Reconstructing activity locations from zone-based trip data for discrete choice modeling

Milos Balac

Sebastian Hörl

STRC 2021 **September 2021**

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Reconstructing activity locations from zone-based trip data for discrete choice modeling

Milos Balac
IVT, ETH Zürich
Zurich, Switzerland
`milos.balac@ivt.baug.ethz.ch`

Sebastian Hörl
IRT SystemX
Palaiseau, France
`sebastian.horl@irt-systemx.fr`

September 2021

Abstract

This paper presents a methodology to disaggregate activity locations from zone-based activity chain data usually reported in the anonymized travel surveys. We propose an algorithm that aims to find a feasible sequence of activity locations, for each individual, that minimizes the maximum error of each trip's Euclidean distance within the activity chain. The reconstructed activity locations are then used to create unchosen alternatives within the choice set for each individual. This is followed by the mode-choice model estimation. We test our approach on three large-scale travel surveys conducted in Switzerland, Île-de-France and São Paulo. We find that with our approach we can reconstruct activity locations that accurately match trip Euclidean distances, but with location errors that still provide location protection. The models estimated on the reconstructed locations perform similarly, in terms of goodness of fit and prediction, to the ones obtained on the original activity locations.

Keywords

anonymization, data privacy, travel survey, choice model, discrete choice

1 Introduction

One of the most critical parts of transport planning is transport modeling. It should be able to support transport planners in anticipating the impacts of policies and infrastructure projects. The collection of various transport-related data supports transport modeling. While today information can be collected through smartphone applications, transit tap-in/tap-out data, or mobile phone data, the traditional approach is to utilize (household) travel surveys. These surveys, also referred to as revealed preference (RP) surveys, usually collect detailed sociodemographic information on individuals living in the area of interest together with their activity and trip behavior on one or multiple days of the week. The activities can frequently be identified by a GPS coordinate or detailed address. Typically, the gathered information on mobility behavior is enriched with unchosen alternatives for each trip based on the choice set for each individual. This serves as a preparatory step for further mode-choice modeling.

Due to privacy concerns and governing laws in many countries, the information in travel surveys has to be anonymized at a level that protects the identity of individuals and their link to the survey data. For this reason, identifying information like first and last name, home address, or coordinates of activities are removed. The location of activities in publicly available versions of surveys is usually published on a zonal level (i.e., traffic analysis zone, census zone). While this protects the interviewed individuals, it is unknown how this aggregation affects the generation of the unchosen alternatives, and subsequently, the modeling of the data and the forecasting power of the created models. Therefore, in this paper, we aim to answer these questions.

The paper is organized as follows. Section 2 goes over the current literature in data anonymization and its application to the field of transportation. Section 3 proposes a heuristic to reconstruct activity locations based on zone-based trip data and explains the subsequently used mode-choice modeling approach. Section 4 explains the used data sets, and Section 5 presents the results. After, Sections 6 and 7 provide discussion and closing remarks.

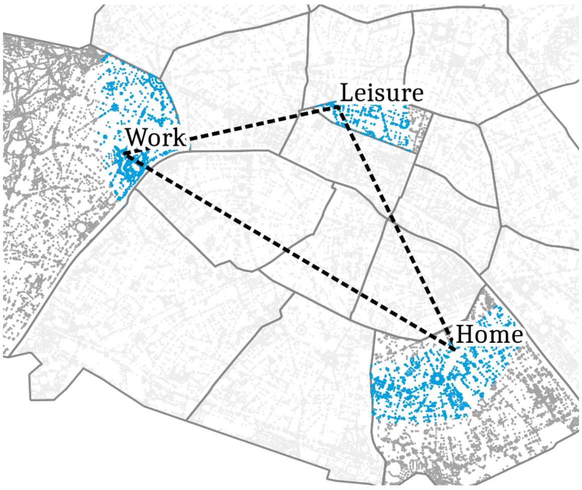
2 Background

With the increasing popularity of the open-data concept, the need to protect the privacy of individuals that provided their data has increased. One of the most usual pieces of information that needs to be anonymized is location. Techniques used to provide location protection aim to obscure the location of activities of individuals. Some of these techniques involve aggregation, spatial cloaking, or random perturbation (for a detailed overview of different mechanisms, please refer to Krumm (2009)). A typical example is perturbation of residential locations of surveyed individuals, where the anonymization procedure aims to maintain the usefulness of the data Badu-Marfo *et al.* (2019). The authors of Badu-Marfo *et al.* (2019) focus on analyzing the performance of different perturbation mechanisms for protecting the privacy of survey respondents. They also point out that current methods mainly deal with the anonymization of single points and that further research is needed in developing methods for multi-point data.

Travel surveys that collect the mobility behavior of respondents over a day or week have to deal with such multi-location data. Since each respondent reports multiple activities, a suitable technique needs to be utilized that protects the privacy of individuals while still maintaining the usefulness of the data. Most surveys utilize zone aggregation mechanisms (i.e., activity locations are provided on a zone level). In the United States, each activity is usually aggregated to the census tract (i.e., California Household Travel Survey Center (2021), or My Daily Travel Survey conducted in the Chicago Metropolitan Region Chicago Metropolitan Agnecy for Planning (2021)). In the case of France, multiple surveys exist. The publicly accessible national survey has a high degree of aggregation on the level of departments, which cover thousands or millions of residents. More local surveys, such as the one for the Île-de-France region around Paris, are only accessible on request and provide locations aggregated to a grid of 100x100 meters. A commonly used aggregation level in French data sets are municipalities with thousands to tens of thousands inhabitants. In São Paulo, the publicly available travel survey does not provide location protection. In constrast, publicly available Brazilian census data is aggregated to a census zone containing between 20 and 55 thousand people.

Even when privacy protection techniques are used, confidential data can be at risk if additional information obtained from other sources can uniquely identify individuals. For example, De Montjoye *et al.* (2013) show that mobile-phone traces provided in hourly intervals and with the spatial resolution provided by antennas can be uniquely identified in 95% of the cases with only four spatio-temporal points. Golle and Partridge (2009) show that by revealing home and work census tract information, the anonymity set (i.e., the number of potential matching individuals) has a median size of 21 for the case of the U.S. working population. This raises a potential privacy concern for anonymized travel or commuting surveys. Nevertheless, identifying the level of privacy that the location protection techniques bring to the respondents in these surveys is not a direct aim of this

Figure 1: Example of a feasible set of candidate points



paper, even though we provide some insights. We, however, aim to show how much the level of aggregation provided by the travel surveys could affect the prediction power of downstream models.

Therefore, to the best of our knowledge, we provide a first documented effort of the following aspects:

- We propose a heuristic that, based on anonymized and aggregated zone-based trip data, creates disaggregated activity locations for all trips conducted by interviewed individuals.
- We perform analyses on the prediction accuracy of discrete choice models estimated on the basis of non-anonymized location information versus reconstructed locations.
- We show the universality of our findings based on survey data from three different countries.

3 Methodology

3.1 Problem statement

Figure 1 shows a motivating example for our approach. It shows an activity chain with four activities, where a person starts the daily travels at home in the 13th arrondissement in

Paris, then goes to work close to the Eiffel tour which is located in the 16th arrondissement, continues to the Opera (2nd arrondissement) in the evening and then goes back home. In an anonymized travel survey, we may only know the Euclidean (and/or routed) distances between the activities, but also the zones in which the activities occur, represented by the arrondissements in this example. In dark gray, a set of possible activity locations in the zones has been obtained (here based on OpenStreetMap data). Furthermore, the Euclidean distances between all activities are known (exemplified by the dotted lines). If one now starts to move the locations of the four activities under the two conditions that (1) both "home" activities need to be at the same place, (2) Euclidean distances between the locations need to deviate no more than 50 meters from the reference distances, we arrive at a feasible set of locations which is colored in blue. The smaller the allowed deviation gets (e.g., 10 meters, 5 meters), the smaller the feasible set of locations will become. Ideally, if our set of possible activity locations represents well the locations used in the survey, one would find the exact locations by reducing the deviation to zero.

3.2 Location search problem

The algorithm to find locations for the activities in a chain of a specific person is described in the following. As input, we know the number of activities in the chain N, as well as whether each of the activities $i \in \{1, ..., N\}$ is a "home" activity. The indices of those activities are noted down in the index set $\mathcal{H}$. Furthermore, reference Euclidean distances are given as $r_i \in \mathbb{R}$.

The potential locations for the ith activity correspond to the potential locations in the respective zone. We denote the set of those locations as $\mathcal{L}_i$ and the set of all potential locations in the activity chain is $\mathcal{L} = \mathcal{L}_1 \cap ... \cap \mathcal{L}_N$. Let $k \in \{1, |\mathcal{L}|\}$ reference the elements of $\mathcal{L}$, then $y_{k,i}$ indicates whether location k is a potential location for the zone of activity i. The Euclidean distance between location k and k' is denoted as $d(k, k')$.

The aim of the algorithm is then to find a sequence $l = (l_1, ..., l_N)$ with $l_i \in \mathcal{L}_i$ such that (1) the location for each activity is located in the respective zone, and (2) "home" activities always take place at the same location. To select among the feasible locations, the maximum deviation of the generated distances along the chain, compared to the reference distances, is minimized. The optimization problem is defined by the following objective function

$$\underset{(l_1,\ldots,l_N)}{\text{minimize}} \quad \max_{i\in\{1,\ldots,N-1\}} \{ \ | \ d(l_i, l_{i+1}) - r_i \ | \ \} \tag{1}$$

with the following constraints:

$$\begin{aligned} y_{l_i,i} = 1 \qquad & \forall i \in \{1, \ldots, N\} \\ l_i = l_{\min \mathcal{H}} \qquad & \forall i \in \mathcal{H} \end{aligned} \tag{2}$$

The first constraint makes sure that activities along the sequence only take place in locations that belong to the respective zone. The second constraint requires that all home activities take place at the same location.

3.3 Solution strategy

The solution strategy aims to find a feasible and optimal sequence $(l_1, \ldots, l_N)$ for each person. The most straightforward approach would use a depth-first branch-and-bound algorithm, where we would start a chain at any location in the first zone, then extend these chains with locations from the second zone and after with succeeding zones until one complete chain is found. The maximum deviation along this chain can then be used to bound further exploration steps of the graph. Additionally, locations for home activities are set to the first occurrence of a home location along the constructed chain.

Our experiments have shown that such an approach causes very long run times if multiple times hundreds of potential locations need to be examined, especially for long activity chains. Hence, we perform a directed search where candidates in the following zones are chosen such that the local error is minimized. While the solutions of such an algorithm are not optimal, they perform well for the following modeling steps, as will be shown further below. Formally, the following depth-first branch-and-bound algorithm is proposed:

Note that location sequences are only extended in a best-response fashion using the closest successor in terms of minimizing the Euclidean distance error, rather than enumerating

ALGORITHM 1: Chain-based location assignment

```
Input:
Location sets $\mathcal{L}_1, ..., \mathcal{L}_N$ and $\mathcal{L}$
Home activity index set $\mathcal{H}$

Initialize:
C = []        $l^* = \emptyset$        $q^* = \infty$

For each $l_1 \in \mathcal{L}_1$
    $C \leftarrow ((l_1), 0)$
Continue

While $|C| > 0$
    $(l_1, ..., l_n), q_n \leftarrow$ pop $C$
    If $q_n < q^*$ Then
        If $n = N$ Then
            $q^*, l^* = q_n, l$
        Else
            If $n \in \mathcal{H}$ and $n > \min\mathcal{H}$
                $l_{n+1} = l_{\min\mathcal{H}}$
            Else
                $l_{n+1} = \arg\min_{l_u}\{|d(l_n, l_u) - r_i| \mid l_u \in \mathcal{L}_{n+1}\}$
            End
            $q_{n+1} = \max\{q_n, |d(l_n, l_{n+1}) - r_i|\}$
            $C \leftarrow ((l_1, ..., l_n, l_{n+1})), q_{n+1})$
        End
    End
Continue
Return $l^*$
```

all possible options. However, the algorithm can be easily modified to perform a complete enumeration if necessary.

3.4 Choice model

To test the impacts of location error on mode-choice model estimates, we make use of a straightforward logistic regression model. We model the mode-choice for trips where car or public transport were a chosen mode. Therefore, the choice set includes only public transport and private car. To obtain relevant characteristics of the two alternatives, we perform a minimum cost path routing for all car trips, based on road networks obtained from OpenStreetMap data and free flow speeds. For public transport, we use an implementation of the RAPTOR algorithm Delling *et al.* (2015) to find routes through the public transport network provided in GTFS format which minimize the total travel time of the trips. The data sets are documented in the scope of the development of synthetic populations for agent-based transport simulation for the three cases of São Paulo Sallard *et al.* (2021), Switzerland Tchervenkov *et al.* (2021) and Île-de-France Hörl and Balac (2021). As for some trips a public transport route cannot be found (i.e., the trip is too short, or public transport is not accessible), those trips are filtered out, which creates some differences in the size of the data set for reconstructed and original coordinates (see also Table 1). The mathematical formulation of the model is as follows:

$$\begin{aligned} \log\left(\frac{p_{car}}{1-p_{car}}\right) =&\alpha+ \\ &\beta_{hascar}\cdot\delta_{car}+\beta_{haslicense}\cdot\delta_{license} \\ &\beta_{invehicle,car}\cdot tt_{invehicle,car}+\beta_{invehicle,pt}\cdot tt_{invehicle,pt}+ \\ &\beta_{access,pt}\cdot tt_{access,pt}+\beta_{egress,pt}\cdot tt_{egress,pt}+ \\ &\beta_{transfer,pt}\cdot tt_{transfer,pt} \end{aligned} \tag{3}$$

where p_{car} is the probability of choosing a car. All independent variables are continuous except δ_{car} and $\delta_{license}$, which are dummy variables representing whether a person has a car or driver's license, respectively. $tt_{invehicle,car}$ represents the travel time by car, and $tt_{invehicle,pt}$, $tt_{access,pt}$, $tt_{egress,pt}$, and $tt_{transfer,pt}$ represent the in-vehicle travel time, access time, egress time, and transfer time of public transport alternative. In São Paulo, driver's license information was not collected and, therefore, is not used in the models for São Paulo.

For each of the three case studies denoted by i, we estimate two models, one based on the original coordinates M_i^o and one based on the reconstructed coordinates M_i^r. To compare the predictive power of these two models, we split both data sets into a training set containing 70% (T_i^o and T_i^r) and a test set containing 30% (V_i^o and V_i^r) of the data by ensuring that the same trips are contained in both (i.e., T_i^o and T_i^r contain the same trips, but with different routing data). We train both M_i^o and M_i^r on the respective training set T_i^o and T_i^r. Finally, we analyze the predictive accuracy of the trained models on V_i^o data.

All models are estimated using the `scikit-learn` package in Python Pedregosa *et al.* (2011).

4 Case study

We make use of the already existing travel surveys from Switzerland Swiss Federal Office of Statistics (BFS) and Federal Office for Spatial Development (ARE) (2018), Île-de-France Île-de-France Mobilités *et al.* (2010), and Greater São Paulo Metropolitan Region Secretaria Estudal dos Transportes Metropolitanos and Companhia do Metropolitano de São Paulo – METRÔ (2019) to create the inputs for the reconstruction algorithm and the downstream mode-choice model estimation.

4.1 Switzerland

The *Mikrozensus Mobilität und Verkehr* (Swiss Federal Office of Statistics (BFS) and Federal Office for Spatial Development (ARE), 2018) is a national travel survey conducted every five years in Switzerland. For the last edition conducted in 2015, about 56 000 persons ($\simeq 0.6\%$ of the total Swiss population) are asked questions about their mobility behavior and their socio-demographic attributes. Disaggregated, coordinate-level information about activities is available to the research community upon request. The aggregated zonal information used in this study comes from the National transport Model Bundesamt für Raumentwicklung (ARE) (2020).

4.2 Île-de-France

The *Enquête globale de transport* (EGT, Île-de-France Mobilités *et al.*, 2010) is a household travel survey conducted in the Île-de-France region, mainly during the year 2010. The EGT contains the trip chains of around 35 000 respondents in 15 000 households in the Île-de-France region. These numbers translate to a sample of around 0.3% of people living in the region. Within Île-de-France, around 122 000 trips are reported of all the members in each household. Unfortunately, EGT is only available on request from the regional authorities and therefore not publicly available. Activity locations are reported on a grid of 100x100 meters. As zoning data, French municipalities are used.

4.3 São Paulo

The last household travel survey in the Greater São Paulo Metropolitan Region was conducted in 2017 and is publicly available Transportes Metropolitanos (2017). It contains

84 889 weighted samples. For each sample, both person and household-level information is provided. Unfortunately, no driver's license information is available. Locations of activities performed by the respondents are reported with coordinate accuracy. The dataset also provides a traffic zone for each of the activities, which are then used to test the performance of the disaggregation algorithm.

4.4 Candidates

For the three cases, multiple sets of candidate points are created, among which the locations of the activities can be chosen. Two different ways of generating such points are looked at.

First, we sample points at random for each zone in the three use cases. To do so, we obtain the bounding box of each zone, sample N points within the bounding box, and then keep those points that fall inside the zone boundaries. The number of points is defined as $N = A \cdot \eta$ with A being the bounding box area and η a configurable density. In the experiments below, densities of 1, 5, 10, and 20 km^{-2} are used.

Second, we obtain OpenStreetMap data for each case. We filter for all road geometries that are included or intersect with the case study area and use the *nodes* of the remaining road shapes as location candidates.

5 Results

5.1 Reconstruction process

First, the results of the reconstruction algorithm are presented. We examine the *distance errors* and the *location errors* produced by the reconstruction algorithm. The *distance error* is defined as the absolute difference between the Euclidean distance of a trip from the original data set and the Euclidean distance between the selected location candidates. It is, hence, a measure of how well the algorithm can recover the reference distances. The *location error* represents the distance between an activity's location in the reference data set and its location. Therefore, it is a measure of how well the algorithm reconstructs the original locations. Note that it is a validation measure, as in the general case (with an anonymized data set), the original locations would not be available.

Figure 2 shows the cumulative distribution function of both error types for the three use cases. In all cases, we observe that the distance error decreases strongly with an increased density of the location candidates, as more options allow a more fine-grained assignment. Furthermore, the OSM-based assignment performs the best in terms of reducing the distance error. For the location error, the same effects can be observed.

Interestingly, using the OSM candidates, the distance error is reduced to zero for almost all trips, i.e., point sequences that match the actual distances can be found in almost every case. The Euclidean distances are, hence, replicated almost perfectly.

The results on the location error are essential in terms of identifying specific activity locations. Even with the high-density OSM data, locations can not be reconstructed perfectly. For Switzerland, however, 90% of activities are located within 1km of the original location. For Île-de-France and São Paulo, this threshold is reached at about 2km. On the contrary, more than 50% of locations in Switzerland can be reconstructed with an accuracy of 300m.

While Figure 2 gives a general impression on the matching performance of the algorithm, it is interesting to analyze how errors are distributed spatially. Figure 3 shows the location error, capped at 2km, for the three use cases. A high matching performance can be observed for Switzerland for the finely zoned and highly populated areas around Zurich in the North and along the Geneva lake in the South-West. On the contrary, the sparsely populated and coarsely zoned areas in the Alps can be identified clearly as a strip of high location errors. For Île-de-France, errors are distributed somewhat randomly across space, especially no increase in accuracy can be observed for Paris and its metropolitan region, which would otherwise stick out in the center of the map. For São Paulo, accuracy is very low in the outer regions, where enormous zones contain large, unpopulated areas. Accuracy, however, increases towards the city center of São Paulo.

5.2 Model estimation

Table 1 presents the models estimated for different study areas and activity locations source. All parameters have the expected sign and are significant at 0.1% level. The

Figure 2: Distance and location errors after the matching process

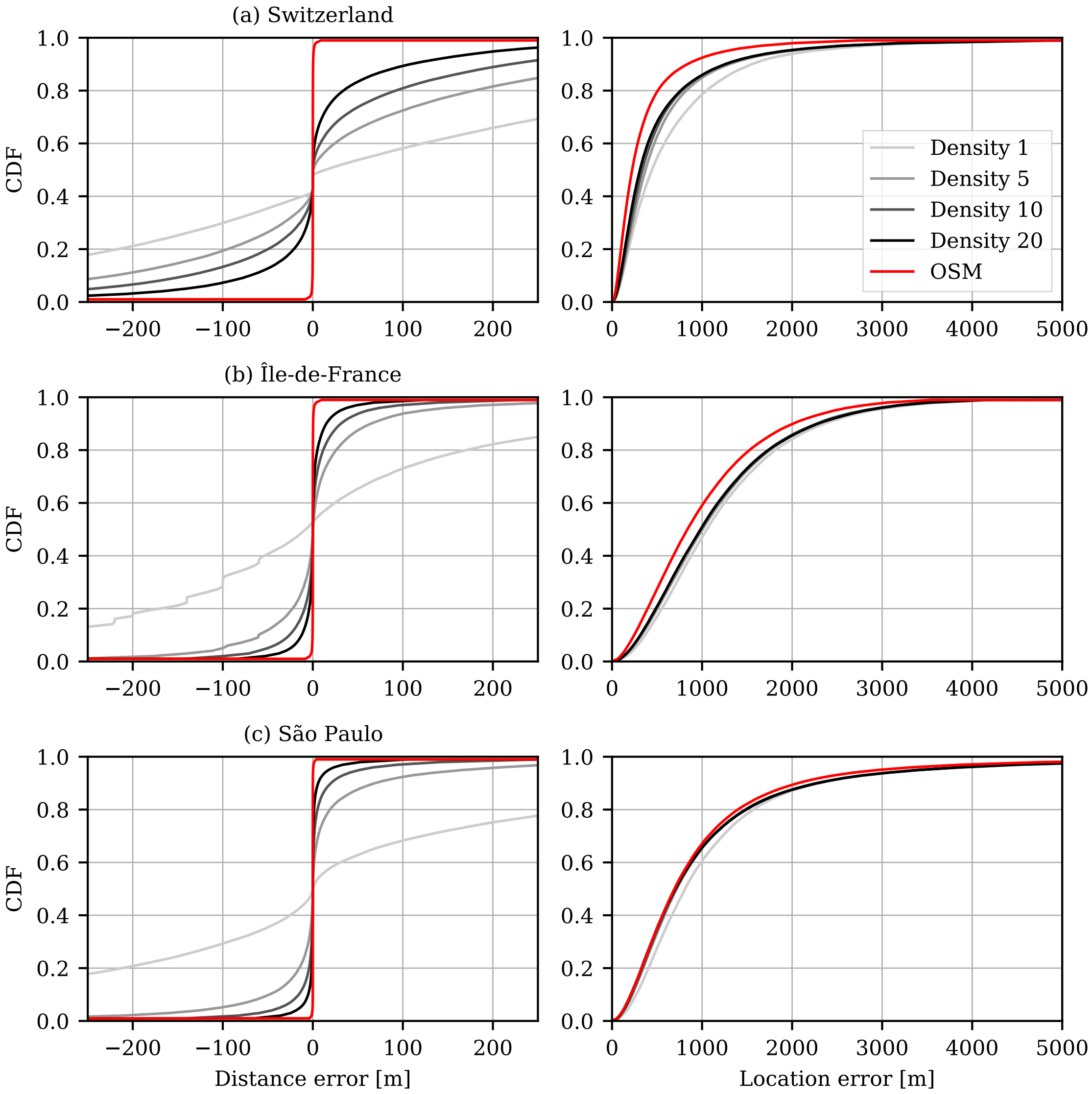


Figure 3: Spatial distribution of the location error (from left to right: Switzerland, Île-de-France, São Paulo)

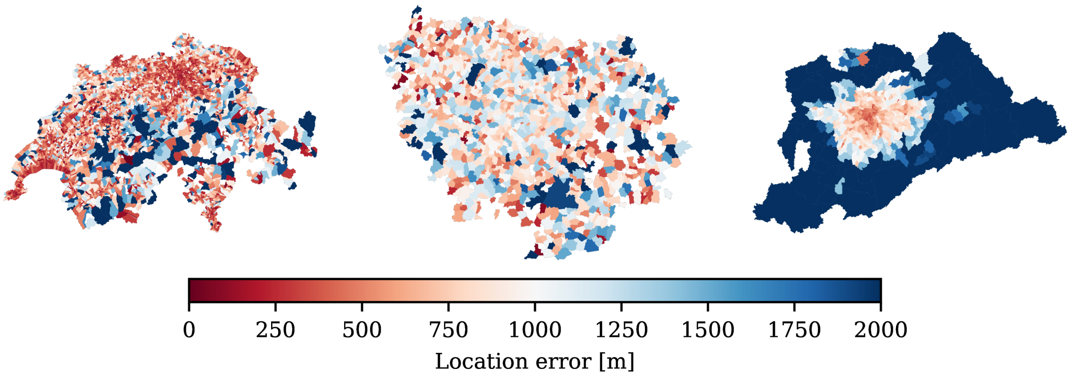


Table 1: Models estimated for three study areas for original and reconstructed coordinates.

Parameter		Switzerland Rec.	 Orig.	Île-de-France Rec.	 Orig.	São Paulo Rec.	 Orig.
α		-2.954	-2.748	-7.989	-7.771	-2.368	-2.379
$\beta_{access,pt}$	$[min^{-1}]$	0.040	0.036	0.039	0.033	0.009	0.010
$\beta_{egress,pt}$	$[min^{-1}]$	0.042	0.040	0.039	0.037	0.008	0.008
$\beta_{waiting,pt}$	$[min^{-1}]$	0.035	0.045	0.040	0.058	0.012	0.013
$\beta_{invehicle,pt}$	$[min^{-1}]$	0.006	0.004	0.017	0.025	0.012	0.016
$\beta_{invehicle,car}$	$[min^{-1}]$	-0.052	-0.052	-0.102	-0.123	-0.077	-0.089
β_{hascar}		1.781	1.757	4.691	4.720	2.933	2.933
$\beta_{license}$		2.642	2.592	4.588	4.491	-	-
Observations:		57589	59329	53869	55506	76867	77764
Pseudo R-squared:		0.339	0.338	0.411	0.422	0.205	0.208

Note: All parameters are significant at the 0.1% level.

Table 2: Prediction accuracy of models estimated based on reconstructed vs. original locations

	Reconstructed	Original
Switzerland	0.854	0.853
Île-de-France	0.821	0.822
São Paulo	0.699	0.700

parameters are in most cases very similar between models estimated on reconstructed and original activity locations. However, some differences are observable, with the most prominent being for $\beta_{car,invehicle}$ and $\beta_{pt,waiting}$ in Île-de-France.

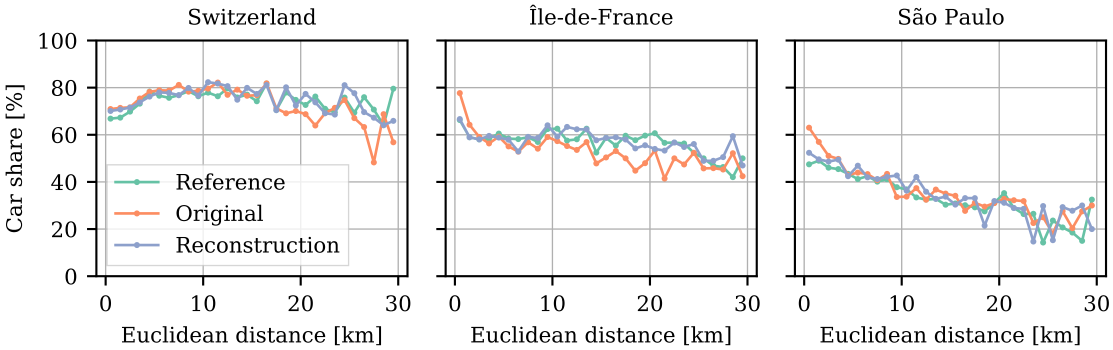
Table 2 shows the prediction accuracy of the models, an evaluation mechanism that is frequently used in machine learning. For this measure, the systematic utilities of the two alternatives are calculated, and the better one is chosen. After, it is evaluated how many choices have been predicted correctly this way. Interestingly, both models perform similarly.

Prediction accuracy assumes that we have perfect knowledge of the individuals and their decision behavior. However, Train (2009) argues that, given the taste variations within the population, it might be more suitable to compare mode-share predictions by sampling from the obtained choice probabilities. The results of this approach can be seen in Table 3. Both models predict mode-shares quite well. However, the models based on the original

Table 3: Forecasted car mode share for models based on reconstructed vs. original location data in comparison to survey reference shares

	Reconstructed	Original	Reference
Switzerland	0.726	0.721	0.716
Île-de-France	0.576	0.570	0.572
São Paulo	0.414	0.409	0.409

Figure 4: Car mode-share in 1km distance bins for reference data from the surveys, and the models based on original and reconstructed locations



coordinates perform slightly better.

Figure 4 shows the car mode share in 1km distance bins for two models and the observed data. Once more, both models show similar patterns and forecasting quality. Towards longer distances, both models start to deviate from the observed mode-share. This could be accredited to the small number of observations for large distances leading higher likelihood of error.

6 Discussion

Based on the three data sets, the results show that the proposed location reconstruction algorithm generates activity locations that match Euclidean trip distances well. Furthermore, models based on reconstructed location provide a prediction quality very similar to the original data. While we only model the binary choice between a car and public transport, the results are promising. Future work should show if the models can still be estimated with a good fit when additional transport modes are added, or more complex models are estimated.

Some of the additional ways that the reconstruction of locations can be improved are:

- For trips made with public transport, origin or destination activity locations with reasonable access to public transport could be sampled within the zones. Consequently, unrealistic locations can be avoided, and higher location precision may be obtained.
- Currently, we only consider Euclidean distances between consecutive activities. Taking into account network distances could potentially improve the accuracy of the algorithm. Even (congested) network travel times could be used to reconstruct activity-to-activity travel times, if available.
- In the current approach, we extract all road nodes from the OSM network. In areas where OSM data has good quality, like in Switzerland or France, one could sample from potential locations based on the origin and destination activity. This way, possible locations for shopping activities would come from the location of shopping facilities present in OSM. More importantly, this could speed up the reconstruction algorithm. On the other hand, it could potentially increase the chances of precisely identifying activity locations of individuals, which would violate the anonymity requirement. If this is the case, suitable measures would need to be taken to further anonymize the data.
- During location reconstruction, we only restrict home activities to happen at the same location. Similarly, we could impose restrictions on education and work activities. However, some individuals perform work activities in different places during the day. If this is the case, we could identify this change in the activity chain by the change of the zone where the work activity is performed.
- Finally, from the location protection perspective, it would be interesting to investigate how knowing the exact location of one of the activities would affect the knowledge about the other activity locations in the chain, which would give insights on the potential vulnerability of the data to outside attacks.

7 Conclusion

This paper demonstrates that discrete choice models estimated from disaggregated zone-based trip data obtained with the proposed reconstruction methodology exhibit similar goodness of fit as those based on non-anonymized data. These results are encouraging as they imply that by using spatial cloaking on the level employed in the three datasets described for Switzerland, Île-de-France, and São Paulo, the usefulness of the data sets for mode-choice modeling can be maintained. The reconstruction algorithm presented in this paper can easily be applied to other data sets (such as California Household Travel Survey Center (2021)), which are spatially anonymized by default.

We observe that anonymity of individuals is not endangered by the methodology we employ. We have highlighted some essential future investigations that can help answer whether additional data could potentially endanger the privacy of the surveyed individuals. As different entities are increasingly collecting data from their users, the possibility to identify individuals from anonymized surveys is increasing, which could have consequences on how future datasets should be anonymized. Therefore, future work should focus on finding the potential weak points of current anonymization techniques, especially when combined with other data sources, to inform on potential risks and vulnerabilities.

8 References

Badu-Marfo, G., B. Farooq and Z. Patterson (2019) Perturbation methods for protection of sensitive location data: Smartphone travel survey case study, *Transportation Research Record*, **2673** (12) 244–255.

Bundesamt für Raumentwicklung (ARE) (2020) Modelletablierung Nationales Personenverkehrsmodell (NPVM) 2017.

Center, T. S. D. (2021) National Renewable Energy Laboratory, `www.nrel.gov/tsdc`. (Accessed: 20.07.2021).

Chicago Metropolitan Agnecy for Planning (2021) My Daily Travel Survey, `https://www.cmap.illinois.gov/data/transportation/travel-survey`. (Accessed: 20.07.2021).

De Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen and V. D. Blondel (2013) Unique in the crowd: The privacy bounds of human mobility, *Scientific reports*, **3** (1) 1–5.

Delling, D., T. Pajor and R. F. Werneck (2015) Round-based public transit routing, *Transportation Science*, **49** (3) 591–604.

Golle, P. and K. Partridge (2009) On the anonymity of home/work location pairs, paper presented at the *Pervasive Computing*, 390–397, Berlin, Heidelberg.

Hörl, S. and M. Balac (2021) Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data, *Transportation Research Part C: Emerging Technologies*, **130**, 103291.

Krumm, J. (2009) A survey of computational location privacy, *Personal and Ubiquitous Computing*, **13** (6) 391–399.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011) Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**, 2825–2830.

Sallard, A., M. Balac and S. Hörl (2021) Synthetic travel demand for the Greater São Paulo Metropolitan Region, based on open data, *Regional Studies, Regional Science*, **In Press**.

Secretaria Estudal dos Transportes Metropolitanos and Companhia do Metropolitano de São Paulo – METRÔ (2019) Pesquisa Origem Destino 2017.

Swiss Federal Office of Statistics (BFS) and Federal Office for Spatial Development (ARE) (2018) Mikrozensus Mobilität und Verkehr. Neuchâtel.

Tchervenkov, C., A. Sallard, G. Kagho, S. Hörl, M. Balac and K. W. Axhausen (2021) Synthetic travel demand for Switzerland, *Working Paper*.

Train, K. E. (2009) *Discrete choice methods with simulation*, Cambridge university press.

Transportes Metropolitanos (2017) Resultados finais da pesquisa origem e destino 2017 (final results of the 2017 origin-destination survey), `http://www.metro.sp.gov.br/pesquisa-od/`.

Île-de-France Mobilités, OMNIL and DRIEA (2010) Enquête Globale Transport 2010.