LEVERAGING OPENSTREETMAP INFORMATION TO IDENTIFY CLUSTER CENTERS IN AGGREGATED MOVEMENT DATA

Maylin Wartenberg[1], Luca Marco Heitmann[1], and Marvin Auf der Landwehr[2]

[1]Hochschule Hannover, Ricklinger Stadtweg 120, 30459 Hannover, GERMANY
[2]FehrAdvice & Partners AG, Binzmühlestrasse 170, 8050 Zurich, SWITZERLAND

ABSTRACT

Aggregated movement data is widely used for traffic simulations, but privacy constraints often limit data granularity, requiring the use of centroids as cluster representatives. However, centroids might locate cluster centers in contextually irrelevant areas, such as an open field, leading to inaccuracies. This paper introduces a method that leverages an aggregation of points of interest (POIs) such as bus stops or buildings from OpenStreetMap as cluster centers. Using trip data from a suburban region in Germany, we evaluate the spatial deviation between centroids, POIs, and real trip origins and destinations. Our findings show that POI-based centers reduce spatial deviation by up to 46% compared to centroids, with the greatest improvements in rural areas. Furthermore, in an agent-based mobility simulation, POI-based centers significantly reduced travel distances. These results demonstrate that POI-based centers offer a context-aware alternative to centroids, with significant implications for mobility modeling, urban planning, and traffic management.

1 INTRODUCTION

Transportation is a fundamental part of modern society, playing a key role in solving ecological, economic, and social challenges. With the ongoing expansion of mobility and the increase in traffic volumes, the need for efficient and sustainable transportation solutions grows. The development and deployment of such solutions frequently require comprehensive pilot testing, a process that is both financially and logistically onerous. Consequently, advanced computational modeling and traffic simulation, driven by the analysis of movement and traffic data, are gaining greater significance (Alghamdi et al. 2022).

These data-driven approaches rely on a variety of data sources, including surveys, GPS tracking, and mobile network data (Phithakkitnukoon 2023). Realistic simulations require correct and contextually relevant data, which is difficult to obtain, especially in light of growing privacy concerns (Krumm 2007). Privacy restrictions often necessitate aggregation, resulting in clusters represented by centroids to protect individual identities (de Montjoye et al. 2018). However, these aggregated datasets present significant challenges for analysis and simulation, as they lack precise geographic information and contextual meaning. Consequently, the use of cluster centroids often fails to accurately represent actual movement patterns, which in turn reduces the quality, explanatory power, and predictive accuracy of simulation models. As an example, Figure 1 shows clusters where the centroids are located in a forest or an open field. For the simulation community, particularly in the field of dynamic simulation and agent-based models, this limitation is highly problematic because the spatial starting points of agents (e.g., vehicles, pedestrians) are critical for realistic modeling of movement dynamics, routing, and interaction patterns within the system. Inaccurate cluster representations propagate into system dynamics, causing biases in load distributions, unrealistic flows, and distorted behavioral patterns that affect model outcomes significantly.

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

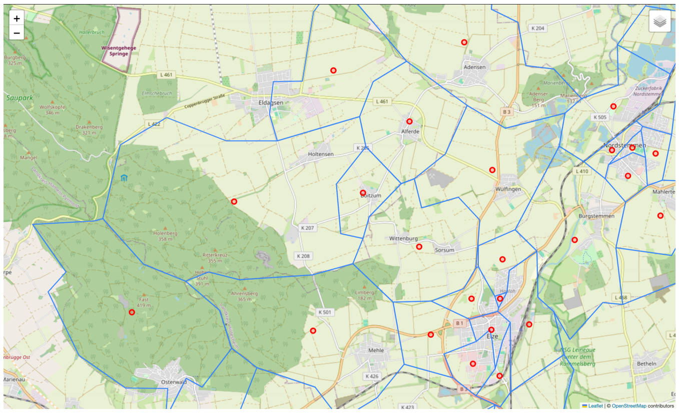

Figure 1: The map shows a portion of the study area with clusters in blue and their centroids in red. In rural areas, the centroid might be located in a forest or an open field.

This study evaluates spatial deviations between cluster centroids and actual travel routes, and explores alternative approaches based on points of interest (POIs) to improve the accuracy of cluster-based movement modeling. POIs represent specific, functionally relevant locations like buildings or bus stops (Kashian 2017). It also compares the effects of using cluster centers versus the actual trips in a simulation and measures the differences between different cluster center choices. In particular, the research addresses the following research questions:

RQ1: How well do the centroids of the clusters represent the actual travel routes in aggregated movement data?

RQ2: How can POIs be used to identify an alternative point within the cluster to better represent actual movement patterns compared to centroids?

RQ3: What are the effects on the total distance traveled in a simulation of a mobility service provider if the pickup location is the exact order location, the centroid of the order cluster, or an alternative point that serves as the center of the order cluster?

The study utilizes mobility data from Sarstedt, a sub-urban region in Germany, employing two datasets: a trip table derived from synthetic persons in a mobility model, which contains precise start and end points of trips (Senozon 2022), and a mobile network dataset that aggregates mobility flows into clusters. The initial step involves a comparison of the distances between the actual travel routes from the trip table and the centroids of the clusters in the mobile network dataset, with the objective of quantifying spatial deviations. While straight-line distances are not fully representative of human mobility patterns, they are sufficient in this study as the deviation to the actual routes is minimal.

Subsequently, an alternative point to better represent the cluster is determined using POIs from OpenStreetMap (OSM) data. To select and weigh the specific POIs, two scenarios were conducted. The first scenario aims to find a simple POI selection. This is achieved by identifying the POI tags that appear in at least 60% of all clusters. These POIs are then used without weighting to calculate the POI center of the clusters. The second scenario focuses on finding a near-optimal POI center for the study area. For this purpose, a Bayesian optimization is used to select and weigh the POI tags in a way that minimizes the distance to the actual trips.

The analysis evaluates the results by calculating the average distance between the actual start and end points from the trip table and the various cluster centers, including centroids and POI-based centers. It is important to note that the result accuracy heavily depends on the quality and completeness of the OSM data used.

In a simulation of a mobility service provider, the impact of different cluster center selection methods on total travel distance is assessed by comparing scenarios where the exact order location, the centroid, or the POI-based center serves as the designated start and end point.

This research aims to assess the approximation accuracy of POI data in representing movement patterns and explore its viability as a substitute for centroids. By addressing these questions, the study offers valuable insights for optimizing data-driven traffic analysis and introduces innovative approaches to model movement flows in privacy-sensitive contexts.

2 RELATED WORK

Existing research has largely focused on improving clustering algorithms or refining centroid placement within aggregated datasets to minimize spatial deviation (Gao et al. 2021; Thuillier et al. 2018). Common approaches include using density-based or hierarchical clustering methods to better approximate movement patterns. Yet, these methods typically rely on geometric principles and do not account for the semantic and functional structure of urban and sub-urban spaces, such as POIs, land use, or transportation hubs (Heredia et al. 2022). While some studies have explored geospatial features like land-use patterns or transportation networks as additional information sources, their integration into mobility modeling remains fragmented as well as case-specific and lacks a systematic and rigorous evaluation of their impact on simulation outcomes (Tran and Draeger 2021). Moreover, these approaches rarely address the specific needs of dynamical simulation models, where accurate initial positioning and flow distribution of agents is vital for capturing emergent phenomena like congestion, demand peaks, or service bottlenecks. Consequently, centroids continue to serve as the de facto standard for cluster representation, despite their known inability to accurately capture the spatial and semantic complexity of human movement, which poses a fundamental limitation for behaviorally accurate and policy-relevant simulations.

Notelaers et al. (2024) present a method to estimate travel demand based on POIs from OSM. This method could also be adapted to use POIs as start and end points for trips. It works by splitting the demand of a cluster into multiple smaller clusters and ultimately models trips at the POI level. It is an extensive method that shows promising results. In this paper, we aim to solve a similar problem while preserving the original clusters as they represent geographical regions that are important for the simulation and for ensuring comparability with existing methods.

A substantial body of research has been dedicated to the examination of OSM data, and the results of these studies indicate that the data can be considered as reliable (Haklay 2010). However, the presence of deviations is observed, contingent upon the geographical area under investigation (Neis et al. 2013). European regions demonstrated particularly robust coverage (Neis et al. 2013), which supports the utilization of these data for the study area in this paper.

3 MATERIALS AND METHOD

This work follows a positivist approach by assuming that the underlying movement patterns of the aggregated data can be objectively described by mathematical and statistical models.

The Design Science Research (DSR) methodology was chosen for this study due to its structured, iterative approach to solving complex, real-world problems through the development and evaluation of innovative artifacts. The core objective of this research is to address the limitations of centroid-based cluster centers in representing aggregated movement data by proposing a novel, context-aware methodology using POIs. The DSR framework aligns well with this goal by emphasizing the iterative cycles of Relevance, Design, and Rigor (Hevner et al. 2004). The Relevance Cycle ensures the research remains grounded in practical needs, linking the artifact design directly to the research questions and real-world challenges in mobility modeling. The Design Cycle facilitates the creation, testing, and refinement of the methodology for POI selection, integrating systematic evaluation to validate its effectiveness. Finally, the Rigor Cycle incorporates established scientific theories and geospatial methods, ensuring the solution is both theoretically sound and practically applicable. Unlike traditional experimental or qualitative approaches, which lack iterative refinement and artifact focus, DSR bridges the gap between theory and practice, enabling the development of a robust methodology adaptable to various environmental contexts. This structured approach ensures that the proposed POI-based methodology contributes both scientifically and practically to mobility analysis, advancing the integration of geospatial data into privacy-sensitive traffic modeling applications.

In the following parts the data sources used and the study area are described. Then the methodology is explained in three stages. Initially, the most impactful POI types are determined and two scenarios with a different selection and weighing of POI types are created. Then, for both scenarios the POI centers of all clusters within the study area are calculated. Finally, it is calculated how much the found centers of the clusters deviate from the actual trips.

3.1 Study Area and Data Sources

The study area encompasses various municipalities in central Germany, including Sarstedt, Nordstemmen, Elze, Gronau (Leine), and a portion of Hildesheim. The study area includes mainly rural and a few urban regions. The Sarstedt trip table, as presented in the 2021 Mobility Model Germany by Senozon Deutschland GmbH (Senozon 2022), is employed for the purpose of analyzing actual trips. This contains synthetic movement patterns for 25% of the population from Sarstedt, Nordstemmen, Elze, and Gronau (Leine). As it also includes trips that cross the study area, sufficient data is available for Hildesheim, which was thus integrated into the study area. The study area covers an expanse of 530 km^2. The trip table is a text file comprising data in a tab-separated values format. It encompasses trips with start and end coordinates, which are represented in the projected coordinate system EPSG:25832 (ETRS89 32N).

The second data source comprises movement data derived from mobile phone data for the same study area, which has been aggregated into clusters. Only the cluster boundaries are employed, not the entries of the movement patterns between clusters. The data set originates from Teralytics Inc. (2022) and encompasses the status from Q1 of 2022. It contains 178 clusters for the study area.

3.2 Data Preprocessing

Data preprocessing involves several steps to ensure consistency and usability of the data. First, the trips from the route table were assigned to the appropriate clusters. This is done by comparing the coordinates of the start and destination points with the cluster boundaries and determining the start and destination cluster IDs. To enable this comparison, both the coordinates of the trips and those of the cluster boundaries must be transformed into the projected coordinate system EPSG:25832. All trips for which no cluster could be identified for the specified start or destination coordinates were excluded from the analysis, namely those that fell outside the defined study area.

Furthermore, the average center is calculated for each cluster based on the mean value of the actual start and end points of trips within the cluster. This center is utilized as a reference point, which should be approached by the POI center.

3.3 Selecting POI Types

The initial step is to ascertain which POI types are optimal indicators of a point within a cluster that is situated at the shortest possible linear distance to all start or end points of the cluster's trips. In order to achieve this objective, all POIs within the study area, classified according to the following categories: aeroway, amenity, building, office, public transportation, store, sport, and tourism, are retrieved from OSM (OpenStreetMap Wiki contributors 2025). In the following steps, distance calculations will be conducted with the POIs, which necessitates the establishment of a coordinate point for each POI. To fulfill this requirement, the centroids of polygon geometries are formed and employed as the POI's coordinates. The POIs are then assigned to the corresponding cluster in which they are located based on their coordinates.

Subsequently, the distance to the average center of the corresponding cluster is calculated for each POI. The POIs are then aggregated based on their tag and the mean value of the distances to the average center is calculated. In addition, the number of different clusters in which the POI type is present and the total number of occurrences of the POI type are calculated. The resulting table is used as the basis for determining the POI types to be selected.

The POI tags, which are distributed across more than 60% of all clusters, are evaluated in two different scenarios. The 60% threshold ensures representativeness by including only those POI tags that appear in the majority of clusters, thereby capturing the most characteristic features of the study area while filtering out potential outliers. The first scenario uses the remaining POI tags without applying any weighting. Since no assumptions are made regarding specific structural characteristics of the region, such as particular amenities, the aim is to achieve a selection of POIs that is as universally applicable and region-independent as possible.

In the second scenario, Bayesian optimization using Gaussian Processes is performed to select and weight the POI tags. Bayesian optimization is particularly suitable for this problem due to its efficiency in optimizing expensive-to-evaluate functions with a limited number of iterations (Yang 2024). It effectively balances exploration and exploitation, making it ideal for finding the optimal configuration of POI tags and their weightings in a high-dimensional space. For each possible POI tag, a dimension is created to determine whether the tag is used or not, and an additional dimension is created for the weighting, ranging from 0 to 10. A function calculates the cluster centers from the configuration determined by the algorithm at each iteration. Subsequently, the average deviation from the actual start and end points of the trips is computed. This average deviation is minimized by the algorithm, allowing for the determination of a configuration that is as close as possible to the actual trips within the study area. The optimization process was configured to run with 150 iterations.

3.4 Calculating POI Centers in the Study Area

The next step is to calculate the POI centers for the clusters in the study area. This is done by first retrieving all POIs that have a tag that exists in the table created in section 3.3. The result is a set of POIs with their tags and the coordinates where they are located and the assignment to the corresponding cluster in the study area.

The POI center is then determined for each cluster by calculating the average of all POI coordinates within the cluster, taking into account the weighting in scenario 2.

3.5 Evaluation of the POI Centers in the Study Area

The POI center determined for each cluster is compared to the distance to the actual start and end points of the route table. In addition, the distance is calculated using the average center and the centroid. The results are compared, and a statement is made about the accuracy of the POI center.

4 RESULTS

The methodology was implemented in the Python programming language as a Jupyter notebook. In addition, GeoPandas, Shapely, OSMnx (Boeing 2017), GeoPy and Folium were used to process the geodata.

4.1 Overview of the Data Provided

The initial stage of the process entails the loading and pre-processing of the data. All coordinates are transformed into the projected coordinate system EPSG:25832, and the route table is filtered for trips that have a start and end point within the study area. Following this, 178 cluster boundaries and 54,965 trips were extracted following the completion of the pre-processing stage.

The mean number of start or end points per cluster is 618. This varies significantly, with clusters ranging from as few as 4 start or end points to as many as 4,210. This variation is primarily due to differences in cluster size and population density, and does not follow a clear urban-rural pattern. The area of the clusters varies considerably, from 0.0098 km^2 to 25.5709 km^2 with a mean size of 2.9792 km^2. The clusters are larger in rural regions in particular. In addition to the trip table and the cluster boundaries, all POIs within the study area are retrieved from OSM that have one of the following tags: aeroway, amenity, building, office, public_transport, store, sport, tourism. The coordinates for each POI are determined in EPSG:25832 and assigned to the corresponding cluster. There are 72,453 POIs in the entire study area. The mean number of POIs per cluster is 407, with a standard deviation of 270. The minimum number of POIs per cluster is 3, while the maximum is 1,582.

The subsequent phase involves creating a register of POIs and classifying them according to their tags. This necessitates the handling of POIs bearing multiple tags. As this applies only to 775 of the 72,453 POIs, these POIs are counted on multiple occasions, with each assigned tag. This increases the number of POIs in the study area to 73,247.

In the final step of preprocessing, the average center of each cluster is determined by calculating the mean of all start or end points of a cluster. Additionally, the geodetic distance from each POI to this average center is computed.

4.2 Selection of POI Types

First, the POIs are grouped by tag, with the mean value of the center distance, the total number of POIs for this tag, and the number of different clusters in which the POI tag occurs forming the basis for this grouping. The result is illustrated in Figure 2.

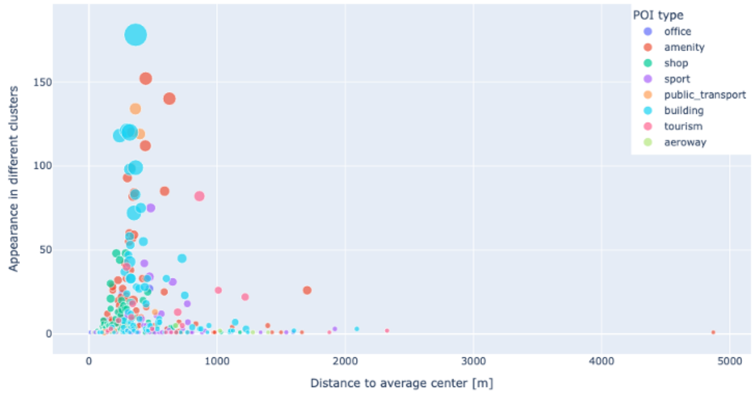

Figure 2: Scatter plot of all POI tags according to their distance from the average center (size of the dots represents the total number of POIs associated with each tag).

It can be observed that a considerable number of POIs are situated near the average center yet exhibit limited coverage across diverse clusters. Consequently, the subsequent analysis will focus on POIs that have been identified in at least 60% of the clusters within the study area. This approach enables the formulation of a comprehensive assertion that can be extrapolated to other geographical contexts.

Table 1: POI tags that have been identified in at least 60% of all clusters within the study area.

tag	Distance to average center [m]	Total number of POIs	Appearance in different clusters
building.apartments	239.69	1,933	118
building.garage	298.64	2,862	121
building.house	318.27	6,156	120
amenity.post_box	333.23	167	120
public_transport.platform	362.62	624	134
building.yes	364.43	43,509	178
public_transport.stop_position	397.43	418	119

amenity.waste_basket	439.86	512	112
amenity.parking	443.07	1,143	152
amenity.bench	627.73	1,024	140

Two scenarios are created for the selection of POIs, which are then used to calculate the POI center of the clusters. Scenario 1 uses all of the POI tags from Table 1 without any weights, while Scenario 2 uses these POI tags as input for a Bayesian optimization, as described in section 3.3, to determine a configuration of POI tags and corresponding weights that minimizes the deviation from the actual trips. Table 2 contains the result of the Bayesian optimization.

Table 2: Selection of POI tags and corresponding weights for scenario 2.

POI tag	weight
public_transport.platform	10.00
building.garage	8.79
amenity.post_box	8.24
building.apartments	3.96
building.yes	1.28
amenity.parking	1.12

4.3 Evaluation

The evaluation calculates the average distance to the actual start and end points of the route table for each cluster center. The smaller the average distance, the closer the cluster center is to the actual trips. Table 3 shows the results. For reference, the table additionally includes how far the average center is from the actual trips.

Table 3: Mean distance between the actual trips and the calculated cluster centers.

scenario	mean distance between actual start/end points and center [m]
centroid	630.21
average center	313.92
1	335.89
2	334.65

The use of cluster centroids in the study area is on average 630 m away from the actual trips, while the calculated POI center in scenario 2 is on average only 335 m away. This suggests that the use of POIs leads to a better choice of cluster center.

Scenario 2 has the lowest mean distance to of actual trips. The disadvantage, however, is that this scenario is hardly transferable to other regions. Through the optimization, the POI center was determined, which is as close as possible to the actual trips specifically for this region.

Scenario 1, which emphasizes simplicity, achieves an average deviation of 336 m—only marginally higher than the 335 m from Scenario 2. Crucially, it avoids the need for weighting POI tags, thus reducing region-specific assumptions and enhancing transferability. The distance distribution across clusters ranges from 48 m to 1,275 m, with only two clusters exceeding one kilometer due to spatially disconnected village structures. This is visualized in Figure 3.

Notably, the simulation results for Scenario 1 and Scenario 2 (Table 4) also show similar total travel distances and success rates, indicating that both approaches are functionally equivalent in practice. While Scenario 2 achieves a lower mean distance between the actual start/end points and the cluster center compared to Scenario 1, Scenario 1 offers a compelling trade-off between simplicity, robustness, and performance. Therefore, it may be the preferable option in settings where model generalizability or computational efficiency is critical.

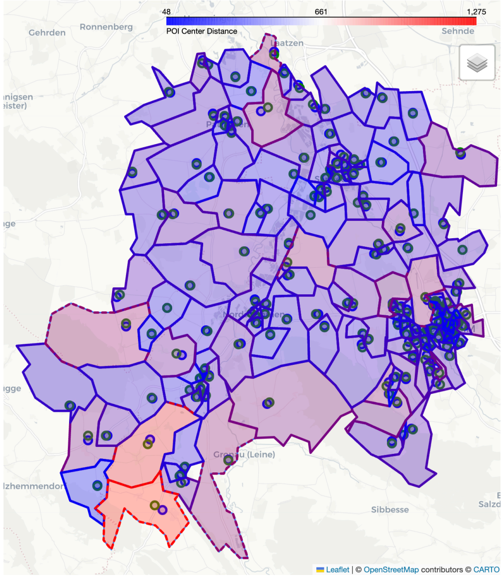

Figure 3: Study area with all clusters, the average center of each cluster in green and the POI center of scenario 1 in blue. The color of the clusters shows the average distance between the actual start and end points of the trips and the POI center of scenario 1.

In brief, while scenario 2 yields slightly better average results, scenario 1 constitutes a simpler and more transferable alternative, making it particularly suitable for application in other geographic regions.

5 SIMULATION-BASED VALIDATION

5.1 Simulation of a Mobility Service Provider

To assess the practical implications of different cluster center representations, we conducted an agent-based simulation of an integrated, demand-responsive passenger transport (IDRT) system based on the model by Staritz (2023). The simulation consists of four agent types: a Main-Agent managing the simulation environment (map, cluster stops, and 640 Shared Autonomous Vehicles [SAVs]), and SAV-Agents representing the vehicles. The Main-Agent processes mobility requests from the trip table, assigning them to SAVs using a routing algorithm executed hourly. SAVs transition between states ("parked," "toNextStop," "atNextStop") and operate under constraints: a maximum of 4 passengers per vehicle and time window constraints. The simulation was implemented in AnyLogic (version 8) and operationalized as a Vehicle Routing Problem with Pickup and Delivery (VRPPD), using Google OR Tools for route optimization. The solver applies an iterative local search approach, prioritizing requests based on time windows and minimizing total travel distance. Mobility requests are queued and matched in batches, prioritizing requests based on earliest booking time. Requests that cannot be fulfilled within the defined time window are considered unsuccessful. All distances are calculated as straight-line (geodesic) distances, and standard average vehicle speeds are applied to derive travel times. SAVs remain idle at their last stop until reassigned, and no warm-up period is applied. This setup aims to reflect a neutral baseline without pre-optimization or artificial constraints. The list of mobility requests is generated from the trip table, whereby no logistics orders are created, only passenger transportation orders. Another adaption was made to adjust the cluster boundaries to the 178 clusters used in this study and the configuration of the number of vehicles is increased to 640 to fulfill all trips. SAVs are initially assigned random starting locations and fulfill as many bookings as possible within a simulated day, adhering to passenger pickup/drop-off time windows.

This simulation is executed three times: once with the actual start and end points of all trips from the trip table, once with the centroids of the start and end clusters in which the trips are located and once with the POI-based center of the clusters. The POI center was calculated based on the configuration of scenario 1 as it is an efficient and simple configuration of POI tags. The total distance driven by all vehicles is compared across scenarios, with aggregation (centroids/POI centers) expected to reduce travel distances by consolidating geographically dispersed trips. An aggregation of the start and end points, as is the case with the centroids and the POI-based center, should lead to shorter journeys.

5.2 Results

The goal of this simulation is to assess how different methods for determining cluster centers impact the operational efficiency of a mobility service provider, measured as total distance traveled and successful order completion.

In addition, each job has precise start and end coordinates. In the centroid and POI scenario, these are mapped to the respective cluster centers in which they are located. For this purpose, 82,311 orders were generated from the trip table for the Staritz (2023) simulation model. Since this number of orders could not be processed efficiently, the order list was executed in three parts and the result values were aggregated accordingly at the end. The routing of the vehicles is based on the orders with the objective to minimize the total distance traveled, which is solved by Google OR Tools (Staritz 2023). While the original model also simulates parcel deliveries, we focus on mobility requests as the trip table does not include freight transport. Table 4 shows the results of the simulation.

Table 4: Results of the simulation with the different start/end points of the trips.

Start/end points	Successful orders	Overall distance [km]
Actual coordinates	79,364	244,128
Mapped to centroid	79,343	246,865
Mapped to POI center 1	79,377	233,461

Contrary to the common assumption that centroid aggregation reduces total travel distance by consolidating trips, the simulation results show that POI-based centers outperform centroids in both total distance and number of successful orders. As shown in Table 4, the POI scenario resulted in 10,667 km less distance traveled and a slightly higher number of completed bookings. This indicates that POI centers better reflect actual demand locations and lead to more efficient vehicle routing. In contrast, centroids seem to not reflect existing mobility patterns in a realistic way, ultimately resulting in inefficient routing.

These results reinforce the importance of semantic accuracy in initial agent placement: even small spatial improvements in cluster center selection can yield substantial gains in large-scale, time-constrained simulations.

6 DISCUSSION

The findings of this study indicate that identifying a central point for clusters of movement data based on POIs is not only a viable approach but also results in greater accuracy compared to traditional centroids. This enhancement underscores the potential of utilizing POI data for more precise movement modeling in aggregated datasets. However, the accuracy of the results is significantly influenced by the quality and completeness of both the OSM data and the trip table employed for validation.

Using POI centers instead of centroids does not compromise privacy, as POIs contain no personal data and cannot be traced back to specific individuals who made the trips. Both methods equally protect privacy since neither reveals identifiable information about travelers. Despite the privacy-neutral shift to POI centers, the trip dataset itself must remain anonymized to prevent re-identification. Techniques like kanonymity or differential privacy should be applied to ensure that the dataset cannot be linked to individuals, thus preserving privacy.

It is notable that this method is particularly effective in rural areas. In these regions, clusters are typically larger, and the number of POIs is limited, which facilitates the identification of a central POI in a more straightforward and meaningful manner. In contrast, the approach is less effective in urban environments, where clusters are typically smaller, and the density of POIs is significantly higher. This abundance reduces the effectiveness of selecting an optimized cluster center, as the variability among potential POI-based centers increases and the results get less representative of actual movement patterns.

A particular challenge emerges when clusters encompass multiple villages. In such instances, the method frequently identifies a POI center situated between the villages. While this may appear to be a balanced solution, it frequently results in considerable distances from the actual travel routes, thereby reducing the overall accuracy of the approach.

The findings of this study provide answers to the research questions:

RQ1: How well do the centroids of the clusters represent the actual travel routes in aggregated movement data?

The analysis revealed that the average spatial deviation between the centroids and the actual travel routes is 630 m which is significant, particularly in rural areas where clusters are larger and more dispersed. While centroids provide a basic approximation, their accuracy is limited due to the lack of contextual information about the nature of movement within clusters.

RQ2: How can POIs be used to identify an alternative point within the cluster to better represent actual movement patterns compared to centroids?

Yes, an alternative cluster center based on POIs can be identified. The results show that POI-based centers provide a more accurate representation of actual movement patterns compared to centroids. The average spatial deviation between the POI center and the actual travel routes is 336 m. This improvement is most pronounced in rural areas, where POI data effectively capture the locations of homes or key destinations. However, in urban areas, the abundance of POIs and smaller cluster sizes reduce the effectiveness of this approach.

RQ3: What are the effects on the total distance traveled in a simulation of a mobility service provider if the pickup location is the exact order location, the centroid of the order cluster, or an alternative point that serves as the center of the order cluster?

Using a central stop in a cluster should reduce the distance compared to the actual start/end points. The simulation results in Table 4 show that this is true for the POI center, with the use of centroids increasing the distances.

For the field of dynamic simulation modeling, particularly agent-based models, our findings offer significant methodological and practical contributions. Accurately modeling mobility systems, traffic flows, or demand-responsive transport services requires realistic spatial representations of where agents (e.g., passengers, vehicles, service points) are located and how they interact over time. The use of POI-based cluster centers provides a context-aware, semantically meaningful alternative to geometric centroids, ensuring that agents are placed in locations that reflect actual human behavior and urban structure. This leads to more realistic agent interactions, including pickup and drop-off processes, traffic congestion formation, and localized demand peaks. Furthermore, dynamical simulation models often require iterative, time-dependent updates of system states, such as adapting service routes in response to real-time demand or simulating policy interventions (e.g., changes in public transport hubs). Here, POI-based centers allow for better initialization of system states and adaptive feedback loops that reflect real-world constraints and opportunities. For example, placing agents near frequently used POIs, such as public transport platforms or residential buildings, ensures that the simulated flow of agents through the system corresponds more closely to actual urban mobility patterns, enhancing the validity of simulation outcomes. By integrating POI-based centers into dynamic simulations, researchers and practitioners gain a tool to enhance the fidelity of their models, reduce structural biases from arbitrary centroids, and improve the explanatory and predictive power of their simulations, especially in privacy-sensitive contexts where raw data is unavailable. This is particularly relevant for smart city simulations, demand-responsive transport, and urban logistics models, where spatial accuracy of agent initialization strongly influences systemic outcomes like congestion, service efficiency, and environmental impacts.

For data researchers, the study opens new avenues to refine and expand POI-based methods. The demonstrated differential effectiveness in rural versus urban environments highlights the necessity for context-specific modeling approaches that account for the unique spatial dynamics of these areas. This research also addresses critical challenges in privacy-sensitive data analysis by presenting a viable alternative for deriving meaningful insights from aggregated datasets. In doing so, it contributes to a growing body of literature focused on balancing data utility with privacy preservation, paving the way for future work to test the methodology with more diverse POI categories, enhanced datasets, or alternative sources of geospatial information. Nevertheless, the study's findings are not limited to the domain of transportation analysis; they also have interdisciplinary applications that can inform tourism planning, accessibility studies, and public health interventions. By offering a methodology to better understand and represent movement patterns, researchers in these fields can enhance their spatial analyses, leading to more informed decision-making and resource allocation.

From a practical perspective, the methodology proposed in this study offers valuable tools for urban and rural planning. For urban planners and transportation authorities, POI-based cluster centers provide a more precise representation of movement patterns, enabling the development of better-targeted infrastructure and resource allocation strategies. In rural areas, where clusters are larger and POI distribution is sparser, this method helps identify critical locations for public services such as healthcare facilities, public transport hubs, and community centers. By improving the accuracy of cluster center identification, practitioners can prioritize interventions where they are most needed, optimizing the delivery of services and infrastructure.

The approach also has clear applications for traffic management and accessibility planning. By accurately identifying key POIs within clusters, practitioners can design strategies to enhance traffic flow, improve parking availability, and optimize pedestrian pathways. Furthermore, as cities increasingly adopt smart technologies and data-driven approaches to urban management, this methodology can be integrated into smart city platforms to support real-time decision-making. The use of POI-based centers that dynamically update with real-time data would allow cities to respond more effectively to changes in movement patterns and urban landscapes.

In contexts where privacy constraints limit access to granular mobility data, this approach offers a scalable and effective alternative. The reliance on OSM data ensures that the methodology can be applied across different geographic regions, particularly in areas with robust OSM coverage. This makes it a cost-effective and replicable solution for analyzing movement patterns in diverse contexts. Moreover, the privacy-sensitive nature of this method is particularly relevant for organizations and agencies operating in data-sensitive environments, where it enables actionable insights without compromising individual privacy.

Despite its virtues and valuable implications, like all research, our study features some limitations that can be addressed in future research. First distances in the simulation are calculated as straight-line distances and standard speeds are used to calculate travel times. However, as all scenarios were calculated identically in the simulation, the significance of the result is not affected. In traffic simulations that go beyond the comparison of different cluster centers, however, realistic road networks should be used. In addition, only one cluster center is used in this comparison. This enables a direct comparison with centroids but has proven to be a challenge for clusters that contain separate villages.

Overall, our study bridges the gap between theoretical models and real-world applications, contributing to the development of more precise, context-aware, and privacy-preserving approaches to understanding movement patterns. By advancing the integration of semantic geospatial data into mobility analysis, it provides a scalable framework for both research and practice. Future work can build upon these findings to further refine and expand the methodology, enhancing its utility across a wide range of fields and applications. In comparison to existing research, this research provides a novel contribution by systematically evaluating POI-based centers against centroids, yielding quantifiable improvements.

REFERENCES

Alghamdi, T., S. Mostafi, G. Abdelkader, and K. Elgazzar. 2022. “A Comparative Study on Traffic Modeling Techniques for Predicting and Simulating Traffic Behavior”. *Future Internet* 14(10):294.

Boeing, G. 2017. “OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks”. *Computers, environment and urban systems* 65: 126–139.

de Montjoye, Y.-A., S. Gambs, V. Blondel, G. Canright, N. de Cordes, S. Deletaille, et al. 2018. “On the Privacy-Conscientious Use of Mobile Phone Data”. Scientific Data 5(1):180286.

Gao, X., H. Wang, and L. Liu. 2021. “Profiling Residents’ Mobility with Grid-Aggregated Mobile Phone Trace Data Using Chengdu as the Case”. *Sustainability* 13(24):13713.

Haklay, M. 2010. “How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets”. *Environment and Planning B: Planning and Design* 37(4):682–703.

Heredia, C., S. Moreno, and W. F. Yushimito. 2022. “Characterization of Mobility Patterns With a Hierarchical Clustering of Origin-Destination GPS Taxi Data”. *IEEE Transactions on Intelligent Transportation Systems* 23(8):12700–12710.

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004. “Design Science in Information Systems Research”. *Management Information Systems Quarterly* 28(1):75-105.

Kashian, A., A. Rajabifard, Y. Chen, and K. Richter. 2017. “OSM POI Analyzer: a Platform For Assessing Position of POIs in OpenStreetMap”. *International Symposium on Spatial Data Quality*, September 18th-22nd, Wuhan, China, 497-504.

Krumm, J. 2007. “Inference Attacks on Location Tracks”. In: *Pervasive Computing*, edited by A. LaMarca, M. Langheinrich, K.N. Truong, 127-128. Berlin, Heidelberg: Springer Berlin Heidelberg.

Neis, P., D. Zielstra, and A. Zipf. 2013. "Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions". *Future Internet* 5(2): 282-300.

Notelaers, L., J. Verstraete, P. Vansteenwegen and C. M. J. Tampère. 2024. „A Travel Demand Modeling Framework Based On OpenStreetMap”. *Discover Civil Engineering* 1:26.

OpenStreetMap Wiki contributors. 2025. “Map features”. *OpenStreetMap Wiki*, https://wiki.openstreetmap.org/w/index.php?title=Map_features&oldid=2804789, accessed 2nd March 2025.

Phithakkitnukoon, S. 2023. *Urban Informatics Using Mobile Network Data*, Singapore: Springer Singapore.

Senozon Deutschland GmbH. 2022. Mobilitätsmodell Deutschland. https://mobilithek.info/offers/524943653221584896 accessed 21th December 2024.

Spengler, L., E. Gößwein, I. Kranefeld, E. Spachtholz, F. E. Kracht, M. Liebherr, et al. 2024. “Aktuelle Herausforderungen der Modellierung von Mobility on Demand-Systemen”. In *Next Chapter in Mobility: Technische und betriebswirtschaftliche Aspekte*, edited by H. Proff, 535–547. Wiesbaden:Springer Fachmedien.

Staritz, J., J. Kütemeier, H. Sand, C. v. Viebahn, and M. Wartenberg. 2023. "Rebalancing Integrated, Demand-Responsive Passenger and Freight Transport – An Agent-Based Simulation Approach". In *2023* Winter Simulation Conference (WSC), 185-196, https://doi.org/10.1109/WSC60868.2023.10408669.

Teralytics Inc. 2022. “Teralytics Matrix - Mobility Trends in Germany” [Unpublished raw data]. accessed via Teralytics platform 15th March 2025.

Thuillier, E., L. Moalic, S. Lamrous, and A. Caminada. 2018. “Clustering Weekly Patterns of Human Mobility Through Mobile Phone Data”. In *IEEE Transactions on Mobile Computing* 17(4):817–830.

Tran, M. and C. Draeger. 2021. “A Data-Driven Complex Network Approach for Planning Sustainable and Inclusive Urban Mobility Hubs and Services” In *Environment and Planning B: Urban Analytics and City Science* 48(9):2726–2742.

Yang, K., L. Liu, and Y. Wen. 2024. “The Impact of Bayesian Optimization on Feature Selection”. *Scientific Reports* 14:3948.

AUTHOR BIOGRAPHIES

MAYLIN WARTENBERG is professor for business informatics at the University of Applied Sciences Hannover. She has a mathematical background and profound experience in the financial and automotive industry. Her research areas include Data Science, Business Intelligence and Artificial Intelligence. She is also active in Science Communication in the field of AI. Her email address is maylin.wartenberg@hs-hannover.de. Her website is http://www.das-hub.de.

LUCA MARCO HEITMANN is a student for business informatics and part-time resesearcher at the University of Applied Sciences Hannover. His research interests focus on Agent-based Simulation in the contexts of Supply Chain Management, Urban Logistics, and Mobility. His e-mail address is mail@luca-heitmann.de.

MARVIN AUF DER LANDWEHR is a senior consultant at FehrAdvice & Partners AG. He holds a Ph.D. in Economics and has published his research in major scientific outlets such as European Journal of Information Systems and Journal of Cleaner Production. His research interests focus on Logistics and Supply Chain Simulation, Information Systems and Behavioral Economis. His e-mail address is marvin.aufderlandwehr@fehradvice.com.