



Discrete choice modeling with anonymized data

Milos Balac, Sebastian Hörl, Basil Schmid

► To cite this version:

Milos Balac, Sebastian Hörl, Basil Schmid. Discrete choice modeling with anonymized data. Transportation, 2022, 10.1007/s11116-022-10337-1 . hal-03778445

HAL Id: hal-03778445

<https://hal.science/hal-03778445>

Submitted on 15 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Discrete choice modeling with anonymized data

Milos Balac¹ · Sebastian Hörl² · Basil Schmid¹

Accepted: 5 September 2022
© The Author(s) 2022

Abstract

This paper presents an approach to estimate mode-choice models from spatially anonymized revealed preference travel survey data. We propose an algorithm 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 synthetic 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 our methodological approach can reconstruct activity locations that accurately match trip Euclidean distances but with location errors that still provide location protection. The discrete mode-choice models estimated on the synthetic locations perform similarly, in terms of goodness of fit and prediction, to the ones obtained from the observed activity locations.

Keywords Anonymization · Data privacy · Travel survey · Discrete choice model

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

✉ Milos Balac
milos.balac@ivt.baug.ethz.ch
Sebastian Hörl
sebastian.horl@irt-systemx.fr
Basil Schmid
basil.schmid@ivt.baug.ethz.ch

¹ IVT, ETHZ, Stefano-Francini-Platz 5, Zurich 8093, Switzerland

² IRT SystemX, 2 Bd Thomas Gobert, Palaiseau 91120, France

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 subsequent 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. "[Background](#)" Section discusses the current literature in data anonymization, its application to the field of transportation and issue associated with data aggregation. "[Methodology](#)" Section proposes a heuristic to reconstruct activity locations based on zone-based trip data and explains the subsequently used mode choice modeling approach. "[Case study](#)" Section explains the data sets used, and "[Results](#)" Section presents the results. Finally, "[Discussion](#) and [Conclusion](#)" Sections provide discussion and closing remarks.

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 common pieces of information that needs to be anonymized is the 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 (California Department of Transportation 2021), or My Daily Travel Survey conducted in the Chicago Metropolitan Region (Chicago Metropolitan Agency 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 contrast, publicly available Brazilian census data is aggregated to a census zone containing between 20 and 55 thousand people.

The problems of spatial data aggregation presented above are also known as the Modifiable Areal Unit Problem (MAUP), which was first formally explained by Openshaw (1981). MAUP arises when data is aggregated and presented on various scales and zoning systems, leading to different interpretations and model estimations. In the case of choice modeling, similar issues arise. Guo and Bhat (2004) show, for the case of residential choice modeling, how MAUP can lead to worse model fit. What is notable in the study by Gao and Bhat, and most of the studies highlighting the impact of MAUP on model estimates is that they mostly rely on aggregated, zonal variables (i.e., average accessibility, safety, income). The modeling problem tackled in this paper uses disaggregated, individual data as choice variables, which avoids MAUP. On the other hand, the potential impact of zoning system employed on mode-choice model estimates remains unclear.

Even when privacy protection techniques, like aggregation 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 paper, even though we provide some insights. However, we 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 the 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.

Methodology

Problem statement

Figure 1 shows a motivating example for our approach. It shows an activity chain with four activities, where a person starts his/her daily travels at home in the 13th arrondissement in Paris, then goes to work close to the Eiffel tower 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)

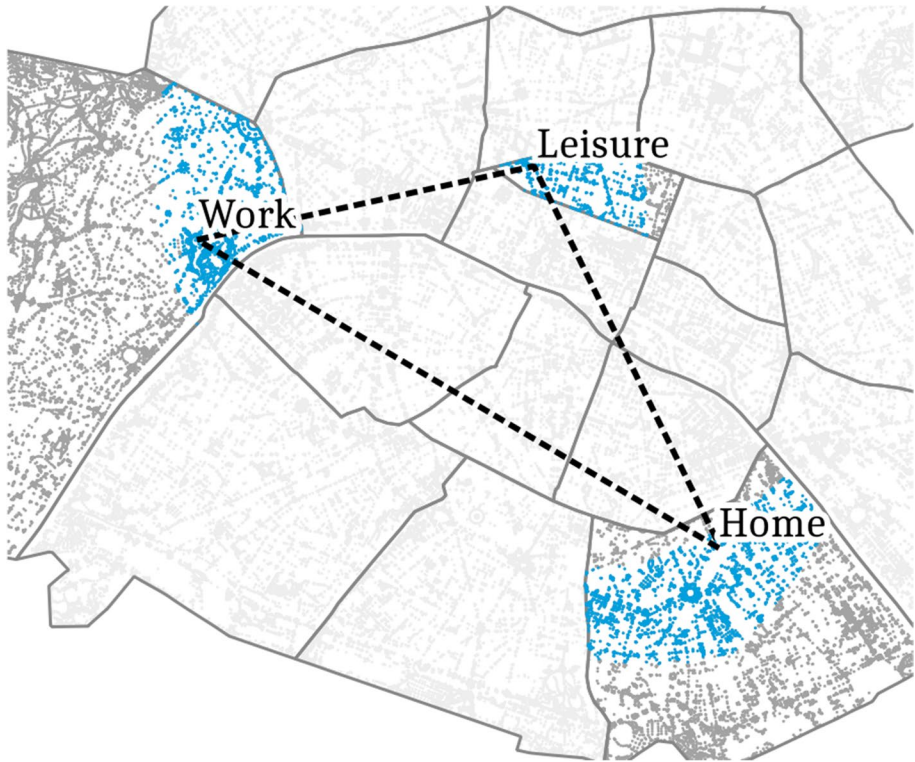


Fig. 1 Example of a feasible set of candidate points

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 is small and contains original coordinates, recovering the exact locations would substantially increase.

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 $i \in \{1, \dots, N\}$ is a “home” activity. The indices of those activities are noted in the index set $\mathcal{H}$. Furthermore, reference Euclidean distances are given as $r_i \in \mathbb{R}$.

The potential locations for the i th 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 $\mathcal{L} = \mathcal{L}_1 \cap \dots \cap \mathcal{L}_N$. Let $k \in \{1, \text{card}(\mathcal{L})\}$ 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 $l = (l_1, \dots, l_N)$ 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

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

with the following constraints:

$$\begin{aligned} y_{l_i, i} &= 1 & \forall i \in \{1, \dots, N\} \\ l_i &= l_{\min \mathcal{H}} & \forall i \in \mathcal{H} \end{aligned} \quad (2)$$

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.

Solution strategy

The solution strategy aims to find a feasible and optimal sequence $(l_1, \dots, l_N)$ 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:

Algorithm 1 Chain-based location assignment

Input:

Location sets $\mathcal{L}_1, \dots, \mathcal{L}_N$ and $\mathcal{L}$

Home activity index set $\mathcal{H}$

Initialize:

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

For each $l_1 \in \mathcal{L}_1$

$C \leftarrow ((l_1), 0)$

Continue

While $C > 0$

$(l_1, \dots, l_n), q_n \leftarrow \text{pop } C$

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

$q_{n+1} = \max\{q_n, |d(l_n, l_{n+1}) - r_i|\}$

$C \leftarrow ((l_1, \dots, l_n, l_{n+1}), q_{n+1})$

End

End

Continue

Return l^*

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

Choice model

To test the impacts of location error on mode choice model estimates, we make use of a multinomial Logit (MNL) model (e.g. McFadden 1986; Train 2009). We model the choice among the three mode alternatives car (C), public transport (PT) or walk (W). To obtain the relevant characteristics of the three alternatives, we perform a minimum generalized cost path routing for car trips, based on road networks obtained from OpenStreetMap (OSM) data and free flow speeds. For public transport, we use an implementation of the RAPTOR algorithm (Delling et al. 2015) to find routes in the public

transport network provided in GTFS format which minimize the total travel time of the trips. The data sets are documented in 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). Since a public transport route cannot be found for some trips (i.e., the trip is too short, or public transport is not accessible), those trips are filtered out, which creates some minor differences in the size of the data set for the reconstructed and original coordinates (see also Table 1 at the bottom). The travel time for the walk alternative is calculated based on the routed distance and an average speed of 5 km/h. Since we are using revealed preference data, note that including travel costs in the model is problematic, as they are highly correlated with distance (Train 2009), providing only limited trade-off information to obtain a robust estimate for the value of travel time (e.g. Schmid et al. 2019). However, this does not affect our general research design, investigating if the reference and reconstructed data sets lead to similar behavioral parameters and predictions.

The utility function of alternative $i \in \{C, PT, W\}$ and individual $n \in \{1, 2, \dots, N\}$ in each choice situation $t \in \{1, 2, \dots, T_n\}$ is given by

$$U_{i,n,t} = \alpha_i + \mathbf{x}_{i,n,t} \beta_i \left(\frac{dist_{n,t}}{dist} \right)^{\lambda_{k,i}} + \epsilon_{i,n,t} \quad (3)$$

where C (car) is the reference alternative for identification purposes (i.e. $\alpha_C = 0$). The utility function $U_{i,n,t}$ includes the following components:

- α_i : Parameter of alternative-specific constant (ASC).
- $\mathbf{x}_{i,n,t}$: Vector of level-of-service (LOS) attributes. Car (C): In-vehicle travel time [h]; public transport (PT): In-vehicle travel time [h], access and egress time [h], number of transfers [#] and transfer waiting time [h]; walk (W): Travel time [h]
- β_i : Alternative-specific parameter vector of LOS attributes.
- $dist_{n,t}$: Crow-fly distance [km]; $dist$ = sample mean [km].
- $\lambda_{k,i}$: Parameter capturing non-linear changes in LOS sensitivity of attribute k according to trip distance (included for car in-vehicle travel time, public transport in-vehicle travel time, sum of access and egress time, and transfer waiting time. $\lambda_{k,i} > 0$: Increasing LOS sensitivity; $\lambda_{k,i} < 0$: Decreasing LOS sensitivity (e.g. Schmid et al. 2021).
- $\epsilon_{i,n,t}$: Remaining IID extreme value type I error term.

The availability of alternative j varies depending on sociodemographic information and network characteristics and is coded as a dummy variable $a_{j,n,t}$ that enters the Logit choice probability, as shown in Eq. (4):

- C Available if a respondent has a driving license and/or stated that he/she has access to a car.
- PT Available if a PT route was identified by the routing algorithm.
- W Available if trip distance < 5 km.

The probability that alternative i among the set of available alternatives $j \in \{C, PT, W\}$ for trip t by individual n is chosen is given by

$$P(i_{n,t}|\Omega) = \frac{a_{i,n,t} \exp(V_{i,n,t})}{\sum_j a_{j,n,t} \exp(V_{j,n,t})} \quad (4)$$

where Ω is the set of all parameters to be estimated and $V_{i,n,t}$ is the deterministic part of utility.

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 (2010), and Greater São Paulo Metropolitan Region (Secretaria Estadual dos Transportes Metropolitanos, Companhia do Metropolitano de São Paulo-METRÔ 2019) to create the inputs for the reconstruction algorithm and the downstream mode choice model estimation.

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 ($\approx 0.6\%$ 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 2020).

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. EGT is only available on request from the regional authorities and therefore not publicly available. Activity locations are reported on a grid of 100×100 meters. As zoning data, French municipalities are used.

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.

Candidates

For the three cases, multiple sets of candidate points are created, among which the locations of the activities can be chosen.

A straightforward approach to obtain candidate locations would be to use centroid points of all zones as potential activity locations. However, in this case all within zone trips would have a same starting and ending locations, which in our experiments show a strong negative effect on the reconstructed location quality and subsequently on model estimates. Therefore, we focus on the following two ways to generate candidate points.

First, we sample points at random for each zone in the three use cases. To do so, we obtain the axis-aligned bounding box around each zone's complex geometry, sample N points within the bounding box, and then keep those points that fall inside the zone shape. 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 (OSM) 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 (this includes all nodes along the road geometry, and not just start and end points) as location candidates.

Estimation and model evaluation

For each of the three case studies denoted by s , we estimate two choice models, one based on the original coordinates M_s^o and one based on the reconstructed (OSM) coordinates M_s^r . It is important to note that in contrast to other studies that used imputation of missing/noisy information (e.g. Steimetz and Brownstone 2005), the reconstructed (imputed) coordinates are deterministic in the sense that the algorithm minimizes the maximum deviation of the Euclidean distance of the trips, such that a multiple imputation approach is not applicable. To compare the predictive power of these three models, we split each data set into a training set containing 70% (T_s^o and T_s^r) and a test set containing 30% (V_s^o and V_s^r) of the data (i.e., T_s^o and T_s^r in the majority of cases contain the same trips, but with different routing data). We re-estimate M_s^o and M_s^r on the respective training sets T_s^o and T_s^r . Finally, we analyze the prediction accuracy (PA) of the trained models on V_s^o data (original 30% test set), investigating to what extent a model based on original coordinates outperforms the one based on anonymized coordinates.

The models are estimated using the `mixl` package in *R* 4.0.4 (Molloy et al. 2021). Cluster-robust (by individual) standard errors are obtained by using the Eicker–Huber–White sandwich estimator.

Results

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.

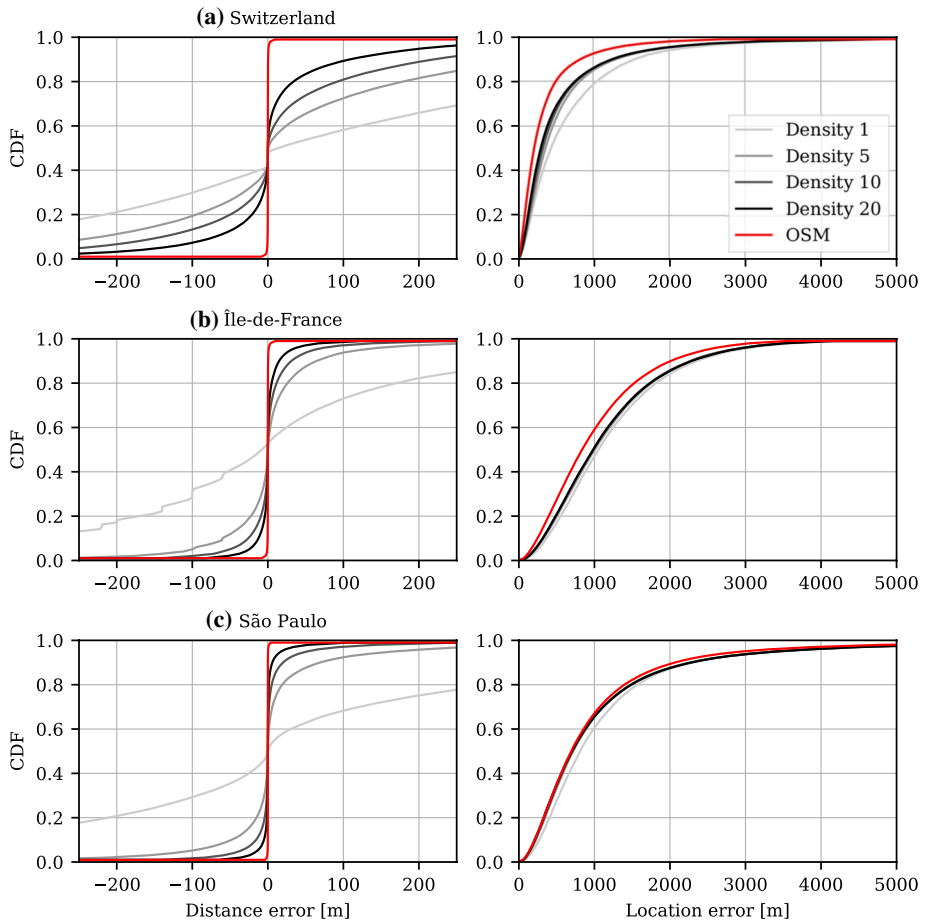


Fig. 2 Distance and location errors after the matching process

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-based 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-based data, locations can not be reconstructed

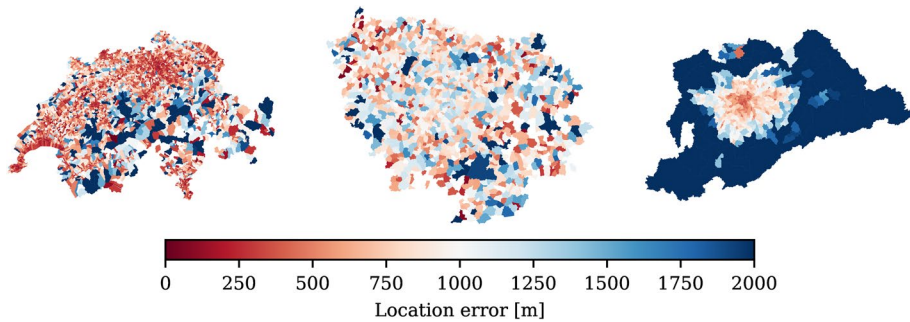


Fig. 3 Spatial distribution of the location error for OSM-based reconstruction (from left to right: Switzerland, Île-de-France, São Paulo)

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, the accuracy is very low in the outer regions, where enormous zones contain large, unpopulated areas. However, the accuracy increases towards the city center of São Paulo.

Model estimation

In what follows are the model estimates based on the reference data and OSM-based reconstructed data, as it has been shown to be superior to random sampling approach. For the sake of comparison, we also present the results using the centroids of zones as activity points.

Table 1 presents the models estimated for the different study areas (i.e., Switzerland, Île-de-France and São Paulo) and complete data sets (i.e., reference, OSM-based and centroid-based data sets). In the reference data set, all parameters have the expected sign and are, in most cases, significant at the 1% level. The parameters are in most cases very similar between models estimated on original and OSM activity locations. Only one substantial and significant difference is observable (i.e., 95% confidence bands not overlapping), which is the case for $\hat{\rho}_{travel\ time, C}$ in Île-de-France. This shows that the OSM-based reconstructed data set would lead to very similar behavioral indicators (e.g., such as elasticities and marginal rates of substitution) and therefore would be an appropriate alternative to original coordinates. After all, the models all perform very similarly in terms of goodness-of-fit, as indicated by the ρ^2 .

Table 1 Models estimated for the three study areas for the reference, OSM-based and centroid-based data sets (100% of observations)

Reference category: C	Switzerland			Île-de-France			São Paulo		
	Ref.	OSM	Cent.	Ref.	OSM	Cent.	Ref.	OSM	Cent.
ASC W ($\hat{\alpha}_W$)	0.67*** (0.11)	0.74*** (0.11)	0.84*** (0.10)	2.06*** (0.06)	2.11*** (0.06)	1.41*** (0.09)	0.05 (0.04)	-0.00 (0.03)	0.24*** (0.07)
ASC PT ($\hat{\alpha}_{PT}$)	-0.95*** (0.04)	-0.86*** (0.04)	-0.89*** (0.06)	-0.00 (0.05)	-0.17*** (0.06)	0.19*** (0.06)	-0.25*** (0.02)	-0.26*** (0.03)	-0.69*** (0.03)
$\hat{\beta}_{travel\ time,W}$	-5.94*** (0.42)	-6.07*** (0.42)	-6.28*** (0.39)	-8.65*** (0.22)	-8.58*** (0.23)	-6.47*** (0.28)	-0.70*** (0.09)	-0.62*** (0.08)	-1.33*** (0.21)
$\hat{\beta}_{travel\ time,PT}$	-0.28*** (0.05)	-0.31*** (0.05)	-0.17*** (0.04)	-1.26*** (0.12)	-0.86*** (0.11)	-1.04*** (0.11)	-0.31*** (0.05)	-0.22*** (0.05)	-0.42*** (0.04)
$\hat{\beta}_{travel\ time,C}$	-1.35*** (0.15)	-1.46*** (0.15)	-0.99*** (0.14)	-3.17*** (0.16)	-2.22*** (0.13)	-2.50*** (0.14)	-2.21*** (0.16)	-1.81*** (0.15)	-3.85*** (0.23)
$\hat{\beta}_{access\ time,PT}$	-2.18*** (0.14)	-2.43*** (0.12)	-1.64*** (0.13)	-1.43*** (0.12)	-0.92*** (0.11)	-1.68*** (0.11)	-0.05*** (0.01)	-0.05*** (0.01)	-0.25*** (0.04)
$\hat{\beta}_{trans.\ wait.\ time,PT}$	-3.97*** (0.21)	-3.30*** (0.21)	-2.26*** (0.18)	-4.55*** (0.32)	-3.25*** (0.27)	-2.90*** (0.25)	-0.98*** (0.16)	-0.77*** (0.13)	-0.45*** (0.11)
$\hat{\beta}_{transfers,PT}$	-0.04 (0.03)	-0.04* (0.03)	-0.10*** (0.03)	-0.04 (0.03)	0.01 (0.03)	0.02 (0.03)	-0.13*** (0.02)	-0.10*** (0.02)	-0.04*** (0.02)
$\hat{\lambda}_{travel\ time,PT}$	0.95*** (0.06)	0.94*** (0.05)	1.14*** (0.07)	0.55*** (0.07)	0.68*** (0.10)	0.67*** (0.10)	1.13*** (0.10)	1.28*** (0.12)	1.10*** (0.08)
$\hat{\lambda}_{travel\ time,C}$	0.59*** (0.05)	0.56*** (0.05)	0.70*** (0.06)	0.44*** (0.04)	0.57*** (0.05)	0.57*** (0.05)	0.74*** (0.05)	0.79*** (0.05)	0.59*** (0.07)
$\hat{\lambda}_{access\ time,PT}$	-0.50*** (0.03)	-0.41*** (0.03)	-0.39*** (0.04)	-0.92*** (0.04)	-1.02*** (0.05)	-0.57*** (0.03)	-2.31*** (0.07)	-2.26*** (0.13)	1.14*** (0.14)

Table 1 (continued)

Reference category: C	Switzerland			Île-de-France			São Paulo		
	Ref.	OSM	Cent.	Ref.	OSM	Cent.	Ref.	OSM	Cent.
	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)	Coef./ (SE)
$\hat{\lambda}_{trans, wait, time, PT}$	-0.42*** (0.04)	-0.42*** (0.05)	-0.32*** (0.06)	-0.64*** (0.06)	-0.61*** (0.07)	-0.57*** (0.06)	-0.17 (0.21)	0.06 (0.20)	0.64*** (0.21)
Number of parameters	12	12	12	12	12	12	12	12	12
Number of respondents	24205	23914	24094	24086	23755	19106	44714	44003	38201
Number of observations	72518	70822	72119	72347	70400	51524	100402	98602	87014
$\mathcal{LL}_{null}$	-54333.12	-52466.96	-53294.64	-55109.37	-52969.54	-35460.81	-67713.01	-66069.15	-57327.06
$\mathcal{LL}_{model}$	-35920.01	-34803.51	-35379.52	-43101.07	-42090.56	-29482.16	-63264.71	-62031.10	-54894.16
ρ^2	0.34	0.34	0.34	0.22	0.21	0.17	0.07	0.06	0.04

Robust standard errors: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2 Marginal rates of substitution (MRS) with the coefficient of car travel time ($\hat{\beta}_{travel\ time,C}$) in the denominator based on the models shown in Table 1

	Reference	OSM	Centroid
<i>Switzerland</i>			
$\widehat{MRS}_{travel\ time,W}$	4.39	4.15	6.32
$\widehat{MRS}_{travel\ time,PT}$	0.21	0.21	0.17
$\widehat{MRS}_{access\ time,PT}$	1.61	1.66	1.65
$\widehat{MRS}_{trans.\ wait.\ time,PT}$	2.93	2.26	2.28
$\widehat{MRS}_{transfers,PT}$	0.03	0.03	0.10
<i>Île-de-France</i>			
$\widehat{MRS}_{travel\ time,W}$	2.72	3.86*	2.58
$\widehat{MRS}_{travel\ time,PT}$	0.40	0.39	0.41
$\widehat{MRS}_{access\ time,PT}$	0.45	0.41	0.67*
$\widehat{MRS}_{trans.\ wait.\ time,PT}$	1.43	1.46	1.16
$\widehat{MRS}_{transfers,PT}$	0.01	-0.01	-0.01
<i>São Paulo</i>			
$\widehat{MRS}_{travel\ time,W}$	0.32	0.34	0.35
$\widehat{MRS}_{travel\ time,PT}$	0.14	0.12	0.11
$\widehat{MRS}_{access\ time,PT}$	0.02	0.03	0.07*
$\widehat{MRS}_{trans.\ wait.\ time,PT}$	0.44	0.43	0.12*
$\widehat{MRS}_{transfers,PT}$	0.06	0.06	0.01*

*95% confidence band not overlapping with reference

Results look considerably different for the centroid-based data set. While in case of Switzerland, the ρ^2 is close to the reference case, it drops considerably for São Paulo (-3%-points) and, most pronounced, for Île-de-France (-5%-points). Also, in all three cases, many parameter estimates are considerably different than in the reference case.

Findings are confirmed when comparing the marginal rates of substitution [MRS; see e.g. Train (2009)], i.e. the ratio of coefficients of level-of-service attributes using $\hat{\beta}_{travel\ time,C}$ as the reference¹, as shown in Table 2. In the case of Switzerland, while the 95% confidence bands of the OSM and centroid always overlap with the reference case, the MRS are substantially different for the centroid data set (most pronounced for PT transfers by more than factor 3, followed by walk travel time). E.g., while one PT transfer is valued 0.03 hours of car travel time in both the reference and OSM case, it is valued 0.1 hours in the centroid case. In the case of Île-de-France, the MRS for walk travel time is significantly different for the OSM case, while the MRS for PT access time is significantly different in the centroid case. Finally, in the case of São Paulo, the reference and OSM case exhibit very similar MRS, while in the centroid case, three of them are significantly different. To summarize, while in only one case, the OSM-based data set performs significantly worse than the reference data set, the MRS obtained based on the centroid data set often differ substantially.

¹ Since our models do not include travel costs, we (arbitrarily) chose the coefficient of car travel time to be in the denominator. Confidence intervals are calculated using the Delta method (Daly et al. 2012).

Table 3 Prediction accuracy (PA; in %) and 95% confidence intervals (in brackets) of estimated models (70% training data sets of original coordinates, reconstructed coordinates using OSM and centroid-based coordinates) for 30% test data set (original coordinates)

	Reference	OSM	Centroid
Switzerland	70.01 (69.27; 70.75)	69.44 (68.69; 70.12)	69.15 (68.40; 69.96)
Île-de-France	61.17 (60.49; 61.84)	60.29 (59.60; 60.98)	57.26* (56.52; 58.15)
São Paulo	58.88 (58.20; 59.57)	58.75 (58.13; 59.37)	58.23 (57.69; 58.77)

*95% confidence band not overlapping with reference

Table 3 shows the prediction accuracy (PA) of the models re-estimated for the 70% training data sets and validated on the 30% test data sets, an evaluation method that is frequently used in machine learning (e.g., Yadav and Shukla 2016; Schmid et al. 2022). For each model and test data set, the PA is obtained by simulating how many choices are, on average, predicted correctly. Specifically, we draw $R = 500$ times from the multivariate normal distribution $\mathcal{N}(\hat{\mathbf{\Omega}}, \hat{\mathbf{\Sigma}})$, where $\hat{\mathbf{\Omega}}$ is the vector of estimated parameters and $\hat{\mathbf{\Sigma}}$ is the robust variance-covariance matrix of a model, to predict the alternative-specific probabilities in the test data set (e.g., Bierlaire 2017). In each repetition, we use a probabilistic calculation of the PA by sampling the predicted choices according to the probabilities of each alternative. As discussed in Train (2009), this measure is more appropriate than the percent of correctly predicted choices according to the highest probabilities (first preference recovery; e.g., Ortúzar and Willumsen 2011; Parady et al. 2021), since it better reproduces the market shares and reflects the probabilistic nature (uncertainty) of the Logit model (see also Palma et al. 2016). If the choices were to be repeated many times, or observed by many individuals with the same attributes, each alternative would be chosen by a certain fraction. Finally, the 95% confidence interval is approximated by calculating the 2.5% and 97.5% quantiles of the resulting distribution as a lower and upper bound, respectively (Bierlaire 2017).

As expected (given by the very similar ρ^2 values and parameter estimates), results indicate that there is only a minor difference between the reference and OSM-based data set, with the latter always performing worse. However, the PAs are not significantly different, as indicated by the always overlapping 95% confidence intervals. When investigating the alternative-specific PA differences (not reported), only the Île-de-France model with original coordinates exhibits a significantly higher PA (1.09%-points; $p < 0.05$) for the car alternative, which goes in line with the substantial difference of the coefficient for car travel time, $\hat{\beta}_{travel\ time, C}$. Finally, as expected, the centroid-based data set again performs worse than the OSM-based data set, and in case of Île-de-France the drop in the PA of more than 3%-points is significant.

As shown in Table 4, the models based on the reference and OSM coordinates are able to predict the observed mode shares in the test data set sufficiently. However, while the models based on the original coordinates (reference) significantly ($p < 0.05$; using the same method as for the PA) under/overestimate the mode share of PT (Switzerland) and Walk (São Paulo and Switzerland, respectively), the models based on the OSM data set under/overestimate the mode shares of Car (Switzerland), PT (Île-de-France) and Walk (São Paulo and Switzerland, respectively), therefore performing only slightly worse. Nevertheless, the absolute differences of mode shares are not substantial.

Table 4 Observed and predicted mode shares (in %; 95% confidence intervals in brackets) of estimated models (70% training data sets of original/reference, OSM- and centroid-based coordinates) for 30% test data set (original coordinates)

Car (C)	Observed	Reference	OSM	Centroid
Switzerland	61.65	61.92 (61.41; 62.37)	60.90* (60.34; 61.40)	60.13* (59.58; 60.70)
Île-de-France	47.49	47.80 (47.10; 48.42)	46.37 (45.58; 47.02)	42.94* (41.99; 43.79)
São Paulo	33.37	33.16 (31.93; 34.38)	34.14 (31.83; 36.45)	34.33* (33.59; 35.03)
Public transport (PT)	Observed	Reference	OSM	Centroid
Switzerland	23.03	21.93* (21.51; 22.32)	23.16 (22.72; 23.67)	24.48* (23.92; 25.02)
Île-de-France	31.00	30.81 (30.30; 31.41)	32.28* (31.75; 32.99)	38.65* (37.74; 39.72)
São Paulo	46.88	48.33 (46.77; 49.88)	47.68 (45.20; 50.17)	48.16* (47.48; 49.03)
Walk (W)	Observed	Reference	OSM	Centroid
Switzerland	15.32	16.15* (15.82; 16.52)	15.94* (15.62; 16.29)	15.39 (15.03; 15.73)
Île-de-France	21.51	21.39 (21.03; 21.74)	21.35 (20.97; 21.69)	18.41* (17.77; 19.04)
São Paulo	19.75	18.52* (17.67; 19.36)	18.18* (17.37; 18.98)	17.51* (17.09; 17.81)

*95% confidence band not overlapping with observed mode share

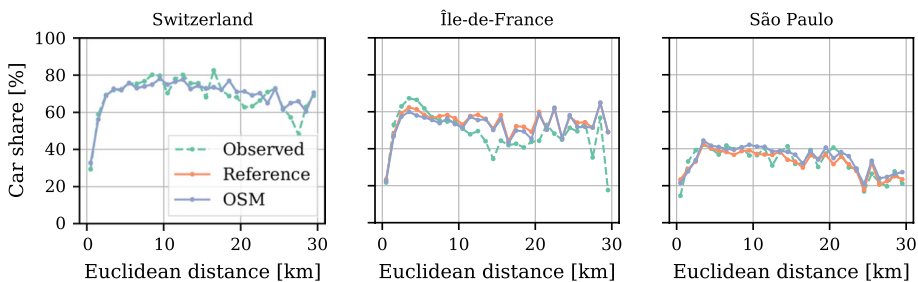


Fig. 4 Car mode-share in 1km distance bins for observed data from the surveys, and the models based on reference and reconstructed locations

After all, we also do not find any substantial differences in the confidence intervals between the two approaches (original and reconstructed coordinates). Finally, the models based on the centroid coordinates perform notably worse (in eight of nine cases, the confidence bands do not overlap with the observed mode share), and in the case of

Île-de-France the deviations are very substantial (e.g., overestimation of 7.65%-points of PT mode share).

Figure 4 shows the car mode share in 1km distance bins for two models and the observed data. Once more, all models show similar patterns and forecasting quality. For Switzerland, reference and OSM models are almost identical.

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 to a higher likelihood of error.

Discussion

Based on the three data sets, the results show that the models estimated from reconstructed activity locations perform similarly as those estimated on the observed activity locations. The proposed reconstruction algorithm together with the OSM data can provide a useful alternative to observed non-anonymized data set. This finding is important as it suggests that anonymized data can be used successfully for mode-choice modelling. However, we demonstrate that just using the centroids of the zones as activity points is not sufficient, as indicated by the substantial differences in marginal rates of substitution, prediction accuracy and mode shares.

The following describes the limitations of our work and possible future directions:

- 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 (to some extent) this change in the activity chain by the change of the zone where the work activity is performed.
- In certain cases the activity location reconstruction problem could have a large number of possible solutions (see also 1). This can potentially also be avoided by better representation of candidate locations, thus limiting the choice set, as explained above.
- The modeling approach used in the study is purposely simple in terms of choice model and variables, as we aim to show the first tradeoffs between models estimated

on anonymized and raw datasets. It must be kept in mind that other, more sophisticated models with different specifications might exhibit different tradeoffs.

Improving the activity location reconstruction could further enhance the model estimates, especially in areas where certain zones are large, like in São Paulo. On the other hand, from the location protection perspective, an improved reconstruction approach could endanger the privacy of the survey respondents. If this is the case, it would signal to the maintainers that a more sophisticated anonymization procedure is needed. Finally, from the aspect of cyber security, 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 into the potential vulnerability of the data to outside attacks.

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 preserved. We also show that using centroids of zones as activity locations is insufficient and leads to considerably different parameters estimates. Finally, the reconstruction algorithm presented in this paper can easily be applied to other data sets (such as California Household Travel Survey (California Department of Transportation 2021)), which are spatially anonymized by default.

We observe that our methodological approach does not endanger the anonymity of individuals. However, we have highlighted some essential future investigations that can help answer whether additional data could threaten the surveyed individuals' privacy. As different entities are increasingly collecting data from their users, the possibility of identifying individuals from anonymized surveys is increasing, which could affect 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.

Authors' contribution MB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing. SH: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing. BS: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute. The survey Enquête Globale Transport 2010 used in this research was co-piloted and co-financed by Île-de-France Mobilités and DRIEA. It was executed by Île-de-France Mobilités.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Badu-Marfo, G., Farooq, B., Patterson, Z.: Perturbation methods for protection of sensitive location data: smartphone travel survey case study. *Transp. Res. Rec.* **2673**(12), 244–255 (2019)
- Bierlaire, M.: Calculating indicators with PythonBiogeme. Tech. rep (2017)
- Bundesamt für Raumentwicklung: Modelletablierung Nationales Personenverkehrsmodell (NPVM) 2017. Tech. rep (2020)
- California Department of Transportation: California Household Travel Survey. www.nrel.gov/tsdc, (Accessed: 20.07.2021) (2021)
- Chicago Metropolitan Agency for Planning: My daily travel survey. <https://www.cmap.illinois.gov/data/transportation/travel-survey>, (Accessed: 20.07.2021) (2021)
- Daly, A., Hess, S., de Jong, G.: Calculating errors for measures derived from choice modelling estimates. *Transp. Res. Part B: Methodol.* **46**(2), 333–341 (2012)
- De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., et al.: Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* **3**(1), 1–5 (2013)
- Delling, D., Pajor, T., Werneck, R.F.: Round-based public transit routing. *Transp. Sci.* **49**(3), 591–604 (2015)
- Golle, P., Partridge, K.: On the anonymity of home/work location pairs. In: Tokuda, H., Beigl, M., Friday, A., et al. (eds.) *Pervasive Computing*, pp. 390–397. Springer, Berlin Heidelberg, Berlin, Heidelberg (2009)
- Guo, J.Y.: Bhat CR (2004) Modifiable areal units: Problem or perception in modeling of residential location choice? *Transp. Res. Rec.* **1**, 138–147 (1898)
- Hörl, S., Balac, M.: Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transp. Res. Part C: Emerg. Technol.* **130**, 103291 (2021)
- Île-de-France Mobilités, OMNIL, DRIEA (2010) *Enquête Globale Transport* (2010)
- Krumm, J.: A survey of computational location privacy. *Pers. Ubiquit. Comput.* **13**(6), 391–399 (2009)
- McFadden, D.: The choice theory approach to market research. *Mark. Sci.* **5**(4), 275–297 (1986)
- Molloy, J., Becker, F., Schmid, B., et al.: mixl: an open-source R package for estimating complex choice models on large datasets. *J. Choice Model.* **39**(100), 284 (2021)
- Openshaw, S.: The modifiable areal unit problem, pp. 60–69. *A British view*, Quantitative geography (1981)
- Ortúzar, Jd.D., Willumsen, L.G.: *Modelling Transport*. John Wiley and Sons, West Sussex (2011)
- Palma, D., de Dios, Ortúzar J., Rizzi, L.I., et al.: Modelling choice when price is a cue for quality: a case study with Chinese consumers. *J. Choice Model.* **19**, 24–39 (2016)
- Parady, G., Ory, D., Walker, J.: The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: a review of validation practices in the transportation academic literature. *J. Choice Model.* **38**(100), 257 (2021)
- Sallard, A., Balac, M., Hörl, S.: An open data-driven approach for travel demand synthesis: an application to São Paulo. *Reg. Stud. Reg. Sci.* **8**(1), 371–386 (2021)
- Schmid, B., Jokubauskaite, S., Aschauer, F., et al.: A pooled RP/SP mode, route and destination choice model to investigate mode and user-type effects in the value of travel time savings. *Transp. Res. Part A: Policy Pract.* **124**, 262–294 (2019)
- Schmid, B., Molloy, J., Peer, S., et al.: The value of travel time savings and the value of leisure in Zurich: estimation, decomposition and policy implications. *Transp. Res. Part A: Policy Pract.* **150**, 186–215 (2021)
- Schmid, B., Becker, F., Molloy, J., et al.: Modeling train route decisions during track works. *J. Rail Transp. Plann. Manag.* **22**(100), 320 (2022)
- Secretaria Estadual dos Transportes Metropolitanos, Companhia do Metropolitano de São Paulo - METRÔ: *Pesquisa Origem Destino 2017*. Tech. rep (2019)

- Steinmetz, S.S., Brownstone, D.: Estimating commuters' value of time with noisy data: a multiple imputation approach. *Transp. Res. Part B: Methodol.* **39**(10), 865–889 (2005)
- Swiss Federal Office of Statistics (BFS), Federal Office for Spatial Development (ARE): *Mikrozensus Mobilität und Verkehr*. Neuchâtel (2018)
- Tchervenkov, C., Sallard, A., Kagho, G. et al.: Synthetic travel demand for Switzerland. Working Paper (2021)
- Train, K.E.: *Discrete choice methods with simulation*. Cambridge University Press, New York, USA (2009)
- Transportes Metropolitanos: Resultados finais da pesquisa origem e destino 2017 (final results of the 2017 origin-destination survey). <http://www.metro.sp.gov.br/pesquisa-od/> (2017)
- Yadav, S., Shukla, S.: Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: 2016 IEEE 6th International conference on advanced computing (IACC), IEEE, pp 78–83 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Milos Balac holds a Ph.D. from ETH and is currently a Senior Research Assistant and lecturer at the institute for transport planning and systems at ETH Zurich. His research is focused on investigating complex mobility systems and their impacts on individual mobility patterns using agent-based methodology.

Sebastian Hörl is a senior researcher at IRT SystemX where he is involved in various projects on agent-based transport simulation for passenger transport and logistics. He received his Ph.D. in Transport Planning at ETH Zurich. His main interests revolve around the topics of replicable use of open data and software in transport planning and applied large-scale transport simulation.

Basil Schmid studied economics at the University of Zurich with a strong focus discrete choice modeling, and finished his Ph.D. in transportation science in 2019. He currently works as a postdoctoral researcher and lecturer at the institute for transport planning and systems at ETH Zurich, where his main research interest lies in the estimation of value of (travel) time components.