This is the accepted version of Malkus et al., "An OpenStreetMap-based approach for generating capacity-restricted POIs for activity-based travel demand modeling" published by Elsevier B.V. in Procedia Computer Science 238 (2024) 420-427. Available online via https://doi.org/10.1016/j.procs.2024.06.043.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 15th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 23-25, 2024, Hasselt, Belgium

An OpenStreetMap-based approach for generating capacity-restricted POIs for activity-based travel demand modeling

Jan-Lukas Malkus[a,*], María López Díaz[a], Alain Schengen[a], Tudor Mocanu[a], Martin J. Kühn[b]

[a]*Institute of Transport Research, German Aerospace Center, Rudower Chaussee 7, 12489 Berlin, Germany*
[b]*Institute for Software Technology, German Aerospace Center, Linder Höhe, 51147 Cologne, Germany*

Abstract

The generation of Points of Interest including capacities for travel demand models is usually costly and laborious. Recently, several methods have been developed to automate this process based on Volunteered Geographic Information data as well as to validate the corresponding results. The methodology presented in this paper takes on the task of setting up such a workflow on a large scale, including capacities and all common activity types, by making use of OpenStreetMap data. As a result, a Points of Interest dataset could be created for the whole of Germany, which maps around 24 million destinations including capacities divided into seven activity types. An exemplary validation based on the generated workplace and shopping locations as well as capacities indicates a varying quality of results depending on the respective regional type. The discussion additionally concludes that assignment of several activities to one location presents challenges that should be addressed in future approaches.

Keywords: OSM; Points of Interest; travel demand modeling; mobility; capacities

1. Introduction

Travel demand models are an important tool to support transportation planning and policy makers. More recently, agent-based travel demand models have been employed for purposes beyond infrastructure and transport planning, e.g. to provide input for epidemiological models, as observed during the COVID-19 pandemic [1]. The quality of travel demand models depends, amongst other factors, on the accuracy, completeness and level of detail of the input data. Data requirements vary depending on the type of model, however, certain categories of input data (e.g. population, travel behavior, network) are relevant for nearly all modeling approaches [2].

One important category of input data for travel demand models is land use data, which is particularly relevant for the trip distribution and destination choice step. Land use data can reflect the availability, attraction potential and

* Corresponding author. Tel.: +49-30-67055-8179
E-mail address: jan-lukas.malkus@dlr.de

capacity of specific destinations, such as workplaces, education, shopping or leisure facilities. In macroscopic travel demand models, aggregated metrics (e.g. total shopping area or total number of employees) are required at the level of traffic analysis zones. In more detailed microscopic models however, it is necessary to include each individual location with its exact position and other relevant attributes, such as maximum capacity. This is the total number of persons that can visit the destination before it becomes overcrowded.

Feasible origins for such detailed location data involve commercial sources, public authorities, surveys or open data. Though, commercial data can be expensive, while (open) government data is limited to narrow topics, and surveys can only concentrate on relatively small study areas or time slots. Furthermore, gathering, harmonizing and processing data from several sources can be extremely time-consuming. For instance, [3] present the development of a microscopic travel demand model for a specific region in Germany. Most of the activity locations for this model are sourced from a commercial data provider, while others are obtained from public authorities and, to a lesser extent, OpenStreetMap (OSM). However, this preparation process is both time- and cost-intensive, even for small regions. A contrasting method, although less precise yet notably more straightforward, is presented in [4]. Here, an open scenario for the city of Berlin is generated, intended for use with the agent-based travel demand software MATSim. Work locations are generated selecting a random coordinate within the zone assigned to the agent, and it is retained only if it aligns with the appropriate land use.

Using OSM data to depict activity locations presents several potential advantages, given that it is free-of-charge, world-wide and uniform. It is therefore not surprising that the use of OSM as data source for travel demand models is widespread, not only its road network for routing but also its Points of Interest (POIs) for destination choice. For instance, in [5], they use OSM to identify activity locations for travel demand models by mapping activity types to OSM tags. Attraction factors are determined through trip generation, and then converted to daily capacity values. After plausibility checks with official data, they find OSM data quality suitable, sometimes even more up-to-date depending on the activity type. In [6], a method to generate a POI dataset from OSM is presented. The dataset is available and could be used in general for urban science research, not for travel modeling only. It covers Europe and is simplified for use on less powerful computers.

OSM, relying on Volunteered Geographic Information (VGI), often sees errors in completeness, tagging or location. Consequently, studies aim to implement plausibility checks to assess the reliability of this data as a source for microscopic travel demand modeling or other applications. The authors of [7] evaluate a subset of OSM POIs by comparing them with data from Foursquare, another VGI. Despite data imperfections, it deems the dataset suitable for research purposes without major issues. In [8], OSM data's suitability for microscopic destination choice models is evaluated. Two main analyses are conducted: a temporal comparison across years and a scrutiny of OSM versus official data for education and work locations. While the results for educational locations are deemed reliable, many work locations are missing in the OSM dataset. In their research, the authors of [9] compare OSM POIs with manually collected data in 49 German areas, focusing on shopping and private businesses. Their findings reveal a lack of POIs in OSM for both activity types, with a more significant gap for private businesses than shopping locations. Furthermore, urban areas show a higher data completeness.

The aim of the methodology presented in this paper is to (semi-)automatically generate a large scaled (in this use case Germany) POI dataset from OSM data that is provided with capacities and covers the usual activity types (in this use case: based on the German National Transport Model *DEMO* [10]). The remaining part of the paper is structured as follows. After this introduction, the second part presents the methodology, focusing on the data extraction and filtering as well as the assignment of capacities to the POIs. The third section is concerned with the results and their validation. Findings are being discussed in the fourth section. Finally, the conclusion summarizes the paper and provides insights for future research.

2. Methods

The structure of OSM data distinguishes between three element types: nodes, ways, and relations. Nodes represent points in space, while ways represent linear or planar objects (start point equals end point). For more complex objects consisting of several parts or areas with recesses, the relation element type is used. In addition to these geometric attributes, each object in OSM also has an (element type-specific) ID. Furthermore, the attributive differentiation of objects is carried out by the help of key-value pairs (KVPs), which can be freely assigned, but which are generally

based on the OSM Wiki[1]. The primary map features are generally used for basic categorization [11]. For further differentiation, other (partially map feature-specific) KVPs are then usually assigned.

2.1. Data extraction

The english OSM Wiki recognizes 29 such (primary map feature) keys, of which the following ten were selected because they are not primarily used to map (transport) infrastructures (such as aerialway, aeroway, highway, public transport, railway, route, power, telecom, waterway), the landscape (geological, natural), various types of borders (barrier, boundary, place) or other objects that are not suitable as POIs or can only be used in conjunction with other primary map features (emergency, man_made, sport):

- amenity
- building
- craft
- healthcare
- landuse
- leisure
- office
- shop
- historic
- tourism

For the extraction, a Germany-wide osm.pbf file (current as of 2023-07-04) was obtained from *Geofabrik* [12] and processed using the Python library *PyOsmium* [13], extracting all objects (nodes, ways, and relations) from the pbf file that had at least one of the aforementioned map feature keys. For each object, the geometry (for linear or areal objects, the centroid), OSM-ID, capacity (if given and readable as an integer) and (primary) map feature KVP were extracted. For objects that had several map feature keys, the first one listed was used, except for key equal to building, as the values associated with this usually provide less information than those of the others, so that in such cases the second listed primary map feature KVP was used. Objects with invalid geometry were not extracted. The result set totalled 88, 077, 367 objects and was extracted in a run-time of around 3h 35min on a personal laptop.

2.2. Data filtering

Due to the free assignment of map feature values to OSM-objects and the thematic broadness of the map feature KVPs themselves, leading to the occurrence of less frequent values and representation of all kinds of real-world objects under one label, the result set included a large number of objects without relevance as destinations for travel demand modeling as well as data points for which the extraction would involve too much effort in contrast to the additional quantity of results obtained. In fact, there are only a small number of map feature KVPs that represent the majority of all objects (according to the selected map features). The first 1 % (n=75) of the most frequently used (map feature) KVPs already represent 97.68 %, the first 5 % (n=375) even 99.83 % of all initially extracted objects. For these reasons, the KVPs were sorted in descending order according to their frequency, all KVPs with a frequency $\geq$ 100 (n=798) and 76 others from the frequencies $\geq$ 10, whose values were previously used in the list with other keys, were identified. Then, the following attributes were assigned manually and based on expert knowledge to them:

- whether objects corresponding to the respective KVP should be used as POIs
- up to three activity types from the set: work, education, shopping, leisure, private business, residential, other
- the binary information as to whether the real-world object is (primarily) an outdoor activity location

resulting in a scheme as shown in Table 1.

The table was imported into the database and based on its Boolean *usage* variable, the extracted locations were then filtered down to 18, 765, 478 objects in a run-time of around 1min 20sec (on the database).

[1] https://wiki.openstreetmap.org/wiki/Main_Page (accessed on: 2023-12-15)

Table 1. Table for filtering, assigning activity types and a boolean outdoor attribute (Exemplary extract; in descending [frequency] order)

OSM primary map feature Key-Value-Pair ({*Key:Value*})	Usage (*Boolean*)	Primary activity type	Secondary activity type	Tertiary activity type	outdoor activity (*Boolean*)	frequency (*n*)
{building:yes}	false					51,888,739
{building:house}	true	residential			false	5,608,155
{landuse:forest}	true	leisure	work		true	1,118,187
{building:terrace}	false					583,320
{building:commercial}	true	shopping	work	education	false	203,714
{amenity:recycling}	true	private business	work		true	91,987
{building:greenhouse}	true	work			false	79,136
{building:cabin}	true	other			false	62,167
{amenity:library}	true	education	work	leisure	false	6,867
{shop:vacant}	false					5,025

2.3. Deriving capacities

As only a very small proportion (around 0.6 % based on the original extract) of the OSM-objects were provided with a capacity (readable as an integer), the so-called Bosserhoff table was used to derive capacities. This source originates from urban land-use planning in the context of traffic volumes and can be obtained for a fee [14], which is why it cannot be published freely. It holds an assumed default capacity for both employees (workplaces) and customers/users for the already shown activity types, further subdivided using two consecutive subcategories per main type (e.g. Shopping - Specialist retailers - Bookstore).

Regarding the activity-type-attribution (see Table 1), in addition to the already assigned activity types (no matter if primary, secondary or tertiary), those refined sub and sub-sub-divisions were additionally assigned per KVP. There were 106 different category-/sub-/sub-sub-category combinations to choose from (based on the Bosserhoff table). For KVPs that were originally assigned several (up to three) activities, this differentiation was made for all of them. For the sake of clarification, let us give an example: For the KVP *amenity:library* (see Table 1), the primary activity type (education) was refined by sub-category 'other', sub-sub-category 'library', the secondary activity type (work) was not differentiated, as the Bosserhoff table does not either and the tertiary activity type (leisure) was supplemented by the sub-category 'event', sub-sub-category 'culture'.

Two changes were made compared to the Bosserhoff table (from 2017) itself: Firstly, the default worker capacity of the activity category work (no sub-categories; used for those KVPs to which work was assigned as the primary activity) was set to 11. This value corresponds to the rounded average, weighted across all destinations, of all those destinations to which work was assigned as a secondary or tertiary activity. In addition, the default worker capacity was set to 1 for the activity-category 'leisure', sub-category 'recreation', sub-sub-category 'forest'.

Using the differentiated activity categories as key, the locations (represented by their KVPs) could be joined with the Bosserhoff table, and the assumed capacities from the latter could be added to the former. Objects whose KVP was assigned to several activities, were mapped several (up to three) times (activity-distinct and therefore with different capacities). Thus, the result set totalled 23,834,222 objects.

3. Results and validation

3.1. Results

The aforementioned key figures of 18,765,478 objects before and 23,834,222 objects after multiple mapping of locations with several activity purposes, result in 5,068,744 objects (around 21 % of the results dataset) for mapping secondary (n=4,456,93) and tertiary (n=611,809) activities at the same location. Speaking in capacities, these make up for 64,238,540 (secondary), respectively 25,486,947 (tertiary) and combined for around 6 % of the total capacity (n=1,507,081,047) of all activity destinations. This leads to average capacities of approximately 76 for primary, 14 for secondary, and 42 for tertiary activity purposes, respectively.

Differentiation by user and worker capacity shows low shares (below 1 %) of worker capacities in total capacity for primary and tertiary activity locations purposes. For secondary purposes though, there are around 6.7 worker capacities per user capacity. The average capacity per destination with work-purpose (worker capacity) is 13.73 while being 75.60 for destinations with non-work purpose (user capacity). The share of work locations in all locations is 4.38 % for primary purposes, 87.81 % for secondary purposes and 0.22 % for tertiary purposes.

Table 2. Absolute and relative counts of destinations and capacities including average capacities depending on activity type

Activity type	Destination count (n)	Destination share (%)	Capacity count (n)	Capacity share (%)	Average capacities (n)
education	976,543	4.10	77,132,118	5.12	78.98
leisure	3,926,773	16.48	860,903,728	57.12	219.24
other	344,706	1.45	1,723,530	0.11	5.00
private business	747,342	3.14	155,345,551	10.31	207.86
shopping	746,655	3.13	223,617,250	14.84	299.49
residential	12,332,878	51.74	123,328,780	8.18	10.00
work	4,759,325	19.97	65,030,090	4.31	13.66

A differentiation by activity types (see Table 2) shows heavily varying shares of destinations and capacities, both among as well as within the activity type categories. This is also reflected in the wide range of average capacities per destination activity type, which extends from 5 (other) to just under 300 (shopping).

3.2. Validation

The validation of locations and corresponding capacities derived from OSM data is no trivial task, since there are no other readily available data sources that can be considered the ground truth and to which we should match. Note that otherwise, there would be no need for the proposed OSM-based approach. Nevertheless, there are several validation approaches that are possible and feasible:

- Aggregation and comparison with (official) data available at a (spatially) aggregated level
- Comparison of individual features for small-scale samples
- Comparison of results for different region types

A comprehensive validation of all results goes beyond the scope of this paper. Instead, in this section we will focus only on certain aspects that should be regarded more as examples.

Fig. 1 shows a comparison of the OSM-based workplace capacity to the official employment data by municipality. Displayed is the ratio between the OSM-based results and the statistical data[2]. As can be seen, the OSM-based values are usually higher, with the total workplace capacity (see Table 2) being almost double compared to the official data. One explanation for this mismatch is that the official data only contains employees subject to social insurance contributions, excluding roughly 30 percent of workers (e.g. freelancers, public officials). Another interesting aspect that becomes apparent in Fig. 1 is that our approach seems to overestimate the employment capacity especially in rural areas, whereas urban areas are probably even slightly under-estimated. This is an indication that there are either too many locations in rural areas (possibly due to doubling) or that the values from the Bosserhoff tables are too high (to cover for rural regional types).

Fig. 2 shows the capacities of shopping locations in relation to the population size of the corresponding municipality. The expectation here is that there will be roughly the same shopping capacity available in both urban and rural regions, since the shopping demand per person is roughly the same across Germany. The finer-grained analysis at municipality level shows a high variability across the country, with roughly 20 percent of municipalities having no shopping locations on their territory at all. The majority of the municipalities without shopping locations are ones

[2] BBSR Bonn 2023: Sozialversicherungspflichtige am Arbeitsort 2020; https://www.inkar.de/; dl-de/by-2-0: http://www.govdata.de/dl-de/by-2-0

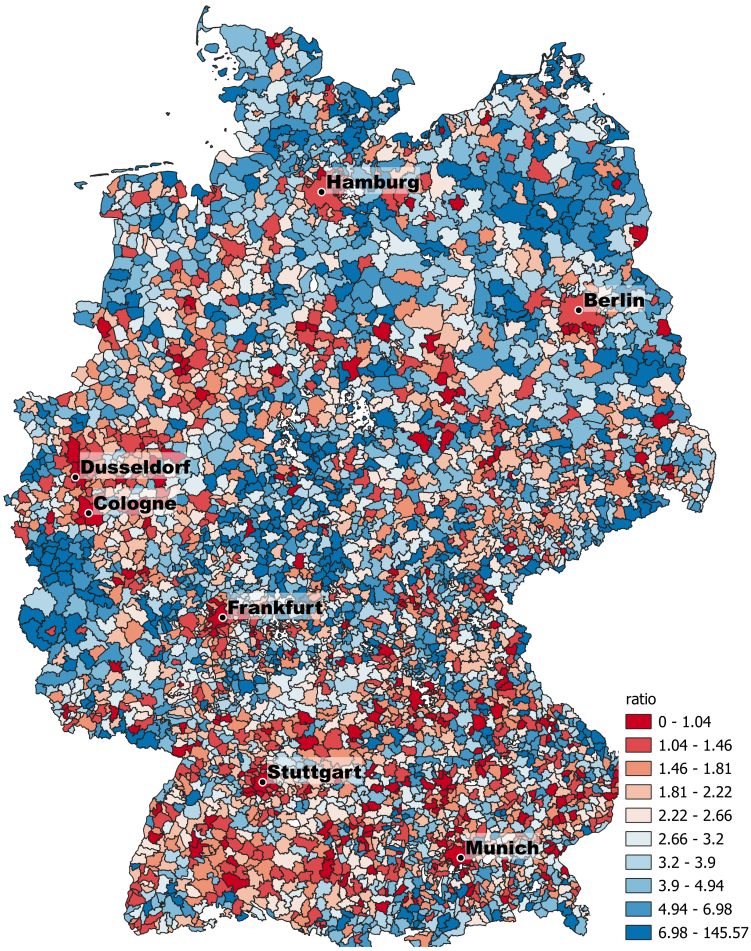

Fig. 1. **Ratio of OSM-based workplace capacity to official employment data**: by municipality.

with small areas in rural regions, which is plausible, but requires further, more detailed analyses. Nevertheless, aggregating at county level shows that these rural regions have relative capacities comparable to larger cities, which seems to validate the original assumption.

4. Discussion

While mapping possibly given secondary and tertiary activity purposes in the same location, the corresponding capacities were not down-weighted according to the lower importance/attractiveness in relation to the primary activity purposes one. Such a weighting would have to be carried out depending on the objects OSM map feature KVP, as well as all assigned activity purposes including sub- and sub-sub-categorisation, taking into account the purposes ranking. In addition, the regional type (e.g. RegioStaR for Germany), local culture and national territory of the location may also have a relevant influence, as types and frequency of use of the same location type tend to differ according to the aforementioned factors.

But even without such weighting, the average capacity for secondary and tertiary purposes is lower then for primary ones. Though, this is only based on the assignment of OSM map feature KVPs and the ranking for several trip purposes, resulting in a higher average capacity for tertiary compared to secondary purpose depictions. Based on the differences between the average capacities of work and non-work related purposes, as well as the share of work-

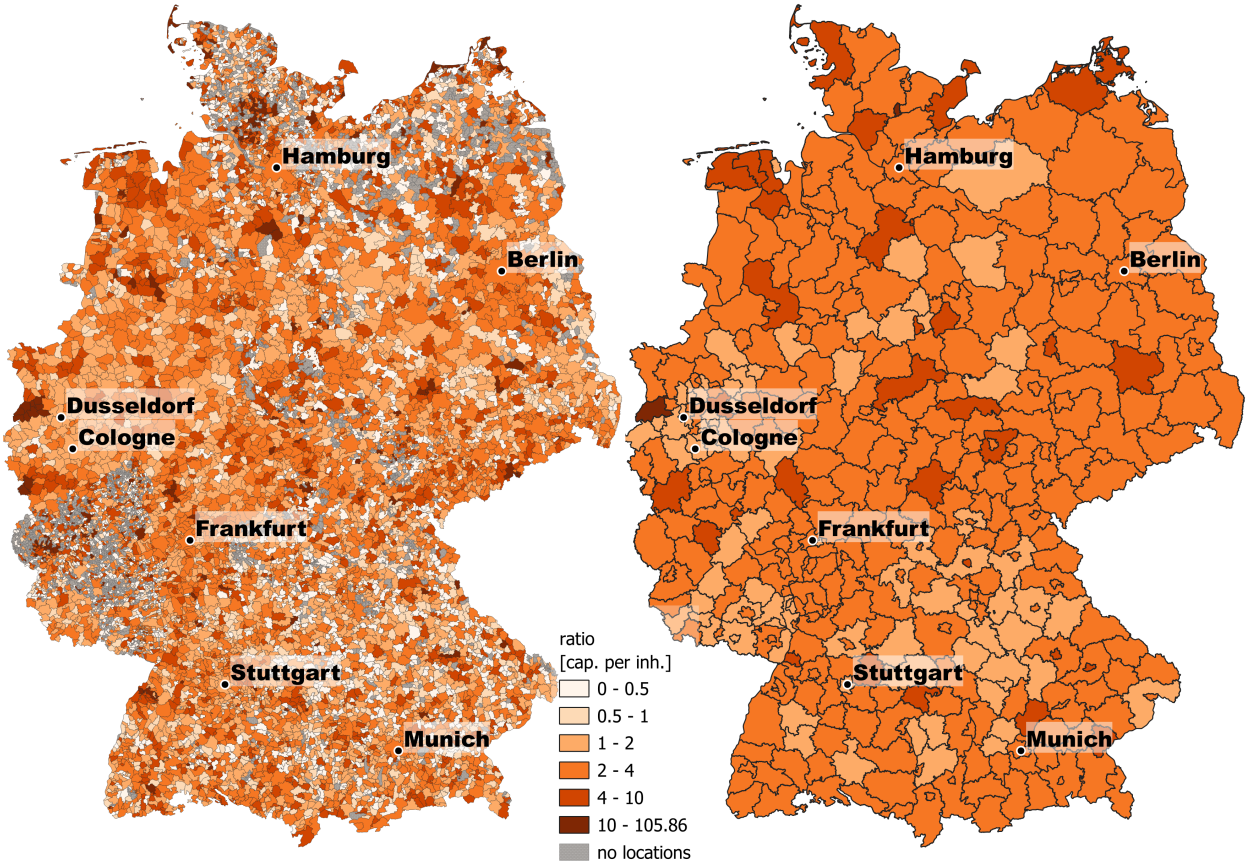

Fig. 2. **Ratio of OSM-based shopping capacity to inhabitants**: by municipality (left) and county (right).

purpose locations in all locations per purpose rank (primary to tertiary), this can be explained by the high share (87.81 %) of work-purposes assigned as secondary location purpose.

Also, the assignment and order of the activity categories (including subcategories) to OSM KVPs is subjective and therefore strongly dependent on the imagination, perception and prior knowledge of the person making the assignment. In addition, for non-specific KVPs (e.g. {*building:commercial*}), categorisation in the Bosserhoff (sub-/sub-sub-) categories can only represent a first approximation. Advanced methods need to be developed in order to achieve a better differentiation and thus categorization of such locations on the basis of additional KVPs, for example. Another method is needed to reduce real-world objects, which are mapped several times in OSM (e.g. as a node and as a way), to a single representative.

5. Conclusion

In this paper we have demonstrated the generation of locations, including their capacities, for all common activities for subsequent usage in both microscopic and macroscopic travel demand models using OSM data. As information on the capacities of locations is available only very sparsely in OSM, we employed default capacities obtained by a widely-used German guideline.

A limitation of the presented study is the rather heterogeneous structure of data quality, as OSM is based on individual user input. For instance, the spatially heterogeneous structure of workplace capacities when compared to official employment data might be explained by better OSM quality in cities. However, official employment data does not represent true capacity data and further research is required here. While the structure of, e.g., shopping capacities

on municipality level seems highly fragmented, the picture on county level looks well equalized and might better represent the interdependence of rural areas. Due to the use of the Bosserhoff table, which is subject to a charge, the method presented here cannot be implemented completely cost-neutrally. The down-weighting of non-primary activity purposes, the differentiation of less specific KVPs, and the reduction of objects mapped multiple times in OSM were identified as topics for future research and development of the methodology presented.

The collection of data as presented here is cumbersome and, to the best of our knowledge, simply not available to the research community yet. The data obtained from OSM is suitable for subsequent use as it builds an excellent and cost-effective base for further development. In order to use it in microscopic and macroscopic travel demand models, an important next step requires a comparison or fitting of the overall destination capacities to the total number of trips generated in either the entire study area or in smaller sub-areas. This step is necessary to ensure that capacity restrictions can be properly implemented in such microscopic agent-based models.

Acknowledgements

The authors JLM, MLD, AS, TM and MJK have received funding from the German Federal Ministry for Digital and Transport under grant agreement FKZ19F2211A. The author MJK was also funded by the Initiative and Networking Fund of the Helmholtz Association (grant agreement number KA1-Co-08) for the project 'Integrated Early Warning System for Local Recognition, Prevention, and Control for Epidemic Outbreaks' (LOKI).

References

[1] Müller, Sebastian A., Michael Balmer, William Charlton, Ricardo Ewert, Andreas Neumann, Christian Rakow, Tilmann Schlenther, and Kai Nagel. (2021) "Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data" *PLoS ONE* (16). DOI: 10.1371/journal.pone.0259037

[2] Kagho, Grace O., Milos Balac, and Kay W. Axhausen. (2020) "Agent-Based Models in Transport Planning: Current State, Issues, and Expectations" *Procedia Computer Science* **170** 726-–732. DOI: 10.1016/j.procs.2020.03.164

[3] von Schmidt, Antje, María López Díaz, and Alain Schengen (2021) "Creating a Baseline Scenario for Simulating Travel Demand: A Case Studyfor Preparing the Region Test Bed Lower Saxony, Germany" *The Thirteenth International Conference on Advances in System Simulation(SIMUL), IARIA, Think Mind* 51–57, ISBN: 978-1-61208-898-3

[4] Ziemke, Dominik, Ihab Kaddoura, and Kai Nagel. (2019) "The MATSim Open Berlin Scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data" *Procedia Computer Science, The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops* **151**, 870–877. DOI: 10.1016/j.procs.2019.04.120

[5] Klinkhardt, Christian, Tim Woerle, Lars Briem, Michael Heilig, Martin Kagerbauer, and Peter Vortisch. (2021) "Using OpenStreetMap as a Data Source for Attractiveness in Travel Demand Models" *Transportation Research Record* **2675** (8): 294–303. DOI: 10.1177/0361198121997415

[6] McCarty, Dakota Aaron, and Hyun Woo Kim. (2023) "A standardized European hexagon gridded dataset based on OpenStreetMap POIs." *Data in Brief* **49**. DOI: 10.1016/j.dib.2023.109315

[7] Zhang, Liming, and Dieter Pfoser. (2019) "Using OpenStreetMap point-of-interest data to model urban change—A feasibility study" *PLoS ONE* **14** (2). DOI: 10.1371/journal.pone.0212606

[8] Briem, Lars, Michael Heilig, Christian Klinkhardt, and Peter Vortisch. (2019) "Analyzing OpenStreetMap as data source for travel demand models A case study in Karlsruhe" *Urban Mobility – Shaping the Future Together* **41** 104–112. DOI: 10.1016/j.trpro.2019.09.021

[9] Klinkhardt, Christian, Fabian Kuehnel, Michael Heilig, Sven Lautenbach, Tim Woerle, Peter Vortisch, and Tobias Kuhnimhof. (2023) "Quality Assessment of OpenStreetMap's Points of Interest with Large-Scale Real Data" *Transportation Research Record* **2677** (12) 661–674. DOI: 10.1177/03611981231169280

[10] Winkler, Christian and Tudor Mocanu (2017) "Methodology and Application of a German National Passenger Transport Model for Future Transport Scenarios". In: *Proceedings of the 45th European Transport Conference*. European Transport Conference, 4.-6. Oct. 2017, Barcelona, Spain.

[11] OpenStreetMap. (2023) "Map features" *Website*. URL: https://wiki.openstreetmap.org/wiki/Map_features (accessed on: 2023-12-15)

[12] Geofabrik. (2023) "OpenStreetMap Data Extracts" *Website*. URL: https://download.geofabrik.de/ (accessed on: 2023-12-15)

[13] PyOsmium. (2023) "PyOsmium – Python bindings to Osmium Library" *Website*. URL: https://osmcode.org/pyosmium/ (accessed on: 2023-12-15)

[14] Bosserhoff, Dietmar. (2023) "Programm Ver_Bau: Abschätzung des Verkehrsaufkommens durch Vorhaben der Bauleitplanung mit Excel-Tabellen am PC" *Website*. URL: https://www.dietmar-bosserhoff.de/Programm.html (accessed on: 2023-12-15)