

Extracting Relevant Points of Interest from Open Street Map to Support E-Mobility Infrastructure Models

Javier Valdes*

Roland Zink*

Matthias Schmidbauer*

Jane Wuth*

Sebastian Schröck*

ABSTRACT

In addition to commercial geodata, Volunteered Geographic Information (VGI) is gaining more and more importance in research. Platforms like Open Street Map (OSM) meanwhile provide an enormous amount of geodata. At the same time, however, new questions arise regarding the quality and the possibilities of using OSM data for research matters. Therefore, in the field of spatial planning, data require further validation processes and data cleaning frameworks. This paper presents a method of data processing in the context of electric mobility (e-mobility) research with a focus on a charging station placement model. The presented methodology is divided into pre-validation, to gather the relevant data set, and data processing, that specifies the relevant Points of Interest (POI) for further research by deleting all possible complications arising in OSM data. The validation process is customized to the model that determines the demand of electric charging by categorizing POIs into the four time slots living, work, shopping and recreation. By processing data in the presented way, the electric vehicle charging model is filled with improved input data, which allows to reduce the bias associated to the particularities of the OSM production process. A case study in the Bavarian-Czech border area demonstrates that the error correction rate through the model is at about 10%.

Neben kommerziell bereitgestellten Geodaten nehmen freiwillig erhobene geographische Daten (VGI, volunteered geographic information) in der Forschung einen wachsenden Stellenwert ein. Plattformen wie OpenStreetMap (OSM) bieten inzwischen eine enorme Menge an Geodaten, deren Qualität und deren Mehrwert für die Forschung zunehmend kritisch betrachtet werden. Besonders dann, wenn es um räumliche Planung geht, müssen die Daten vor der Anwendung angemessen validiert und bereinigt werden. In diesem Artikel wird ein Datenverarbeitungsmodell vorgestellt, mit dem OSM-Daten so aufbereitet werden können, dass der Bedarf an Ladeinfrastruktur für Elektroautos über von Nutzern eingetragene Point of Interests (POI) räumlich möglichst genau erfasst und abgebildet werden kann. Zunächst erfolgt eine Selektion relevanter Datensets durch eine Prävalidierung. Im zweiten Schritt werden die wichtigsten POI selektiert und spezifiziert. In einem dritten Schritt werden alle OSM-immanenten Schwierigkeiten beseitigt. Das hier vorgestellte Datenverarbeitungsmodell ist auf das Thema Elektromobilität zugeschnitten, für welches POIs in die vier Zeitkategorien Arbeit, Leben, Einkaufen und Freizeit eingeteilt werden, um so den Bedarf an Ladeinfrastruktur zu erheben. Durch dieses Modell der Datenverarbeitung soll die Energienachfrage für Elektroautos räumlich möglichst realistisch dargestellt werden und Rohdaten mit ihren bekannten Fehlern durch den OSM-Datensammelprozess bereinigt aufbereitet werden. Anhand einer Fallstudie, die im bayerisch-tschechischen Grenzraum durchgeführt wurde, wird gezeigt, dass die Fehlerrate durch die Implementierung validierter Daten um ca 10% reduziert werden kann.

* Deggendorf Institute of Technology

KEYWORDS

Electric vehicle charging station, spatial modelling, Open Street Map, data validation

Ladestationen für Elektrofahrzeuge, räumliche Modellierung, Open Street Map, Datenvalidierung

1. Introduction

Recent technological progress, especially in battery research, increases the electric vehicles' (EVs) competitiveness compared to conventional vehicles. Many governments are promoting EVs in order to reduce the carbon footprint of their respective country, which is pushed, among others, by the European Commission. Different subsidies and incentives have been established at several levels to raise the awareness for EVs and to increase their popularity in the public. In Germany, for instance, the government is paying a sustainability bonus ("Umweltbonus") to private users purchasing an EV and is granting tax deductions on company cars, among other measures [1]. In addition to the financial incentives, municipalities and regional actors also try to increase the attractiveness of e-mobility through economic and regional development concepts such as purchase subsidies, free parking or reserved parking spots and free bus lane use for EVs [2]. Currently, the focus of these concepts is put more on purposefully integrating e-mobility infrastructure into existing structures and less on substituting the latter by entirely new forms of mobility.

Given the objective of integrating EVs into the existing transportation and energy distribution network, creating the required infrastructures for recharging EV batteries poses new challenges to regional governments, municipalities and individuals, as well as to distribution grid operators. These challenges result from the high importance of implementing a charging infrastructure fitting to the regional needs of customers. This is due to the significant differences from conventional refueling as charging EV batteries takes considerably longer. Compared to fossil fuels, the high time expenditure on EV charging means that charging processes have to be embedded into the daily schedule and route planning of users, but also into the already existing regional energy supply structures. This is why it is highly important to find answers to the question of where charging stations (CS) have to be set up.

With the ability of spatiotemporal modeling, Geographic Information Systems (GIS) can be utilized to support answering this question. In addition to the GIS functionalities, the spatiotemporal approach requires a comprehensive database for finding suitable sites for CS. Nevertheless, due to the complexity of the collection methods and the high amount of information needed, there is still no database available that has been designed to assist the planning of charging infrastructure. The methodology presented in this paper draws the attention on how to use and ensure the quality of collaborative mapping platforms data for spatiotemporal modeling in the context of e-mobility. A big advantage of OSM compared to commercial geodata and other collaborative mapping platforms is that OSM is the most complete free open source database available for everyone. Moreover, OSM is constantly being updated, constantly growing and open to be edited by all potential actors involved.

Using OSM data offers a broad database with great potential, however, also involves some inconveniences. Even though collaborative mapping platform data is versatile enough to be used in a variety of applications, it has not primarily been designed to be used for scientific purposes. OSM, and other crowdsourced geographic data sources in general, are known to be partly incomplete or incorrect [3–8]. This study focuses on the extraction of relevant Points of Interest (POIs) from OSM to support e-mobility infrastructure models as in Zink et al. [9]. Unlike earlier quality assessments based on the quantitative and statistical characteristics of OSM data, this study contributes to the literature about OSM data, presenting an innovative method for extracting valuable information based on qualitative criteria and a data processing methodology. The results show the potential of qualitative studies and their complementarity with conventional statistical approaches to VGI quality assessments. The second contribution of this study is related to the literature on CS placement. Unlike most of the previous studies

on CS locations, the reference regions are rural areas with a focus on the road from Písek (Czech Republic) to Deggendorf (Germany), whose future suitability for EVs is currently being evaluated. As previous studies show that OSM biases are greater in areas with lower population density, focusing on this area allows to assess the robustness of the methodology under diverse conditions – it contains both rural and small urban areas – and to open up an avenue for research on CS locations models based on VGI.

The rest of the article is structured as follows: the next section presents an overview of rural e-mobility models and the possibilities for using OSM. Section 3 presents a methodology for selecting and processing OSM data for e-mobility projects, and section 4 shows results of applying the methodology in rural areas stretching over two countries, currently facing demographic exodus. In section 5 the findings are concluded.

2. Spatial localization for charging stations and OSM-based studies

Overall, models for the optimal EV charging infrastructure can be studied in many different directions, starting from the inside (battery lifetime analyses, driving distance enhancements or consumption needs of single tools) and leading to outside factors (CS infrastructure, willingness to buy EV or awareness of natural impact improvements). Although EVs have already become a reality in many countries and cities, the research effort devoted to generating optimal models for CS infrastructure planning is still in its infancy. Developing new approaches is of primary importance, as several studies indicate that range anxiety is a central inhibitor to going electric [10–12]. A sufficiently developed CS infrastructure could contribute to diminishing this inhibitor [13] and foster the trust in EV technology.

The lack of studies on the optimal placement of CS in rural areas is in strong contrast to the increasing attention on optimal allocation and modelling in urban areas. In a recent review of the literature, Shareef et al. [14] found that more than 100 articles with a focus on the optimal placement and sizing are published every year in specialized research journals, all aiming at urban areas. These studies are in many cases based on geographical information of different complexity.

For instance, Huang et al. [15] proposed a model for the placement of CS in urban areas based merely on geographical correctness. This means that CS are not located depending on the actual demand but on the geographic distances from each other to tackle possible driving/battery range anxieties. Frade et al. [16] define the demand for CS on the basis of the 2001 census data and apply their demand model to an area in Lisbon, Portugal. In comparison, Chen et al. [17] base their CS location model on the travel distance while bearing in mind the re-charging requirements of consumers with respect to the existing transportation network and its efficiency. Funke et al. [18] compare demand-driven localization models with a coverage-driven one. They state in their results that planners should first find out on what matter to put their focus: on the demand-driven amount of needed CS or the distribution of CS pursuing the coverage aspect. All mentioned studies propose CS locations depending on different factors, however, always in urban, very densely populated areas.

The vast availability of studies shows the wide range of options and possibilities for locating CS depending on the characteristics of drivers, EVs, locations, needs and charging technologies. However, to the authors' knowledge, only a few number of studies have been carried out in rural areas to describe important characteristics and they are based on GPS recorded data. For instance, Triebke et al. [19] collected data on more than 15.000 charging events in both rural and urban areas and their statistical analysis shows that although the charging behavior in urban areas is predictable, it is rather random in rural environments and occupancy rates are low compared to urban areas. In a more recent study, Gasde et al. [20] also found that although EVs offer socially and environmentally promising perspectives in rural areas, the economic drawbacks of current e-mobility solutions are a barrier for EV expansion. These case studies reveal the difficulties in implementing EV solutions in rural areas. The challenge is even greater when the model's scope is extended to covering multiple municipalities given the complexity of GPS-based collection methods and the recourse-intensive data collection method.

Data collection, or more specifically, the collection, assembling and dissemination of geographic data, is the focus of collective mapping and VGI databases. The increasing availability of data sets of freely available VGI has led to strong

interest from researchers and practitioners in the usability of these data, both their limitations and their potential. Although VGI offers an alternative mechanism for the acquisition and compilation of geographic information, it requires many individuals to participate. Some authors argue that researchers should analyze but not complement VGI [21]. The collection of data should be done by the overall public which leaves researchers more space and time to focus on quality checks. Those quality checks have been realized by many studies in a variety of ways. The unconventional way in which data are being produced as well as their richness and heterogeneity have resulted in a range of different research questions on how to assess, mine, enrich, or just use these data in different domains and for a wide range of applications. In the area of e-mobility, to the authors' best knowledge, only the study of Zink et al. [9] is based on OSM data. Similarly to Frade et al. [16], they locate CS infrastructure at places where the potential demand for electricity for transport reasons is the highest, but instead of using census data they used OSM data to allocate the demand. Using OSM allowed Zink et al. [9] to generate a CS location model for a large low-density area like the studies of Triebke et al. [19] and Fournier et al. [20]. Finally, Wanger et al. [22] also developed a model for demand location for car sharing, in this case based on Google API; they included 180,000 POIs as well as demographic data in order to identify POI categories that substantially influence variation in car sharing activity.

OSM and crowdsourced geographic data sources in general are known to be partly incomplete or incorrect, and many studies were performed to analyze data characteristics of different VGI platforms [3–8]. In the following, we focus on OSM which today is the platform with the highest number of registered users and the largest crowdsource-based data volume [3]. The methodology discussed in this paper attempts to make it possible to use these data for research, that is to say, they need to be cleaned and verified. Senaratne et al. [3] provide a summary of articles in the area of VGI Data Quality Analysis. Quality measures for VGI as defined by the International Organization for Standardizations (ISO) are completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy. Besides these qualitative criteria, quantitative criteria are introduced, namely purpose, usage and lineage. For all these different criteria, Senaratne et

al. [3] have executed analyses in order to define the level of correctness of different data sources. However, these authors did not introduce data cleaning processes on a larger scale.

Regarding specific OSM data studies, several research projects focus on different OSM data sets and analyze different layers. For Germany, Fan et al. [23] compare road data in OSM with official data while Zielstra and Zipf [5] compare road data with commercial data acquired by navigation service providers. The results of both studies indicate medium to high levels of completeness and quality of the examined data, depending on the population density. Moreover, they show how fast the OSM data set is increasing over time. Between July and December 2009, the road data provided on OSM in the specified research regions increased by 20%. Neis et al. [6] show that OSM data in several cases even exceed the data pool used by car navigation systems in 2011. Arsanjani et al. [7] compared OSM data with pan-European GMESUA data and Dorn et al. [8] compared OSM data with official German administrative data updated every 3 months, both putting a focus on land use and land cover (LU/LC) data. Arsanjani et al. [7] conclude that LU/LC OSM data lie within a correctness and completeness level of 40%-60%. Dorn et al. [8] reveal that concerning LU/LC data, the quality and completeness of OSM data strongly depends on the feature classes observed. Forests, for example, show a large completeness and correctness level in the study area in southern Germany. Other classes of data such as farmland or urban areas/characteristics are often incomplete although the information provided is of high quality. In other cases, the quality of data is low while the level of completeness is high [8]. Barron et al. [24] focus their research on quality assessments of OSM by reviewing the history of OSM data. They state that it is possible to evaluate the quality of the observed data solely by assessing the history of the data generated and without using external sources.

These examples demonstrate that many research efforts have been realized in different use cases to define the degree of quality of OSM data. A variably applicable cleaning and verification process for a better usage of OSM data, however, has not yet been introduced. The next section introduces a validation methodology in order to fill this gap and to increase the application of these freely available data for specific research topics.

3. Methodology for data validation

The data validation methodology presented herein was developed for an e-mobility demand model based on OSM data which was developed by Zink et al. [9], however, it may also be tailored to other mobility-related models as the one by Wagner et al. [22]. To facilitate the logic behind the data validation process presented in the following section, figure 1 displays the modeling approach by Zink et al. [9]. As described in the previous section, the aim of this site planning model for e-mobility is to place CS infrastructure at locations where the potential demand for electricity for transport reasons is the highest. It is therefore based on existing transportation patterns and current transport infrastructure and not on the current system based on gas stations. Due to the long periods needed for charging EV, the main idea behind the model is to create

charging infrastructure at existing parking locations instead of using the current model of isolated gas stations. The charging demand calculation is based on a spatial statistical methodology considering demographic parameters and maximum walking distances from the CS to selected POIs according to four age groups. It furthermore contains the average time use per activity per day, dwelling time at one POI, the penetration rate of EVs and POIs from OSM. The walking distance determines the optimal location within a region with high energy demand by minimizing the average walking distance to the surrounding POIs. Therefore, the optimal location is found while simultaneously considering high parking demand and short walking distances. If there is only one POI within the user's maximum walking distance, the charging location is as close to the POI as possible, but always bound to the street.

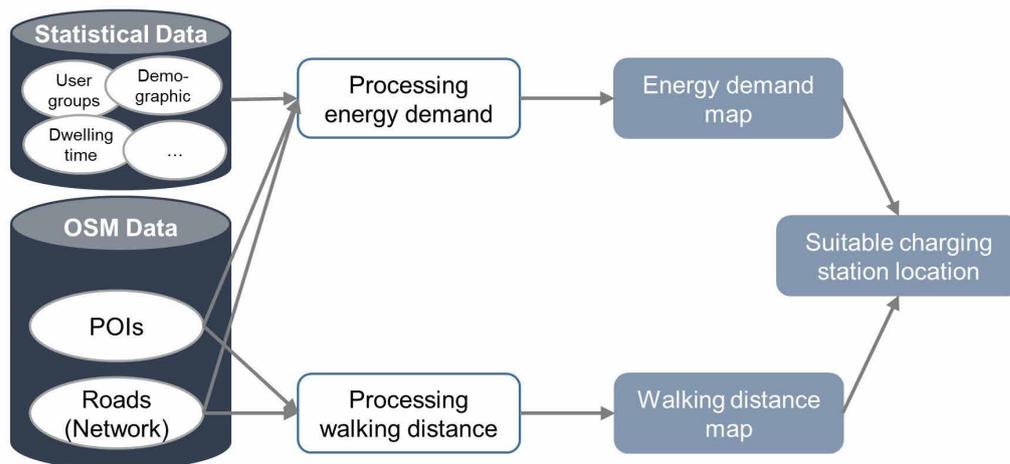


Figure 1. Electricity demand and site planning modeling approach

POIs from OSM are included in the model to define the areas where demand is highest. As already described, in OSM, gathering data depends on the users, this means that anyone who finds a certain spot on the map that is not yet classified, such as a fuel station, can add it accordingly and tag it with the category of their choice. Consequently, a personal bias might be in place for each single location defined in OSM. Thus, a qualitative review and data validation process is necessary to assess the quality and relevance of the input data.

Furthermore, POIs like, for instance, kindergartens or schools can have a point-like or a polygon shape in OSM. As an example, figure 2 shows a screenshot from the OSM web

map with a query about schools. On the left hand side, a point is selected that represents *Staatliche Berufsschule Deggendorf* (public vocational training school). The screenshot on the right hand side visualizes the selection of the entire building (polygon), including four other schools (so 4 points in one polygon). For modeling, this raises the question as to which dataset the location search should be based on: on one school or on four schools. In the electric demand modeling approach, the number of POIs is the determinant of EV's charging demand, which induces the need for inserting the number of POIs into the model that is closest to reality. The methodology presented herein reviews necessary processes for validating OSM data before including it into an e-mobility model.

Query about school: Point shape



Screenshot from OSM <https://www.openstreetmap.org>, October 2018

Query about school: Polygon shape



Figure 2. Screenshot of OSM web map

3.1 OSM data characteristics

In OSM data, different POI categories like buildings, roads, water, etc. are classified and they are further divided into sub-categories, which are also called feature classes (f-classes) by Geofabrik, the provider of OSM data used in this study. F-classes are, for instance, tree, school or bank¹. The decision to add a new location to a specific category in the OSM data depends on the judgement of the volunteers. As it is possible for everyone to participate in adding data to OSM, the map is highly user-reliant, depending heavily on individual preferences as well as on the socio-economic and geographic characteristics of each region. In order to reduce this source of bias, the OSM community maintains a large OSM wiki page and implemented a series of quality checks as well as reviews of the dataset. However, users mainly add new features to the map for regions they know well. As a consequence, every tree and every bench in a park have been added for some regions, while the mapping for other regions is rather poor. Regions and POIs attracting more people are mapped with more detail and are more complete while, for instance, important POIs in rural regions are missing. This pattern is also observable with regard to the distribution of mapped parking spaces. A decrease in the percentage of mapped parking areas is observable the further away the parking space is from regions that are highly interesting. The propensity for higher data correctness and completeness depending on

population density is also very well represented in the analysis of Dorn et al. [8]. Due to these patterns, it is necessary to clean and validate the data before using OSM data and in order to assess the level of correctness. The validation process presented herein is specifically designed for the requirements of the OSM data and can be applied to OSM data from other regions with minor adaptations. It consists of two main steps, namely (1) pre-validation, which includes selecting relevant data for the project, and (2) data processing, which considers the cleaning of data and the preparation of one complete data set.

3.2 Pre-validation

In the first step of the validation process, the main goal is to extract only data that are relevant for the e-mobility model described in figure 1. The first questions that needs to be answered is whether (1) the feature class (f-class) is relevant for the model's goal. The second question is whether (2) the f-class is distinguishable enough for use within the model. The criteria for the f-classes to be chosen are dependent on the model specification and the research objective. In this model, as explained above, different time categories were defined. The POIs need to be classified into one of these classes according to their main purpose. It is therefore necessary to ask (3) whether the f-class is defined narrowly enough so that it can be categorized into one of the four time classes. Figure 3 lists these questions including criteria that need to be met in order to fit into the model.

¹ See: <https://www.geofabrik.de/data/geofabrik-osm-gis-standard-0.7.pdf>

Extracting Relevant Points of Interest from Open Street Map
to Support E-Mobility Infrastructure Models

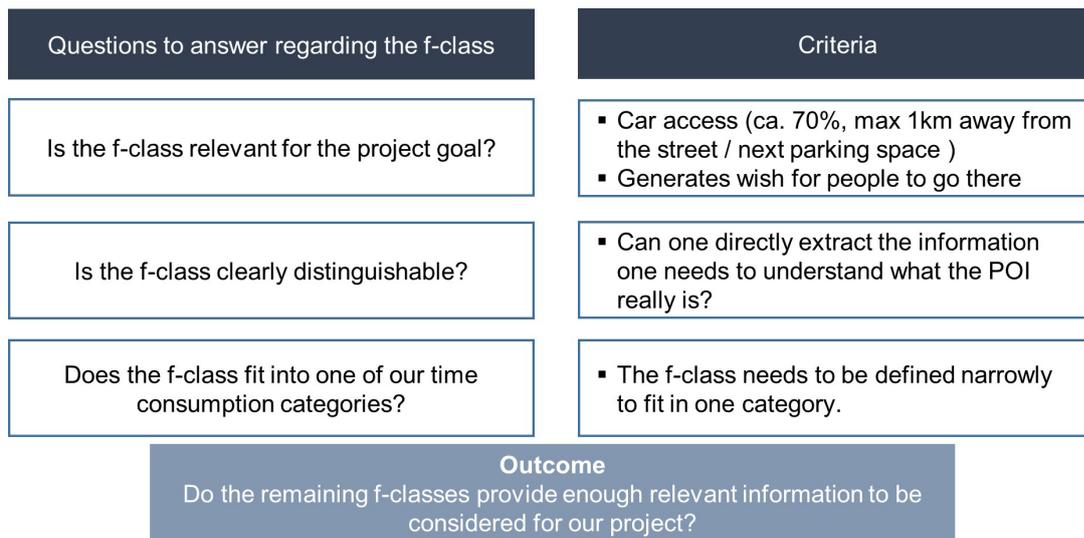


Figure 3. Pre-validation process

Besides these three initial questions for data selection, additional qualitative criteria have to be met in order to further process the relevant data. These criteria do not apply directly to the entire f-class but to the single features within f-classes. The first criterion is related to the location’s geographical characteristics. In the model it is assumed that at least 70% of the POIs per f-class needed to be close to streets or parking spaces. A range of 1km is defined as the maximum distance. The maximum distance refers to 1km on paths or tracks, not necessarily passable by car, however, accessible by foot, bicycle or other means of transport. Additionally, the f-class considered in the model needs to be exactly categorized in its additional information. For example, the POI category “building” can be any type of building, ranging from family houses over public administrative buildings to convenience stores. Thus, if additional information about the type of the building is missing on the platform, it is neither possible to define the kind of demand for CS infrastructure for this f-class, nor can the POI be defined according to the model’s needs. The correct pre-validation yields a list of f-classes which should be relevant for the model and can be further evaluated in the second validation step.

The following quantitative and qualitative criteria are developed in order to define the OSM data that are relevant for the EV CS demand-model:

i. Distance to roads or parking spaces

Walking distances is a central concept in a model for transportation and public health and it is the major factor in the context of parking choice [25]. Overall, research on the exact distance and duration of walking trips for different purposes and across diverse population groups remains limited. For Germany, it is difficult to obtain estimates of the distribution of walking distances from parking places to final destinations. As to research in other countries, it is difficult to assess the validity of the results based on survey data. These limitations arise because of the discrepancy between perceived distances and objective distances [26, 27]. Based on these studies and for the purpose of the model applied, the maximum walking distance to the next parking place or street is set to 1km.

ii. Relevance

Different time categories for weighting the relevance of POIs are applied in the e-mobility model. These time categories are based on surveys that collected information on the time spent on daily activities in Germany and the Czech Republic [9]. The f-classes are considered relevant for the model if they can be classified into one of the activities depicted in the time surveys. For instance, the f-classes “school” or “courthouse” can be classified within activities concerning the time category “work”, other f-classes such as “tree” or “bench” cannot easily be allocated to one certain activity as they describe street furniture and vegetation. In order to assess the significance of this criterion,

a literature review was conducted to see whether the charging events recorded on pilot projects and used to model EV driving behaviors are consistent with the time usage model applied in this project. The results of the literature review show that current charging behavior in urban areas is strongly determined by the activities of EV drivers [19]. Studies point out that the distribution of charging events in function of the location of the CSs (for instance: working places, homes and recreation locations) are consistent with the users' use of time [12, 27].

iii. Differentiability

Differentiability is a key element to the purpose of e-mobility models and is related to the way in which data are recorded and classified. It accounts for the capacity of distinguishing within and between the classes. The f-class "swimming pools" provides an idea of the capacity to differentiate within classes. It characterizes whether the data recorded respond to a common structure and whether their characteristics are identifiable. In the case of the swimming pools, the data is considered to be non-differentiable as it does not allow to draw a line between public or private swimming pools. Another issue can be displayed by the f-class "commercial". This class does not allow to be differentiated from other f-classes such as shops, restaurants or clothes. Therefore, such f-classes are also not considered to be differentiable among f-classes.

iv. Classifiability

As OSM data are integrated into the model with data of different origins, the classifiability of each class is paramount. In the OSM e-mobility model, instead of basing demand for each CS on the current behavior of EV users, the model uses the amount of time that both members of the German and the Czech population spend on different activities. The classifiability criterion represents the possibility to allocate the f-classes to those activities. Therefore, an enterprise or a city hall are classified as places of work, a restaurant is classified as a recreation location and any kind of shop is allocated to shopping activities.

3.3 Data processing

The data processing phase consists of several steps, all building on one another in order to identify whether the locations represented in the OSM data correspond to real sites and to make sure that duplicates are erased. Figure

4 displays the different steps of this process and the resulting actions. In order to do so, it is first necessary to check the geographical intersections between points and polygons to see whether there are points that are as well included in the form of a polygon, both referring to the same POI. This first step in the analysis is done across all f-classes, simultaneously checking for duplicates within and between the f-classes. There are some classes that are likely to be used as substitutes, like "observation tower" and "viewpoint". The extracted list of duplicates is further separated into three categories, namely "exact matches", "approximate matches" and "possible coexistence". In the first two groups, the corresponding polygons are erased in order to reduce the amount of duplicates. In the group containing the inconsistencies, both the points and the polygons indicate two different sites and thus must remain in the data set. The "possible coexistences" are left aside in this step, as it is liable to believe both OSM tags are relevant.

After erasing all irrelevant polygons or points identified in this step, polygons get transformed into points and one large data set is created for further analysis. A common issue with OSM data is that different users may consider the same location for different f-classes. This results in one data point, with one OSM data ID, represented in different f-classes. Therefore, when considering more than one f-class, two or more data points with the same OSM ID at the same location but in different f-classes are possible to arise. Thus, the intermediate step is to erase points that have the same OSM ID and indicate the same spot. As already explained, some f-classes have a broader definition than others, a point "commercial" can cover various topics that are specified in other f-classes like "clothes", "hairdresser", "supermarket" or "restaurant", just to name few. Thus, a list of priorities is introduced that deletes the second or third data point with the same OSM ID, always leaving the most specific point in the data set. For example, a religious location with the f-classes church, Catholic, cemetery, work-art, Christian and garden is reduced to church.

For further data processing, data points are compared considering the belonging to an f-class. Further duplicates may exist even after comparing points directly situated in polygons and deleting double OSM IDs. This can happen if the same location appears as a point and as a polygon without any overlap, for instance, it

Extracting Relevant Points of Interest from Open Street Map to Support E-Mobility Infrastructure Models

may be the case of a polygon representing an edifice that is surrounded by a garden and also of a point, representing the street number of the edifice and being located on the sidewalk. These duplicates can best be found by geographically estimating rational distances between points of interest. A wastewater plant, for example, is usually not located directly next to another one, clothing stores, however, are mostly grouped together in a particular part of the city center. Thus, a special distance is allocated to each feature class in which one point would normally not occur twice. For that, f-classes representing features that could be used equivalently are grouped. A common example would be the f-classes graveyard and cemetery. Simultaneously, to make sure former polygons are represented according to their size, this analysis was made by introducing three different polygon sizes. Either 0, 50% or 100% of the former shape is represented in the form of a buffer around the points. A comparison between

the three different results are made in order to estimate the best fitting solution for each f-class. Double entries are thus deleted accordingly while points of different origin are not erased by mistake.

In the last step of data processing, buffers of different sizes are laid around each remaining point in the data set to analyze in purely geographical manner whether points are close to each other. This final step is done without limiting the comparison to inter-f-class consolidations. Results indicate the most mapped areas within the data sets. This step is undertaken to find and manually delete remaining duplicates. Besides being the final step of data processing, it simultaneously represents a quality assessment of the previous steps. If many duplicates occur that still need to be erased manually, it is possible to identify which parameters of the former steps need to be adjusted in a way that allows for a more qualitative result.

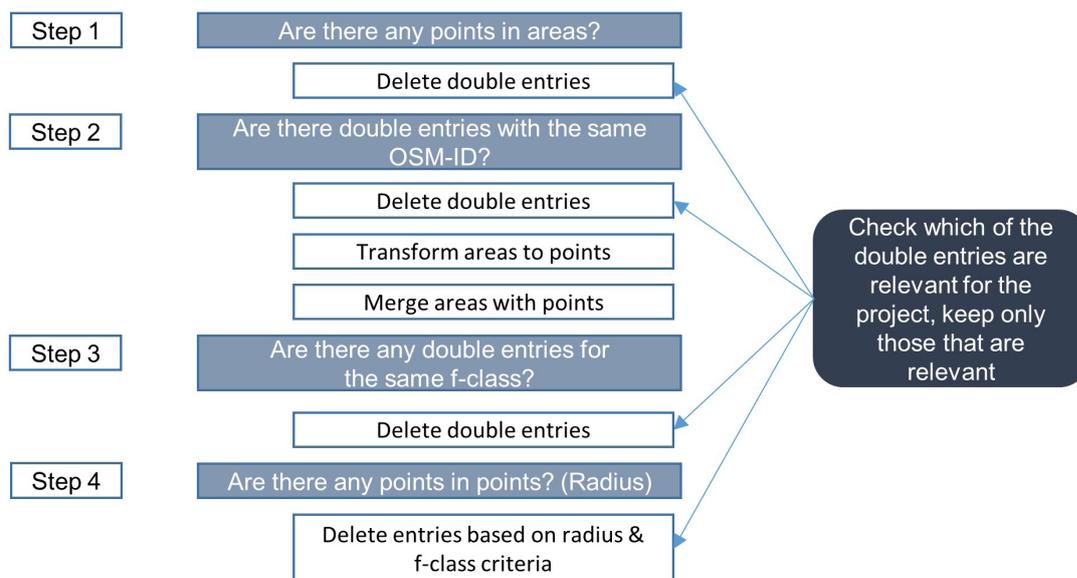


Figure 4. Data processing

4. Case study eRoad: Deggendorf – Písek

The presented methodology is applied to an area at the German-Czech border (see Figure 5). The regions involved around Deggendorf and Písek are facing particularly large challenges: on the one hand, population numbers are decreasing due to the absence of basic amenities or the general lack of employment opportunities in rural areas. On the other hand, there is a flow of migration of new citizens coming from towns

in a process of deconcentrating population also called “counterurbanization” [29]. Under such a scenario, municipalities are lacking fitting infrastructure and their competence in areas related to mobility is often not sufficient in order to react to the ongoing e-mobility transition with a mature concept. As preferences and needs of habitants from these regions do not correspond to those of urban citizens, there is a need to adapt existing e-mobility solutions to the characteristics of this area. A concept covering all relevant topics can only be realized when municipalities ensure

basic infrastructure for their population that satisfies current and future demands. The concept moreover needs to provide attractive incentives and possibilities for the economic development of the area and, at the same time, offer an attractive

quality of living. Therefore, the selection of POIs to be implemented in the CS demand model should be based on such characteristics and present an appropriate response to the potential citizens' needs.

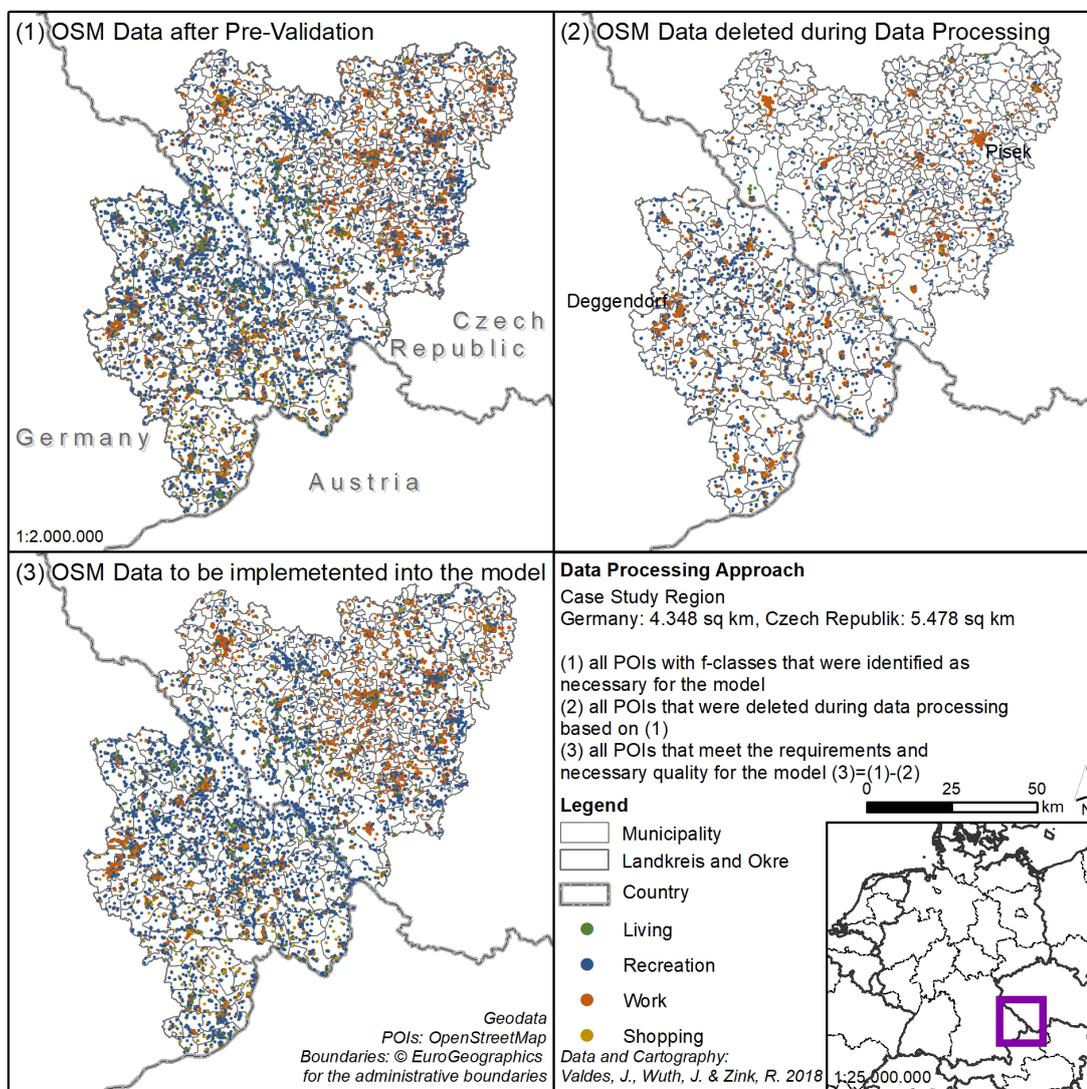


Figure 5. eRoad project region and POIs from OSM data

The full OSM dataset was downloaded from Geofabrik GmbH (<https://www.geofabrik.de/>) covering Lower Bavaria, and the Czech Republic in June 2018 and in August 2018, respectively. Following the first quality criterion, POIs were selected based on their proximity to roads and parking spaces. For the regions of the study, OSM data are mainly composed by polygon layers and f-classes associated with buildings and land use, which account for almost 80% of the whole data set followed by data on transport infrastructure, mainly roads and railways (12%) and points signaling very different classes of locations, from natural parks to cafés and bus stations (8%).

In the pre-validation phase, the georeferenced f-classes were evaluated in relation to the explained criteria: relevance, differentiability and classifiability. Using a scale from 1 to 3, each f-class received a rating on how well they covered the queried need for the model. For instance, a bus station is relevant for the e-mobility model as some users may drive to the station by car and then continue their journey by bus, and therefore may score a 3 in relevance. It is, however, not differentiable, as the f-class contains small bus stops that cannot be considered to have a car transport demand but also central bus stations which may raise

Extracting Relevant Points of Interest from Open Street Map to Support E-Mobility Infrastructure Models

a large demand. It thus would also be difficult to classify and therefore would score a 2 in the classifiability and a 1 in the differentiability criterion. On the other hand, a waterfall may also be relevant as it could generate interest to visit it, but not as relevant as a supermarket that would score a 3. Moreover, a waterfall is highly differentiable and may score a 3 there, unlike a building with no information on its use, which scores a 1 or 2. Finally, all f-classes containing a 1 were discarded as they were not at all considered to be relevant, differentiable or classifiable. In this process, the number of f-classes was reduced from 313 to 130 in the case of Germany, and from 270 to 113 in the case of the Czech Republic.

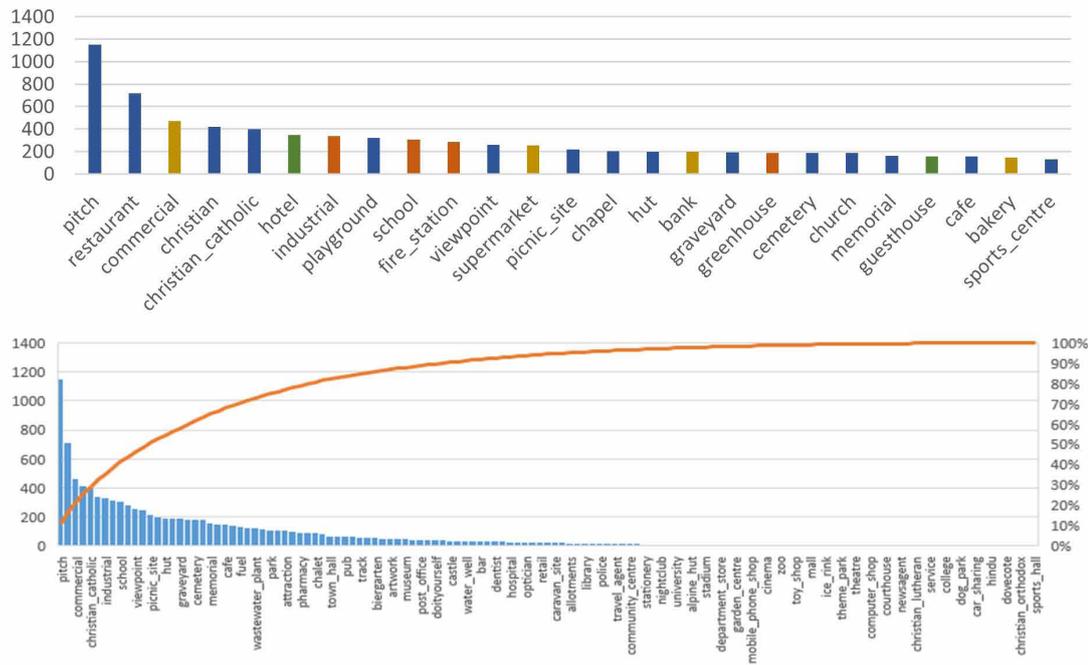
Based on the model classifications of Zink et al. [9], the f-classes were organized into the four different time categories “work”, “home”, “recreation” and “shopping” in order to introduce them to the model (figure 6). The f-class categorization was done equally for both countries. The results of the remaining points per f-class are demonstrated in figure 7. It shows that the category “recreation” is the largest one in Germany, including 4,407 points, directly followed by the category “shopping” with 2,126 points. In the Czech Republic, the category “shopping” is the largest with 3,126 points, followed by “recreation” with 2,982 points. The different order of time categories between the countries may indicate that users mapping the Czech Republic have concentrated their efforts on different f-classes compared to users mapping Germany, or that the structures of the two countries are differently specified, even though the regions observed are directly neighboring each other.

In figure 7, an overview of the points eliminated by demand category in each methodology step is presented. After allocating every f-class to a time category, the data processing started with the point in polygon comparison. 677 of the 10,620 locations in the German case, respectively 101 of the 9,613 locations in the Czech region observed show a problematic entanglement. Several points were found in one single polygon. For example, 18 attraction points were located in one polygon belonging to the f-class zoo. One OSM user named each attraction according to the animal that could be seen from that exact spot in the zoo. Simultaneously, several large and broader classified f-classes like, for instance, industrial contained more specified

points like car wash, hairdresser or butcher. As a result, the f-classes industrial and commercial of the German data set were removed as more than 50% of the points and polygons in these f-classes were erroneous. The other point-in-polygon incidents were deleted depending on the more precise explanation of the f-class.

After cleaning the data set according to the findings, all remaining polygons were transformed into points and merged with the shapefile including all locations that were marked as points from the beginning. Step 2 in the analysis deleted 958 German and 520 Czech points of the datasets, as they were duplicates by OSM ID. In both regions, these points were mainly recreation places related to religious activities. Due to the proximity of churches, cemeteries, chapels and locations with the reference Christian, which obviously in many cases made reference to the same location, many of these points had to be deleted in order to avoid a non-existing demand for CS infrastructure. Moreover, in the Czech regions, a high number of camping sites were referenced in the OSM data as different locations, even if they made reference to buildings and land areas belonging to the same camping area. On the other hand, in the German regions many duplicates were locations belonging to the categories chalet, gift shop and car wash. These differences give a further idea of the dependence of OSM data to regional and cultural characteristics and preferences.

OSM data for Germany after Pre-Validation



OSM data for Czech Republic after Pre-Validation

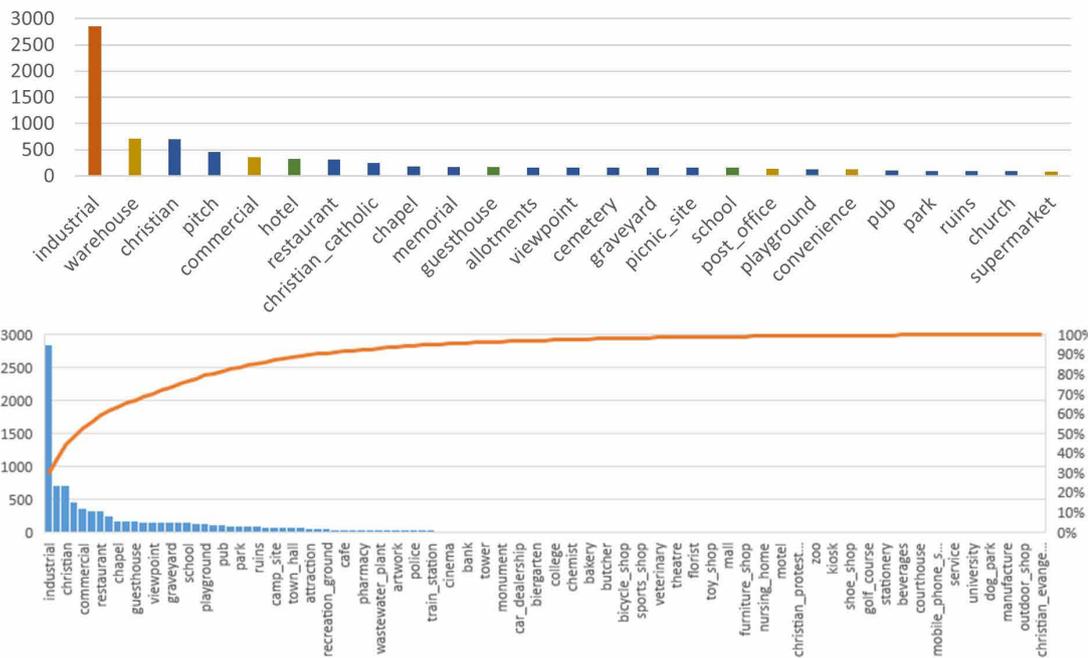


Figure 6: Number of POIs by f-class after Pre-validation (top 25 f-classes)

The last step of the methodology presented herein is responsible for most of the discarded data. This step is to a strong degree regionally dependent as it has to do with the way OSM users insert their data. A large number of locations were eliminated in the Czech Republic (1,907), of which the most were industrial locations (1,212 out of 1,625), schools (51 out of 93) and pitches (130 out of 329). In Germany, 1,275 locations were eliminated

including a large number of chapels (52 out of 146), huts (64 out of 134) and also pitches (367 out of 782). This last step reduced the dataset to the final 8,010 locations in the German regions and 7,085 individual locations in the Czech Republic. 5,438 points were eliminated during the process (Figure 7). In both countries, around 73% of the pre-validated data are left after data are left to be implemented into the model after data processing.

Extracting Relevant Points of Interest from Open Street Map to Support E-Mobility Infrastructure Models

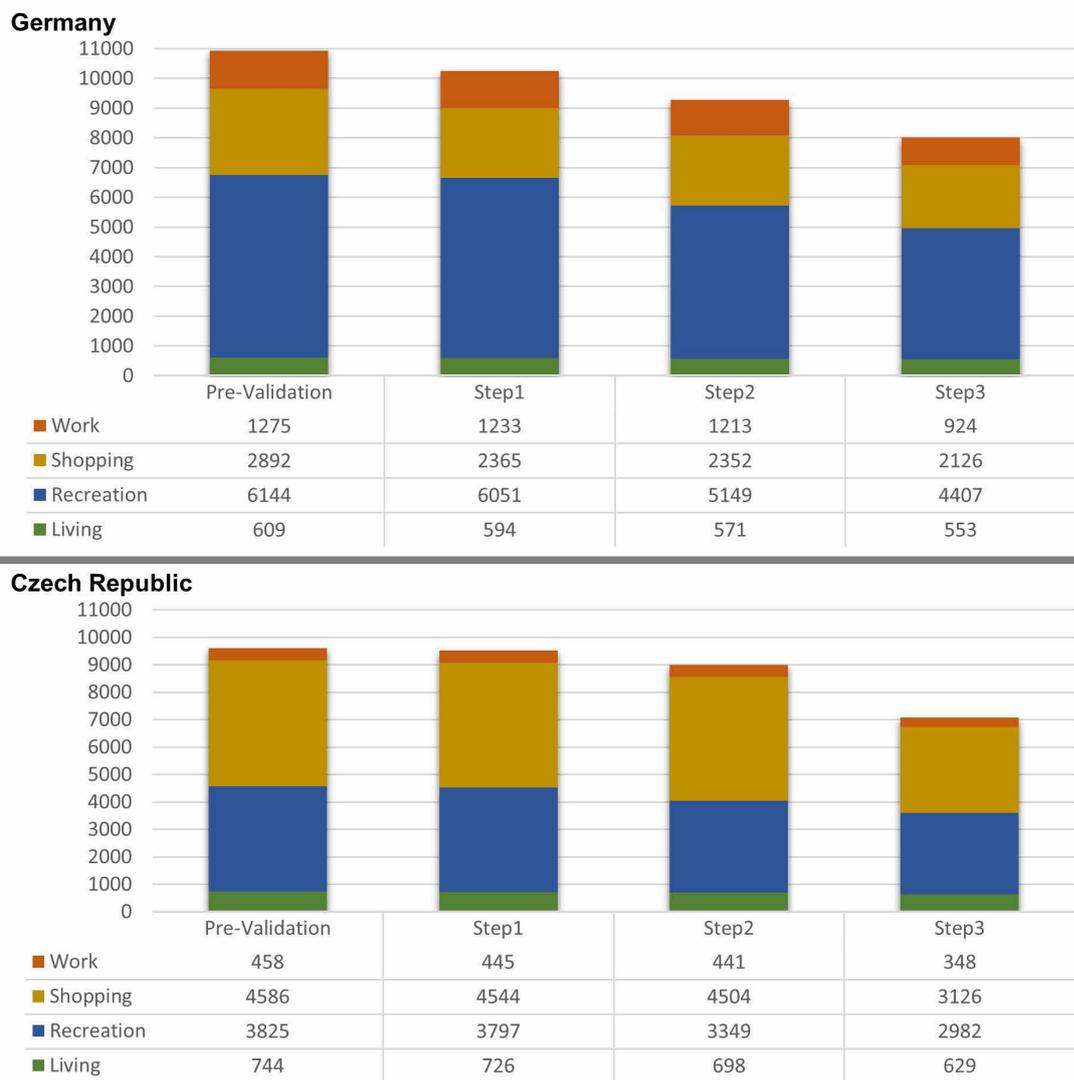


Figure 7: Number of POIs by demand category by data pre-validation and data processing step

5. Discussion and outlook

This study focused on the potential of OSM data for locating demand for modeling EV energy demand, presenting a novel approach for OSM data quality validation. Instead of conventional statistical approaches, the performed analysis is based on qualitative criteria which are related both to the content and the structure of the data. The results of the analysis performed show that OSM data cannot be included in e-mobility models as they are because of the way in which they are generated. In particular, the OSM data analyzed in this project show that the features in many cases lack adequate descriptions that allow to identify basic characteristics of the locations as, for instance, their names. This particularly applies to the f-classes “industrial”,

“commercial” and “buildings”, which in many cases do not contain any description of the industrial and commercial activities they made reference to. In the case of buildings, the purpose of those buildings and their typology is mostly missing.

The results show that the high degree of subjectivity of the users and the administration of the platform are other issues related to data quality. There is a lack of quality assessments by the platform administrators of common errors like duplications of POIs. These errors can occur in different situations, like when mapping an already existing site or adding an already existing POI by allocating it to a different POI f-class with a slightly different definition. A common mistake, for example, is

to allocate one church to the f-class Catholic, Catholic church or church. What's more, the mapping of data is very user and POI-specific. Some users might map "swimming pools" not only as publicly available swimming pools, but define every private pool as a swimming pool. It would therefore be interesting if the tag for public and private locations was better controlled. Moreover, another common mistake related to classification stems from confusing different classes, for example not all participants may know the differences between the f-classes graveyard and cemetery.

Adding these feature class errors to a model may not have an impact at the aggregated level, but it can heavily distort the outcomes for some specific location to an even greater extent than if the erroneous feature class were left out completely. Taking once more the example of the swimming pools, defining one POI as several ones can be another area for errors, because the institution "swimming pool" consists of more than one pool. Thus, a swimming pool with 3 different pools would be marked in the map as 3 different POIs, which falsifies the data set and triples the calculated energy demand when using three POIs instead of the single one existing. An avenue of future research will be the evaluation of these biases through case studies in different regions.

As the f-classes function as tags for POIs, different POI categories sometimes contain very similar f-classes. This may lead to further POI duplicates as some users may map the viewpoint as a "watchtower", some as an "observation point" and then again, others may just use the category "viewpoint". In this example, one POI would turn out to be three on the map with different names but referring to the same spot. Moreover, all f-classes consist of point and polygon data sets. As a result, many, however not all of the points are additionally represented as polygons, inducing more duplications.

In order to cope with these problems, one of the potential possibilities is to complement OSM data with data from other sources. For example, in the case of Germany and the Czech Republic, the data can be enriched by firm-level data obtained through the Czech Statistical Office (CZSO) and the German Chambers of Industry and Commerce (IHK) company registers in order to locate industries. Both institutions collect data on firm locations and the numbers

of workers which allow to differentiate and calibrate the model in function of the affluence to work places based on the numbers of workers. Nevertheless, in order to include such data sets, it would be necessary to ensure that there are no duplicates between the different data sources as the official registers may contain enterprises that are already included in the OSM dataset. Moreover, such data sets are often very costly and the large advantage of OSM data, besides its richness, is that it is free of charge.

The methodology presented in this paper shows the relevance of conducting studies to increase the use of OSM data for geographic modeling. Nevertheless, the current classification methodology based on tags should be improved in order to make OSM more user-friendly. To this end, one option would be to re-define the tag classification methodology in order to simplify the complexity of the data. This option, although possible, would probably impact one of the main advantages of OSM data: its richness. Another option would be to generate a parallel or multiscale tag classification that allows to identify which data is susceptible to be used in different contexts.

The analyzed OSM data in the scope of a model for CS placement show that most of the selected POIs are not suitable for the model. Nevertheless, it is necessary to note that OSM is not complete and therefore there may be locations that have not been included in the model. This bias is normal, taking into account the relative novelty of the OSM project and the constant nature of OSM data. For this reason, the major part of the methodology presented herein is based on automatic processes and algorithms specifically designed to reduce human supervision of the results. Moreover, the data processing also includes a validation step in order to acknowledge for any bias associated to the f-class classification during the validation process itself.

Finally, it should however be noted that many research projects could not have been carried out in the way they are currently done if OSM data were not available. The availability of such a rich and diverse data set can despite large error fields provide insights that would normally have not been possible without considering costly datasets. Therefore, there is a need to continue the research in OSM data quality control in order to increase the range of applications of such a rich source

Extracting Relevant Points of Interest from Open Street Map to Support E-Mobility Infrastructure Models

of information. For that, it may be necessary to create specific classification methods. These classifications would allow to identify which data are susceptible to be used within different contexts. This would make it possible for a region or city interested in, for instance, applications of OSM data for tourism to just download a set of data already validated. Further work should focus on approaches to quality assurance of such data for accuracy assessment and to extend the analyses to other regions. This may impact the quality and accuracy of the overall model and reduce biases, especially in rural areas with lower overall numbers of POIs.

6. Acknowledgement

The presented methodology and results are extracted from the research and development project “eRoad Písek - Deggendorf”, which is funded by the European Union and the European Regional Development Fund through the program “ETZ Freistaat Bayern – Tschechische Republik 2014 – 2020” (Interreg V). Project partners are Technologické centrum Písek (lead partner), E-WALD GmbH (project partner), and the associated partners, the City of Písek and e-Šumava.cz.

7. References

- [1] Taefi, Tessa T.; Kreutzfeldt, Jochen; Held, Tobias; Fink, Andreas (2016): Supporting the adoption of electric vehicles in urban road freight transport – A multi-criteria analysis of policy measures in Germany. In: *Transportation Research Part A: Policy and Practice* 91 (September), S. 61–79.
- [2] European Alternative Fuels Observatory (2018): Incentives & Legislation. Online verfügbar unter <https://www.eafo.eu/incentives-legislation>, zuletzt geprüft am 03.08.2018.
- [3] Senaratne, Hansi; Mobasheri, Amin; Ali, Ahmed, Loai; Capineri, Cristina; Haklay, Mordechai (Muki) (2017): A review of volunteered geographic information quality assessment methods. In: *International Journal of Geographical Information Science* 31 (1), S. 139–167.
- [4] Goodchild, Michael F.; Li, Linna (2012): Assuring the quality of volunteered geographic information. In: *Spatial Statistics* 1 (May), S. 110–120.
- [5] Zielstra, Dennis; Zipf, Alexander (2010): A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. Universität Heidelberg (Geographisches Institut, Geoinformatik).
- [6] Neis, Pascal; Zielstra, Dennis; Zipf, Alexander (2012): The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011. In: *Future Internet* 4 (1), S. 1–21.
- [7] Arsanjani Jamal, JOKAR; Zipf, Alexander; Mooney, Peter; Helbich, Marco (2015): Openstreetmap in Giscience. Cham: Springer International Publishing AG (Lecture Notes in Geoinformation and Cartography).
- [8] Dorn, Helen; Törnros, Tobias; Zipf, Alexander (2015): Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. In: *International Journal of Geo-Information* 4 (3), S. 1657–1671.
- [9] Zink, Roland; Pagany, Raphaela; Marquardt, Anna (2017): OSM-basierte Standortmodellierung von Ladesäulen für Elektromobilität am Beispiel des Bayerischen Waldes. In: FOSSGISS e.V. (Hg.): Tagungsband der FOSSGISS 2017: Anwenderkonferenz für Freie und Open Source Software für Geoinformationssysteme (Universität Passau; 22.-25.03.2017), S. 59–64.
- [10] Akbari, Hesam; Fernando, Xavier (2016): Futuristic model of Electric Vehicle charging queues. In: 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), S. 789–794.
- [11] Ahn, Yongjun; Yeo, Hwasoo (2015): An analytical planning model to estimate the optimal density of charging stations for electric vehicles. In: *PLoS ONE* 10 (11 (e0141307)).
- [12] Kaiser, Oliver S.; Meyer, Sarah; Schippl, Jens (2011): ITA-Monitoring "Elektromobilität". Kurzstudie. Karlsruhe: ITAS-EPP (Elektronische Pre-Prints). Online verfügbar unter <http://www.itas.kit.edu/pub/v/2011/epp/kaua11-pre01.pdf>, zuletzt geprüft am 25.01.2016.
- [13] Liao, Fanchao; Molin, Eric; van Wee, Bert (2017): Consumer preferences for electric vehicles: a literature review. In: *Transport Reviews* 37 (3), S. 252–275.
- [14] Shareef, Hussain; Islam, Md. Mainul; Mohamed, Azah (2016): A review of the stage-of-the-art charging technologies, placement methodologies, and impacts of electric vehicles. In: *Renewable and Sustainable Energy Reviews* 64 (October), S. 403–420.

- [15] Huang, Kai; Kanaroglou, Pavlos; Zhang, Xiaozhou (2016): The design of electric vehicle charging network. In: *Transportation Research Part D: Transport and Environment* 49 (December), S. 1–17.
- [16] Frade, Inês; Ribeiro, Anabela; Gonçalves, Gonçalo; Antunes, António Pais (2011): Optimal Location of Charging Stations for Electric Vehicles in a Neighborhood in Lisbon, Portugal. In: *Transportation Research Record: Journal of the Transportation Research Board* 2252 (1), S. 91–98.
- [17] Chen, Nan; Gardner, Lauren; Duell, Melissa; Waller, S. Travis (2014): Evaluating Location Alternatives for Electric Vehicle Re-charging Infrastructure Using a Distance Constrained Equilibrium Assignment Model. (Transportation Research Board 93rd Annual Meeting; Jan 12-16, 2014; Washington DC): The Transportation Research Board (TRB). Online verfügbar unter <https://trid.trb.org/view/1288010>.
- [18] Funke, Simon Árpád; Gnann, Till; Plötz, Patrick (2015): Addressing the Different Needs for Charging Infrastructure: An Analysis of Some Criteria for Charging Infrastructure Set-up. In: Walter Leal Filho und Richard Kotter (Hg.): *E-mobility in Europe. Trends and good practice*. Cham: Springer International Publishing (Green energy and technology), S. 73–90.
- [19] Triebke, Henriette; Göhler, Georg; Wagner, Sabine (2016): Data analysis of PEV charging events in rural and business environments – a load behaviour comparison. In: 13. Symposium on Hybrid and Electric Vehicles, S. 83–102. Online verfügbar unter https://www.muse.iao.fraunhofer.de/content/dam/iao/muse/de/documents/projekte/Data_analysis_of_PEV_charging_events.pdf.
- [20] Fournier, Guy; Baumann, Manuel; Gasde, Johannes; Kilian-Yasin, Katharina (2018): Innovative mobility in rural areas - the case of the Black Forest. In: *International Journal of Automotive Technology and Management (IJATM)* 18 (3).
- [21] Gitanjali; Sehra, Sukhjit; Singh, Jaiteg (2013): Policy Specification in Role based Access Control on Clouds. In: *International Journal of Computer Applications* 75 (1), S. 39–43.
- [22] Wagner, Sebastian; Brandt, Tobias; Neumann, Dirk (2016): In free float: Developing Business Analytics support for carsharing providers. In: *Omega* Volume 59, Part A (March), S. 4–14.
- [23] Fan, Hongchao; Yang, Bisheng; Zipf, Alexander; Rousell, Adam (2016): A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. In: *International Journal of Geographical Information Science* 30 (4), S. 748–764.
- [24] Barron, Christopher; Neis, Pascal; Zipf, Alexander (2014): A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. In: *Transactions in GIS* 18, S. 877–895.
- [25] Ji, Yanjie; Wei, Deng; Wang, Wei; Lui, Guangrong (2007): Two-Phased Parking Choice Model for Pre-trip Parking Guidance System. In: Transportation Research Board 86th Annual Meeting (Washington DC, USA; Jan 21-25, 2007).
- [26] van der Waerden, Petrus; Timmermans, Harry; Bruin-Verhoeven, Marloes de (2017): Car drivers' characteristics and the maximum walking distance between parking facility and final destination. In: *Journal of Transport and Land Use (JTLU)* 10 (1).
- [27] Yang, Yong; Diez-Roux, Ana V. (2012): Walking Distance by Trip Purpose and Population Subgroups. In: *American Journal of Preventive Medicine* 43 (1), S. 11–19.
- [28] Wagner, Sabine; Klausmann, Florian; Lange, Ralph; Göhler, Georg; Triebke, Henriette; Schmidt, Sarah; Edel, Fabian (2016): PEV-charging in rural and business environments. A user behaviour and demand comparison. In: Conference on Future Automotive Technology, CoFAT 2016, Focus Electromobility; Veranstaltungsforum Fürstenfeld, 03.-04. Mai 2016.
- [29] Mitchell, Claire J.A. (2004): Making sense of counterurbanization. In: *Journal of Rural Studies* 20 (1), S. 15–34.



Dr. Javier Valdes

Dr. Javier Valdes is senior scientist at DIT. He studied economics at the Universidad Autónoma de Madrid, holds a PhD from the University of Ferrara and the University of Parma and a Master of Economics, Law and Business from the Université Grenoble Alpes. His research focuses on applied economics, and spatio-temporal analysis with a focus on energy economics and economics of innovation.

Dr. Javier Valdes ist wissenschaftlicher Mitarbeiter an der THD. Er studierte Wirtschaftswissenschaften an der Autonomen Universität Madrid, erwarb seinen Doktorgrad (PhD) an den Universitäten von Ferrara und Parma sowie einen Master in Wirtschaftswissenschaften, Jura und Betriebswirtschaft an der Universität Grenoble Alpes. Im Mittelpunkt seiner Forschungsaktivität stehen Themen der angewandten Wirtschaftswissenschaft sowie die räumlich-zeitliche Analyse mit einem Schwerpunkt auf Energiewirtschaft und Innovationsökonomik.

Contact / Kontakt

✉ javier.valdes@th-deg.de



Jane Wuth, M.A.

Jane Wuth studied International Economics and Business with the main focus on Behavioral Economics. Since 2018, she has been working as a researcher at the Institute of Applied Informatics, DIT's Technology Campus in Freyung. She is currently involved in the Interreg-funded research project eRoad: Pisek-Deggendorf where her focus lies on the user-centered placement of charging stations for electric vehicles.

Jane Wuth studierte an der Universität Passau International Economics and Business mit dem Schwerpunkt auf Verhaltensökonomie. Seit 2018 ist sie wissenschaftliche Mitarbeiterin am Institut für angewandte Informatik, dem Technologie Campus Freyung der Technischen Hochschule Deggendorf (THD), und arbeitet im Interreg-finanzierten Forschungsprojekt eRoad: Pisek-Deggendorf. Schwerpunktmäßig beschäftigt sie sich hierbei insbesondere mit der nutzerbezogenen Platzierung von Ladestationen für Elektroautos.

Contact / Kontakt

✉ jane.wuth@th-deg.de



Prof. Dr. Roland Zink

Roland Zink received his doctorate at the University of Passau on the subject of the transition of the energy sector and the designing of a sustainable energy supply under a spatial perspective. Since 2013, he has been Professor of Spatial Sciences and Information Systems at DIT and scientific head of the research group “Applied Energy Research” at the Technology Campus Freyung, as well as spokesman of the research group “Energy Systems and Technologies”. In addition to further developing spatio-temporal energy models, his current research focuses on innovative opportunities for public participation in spatial planning processes through new media.

Roland Zink promovierte an der Universität Passau zum Thema der Transformation des Energiesektors und der Gestaltung nachhaltiger Energieversorgungsstrukturen aus einer räumlichen Perspektive. Seit 2013 ist er Professor für Raumwissenschaften und Informationssysteme an der THD und wissenschaftlicher Leiter der Forschungsgruppe „Angewandte Energieforschung“ am Technologie Campus Freyung sowie Sprecher der Forschungsgruppe „Energiesysteme und Technologien“. Neben der weiteren Entwicklung raum-zeitlicher Energiemodelle konzentriert sich seine aktuelle Forschungstätigkeit auf innovative Partizipationsmöglichkeiten von Akteuren in räumlichen Planungsprozessen durch neue Medien.

Contact / Kontakt

✉ roland.zink@th-deg.de



Sebastian Schröck

Sebastian Schröck is a technical specialist at DIT’s Institute of Applied Informatics (Technology Campus Freyung). In 2014, he received his Bachelor of Engineering degree in Electrical Engineering and Information Technology with a focus on automation technology from DIT. Ever since then, he has been developing software solutions for spatial modeling at TC Freyung.

Sebastian Schröck ist technischer Mitarbeiter am Institut für angewandte Informatik (Technologie Campus Freyung) der THD. 2014 schloss er sein Elektro- und Informationstechnik-Studium (B. Eng.) mit dem Schwerpunkt Automatisierungstechnik an der THD ab und ist seitdem in der Entwicklung von Softwarelösungen zu räumlicher Modellierung tätig.

Contact / Kontakt

✉ sebastian.schroeck@th-deg.de



Matthias Schmidbauer

Matthias Schmidbauer studies in the Bachelor degree program “Renewable Resources” at the Technical University of Munich. In September and October 2018, he did a research internship in the Interreg project eRoad: Pisek-Deggendorf at DIT’s Technology Campus Freyung. From November 2018 until January 2019, he worked as a student assistant in DIT’s research group “Applied Energy Research” at TC Freyung.

Matthias Schmidbauer studiert im B.Sc.-Studiengang „Nachwachsende Rohstoffe“ an der Technischen Universität München. Im September und Oktober 2018 absolvierte er ein Forschungspraktikum in dem Interreg-Projekt eRoad: Pisek-Deggendorf der THD am Technologie Campus Freyung. Im Anschluss war er von November 2018 bis Januar 2019 als studentische Hilfskraft in der Forschungsgruppe „Angewandte Energieforschung“ der THD beschäftigt.