken from tables provided by the BA on its website. These include the proportion of employees in the primary, secondary and tertiary sectors, population density, the average age of the population and the number of employees per company. Table 1 provides an overview of the key figures per AAB used in the following. Most employees work in the tertiary sector. The majority of people are over 50 years old. On average (self-calculated index from the age groups), people are 44 years old. The population density varies greatly between the AAB, ranging from 47 to 7339.5 inhabitants/km2. On average, 16 people are employed per firm.

C. Model

A linear regression model (OLS) is used for the analysis in order to quantify the influence of these predictors on the recording rate (log. number of firms according to osm per aab in Germany) in order to identify systematic differences in the recording quality between the AAB. M1 includes Log. Number of employees in official BA-statistics (1), Number of employees in primary sector, Number of employees in secondary sector and Pop. Density (2). M2 additionally includes the Index of Average age per AAB. M3 adds an interaction term between (1) and (2).

V. Results

A. Descriptive results

In order to assess the suitability of OpenStreetMap (OSM) as a data source for recording businesses, a quantitative comparison of the businesses recorded in OSM with the official register data of the Federal Employment Agency (BA) is first carried out, supplemented by an analysis of regional differences and a modelling of relevant influencing factors.

TABLE 1: KEY FIGURES FOR THE INDIVIDUAL AGENCY DISTRICTS IN GERMANY
SOURCE: OWN CREATION. BASED ON: BA STATISTICS ON THE INDIVIDUAL AGENCY DISTRICTS. N = 147 AGENCY DISTRICTS (BERLIN IS MISSING)

Variable	Mean	Std. dev.	Min	Max
% Companies prim. sector	1,0	0,8	0,0	3,6
% Secondary sector Sector	29,8	8,6	9,5	51,4
% Companies tert. Sector	69,2	8,6	48,4	90,4
Persons under 25 years of age	136361	65758	44806	476924
Persons aged 25 to under 50	177142	100968	55786	716246
Persons 50 years and older	250959	99757	99536	716990
Index Average age	43,7	1,1	40,6	46,4
Population density (inhabitants/km2)	588,6	936,5	47,2	7338,5
Employees per company	15,8	2,7	10,8	26,6

TABLE 2: COMPARISON OF THE NUMBER OF REGISTERED COMPANIES (PER DISTRICT) ACCORDING TO BA AND OSM IN GERMANY. SOURCE: OWN COMPILATION: OFFICIAL BA STATISTICS (CUT-OFF DATE 31/12/2023) AND OSM QUERY (CUT-OFF DATE 25/05/2025). N = 147 AGENCY DISTRICTS.

Variable	Mean	Std. dev.	Min	Max	Total
Number of companies according to official (BA) statistics	14443,94	10049,79	5350	99550	2.137.703
Number of businesses according to OSM	5503,13	2745,02	2143	20471	824.738

In total, approximately 39% of all firms in Germany can be identified in OpenStreetMap (OSM), assuming that the matched entries refer to the same entities. (see Table 2). This shows that, as expected, less establishments are found than are recorded in the official statistics.

However, it must be noted that this comparison is conducted at the macro level. It remains unclear whether there are false positives or false negatives in the assignments of firms at the employment agency district level. For example, establishments may be recorded that have no employees subject to social security contributions.

To assess the potential of using OSM data for social science research, a more detailed examination of the identified firms is undertaken. For follow-up surveys, the availability of identifying information – such as postal address, email address, or telephone number – is particularly important. In addition, company websites play a key role in enabling further data enrichment. The following table 3 provides an overview of the proportion of firms for which these attributes could be retrieved.

TABLE 3: % OF FIRMS IN OSM WITH. N = 824.738 FIRMS TOTAL = NUMBER OF REGISTERED COMPANIES IN GERMANY. SOURCE: OWN COMPILATION: OSM QUERY (CUT-OFF DATE 25/05/2025).

	% firms with…
Name	95
Address: city	56
Address: postcode	57
Address: street	59
Address: house number	59
Phone	27
Email	13
Fax	4
Website	29

The analysis of business entries in OpenStreetMap (OSM) reveals varying levels of detail across different attributes.

Nearly all firms (95%) are identified by name, indicating a high degree of basic recognition in the dataset. Address-related information is moderately well-represented, with 56–59% of firms including details such as city, postcode, street, and house number. In contrast, contact information is less frequently available: only 27% of firms list a phone number, 13% provide an email address, and a mere 4% include a fax number. Websites are linked in 29% of cases. These findings suggest that while core identification and location data are commonly present in OSM business entries, contact details are comparatively sparse.

In the next step, the number of firms recorded in OSM will be compared, at the level of employment agency districts, with the number of firms that should be present according to official statistics. If one looks at the correlation between the values from the official statistics and the data from OSM, you can see a clear linear correlation across the AAB (see Figure 1). The dispersion is moderate, which indicates a good mapping.

A graphical analysis shows that regions in west Germany, central Germany and north-east Germany are more likely to ikely have a higher coverage rate of firms in OSM. No data is available for Berlin due to a different data structure (see Figure 2). Some of these regions are particularly rural.

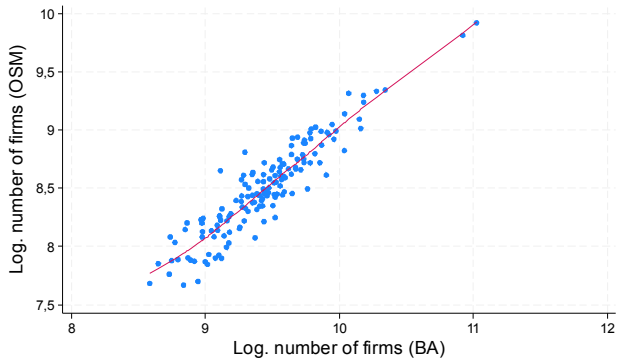

Fig. 1: Log. Number of firms per AAB according to OSM and BA in Germany, Source: Own compilation. Variables logarithmised due to skewness. n=147 agency districts.

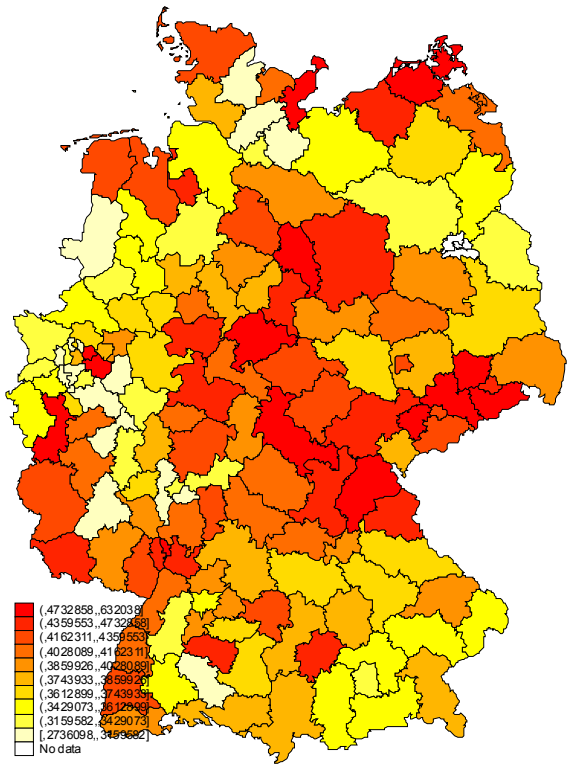

Fig. 2: higher (red) and lower coverage rate of firms (white) in OSM compared to BA in the agency districts in Germany.
Source: own compilation. The totals of the figures per agency district are compared. Berlin: no data.

B. Multivariate results

A regression model is calculated for further analysis. The dependent variable for each AAB is the logarithmised number of firms according to the OSM. The number of firms according to the BA is controlled for (Table 4).

Overall, the results show that both demographic and economic-structural characteristics have no significant influence on the completeness of voluntarily created geodata. The analysis shows no regional differences in the recording of businesses in OSM. Firstly (H1), a higher population size in the AAB does not improve the capture rate. Secondly (H2), a high proportion of firms in the primary sector, such as agriculture and forestry, does not correlate with the capture quality. Thirdly (H3), there is no correlation for the secondary sector. Fourthly (H4), a higher population density does not indicate a poorer coverage rate. Finally, there is no significant correlation between the average age of the population in the AAB and a better coverage rate.

VI. CONCLUSION

This study shows that OpenStreetMap (OSM) offers moderate coverage of businesses in Germany (about 39%). The total number of establishments recorded in OSM is not comparable to the official establishment file of the Federal Employment Agency (BA), although not necessarily the same establishments are recorded.

It was all the more important to examine whether the coverage is randomly distributed or follows a systematic pattern. The analysis shows that there are no significant regional and sectoral differences: Larger populations and lower average ages seems not to be associated with better coverage quality, while establishments in the primary and secondary sectors and regions with high population density seems not to tend to be better represented. These results illustrate both the potential and limitations of OSM for firm-related research.

Despite the limitations observed, OSM offers a fundamental advantage over traditional official registers: Because the data from the Federal Employment Agency's establishment file (BA establishment file) is based on mandatory social security data, it is generally subject to the requirements of the GDPR. Data on companies from OSM generally has no personal reference and can therefore be linked more easily with other digital resources. In addition, OSM opens up new opportunities for analysing "found data", for example by linking it to company websites, job advertisements or company valuations. These approaches enable deeper insights into company structures and dynamics that are difficult to capture with traditional surveys or register data.

For future research, an in-depth analysis of the validity and accessibility of the businesses recorded in OSM is recommended, for example by enriching the datasets with business email addresses, business websites, job adverts or telephone numbers - if these are missing. Companies could then be contacted for company surveys using contact information. Enrichment with qualitative information and text data can provide a new in-depth insight into the behaviour of companies. Improved procedures could help to systematically identify false-positive entries or the allocation of establishments as such and further increase data quality. In particular, approaches for the automated linking of OSM data with other open sources - such as knowledge graphs or web mining/web scraping of company websites - should be developed. At the same time, critical reflection is essential, particularly with regard to data protection, anonymisation and potential distortions, in order to ensure the validity and

TABLE 4: REGRESSION MODEL: LOG. NUMBER OF FIRMS ACCORDING TO OSM PER AAB IN GERMANY. SOURCE: OWN COMPILATION. N=147. MISSING AAB (MAX. 150): BERLIN (OTHER DATA COLLECTION) MODEL M2 ADDITIONALLY WITH AGE INDEX, MODEL M3 ADDITIONALLY WITH INTERACTION TERM.

	M1	M2	M3
Log. Number of employees in official BA-statistics (1)	0,931***	0,945***	1,031***
% of employees in primary sector	-0,029	-0,035	-0,034
% of employees in secondary sector	-0,003	-0,003	-0,003
Pop. Density (2)	-0,04	-0,03	0,005
Index average age		0,019	0,019
(1) x (2)			-0,014
Constant	6,469***	5,530***	5,336***
n	147	147	147
Adj. R^2	0,861	0,861	0,861

reliability of the results. In view of the increasing importance of flexible and up-to-date data sources in the social sciences, economics and humanities, OSM is therefore a promising tool that should be further utilised and developed in interdisciplinary research approaches.

VII. REFERENCES

[1] D. Zielstra, H. H. Hochmair and P. Neis, "Assessing the effect of data imports on the completeness of OpenStreetMap – A United States case study", Transactions in GIS, vol. 17, no. 3, pp. 315–334, 2013.

[2] J. E. Vargas-Munoz, S. Srivastava, D. Tuia and A. X. Falcão, "OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing", IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 1, pp. 184–199, March 2021, doi: 10.1109/MGRS.2020.2994107.

[3] B. Herfort, S. Lautenbach, J. P. de Albuquerque and A. Zipf, "A spatio-temporal analysis investigating completeness and inequalities of OpenStreetMap building data", Nature Communications, vol. 14, 39698, 2023, doi: 10.1038/s41467-023-39698-6.

[4] D. Zielstra and A. Zipf, "A comparative study of proprietary geodata and volunteered geographic information for Germany", in 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 2010, pp. 1–15.

[5] P. Neis, D. Zielstra, A. Zipf and A. Struck, "Empirical analyses of the data quality of OpenStreetMap – Experiences from two years of operating several OSM online services", in Applied Geoinformatics 2010, 2010.

[6] A. Klonner et al., "'Ohsome' OpenStreetMap Data Evaluation: Fitness of Field Papers for Participatory Mapping", in Proceedings of the Academic Track at the State of the Map 2019, M. Minghini, A. Y. Grinberger, L. Juhász, G. Yeboah and P. Mooney, Eds., pp. 35–36, Heidelberg, Germany, September 21–23, 2019. [Online]. Available: https://zenodo.org/communities/sotm-2019, doi: 10.5281/zenodo.3387725.

[7] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography", GeoJournal, vol. 69, pp. 211–221, 2007.