Generating Synthetic mobile phone datasets using MATSim

Joseph Molloy

Michael Cik

Martin Fellendorf

Kay W. Axhausen

Conference paper STRC 2020

STRC 20th Swiss Transport Research Conference
Monte Verità / Ascona, May 13 – 15, 2020

Generating Synthetic mobile phone datasets using MATSim

Joseph Molloy
IVT
ETH Zurich
CH-8093 Zurich
joseph.molloy@ivt.baug.ethz.ch

Michael Cik
ISV
TU Graz

Martin Fellendorf
ISV
TU Graz

Kay W. Axhausen
IVT
ETH Zurich

Abstract

Passively collected datasets of mobile phone traces are increasingly used for the generation of transportation models. Datasets can contain more than 2000 location events per person per day and can observe hundreds of thousands of participants with no response burden. Hence, such datasets are very attractive for transport modelling, particularly on a regional level. However, privacy regulations make accessing, working with, and sharing such data challenging. We propose an approach for the generation of open, synthetic mobile phone traces, based on a small sample of network traces, information on the location of the network antennas, and activity patterns from a MATSim scenario. Such datasets will allow for better collaboration between researchers on the development of new algorithms for extracting travel plans and other indicators. Previous approaches only generated synthetic traces for CDR (Call detail record) data, which contains many less data points than traces from 3G and 4G networks. The method accommodates different network types (GSM, 3G, 4G etc), and the introduction of important data artefacts such as pinging and loss of reception. Using the proposed method, a the first steps towards a synthetic network trace dataset for Switzerland calibrated from Austrian network traces is presented.

Keywords

Mobile phone data, synthetic data, open data, anonymisation, transport modelling, MATSim

1 Introduction

Passively collected data from mobile phone networks has allowed mobility research on a scale that was unprecedented a decade ago. Previously, the main data sources for transport models were traffic counts, travel diaries or small samples of GPS data. Traffic counts (excluding number plate recognition) is link-based and not person-based, while travel diaries are expensive and time consuming to conduct, as are GPS surveys. Network traces - also known as Call Detail Records (CDR) - can be collected for millions of unsuspecting persons over a period of time, giving a low resolution picture of mobility patterns within a ciy, region or country. Blondel *et al.* (2015) reported that market penetration had reached 128% in the developed work, and 90% in developing countries. However, the revealing nature of these datasets means that they can normally not be shared openly, even when anonymised. The first mobile phone datasets only recorded a rough location for each phone call or SMS, with between 2 and 6 location points per day. More recently, with the advent of mobile internet and low-range 4/5G antennae, the resolution of CDR data can be as low as 50m in urban areas, with thousands of data points per person per day. This opens up new applications for CDR data, but reduces the likelyhood even further that such data will be made publicly available. As such, there is a need for methods to generate open CDR datasets that allow collaboration and open research.

2 Related Work

2.1 Mobile data in research

The use of moible phone data is increasingly seen in different research areas in recent years. However they also emphasise the difficulties in collaboratively working with such data. Focusing on the mobility space, Gonzalez *et al.* (2008) using CDR data to fit a power law to human mobility patterns for 100,000 individuals over a six-month period, showing that human trajectories show a high degree of both temporal and spatial regularity. Janzen *et al.* (2018) examined long distance travel behaviour using mobile billing data (CDRs) in France, showing that travel surveys can under-represent the number of long distance trips by a factor of 2. Furthermore, approaches for generating OD-matrices for transport planning from CDR data are well estabilised (Alexander *et al.*, 2015; Calabrese *et al.*, 2011; Friedrich *et al.*, 2010)

Most of these papers rely on mobile billing data in their work. In such data, each row repsents a call or sms, along with the timestamp and location. The location can either be a cell tower or a more accurate estimated location. For analysising social networks and communities, sometimes the Id of the receiving party is also included. More recently, the surge in mobile internet usage and move towards internet based services such as Whatsapp and Facebook messenger has led to a decline in the use of analog voice-calling and SMS, but a massive increase in the number of data points collected per day by the network operators. Work using high frequency network trace data is much less common. Nachbagauer *et al.* (2012) modelled traffic volume using cellular network data (CNA) rather than CDRs. Also in Austria, Horn *et al.* (2014) improved the approach by developing a method to remove outliers in data. Cik *et al.* (2020) used the same datasets as this paper to develop methods for trip purpose imputation on cellular network data.

2.2 Privacy and Anonyminity

Mobile phone traces are passively collected data on the network operator side, from the interactions between a person's device and the communications network. As such, even more so that other location-based datasets, their use is strictly regulated, in part by the GDPR (de Montjoye *et al.*, 2018). The data can often only be accessed on-site at the network operator, and sharing of the data is often forbidden. Firstly, this goes against the emerging principles of open research, where the data behind the science is made available, and limits the opportunities to reproduce and validate the results of colleagues. Secondly, it restricts the advancement of methods for processing mobile trace data, including the development of methods for both cleaning and verifying the data. While it is understandable that each operator will develop their own proprietary algorithms to process the raw signalling data into location data, accepted methods for evaluating the data for systematic errors are few and far between.

Naturally, CNA datasets cannot simply be made public, and research has shown that even strict anonymisation procedures struggle to remove all identifying features from the dataset (de Montjoye *et al.*, 2013). As such, this paper proposes an approach for generating a synthetic CNA that can be used to support open collaboration and further research in the field when using such data.

2.3 Mobile data for Covid-19 research

More recently, mobile phone data has been an important data source for governments around the world to monitor the both the spread of Covid-19 and the adherence to lockdown measures (Oliver *et al.*, 2020). With concerns about the privacy implications of such a use of mobile phone data, the European Union has released guidelines for the use of mobile phone data during the pandemic, indicating the importance of privacy considerations when working with such data (European Data Protection Board, 2020).

2.4 MATSim and the Switzerland scenario

This paper uses the Agent-based simulation framework MATSim (Horni *et al.*, 2016) to generate a synthetic mobile-phone dataset. Agent based frameworks are uniquely suited to this task, as a population is represented by a set of agents who move about, performing activities at certain locations. The combination of a population, the transport network, and other components such as public transport schedule, is called a scenario. Traditionally, these scenarios were generated from travel diaries and calibrated against traffic counts (Balmer *et al.*, 2008). One particular scenario is the MATSim Switzerland scenario (Bösch *et al.*, 2016; Hörl and Balac, 2020) which we use in this paper.

2.5 Generating MATSim scenarios from CDR Data

CDR data has also been used to create Scenarios for transport modelling MATSim. Anda *et al.* (2018) generated hourly-aggregated OD matrices from mobile phone data and used them to build a large-scale scenario for Singapore, while sidestepping the privacy challenges by using OD Matrices as the input. Yin *et al.* (2017) used machine learning methods to generate synthetic activity chains and integrated them into the MATSim framework to create a scenario for the San Francisco Bay Area. Bassolas *et al.* (2019) used CDRs to generate a scenario for Barcelona, using the trace data itself, rather than aggregated OD Matrices.

2.6 Earlier attempts at synthetic CDR dataset generation

While much work has focused on generating matsim scenarios from mobile network data, to date, Zilske and Nagel (2014) is the only work to investigate building a mobile phone usage model on top of matsim to generate artificial traces. They divide the study area into antenna cells and generate synthetic CDRs for the agents in the simulation based on a call rate. The output of the CDR generation is $p_i, t_i, c_i)$ where p_i is a person identifier, t_i a timestamp, and c_i a cell. However, more recently, the number overlapping cells and the ubiquitousness of mobile internet usage, especially on 4G and soon 5G networks means that there are many overlapping antenna cells, limiting the usefulness of a antenna cell-based approach. Zilske et al. test their approach with a uniform call rate throughout the day, at the rate λ. In a congested MATSim Berlin scenario, they reproduce 95% of the total travelled distance with 50 network trace points per day. However, they assign a network tower to every link, which is unrealistic. In reality, there will be multiple towers overlapping multiple links and significant spatial differences in network coverage. Artifacts in the data such as pinging behaviour between antennae are also not considered.

3 Input datasets

As introduced in Section 1, mobile network operators can geolocate a device on their network using triangulation to the antenna or antennae to which the device is connected. As such, to generate an appropriate synthetic dataset, the location of the cell towers and their approximate strength is needed, as well a model of the trace behaviour - in particular positioning accuracy and signalling rate.

For this paper, due to restrictions on access to an appropriate dataset in Switzerland, data was kindly made available from a major network provider in Austria, accessed only on-site at the Technical University of Graz. Austria has a similar mountainous geography, comparable population size and network of smaller cities with good transport links to Switzerland, making such a dataset appropriate for developing transmission model transferable to Switzerland and other alpen European areas.

3.1 Austrian Datasets

Two Austrian datasets were made available for developing this synthetic network trace approach. In the first, GPS and CNA data were collected simultaneously from 8 devices attached to service personnel working for the telecom provider, who travelled widely around Austria. This included two persons carrying backpacks, 5 with a car, and one on the rail network. These devices collected a total of 903,992 GPS points and 368,209 mobile trace locations, covering 10 overlapping days.

The second dataset consists of a complete month of anonymised mobile network traces for approximately 3 million persons in October 2017. The unique identifiers are randomised each day, meaning that a single person cannot be tracked over multiple days. This dataset does not have accompanying GPS traces, with which to determine the accuracy of the traces.

In both datasets the locations recorded from the mobile network are estimated using the same proprietory algorithm, based on the recorded interactions with the network antennas and the detected transmission delay between the antenna and the device, along with information about the direction and power of the antenna.

3.2 Swiss network trace sample

A small sample of network trace data was provided by a network operator in Switzerland for one unspecified day. It is in similar format to the second larger Austrian dataset.

3.3 Cell Tower dataset

The OpenCellid project is a open data project to collect the locations of cell towers (antenne) all around the world, along with thier estimated range. It is a crowd sourced project, which is mainly used for geolocating devices without using GPS and monitoring mobile provider coverage. The tower locations are estimated based on the reported data from millions of users. The dataset includes records for all towers observed by collaborators of the project over a period spanning the last 14 years. Hence, only LTE and UMTS (3G) towers which were not observed in the last 3 years were excluded. Additionally, towers and have less than 10 observations for LTE and 20 observations for 3G were respectively excluded.

The Federal Office of Communications in Switzerland makes the locations of all mobile phone atennas public avaible as open data for all network providers. Hence, as an official datasource, it will be used instead of the opencellID project as the antenna locations in generating the simulated dataset. As it is extremely rare to work with network traces from multiple network providers simultaneously, we select only those cell tower locations from the largest telecom provider in Switzerland, Swisscom, which has has around 60% market share.

3.4 Terminology

There is a need to clarify the distinction between these types of data. CDRs include additional information on the duration of the call or receiver, but have a lower temporal resolution. CNA datasets, on the other-hand, are records of the individual interactions of the mobile device with the communication network. The spatial resolution between datasets can also vary greatly, depending on whether the cell tower location is used for the position, or if triangulation and sector information is used to improve the accuracy. Especially as network operators move towards 5G networks, mobile data datasets will start to resemble low resolution GPS datasets moreso than CDR data, hence the importance of the distinction between classic CDRs and CNA.

Additionally, out of necessity we refer to two types of events - MATSim events generated by agents in the MATSim scenario, and network events - single records in a CNA dataset representing an interaction between a mobile device and the communication network. To avoid confusion they will be referred to as agent-events and network-events respectively.

4 Methodology

In this section, we present an approach to generate a model of mobile phone data behaviour from real mobile phone data, which can be overlaid on a MATSim simulation to generate a synthetic dataset that replicates the important features of the original data. There are a few key variables that need to be replicated. $\Delta T_{i,t}$, the time delay since the previous event $T_{i,\boldsymbol{x},t-1}$ and the current network-event T_t at time t for a particular person i and location $\boldsymbol{x}$ is the vector representing the location. The location accuracy can be represented by $\hat{\boldsymbol{l}}$, where $\boldsymbol{x}$ is adjusted by an error $\boldsymbol{\epsilon}$ taken from the spatially dependent distribution $f(\boldsymbol{x})$ of the errors in the original data.

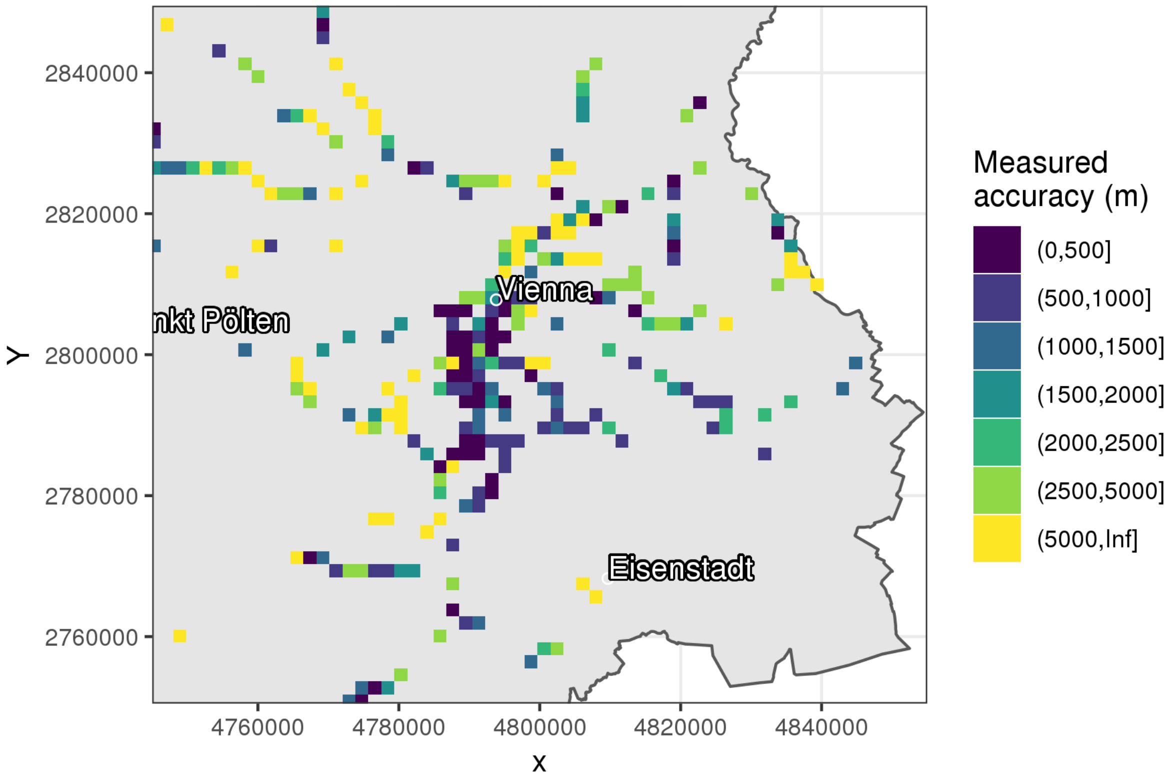
From experience working with CNA datasets from different countries, it is clear that each network operator uses different methods to generate CNA data from the network interactions of the mobile phone users. Rarely is the raw signaling data provided. Instead, the data is pre-processed to either a periodic location estimate, or a trip-stay based structure, similar to that of a travel diary. Hence, in this work, a method for generating raw signaling traces is presented, where the output can be further processed to replicate the expected output of the network provider. In the Austrian case above, this would be a grid-based location estimate on a 50x50m resolution at 15 minute intervals.

4.1 Signally dependant location accuracy

Using the combined GPS-CNA dataset, the location accuracy was determined for a 1kmx1km grid covering Austria, taking the GPS location as the approximate ground-truth. Since the GPS-events and network events were not simultaneous - the true location of a network-event in the CNA dataset is calculated as distance to the temporally closest GPS point from the same device, within a 5 second interval. Values where the GPS or mobile data point occur outside of Austria are excluded. Cells where there were less than 10 accuracy observations were removed.

This gives accuracy measurements for only 2.6% of the Austrian surface area, but 38.1% of the population. As such it was determined that a aggregated raster grid was not required. The average location error for a 1km grid covering Austria is shown in Figure 1. As can be seen in the map, values are generally only available for urban areas and those along main road arterials. A visual inspection of Fig. 2 shows that the distribution of the accuracy measurements are roughly log-normally distributed, with outliers above 10km. The accuracy is also clearly for different urbanity classifications, which we will aim to capture in the model.

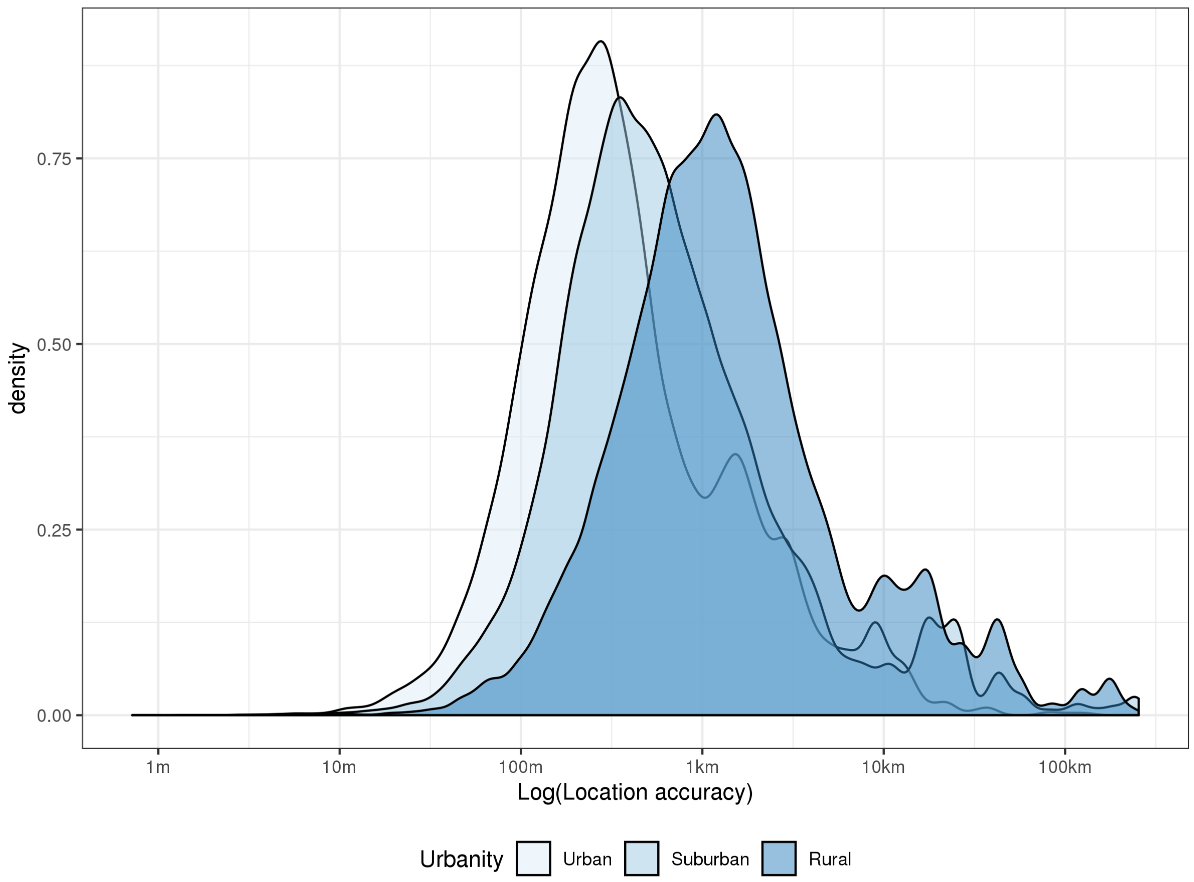
Figure 1: Measured mean location accuracy of the Austrian CNA dataset



4.2 Predictive Model

A predictive model of the mean location accuracy in each cell is constructed as log linear regression. To control for spatial auto-correlation present in the data, simultaneous autoregressive error approach is applied, which assumes that the response variable is a function of both the explanatory variables and the neighbouring locations (Kissling and Carl, 2008; LeSage, 2008), where spatial dependence in the error terms is accounted for. Two model types are used, a spatial error model (SEM), specified as:

Figure 2: Distribution of accuracy measures for all of Austria for the CNA dataset



$$y = X\beta + u, u = \lambda Wu + \epsilon$$

where W is the spatial weights matrix of the neighbours and u are the observed errors. Additionally, we also estimate a spatial Durbin error model (SDEM), which includes exogenous interaction effects as θ in the model. (LeSage, 2008),

$$y = X\beta + WZ\theta + u, u = \lambda Wu + \epsilon$$

In setting up the input data, the value of each cell was weighted by the number of observations, to add more weight to those with more measurements. To prevent negative accuracy measurements, the dependent variable (accuracy in meters) is log transformed.

In model P, only population and the number of cell towers in range was considered. Both log transformed. Interestingly, the coefficient for the tower count is positive. Even though more towers should lead to better triangulation, when more towers are visible to the device, some of these towers are likely to be high-powered long range antennae, which don't aid the triangulation and give a poor location estimation. The model PU includes the European urban codes, which are available on a municipality level,and assigned to the 1km x 1km grid. These are:

- (1) Densely populated area (cities/urban centres/urban areas)
- (2) Intermediate density area (towns, suburbs)
- (3) Thinly-populated area (rural area)

Although these are calculated from the same population density values used for $log(population)$, they improve the model fit in model PU. As expected, the coefficients indicate that the location accuracy improves in the suburbs, and best in urban centers.

In a third model, PUT, the distance to the nearest cell tower (in meters) was included. This was sound to be highly significant, and improved model fit, reducing the AIC by 16 units. The number of towers within range was also tested, and neither it or its log transformation was found not to be significant for any of the above models.

Finally a model with spatial lags on all variables (PUT-SDEM) was estimated, which improved the model fit further. The lag coefficients are significant for $log(population)$ but not for the urban codes. Interestingly, in the SDEM model, the introduction of the lag coefficients reduces the significance of the urban code coefficients. This reduction is stronger for highly urbanised areas, like due to the stronger spatial correlations in population in these ares. The lag coefficient for $dist\_to\_tower$ was also highly significant. The Wald and Likelihood-ratio tests were significant at the p<0.01 level for all models.

A Monte-Carlo test of Moran's I on the residuals gives a p-value of 0.998, indicating that the spatial correlations in the residuals are no longer significant. The prediction accuracy for Austria is presented in Fig. 3. As expected, we see a better accuracy in city centers, and the accuracy values are reasonable, even where the population count provided by the data is 0.

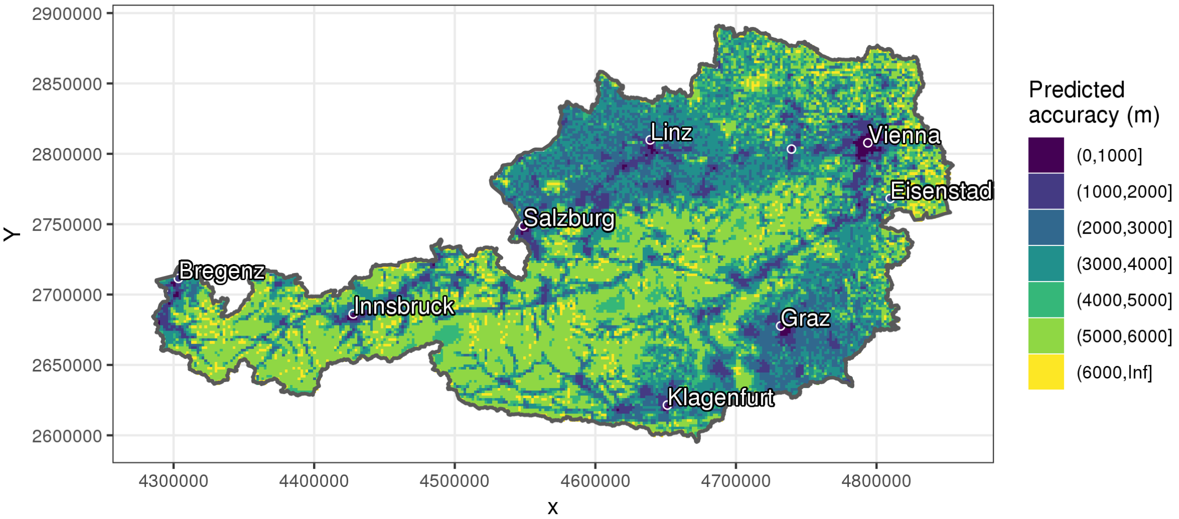
Table 1: Spatial regression results

	Dependent variable:			
	$log(location\_accuracy)$			
	P	PU	PUT	PUT-SDEM
	(1)	(2)	(3)	(4)
$constant$	8.134^{***}	8.297^{***}	8.201^{***}	8.725^{***}
	(0.064)	(0.065)	(0.071)	(0.099)
$log(population)$	-0.120^{***}	-0.093^{***}	-0.088^{***}	-0.091^{***}
	(0.008)	(0.008)	(0.009)	(0.009)
$urban\_code = 1$		-0.896^{***}	-0.861^{***}	-0.525^{***}
		(0.081)	(0.082)	(0.132)
$urban\_code = 2$		-0.421^{***}	-0.398^{***}	-0.228^{***}
		(0.059)	(0.059)	(0.066)
$dist\_to\_tower(km)$			0.205^{***}	0.156^{**}
			(0.060)	(0.061)
$lag(log(population))$				-0.087^{***}
				(0.014)
$lag(urban\_code = 1)$				-0.120
				(0.162)
$lag(urban\_code = 2)$				-0.263^{***}
				(0.090)
$lag(dist\_to\_tower)$				-0.186^{**}
				(0.090)
λ	0.66	0.64	0.64	0.64
Model Type	SEM	SEM	SEM	SDEM
Observations	4,456	4,453	4,453	4,453
Log Likelihood	−7,878.980	−7,811.046	−7,805.295	−7,769.351
σ^2	51.157	50.359	50.248	49.327
Akaike Inf. Crit.	15,765.960	15,634.090	15,624.590	15,560.700
Wald Test (df = 1)	$5{,}048.554^{***}$	$4{,}512.834^{***}$	$4{,}497.601^{***}$	$4{,}597.204^{***}$
LR Test (df = 1)	$2{,}342.888^{***}$	$2{,}250.927^{***}$	$2{,}246.324^{***}$	$2{,}296.528^{***}$
Note:				*p<0.1; **p<0.05; ***p<0.01

4.3 Transferring the spatial model to Switzerland

The urban codes are also available on the municipality level for Switzerland using the same EU methodology. These were assigned in the same way to the 1km grid as was done for the Austrian data. The *distance_to_tower* variable was calculated in kilometers, using the set of UMTS and LTE towers in the Swisscom dataset, described in Section 3.3. The spatial weight matrix was also generated for the 1km grid of Switzerland.

Figure 3: Predicted mean location accuracy for the model PUT-SDEM

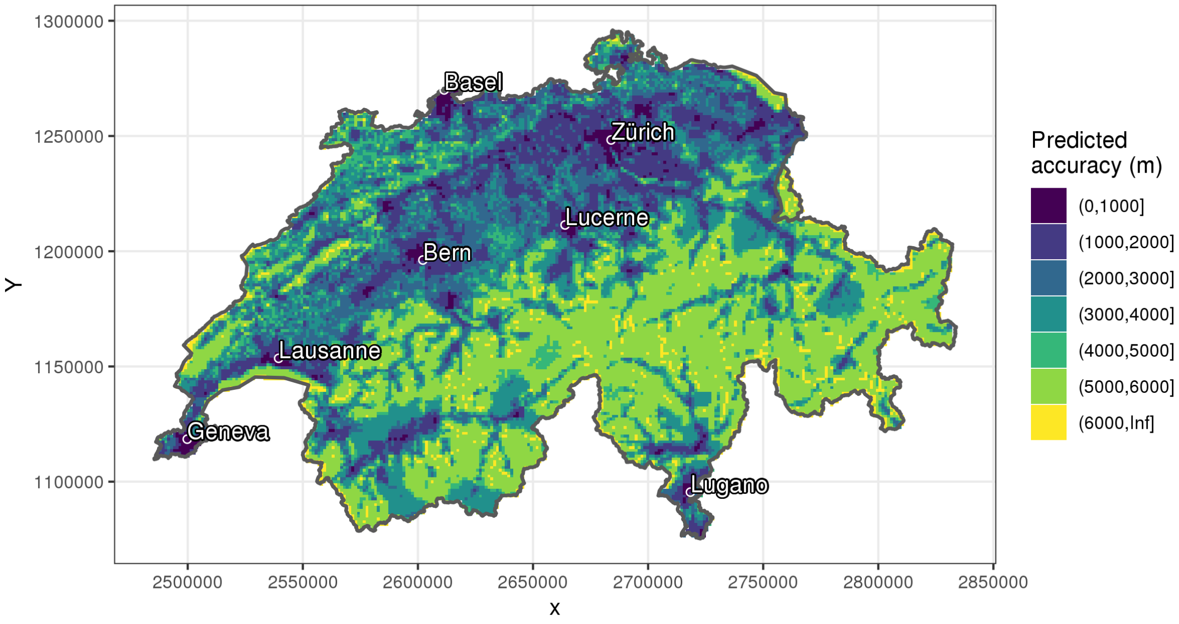


As such, we can reasonably transfer the PUT-SDEM model directly. Fig. 4 shows the results of the predicted mean accuracy for Switzerland using the PUT-SDEM model. As in the Austrian predictive model, we see poorer location-accuracy in the central mountainous axis, and better accuracy in the main cities such as Zurich and Geneva, as one would expect.

4.4 Replicating realistic signaling patterns

Secondly, a model for the event frequency, $\Delta T_{i,t}$, needs to be constructed. Figure **??** shows the average delay between network-events in the second dataset, fitted with a Poisson distribution. The lambda is ???, indicating a mean of ??? events per day. However, both temporal and spatial dependencies can be observed. In Fig. 5, the relationship between $\Delta T_{i,t-1}$ and $\Delta T_{i,t}$ using the Swiss dataset is illustrated. One sees clear clusters here, indicating the existence of multiple stable states, where the network-event frequency stays stable over a period of time. As such, one can infer that there are likely periods where the signaling rate is high for a period of time (i.e. continuous internet usages), and others where it is low, indicating background activity on the mobile phone. It is important to to include these patterns in the model, to replicate the variability in signalling activity, which

Figure 4: Predicted mean location accuracy for the model PUT-SDEM when transferred to Switzerland

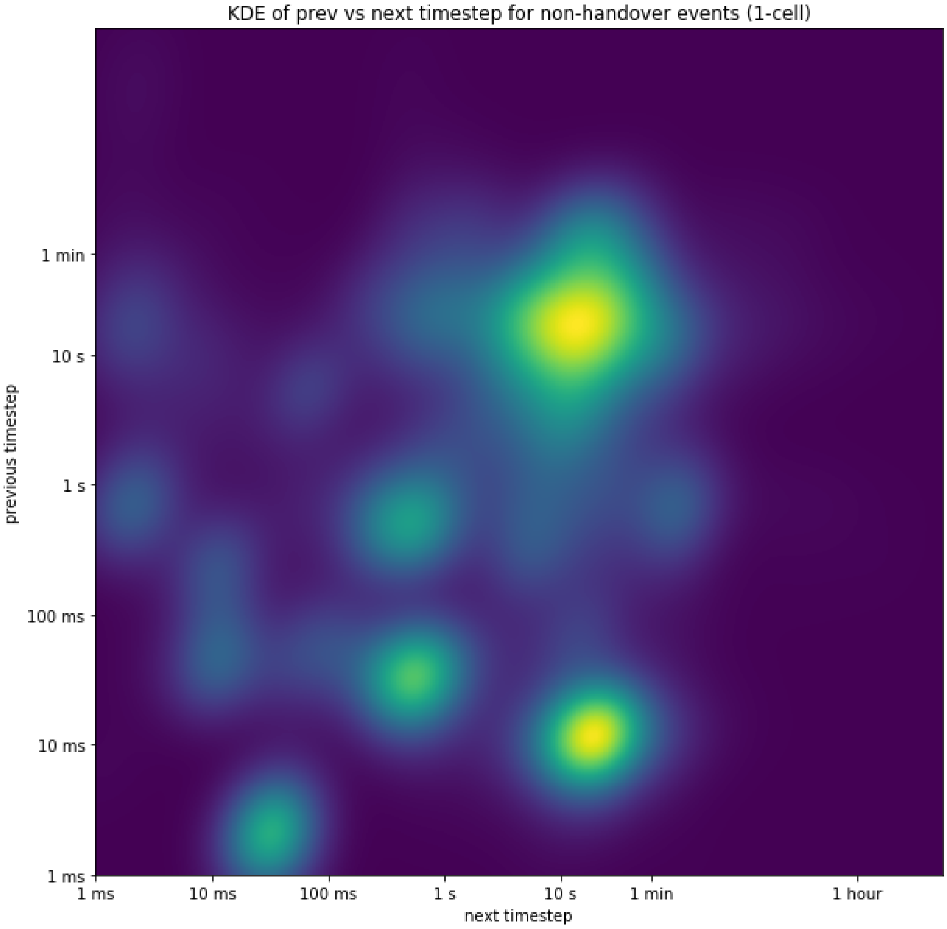


makes the design of robust algorithms (for example those to perform mode or activity detection) challenging.

4.5 Including artifacts

Mobile phone data has been widely recognised as ‘noisy’ data (Zilske and Nagel, 2014; Alexander *et al.*, 2015), Primarily this refers the the propensity for the location to ’jump’ around as the mobile device moves between cell towers. This is particularly prominent when a device switches from a low-powered short-range tower to a stronger long-range one which naturally gives a much poorer estimate of the location of the device. In particular, this leads to a phenomena called ‘pinging’, where a device jumps back and forward between multiple towers in quick succession, giving the impression of movement, even though none has actually occurred. Since we model the accuracy of the location precision using a non-parametric distribution, pings as extreme events are already included as generated network-events where the point accuracy is very poor.

Figure 5: Temporal signalling patterns for Switzerland



5 Proposed Integration into MATSim

Previous work developed routines for MATSim to generate the synthetic CDR data, based on a simple lambda parameter, without spatial or temporal parameters in the model (Zilske and Nagel, 2014). Since the aim here is only to generate a synthetic mobile data dataset, the routines do not need be integrated into the MATSim simulation itself; it is enough to process the agent-based behaviour generated as output from the simulation, recorded as events of the agents.

The module to generate synthetic traces is structured around an implementation of the event listener interface in MATSim. Each unique mobile phone subscriber is independent, in that the density of agents in an area does not affect the resulting mobile locating behaviour. In reality, network operators perform extensive load balancing on their networks by moving connected devices between antennae to optimise call quality, internet speeds and power consumption. However, without knowledge of the proprietary algorithms used to do so, or even the capacity of each antennae, simulating such processes is infeasible.

5.1 Network-event synthesis algorithm

By default, MATSim simulation steps have 1-second frequency with agent locations observed at event occurrences, such as activity start and ends, or link entry and exits. When an agent is stationary, i.e. performing an activity, the location $\boldsymbol{x_{i,k}}$ is known. However, when the agent is traversing a link, the location and timestamp are only known when the agent enters or exits a link. In between these events, the location is imputed from the travel speed (taken as constant) on the link and the timestamp $t_{i,k}$ of the network-event being generated. When an mobile network event $e_{\boldsymbol{x},i,t}$ is generated, the delay $\Delta T_{i,k}$ until the next signaling event is calculated. When $t + \Delta T_{i,k} > t(E)$, where $t(E)$ is the time at the current MATSim-event, $\Delta' T_{i,k}$ is calculated as the remainder $\Delta T_{i,k} - t(E)$ and stored to process the next activity or link traversal.

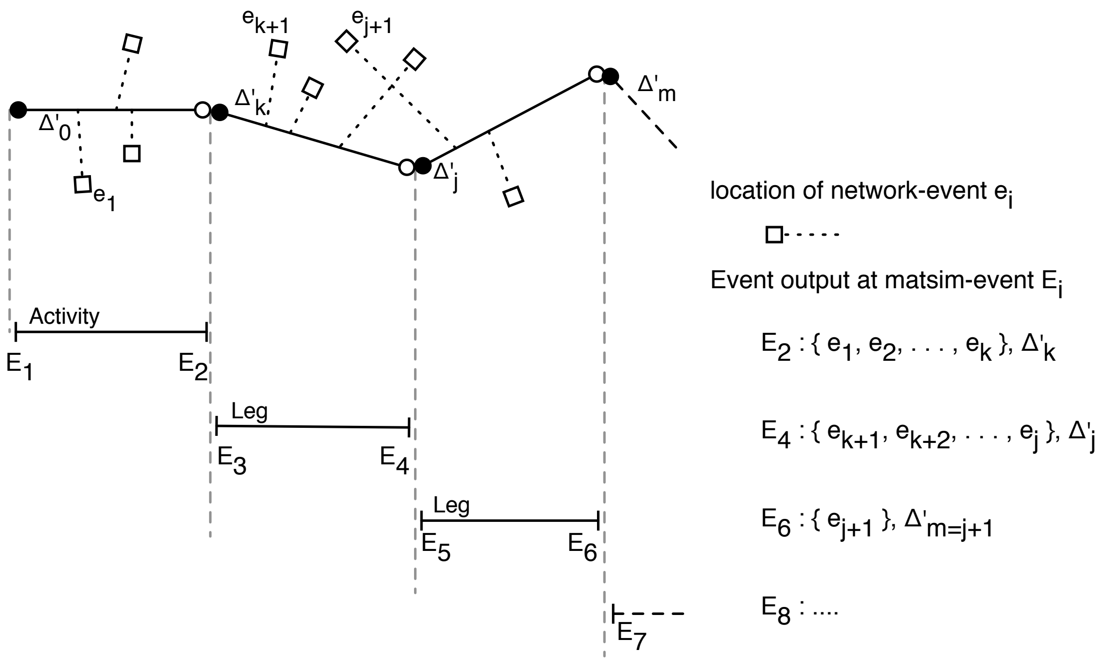
To accommodate the difference in timescales between the MATSim agent events and the simulated network events, the network events are computed iteratively for the time period between the preceding MATSim event $E_{i,k-1}$ and the current one, $E_{i,k}$. This process is illustrated in Figure 6.

MATsim events are processed sequentially, but this processed runs concurrently for each agent in the simulation, with a record being kept of the last event and the current $\Delta' T_{i,k}$. Initially, a random $\Delta' T_{i,0}$ is sampled from the distribution for each participant.

5.2 Other considerations

Each agent in the simulation is assumed to have a single mobile device. To replicate the over-saturation observed in reality, where many users have more than one device, the agents in the scenario would simply have to replicated in the MATSim output.

Figure 6: The network-event generation algorithm. Δ'_k is the remaining duration of ΔT after the MATSim-event E_i


e_{k+1}
e_{j+1}
Δ'_0
e_1
Δ'_k
Δ'_j
Δ'_m
location of network-event e_i
Event output at matsim-event E_i
E_2 : { e_1, e_2, . . . , e_k }, Δ'_k
E_4 : { e_{k+1}, e_{k+2}, . . . , e_j }, Δ'_j
E_6 : { e_{j+1} }, Δ'_{m=j+1}
E_8 : ....
Activity
E_1
E_2
Leg
E_3
E_4
Leg
E_5
E_6
E_7

6 Discussion

There are many factors that influence the localized accuracy of mobile phone positioning, and these would vary between datasets. Different Network operators may use different systems, and focus on improving service in certain areas, decisions which are hard to replicate in such a synthetic dataset without more information from the operator. Furthermore, geographical and urban features are important. The presence of large bodies of water or hills and mountains are also known to affect the location accuracy. In urban areas, skyscrapers and dense urban areas also play a role. Another man-made feature, which is particularly relevant in Switzerland is the loss of reception when travelling through a tunnel. This is not included in the model, but could be added to the MATSim module, deactivating the devices of agents when they enter a tunnel in the network.

While it would be ideal to include many of these features into the model, on a regional scale, when modelling the location tracking accuracy for a whole country it is more important to replicate the general spatial patterns, which we have done, and include some variation in urban areas. We showed that taking a singular accuracy value is not sufficient, as the location accuracy fits a log-normal distribution. Even for the same location, the accuracy can vary between 100m and over 10km, depending on the cell tower to which the device is connected.

A visual inspection of the Swiss predictive model for location accuracy suggests that the model is indeed transferable, and would give usable results. However, ideally, the predictions would be validated against a dataset from the network provider themselves, if one were made available.

Temporally, there are clear second-order states in the timing delta between network events, which are also important to replicate. More work needs to be done to see if these patterns vary among socio-economic groups. This would require a dataset with such information, or at least a model of mobile-phone use for different socio-economic groups. For example, one could expect younger persons may be heavier mobile internet users and therefore generate more network events. Similarly, those who use public transport and can take advantage of their travel time may generate more network events than cyclists or car drivers, who need to be in control of their vehicle.

7 Conclusion

This paper showed that it is possible to generate more realistic synthetic datasets of CNA data using both unanonymised and anonymised data, which can be transferred from one study area to another. The work identified key spatial and temporal patterns that should be included in such models. Further work would use such models to develop and test the methods for working with mobile phone data, where the both the methods and data used can be shared publicly, and investigate how methods calibrated on such a synthetic dataset perform on real data from the same study area.

8 References

Alexander, L., S. Jiang, M. Murga and M. C. González (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data, *Transportation Research Part C: Emerging Technologies*, **58**, 240–250.

Anda, C., S. A. O. Medina and P. Fourie (2018) Multi-agent urban transport simulations using od matrices from mobile phone data, *Procedia computer science*, **130**, 803–809.

Balmer, M., K. Meister, M. Rieser, K. Nagel and K. W. Axhausen (2008) Agent-based simulation of travel demand: Structure and computational performance of matsim-t, *Arbeitsberichte Verkehrs-und Raumplanung*, **504**.

Bassolas, A., J. J. Ramasco, R. Herranz and O. G. Cantú-Ros (2019) Mobile phone records to feed activity-based travel demand models: Matsim for studying a cordon toll policy in barcelona, *Transportation Research Part A: Policy and Practice*, **121**, 56–74.

Blondel, V. D., A. Decuyper and G. Krings (2015) A survey of results on mobile phone datasets analysis, *EPJ data science*, **4** (1) 10.

Bösch, P. M., K. Müller and F. Ciari (2016) The ivt 2015 baseline scenario, paper presented at the *16th Swiss Transport Research Conference (STRC 2016)*.

Calabrese, F., G. Di Lorenzo, L. Liu and C. Ratti (2011) Estimating origin-destination flows using mobile phone location data, *IEEE Pervasive Computing*, (4) 36–44.

Cik, M., A. Lechner, C. Hebenstreit and M. Fellendorf (2020) Activity estimation from mobile phone data, paper presented at the *Transportation Research Board Annual Meeting*.

de Montjoye, Y.-A., S. Gambs, V. Blondel, G. Canright, N. De Cordes, S. Deletaille, K. Engø-Monsen, M. Garcia-Herranz, J. Kendall, C. Kerry *et al.* (2018) On the privacy-conscientious use of mobile phone data, *Scientific data*, **5** (1) 1–6.

de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen and V. D. Blondel (2013) Unique in the crowd: The privacy bounds of human mobility, *Scientific reports*, **3**, 1376.

European Data Protection Board (2020) Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, *Guidelines*, May 2020.

Friedrich, M., K. Immisch, P. Jehlicka, T. Otterstätter and J. Schlaich (2010) Generating origin–destination matrices from mobile phone trajectories, *Transportation research record*, **2196** (1) 93–101.

Gonzalez, M. C., C. A. Hidalgo and A.-L. Barabasi (2008) Understanding individual human mobility patterns, *nature*, **453** (7196) 779–782.

Horn, C., S. Klampfl, M. Cik and T. Reiter (2014) Detecting outliers in cell phone data: correcting trajectories to improve traffic modeling, *Transportation research record*, **2405** (1) 49–56.

Horni, A., K. Nagel and K. W. Axhausen (2016) *The multi-agent transport simulation MATSim*, Ubiquity Press, London.

Hörl, S. and M. Balac (2020) Reproducible scenarios for agent-based transport simulation: A case study for paris and Île-de-france, 05 2020.

Janzen, M., M. Vanhoof, Z. Smoreda and K. W. Axhausen (2018) Closer to the total? long-distance travel of french mobile phone users, *Travel Behaviour and Society*, **11**, 31–42.

Kissling, W. D. and G. Carl (2008) Spatial autocorrelation and the selection of simultaneous autoregressive models, *Global Ecology and Biogeography*, **17** (1) 59–71.

LeSage, J. P. (2008) An introduction to spatial econometrics, *Revue d'économie industrielle*, (123) 19–44.

Nachbagauer, G., P. Schosteritsch, T. Reiter, R. Scherer, M. Cik and M. Fellendorf (2012) Traffic analysis using cellular network data, paper presented at the *19th ITS World Congress*.

Oliver, N., B. Lepri, H. Sterly, R. Lambiotte, S. Delataille, M. D. Nadai, E. Letouzé, A. A. Salah, R. Benjamins, C. Cattuto, V. Colizza, N. de Cordes, S. P. Fraiberger, T. Koebe, S. Lehmann, J. Murillo, A. Pentland, P. N. Pham, F. Pivetta, J. Saramäki, S. V. Scarpino, M. Tizzoni, S. Verhulst2 and P. Vinck (2020) Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, *Scientific Advances*.

Yin, M., M. Sheehan, S. Feygin, J.-F. Paiement and A. Pozdnoukhov (2017) A generative model of urban activities from cellular data, *IEEE Transactions on Intelligent Transportation Systems*, **19** (6) 1682–1696.

Zilske, M. and K. Nagel (2014) Studying the accuracy of demand generation from mobile phone trajectories with synthetic data, *Procedia Computer Science*, **32**, 802–807.