International Journal of Geo-Information

MDPI

*Article*

# Privacy Preserving Human Mobility Generation Using Grid-Based Data and Graph Autoencoders

**Fabian Netzler * and Markus Lienkamp**

Chair of Automotive Technology, Technical University of Munich, Boltzmannstr 15, D-85748 Garching, Germany
* Correspondence: f.netzler@tum.de; Tel.: +49-89-289-16598

**Abstract:** This paper proposes a one-to-one trajectory synthetization method with stable long-term individual mobility behavior based on a generalizable area embedding. Previous methods concentrate on producing highly detailed data on short-term and restricted areas for, e.g., autonomous driving scenarios. Another possibility consists of city-wide and beyond scales that can be used to predict general traffic flows. The now-presented approach takes the tracked mobility behavior of individuals and creates coherent synthetic mobility data. These generated data reflect the person's long-term mobility behavior, guaranteeing location persistency and sound embedding within the point-of-interest structure of the observed area. After an analysis and clustering step of the original data, the area is distributed into a geospatial grid structure (H3 is used here). The neighborhood relationships between the grids are interpreted as a graph. A feed-forward autoencoder and a graph encoding–decoding network generate a latent space representation of the area. The original clustered data are associated with their respective H3 grids. With a greedy algorithm approach and concerning privacy strategies, new combinations of grids are generated as top-level patterns for individual mobility behavior. Based on the original data, concrete locations within the new grids are found and connected to ways. The goal is to generate a dataset that shows equivalence in aggregated characteristics and distances in comparison with the original data. The described method is applied to a sample of 120 from a study with 1000 participants whose mobility data were generated in the city of Munich in Germany. The results show the applicability of the approach in generating synthetic data, enabling further research on individual mobility behavior and patterns. The result comprises a sharable dataset on the same abstraction level as the input data, which can be beneficial for different applications, particularly for machine learning.

**Keywords:** mobility data; synthetic data generation; mobility data analytics



**Citation:** Netzler, F.; Lienkamp, M. Privacy Preserving Human Mobility Generation Using Grid-Based Data and Graph Autoencoders. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 245. https://doi.org/10.3390/ijgi13070245

Academic Editors: Wolfgang Kainz, Jun Feng, Changqing Luo and Mamoun Alazab

Received: 15 April 2024
Revised: 13 June 2024
Accepted: 30 June 2024
Published: 9 July 2024



## 1. Introduction

Individual trajectory data play an increasing role in different application fields. The field reaches from local trajectories with a short duration period used for planning algorithms for autonomous driving applications [1] to large-scale city scales. The first application deals with short-term trajectories, lasting from seconds to a few minutes, and involves local geospatial dimensions measured in meters. The second type of application, e.g., city planning activities or location-based-services, requires data covering days to years and distances from kilometers to entire metropolitan areas and regions.

The need to create artificial datasets arises from two main arguments: First is the generalizability. This is required due to different local restrictions and limited resources since generating original data is time- and resource-consuming. The ability to generalize already collected data and reapply them to the same area is useful in leveraging small data samples. This holds true, especially with added noise to enrich the dataset or use it for a completely new area while keeping the semantic and geographical distribution [2]. Second is the need for privacy. Highly accurate personal tracking data have become more accessible with smartphones and other GPS tracking solutions. Precise information

about a person's position and activities is tracked for an extended period. These data are applied for model training for various services, e.g., for destination prediction [3] or contact tracing during COVID-19 [4]. In these applications, critical personal information about the user, such as home or work locations, can be exposed. Subsequently, several methods have recently been developed to generate artificial and personalized trajectory data, with distinctions regarding the domain (local/short-term and area/long-term) and method (data-driven versus knowledge-driven) [5].

The data-driven methods found an enabling factor especially within the geospatial data processing and geospatial embedding techniques. These techniques leverage information about POIs and transportation structures into state vectors that can be used to enrich machine learning models. Many methods in this category [2,6–9] excel in predicting large-scale person movement, e.g., to assess city planning regarding transportation capabilities. Others concentrate on individual POI-check-in synthetization.

Both categories need two significant requirements. First is the possibility of having a one-to-one matching between a real and an artificially generated person's mobility behavior. Therefore, the exact activity order needs to be respected and location stability is required. This means, for example, that the tracked person has the exact location as "home" over the whole dataset. The second is the explicit characterization of the neighborhood of POIs. Not just a point-of-interest (POI) category or a general statement like a city center is taken into account. Instead, a more explicit embedding, adaptable to the significant characteristics of the individual use case, is generated.

While the second requirement is covered by some of the existing methods, e.g., by Choi et al. [8], the topic of long-term consistency is barely addressed. This holds true especially in combination with the correct POI-check-in synthetization and correct regional embedding into a city structure.

This paper proposes a one-to-one trajectory synthetization method with stable long-term individual mobility behavior based on a generalizable area embedding. In contrast to the state-of-the-art, the focus lies on long-term origin location consistency while guaranteeing a meaningful neighborhood selection for the POI choice of each way.

The underlying data flow and principal components for the proposed method are illustrated in Figure 1. The flowchart emphasizes the three stages that are discussed in this paper: pre-processing, data generation, and post-processing. During pre-processing, the tracking data are divided into tracking data for a single user and for a whole study. Participants are analyzed individually, and the results are used for the data generator in the second phase. In contrast metrics, for example, travel distances generated in the study are used to validate the results of the synthetization process. For each participant, new synthetical data are generated based on a similarity function that compares possible tracking data candidates with the original data and a privacy-ensuring mechanism that excludes possible artificial data points that endanger the participant's privacy. The metrics from the artificial tracking data are calculated and aggregated among all participants. They are compared to the original metrics in the post-processing phase to produce a quality measure of the synthetization method.

This paper is structured as follows: Section 2 discusses the state-of-the-art geospatial embedding and synthetic trajectory generation methods. Afterwards, Section 3 introduces the proposed method of autoencoder-based neighborhood respective trajectory generation. Finally, Section 4 shows the result of the proposed method with its application to a dataset from the Mobilitaet.Leben study from Munich.

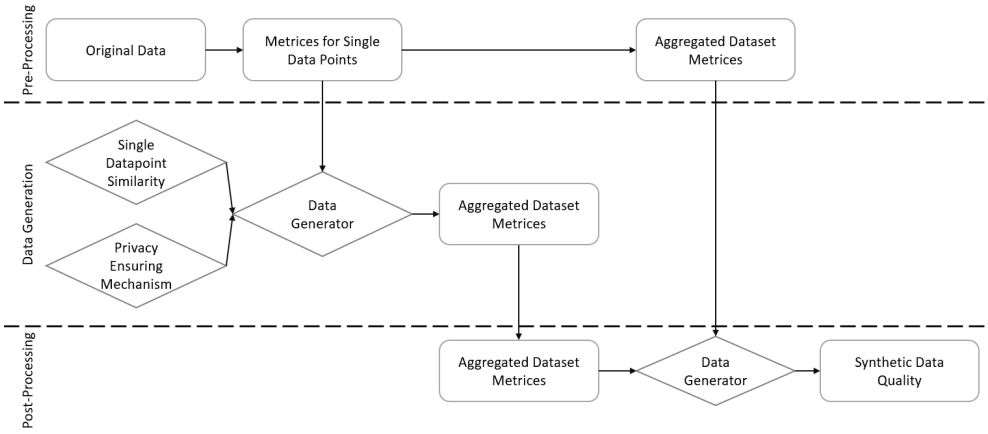


**Figure 1.** The underlying data flowchart for the proposed method.

## 2. Related Works

The proposed method for synthetic trajectory generation is based on geospatial feature embedding. The following chapter starts with the state-of-the-art of this topic before discussing the main techniques of synthetic data generation.

### *2.1. Geospatial Embedding*

Typical methods for generating synthetic trajectories use point-of-interest embedding to generalize and represent the geographical region and the underlying use of the trajectories. Embedding is a form of representation learning that aims to learn a mapping function from a generic object (e.g., sequential data like a sentence) to a vector representation. Many of these methods stem from the advancements in Natural Language Processing, as seen in [10]. Table 1 indicates the general stages for geoembedding and the current state-of-the-art for the different areas.

**Table 1.** Structure to distinguish different geoembedding methods. The methods in the state-of-the-art can be divided by the data they use, how the embedded regions are defined, by which algorithm the embedding is realized, and for what purpose.

| | | |
|---|---|---|
| Data | OpenStreetMap | [7,11–14] |
| | OpenStreetMap—Network | [13,15] |
| | Trajectories | [7,12,16,17] |
| | Aerial Images | [12,18] |
| Regions | H3 | [11,13,14,18] |
| | Semantic Regions | [12,15,16] |
| | Administrative Boundaries | [12,17] |
| | Pixel based | [7,17,18] |
| Algorithms | Skip-Gram | [11] |
| | Autoencoder | [12–14] |
| | Graph-based | [12,15,17] |
| | CNN-LSTM-Architecture | [7] |
| | Stochastic | [16,18] |
| Purpose | Transportation Network Embedding | [13,15] |
| | General POI Embedding | [11,12,14,16,18] |
| | Mobility Pattern Recognition | [7] |
| | Specialized Urban Features | [12,17,18] |

In geospatial data analysis, the embedded objects primarily consist of regions, often defined semantically [16] to capture human interactions within the region or based on strict geographical features, such as feature aggregation over H3 Grids [11]. Du et al. [16] implemented a method that utilizes recorded trajectories to create zones and generates

an embedding based on the zone properties, effectively establishing connections between these zones.

The method Tile2Vec, proposed by Jean et al. [18], uses the distributional hypothesis from natural language processing to learn semantically meaningful embedding from aerial imagery data. Empirically, it is shown that visual analogies can be obtained with simple operations within the calculated latent space. Jenkins et al. [12] introduced multimodal data as a basis for advanced embedding, leveraging aerial images, human mobility traces, and point-of-interest data. The developed end-to-end framework achieved semantic embedding over discrete regions and was discussed mainly for urban areas. The author describes this approach as Learning an Embedding Space for Regions (LESR). The linkage between human mobility and each trip's semantic meaning is again evaluated in [7], in combination with the use case of transferring mobility knowledge from one city to another. Jiang et al. [7] combines a word-like embedding using POI categories to create a POI-based image-like data structure while using Convolutional Neural Networks and Long Short-Term Memory (LSTM) architectures taken from the field of image recognition. The first approach towards representation learning over OpenStreetMap data within microgrids concerning urban functions and land use is proposed in [11].

Uber's H3 index is used to define microgrids, and the "hex2vec" method employs a SkipGram Model to calculate vector representations of each H3 grid, displaying semantic properties of the map. This method allows for simple arithmetic functions to compare areas and find those with selected features, particularly focusing on the road-based representation introduced in [13]. Leveraging OpenStreetMap as the underlying data source, a vector representation is learned, enabling clustering and arithmetic operations over the latent space. The outcome is a high-level, scalable typology related to the underlying road network in the observed area.

In their work, Zhang [15] and Shin et al. [17] introduced the concept of graph neural networks. Zhang et al. [15] also addressed the challenge of road embedding, termed "road network representation learning (RNRL)". Their method emphasizes the high-order relationships between roads, focusing on deriving regions and identifying the central connecting roads between them. The key aspects of their approach include constructing a hypergraph over the road network and implementing an information propagation mechanism within this hypergraph. Shin et al. [17] pursued the goal of obtaining an urban representational embedding, which was used for predicting house prices and employee rates. The input data comprised taxi trips and subway rides, and their architecture was built on a Graph Attention Network. Their study also delved into embedding dimensions and utilizing the urban mobility network for various tasks.

An alternative method and advancement rooted in H3 grids and OpenStreetMap information is GeoVeX, introduced by Donghi et al. in 2023 [14]. This approach utilizes autoencoders to manage geographic count data by amalgamating neighboring hexagon attributes into a task-agnostic latent space.

### *2.2. Synthetic Trajectory Generation*

The generation of synthetic trajectories can be categorized based on the used methods (data- or knowledge-driven) and the specific scope. The distinction lies in whether the focus is on generating overall trajectories within a specific area or finding realistic trajectories for individuals with their respective attributes. These scopes are crucial for different applications. While the former is useful for city planning tasks, the latter can be applied in privacy-preserving methods or to create synthetic data for various machine learning applications. Table 2 shows the main methods used for synthetic data generation and the primary corresponding sources, as discussed in the subsequent part.

Many data-driven approaches are developed, especially for generating general trajectories, without including personal information or biases between groups. These approaches are scalable, from generating pedestrian trajectories [19] to city-wide traffic counts [8]. These approaches to generate pedestrian trajectories consider both temporal and spatial

relations and a social aspect, like the interaction between different pedestrians, bicycles, and traffic members in the local area [20]. The focus of this work is on generating individual vehicle-based trajectories; therefore, this factor needs to be further discussed.

The task can be distinguished between short-term trajectories, for example, by Park et al. [21] or Messaoud et al. [22] with a prediction horizon under 10 s, and large-scale trajectories over a whole geographical region and a time horizon around hours to days. The short-term solutions use LSTM architectures, mainly used to prevent collisions. Long-term trajectory generation, associated with privacy-preserving data publishing, is addressed by Rao et al. in 2020 [23].

The proposed LSTM-TrajGAN is a deep learning framework using a generative adversarial network to create synthetic trajectories on a city scale. Cao et al. [24] developed the TrajGen method that generates one-to-one real-to-artificial trajectories while preserving some characteristics, like aggregated POI and distance statistics. The method was tested on a taxi dataset, demonstrating its capabilities. With a focus on $\epsilon$-differential privacy, Alatrista-Salas et al. [25] conducted a study showing the applicability of Differential Privacy Generative Adversarial Networks on mobility data on a GPS level, but also that the risk of re-identification persists. If such methods are used for sharing, the original data must not be merged and shared with the synthetic data. This requires that all significant data characteristics are contained within the artificial data.

Wu et al. [26] gave a deeper insight into guaranteeing POI sequences. While using hierarchical POI categories, the method based on pairwise location reorganization persists in the exact POI categories within a trajectory sequence. The experiments conducted are only in a limited area due to the complexity of the solution space, and the POIs are also characterized as scalable and precise; the neighborhood where those trajectories are embedded is widely ignored.

While those methods generate plausible results on the aggregated micro- and macroscopic dimensions, they lack some attributes that may be needed for higher-tier data analytics. They are part of mobility studies like the Mobilität in Deutschland [27] or Mobilitätspanel [28]. Especially when handling personal tracking data, a plausible and stable location association is needed. If a location remains the same in the original data, it has to be guaranteed that this is the same for the synthetic dataset. A plausible set of locations and interactions for every participant is set even for long-term (months to years) studies.

A survey can be found in [5], giving an extended overview of additional approaches, such as knowledge-driven methods using simulation software such as SUMO or MATSIM and mobility demand generators for them like SUMOPY [29], MiTo [30], or ABIT [31]. These methods are more applicable to creating general demand data from a statistical population average and not reproducing exact studies with strong behavioral biases. This differs from the scope of this work and will not be further evaluated.

**Table 2.** Key methods used for synthetic trajectory generation divided by possible application scopes.

| **Knowledge-Driven** | **Data-Driven** | |
|---|---|---|
| **Potential City-Wide** | **Local** | **Potential City-Wide** |
| Socio-Demographic Travel Demand Generation [32–34] | LSTM [22–25] | LSTM [26]<br>GAN [12,26–28]<br>Pairwise Reorganization [29] |

### *2.3. Scope of This Work*

The scope of this work is to create a synthetic dataset from long-time mobility behavior studies synthesizing the personal tracking data of the participants, e.g., Mobilitaet.Leben [32]. The generated artificial data should ensure that the data can be shared afterward without violating privacy requirements.

Reviewing the state-of-the-art for synthetic trajectory generation shows recent achievement in three categories, as illustrated in Table 2. First, knowledge-based methods that

create synthetic populations. These consist of single agents with consistent mobility behavior over the whole observation period. As these methods are based on population statistics, they cannot represent highly biased data of single individuals and therefore cannot be applied to this kind of study.

Second and third are data-driven approaches described above, but only the at-least-city-wide approaches are applicable since the focused tracking studies require an extended region.

The applicability of those remaining methods [12,26–29] on the problem is further evaluated on a qualitative basis by different dimensions in Table 3. The dimensions are:

1. Routes: How strong the characteristics of the taken routes resembled are in the synthetic dataset. Exemplary characteristics are: distance, speed profile, and traversed regions.
2. POIs: Respective points of interest at the origin and destination of ways are aligned between original and synthetic data.
3. Time: The duration of ways and stay times during activities between ways are considered.
4. Regions: Not only the exact point of origin and destination of a way has influence on the way generation, but also the region around those points, defined by characteristics, like nearby points of interest.
5. Consistency: Especially for long-term studies, the same locations, e.g., home, are visited multiple times. If the synthetic data of individuals keeps the characteristic, it is necessary to ensure that this visit frequency to the same location is the same in the original and the synthetic data.

**Table 3.** Overview over capability of data-driven and city-wide synthetic way generation methods.

| Literature | Routes | POIs | Time | Regions | |
|---|---|---|---|---|---|
| Choi et al. (2021) [8] | ● | ○ | ○ | ● | ○ |
| Rao et al. (2020) [23] | ○ | ● | ○ | ◔ | ○ |
| Cao et al. (2021) [24] | ● | ○ | ◑ | ◔ | ○ |
| Alastrista-Salas et al. (2022) [25] | ◔ | ◑ | ○ | ◔ | ◔ |
| Wu et al. (2022) [26] | ◔ | ● | ○ | ○ | ◔ |
| Proposed Method | ◔ | ● | ○ | ● | ● |

Table 3 shows the missing capabilities of current methods when dealing with the time and location consistency of activities, if applied to long-term studies. The proposed method focuses on the topic of way consistency for synthetization. Comparable to diary-based studies like Mobilität in Deutschland [27] or Mobilitätspanel [28], the POI information, but also the surrounding region information, needs to be conserved too.

Based on the reviewed literature, the proposed method fulfills three main hypotheses:

1. The mobility behavior of each participant is reflected accurately.
2. There is no personal information within the artificial dataset that allows a reidentification.
3. The main characteristics of the data remain preserved within the single participant's data, but especially when aggregating the data of all participants.

## 3. Method

This section illustrates the proposed method. The implementation can be found on GitHub (https://github.com/TUMFTM/way_chain_generator—accessed on 5 July 2024).

### 3.1. Requirements towards the Approach

The approach differs in its requirement from previous work by focusing on the individual long-term mobility behavior of people. For this, the geographic area is at least city-wide or, optionally, a whole metropolitan area, and the period expands over several days to months. The goal is to achieve consistency regarding the POIs and the area surrounding those. The observed region includes the inner city, suburbs, and industrial districts. Regarding recurrent visited locations, the mobility behavior must also be reflected as precisely as possible so that the place of work and home is the same throughout the synthetic trajectories of a person. The mobility behavior should be as close as possible to the original dataset without revealing personal location data, which enables a reidentification. Only location-based reidentification attacks are to be prevented by the proposed method; frequencies of visits per user are explicitly not masked [34].

### 3.2. Approach Overview

Four core challenges arise from the requirements and are reflected in some design decisions regarding the method.

1. Challenge: Reflect each participant's specialties in the respective mobility behavior.
   Solution: Conduct a one-to-one synthesis, meaning every person's results in exactly one artificial way chain. This reflects the characteristics of this person and is not learned from a group of people.
2. Challenge: If the person returns to a previous location within the data, the respective dependencies need to be found within the individual dataset.
   Solution: Include a pre-processing step analyzing the complete location-to-location dependencies of each participant.
3. Challenge: The geographical space is large-scale, with over a million potential start-stop points.
   Solution: Implement a region-based pre-selection. As a regionalizer algorithm, Uber's H3 grid system is used.
4. Challenge: Complexity of the solution space with OSM having over 300 usable features just for buildings, and the need to include the surroundings and neighborhood for the characteristics of potential target points.
   Solution: Create a latent space embedding for the regions derived from Challenge 3. The regions are embedded using a low-dimensional feature vector, combining the features of the specific region with its neighbors.

Resulting from the identified challenges, the architecture shown in Figure 2 was derived. As a starting point, the original tracking data are analyzed and the start and end points of each track of the user are clustered with a DBSCAN algorithm. These clusters and their connection, given through the user's trajectories, are analyzed, and the most central cluster, further called the gravitational center, is identified. The clusters are associated with their H3 grids; the details are described in Section 3.3.

Next, all H3 grids within the overall study area are embedded into a latent space, combining their own and the features of their surrounding grids into a low-dimensional representation. The used methods are further described in Section 3.4. New combinations of H3 grids are found within this latent space, according to the "loss" value of an application-specific cost function with a greedy approach. This value combines the difference between the features of the original grids and potential candidate grids' features and the distances within the way chain. More details can be found in Section 3.5.

The precise cluster positions are located within the selected grids by identifying applicable buildings. The input data structure, georeferenced linestrings, is imitated by routing over the street network within the new-found clusters in the same order as the original trajectory. Possible privacy-ensuring mechanisms are described in Section 3.6.

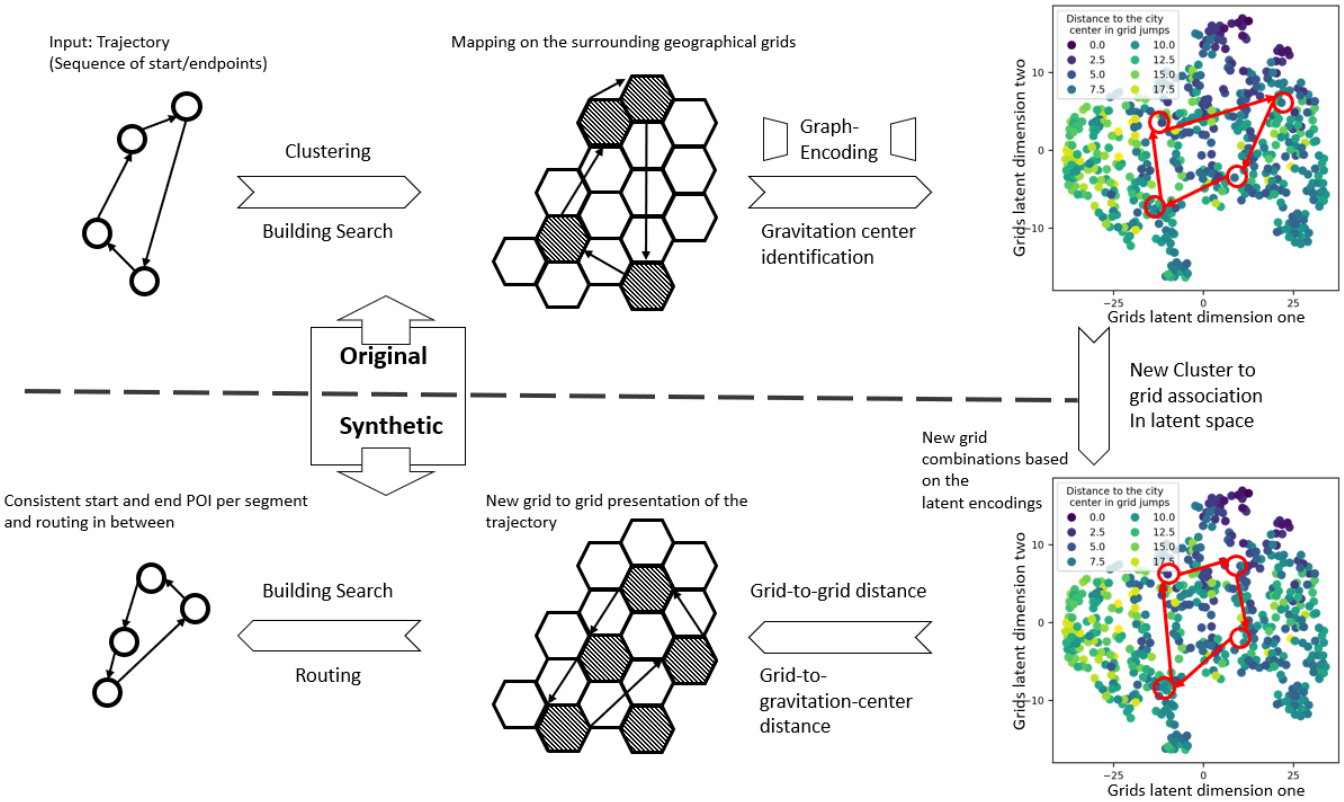


**Figure 2.** Schematic representation of the methodology, showing the sequence to generate synthetic way chains from tracked mobility behavior.

### *3.3. Data Analysis and Filtering*

The trajectory data of each user are loaded and given as a linestring with WGS84 coordinates. From this collection of ways, only the start- and endpoints and their connections are considered, as shown in the first picture of Figure 3. It is assumed that the data collection frequency and the level of noise within the position data allow for a guaranteed correct POI identification from those points. After conversion to a metric coordinate system, a DBSCAN algorithm is used to identify clusters of start- and endpoints. The result indicates which ways are aimed at the same place; for example, all the tracks that can be associated with the user's home.

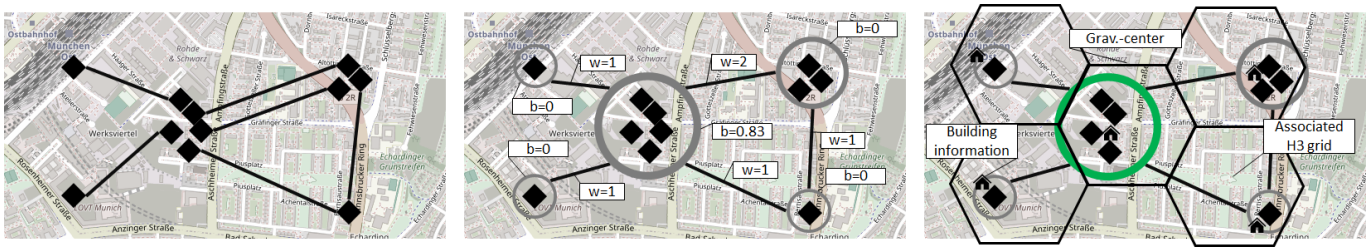


**Figure 3.** Symbolic representation of the main analyzing steps. The left image shows the exemplary start- and endpoints. The middle illustrates the generated mobility graph. The visited clusters are nodes and the recorded ways between give the edges. The centrality value of each cluster (value "b") is associated with the nodes and the edge weights are derived from the frequency of ways between the nodes. The right one shows the association of clusters with H3 grids and information about the building nearest to each cluster centroid.

After the clustering, the centroid of the shape made from points related to one cluster is used. Following this first analyzing step, the tracking data consist of buildings connected with each other that correlate with the user's movement. The data are interpreted as a graph

for the next part of the process, with the clusters as nodes and the ways between them as edges. This graph is enriched by finding the node with the highest betweenness value [33]. This node is marked as the "gravitation center" of the user's mobility behavior. The edge weight is given by the number of connections between two nodes of the mobility graph, and used to calculate the betweenness value, shown in the middle image of Figure 3.

OpenStreetMap (OSM) is now used to select the next building together with its properties to each origin or respective destination of a track. Every cluster is associated with the corresponding H3 grid. For the mobility behavior of a person, it is essential to characterize not only the concrete building in the specific way it leads to, e.g., a grocery store, but also the surrounding area. The end result of the analyzing stage can be seen in the figure's right image. The data model at this stage is depicted in Figure 4. The three data sources are represented in cursive: a database containing buildings from OpenStreetMap, the H3-grids of the region, and the raw trajectory data. The steps shown in Section 3.3 demonstrate the impact on the data model.

Clusters are generated from the input trajectories. The connections between clusters are tallied and utilized as weights for the edges in the mobility graph. These data enable the calculation of the betweenness value, thus completing the "Output Data".

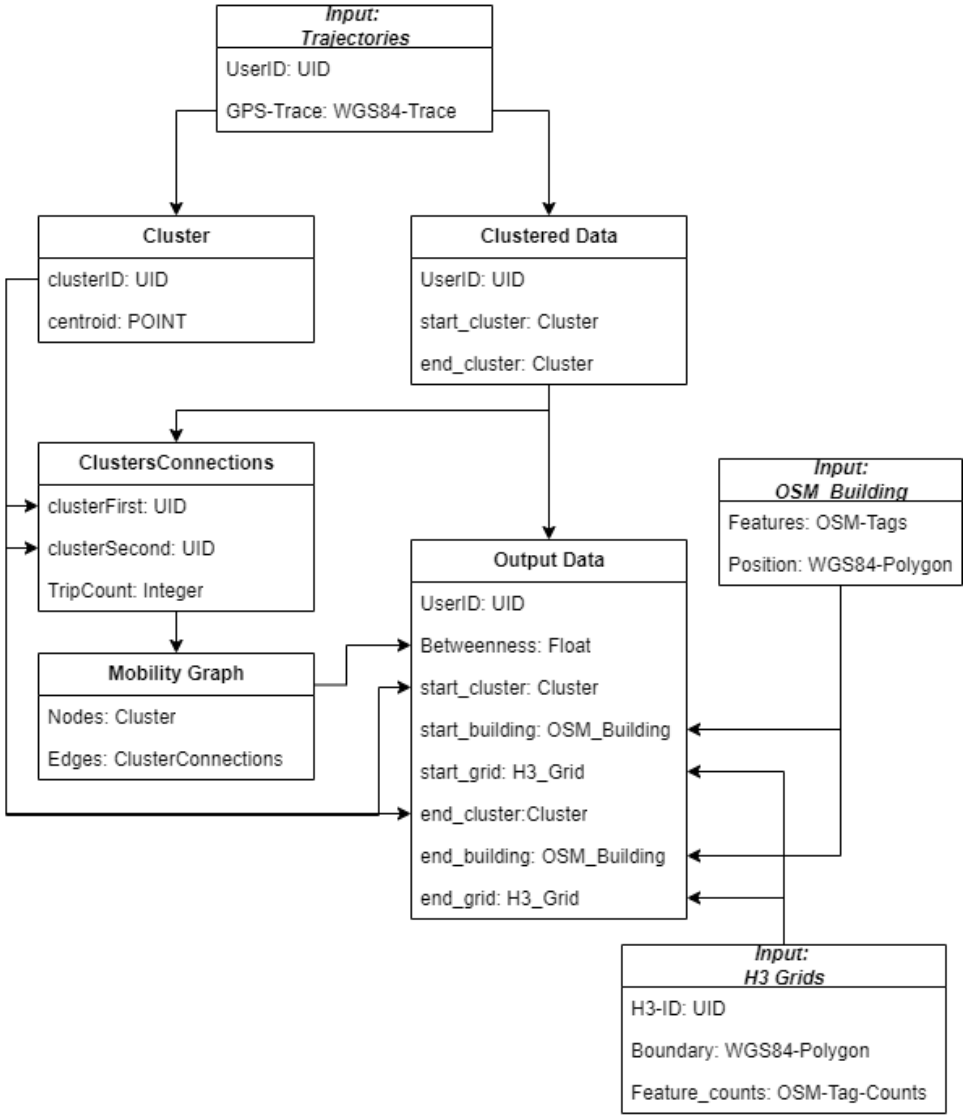


**Figure 4.** The data model during the data analysis stage. In cursive are the input data.

### 3.4. Latent Space Creation

The area where the trajectories are generated is discretized by the UBER H3 grids system [35]. Every grid is associated with several OSM tags and the count of how those are present in the grid. The tags can be chosen according to the attributes the synthetic trajectories should consider. The parameter used for this paper can be found in Section 4.1.

A latent space for the grids is constructed to reduce the complexity of the feature space and balance features. To gain a characterization of not only the single H3 grid, but also its connection to the surrounding area structure, a graph-based encoding–decoding network is used to create a latent space representation of the areas. The latent space representation of a grid is divided into two parts. The first component of the latent space is an encoding of its own OpenStreetMap features; the second reflects its connection to the surrounding grids.

The OpenStreetMap properties are used as a database and reflect the attributes that shall stay consistent from the original to the synthetically generated tracks. The number of feature occurrences is counted for each grid. By concatenating these counters, the feature vector for each grid is constructed. This implementation used the SRAI framework [36]. For the varying number of features, an autoencoding network is used to create a low-dimension representation for the counted properties of each grid. A graph encoding–decoding network is used for the latent space representation of the structure around the grid. Figure 5 shows the underlying basic principle behind the architecture.

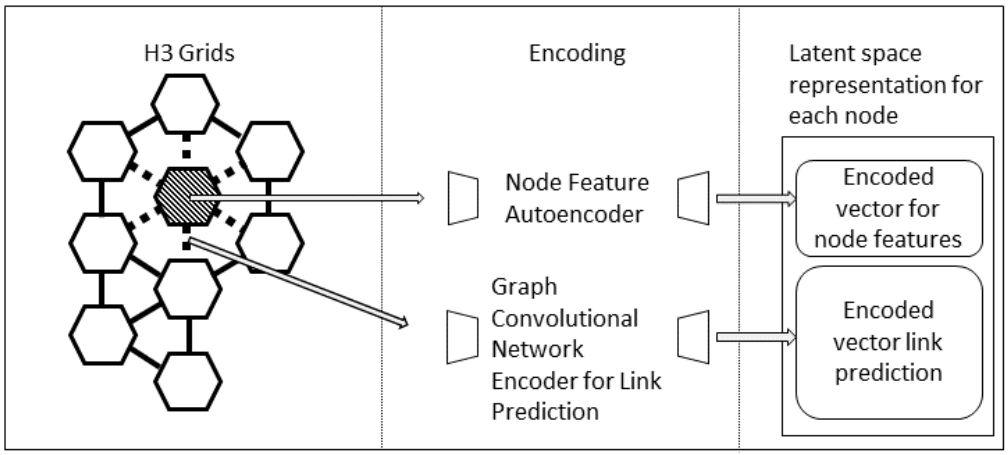


**Figure 5.** Creation of the latent space representation of the H3 grids. Every node gets an embedding vector used for link prediction to its neighboring nodes. It acts as area characterization for the node. In addition, an embedding vector from the encoding of its own OpenStreetMap tag counts is calculated.

This graph consists of the nodes as the H3 grids within the area and the edges connecting each grid to its neighbors. The edge weight is given as 1, and the attributes of each node are the counted OSM features. The encoder consists of two Convolutional Layers from [37], which can process graph-structured data. The encoding is the cross-product between two latent spaces of nodes, resulting in the probability that these two nodes are connected with an edge. The entire latent space of a node is then given by concatenating the two encoding vectors. The complete layout of the neural network with the parameters used in Section 4.2 is illustrated in Figure 6.

The architecture used for this paper considers nine OpenStreetMap tags (“shop”, “residential”, “office”, “building”, “leisure”, “restaurant”, “medical”, “train_station”, “school”) and constructs a 3-dimensional embedding space for the feature encoding space and a 2-dimensional space for the link prediction autoencoder.

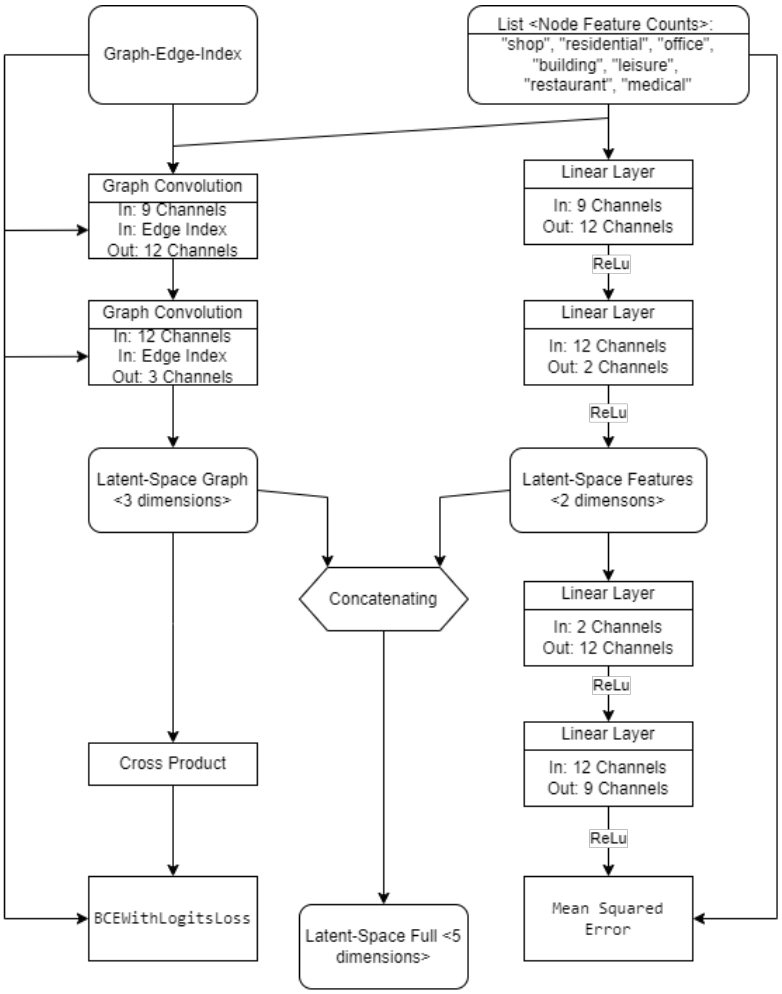


**Figure 6.** Neural Network Layer architecture for the latent space creation. The included values reflect the parameters as implemented in Section 4.2.

### *3.5. Trajectory Generation*

The synthetic way chain is built using a greedy approach. The basic principle is shown in Listing 1. The algorithm starts by selecting the best-fitting H3 grid within the latent space; compared to the H3 grid, the user begins its first way with. Afterwards, the according endpoint of the trajectory is found. This happens based on the cost function:

$$\forall\, node\ in\ LatentSpace \tag{1}$$

$$C_{node} = C_{latentSpaceEuklideanDistance} * \Delta_{distance} * \Delta_{gravitation} + C_{distance} + C_{gravitation} \tag{2}$$

$$C_{distance} = \Delta_{distance} * C_{latentSpaceEuklideanDistance}[Q_{0.1}] \tag{3}$$

$$C_{gravitation} = \Delta_{gravitation} * C_{latentSpaceEuklideanDistance}[Q_{0.1}] \tag{4}$$

This function is applied in the "sort_loss" call in Listing 1. It takes the H3 grid from the original trajectory and the whole latent space as parameters. The method uses differences in the latent space. It multiplies these differences by factors that are based on the difference in the distance of the route and the distance from the trajectory's gravitational center. Formulas (2) and (3) add error terms based on the distance error and the distance error from the gravitational center of the entire route. This is carried out by multiplying the corresponding delta+1 with the 10th percentile of the latent space distances across all grids. These additional terms are essential for selecting a nearly perfect matching grid in the latent space, regardless of the distance that needs to be covered, in order to avoid unrealistic behavior.

Among the identified grids, the buildings most similar to those in the original trajectory are identified based on their OSM properties. These coordinates are associated with the clusters identified in the analysis step. Every subsequent trajectory is treated in the same way, but each starting or ending point is checked to see if the corresponding cluster has already been defined. This approach ensures that the characteristics of individual trajectories remain plausible. If a user visits a point once, it refers to the same point if seen later, and the trajectory can return to the most central point without implausible distances in between. Once the trajectory is identified as a sequence of start- and endpoints, these are routed to the next element on the street graph. The routing algorithm is a simple free-flow routing. It assumes that no congested roads, traffic jams, or any other conditions influence the route choice. To match a more realistic routing, additionally, the traffic conditions that were recorded in the original data need to be implemented.

**Listing 1.** Code for trajectory generation.

```
Input:
    original_tracks – pre filtered per user:
        <start_point, end_point, start_grid,
        end_grid, start_cluster, end_cluster>,
    grids_latent_space: <h3_grid_id, [5 dimensional latent space]>
Result:
    synth_tracks:  <start_point, end_point>

grav_center_cluster = get_grav_center(original_tracks)
For track in original_tracks:
    if track["start_cluster] in clusters:
        syn_track_start = clusters[track_start]
    else:
        grids = sort_loss(grids_latent_space, track["start_grid])
        start = best_grid_anonymized(grids)
        syn_track_start = get_best_fitting_building(start)
        clusters.add(track["start_cluster], syn_track_start)
    if track["end_cluster] in clusters:
        syn_track_end = clusters[track["end_cluster]]
    else:
        grids = sort_loss(grids_latent_space, track["end_grid])
        grids = add_gravitational_loss(grav_center_distance)
        end = best_grid_anonymized(grids)
        syn_track_end = get_buildings_in_grid(end)
        clusters.add(track["end_cluster], syn_track_end)
    synth_tracks.append((syn_track_start, syn_track_end))
synth_tracks = do_routing(synth_tracks)
```

### *3.6. Possible Privacy-Ensuring Mechanism*

The proposed method will generate waypoints that refer to the original grids and even the same buildings as the input trajectory. This behavior must be prohibited to preserve the study participants' privacy. The privacy requirements are implemented on the grid selection level to guarantee no overlaps in the location–activity tuple.

The discussed approaches should ensure a protection against simple location matching attacks [34]. Complex context linking or frequency-based attacks are not further elaborated and are the subject of future research.

The naive method to ensure that the synthetic trajectories are correct is to leave out the original grid while synthesizing a single way. It can be forced by taking a grid with a latent space error more significant than 0. This step is taken when selecting start- and endpoints.

The approach can be extended by defining an epsilon that describes the minimal latent space distance from the original grid to the equivalent grid within the synthetic trajectory. This results in a trade-off between similarities in area characteristics and accuracy in reproducing the actual behavior. It is important, that the latent space or all the information to create it in a deterministic way must not be shared with the dataset. This represents a major weakness in this naive approach.

$$\epsilon > 0 \tag{5}$$

$$C_{latentSpaceEuklideanDistance} > \epsilon \tag{6}$$

Another possibility is the introduction of noise to the trajectories' latent space representation. This also allows the sharing of data on a more generalized level. A comparable concept is introduced in [38]. Here, the first part of the algorithm is executed, and the input trajectories are brought into the form sequence (cluster, building information, grid latent space).

$$W(start\_grid, end\_grid, start\_building, end\_building) = Way\,Chain \tag{7}$$

$$nodes = H3 - Grids\,with\,feature\,counts \tag{8}$$

$$LatentSpace = N(nodes) \tag{9}$$

$$LatentSpace' = LatentSpace + Noise(W[start\_grid], W[end\_grid]) \quad for\,\forall W \tag{10}$$

The cluster information is used to guarantee the long-term spatial consistency of the behavior, and the building information is for finding semantically suitable buildings within a grid. The grid latent space is used for identifying the suitable search space for the building and defining the general area for the cluster point. It is important to note: if the data are shared on this abstraction level, it has to be ensured that the building information is selected so that an appropriate k-anonymity over the entirety of buildings in the observed area is guaranteed, especially when used in combination with the information from the grid latent space. For example, the precise number of floors and area of a building can lead to a simple reidentification, other than an activity-focused description like an OSM tag, such as: shop–groceries. Then, a noise value can be applied to the latent space representation, masking its original values.

The most computationally expensive approach would be adding the privacy requirement to the latent space in which the grids are searched. The encoding–decoding network is trained over the original features. When a trajectory must be generated, the network is again used on the H3 grid graph, but the grids' features within the trajectory are inverted. With this, not only do the original grids move within the latent space, but also, due to the convolutions in the graph neural network, the neighboring nodes change their latent space representation, making it less probable for those to be selected during the trajectory generation step. The encoding–decoding network needs retraining but must be reapplied over the area for every user. This approach may lead to the most indeterministic results, protecting the synthetic trajectories against attacks. Especially for long-term studies, where users are visiting the entirety of the observed area, the results may lead to unplausible behavior.

$$N = Neural\,Network\,for\,Latent\,Space\,Creation \tag{11}$$

$$During\,Training:\ N = f(nodes) \tag{12}$$

$$nodes' = Nodes \cup FeatureInverted(nodes_{OriginalWayChain}) \tag{13}$$

$$During\,Trajectory\,Construction:\ LatentSpace = f(nodes') \tag{14}$$

## 4. Results and Discussion

### *4.1. Data and Parameters*

To assess the proposed method, we use an app-based tracking dataset comprising 1096 participants tracked over 12 months (May 2022 to May 2023), within the scope of the

Mobilität.Leben study [32]. The tracking data are limited to trips within the Munich area and tracks are covered by individual, motorized vehicles. A random sample of 120 people is drawn. The key facts are shown in Table 4.

**Table 4.** Key data attributes from the MobilitaetLeben study, random draw, as used for the results.

| Attribute | Total Number | Average per Person | Median per Person |
|---|---|---|---|
| People | 120 | – | – |
| Area | 310.7 km$^2$ | – | – |
| Number of Tracks | 3151 | 25.2 | 10 |
| Distance of Tracks | 628.3 km | 5235.8 m | 3990.5 m |

The data are not representative for the average population sample from Munich but are biased toward university-related people. The data are exported with the following properties: [user id, linestring (WGS84 coordinates), timestamp start, timestamp end].

The H3-grids were enriched with the counts over several OSM tags. These OSM tags were: "shop", "residential", "office", "building", "leisure", "restaurant", "medical", "train_station", and "school". The H3 indexing provides grids in various sizes, dependent on a zoom level; an H3-resolution of 9 was chosen, leading to 511 grids in the observed area with an average area of 0.711 square kilometers per grid. The dataset consists of a distinct user id and multiple trajectories per person. Each trajectory is given a timestamp and a WGS 84 linestring. The clustering for the grid generation is conducted with a DBSCAN algorithm. Because every endpoint of a trajectory can be treated as its cluster if no other point is near, the minimal samples per cluster are set to one. In contrast, the epsilon parameter for the distance of two core points is set to 15 m. The coordinate system for metric comparisons is EPSG:31468 (Grauss–Kruger Zone 3). The latent space is a five-dimensional vector with two values for the encoded features of the specific grid and three values from the graph-link autoencoder. The hidden layer of each network uses 12 channels each.

### *4.2. Result on Example Track*

The results on a subset of tracks of an exemplary user are shown first. Figure 7 shows a subset of the trajectories on the left side and the corresponding clusters and their connection on the right side. The color of the cluster center indicates the betweenness value with the gravitational center; the node with the highest value is visible.

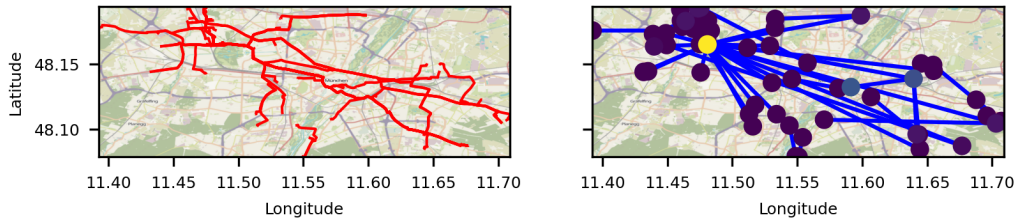


**Figure 7.** The original tracked data for car usage and the resulting clustering. The node that is selected as gravitational center is colored yellow.

Figure 8 shows the nodes in their latent space. A t-SNE [39] algorithm is used for the two-dimensional visualization over the described five-dimensional encoding vector. The coloring shows the distance from the grid in the city center by the number of grids in between. The figure shows several clustered grids on the left side, which are, on average, closer to the city center than on the right. More distinction is visible there. This indicates that a more complex distinction is needed in comparison to a naive approach. Such an approach could consist of dividing the city into a low number of city districts or just using terms like "city center" or "suburbs" as characterizing categories for neighborhoods. In this latent space, the selection of the areas to begin and end ways, is taken. The naive approach is

chosen for privacy measurement. This means a difference in the latent space between an original and a chosen node for the synthetic trajectories is required. The two sets of nodes in the original ways and those used for the synthesized mobility are not always direct neighbors. This results from the distance-based loss terms in Equations (2) and (3).

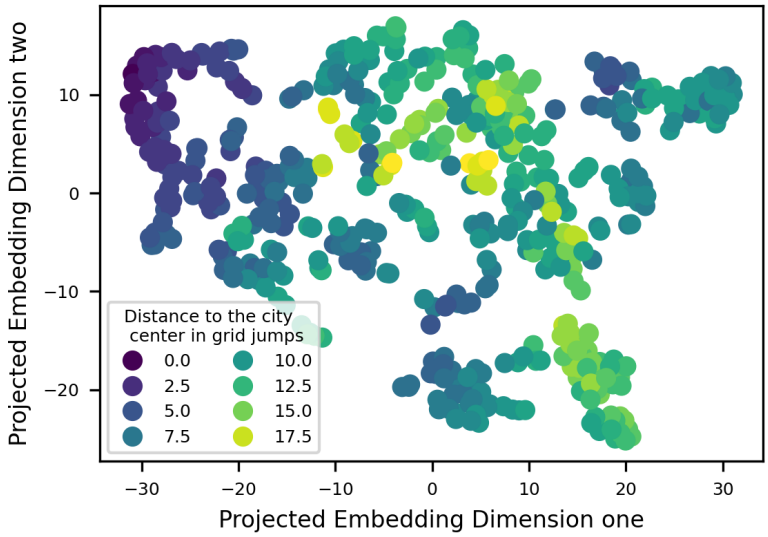


**Figure 8.** t-SNE projection of the grid latent space from the 5-dimensional encoding. The colors indicate the distance in H3 grids from the city center.

Figure 9 shows the result after applying the synthetic trajectory generation. The synthetically generated tracking data are shown on the right side, and the cluster view is on the left. Like in Figure 7, the betweenness of the clusters is color-coded, allowing for an easier reidentification and comparison between the figures. The node, marked in yellow, symbolizes the gravitational center, which is part of the distance loss function of every found cluster.

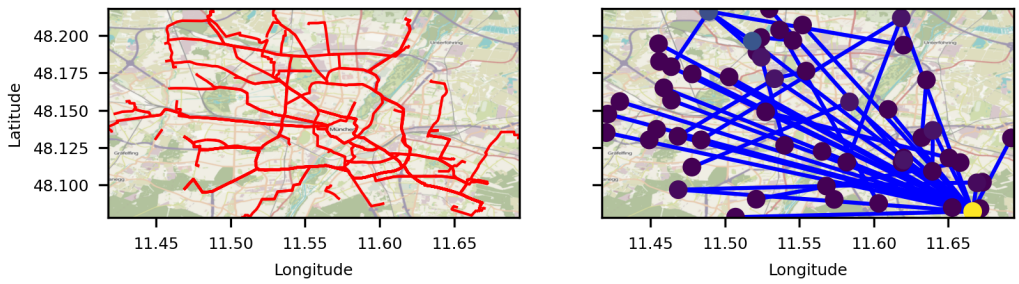


**Figure 9.** The synthetic data and the resulting clustering. Again, the center node is colored yellow.

The boxplot in Figure 10 shows the distribution of track distances between the original and the synthetic data and the distance error distribution from cluster to cluster. The plot giving the distance between the original and synthetic data shows the distance of an end or starting point of an original way compared to the same way in the artificial dataset. The algorithm only compares a point to its counterpart. Thus, it is still possible that the trajectories are nearer without breaking the privacy requirements, for example, if the working place in the synthetic dataset is near the home area in the original data. The minimum distance between the two points is 0.86 km, and the maximum is 12.1 km; in comparison, the area where the data were recorded has a total length from west to east of 26 km. The generated way chain has no simple shift as if only the neighboring grids to the original way chain were used.

The second set of plots shows the difference in the length of the routed linestrings in the newly generated ways against the length of the original tracked ways. Naturally, there is no way in the tracked dataset with a length of 0 m; the minimum is 0.3 km. Artificial generation identifies clusters and has every step start and end at one. For example, if the tracked movement of a person includes somebody going around a building and ending adjacent to the start, both start and end are associated with the same cluster. If this trajectory is synthesized, it will include the cluster as start- and endpoint and a length of 0.

The median error lies at −1.6 km. Negative error means the synthetic way is shorter than the original, and positive means that the artificial way is more prolonged. If the absolute error is taken, the median lies at 2.2 km. The median non-absolute distance error is positive, meaning the synthetic generator is biased, so longer ways are more probable to be generated. Additionally, the original data contain some measurement errors. For some trajectories, only the start- and endpoints exist. This leads to a linear distance between these two points in the original data but a routed way in the synthetized dataset.

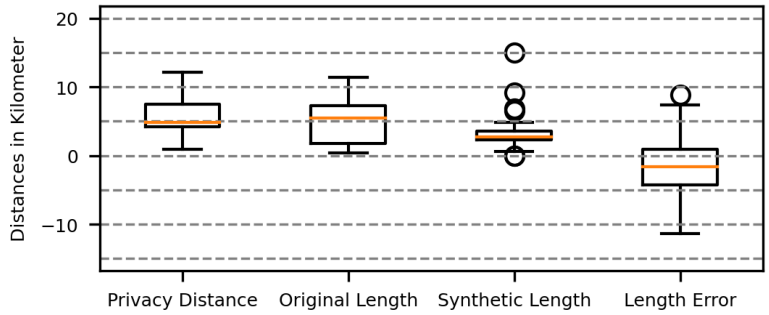


**Figure 10.** Distance from the original ways to the artificial ones and differences in the track lengths for the example ways.

### *4.3. Difference in Lengths over the Whole Dataset*

The following results are generated using all data described in Table 4. Figure 11 shows the differences over all 3151 ways of the 120 people in the same format as Figure 10.

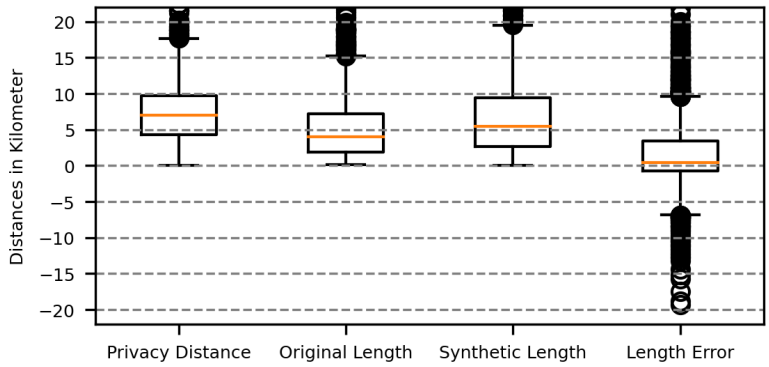


**Figure 11.** Distance from the original ways to the artificial ones and differences in the track lengths aggregated over the whole dataset.

The overall median of all track lengths in the tracked data is 3990 m and within the synthetic data 5496 m leading to a median error of 1473 m. When not considering positive and negative errors, the median absolute error is 1617 m. Like the single person's data, the artificial ways are biased to be longer than the original ones.

There are several outliers, especially with distances over 30 km. With a maximum west–east axis of 26 km and a north–south axis, long drives are abnormal within this area. These single tracks are ways that have no detected stops over a long period of over 8 h and

start and end at the exact same location. During this period no stops longer than 5 min were detected, leading to the assumption that the person is driving for professional reasons and not to a single destination. The algorithm does not yet cover this case.

### *4.4. Difference in Activity per Grid over the Whole Dataset*

The plausibility of the semantics of the generated ways is more complex. A main concern is connecting buildings with similar attributes and their geographical location and surroundings. One hypothesis is that the synthetic dataset should have comparable characteristics to the original. A possible plausibility check is the aggregated activity value per region. Figure 12 shows the activity in the original data aggregated over an H3 grid of resolution 7. The resolution is two steps coarser than the one used within the way generation algorithm. The activity value is calculated as the sum of all start- and endpoints within the area of each H3 grid.

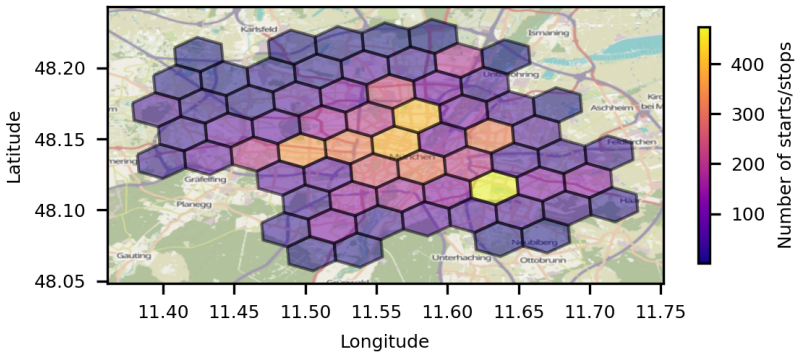


**Figure 12.** Start- and endpoints of the original ways aggregated over H3 grids with resolution 7.

Figure 13 shows the same metric generated from the synthetic ways. To some extent, the concentration within the inner city is still visible, as well as the extended activity to the northern and southern parts of the region.

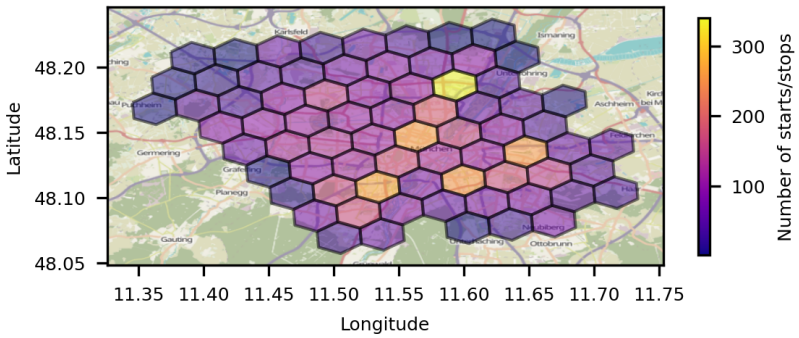


**Figure 13.** Start- and endpoints of the artificial ways aggregated over H3 grids with resolution 7.

To give some numerical context to the metric of area activity matching, Table 5 summarizes the critical values for the experiment. The value Intersection over Union (IoU), also known as the Jaccard Index [40], is comparable with its application in image processing. However, the exact calculation is refined for this use case. Every start- and endpoint of a way per grid is meant as one activity point. The union of these points between the original and synthetic ways is the maximum of those counts per grid. The intersection is the minimum value per grid.

**Table 5.** Area-dependent metrics from the synthetic tracks.

| | Intersection | Union | IoU in Percent |
|---|---|---|---|
| Participant from the example | 23 | 217 | 10.6 |
| Single participant with peak IoU | 50 | 118 | 42.5 |
| All ways | 3945 | 8656 | 45.6 |

Compared to image recognition tasks, where in most cases, the number of classes is limited, and for each class, each pixel can only have a value of zero or one; due to the calculation of the value, non-binary results are typical. For example, a particular grid may have in the original dataset 50 points, but in the synthetic datasets, only 40 start- and endpoints are within the grid border. This would result in a union value of 50, an intersection value of 40, and an IoU of only 80 percent. This leads to lower values than typically achieved in image recognition tasks because instead of the Boolean fact, that there is an intersection, there is an additional quality measure of this intersection.

Figure 14 indicates the value development over the experiment. The single values for each participant are in red, while the blue plot shows the aggregated metric. The dotted red line shows the maximum IoU of a single person's way syncretization, and the blue line is the maximum of the aggregated analysis. Every artificial way chain has a worse value than the experiment-wide aggregated evaluation.

Table 5 displays the IoU and connected values for the entire dataset, the individual with the highest single IoU resulting from the specific synthesis method, and the person referenced in Section 4.2.

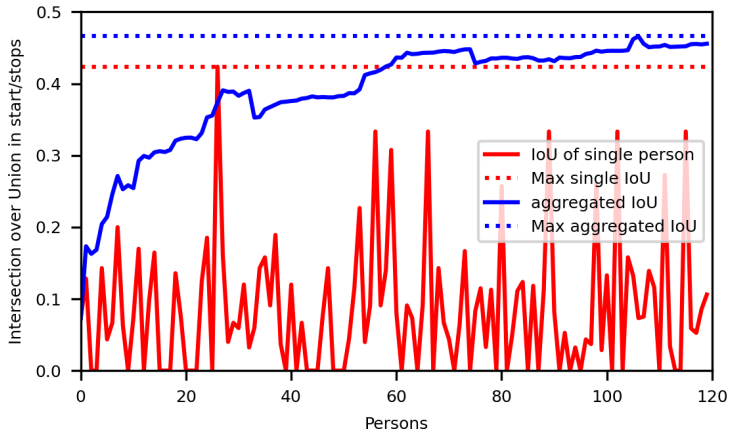


**Figure 14.** Intersection over union of start- and endpoints between the original and the artificial dataset aggregated over H3 grids with resolution 7.

### *4.5. Privacy*

The proposed method aims to share data on the original abstraction level without revealing personal details such as home or work locations. While the previous chapter explained how original characteristics are derived, the specific original locations need to be concealed. To achieve this, the start- and endpoints of an artificial path must not fall within the same H3 grid as the original data, ensuring that the exact locations are not disclosed. Figure 11 displays the distances from the original to the artificial dataset for each start-/endpoint of every track. The histogram highlights another important aspect—with a minimum distance of 17 m, no original point was exposed. Although an original and a synthetic waypoint cannot overlap, the distance can be as close as neighboring buildings if the algorithm selects adjacent grids. However, the generator's average behavior must not involve simply biasing or adding noise within a small radius to waypoints. This is not the case in this study, as the median distance is 7025 m, with the deviation between the q1 and q3 markers being 5353 meters.

The shown metrics prevent simple location-based attacks, where an attacker knows a single location of a person, e.g., the workplace or home. The attacker then tries to reidentify a person based on a single- or multipoint matching. The frequency of visits at specific locations remains unchanged even if those are not at the same spot as in the original dataset. This may expose the artificial datasets if mixed with the original data. This reidentification is the scope of most traditional privacy penetration, e.g., shown in [26,41].

In order to carry out these types of attacks, some information about the person being tracked is still necessary, such as how often they come to work. This information could potentially be gathered from other data sources. For instance, even if the specific visited locations are different from those in the original dataset, the visit frequency remains consistent. This could potentially allow a reidentification of individuals in the synthetic dataset when combined with context from the original data. This type of reidentification is the focus of most traditional privacy breaches, as discussed in previous works [26,41].

## 5. Summary, Limitations, and Further Applicability

In this paper, the possibility of sharing personal data obtained by a mobility tracking solution by creating an artificial dataset is discussed. It reflects some of the main characteristics but masks the critical information about where the single participant was during the study. This approach focuses on typical applications from the term of mobility research. The main goal is to keep the abstraction level of the data. This allows models that are developed and pretrained on such sharable data to be applied to the original data without further adjustments.

Another unique attribute of the method is that the mobility behavior per person is guaranteed to be stable regarding reoccurring locations within a single person's way chain. This means that if the person visits a specific workplace several times, it remains the same place within the artificial data. This is achieved by clustering the target and start points, creating a graph, representing the mobility behavior by taking the clusters as nodes and the ways between them as edges. The node with the highest betweenness value is taken as the "middle point" or "gravitational center" of the whole mobility behavior of the person.

A latent space of the H3 grids within the region is created by combining the results of a feature autoencoder and a link prediction autoencoder. Then, every grid has values representing its features and a vector representing the connection within the area. Within this latent space are new grids, representing the original H3 grids, where the ways started and ended are searched. This happens by sorting all grids based on the similarity within the latent space in relation to the grid where the original data are set. Additionally, the distance in the geographical space from the grid in the way chain before and the geographical distance from the mobility gravitational center is considered. Enforcing rules over the grid selection guarantees privacy criteria, and the concrete locations are found within the selected grids.

As an exemplary use-case, this method is applied to a dataset from the study of Moblitaet.Leben. The results show that the method is functional. Some current limitations include that only location-based attacks are prohibited. A reidentification is still possible when using frequency-based attacks, meaning the original data must not be shared. Another limitation of the current implementation is that all points of the participant's way chain are relocated, even those that are not compromising. If a mobility tracking study with company employees was used as input data, the location of such a company might be allowed to be in the synthetic result. In this case, a significant k-anonymity over all participants is still guaranteed in company-related locations.

The global availability of OpenStreetMap makes this method useful worldwide. The scalability of H3 grids over different zoom levels allows for further customization of the underlying geographic information structure and covered region. However, the requirements for the dataset limit the number of useful applications. These requirements include:

1. Long-term user tracking.
2. Unique user ID for extended periods.
3. Possible trip-to-POI matching, including exact origin and destination points.

These requirements should be valid for privacy-sensitive tracking data, which are typically not publicly available.

Another future research focus lies within the embedding and route selection; at the moment only driveways are usable. Future applications need to be multimodal mobility-ready. Thus, the embedding and routing need to incorporate more transportation networks like bikes and public transport and use non-point-of-interest data like demographic factors over residential areas. Moreover, the state of the traffic during the original recording needs to be taken into account to improve the routing. Once this is carried out, the data can be further enhanced by generating GPS data along the route to match the recorded data structure better. Then, the recording noise and data collection frequency can be included.

## 6. Conclusions

In conclusion, this paper presents an innovative approach to data anonymization in long-term mobility tracking studies. An artificial dataset is generated that maintains the core characteristics of the original data, while the privacy of participants is kept. By this method, key challenges in sharing personal mobility data for research purposes are addressed.

The integrity of mobility patterns are preserved by the technique of creating a latent space through the combination of feature and link prediction autoencoders. Afterward, the intelligent selection of grids based on similarity and geographical criteria is executed. This leads to synthesized way chains without compromising personal privacy. The proposed method concentrates on the recorded long-term mobility behavior. It extends previous methods by providing a possibility to keep the consistency of individual mobility over long recording periods.

In addition, the main characteristics of the single participants are preserved, and basic personal information is hidden. However, while effective against location-based attacks, the potential vulnerability to frequency-based reidentification attacks highlights the need for further refinement.

Additionally, the current limitations regarding non-compromising locations and the reliance on driveway information suggest areas for future development. Ensuring that the methodology is adaptable to multimodal mobility patterns and incorporates a broader range of data types will significantly enhance its applicability and utility in advancing mobility research while safeguarding participant anonymity. Another future research field is the incomprehension of trajectory attributes, like street types. Following the construction of not only start-/endpoints but the artificial generation of complete trajectories. This means the construction of not only start-/endpoint tuples but the creation of a complete artificial trajectory with GPS waypoints every 3 s, leading to a closer data density like real data recordings.

**Author Contributions:** Conceptualization, Fabian Netzler and Markus Lienkamp; methodology, Fabian Netzler; software, Fabian Netzler; validation, Fabian Netzler; data curation, Fabian Netzler; writing—original draft preparation, Fabian Netzler; writing—review and editing, Fabian Netzler and Markus Lienkamp; visualization, Fabian Netzler; supervision, Markus Lienkamp; funding acquisition, Markus Lienkamp All authors have read and agreed to the published version of the manuscript.

**Funding:** The research was conducted with basic research funds from the Institute of Automotive Technology, Technical University of Munich.

**Data Availability Statement:** For more information about the used data, please refer to https://www.hfp.tum.de/hfp/tum-think-tank/mobilitaet-leben/ (accessed on 14 April 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the study's design, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| OSM | OpenStreetMap |
| LSTM | Long Short-Term-Memory |
| POI | Point of Interest |

## References


1. Geisslinger, M.; Karle, P.; Betz, J.; Lienkamp, M. Watch-and-Learn-Net: Self-supervised Online Learning for Probabilistic Vehicle Trajectory Prediction. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 869–875. [CrossRef]
2. He, T.; Bao, J.; Li, R.; Ruan, S.; Li, Y.; Song, L.; He, H.; Zheng, Y. What is the Human Mobility in a New City: Transfer Mobility Knowledge Across Cities. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; Huang, Y., King, I., Liu, T.Y., van Steen, M., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2020; pp. 1355–1365. [CrossRef]
3. Xue, A.Y.; Zhang, R.; Zheng, Y.; Xie, X.; Huang, J.; Xu, Z. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE 2013), Brisbane, QLD, Australia, 8–12 April 2013; Jensen, C.S., Ed.; IEEE: Piscataway, NJ, USA, 2013; pp. 254–265. [CrossRef]
4. Monroe, C.; Tazi, F.; Das, S. Location Data and COVID-19 Contact Tracing: How Data Privacy Regulations and Cell Service Providers Work In Tandem. In Proceedings of the Workshop on Usable Security and Privacy (USEC), Auckland, New Zealand (Virtual), 7 May 2021.
5. Kong, X.; Chen, Q.; Hou, M.; Wang, H.; Xia, F. Mobility trajectory generation: A survey. *Artif. Intell. Rev.* **2023**, *56*, 3057–3098. [CrossRef]
6. Guo, X.; Li, G.; Chen, Z.; Zhang, H.; Ding, Y.; Wang, J.; Zhao, Z.; Tang, L. Large-Scale Human Mobility Prediction Based on Periodic Attenuation and Local Feature Match. In Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge, New York, NY, USA, 13 November 2023; pp. 16–21. [CrossRef]
7. Jiang, R.; Song, X.; Fan, Z.; Xia, T.; Wang, Z.; Chen, Q.; Cai, Z.; Shibasaki, R. Transfer Urban Human Mobility via POI Embedding over Multiple Cities. *ACM/IMS Trans. Data Sci.* **2021**, *2*, 1–26. [CrossRef]
8. Choi, S.; Kim, J.; Yeo, H. TrajGAIL: Generating urban vehicle trajectories using generative adversarial imitation learning. *Transp. Res. Part C Emerg. Technol.* **2021**, *128*, 103091. [CrossRef]
9. Raczycki, K.; Szymański, P. Transfer learning approach to bicycle-sharing systems' station location planning using OpenStreetMap data. In Proceedings of the 4th ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities, Virtual, 2 November 2021; Kar, B., Fu, G., Mohebbi, S., Ye, X., Omitaomu, O.A., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2021; pp. 1–12. [CrossRef]
10. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
11. Woźniak, S.; Szymański, P. hex2vec. In Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Beijing China, 2–5 November 2021; Lunga, D., Yang, L., Gao, S., Martins, B., Hu, Y., Deng, X., Newsam, S., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2021; pp. 61–71. [CrossRef]
12. Jenkins, P.; Farag, A.; Wang, S.; Li, Z. Unsupervised Representation Learning of Spatial Data via Multimodal Embedding. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; Zhu, W., Tao, D., Cheng, X., Cui, P., Rundensteiner, E., Carmel, D., He, Q., Xu Yu, J., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2019; pp. 1993–2002. [CrossRef]
13. Leśniara, K.; Szymański, P. highway2vec. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Seattle, WA, USA, 1 November 2022; Lunga, D., Newsam, S., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2022; pp. 18–29. [CrossRef]
14. Donghi, D.; Morvan, A. GeoVeX. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Hamburg, Germany, 13 November 2023; Newsam, S., Yang, L., Mai, G., Martins, B., Lunga, D., Gao, S., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2023; pp. 3–13. [CrossRef]
15. Zhang, L.; Long, C. Road Network Representation Learning: A Dual Graph-based Approach. *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–25. [CrossRef]
16. Du, J.; Chen, Y.; Wang, Y.; Pu, J. Zone2Vec: Distributed Representation Learning of Urban Zones. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 880–885. [CrossRef]
17. Shin, Y.; Seong, G.; Kim, N.; Kim, S.; Yoon, Y. Understanding Urban Economic Status through GNN-Based Urban Representation Learning Using Mobility Data. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Advances in Urban-AI, New York, NY, USA, 13 November 2023; pp. 71–80. [CrossRef]
18. Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; Ermon, S. Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3967–3974. [CrossRef]

19. Lao, L.; Du, D.; Chen, P. Predicting Pedestrian Trajectories with Deep Adversarial Networks Considering Motion and Spatial Information. *Algorithms* **2023**, *16*, 566. [CrossRef]
20. Zhu, Y.; Ren, D.; Xu, Y.; Qian, D.; Fan, M.; Li, X.; Xia, H. Simultaneous Past and Current Social Interaction-Aware Trajectory Prediction for Multiple Intelligent Agents in Dynamic Scenes. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 1–16. [CrossRef]
21. Park, S.H.; Kim, B.; Kang, C.M.; Chung, C.C.; Choi, J.W. Sequence-to-Sequence Prediction of Vehicle Trajectory via LSTM Encoder-Decoder Architecture. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1672–1678. [CrossRef]
22. Messaoud, K.; Yahiaoui, I.; Verroust-Blondet, A.; Nashashibi, F. Relational Recurrent Neural Networks For Vehicle Trajectory Prediction. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1813–1818. [CrossRef]
23. Rao, J.; Gao, S.; Kang, Y.; Huang, Q. LSTM-TrajGAN: A Deep Learning Approach to Trajectory Privacy Protection. In Proceedings of the International Conference Geographic Information Science, Seattle, WA, USA, 3–6 November 2020. [CrossRef]
24. Cao, C.; Li, M. Generating Mobility Trajectories with Retained Data Utility. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; Zhu, F., Ooi, C., Miao, C.B., Wang, H., Skrypnyk, I., Hsu, W., Chawla, S., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2021; pp. 2610–2620. [CrossRef]
25. Alatrista-Salas, H.; Montalvo-Garcia, P.; Nunez-del Prado, M.; Salas, J. Geolocated Data Generation and Protection Using Generative Adversarial Networks. In *Modeling Decisions for Artificial Intelligence*; Lecture Notes in Computer Science; Torra, V., Narukawa, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2022; Volume 13408, pp. 80–91. [CrossRef]
26. Wu, W.; Shang, W.; Lei, R.; Yang, X. A Trajectory Privacy Protect Method Based on Location Pair Reorganization. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 8635275. [CrossRef]
27. Bundesministerium für Verkehr und digitale Infrastruktur. *Mobilität in Deutschland Ergebnisbericht: Technical*; Bundesministerium für Verkehr und digitale Infrastruktur: Berlin, Germany, 2018.
28. Ecke, L.; Vallee, J.; Chlond, B.; Vortisch, P. *Deutsches Mobilitätspanel (MOP)—Wissenschaftliche Begleitung und Auswertungen Bericht 2022/2023: Alltagsmobilität und Fahrleistung*; Karlsruher Institut für Technologie (KIT): Karlsruhe, Germany, 2023. [CrossRef]
29. Schweizer, J.; Poliziani, C.; Rupi, F.; Morgano, D.; Magi, M. Building a Large-Scale Micro-Simulation Transport Scenario Using Big Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 165. [CrossRef]
30. Moeckel, R.; Kuehnel, N.; Llorca, C.; Moreno, A.T.; Rayaprolu, H. Agent-Based Simulation to Improve Policy Sensitivity of Trip-Based Models. *J. Adv. Transp.* **2020**, *2020*, 1902162. [CrossRef]
31. Moeckel, R.; Huang, W.C.; Ji, J.; Moreno, A.T.; Llorca, C.; Staves, C.; Zhang, Q.; Erhardt, G. The Activity-Based Incremental Model (ABIT): Modeling 24 Hours, 7 Days per Week. In Proceedings of the Euro Working Group Transportation, Santander, Spain, 6–8 September 2023.
32. Loder, A.; Cantner, F.; Adenaw, L.; Nachtigall, N.; Ziegler, D.; Gotzler, F.; Siewert, M.B.; Wurster, S.; Goerg, S.; Lienkamp, M.; et al. Observing Germany's nationwide public transport fare policy experiment "9-Euro-Ticket"—Empirical findings from a panel study. *Case Stud. Transp. Policy* **2024**, *15*, 101148. [CrossRef]
33. Freeman, L.C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* **1977**, *40*, 35. [CrossRef]
34. Pellungrini, R.; Pappalardo, L.; Pratesi, F.; Monreale, A. A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. *ACM Trans. Intell. Syst. Technol.* **2017**, *9*, 1–27. [CrossRef]
35. Uber Technologies Inc. [n.d.]. H3: Uber's Hexagonal Hierarchical Spatial Index. Available online: https://eng.uber.com/h3 (accessed on 18 March 2024).
36. Gramacki, P.; Leśniara, K.; Raczycki, K.; Woźniak, S.; Przymus, M.; Szymański, P. SRAI. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Hamburg, Germany, 13 November 2023; Newsam, S., Yang, L., Mai, G., Martins, B., Lunga, D., Gao, S., Eds.; Association for Computing Machinery (ACM): New York, NY, USA, 2023; pp. 43–52. [CrossRef]
37. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
38. Sakuma, Y.; Tran, T.P.; Iwai, T.; Nishikawa, A.; Nishi, H. Trajectory Anonymization through Laplace Noise Addition in Latent Space. In Proceedings of the 2021 Ninth International Symposium on Computing and Networking (CANDAR), Matsue, Japan, 23–26 November 2021; pp. 65–73. [CrossRef]
39. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
40. Jaccard, P. Lois de distribution florale dans la zone alpine. *Bull. Soc. Vaudoise Sci. Nat.* **1902**, *38*, 69–130. [CrossRef]
41. Xu, Z.; Zhang, J.; Tsai, P.W.; Lin, L.; Zhuo, C. Spatiotemporal Mobility Based Trajectory Privacy-Preserving Algorithm in Location-Based Services. *Sensors* **2021**, *21*, 2021. [CrossRef] [PubMed]