Lay Summary

Using data traces to improve transport systems

Project team
Prof. Kay W. Axhausen, Prof. Andreas Krause, Prof. Martin Fellendorf, Dr. Ing. Dominik Ziemke, Joseph Molloy, Anastasiia Makarova, Billy Charlton

Contact address
Prof. Kay W. Axhausen
Institut für Verkehrsplanung und Transportsysteme
ETH Zürich
Gebäude HIL / F 31.3
Stefano-Franscini-Platz 5
8093 Zürich
axhausen@ethz.ch

12.04.2021

Swiss National Science Foundation
Big Data
National Research Programme

1. Background

The continuous tracing of travellers via their wireless interactions i.e. bluetooth, wifi, GPS and mobile phone networks (GSM, 3G, 4G) systems has opened a new world to transport system planning and management, in which data is not scarce anymore. The data is not only not scarce anymore, but is generated by the vast majority of the population through the near universal ownership of mobile phones and in particular smartphones with their GSM, wifi and Bluetooth antennas.

The usual time frame for the implementation of transport models by local, regional and national authorities is between 5 to 10 years, maybe with some intermediate updating. While this was already slow in the past, in today's phase of dynamic change it is threatening the credibility of the planning and later actions of the authorities. The need for long-term planning is even more obvious given that new services, such as Uber, bike-sharing, carsharing etc. and new technologies, such as autonomous vehicles, fundamentally challenge the regulatory framework, change the cost structures and capacities of the transport system. In such an environment any long term investment in rolling stock and infrastructure needs to be carefully analysed.

Specifically the goal was to link GSM based traces to a state-of-the-art agent-based modelling system, MATSim. The project was able to combine the strengths of each institute within the theoretical frameworks of their particular research areas. The project was made up of 9 various work packages, which can be succinctly grouped here into four strands:

a. Long distance travel using aggregate GSM data
b. The behavioural analysis on GPS and GSM data using individual recorded traces of movement
c. Third, contributions to the process of building transport models from such data
d. Methods for the calibration of these transport models

Agent-based transport simulation models are a particularly useful tool to analyze demand-oriented transport policies, pricing, and new mobility services. Since travel diaries (a traditional source to create the transport demand in agent-based transport models) are often hard to procure and not policy-sensitive, alternative approaches to creating travel demand representations for simulation scenarios are required. In this study, a particularly efficient approach is established based on mobile phone records and a new aspatial activity-based demand model with comparatively low input data requirements. Home, work, and education locations are generated based on mobile-phone-based origin-destination matrices. Other activity locations and suitable travel options are modeled within the scope of the agent-based transport model. As a result, a novel and comparatively lightweight process to create an agent-based transport simulation scenario was to be developed.

2. Goals of the project

As a large collaboration between 4 research groups, this project's main aim was to make contributions to the use of Big Data in transport modelling, with the application to the study of mobility pricing. In particular the focus was to be on the analysis on mobile data traces from a large Swiss mobile network operator. Four datasets were to be provided, at various levels of aggregation and size, which would support the development of methods for working with mobile trace data in transportation research.

In strand A, the aim was to generate origin-destination matrices of long-distance travel from GSM data, and compare them to official Swiss datasets, and investigate how they may differ. This is an important first step in application of GSM data to transport modelling in the long-distance travel context, where available data is generally scarce.

Originally, for the work in strand B the network operator was to provide secure access to datasets of anonymised mobile phone traces. However, arranging secure access in a way suitable to the network operator turned out to be especially challenging. Futhermore, just as the work with the mobile network traces got underway, the remote working requirements necessitated by the Covid-19 pandemic meant that the work had to be paused indefinitely. However, the work could proceed with GPS traces collected from willing participants in the context of another study.

In strand C, the goal was to eschew expensive and difficult to use travel survey datasets in the generation of transport modelling simulations, and instead explore the use of mobile phone data as a partial substitute.

In strand D, methods for the calibration of these transport models were explored. The goal was to apply a method called Multi-fidelity Bayesian Optimisation (MF-BO) to calibration process, by treating it as an optimisation problem.

3. Methods

Strand A – Long Distance

To analyse long distance behaviour, an anonymised and aggregated dataset was generated by the network operator for trips over 50km in length (a common definition of long distance). This dataset was provided in the form of an Origin-Destination Matrix (OD-Matrix) with each cell in the matrix representing the number of long-distance trips per month over a year between each pair of gemeinde in Switzerland. Due to privacy regulations mobile phone data has to be aggregated. The aggregation level used was the month and municipality, i.e. for each municipality and each month the number of trips per capita is calculated. We chose in our sample 36 municipalities of different population size. Within these municipalities 500 mobile customers were selected and their long-distance trips were calculated for each month. A mobile person was defined as a person that performs long-distance trips. In addition, the share of customers not travelling long-distances was known

for each month, which allowed us to calculate the number of long-distance trips per capita on a monthly basis.

Strand B – Individual behaviour

In this strand, choice models were to be used to analyse the important factors of travel behaviour. Choice models are form of regression analysis that allow the analyse of the choice of one alternative from a selection. The main model form used is called “multinomial logit regression”. However, estimating such models on large datasets is extremely computationally expensive using the currently available methods. As such, the software package ‘MIXL’ was created as part of this project, with provides dramatic performance improvements when estimating such models in the statistical software R. These performance advantages are obtained by utilising both parallel commuting over multiple processing cores, and the cross-compilation of the model code to the programming language C++.

Strand C – Model generation from GSM data

Our approach for creating a policy-sensitive transport simulation scenario based on mobile phone records instead of travel surveys consists of four steps:

1. Acquisition of an appropriate population dataset depicting the residents of the region (Zürich, Switzerland).
2. Using cell-phone-based origin-destination matrices to derive work and education locations for all members of the population who are workers or take part in the education system.
3. Applying a new, aspatial activity-scheduling model with comparatively frugal input data requirements.
4. Taking advantage of MATSim agent-based transport simulation software for the selection of locations for discretionary activities and for simulating the travel between all activities. In particular, the method in MATSim for performing destination choice was modernized and applied.

This approach is put into practice as a prototype for the metropolitan area of Zurich in Switzerland. This region is particularly suitable because multiple MATSim scenarios have been created for this region with a high standard of scenario validation, such that scenario-to-scenario comparisons can be carried out. The cell-phone network operator Swisscom provided mobile-phone-based origin-destination matrices for Switzerland within the context of this project.

The STATPOP population dataset was readily available for all of Switzerland, making this a good test of the new method.

Strand D – Bayesian calibration of transport models

We cast the calibration problem as an optimization problem, where we minimize the mismatch of aggregated measures of traffic events, such as travel time distribution, between the output of a transport system simulation and reference observations. This optimization problem can not be shown to be convex and does not have an analytical form. Furthermore, the evaluation of its objective function is computationally expensive.

Bayesian Optimization is a black-box, sample efficient, global optimization algorithm that is well suited for these kind of problems. We use different levels of approximation of the objective function, which can be obtained from transport system simulations to speed up the optimization. We demonstrate the effectiveness of MF-BO for the calibration of an agent-based transport simulation.

4. Results

Strand A – Long Distance

This paper has shown that mobile phone data is a useful data source for the field of travel demand research. Especially, long-distance travel estimators can benefit from this particular data. However, it was not possible to show in this paper that a household travel survey per se under-reports long-distance travel demand. A survey with a detailed questionnaire and a large sample size like the Microzensus can lead to similar results as mobile phone data.

Strand B – Individual behaviour

The software package mixl for choice modelling on large datasets is now public available on the R software archive, CRAN. The work using to analyse individual behaviour is still ongoing.

Strand C – Model generation from GSM data

As a reference, an existing MATSim simulation scenario for the Zurich metropolitan area developed at the Institute for Transport Planning and Systems at ETH Zurich is used.

Based on this well-validated reference, overall comparisons show that average duration of activities deviate less than 10% for all activity types except education activities. Locations of trip origins throughout the day also look very similar. This confirms that the mobile-phone-based approach is selecting realistic work and education facilities. The locations for discretionary activities seem to be too close together, perhaps because the method for choosing candidate facilities was too simplistic. This can be confirmed and improved in later studies.

The goal of creating a region-wide model which is sensitive to policy-level analysis including pricing and road/transit modifications using data from easily-obtainable sources has been achieved.

Strand D – Bayesian calibration of transport models

We cast the calibration problem as an optimization problem, where we minimize the mismatch of aggregate mea- sures of traffic events, such as travel time distribution, between the output of MATsim and real world observations. We show that Bayesian Optimization is well suited for problem. Moreover, it can be adapted to parallel evaluations and it naturally accommodates different levels of approximation of the objective function, which can easily be obtained from MATsim and can speed up the optimization. We demonstrate the effectiveness of MF-BO for the calibration of MAStim on both small (Sioux Falls) and large scale (Zurich) scenarios.

5. Significance of the results for science and practice

- For analytical models that can be specified at various fidelities, the MF-BO approach developed in this project is a useful technique for calibrating the model for a set of parameters. Transport models are a good example of this, where large models are scaled down and run on a sub-sample of the population to reduce runtime. This is especially useful where the model needs to be considered as a 'black box', that is, with no knowledge of the internal workings.

- In this study, a new scenario creation process was shown which intends to make the creation of policy-sensitive agent-based transport simulation scenarios simpler. Only highly standardized input data sources are used and those very sparsely. The created procedure makes use of mobile-phone-sources data which can be assumed to exist in almost any location worldwide. These data matrices have been used to derive work location, but could also be used to derive other types of locations. For activity scheduling, the new, aspatial activity-based demand model actiTopp was used, which – based on its input specifications and modeling scopes – complements MATSim's modeling capabilities very well and helps maintain a toolchain with low data requirements. Given the low data requirements and the initial success in this test of the new method, this methodology likely could be transferred to many other regions, including regions without access to expensive household survey datasets.

- The work on activity purpose identification, synthetic dataset generation and mode detection demonstrates that individual mobility patterns can be extracted from GSM traces. Until now, much of the work with GSM focused on the analysis of aggregated datasets. If the challenges around data access and industry-academic collaboration can be addressed, such anonymised trace datasets could be used to great effect in transportation modelling, as a much larger proportion of the population can be surveyed over a longer period of time, without the response burden of active data collection methods.

- With the shift towards micro-mobility, automated vehicles, and hence a more varied transport landscape, better transport models are more important than ever for understanding how new technological developments driven by digitalisation will affect our lives in the future. The progress made in this project on automated scenario generation and tools for behavioural analysis have contributed in the progress towards this goal.