Discrete Choice Modeling for Modal Disaggregation of Mobile OD Flows: Comparison with Smart Card Data

Paul de Nailly*[1], Etienne Côme[1], Rico Krueger[2], and Latifa Oukhellou[1]

[1]COSYS-GRETTIA, Gustave Eiffel University, France
[2]Department of Technology, Management and Economics, Technical University of Denmark, Denmark

Short summary

Estimating travellers' preferences for different transport modes can be achieved using Discrete Choice Models (DCMs), which often rely on outdated travel surveys (HTSs) that misrepresent current urban mobility. Digital data can address this problem by complementing the survey, but challenges arise when HTS and digital sources are from distant periods and due to the incompleteness and bias inherent in digital data. In this paper, we propose a first step in a broader effort to update mode choice probabilities using recent digital traces. We train a DCM on the HTS to estimate mode choice probabilities, which are then used to disaggregate mobile phone OD matrices by mode. We investigate the extent to which preprocessed mobile phone data disaggregated with a DCM model is comparable to smart card data used to estimate OD matrices for the public transport mode. Experiments are conducted using real data collected in the Métropole de Lyon, France.

Keywords: discrete choice model, travel survey, mobile phone data, smart card data

1 Introduction

Studying mobility patterns in urban areas offers valuable insights into the accessibility and popularity of different modes of transportation, helping determine whether new infrastructures are needed. Travel habits may significantly change over time and as a result of impacting events (e.g., the COVID-19 pandemic). Discrete choice models (DCMs) can compute the utility for travelers to use each mode of transportation, taking into account various alternative-specific (e.g., travel cost, time) and socio-economic (e.g., median income) attributes.

These models are based on Household Travel Surveys (HTS) to provide detailed insights into the mobility of a representative sample of people (Heinen & Chatterjee, 2015). However, due to their high cost, HTS are infrequent and cannot provide a dynamic view of modal shares. In contrast, digital data sources such as smart card and mobile phone data allow for continuous, fine-grained analysis across geographic and temporal scales. However, these data in isolation are partial and their ability to capture complex and interrelated phenomena such as modal shares is still limited (Ounoughi & Yahia (2023)). Several studies have explored integrating digital data alongside HTS in estimating DCMs. Zhang et al. (2019) combine HTS with smart card and car GPS data in a joint DCM model, allowing the inclusion of variables available in either or both data sources in the utility functions (e.g., travelers age and income available only in HTS, travel time available in both). Krueger et al. (2023) create an enriched database for training a DCM based on a HTS and including ride-sourcing and taxi trips. The authors handled the sampling biases due to the inclusion of specific new modes in the database by adding a sampling factor in the DCM. Following Zhang et al. (2019), the authors in Andersson et al. (2024) develop a DCM that combines data from HTS and mobile phone data. They highlight several limitations of using mobile phone data for travel demand modeling, which can be mitigated by incorporating detailed information from HTS. These studies generally utilize both HTS and digital data sources to generate attributes that enrich the training datasets for DCMs. However, this approach may introduce bias into the model when the data comes from significantly different periods. Bwambale et al. (2021) combined HTS with mobile phone data for trip generation modeling by first learning a disaggregated model on the HTS and then by up-(or down-) scaling the learned factors thanks to a model learned on the aggregated mobile phone data. They aim to estimate the number of trips per person and not the preferred transport mode as in the present study.

The present work is related to the problem of estimating recent modal shares based on a detailed but outdated survey and recent but aggregated, partial and biased digital data.
We include mobile phone data, which enables the analysis of people's movements but lacks information on the modes of transport used. Smart card data are also used for comparison purpose. They provide insights into public transport flows, but are limited to this mode and require enrichment to infer destinations. In addition, the definition of a trip differs between OD matrices built from mobile phones and smart card data. Mobile phone data define a trip as a journey between zones with at least one hour's stay in each zone, whereas smart card data capture public transport use between zones without ensuring that it's a complete journey. This distinction needs to be taken into account when using these data sources.
Updated modal shares can provide valuable information for public decision-makers and offer insights into evolving mobility patterns. The present study represents a first step toward this objective. We propose a two-step methodology. First, we estimate a DCM using HTS data, considering four transportation modes: car, transit, bicycle, and walking. The DCM computes the probabilities for each trip from the HTS to use each mode based on various environmental and mode-specific attributes. In the second step, these probabilities are applied to disaggregate recent OD matrices derived from mobile phone data and describe total flows between multiple ODs. This step allows us to estimate mode-specific OD matrices. The methodology is evaluated by comparing the estimated public transport OD matrices with those obtained from smart card data.
The remainder of this paper is organized as follows: Section 2 presents the proposed methodology. Section 3 details the setup of the experiments and the obtained results, with a particular focus on the construction of the training dataset and the comparison of several DCMs using different attributes. Section 4 concludes the paper and provides perspectives for future research.

2 Methodology

We divide the methodology into three sections: (1) estimation of the DCM from HTS, (2) disaggregation of total OD matrices, and (3) comparison with smart card OD matrices. The HTS consists of trips (e) made within a specified territory. A critical condition of this study is that the OD flows derived from mobile data correspond to OD pairs k, l located within the same territory as the HTS. Similarly, smart card data must also originate from this shared territory.

DCM estimation from HTS

From the HTS, the first step consists of building a DCM based on built attributes X. We only work on subsets of trips where a zone change is effective ($k \neq l$), as we only work on these zone-to-zone trips with mobile phone data. For each OD pair k, l, there is a number of trips N made by each transport mode m, computed as:

$$N_{k,l,m} = \sum_{e} D^{e}_{k,l}.C^{e}_{m}$$

with $D^{e}_{k,l} = 1$ if trip e is made on OD k, l, 0 otherwise. $C^{e}_{m} = 1$ if trip e is made with mode m, 0 otherwise. We also compute a set of attributes $X_{k,l}$ based on the characteristics of the corresponding OD pair k, l (e.g., travel time per mode, distance, cost, etc.). The utility for each OD k, l to be made by a transport mode m is then:

$$U_{k,l,m} = \beta_m X_{k,l} + \epsilon_{k,l,m}$$

Where β_m is a set of parameters to be estimated for the m^{th} mode. For an OD k, l, the probability to use a transport mode m is:

$$P_{k,l,m} = \frac{Y_{k,l,m} e^{\beta_m X_{k,l}}}{\sum_{m'} Y_{k,l,m'} e^{\beta_{m'} X_{k,l}}}.$$

with $Y_{k,l,m} = 1$ if the mode m is available on OD k, l, 0 otherwise. The parameters $\beta = \{\beta_{m1}, \beta_{m2}, ...\}$ of the DCM are then learned by maximizing the log-likelihood

$$LL(\beta) = \sum_{k,l} \sum_{m} N_{k,l,m} log(P_{k,l,m})$$

This model remains valid for periods close to when the survey was conducted, as travel dynamics may have evolved. Consequently, it may become biased and outdated in reflecting current travel patterns.

Disaggregate and adjust mobile OD flows

The second step consists of using this DCM to disaggregate the global OD matrices derived from mobile phone data. For each OD pair k, l, the estimated flows split by transport mode are given by:

$$\phi^m_{k,l,t} = \phi_{k,l,t} \times P_{k,l,m}$$

Where $\phi_{k,l,t}$ represents the number of trips between OD k, l at time t as provided by the mobile phone data, while $P_{k,l,m}$ denotes the computed probability of using mode m, based on the attributes $X_{k,l}$ that characterize the OD pair (k, l).
At the same time, the bias inherent in OD matrices constructed from mobile data must be addressed. The mobile phone operator's definition of a trip implies that a person must stay at least one hour in origin and destination zones (Casassa et al. (2024)). We thus define $s^q_{k,l}$, the detection status of the (q-th) trip in the HTS, which occurs between the OD pair k, l as:

$$s^q_{k,l} = \begin{cases} 1 & \text{if trip } q\text{, occurring between } k \text{ and } l\text{, would have been detected in mobile data;} \\ 0 & \text{otherwise.} \end{cases}$$

We then quantify in the HTS the probability for a trip to be detected with mobile data $\tau_{k,l}(t, m)$ between the OD pair k, l. To estimate this probability a logit model is trained, including the following attributes: the distance between OD k, l, the time of day and the transport mode. Subsequently, the biased OD flows from the mobile phone data are adjusted accordingly as:

$$\hat{\phi}^m_{k,l,t} = \phi^m_{k,l,t} / \tau_{k,l}(t, m)$$

Comparison with smart card OD flows

For comparison purposes, we need to introduce a third data source: the smart card data. Let us introduce $\chi_{a,b,t}$, the number of people entering the public transport network at station a at time t and leaving it at station b. This number can be obtained from the tap-in only smart card data on which we apply the trip chaining method (Trépanier et al. (2007)), allowing us to estimate the most probable exit station for each user. As ODs k, l may be composed of small areas, disaggregated mobile OD flows are compared with public transport OD flows at the level of communes and arrondissements for Lyon, by defining $\chi_{i,j,t}$ the smart card OD flows and $\hat{\phi}^m_{i,j,t}$ the disaggregated mobile OD flows between communes i, j at time t:

$$\chi_{i,j,t} = \sum_{a \in I, b \in J} \chi_{a,b,t}$$

$$\hat{\phi}^m_{i,j,t} = \sum_{k \in I, l \in J} \hat{\phi}^m_{k,l,t}$$

with I and J the spatial coverages for communes (or arrondissements) i and j respectively. Given the substantial differences in trip definitions between the two types of OD matrices (mobile phone data and smart card data), it may be necessary to adjust the disaggregated OD flow values accordingly. For comparison we introduce $\chi_{i,j}$ and $\hat{\phi}^{m=\text{`transit'}}_{i,j}$, the daily aggregated volumes. We use a simple linear regression between daily public transport volumes from smart cards $\chi_{i,j}$ and disaggregated mobile data volumes $\hat{\phi}^{m=\text{`transit'}}_{i,j}$ of the form

$$\chi_{i,j} = \delta \hat{\phi}^{m=\text{`transit'}}_{i,j} + \varepsilon_{i,j}$$

where δ is a factor to be estimated and $\varepsilon_{i,j}$ is an error term. We call $\bar{\phi}^{m=\text{`transit'}}_{i,j} = \delta \hat{\phi}^{m=\text{`transit'}}_{i,j}$, the adjusted public transport flows from mobile OD matrices.
Then, it is possible to compare, for the public transit, the values of smart card ODs $\chi_{i,j,.}$ and estimated public transport ODs $\bar{\phi}^{m=\text{`transit'}}_{i,j,.}$. This comparison can help determine whether our methodology successfully inferred public transport shares from the HTS that align closely with those calculated from recent data.

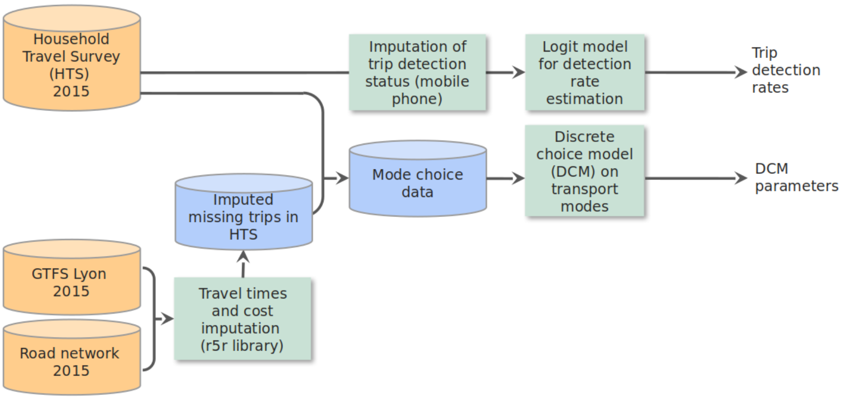
Household Travel Survey (HTS) 2015
Imputation of trip detection status (mobile phone)
Logit model for detection rate estimation
Trip detection rates
Mode choice data
Discrete choice model (DCM) on transport modes
DCM parameters
Imputed missing trips in HTS
GTFS Lyon 2015
Road network 2015
Travel times and cost imputation (r5r library)

Figure 1: Data construction process and mode choice model

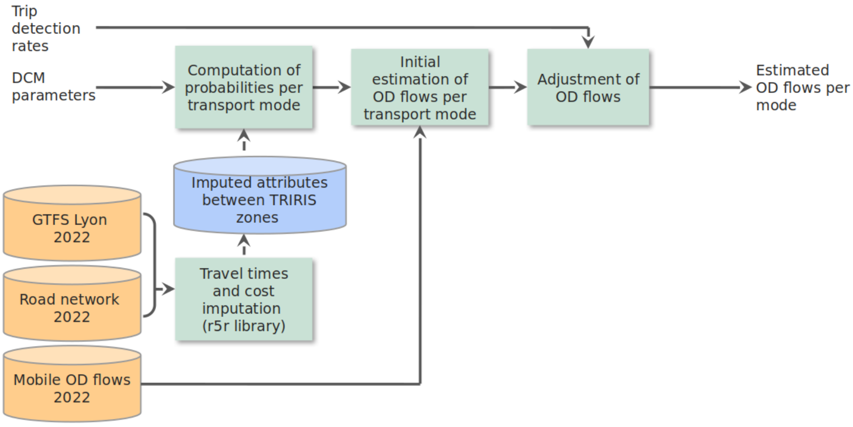
Trip detection rates
DCM parameters
Computation of probabilities per transport mode
Initial estimation of OD flows per transport mode
Adjustment of OD flows
Estimated OD flows per mode
Imputed attributes between TRIRIS zones
GTFS Lyon 2022
Road network 2022
Travel times and cost imputation (r5r library)
Mobile OD flows 2022

Figure 2: Disaggregation process of the total ODs flows into mode-specific OD flows

3 Results and discussion

We illustrate the methodology using the Lyon metropolitan area in France as a case study. The data comes from a HTS conducted on a typical weekday in 2015, covering a representative sample of $28,230$ individuals from the Lyon metropolitan area and its surroundings. The survey includes trip records with information on the chosen transport mode (car, transit, bike, walking) and the origin and destination census tracts. Additionally, we have recent origin-destination (OD) matrices of total flows between zones, called "TRIRIS," based on 2-hour timestamps from September 2022. These matrices were generated from mobile phone data by a telecom operator. As previously stated, these matrices are inherently biased, as trips are only recorded when the origin and destination TRIRIS differ, and individuals must spend at least one hour in each area. Finally, we also have tap-in validation data for the public transport network in Lyon from the same period, which is used to estimate transit OD matrices between public transport stations.

In the following sections, we describe the process of building a training dataset and detail the results of our two-step method for estimating OD flows split by mode. Figure 1 shows the overall dataset construction and the DCM training process. Figure 2 shows the disaggregation process of the total ODs flows.

Training dataset construction

The HTS does not provide all possible routes. For instance, it may include a few examples of car trips between two zones but lack examples of the same trip by bike. Therefore, we need to impute missing travel mode choices and their attributes in the HTS. Travel times, distances, and transit connections between TRIRIS zones are obtained using the r5r library in R. These attributes are computed based on the shortest paths between each zone centroïds for each transport mode, with a maximum trip duration of 180 minutes. As a result, public transport, walking, and biking are unavailable for certain OD pairs due to excessive travel times or insufficient infrastructure. The process relies on the road network (OpenStreetMap data) and public transport networks (GTFS) relevant to the survey period. Since the HTS is from 2015, we use 2015 road network data. We use the current GTFS data for public transport and exclude any new lines or extensions added after 2015. From the computed distances, we estimate travel costs for transit and car modes. Car cost per kilometer values are based on Sivardière (2013) and are set at 24.4 cents/km.We assume zero costs for walk and bicycle modes. Moreover, we don't use any travel cost variable for public transport because it is a fixed fare that does not depend on the distance traveled (ticket price or pass). This factor is expected to be considered in the alternative spectific constant factor related to public transport.

DCM training

Several DCMs can be built with an increasing number of attributes X. In this section, we compare six models described below. The baseline model (model 1) includes the travel time per mode the travel costs for the car. Model 1 also includes the distances between each OD and the number of needed public transport interchanges from rail mode (subway, tramway) to bus. Model 2 is an enrichment of model 1 in which we included some time of day attributes. Two binary variables specify whether the trip is made during the morning peak (7am-10am), evening peak (4pm-7pm), or none. In model 3, we added two binary variables for car and public transportation, specifying if the trip begins (and ends) in the dense urban area (Lyon and Villeurbanne). Due to traffic jams and more accessibility to public transport, these coefficients may penalize the car and give weight to public transport. We add some specificities related to public transport accessibility in model 4: the percentage of departure zones less than 300 meters from a metro station and less than 300 meters from a tram station. In model 5, we add a variable for the percentage of free parking at the destination. Finally, model 6 includes some variables related to the origin and destination median revenues. Trips between rich zones are expected to give more weight to the car.
To compare the models, we build a 10-fold cross validation process. At each iteration, we randomly select 80% of the mode choice dataset for training and the remaining 20% for testing the learned model. We compute the log-likelihood (LL) for the train set (LL train), the test set (LL test), and the Brier score for the test set. Mean results are shown in table 1.

Model	**Attributes**	**LL train**	**LL test**	**BS test**
1	Travel time, Car cost, OD distances, Number buses transfers	-17667	-4380	0.138
2	Attributes model 1 + Trip period	-17621	-4372	0.137
3	Attributes model 2 + Trip inside/outside dense urban area	-17031	-4300	0.134
4	Attributes model 3 + Origin subway/tramway accessibility	-16982	-4247	0.132
5	Attributes model 4 + Destination percentage free parkings	-16977	**-4203**	0.131
6	Attributes model 5 + Origin/Destination median revenues	**-16934**	-4227	**0.130**

Table 1: Comparison of several mode choice models with different attributes. The mean log-likelihood on the training and test sets and Brier scores from 10 iterations are presented.

From table 1 we can observe a good improving with model 3 in which we included the binary variable specifying if the trip departured/ended in the dense urban area. Model 6 is the best model regarding the results obtained on the train set and the Brier score. We thus visualize the results of this model in the remaining study.
As the next section focuses on the comparison of the transit mode only, we need to quantify how much the DCM was able or not to predict the transit mode choice in HTS. We predict modes probabilities from the chosen DCM on the whole HTS and compute for each OD i, j the ratio between the estimated number of public transport trips and the real number as:

$$S_{i,j}^{HTS} = \frac{\sum_e D_{i,j}^e P_{e,m='transit'}}{\sum_e C_e^m D_e^{i,j}}$$

with $C_e^m = 1$ if the trip e has the option $m =' transit'$ and $D_e^{i,j} = 1$ if the trip e is made between the OD pair i, j. The results are shown in figure 3. Overall, scores are around 1, meaning there is a perfect match between the observed and estimated transit OD flows. The two ODs 69120 → SAINT-PRIEST and SAINT-PRIEST → 69120 show scores > 1.5, which depicts an over-estimate of the use of public transport between this suburban city and this district (69120). SAINT-PRIEST is equipped with tramway infrastructures, but this may not be sufficient for the attractiveness of public transport in this city compared with other suburban cities.
The next section compares the disaggregated and adjusted mobile OD matrices with smart card OD flows from the best DCM parameters.

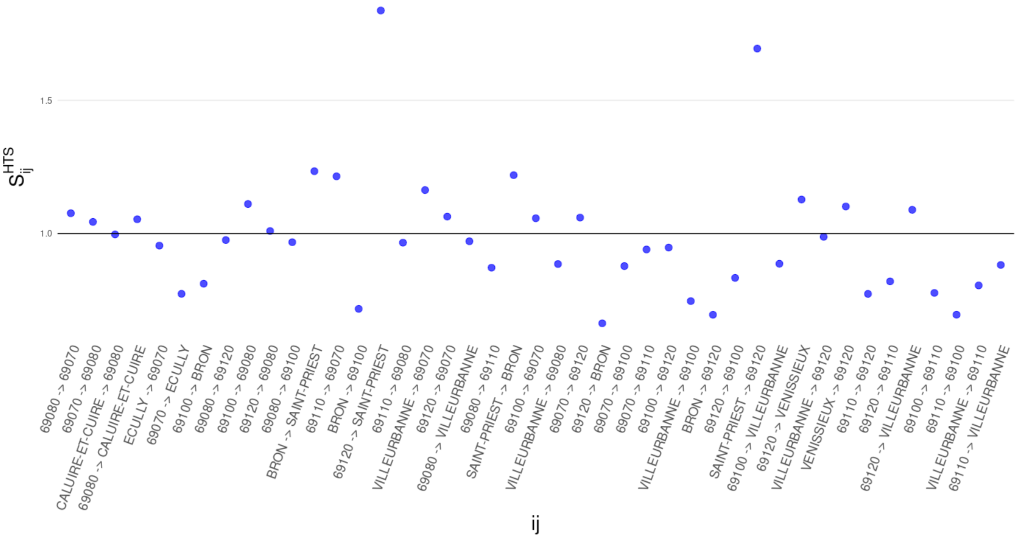

Figure 3: Scores S_{ij}^{HTS} computed for ij where there were more than 100 trips made by public transport from HTS. A score equal to 1 shows a perfect match between the observations and estimations.

Comparison with smart card OD matrices

We compare the three quantities $\chi_{i,j,t}$, $\hat{\phi}_{i,j,t}^{m=transit}$ and $\bar{\phi}_{i,j,t}^{m=transit}$ to check the extent to which the HTS-based model is accurate in estimating the current use of public transport. We represent the scatter plots of daily flows computed for the three quantities in figure 4.

This comparison reveals that the observed $\chi_{i,j}$ are most of the time much higher than the disaggregated public transport OD flows $\hat{\phi}_{i,j}^{m=transit}$ (figure 4(a)). Consequently, we underestimate public transport flows for most ODs. We also compute a $RMSE = 9593$ in this case, which is high.

The value computed for δ factor was 2.78, allowing to adjust the disaggregated public transport OD flows. We obtained the comparison shown in figure 4(b) and a $RMSE = 5932$, which is still high but much better than a model without adjustment.

4 Conclusion and perspectives

With this work, we investigate how to estimate mode choice probabilities given computable attributes from ODs' characteristics. This problem is difficult as the HTS used to fit the DCM is outdated. Moreover, the differences between definitions of a trip for ODs based on smart card data, mobile phone data, and HTS are challenging to handle. To quantify to what extent this model can estimate good probabilities, we disaggregate mobile OD flows using the DCM and compare the obtained public transport flows with OD flows estimated from smart card data. For this comparison to be effective, adjustments had to be made to the disaggregated mobile data to account for significant differences in total volumes between the two sources (smart card and mobile data) and the differences in definitions of what constitutes a trip. These elements could serve as the basis for a model combining these different sources simultaneously, which would better match current mobility trends.

The proposed approach can be extended by developing a joint model that integrates recent digital data into an enriched DCM framework, allowing parameter estimation from both data sources. A potential inspiration for this methodology can be found in Bwambale et al. (2021), where DCM parameters are initially derived from HTS data and subsequently rescaled using digital data. Their proposed method is a joint model that looks relevant when disaggregated HTS and aggregated digital data come from distinct periods because the different sources do not include the same database for learning. On the other hand, initial training is carried out on the HTS, followed by a second on more recent digital data. Moreover, the sensitivity of the DCM parameters to different factors when choosing one or other modes is estimated on the HTS. The parameters are then rescaled using more recent digital data.

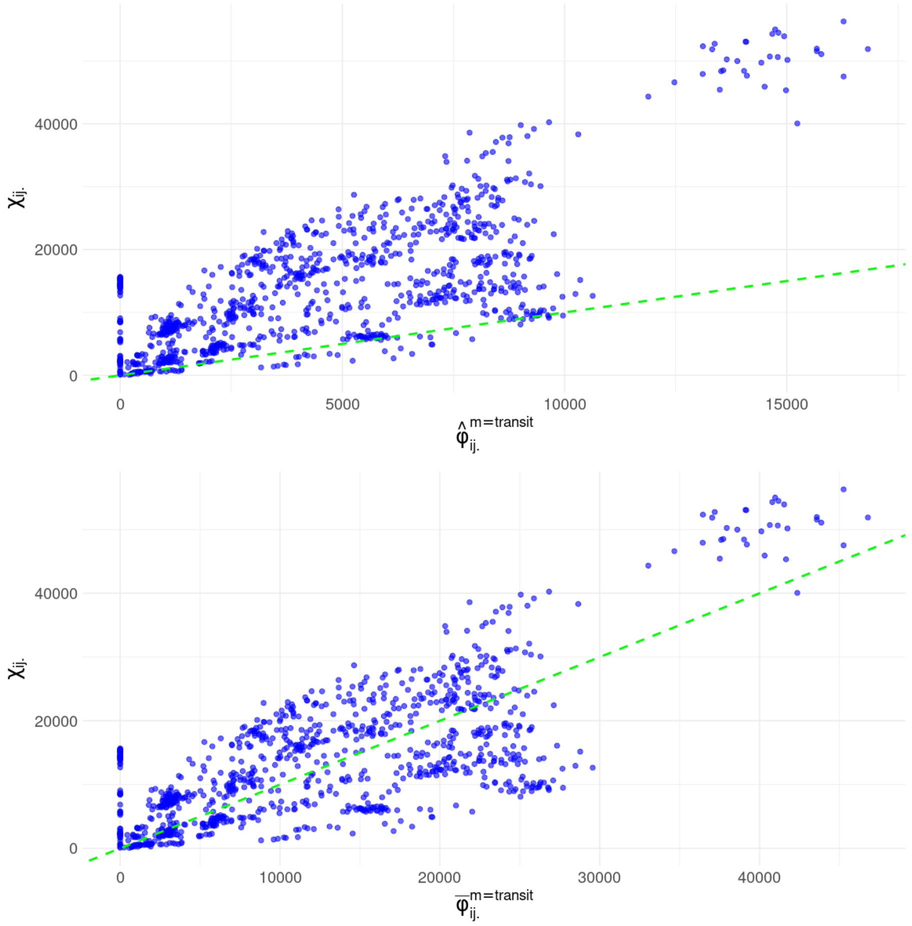

Figure 4: Comparison of daily OD flows in September 2022 between disaggregated mobile data volumes before adjustment $\hat{\phi}_{ij}^{m='transit'}$ (top) and after adjustment $\bar{\phi}_{ij}^{m='transit'}$ (bottom) with smart card OD flows.

Acknowledgement

This research is conducted within the MobiTIC research project (grant number ANR-19-CE22-0010), funded by the Agence Nationale de la Recherche in France.

References

Andersson, A., Kristoffersson, I., Daly, A., & Börjesson, M. (2024). Long-distance mode choice estimation on joint travel survey and mobile phone network data. *Transportation Research Part A: Policy and Practice*, *190*, 104293.

Bwambale, A., Choudhury, C. F., Hess, S., & Iqbal, M. S. (2021). Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. *Transportation*, *48*, 2287–2314.

Casassa, E., Come, E., & Oukhellou, L. (2024). Detected or undetected, which trips are seen in mobile phone od data? a case study of the lyon region (france). *TRC 30*.

Heinen, E., & Chatterjee, K. (2015). The same mode again? an exploration of mode choice variability in great britain using the national travel survey. *Transportation Research Part A: Policy and Practice*, *78*, 266–282.

Krueger, R., Bierlaire, M., & Bansal, P. (2023). A data fusion approach for ride-sourcing demand estimation: A discrete choice model with sampling and endogeneity corrections. *Transportation Research Part C: Emerging Technologies*, *152*, 104180.

Ounoughi, C., & Yahia, S. B. (2023). Data fusion for its: A systematic literature review. *Information Fusion*, *89*, 267–291.

Sivardière, J. (2013). Déplacements de proximité: les coûts pour le consommateur. *Transports Urbains*(2), 12–17.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, *11*(1), 1–14.

Zhang, R., Ye, X., Wang, K., Li, D., & Zhu, J. (2019). Development of commute mode choice model by integrating actively and passively collected travel data. *Sustainability*, *11*(10), 2730.