TRANSPORT MODELS AND BIG DATA FUSION: LESSONS FROM EXPERIENCE

Miguel Picornell
Luis Willumsen
Kineo Mobility Analytics

1 INTRODUCTION

Transport planners and modellers have managed to work with models based on increasingly better theory supported by rather poor and noisy data. We have been using fairly small sample observations at Household Travel Surveys (1%-3% samples), collected for a day or two every 5-10 years, admittedly larger samples at intercept surveys (roadside and on board interviews) and sometimes a good number of traffic counts and travel time measurements. We have blended data from different days (and sometimes even different years) to obtain trip matrices for a mythical “average day and time period” and aspire that these matrices will allocate tips to modes and routes that we can validate using similar observations.

This situation is changing just at the time when the profession is facing probably the greatest challenge in estimating impacts and delivering advice for a future with significant disruptive technologies and behavioural shifts.

Recent years have seen an increase in the number and variety of sensors available to detect the movement of people and vehicles: CCTV, WiFi networks, Bluetooth sensors, smart cards, mobile phones and other devices. All of them leave an electronic trace that may be used to characterise levels of activity in space and time. For example, some of these data sources have been used to characterise urban mobility with “heat maps” displaying densities of use in time and space.

To be useful, these electronic traces have to be stamped with their (geo) location and time. A collection of these data points can then be put to a number of different uses. The one of interest here is applications in transport modelling.

This paper is organised as follows. The next section outlines the different sensors and data sources and the challenges and opportunities that they can offer. Section 3 presents in more detail the specific case of mobile phone data while section 4 discusses a particular application of anonymised mobile phone data to support a transport model to study a toll road; finally, section 5 attempts to summarise the lessons learnt from this and other studies attempting to apply this data to real problems.

2 NEW DATA SOURCES: OPPORTUNITIES AND LIMITATIONS

2.1 Active and passive data collection

One of the advantages of the new data sources, compared to conventional surveys, is that the data is collected passively, that is there is no intervention on the part of the subject. This is important because it is increasingly difficult to persuade people to answer “yet another questionnaire” and they sometimes tend to simplify responses to reduce the time spent on a survey. Another advantage of passive data collection is that the data points can be continuously saved and retained, so that data is permanently up to date. The counterpart of this advantage is that we cannot, in general, ask individuals for their perceptions, attitudes or what they would have done under different conditions; we can only infer what they have actually done.

In the majority of cases these new data sources result from other specific objectives, for example, mobile phone records have traditionally been collected for billing purposes, and nowadays, this information is also used to estimate mobility patterns. This would imply that the cost of collecting these data should be low. Nevertheless, the task of making this data useful is seldom cost-free.

2.2 New data sources

The new data sources offer advantages but have limitations as well:

- Bluetooth traces. These are very similar to using video for automatic number plate recognition (ANPR) and matching. The general availability of low cost Bluetooth sensors makes it possible to identify the unique Media Access Control (MAC) address of any device with Bluetooth on. The appropriate location of these sensors, for example at the entry and exit slip roads of a motorway should enable to match MACs and create local trip matrices and estimate trip duration between devices. The sample size depends on the proportion of devices with Bluetooth activated. The resulting data is (as with ANPR) not error free and the resulting matrices are inevitably local rather than from true Origin to true Destination.

- Smartcards, used to pay for public transport and other services, combined with GPS devices on buses to provide time and location of the transaction, can be used to obtain trip matrices from stop-to-stop or station. These are better when the user validates both on entry and exit (as in the London tube or Singapore buses). When the smart card is validated only on entry the exit point may be estimated by the next transaction, see for example Munizaga and Palma (2012) and Wang et al (2011). Inevitably, this data source cannot say much about the use of other modes or the true origin/destination of each trip.

- Mobile phone data has been used to estimate more comprehensive origin destination matrices and other indicators of mobility. For example, they have been used to obtain urban patterns (Louail et al. 2014), population mobility (Picornell et al. 2015, Cáceres et al. 2012, Calabrese et al. 2011, González et al. 2008, Lenormand et al. 2014) and to study the relationship between social networks and travel behaviour (Picornell, et al 2015). This type of data has the advantage of providing better information on origins and destinations; moreover, longitudinal analysis permits identifying some journey purposes or the frequency and regularity of trips, something not available with conventional methods.

A reasonable conclusion from these considerations is that no data source can provide the comprehensive and rich information required to develop good transport models. It is necessary to combine them with other data sources for example traffic counts and Stated Preference. It is important here to distinguish between “data fusion” and “data combination”. Data fusion refers to the treatment of different data sources in their original, unprocessed form. Data combination refers to the use of different sources in their already processed form; this is probably less demanding but some loss of information is to be expected as the processing may not be the most appropriate to the application in mind.

The table overleaf shows a comparison of different data collection methods and data sources based on our own experience.

These data sources can certainly be used to complement other surveys and in many cases replace conventional ones. We believe there will always be reasons to undertake at least some Household Travel Surveys as the information they provide is very rich. However, if they are not used to develop trip matrices the sample sizes may be smaller, for example 3,000 households.

Nevertheless, these new data sources will offer no information on what would happen if conditions change, new policies or new projects are considered. In order to provide support for decision making we still need models.

Method	**Approximate typical costs**	**Time to results**	**Coverage and precision**	**Level of difficult for a practical application**
Household Travel Surveys	1 to 4 million euros €	1 to 2 years	The whole city, samples 1%-3%, possible non-response/contact bias, incomplete answers	Difficult, costly, quality assurance essential
Intercept Surveys (RSI/on-board)	5,000 € per location, plus delays	1 to 2 months	Samples 10% to 50% on a day. Accuracy depends on coverage (leaks) and responses	Increasingly difficult and expensive
Bluetooth matching	~20,000 € per km (approx.).	Weeks	Accuracy depends on coverage, proportion of Bluetooth devices on; only local ODs	Medium difficulty; requires installation of sensors and their maintenance
Mobile phone traces	25,000 – 50,000 € + depending on the analysis	Days / weeks	Large sample. Accuracy depends on cell density and frequency of interaction with them.	Easy in principle, medium difficulty in the adaptation to practical use.
Smart Card for Public Transport	50,000 €+ depending on the analysis	Days/ weeks	Sample depends on penetration (>90% London and Santiago de Chile) and the accuracy of the equipment.	Easy in principle, medium difficulty in the adaptation to practical use.

3 MOBILE PHONE DATA FOR TRANSPORT MODELS

3.1 Anonymised mobile phone data

The type of data provided by mobile phone operators originates on a contact, an exchange of information, between a device with a SIM (no only mobile phones) and an antenna or cell; it can be geo-located and stamped with the time of the interaction. The most common record of these interactions is a Call Detail Record (CDR); these are generated when a call is initiated and finished, when text messages are sent, any transfer of data and when the device changes zone of coverage.

There are at least two dimensions to the accuracy of mobile phone data: geographic and temporal resolution. The location of the antenna, and in good cases the coverage area of the cells, provides a proxy for the location of the device. Cells are quite small in urban areas but much larger in rural areas, depending on the network capacity needs. The frequency of interactions provides the temporal resolution; this has increased significantly with the use of data under 3G and 4G coverage. Mobile phone operators have started using probes that create a more frequent contact with the device to improve operations; they generate a better temporal resolution and improve performance of this type of data.

Privacy must be protected at all times and therefore it is essential to anonymised all sensible data and to provide only aggregate information that cannot be used to identify a user.

3.2 Advantages of mobile phone data

These have been known for some time:

- Relatively low acquisition costs
- Large sample size, often available for several days
- Passive data collection
- Up to date information
- Opportunity to explore how travellers changed their behaviour when an unplanned disruption takes place: flooding, temporary closure of a service, etc.

On the other hand, using this type of date presents several challenges:

- Data processing and storage: it is necessary to store and process millions of registers per day that cannot be analysed using conventional database tools.
- Data cleaning is essential: many registers contain errors that need to be detected and cleaned.

- Data analysis: the literature presents several different algorithms to process and analyses mobile phone data; some produce very different results. Therefore, data validation is a critical component of any processing.

- Adaptation to the problem in hand: mobile phone data exist to be able to charge customers and improve operations; it has not been designed to obtain origin destination matrices. We have found it essential to understand the application well before processing the raw data as the only way to deliver usable results.

- Data fusion: this is required to enrich the data with attributes of interest to the application, for example socio-economic group based on census data.

- Integration with transport models: if the data is going to be used in a model it is best if the processing and conversion is undertaken with full understanding of the original nature of the data, the best processing method and how the model will be calibrated and used.

To address these challenges it is better to incorporate transport modelling and planning knowledge throughout the process.

4 A CASE STUDY: DEMAND FOR A TOLL ROAD

4.1 The problem

A toll road concessionaire in Spain wanted to study whether a different pricing schedule would be able to attract new customers, in particular during the off-peak seasons or time periods. This required undertaking conventional RSIs outside the toll road, something that was difficult and expensive to do. They approached Kineo to obtain estimates of trip matrices, at different times of the day and eventually other information, from anonymised mobile phone data. This would be used on a multi-class demand model to perform sensitivity tests before implementing a new pricing scheme.

4.2 Data sources used

The following data sources were used in this project:

a) Mobile phone data. Anonymised CDRs for calls, text and data from Orange for the whole of Spain and during more than 40 days.

b) Network data: obtained from different sources; it was important to make this consistent in both the model and the processing of CDRs.

c) Traffic counts, vehicle occupancy: classified counts were available from the toll plazas and from other permanent or recurrent sites from the Spanish Government.

d) Socio-demographic data: census data from the Instituto Nacional de Estadística

4.3 Matrix estimation

We used algorithms developed and validated by Kineo from earlier projects to provide an initial estimate of the required trip matrices. There was a process of iterative and incremental improvement of matrix validation. The first set of matrices was transferred to the traffic model (in Vissum) with some 200 zones and major road links. It was soon apparent that there was a mismatch between how mobile phone movements were recorded and how the model was allocating trips to zones and networks. Indeed, the mobile phone movements were originally coded to some 15,000 "zones" and even in the immediate area of the toll road there were more than double the zones than in the model. Also, there were trips recorded on mobile phone that used secondary roads not in the model; when loaded onto the model produced strange, unlikely loads as they were forced onto the main links.

These problems would not have occurred with conventional data collection as these trips would not have been intercepted on the main road network. This shows the importance of understanding the different nature of the data and how exactly it is going to be used in a model in practice. The model was improved in its granularity to solve the aforementioned problems.

In practice, the number of vehicles in the trip matrices depends not only on market shares and levels of usage but also on vehicle occupancy and the number of SIM active devices. Adjustments had to be made for all of them.

4.4 Matrix validation

This involved looking more closely at three screenlines that could be formed using the traffic counts available. Given that the number of vehicles on the screenlines depended partially on the routes assumed and chosen and this depends on, for example, willingness to pay tolls (values of time),we produced upper and lower estimates of flows on these relatively short screenlines:

Figure 1 shows, for each of the three screenlines and for each of the hours of the day the observed screenline flow (solid) and the upper and lower limits. It can be seen that in the large majority of cases the actual flows are within these limits. With additional adjustments by Kineo the discrepancies between screenline observations and estimates were reduced even further.

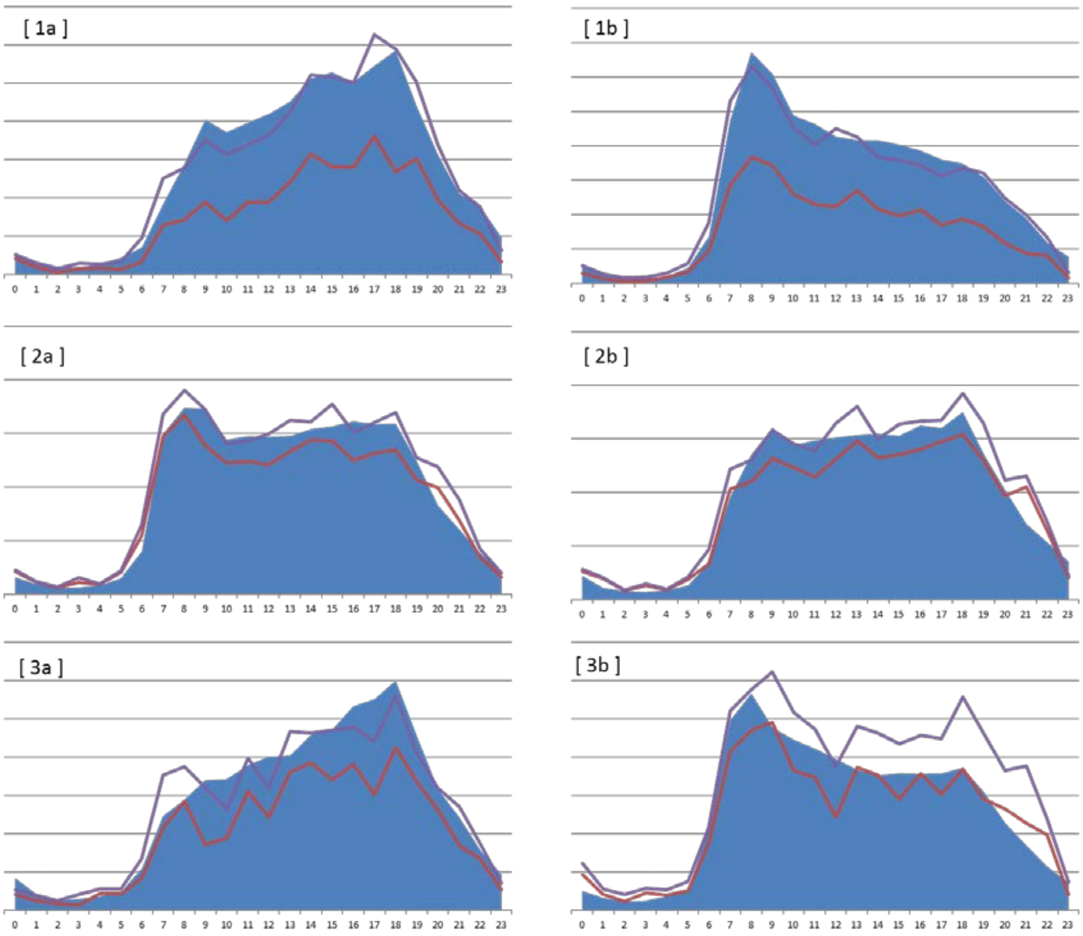

Figure 1 Comparison of observed screenline flows (solid blue) against upper and lower limits estimated from mobile phone data.

4.5 Matrix adjustments in the model

Once we obtained a good match between mobile phone data and model we used the model to refine the trip matrices further. With the data available it was not possible to reliably differentiate between cars, buses and trucks; it was also risky to attempt to estimate values of time, an essential feature of a toll road model. The disaggregate by mode was achieved in the model itself using matrix estimation techniques starting from an initial split obtained from the toll plaza counts. This adjustment also served to make matrices and model more compatible.

We also attempted to use mobile phone data to identify with some precision the route taken by each vehicle. This was necessary in order to advance the task of estimating a distribution of willingness to pay in the area of influence of the toll road. In some locations the cells were sufficiently apart to uniquely identify the route taken. In other parts of the network this was not the case as tolled and untolled roads were very close. In the end, we managed to achieve a clear identification of routes for about 75% of the cases and this information is now being used to estimate values of time.

It must be recognised that in dense urban areas the identification of unique routes will be much more difficult.

5 CONCLUSIONS

The main lessons learnt from this work can be summarised as follows:

- Mobile phone travel data and that obtained from conventional surveys are different in nature and this difference must be allowed for.
- The appropriate fusion of anonymised mobile phone data with other sources permits the estimation of fairly reliable trip matrices and other data for transport models.
- To achieve this consistency it is important to process the data in the full understanding of how it will be used in the transport model.
- Close collaboration between mobile phone operators, data scientists and transport modellers is important to achieve satisfactory results.
- The trip matrices obtained from mobile phone data still require adjustments when adopted into a conventional model; if the processing has considered the points above, these adjustments are smaller than when using conventional data sources.
- It is possible to use information from mobile phone data to improve the estimation of other model parameters.

BIBLIOGRAPHY

Cáceres, N., Romero, L. M., Benítez, F. G. & Castillo, J. M. D. (2012). "Traffic flow estimation models using cellular phone data". *IEEE Transactions on Intelligent Transportation Systems*, 13, 3, 2012, 1430-1441.

Calabrese, F., Lorenzo, G.D., Liu, L. & Ratti, C. (2011). “Estimating origin-destination flows using mobile phone location data”. *Pervasive Computing, IEEE*, 10, 4, 2011, 36– 44.

González, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). “Understanding individual human mobility patterns”. *Nature,* 453(7196), 779-782.

Lenormand M, Picornell M, Cantú-Ros OG, Tugores A, Louail T, Herranz R, Barthelemy, M., Frías-Martínez, & E., Ramasco, J. (2014) “Cross-checking different sources of mobility information”. PLoS ONE 9, e105184

Louail T, Lenormand M, García Cantú OG, Picornell M, Herranz R, Frías-Martínez, E., Ramaswco, J. & Barthelemy, M. (2014) “From mobile phone data to the spatial structure of cities”. *Scientific Reports 4*, 5276

Munizaga, M., Palma, C. (2012) Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C* 24 (2012) 9

Picornell M, Ruiz T, Lenormand M, Ramasco J, Dubernet T & Frías-Martínez E (2015). “Exploring the potential of phone call to characterize the relationship between social network and travel behavior”. *Transportation,* 42, pp. 647-668.

Wang, W., Attanucci, J., & Wilson, N.H. (2011). “Study of Bus Passenger Origin Destination and Travel Behavior Using Automated Data Collection Systems in London”. *90th TRB Annual Meeting*, Washington, D.C.

.