Use of Origin-Destination data for calibration and spatialization of synthetic travel demand

Benoît Matet*[1,2], Etienne Côme[1], Angelo Furno[2], Sebastian Hörl[1,3], and Latifa Oukhellou[1]

[1]Université Gustave Eiffel, COSYS, GRETTIA, France
[2]Université Gustave Eiffel, COSYS, LICIT-ECO7, France
[3]IRT SystemX, France

Short summary

The dynamics of urban transportation can be understood with activity-based models, which rely on synthetic travel demand data to get a comprehensive understanding of urban mobility. These data are usually derived from small population samples and surveys, which may be expensive and do not adequately cover the spatial trajectories of the users. In this paper, we explore the use of a time-dependent origin-destination (OD) matrix derived from mobile phone data for the attribution of locations in a synthetic population for the city of Lyon, France. OD matrix data can also mitigate uncertainties or outdated information in travel surveys regarding flows by time of day and between zones. The resulting population enrichment is measured in terms of fit to the input mobility data.

Keywords: activity-based modeling, data fusion, multi-modal transportation, origin-destination matrix, synthetic demand

1 Introduction

One way of acquiring insight into a city's transportation system is to simulate an adequately generated synthetic population of agents. Assuming the synthetic population is representative of the real one and the simulator accurately describes how travelers make decisions, the result of the simulation should be a comprehensive description of the city's mobility dynamics. The synthetic population should match the known marginal distributions of the real population while staying as close as possible to the joint distribution of variables observed from an input population sample. In activity-based models, each agent must also have an agenda, stating a chain of activities (e.g. home, work, study, shopping, other), corresponding times and transport modes, and a chain of locations for each activity. The agendas should be likely at the individual level and ultimately match flows observed from other data sources.

Hörl & Balac (2021) propose a general pipeline to create a synthetic population with socioeconomic variables and activity chains from multiple data sources. In a first step, a population sample is used as input to set the socioeconomic variables of the agents. When the input sample is small, the approaches proposed by Sun & Erath (2015); Sun et al. (2018) allow modeling the underlying probability law and sampling the desired number of agents from it. In a second step, these agents can be extended with attributes that are not in the population sample, ensuring that the available marginal statistics are matched. Activity chains, without locations, are then assigned to the agents. These activity chains can either be available from a separate mobility survey or generated by approaches such as the ones proposed by Joubert & de Waal (2020); Anda et al. (2020). Activity locations are usually separated into primary locations, such as home or workplace, and secondary locations for other types of activity. While primary locations can be drawn in the first steps of the generation, secondary locations are usually sampled through a process aiming at mimicking the decision rules of an individual choosing where to go for a particular task (Ma & Klein (2017); Hörl & Axhausen (2021)).

In this work, we propose to expand the state-of-the-art pipeline with a novel data-driven approach for the location of activities, leveraging mobile data from the telecommunications provider Orange. Mobile data have recently become an interesting alternative to mobility surveys, as they are much cheaper and faster to generate. They can also be more representative regarding spatial features,

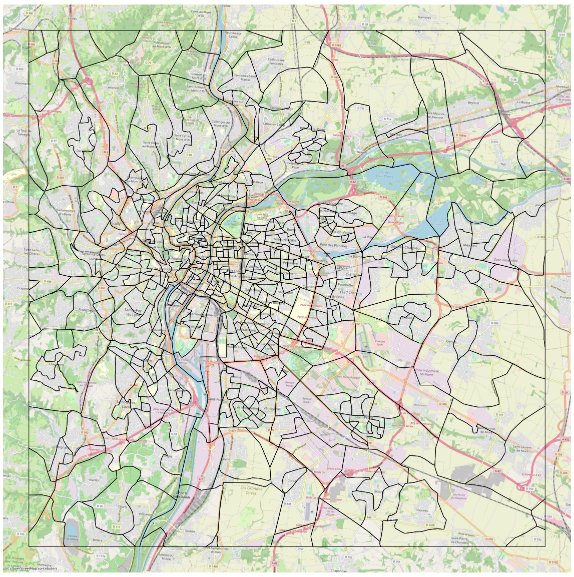

Figure 1: Map of the partitioning of the study area of the city of Lyon (France), Background: OpenStreetMap

as around 30% of people in France are customers of Orange, compared with 1% of persons reached by mobility surveys. Once re-scaled, the data gives us a reliable estimation of the trips of the total population. However, they lack the level of detail offered by surveys, which makes it interesting to use the data sources together in order to obtain a dataset that would be both detailed and easily adaptable to new observations.

Mobile data can take the form of full trajectories of users, on which rely approaches proposed by Zilske & Nagel (2015); M. Yin et al. (2017). However, full trajectories are computationally heavy and feature a significant privacy risk. We thus derive a time-dependent Origin-Destination (OD) matrix from mobile phone reconstructed trajectories (Bonnetain et al. (2021)), giving only a map of flows between the zones of our study area for each time step of the day. OD matrices are lighter and more manageable than whole trajectory datasets and can be fully anonymized as shown in L. Yin et al. (2015); Matet et al. (2021). This makes them more readily available and safer regarding the privacy of transportation users. As our first contribution, we interpret this OD matrix as a transition probability between two locations, which we use to sample the location chain. Second, we use the total number of trips in the OD matrix for each time of the day as a target that the agendas of the population should match.

In the case of France, valuable data sources for the complete generation process are available to researchers upon request. The census performed by the French statistics institute INSEE features socioeconomic variables of a population sample large enough to be representative of the whole population. The activity chains are taken from the Enquête Ménage Déplacements (EMD) performed by the French agency for urban planning Cerema (2015), which details the complete agendas of 25,203 persons in the Lyon region.

2 Methodology

The first steps of the synthetic travel demand generation are inspired from Hörl & Balac (2021), which focuses on Île-de-France and Paris, now with the target area being Lyon and its surroundings. Our study area takes the form of a square 25,600 meters wide centered around Lyon, divided into 515 distinct zones which either represent municipalities or spatial units (IRIS) common to many statistical analyses in France. Figure 1 shows the partitioning of the study area.

In the case of France, the census performed by the French statistics institute INSEE constitutes a population sample extensive enough so that no additional synthesis is required. Each of the 487,628 rows for our study zone features socioeconomic variables and a non-integer scaling coefficient for a total of 1,366,072 persons. To interpret the data as an integer number of agents, we apply the TRS approach (Truncate, Replicate, Sample) introduced by Lovelace & Ballas (2013). Each coefficient is stochastically rounded up with probability equal to its decimal part, and rounded down otherwise. The result is a population of 1,372,566 agents described by the variables detailed in Table 1. Note that we choose not to retain the household structures of the population. In a generative approach, this information would be necessary in the downstream modeling steps. In our approach, activity chains are assigned by sociodemographic attributes, hence only individuals are relevant.

Table 1: Socioeconomic variables for the synthetic population

Variable	**Modalities**
Home zone	515 zones in the study area
Age	0-17; 18-29; 30-59; 60+
Gender	male or female
Occupation	jobless; farmer; independant; executive; employee; intermediate professions; worker; student; retired
Has car	yes or no
Home status	Owner; Tenant; Social housing

The assignment of activity chains from the EMD travel survey to agents from the census follows a process inspired from statistical matching (D'Orazio et al. (2006)). As in Hörl & Axhausen (2021), this process is close to a regular join between relational databases with a join key of multiple variables, except that each census row is required to match at least 20 EMD rows. For census rows for which this is not the case, we successively remove the last variable from the join key until more than 20 EMD rows match. Then, one of these EMD rows is randomly drawn, and its activity chain is assigned to the person described by the census row. This process guarantees diversity in the assignment of chains to population agents while joining as many variables as possible so that the agents have activity chains relevant to their characteristics. The variables used in the join key are described in Table 2.

Table 2: Variables used in the statistical matching to assign activity chains from EMD to agents from census

Variable	**Number of modalities**	**% of agents with more than 20 matches**
Age	4	100%
Gender	2	100%
Occupation	9	95%
Has car	2	94%
Home status	3	90%
Canton	8	75%

Note that although the census is the same, the EMD differs between Paris and Lyon. We cannot use the income in the join key as is done by Hörl & Balac (2021) because the EMD for Lyon does not feature it. We replace it with what is arguably a proxy for one's wealth, i.e., whether the person is the owner, tenant of their home, or lives in social housing. Similarly, instead of the department used in Île-de-France, we use the canton, an intermediary administrative zoning system between our zoning and the department. The resulting population comprises agents, each equipped with a chain of trips specifying the purposes, time of day, and transport modes, as is detailed in Table 3. Note that the day is divided into time steps of one or more hours to be consistent with the division used for anonymization in the mobile data. The activity chains contained in the EMD feature up to 12 trips during the day.

Table 3: Variables describing trips

Variable	**Modalities**
Trip purpose	home; work; study; shopping; personal
Time step	0h-2h; 2h-5h; 5h-7h; 7h; 8h; 9h; 10h-12h; 12h-14h; 14h-16h; 16h; 17h; 18h; 19h; 20h-22h; 22h-0h;
Transport mode	foot; bicycle; car or motorcycle; public transport

Rescaling

While the previous steps correspond to existing approaches for travel demand synthesis, we propose the following steps that make direct use of a large-scale time-dependent Origin-Destination matrix that is featured in this research. Due to data discrepancies, the number of trips performed in each time step of the day by our agent population does not match the number of trips observed from mobile data. This is illustrated in Figure 2, where each point represents the marginal value of a joint socioeconomic attribute (in blue) or the total number of trips taken for a given time step (in yellow). The x-coordinate of the point is the truth value from the available data, i.e., the marginals from the census in the case of blue points and the number of trips from mobile data for yellow points. The y-coordinate is the total measured in our synthetic population. We perform rescaling via Iterative Proportional Updating (IPU) as introduced by Ye et al. (2009), with the adaptation that instead of a population composed of households divided into persons, we have a population of persons divided into trips. The resulting population is consistent both with the socioeconomic composition of the real population and the number of trips they take at each time step of the day.

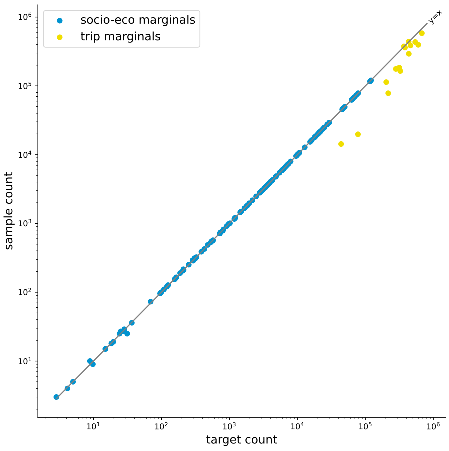

Figure 2: Marginals of the synthetic population versus what is expected from the data sources. Each blue dot is the volume for a socioeconomic modality. Each yellow dot is volume of trips for a given time step.

Spatialization

To be complete, the synthetic population requires locations for each activity in their agenda. We interpret the OD matrix as a probability table of destinations D given the origin O and time T: $P(D|O,T)$. For each agent, we model the chain of locations during the day as a Markov chain with transition probability $P(D|O,T)$.

The activity purpose may specify that the agent is going home, in which case the home area defined in the first step of the generation from the census is deterministically assigned as a location. In contrast with state-of-the-art methods, we do not rely on pre-defined work or study places, as we do for home. This is justified by the fact that, as 45% of trips in the activity chains of our population are commute trips, pre-defining the work and study places would not only half the use of the data we aim at exploiting but also invalidate it as our mobile data would still contain the distribution of the sum of commute and non-commute trips but be used only for non-commute trips.

However, even without pre-defining them, we still want to ensure that work and study activities have the same location when they appear more than once in the activity chain. The resulting probability law for the chain of locations is described by a Markov chain with fixed states for home and linked states for work and study. Because of these dependencies between states, we cannot sample from it as we would a regular Markov chain. We resort to Gibbs sampling to generate a chain of locations. This drawing process can only manage a fixed activity chain with a fixed pre-defined home zone, meaning that two agents with either a different activity chain or a different home require each a distinct Gibbs sampling. We obtain 478,963 distinct processes to run, which is computationally expensive because of the required warm-up phase: Gibbs sampling is considered

to yield the true joint distribution only after having discarded the first thousand samples. We reduce the computing time by observing that most activity chains have a low autonomy, i.e. they feature a limited number of unknown locations between two fixed home activities. Segments of the activity chain that are separated by a fixed state are independent of each other and follow a distribution of much lesser dimension. In fact, segments of only one unknown activity do not require Gibbs sampling, and segments of two activities only require a limited warm-up.
As a refinement, we consider the transport modes M from trips in the activity chains to obtain a transition probability $P(D|O,T,M)$. The transport mode is not specified in the mobile OD data, but can be integrated into the sampling process using Bayes rule: $P(D|O,T,M) \propto P(M|O,D,T) \times P(D|O,T)$. As an estimation of $P(M|O,D,T)$, we consider that the mode M depends only on the distance L between origin O and destination D. Using all trips from the EMD, we obtain an empirical probability distribution $P(M|L(O,D))$ discretized to individual 1 km bins.

3 Results

As the synthetic population is expected to behave like the real population, we evaluate it on all the available indicators about the composition or mobility of the real population. Note that our various input sources can be inconsistent between themselves, e.g. on the number of trips for each time step or even the socioeconomic composition of the population between the census and the EMD. As such, it is already a satisfying result for the synthetic population to match indicators derived from our own input data.

Socioeconomic composition of the population

In Figure 3, we illustrate how the socioeconomic composition of our population matches the official census better than the transport survey EMD. As the EMD represents only a small number of people, it is normal that the totals have a higher variance. Our synthetic population can then be seen as a version of the EMD that agrees with the census on the socioeconomic distribution of the population.

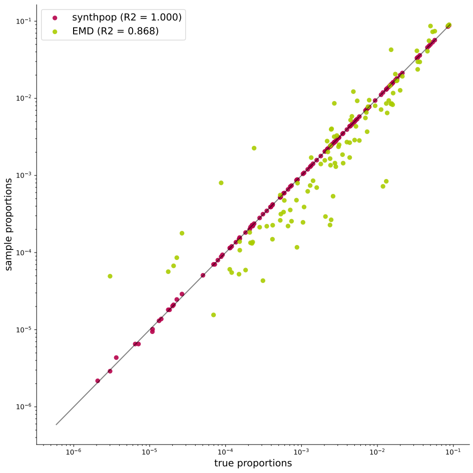

Figure 3: Matching of the socioeconomic marginals of our rescaled synthetic population (in red) compared to the transport survey EMD (in green).

Number of trips by hour

In Figure 4, we illustrate the distribution of trips during a typical day as measured from mobile data (in orange), our synthetic data (in magenta), and the survey EMD (in green). In this case, we consider the observed volumes from the mobile data to be the ground truth. We see that our population can also fit the ground truth source better than the official survey EMD. In particular, the EMD seems to overestimate the volumes of the morning, midday, and evening peaks while underestimating the volumes during the rest of the day.

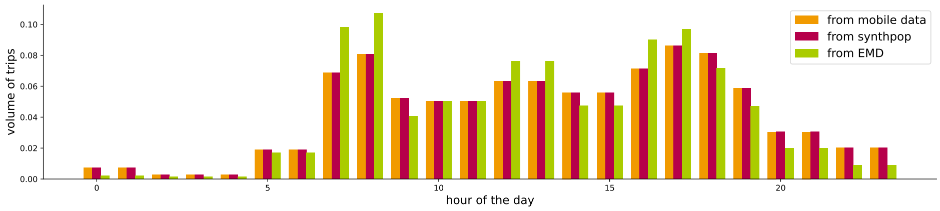

Figure 4: Distribution of trips taken during the day.

Conditional probability of destinations

In Figure 5, we illustrate how the probability of destinations given origin and hour fits the probability table derived from the OD matrix. Each point corresponds to a combination of destination, origin, and time of the day. Its x-coordinate is the probability $P(D|O,T)$ observed in the OD matrix, while its y-coordinate is the same probability as observed in the trips of our synthetic population. Interestingly, we observe a bi-modality in the ground truth: it seems that each origin has a restricted list of favorite destinations forming the upper-right cloud and a list of secondary destinations in the lower-left cloud. We see that this bi-modality is retrieved in our synthetic population and that, overall, our synthetic demand is highly correlated to the actual observations.

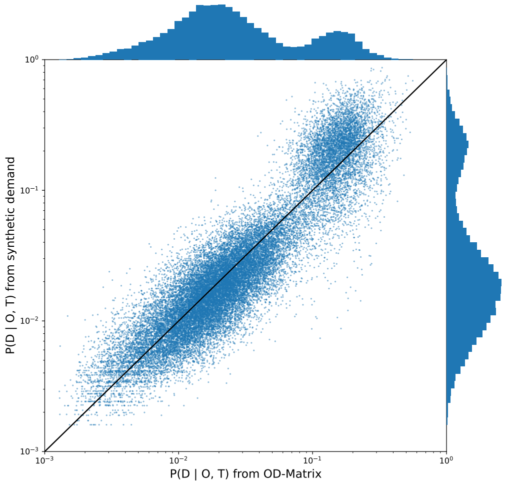

Figure 5: $P(D|O,T)$, from our synthetic data w.r.t. ground truth value from the OD matrix. The black line represents $y = x$.

Limitations

While it can be expected that our synthetic demand data matches well the distribution $P(D|O,T)$, we observe that the distribution of trips $P(O,D|T)$ is not the same as in the input OD matrix. This is because although we sample the destinations using $P(D|O,T)$ as a transition probability, we have no mechanism to make sure that the right number of agents leave each individual origin O at each time step T. This problem can be addressed by decomposing each time slice of the transition matrix into a sum of transition matrices depending on O, T, and on the time step of the next trip of the agent. This new decomposition of the OD matrix in our future work amounts to adding explanatory variables to the mobile data in the same fashion as we added the transport mode. By carefully choosing such a decomposition, we can make sure the agents taking a trip before time step T are assigned to destinations such that the map of agents leaving for a trip at time step T corresponds to the map of origins of trips in the OD matrix.

4 Conclusions and perspectives

This paper illustrates how a time-dependent OD matrix from mobile data can improve the generation process for synthetic travel demand. As we consider mobile data to be closer to the ground truth than surveys regarding the number of trips by the time of day and between zones, they are a valuable asset to take into account in addition to already available sources. Our approach makes use of the highly valuable activity chain structures that are already featured in the surveys and successfully improves on them with mobile data, using an OD matrix both as a rescaling target and a basis for spatialization.

As OD matrices can be anonymized, our approach is also relevant to leverage the richness of mobile data without the computational cost nor the privacy hazard of full trajectory data. However, a huge challenge lies in the fact that OD matrices derived from mobile phones are not dependent on anything other than time. As such, correlations between the spatial characteristics of the trips (such as average commute distance) and other variables are hard to capture. A more detailed OD matrix could be derived from the initial trajectories without additional data: for example, recurrent daily trips could be identified as commute patterns, resulting in two separated OD matrices for commute and non-commute.

Acknowledgements

This research is supported by the French ANR research projects MOBITIC (grant number ANR-19-CE22-0010) and PROMENADE (grant number ANR-18-CE22-0008).

References

Anda, C., Ordonez Medina, S. A., & Axhausen, K. (2020, 09). Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data. doi: 10.3929/ethz-b-000442517

Bonnetain, L., Furno, A., El Faouzi, N.-E., Fiore, M., Stanica, R., Smoreda, Z., & Ziemlicki, C. (2021). Transit: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. *Transportation Research Part C: Emerging Technologies*, *130*, 103257.

Cerema. (2015). lil-1023: Enquête ménage déplacement, lyon / aire métropolitaine lyonnaise.

D'Orazio, M., Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* doi: 10.1002/0470023554

Hörl, S., & Axhausen, K. (2021, 10). Relaxation–discretization algorithm for spatially constrained secondary location assignment. *Transportmetrica A: Transport Science*. doi: 10.1080/23249935.2021.1982068

Hörl, S., & Balac, M. (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. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0968090X21003016` doi: https://doi.org/10.1016/j.trc.2021.103291

Joubert, J., & de Waal, A. (2020, 09). Activity-based travel demand generation using bayesian networks. *Transportation Research Part C Emerging Technologies*, *120*.

Lovelace, R., & Ballas, D. (2013, 09). 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers Environment and Urban Systems*, *41*. doi: 10.1016/j.compenvurbsys.2013.03.004

Ma, T.-y., & Klein, S. (2017, 09). Bayesian networks for constrained location choice modeling using structural restrictions and model averaging. *European Journal of Transport and Infrastructure Research*, *18*.

Matet, B., Côme, E., Furno, A., Bonnetain, L., Oukhellou, L., & El Faouzi, N.-E. (2021, 01). A lightweight approach for origin-destination matrix anonymization. In (p. 487-492). doi: 10.14428/esann/2021.ES2021-56

Sun, L., & Erath, A. (2015, 10). A bayesian network approach for population synthesis. *Transportation Research Part C Emerging Technologies*, *61*, 49-62.

Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, *114*, 199-212.

Ye, X., Konduri, K., Pendyala, R., Sana, B., & Waddell, P. (2009, 01). Methodology to match distributions of both household and person attributes in generation of synthetic populations.

Yin, L., Wang, Q., Shaw, S.-L., Fang, Z., Hu, J., Tao, Y., & Wang, W. (2015, 10). Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PLOS ONE*, *10*(10), 1-23. Retrieved from `https://doi.org/10.1371/journal.pone.0140589` doi: 10.1371/journal.pone.0140589

Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F., & Pozdnoukhov, A. (2017, 05). A generative model of urban activities from cellular data. *IEEE Transactions on Intelligent Transportation Systems*, *PP*, 1-15. doi: 10.1109/TITS.2017.2695438

Zilske, M., & Nagel, K. (2015). A simulation-based approach for constructing all-day travel chains from mobile phone data. *Procedia Computer Science*, *52*, 468–475.