Willumsen, Luis G.

Working Paper

Use of big data in transport modelling

International Transport Forum Discussion Paper, No. 2021/05

Provided in Cooperation with:
International Transport Forum (ITF), OECD

Suggested Citation: Willumsen, Luis G. (2021) : Use of big data in transport modelling, International Transport Forum Discussion Paper, No. 2021/05, Organisation for Economic Co-operation and Development (OECD), International Transport Forum, Paris, https://doi.org/10.1787/86a128c7-en

This Version is available at:
https://hdl.handle.net/10419/245859

International Transport Forum

Use of Big Data in Transport Modelling

Discussion Paper

186 Roundtable

Luis Willumsen

Nommon Solutions and Technologies,
London & Madrid

Use of Big Data in Transport Modelling

Discussion Paper

186 Roundtable

Luis Willumsen

Nommon Solutions and Technologies,
London & Madrid

The International Transport Forum

The International Transport Forum is an intergovernmental organisation with 62 member countries. It acts as a think tank for transport policy and organises the Annual Summit of transport ministers. ITF is the only global body that covers all transport modes. The ITF is politically autonomous and administratively integrated with the OECD.

The ITF works for transport policies that improve peoples' lives. Our mission is to foster a deeper understanding of the role of transport in economic growth, environmental sustainability and social inclusion and to raise the public profile of transport policy.

The ITF organises global dialogue for better transport. We act as a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision makers and civil society. The ITF's Annual Summit is the world's largest gathering of transport ministers and the leading global platform for dialogue on transport policy.

The Members of the Forum are: Albania, Armenia, Argentina, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, China (People's Republic of), Croatia, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Israel, Italy, Japan, Kazakhstan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Mongolia, Montenegro, Morocco, the Netherlands, New Zealand, North Macedonia, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Tunisia, Turkey, Ukraine, the United Arab Emirates, the United Kingdom, the United States and Uzbekistan.

International Transport Forum
2 rue André Pascal
F-75775 Paris Cedex 16
contact@itf-oecd.org
www.itf-oecd.org

ITF Discussion Papers

ITF Discussion Papers make economic research, commissioned or carried out in-house at ITF, available to researchers and practitioners. They describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the ITF works. Any findings, interpretations and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the International Transport Forum or the OECD. Neither the OECD, ITF nor the authors guarantee the accuracy of any data or other information contained in this publication and accept no responsibility whatsoever for any consequence of their use. This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Comments on Discussion Papers are welcome.

Cite this work as: Willumsen, L. (2021), "Use of Big Data in Transport Modelling", *International Transport Forum Discussion Papers*, No. 2021/05, OECD Publishing, Paris.

Acknowledgements

Miguel Picornell and Ricardo Herranz from Nommon Solutions and Technologies made valuable contributions to this document, in particular in terms of the more technical details of processing mobile phone records. Marta Ramírez and Javier Torres, from the same company, helped with the figures in the paper. The author is also grateful for valuable comments from ITF staff.

Table of contents

Figures

Old data, new data

Good decision making requires good data and good models, especially for transport planning. This, in turn, requires good data to establish the model base year, a sound representation of travel choice behaviour by users in its formulations and a good estimation of how the future will evolve. Data collection is therefore essential, but it is often perceived by modellers as an endless source of frustration.

Travel surveys provide the best quantitative link between model and reality but despite the best efforts more often than not the results are far less than satisfactory. No amount of quality assurance and supervision seems to be sufficient to avoid the need for adjustments, exclusion of "outliers" and the filling of gaps from other sources. Any new data offering the promise of reducing this dissatisfaction is welcome. Solutions to this conundrum could be found in new data sources obtained from the digital traces generated by the technology embedded in everyday life: mobile phones and smartphone apps, Bluetooth, GPS and smart cards. Mobile network and smartphone app data have been in use for some years now but not always without their own share of frustrations. Will this new data just replace one set of irritations with another?

This discussion paper considers first conventional and then new data sources with a closer focus on the most promising: mobile network data and GPS traces from smartphone apps. It discusses their key characteristics, in particular precision, accuracy and coverage; they help to understand their strengths and limitations. The paper then suggests how to make the best use of these new data sources and ensure that transport models are better and more reliable. The paper concludes discussing how the new data sources can help to address some new issues in transport planning and management. These include the handling of uncertainty including that generated by the prospect of new technologies and modes of transport.

Classic data collection and modelling

For nearly 50 years transport modellers have been using the same data collection methods. This is the case whether the models are classic aggregate versions or more recent Agent Based Models (ABM[1]). The same data collection methods apply to both strategic and tactical models with some minor adaptations.

The most important data collection effort is expressed in the Household Travel Survey (HTS), an expensive exercise traditionally involving careful sampling design and control, lengthy interviews at home and extensive data cleaning and processing. Despite improvements allowing a degree of self-completion and/or support through telephone interviews the time to completion and error elimination is significant. Because of their cost, they usually cover less than 2% of the population in an area. A good HTS generates plenty of useful information on vehicle ownership, income level, activities and travel patterns (trips and tours) at least for the survey day. HTS are a key source of data to estimate mode and other travel choices. However, their small sample size does not reveal the broad pattern of movements in the form of origin destination matrices.

The small sample of trips obtained from an HTS needs to be augmented through Intercept surveys. These are collected by survey agents at the roadside, on-board public transport and/or at interchanges, stations

and stops. Good sampling is difficult, as data can only be collected at a selection of roads or service points. As these surveys involve fewer questions they are a cost-efficient method for collecting trips by purpose, mode and time of day. Roadside Interviews (RSI) are often extended to cover goods movements.

Stated Choice/Preference methods were developed in the eighties to facilitate estimation of choice models when one or more of the alternatives was non-existent in a location, for example when considering investment in a Light Rail or Bus Rapid Transit system. They involve asking people to make repeated choices between hypothetical alternatives presented as packages of attributes like Time, Money and Frequency.

Stated Choice, Intercept and Household surveys require "active" interaction with the traveller to elicit responses to questions. Practitioners have encountered an increasing number of problems with travel surveys. Access to some households can become problematic as people mistrust and resent the intrusion. People have become reluctant to answer yet another lengthy questionnaire (particularly with the proliferation of in-app surveys to rate services) and often skip or simplify replies to lighten the burden. Under-reporting of trips has become a significant issue.

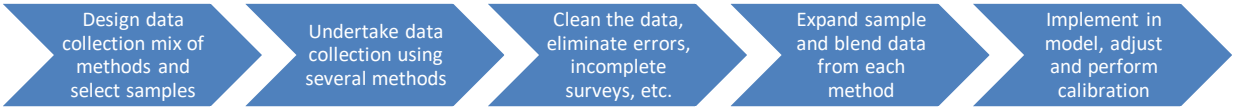
Passive data collection methods do not suffer from this particular limitation, but they do not offer the rich behavioural data of Household Travel Surveys. From the perspective of transport modelling, we need both a good representation of movement in the base year and a deeper understanding of travel behaviour embedded in the functional forms and parameters in the choice models.

Traffic and passenger counts are required to provide expansion factors to represent average or annual conditions and to help detect and account for under-reporting of trips. These are complemented with travel time and supply-side surveys necessary to develop transport models. Number plate surveys, currently using Automatic Number Plate Recognition (ANPR), are sometimes used when only a local vehicle origin destination survey is needed for a micro-simulation model.

The conventional data collection methods above can take a big portion of the study budget and a good deal of the time available to complete it. Data is expensive and has a limited useful life becoming obsolete very quickly. The risks of bias and imperfect sampling are real and require careful quality assurance, detection and correction treatment.

The process of progressing from conventional data collection to valid model is not without adjustments and elimination of errors. It can be summarised in Figure 1.

Figure 1. From data collection to model



The target transport model is usually a slightly mythical "average day of a neutral month" producing average flows and benefits. This is notable as the variability of travel patterns has been known for some time and considered a source of travel time irregularity, a significant concern in large cities.

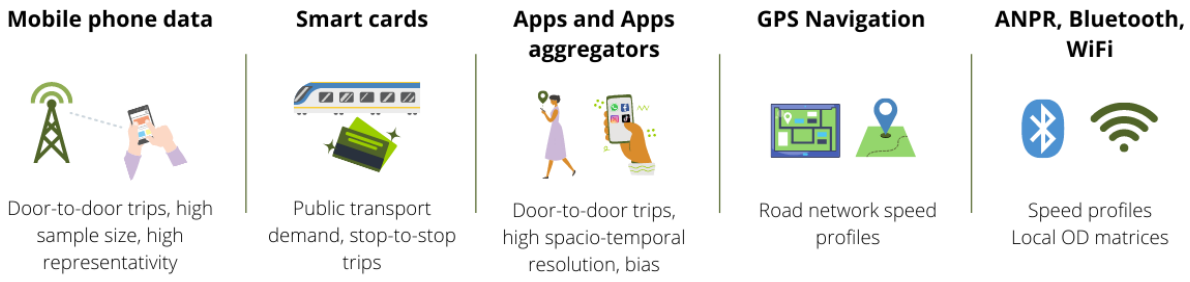
For a multi-modal city model the process from data collection to implementation and forecast may take a couple of years. The model will then be used to develop plans and select projects for implementation. It is generally recommended not to use a model based on data that is older than five years so the model will need updating and this will be achieved mostly using limited surveys plus aggregate data like ticket sales and traffic counts.

The opportunity exists today to use new data sources to replace or support conventional data collection methods, facilitating model updates and exploring new dimensions of travel demand.

The abundance of new data sources should help

New data sources and collection techniques have become available that offer larger sample sizes and longer timespan coverage. Some of them are based on sensors for Bluetooth and WiFi signals, others on public transport smart cards, smartphone apps, mobile network data and vehicle telematics. These are all passive, non-disruptive, forms of data collection that take advantage of the digital traces generated by a range of devices.

Figure 2. New mobility data sources



Not all of these sources are able to identify true origin to destination movements. Well positioned Bluetooth and WiFi surveys pick up data mostly from smartphones and vehicles but suffer the same limitation of number plate surveys as only local movements are collected. Public transport smart card data covers personal movements from stop/station to stop/station. On rail origin and destination stations are usually recorded but on buses only the boarding is identified, the alighting location has to be estimated from other movements. The use of vehicle navigation GPS traces is useful, but the nature of the sample is probably biased and covers only vehicle, not person, movements.

There are two sources of data that offer the prospect of providing true door-to-door movements because they are based on mobile phones: devices which most people carry with them at all times. One of them re-purposes some of the data generated by the Mobile Network Operator (MNO) during the normal operation of the mobile phone network. The other uses the data generated by a Global Navigation Satellite System (GNSS) microchip integrated in all smartphones; the most common of these is GPS. Most smartphones use Assisted-GPS (A-GPS) to increase location accuracy by combining GPS location with cellular data and Wi-Fi network signals.

In this document the first source is referred to as “mobile network data” or “MNO data” and the second as “smartphone data”. When both are considered the paper uses “mobile phone data”.

Box 1. Mobile phone digital traces

Mobile phones generate two broad types of digital traces. One set is produced as part of the normal operation of a mobile phone network and this data is just repurposed to deliver mobility and other indicators. Smartphones also have a GPS capability that is accessed by several apps and they generate additional data.

There are several types of MNO data, collected for different purposes and containing different kinds of information. These are generated by the interaction between the mobile phone and the antenna or Base Transceiver Station (BTS). MNOs collect call detail records (CDRs) generated during calls and messages, and internet protocol detail records (IPDRs) produced when data is transferred to and from smartphones in order to charge users. The acronym CDR is often used in practice to cover these two types of data.

MNOs also generate data to monitor its services and ensure the phone is found quickly when a connection is required. One such data is generated when the phone moves from one group of cells (a Local Area) to another. Another type of data results from **probes** used by the MNO to improve the quality of the service further. In this case each device is "probed" at regular intervals to provide a faster response and monitor other aspects of network quality. These two types of data are often combined with CDRs to deliver more frequent data points identified in this document as mobile network data. Each CDR, Local Area update and probe contact generates a time stamp and an identifier of the cell contacted; the spatial precision of this data point is, therefore, that of the cell involved in the contact.

Many smartphone apps use location data to enhance the service they deliver; typical examples are Trip Planners, Route Guidance and Weather Forecasting. Location data generated in this way is then collected and consolidated by companies supporting device identity (ID) advertising. Mobile phone operating systems provide a device Marketing ID to support advertising; Google provides an Android Advertising ID and Apple an Identifier for Advertiser (IDFA). They are employed to link location and time stamps (and other user data) from different apps on the same phone. Marketing IDs are used to target advertising to a specific device as a function of location and user profile and can be re-purposed to deliver movement data. The user can restrict and reset these identifiers, but few people would do this. The spatial precision of this data is that of the location service on the phone, usually a combination of GPS and other signals.

These mobile phone data sources are processed to deliver commercial mobility products by different businesses. There are at least two types of suppliers for trip matrices from mobile network data. One is the Mobile Network Operator (MNO) itself, seeking to monetise directly the data they hold; this is the case for example for Telefónica under the brand name Luca[2] providing trip matrices and other mobility indicators. Sometimes consultants take these MNO trip matrices and adapt them further to fit more closely the needs of end-users. There are also a few data analytics companies like Airsage[3] , Teralytics[4] and Nommon[5]; they process the raw data from MNOs, combine it with other data sources, add context and deliver trip matrices and other mobility indicators directly adapted to the requirements of the final user.

Smartphone app data is usually aggregated from a range of applications. Aggregators include the companies Unacast[6] or Cuebiq[7]. Their data is then used by other companies to generate more specific mobility indicators, for example Carto[8] and Streetlight[9]; these companies add contextual information and other data to provide a range of mobility and presence indicators.

Trip matrices from both MNO data and app data have been available and used for well over five years. However, their acceptance has not been universal. There are many stories of frustrated users disappointed by the quality of the matrices they get. They question how close the representation of mobility given by those matrices is to reality, illustrating the inherent limitations of the technique; it is almost impossible to

distinguish the mode of transport used in dense urban areas. On the other hand, some data users seem to be very happy employing them and come back for more. It is important to explore the reasons behind this mixed appreciation, attempt to understand the strengths and limitations of these data sources, and produce some recommendations on their specification and use. To do this it is necessary to explore more closely the nature and characteristics of these two data sources.

The nature of the new data sources

The simplest MNO data source is the CDR that logs the use of mobile phones whenever a connection occurs between the device and the telecommunications network. Originally limited to phone calls and text messages, the term is now taken to include the transmission and reception of messages and data connections for emails, route guidance requests and so on. It includes an identifier of the mobile phone, a timestamp, and the identifier of the tower and antenna involved in the contact. A common configuration is a tower or Base Transceiver Station (BTS) with three antennas, each covering 120 degrees generating three cells. Cell boundaries are not crisply defined as there is a degree of overlap between cells, in particular in dense urban areas where towers and cells are closer together.

It is clear from this description that the temporal and spatial granularity of mobile network data depends on the density of towers and cells and the frequency of the interactions between phone and BTS. In densely populated and active areas cells may allow location precisions down to 100 metres where this drops to kilometres in sparsely occupied rural areas (ITF, 2015 [Box 8]).

The other type of data comes from smartphone apps accessing the actual location (mostly through GPS) in order to "provide a better personalised service". The user can specify, under "Privacy/Location Services" which applications have permission to use the location of the phone; the options are "Never", "Only While Using" the app and "Always". Where the user chooses "Always", the app continues collecting location data in the background, for example to record a distance travelled. Many, if not all, users opt for the "While Using" option. The app data aggregator would use the marketing ID to link location data from different apps to fill any gaps. The precision achievable is in the range 10 to 50 metres (ITF, 2015).

These two data sources provide door-to-door account of movement and can be used to identify stops or stays in particular locations when the user does not move for a specific period of time. The passive collection of this data may reduce the potential mismatch, in travel surveys, between what people say and do.

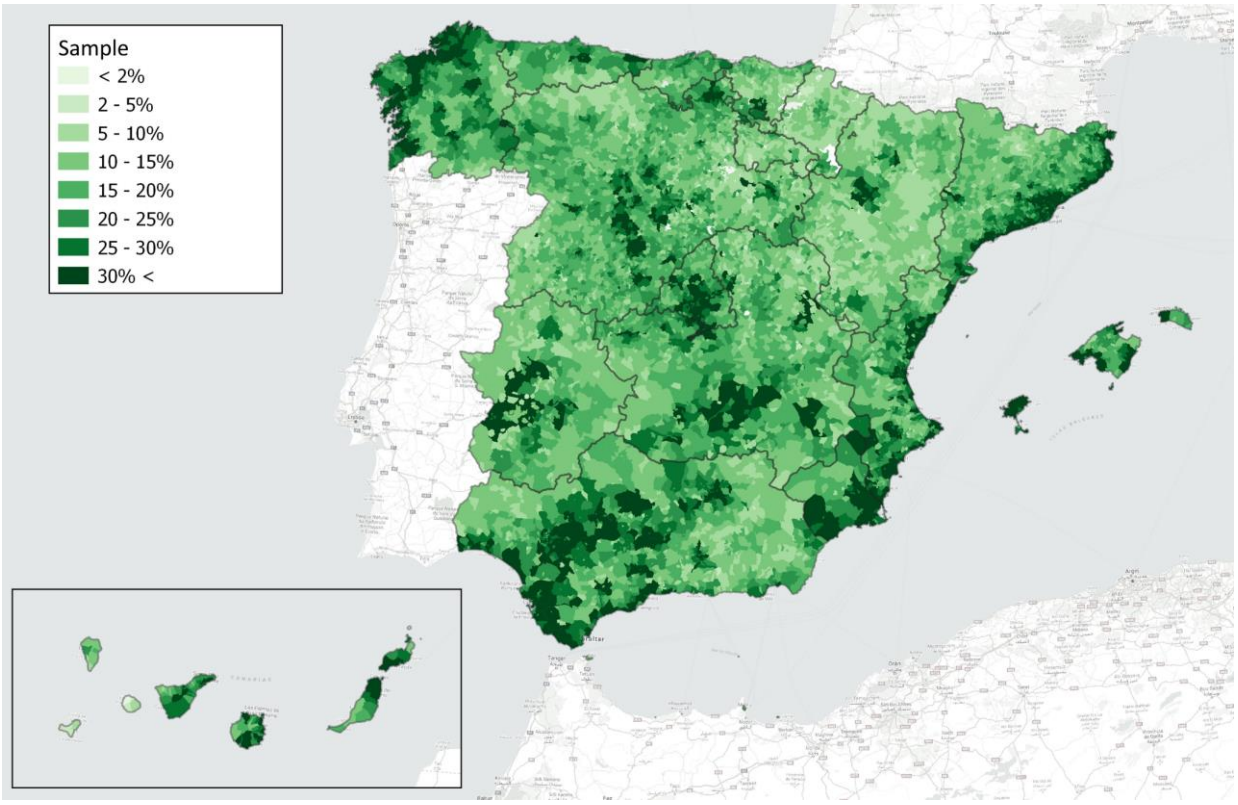
Alas, no data source is error-free and eliminating or compensating for errors is key to the quality of the result. In the same way as Household Travel Surveys suffer from underreporting, mobile network data and smartphone app data could also miss some trips.

When discussing the quality of data sources, the following aspects are usually taken into account:

1. Sample size. MNOs generate some of the largest data samples, only limited by the market share of a given telecommunication company. In most countries the sample can represent between 15% and 40% of the population not just for one day but on 365 days a year. For example, Figure 3 depicts the sample rate achieved with MNO data in Spain. As can be seen this is not uniform but

generally much higher than the sample achieved with household travel surveys. Data from smartphone apps also offer a large sample but one usually smaller than that of MNO data. This may be because not all apps in the bundle are active at all times and not even every day.

Figure 3. Map of sample rates using data from one mobile telecommunication operator in Spain



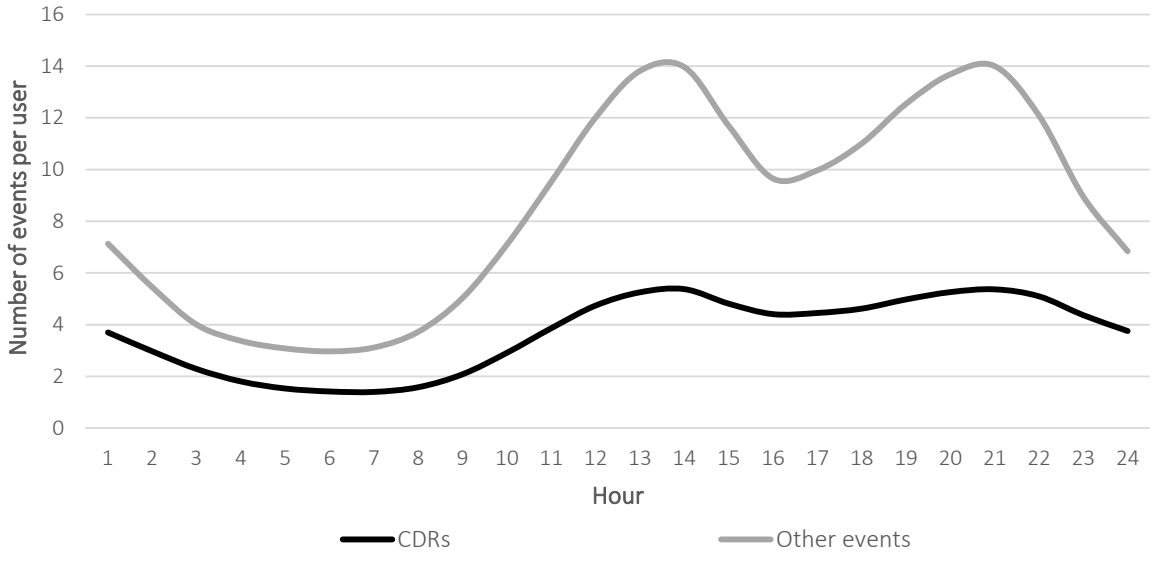
Source: Nommon Solutions and Technologies.

2. Bias. Most data from MNOs is considered to be sufficiently well spread over space and socio-economic groups to be able to correct for bias through appropriate sample expansion techniques; this is usually achieved taking into account the socio-economic characteristics of the population in each zone. The bias present in app data is more difficult to know and to correct. App consolidators merge data from different apps and do not disclose them; in general it is more likely that certain population groups are underrepresented or even absent from the sample than in the case of MNO data.

3. Coverage. Both MNO and app data provide coverage for a whole country. App data allow the tracking of the users in any country, and therefore are not limited by country limits; however, it is often the case that the apps that are more popular in one country are less spread in others, so the sample size may be heterogeneous depending on the target markets of the app aggregator. MNO data can also be used to indicate whether the user is a visitor from abroad (international roaming) identifying the country of origin of the mobile phone contract; some MNOs also provide roaming-out registers, which allows the identification of international trips.

4. Precision. There are two dimensions to geocoded data precision: time and space granularity; that is how often a transaction identifies a trace and how accurate is the location obtained for that trace. MNO data is providing more and more traces as people intensify the use of mobile data; Figure 4 shows the average number of data points per user generated by an MNO, in this case in Spain. It shows the value of using more than CDR data and including other events such as probes and Location Area Updates (LAU). App data usually provides more frequent data points depending on the number and duration of the interruptions in the use of the apps. In terms of spatial precision, smartphone apps offer better geolocation than mobile network data. MNO data is dependent on the size of the cells; this can be quite small in dense urban areas and much bigger in rural underpopulated regions.

Figure 4. Number of data points per user and per hour in mobile network data



Source: Nommon Solutions and Technologies.

5. Accuracy refers to whether the identification of trips and trip chains obtained from MNO or app data correspond well with reality in terms of the location of origin and destination and time of travel.

6. Continuity. The lack of continuity over time degrades the usefulness of the data. This may happen in different ways. Some MNOs re-anonymise the data with some frequency, for example once a day or even a couple of hours. This makes it impossible to track sufficient movements and stays to identify the place of residence with confidence; this, in turn, makes correcting for bias and expansion of the sample much more difficult. Smartphone app data seems to suffer less from this issue as the Marketing IDs used are more persistent. However, people install and uninstall apps much more frequently than they change their mobile network operator, so the temporal series available tend to be shorter, which limits the possibilities for certain types of longitudinal analyses.

Ultimately, both conventional data collection methods and those based on new sources need additional data for expansion and refinement. This will be discussed later in this document.

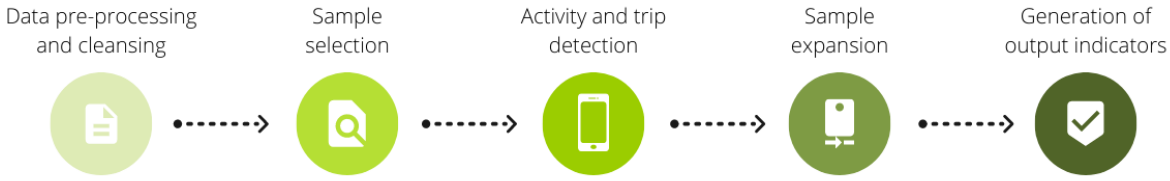
The nature of mobile network and app data

The CDRs and probe data available for processing contain an anonymised (and irreversible) identifier of the user, a time-mark, and the identifier of the antenna and cell used in the connection; plus some additional information not used in mobility studies. The cell map from the MNO provides the rough geographical coverage of each antenna; this is an area rather than an accurate longitude-latitude spot location. There are ways to refine positioning, such as signal triangulation, but they are all process intensive and still not regularly used by most MNOs. In any case, the size of a 3G/4G cell is usually compatible with most transport model zones and 5G cells are even smaller. Each event in the MNO dataset then provides an approximate location and a time stamp. The granularity of these geo-located data points depends on the size of cells and the frequency of phone-antenna connections.

Smartphone apps data also contains an anonymised identifier of the user, a time stamp and approximate longitude-latitude co-ordinates for the location at the time of the contact.

The general approach to estimate trip matrices (and other mobility indicators) from passive mobile network and smartphone app data is illustrated in Figure 5:

Figure 5. Outline of the process to generate trip matrices from mobile network operator and app data



There are five important steps in the processing of mobile phone data, MNO and smartphone apps, and each one of them influences the final result.

The first step is to pre-process and clean the data. The pre-processing usually involves selecting the useful data points and organising them for efficient processing. Mobile network and app data are not error free and must be cleansed before further analysis. Analytics companies providing these indicators have developed their own filters and algorithms to detect errors and eliminate registers that are considered to be unreliable or faulty. The quality of this first step affects, of course, all further processing of the data.

The second step is to select a useful sample as not all phones generate useful data (for example there are SIMs that do not move at all as part of the Internet of Things). The potential effective sample comprises those mobile devices and apps generating useful information about their users' daily activities and trips. The selection of these users is based on a set of criteria related with their mobile phone activity, which shall be enough to determine their mobility patterns with an adequate level of accuracy and reliability. This is less critical when using also network probe data, but still there may be users that, for example, switch off their mobile phone or run out of battery for a significant part of a particular day and may need to be removed from the sample. Again, a sensible selection of this sample and subsequent bias correction affects the quality of the resulting trip matrix.

The third step is the identification of the sequence of stays (dwells) and activities undertaken by the user. Trips are recognised as the connection between two stays. Stays are identified by slightly different algorithms in the case on MNO and app data. In the case of MNO data, a stay is identified when there are two or more CDRs/probes in the same location during a particular period of time; many algorithms use a threshold of the order of 30 minutes to this end, but the exact value depends on the frequency of contact

events with the BTS. In the case of smartphone app data GPS "pings" within a particular spatial range in a particular time threshold are used: for example pings within 80 metres of each other and over 15 minutes. These criteria are very important in the identification of true stays and trips; too short a time threshold may not be detected accurately and one that is too long may miss some activities and short trips.

The algorithms developed by different analytics companies combine different criteria based on stay times, itineraries and longitudinal behavioural patterns to identify activities, trips, intermediate stops subordinate to the trip (e.g. a stop to change to another transport mode in a multimodal trip, etc.) and the tours connecting different trips. An "activity" is defined as an interaction or set of interactions with the environment that takes place in the same location and motivates an individual to reach that location. A "trip" is defined as a sequence of one or more "stages" or "legs" between two consecutive activities, each stage involving a single mode of transport and no interchange. This way, a trip has a main purpose identified by the activity at the destination.

The result of this process is the sequence of activities, trips and tours performed by each user in the sample and for the period of study. The parameters used to define them are usually refined over time and may be project dependent; they will also influence the final trip matrix.

Having identified stays the next step is to infer the place of residence and work/study based on the user's longitudinal behavioural patterns during several days/weeks. This is usually achieved identifying longer stays, overnight for residence and during the day for work, although this can be improved upon by allowing night shifts. The location of residence is later on used to select the sample for a project and to expand such sample to the total population.

The fourth step is the expansion of the sample to the total population while at the same time compensating for any bias in the data: the approach used for the upscaling of the sample depends on the characteristics of the study. In the case of the residents in the country for which the mobile network data is available, the expansion of the sample involves the use of factors based on the home location, typically at the level of census tracts. In the case of tourists and visitors from abroad (roamers), other frameworks are needed (e.g., official statistics on number of tourists and visitors, ports of entry, etc.). In any case, the sample expansion process shall correct for the different level of penetration of the MNO in different areas of the territory and different population groups. Simpler approaches based on overall figures about the market share of the MNO are therefore prone to errors and bias and are not recommended.

Finally, once the sample has been expanded to the total population, the activity-travel diaries are post-processed to produce the information requested by the project (e.g. origin-destination matrices) with the required level of spatial and temporal aggregation. Note that at this stage the data provided to end users is always aggregate. For example, the processing just outlined means that the number of trips in each cell of the trip matrix will be a decimal number of trips, for example 8.245; this number may be the result of one expanded trip or the average of different individual trips over a month of working days.

The transport modelling market is the most demanding of the accuracy of the data generated by mobile network operators and apps. Understanding how this data will be used is therefore of paramount importance in refining the five steps above, in particular the last three. Perhaps, one of the reasons why some customers have found the trip matrices from mobile phone data frustrating is because the specific implementations of steps three to five were too generic rather than specific to transport modelling.

Overall, trip matrices from mobile phone data should be able to provide good quality information about movements in a study area not just for an average "normal day" and its variability but also for a whole range of days including specific dates. Having said that, mobile phone data still has significant limitations that must be overcome by other means.

Limitations of mobile phone data

This section identifies some of the limitations of mobile phone data for trip matrix estimation. It comments on ways that analytics companies have found helpful in overcoming at least some of these limitations. The granularity of the geo-location is limited by the size of the mobile network cells and the accuracy of app GPS data; the granularity of the time-stamps is limited by the frequency of the interactions between mobile phone and antennas (BTS) and active apps. These two constraints make it impossible to accurately identify the start and end time of any particular trip. Therefore, the estimation of the exact start and end of a trip and the duration of an activity must be established through probabilistic models.

Another consequence of this limited granularity is an inability to identify short trips that take place within a mobile network cell in the case of MNO data and within the time and spatial thresholds for app data. These are mostly walking trips and will become intra-zonal trips in a transport model. It is also important that the algorithms used are able to identify "cell jumps". These apparent movements between adjacent cells may result simply because of the way the mobile network operates: if one antenna becomes overloaded a call may be transferred to a nearby cell (even if the phone has not moved) creating a "phantom trip". These must be detected (using knowledge about travel and networks) and eliminated from a trip table.

The identification of the Home location for each user helps associating socio-economic characteristics to the traveller. This may be assisted by some additional information from the contract with the Mobile Network Operator (MNO), for example gender and type of contract. Recognising activities other than Home and Work/Study is made more difficult as most non-residential locations cover areas of mixed use. The most reliable purposes identifiable from mobile phone data are Home to Work/Study, Back Home, Other Recurrent and Other Non-recurrent trips (for all other remaining trips). This distinction between recurrent and non-recurrent trips is new and so far, not well exploited given that we know that most trips in any study area are, in effect, non-recurrent.

Protection of privacy

The protection of privacy is a central concern of both Mobile Network Operators, app aggregators and data analytics companies. The key idea is to provide this protection by design and never to offer individual data that could possibly be traced back to an individual.

CDR and network probe raw data is always anonymised (it would be more accurate to say the data is pseudonymised) using a hash functions such as SHA-224 or SHA-256, which are implemented in many security and cryptography frameworks. Then, only aggregate output is provided to the end-user and only aggregate data leaves the firewall of the MNO. This protection is sometimes strengthened by some providers using K-Anonymity, that is no cell data is provided if the real number is smaller than K. It is sometimes the case that two providers in the same country apply different criteria for selecting the value of K. In general terms, the goal of the anonymisation process shall be preventing any reasonable risk of re-identification.

There are other ways to achieve this by generating an entirely synthetic population that perform activities and trips modelled on the observations. This is, in essence, what is done when using Household Travel Surveys to produce disaggregate models except that in this case the sample size is at least an order of magnitude bigger and the level of detail includes variability of activities and behaviour on different days and times of the year.

Data fusion

The granularity of mobile phone data also makes it difficult to distinguish car, bus and even bicycle movements in dense urban areas and the same is true with the precise routes taken by them. Some of these limitations can be overcome through data fusion. There are two main sources of additional data: context and aggregate mobility information.

Good context data include:

- Census data, essential to expand and correct the sample
- Land use data, very useful to identify the type of activity in each location
- Points of interest, including attraction centres like stadia and logistic hubs
- Data from existing surveys, for example recent household travel and consumer surveys

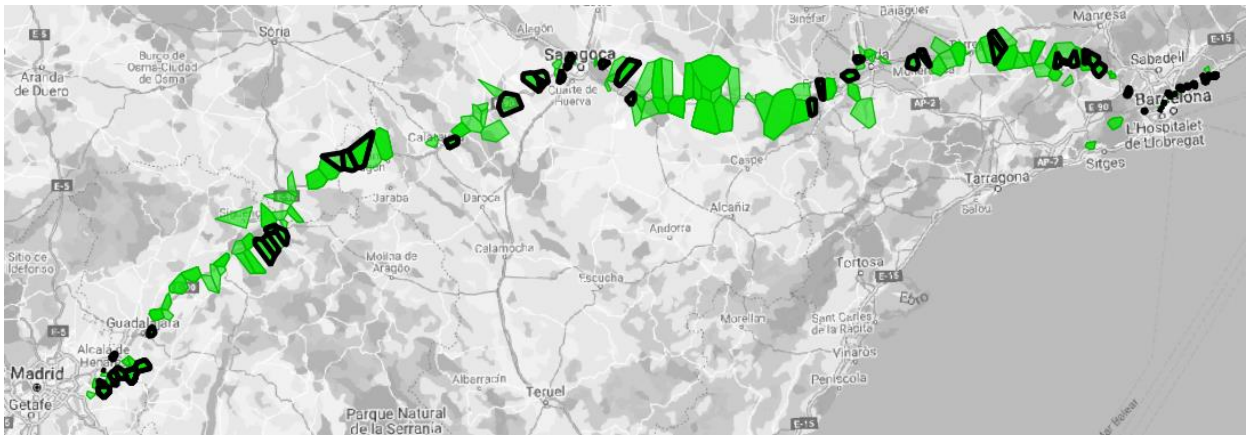
Good candidates for mobility data are:

- Traffic and, where available, person counts
- Public transport routes and frequencies
- Smart Card data, for example Oyster in London, or simply ticketing information

The availability of one or more of these data sources will certainly help to process and improve the accuracy of the trip matrices generated from mobile phone data. The precise approach adopted in each case will depend on the objective of the study, its context and the data available. For example, map data including points of interest can be used to snap trajectories to viable infrastructure and identify routes. Bus ticketing and level of service data may be used to start segmenting trips by mode and corridor; this mode split can be further refined using classified traffic counts. Data science principle should guide this fusion to make it rigorous and reliable.

Figure 6 shows a route taken between Madrid and Barcelona identified with the help of CDRs (in green), probe data (black border) and map data.

Figure 6. Route identified using call detail records, probes and map data



Source: Nommon Solutions and Technologies.

One must bear in mind that different data sources have different error distributions, and these must be considered during data fusion. For example, data from smart cards and bus movements also require cleansing, error and bias correction and sample selection before combining it with mobile phone data in urban areas.

Some of the advantages of data fusion to help develop trip matrices have been explored by Montero Mercadé et al. (2019). Data fusion has also been used to enhance information about travellers, for example by Picornell et al. (2015) and Burrieza et al. (2019) both indicating significant progress.

Overall, data fusion offers the best approach so far to overcome the inherent limitations specific to mobile network and app data. The quality of the resulting trip matrices will depend, of course, on the type of complementary data available and the actual process of data fusion. A good understanding of the domain of these sources and the use that will be made of the resulting mobility indicators is important in devising the most appropriate data fusion strategy in each case.

Reliability and accuracy

For the purposes of this paper, it is assumed that the main use of these new data sources will be in the development of transport models. Therefore, the main interest is in the generation of trip matrices with as much useful segmentation and granularity as possible. Ideally, the trip matrices would be made available with the same segmentation used in most models, that is by:

- origin and destination zones;
- time of day;
- trip purposes, at least Journey to Work and Education, Work trips, Shopping, Non-home based and other purposes;
- mode of transport, usually a main mode or a significant mode combination, for example park&ride;
- person type, usually associated to income levels.

Mobile phone trip matrices can be provided in zone system of most transport models as well as time periods of one hour or more. The number of trip purposes is, however, more limited; usually Home to Work and Education, Back Home, Other trips, home and non-home based. Further refinement requires data fusion, most likely with other surveys.

As mentioned before, the identification of mode of transport is very difficult in urban areas and therefore its discrimination requires fusion with other data sources, in particular ticketing and smart card for public transport and classified counts.

The identification of person type by income can be obtained using probabilities associated to the population in each zone.

There are two basic approaches at using this data bearing in mind its limitations. One approach is to use only the general movement trips matrices as a 'seed' origin destination matrix and use the model and context data in it to generate disaggregated trip matrices following the pattern above. This can be achieved using prior information from other surveys and perhaps legacy models; in essence, the mobile phone trip matrix is used as part of a process for enhancing and updating a model.

An alternative is to require the data analytics firm to achieve as much segmentation as possible using data fusion. This has the advantage that this process can be made more effective when fusing data at a disaggregate level, for example when using public transport service data to help identify modes. This approach is more process intensive than the simple generation of general movement trip matrices but improves confidence in the data. The calibration of the transport model then performs the final adjustment to the trip matrices.

Validation

A confirmation of the actual accuracy of mobile phone trip matrices is highly desirable, especially as it is a new potential contribution to better models. There have been a number of studies comparing mobile phone-based trip matrices with those obtained from other sources, mostly household and intercept travel surveys. Just searching "validation of mobile phone trip matrices" yields a large number of papers and

reports. Most of them praise the advantages of mobile phone data to generate mobility information: passive data collection, large sample size, short time to results, capture of specific dates and also day-to-day variability.

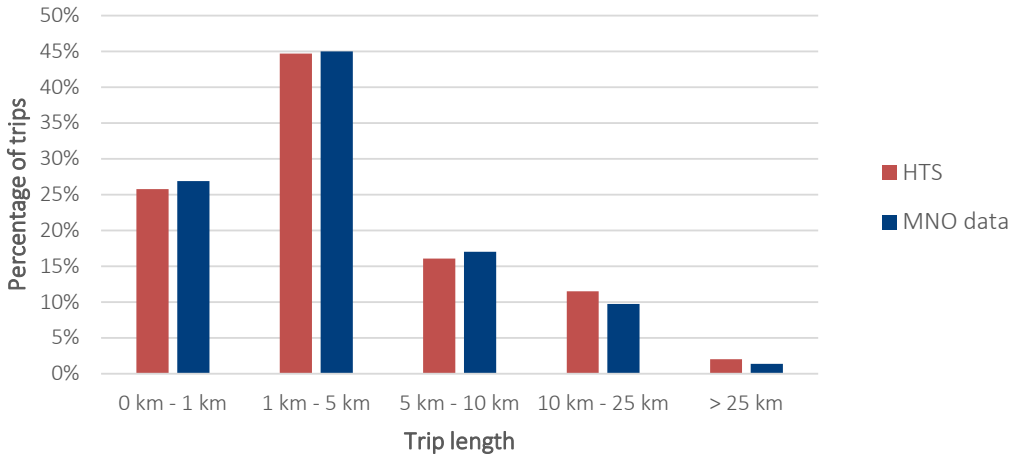
For example, Bonnel, Fekih and Smoreda (2018) compared trip matrices for the Rhone Alpes region in France generated from mobile network data and from a household travel survey. They tested different criteria for identifying a "stay" and found that the use of a value between 30 and 40 minutes produced the best match with the survey. The correlation between the two data sets was high, in particular after eliminating an outlier.

Tolouei, Psarras and Prince (2017) performed a comparison of trip matrices from mobile network data and from roadside interviews (RSI) and concluded that "*…using the mobile phone data, when systematically refined and adjusted using independent data sources to address various known limitations and biases, does not seem to be either biased or less accurate than conventional* methods*. It has also been observed that areas of the model where no RSI data or other similar observed data are available, use of mobile data could result in a more consistent estimate of trips, benefiting from a significantly larger sample size.*"

One of the issues in validation is, of course, the accuracy of the contrasting trip matrices. In other words, how valid is a comparison between a 2% sample Household Travel Surveys against a 20% sample from mobile phone data. As none of them is the "ground truth" most comparisons employ some aggregation of the trip data from both sources.

Analysts often compare trip length distributions. For example, the figure below shows the proportion of trips of a particular length range (in kilometres) in the case of Valencia, Spain, contrasting the results from mobile network data ("MNO data") with those from the household travel survey ("HTS").

Figure 7. Comparing trip lengths between household travel surveys and mobile network data

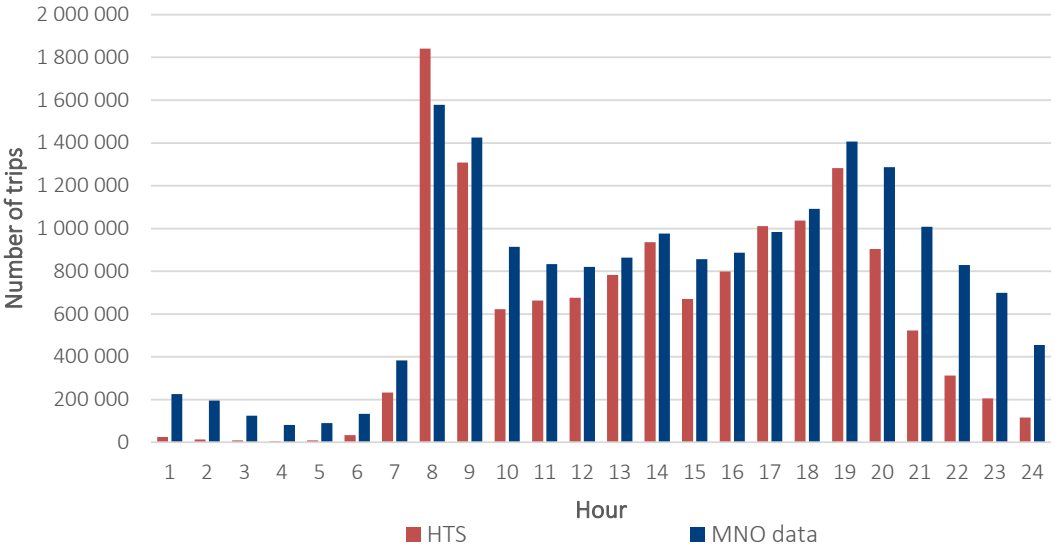


Source: Nommon Solutions and Technology.

As can be seen, the trip length distributions match reasonably well.

Another test is to compare the number of trips generated at different times of the day from both sources. These are shown, for example for Santiago, Chile, comparing mobile network data with a household travel survey.

Figure 8. Comparing number of trips by time of day in Santiago, Chile



Source: Nommon Solutions and Technology.

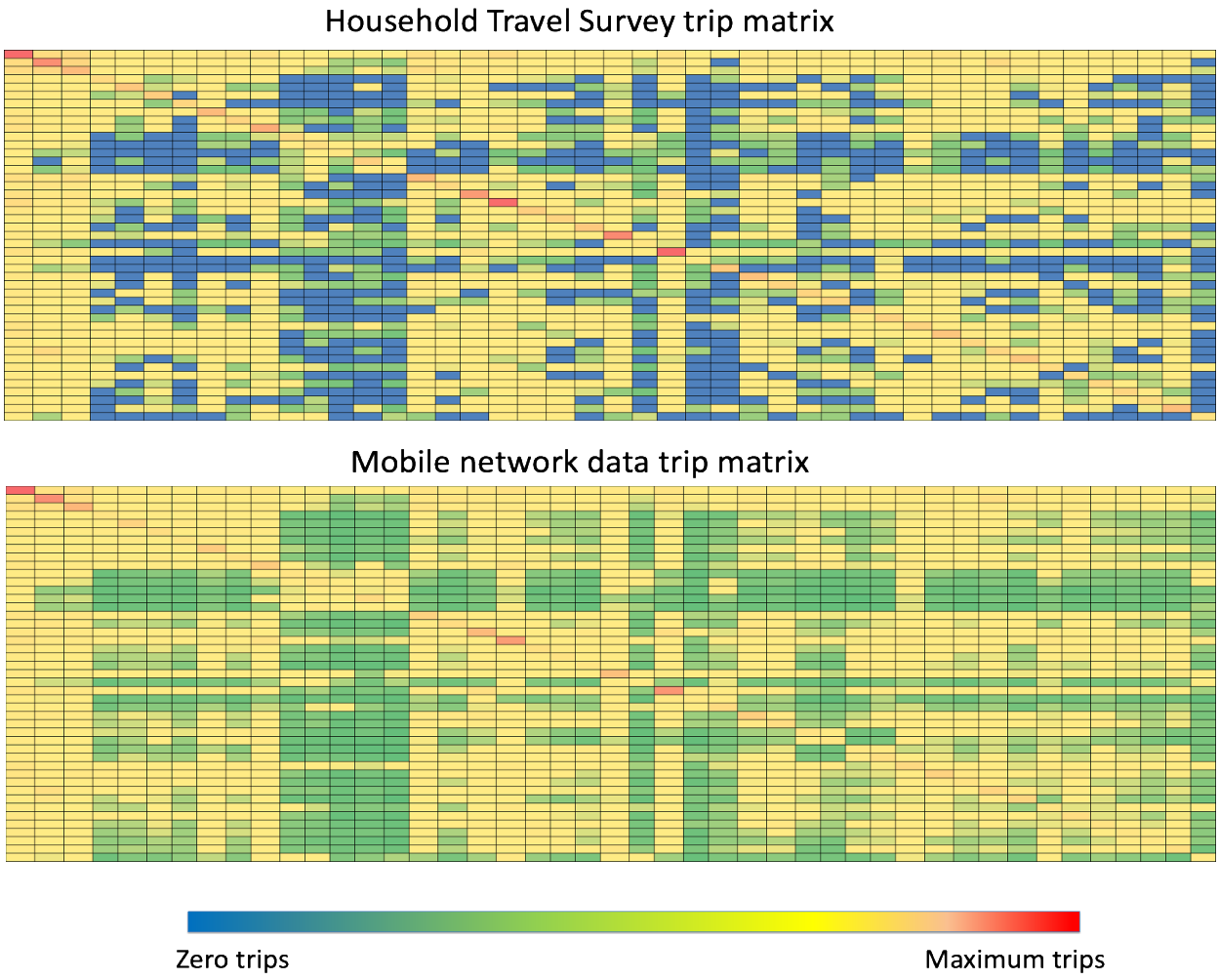
It is interesting to note that the number of trips is similar in the morning peak but differ in the evening and night. Mobile network data shows more and later trips than reported which could be due to the difficulty of recalling these times accurately.

Comparing trip matrices

In the case of Santiago, Chile, the author also compared trip matrices from mobile network data against the 2012 household travel survey using 45 macro-zones. The next figure paints the number of trips in each cell so that blue indicates zero trips and red the maximum number of trips (usually intra-zonal). It clearly shows how the travel survey has plenty of empty cells where no trips have been observed whereas the mobile network data based trip matrix has values in all cells. Therefore, it can be argued that trip matrices generated from mobile phone data provide a more realistic representation of average trip matrices in an area.

Overall, efforts to contrast mobile phone trip matrices against experience and other data sources support the idea that good data processing algorithms and an in-depth understanding of how the trip matrices will be used can result in high-quality and usable movement patterns.

Figure 9. Comparing trip matrices from household travel survey and mobile network data in Santiago, Chile



Source: Nommon Solutions and Technology.

Guidance

This section deals with existing guidance on the estimation of trip matrices using mobile phone data and on how best to specify the desired product. Authorities in different countries have explored the use of MNO data in this context and offered some guidance on their adoption. For example, the UK Department for Transport published "Utilising Mobile Network Data for Transport Modelling" (Transport Systems Catapult, 2016) as a recommendations paper. The same year the US Federal Highway Administration published "Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data" (FHWA, 2016). The Inter-American Development Bank also published a guide on the use of Big Data in urban transport planning (BID, 2019).

All of these documents stress the importance of a careful specification of what transport planners need. All suppliers of this type of trip matrices calculate the cost of delivering them mostly as a function of the granularity and the amount of information requested: the level of detail, the number of zones, days and time periods required by the customer.

Zoning system

It is recommended to start with the zoning system required for transport modelling purposes and discuss this with the trip matrix provider. Any data analytics company should be able to comment on the level of detail in case this is too granular for the direct use of mobile phone data. If the zone details are suitable, the analytics company should be able to adapt its processes to any particular zoning system.

Some app data consolidators and MNOs generate trip matrices on the Quadkey[10] tile system. The most common size for these tiles for the purpose of storing trips matrices is either level 16 (about 600 by 600 metres) or 17 (around 300 by 300 metres). The choice depends on the number and quality of the data points. The user can then adapt this tiled trip matrices to their own transport model zones.

This author prefers that this adaptation is done by the data analytics company as it is better placed to adjust the sampling, processing and expansion stages to any suitable zoning system.

Nevertheless, one must take care in dealing with external zones that any model needs to account for. In some cases, it may be difficult to provide data for external zones that represent "the rest of the country through a gateway"; this is because the way some MNOs select and pre-process the data. In this case, it is better to specify some large zones adjacent to each gateway knowing than some trips from further away will be missing and may have to be adjusted using gateway counts.

Trip and activity segmentation

As mentioned above, only Home to Work/Study, Back Home, Other Recurrent and Other Non-recurrent trips are naturally identified using mobile phone data. Other purposes can be inferred with much lower accuracy, but these are best segmented through data fusion with other sources. Moreover, one could argue that the need for finer segmentation is partially driven by the need to develop different demand models by purpose to generate better synthetic trip matrices.

If the aim is to develop agent-based models it is best to request data on the sequence of stays, activities and trips. This is a more complex data structure than a trip matrix but provides the full linkage of activities and trips required for such a model. This is an area only recently explored by modellers but one that is likely to be particularly rich in insights.

There is at least one example where MNO data has been used to support an activity-based model for Barcelona, Spain (Bassolas et al., 2019). They found that mobile phone data was useful in delivering an activity based model using the MATSim simulation package.

Time of day and day of the week

It is possible to specify a range of time of day slots but one should bear in mind two features of mobile phone data. First, the exact time of starting and ending an activity is only statistically approximated therefore very short time slots are currently not recommended for strategic modelling. Second, the user must specify what part of the trip should be in the time slot and the core study area: the beginning, the end or its middle. Mobile phone trip matrices are, in this sense, different from those obtained through intercept roadside interviews in the study area where this question does not arise.

Most data analytics companies use a month of data to generate trip matrices or tour sequences for particular days or combinations of days. Therefore, the client may specify the month (or months) required and the type of days of interest: average working day, average Tuesday to Thursday day, specific day, etc. The greater the number of these days and time slots the more extensive the processing and cost of the resulting matrices.

Modes of travel

Map matching and statistical inference techniques can be used to infer transport mode and route choice from the overlapping between: i) the spatio-temporal trajectory of the user observed from mobile phone records; ii) the transport network (road network, location of transport hubs, etc.); iii) the travel times for different transport modes and route choices, obtained from diverse data sources such as online travel planners APIs. The range of applicability of map matching approaches is limited by the spatio-temporal resolution of the mobile phone data: while this type of approach is very robust and reliable for medium- and long-distance travel, mode and route identification is not always possible for short trips. Additionally, even for long-distance trips, distinguishing the type of vehicle used for road trips can be challenging. Data analytics companies have developed approaches to estimate freight movements inferred from patterns of usage, but this still benefits from classified counts where available. Distinguishing the type of vehicle for passenger movements (e.g., car vs bus) typically requires other additional data, such as detailed information on the supply of bus services.

Mode identification is particularly problematic in urban areas, due to the density of the transport network and the coexistence of different transport services. For certain origin destination pairs the travel times and user trajectories for different transportation modes (e.g., car, bus and bicycle) can be very similar, making it practically impossible to reliably identify transport mode and route. In these cases, other approaches based on data fusion and transport modelling are employed to segment travellers by transport mode.

The normal output will be person-trips which for some projects may need converting into vehicle-trips using traffic counts and/or occupancy rates obtained from other sources. This type of task must be discussed with the data analytics company at an early stage.

The contribution to transport modelling

Mobile phone data will not replace all data collection required to develop useful transport models. Nevertheless, trip matrices derived from mobile phone data can make important contributions to improved models and demand forecasting.

First of all, by providing an all modes and purposes general mobility trip matrix they reduce the need for large (and expensive) Household Travel Surveys. Travel surveys can then be focussed on obtaining good trip/tour/activity generation, destination and mode choice models; for this, a smaller sample should suffice depending on the segmentation required. Stated Preference surveys, where needed, may still play a useful role.

The role of intercept surveys may well be replaced by a combination of mobile phone trip matrices, smart card data and counts depending on local conditions and the complexity of the public transport network. Third party vehicle GPS data can provide average speeds.

There is scope for investigating methodologies for specifying optimal sampling for household travel and intercept surveys when good mobile phone data is available. Some research in this direction has already been proposed, for example by Bonnel and Munizaga (2018).

Another advantage of mobile phone data is the short time required for its generation and validation. This should enable faster model development even considering the need to complement it with a smaller sample household travel survey.

Then, there is the opportunity to see travel and mobility in a richer form than possible with conventional survey tools. For example, the distinction between recurrent and non-recurrent trips mentioned before opens the opportunity to explore the day-to-day variability of travel demand as well as its evolution over time (Box 2).

Mobile phone data can help monitor travel behaviour in response to natural experiments and changes in transport policy. Whilst not immediately related to transport model development, this aspect is nonetheless important to understand the dynamics of travel demand.

Modellers have been operating very consistent types of transport models for the last 50 years with some more recent modest innovations. However, there are some new challenges to transport planning that require updating the supporting models. Two of them, dealing with equity impacts and uncertainty, can be addressed using improved aggregate model approaches. Nevertheless, the prospects of new technologies and new modes of transport require a different type of model.

The future of modelling

There are three main mobility challenges that future transport modelling must address: Electric Vehicles (EVs), Mobility as a Service (MaaS) and Connected and Automated Vehicles (CAVs). EVs behave in traffic like other vehicles but need charging points at specific locations. This is critical for a good EV taxi service for instance. The optimal location of these charging points depends on their patterns of use and can still be estimated using aggregate models.

Box 2. Variability of trip matrices

One of the common complains of travellers is the poor reliability of services and travel times. Policy makers are therefore committed to improve this reliability. The predictability of travel times is hampered by the variability of demand, in particular under congested conditions. Because conventional data collection is expensive and static, current transport models implicitly assume that origin destination matrices are stable and trips recurrent.

That trip patterns are not that recurrent has been noted for some time. For example, vehicle number plates can be recorded at the same location on several days and to identify how many are "unique" (appear only once) and how many reappear on other days. Cherrett and McDonald (2002) surveyed four different locations during the peak period around Southampton and found that, depending on the road only between 25% and 49% of the vehicles would reappear on subsequent days. Del Mistro and Behrens (2008) undertook a similar survey over three weeks and several roads in Cape Town; they found that in arterial roads only 34% to 43% of the vehicles would reappear on subsequent on one or more subsequent days; this percentage increased to around 54% for residential streets.

It can be expected that the new data sources shed some new light on the variability of travel demand and provide the additional data that conventional methods could not provide. This is an area of research and modelling improvements that has become feasible only because of new longitudinal data streams.

This is not the case for MaaS and CAVs. Autonomous vehicles may travel empty to serve other members of the family and friends or join and be part of a MaaS fleet. In modelling terms, this requires tracking not just passengers but also vehicles. Indeed, all demand responsive modes generate a number of empty vehicles kilometres; this is true from conventional taxis to the most ambitious ride sharing systems.

Customers of demand responsive services may or may not wish to share rides with others, entailing efforts to allocate vehicles to travellers. Estimating the efficiency of this vehicle management, the level of service provided to users and the impact on congestion and emissions requires identifying individual timed requests and how well they can be served with a particular fleet.

A key challenge with the new mobility technologies is that not all will contribute to reduce congestion and emissions. In effect, all new modes that shift passengers from a car to another single use vehicle, be it autonomous or chauffeured, adds a few additional empty vehicle-kilometres travelled with the associated emissions. If the transfer is from public transport the increase in emissions and congestion will be greater. Only ride-sharing services offer the potential to reduce the environmental impact of travel. This conundrum makes the development of modelling tools more urgent to explore which new modes should be supported, and how they should be regulated to align individual, operator and societal objectives.

Attempting to model these aspects of future of mobility with aggregate models can only be done as a very coarse approximation. A serious treatment of these features practically requires an agent-based micro-simulation approach tracking individual vehicles and trips. Such models seek how best to attend requests at specific locations, not centroid connectors. This has been modelled, for example, by the International Transport Forum, modelling ride-sharing and single use MaaS for Lisbon, Auckland, Dublin, Helsinki and Lyon[11] (ITF, 2020). To populate their agent-based models the ITF expanded a sample of the population from household travel surveys creating a synthetic population with specific (time and location) requests for trips; the ITF team then modelled the supply of ride-sharing services to serve them. These simulations have consistently shown the significant advantages to society of ride-sharing services complementing mass rapid transit services.

Agent based modelling

The basic outcome of processing mobile phone records is a set of activity-travel diaries for a real population. The next step would be to set up a full artificial population that mimics the total population under study. To generate this, one can start from the activity-travel diaries of the mobile phone users and use census data to expand the sample by means of classical synthetic population techniques. In that way agents are created to represent the total population each with attributes like age, gender, income and the need to perform certain activities and trips. In many projects these individual trips are later on aggregated to generate origin-destination matrices. However, direct use of synthetic individual activity-travel diaries opens the door to more interesting modelling approaches.

The idea is that the timing and (x,y) co-ordinates of each activity in each individual diary of the total synthetic population are randomly allocated within the coverage area of the corresponding cell or zone based on probabilities that take into account land use data, the population density of the different census tracts intersecting that particular cell, etc. This should not be considered as personal information anymore, as no travel diary corresponds to the actual profile and the exact co-ordinates of any real person.

An obvious limitation is the number of different activities that mobile phone data can identify. The most important (and recurrent) ones of Home, Work, Study are easy but others are not; therefore, opportunities for combining other activities in new ways in the future will be limited. However, new complementary data sources (for example weblogs[12] and pico-cell data [13]) are expanding the range of activities that can be identified and therefore this limitation may be relaxed in the future.

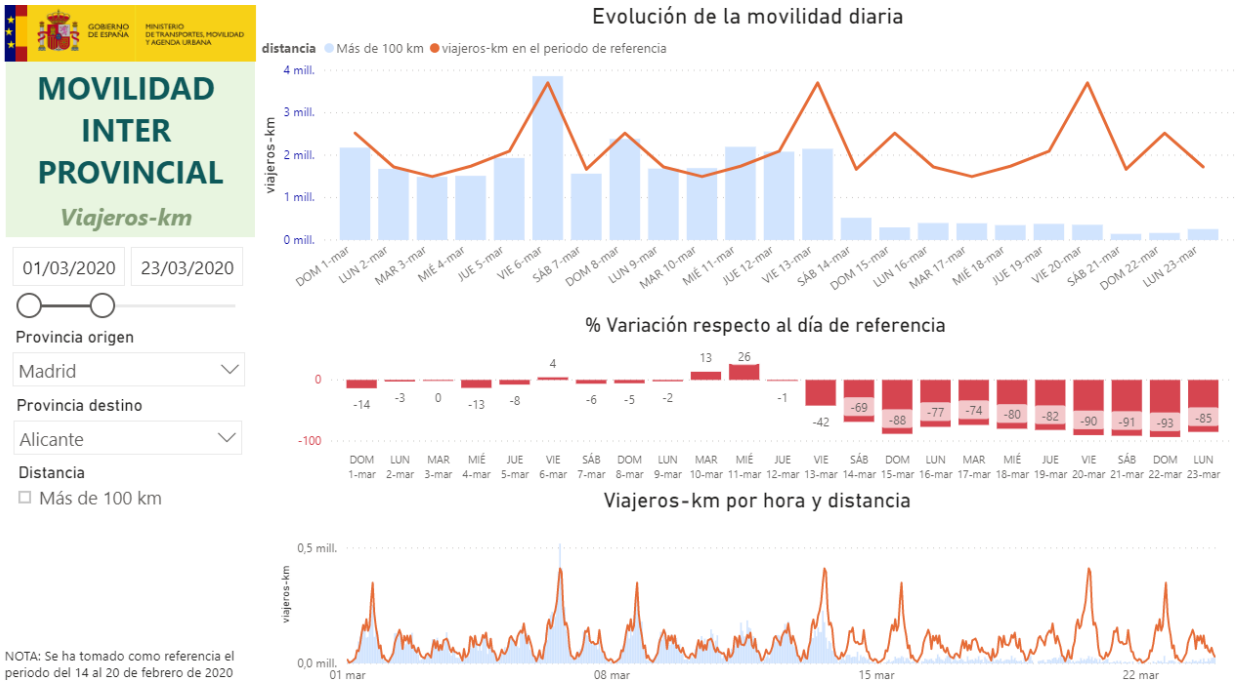
On the other hand, there are reasons to doubt the value of distinguishing many activities. First, only a few of them can be considered recurrent, the most likely being Work and Study. Interestingly, the 2020 pandemic has shown how many of these recurrent activities may be performed remotely thus reducing the need to travel. Second, many activities do not remain the same over the years. Examples that come to mind are video rentals and supermarket shopping. These two characteristics of activities put in question to what extent we can reliably forecast them into future years.

Learning from quasi-experiments

Some knowledge domains benefit from opportunities to do experiments and learn from them, for example agriculture or metallurgy. Because of the cost and disruption of experimentation this is seldom possible in the field of transport. Given that mobile phone data provides a continuous stream of data to monitor movements this creates an interesting opportunity to learn from experience, in particular from planned and un-planned quasi-experiments. For example, monitoring what people do when a metro line is closed for a period should be illuminating in terms of preferences and adaptation of travel patterns. Bad weather could disrupt travel and force people to change routes, modes and even destinations. Research on these adaptations could result in improved and empirically supported choice models.

This advantage became prominent during the 2020 pandemic. Authorities in different parts of the world have used big data sources to monitor how people move under different levels of restrictions and enforcement. For example, Spain has been tracking these movements using mobile phone data provided by Nommon[14] (Figure 10).

Figure 10. Mobility monitoring service by the Spanish Ministry of Transport



Source: Ministerio de Transportes, Movilidad y Agencia Urbana de España [Ministry of Transport, Mobility and Urban Agency of Spain].

Many cities are currently experimenting with pilot services using autonomous and semi-autonomous vehicles. They are a key testing ground for the technology but also a valuable source of data on user response to the characteristics of these new modes. Monitoring their take up and sustained adoption needs to consider other changes that may be happening at the same time. Mobile phone data can have a role in both identifying the evolution of mobility in areas with and without the intervention of the new mode. Carefully designed experiments of this nature can provide useful insights for policy development.

Mobile phone data thus provides new and rich openings for further research into travel behaviour and transport models strengthening the ground for practical applications and more reliable forecasting.

Adapting to change

A rapidly evolving future places additional demands on planners and decision makers. It requires a more rapid updating of plans and a more sensitive detection of changing trends and technologies. Moreover, the work environment is becoming more flexible and responsive to changes in behaviour and this creates more variable trip making patterns; this may well require the analysis of travel on different days rather than on the "average and neutral" one. The usual five to ten years between household travel surveys diminishes their value as a trend monitoring tool. The greater variety of mobility options, including micro-mobility and demand-responsive services, requires much larger samples than feasible with conventional data collection methods.

This context makes the case for passive data collection methods that rely on traces naturally generated as part of daily life. Anonymised mobile phone data with suitable measures to protect privacy and algorithms refined in practice appear to be an ideal source of data to address the challenges of a rapidly evolving mobility landscape.

Conclusions and recommendations

Big data, and in particular data from mobile network operators and from smartphone apps, offer a significant potential to enhance transport models and therefore support better decision making. This potential is not necessarily realised in practice. Not all data processing and data analytics generate high quality and useful trip matrices and mobility indicators.

The practical use of mobile phone data to generate trip matrices is already reasonably well established in many countries, including the USA, Spain and the United Kingdom, and is being adopted in other countries. However, guidelines on their specification and quality assurance are still under development and remain to be tested in large scale projects.

Mobile phone data also offers the opportunity to explore travel demand in ways impossible with conventional surveys. New insights should emanate from research into the recurrence of some trips, variability of demand between days and over time and the stability of activities. The disruption caused by the 2020 pandemic and the gradual introduction of new modes of travel should be monitored by governments; mobile phone data has a useful role to play in this task.

This paper suggests the following implications for policy development:

- Authorities should support investigations to establish the accuracy, realism and reliability of using big data to generate trip matrices and other mobility indicators
- Authorities should prepare and update, where necessary, modelling guidance to take advantage of new big data sources; these should provide guidelines to ensure regular validation of the product
- The specification and use of conventional surveys must be revised to make the best of the new data sources and reducing the burden of conventional surveys
- Countries could use big data to learn from quasi-experiments and better predict changes in travel behaviour
- Promising research on how best to exploit the opportunities opened by new big data sources include:
    - Optimising a combination of conventional and new data sources to improve transport model development and policy making
    - Development of new model forms to simulate the introduction of new mobility technologies and their implications for users, operators, citizens and the environment
    - Gaining a better understanding of the variability of travel demand day to day and over time using big data
    - Monitoring travel demand to gain insights into changing preferences
    - Exploring how big data can help understand uncertainty in travel demand forecasting and develop better decision making approaches.

Notes

1 An Activity Based Model is a model based on the microsimulation of agents representing each individual in the population and where the main focus is in their activities and the tours connecting them.

2 https://luca-d3.com/.

3 https://www.airsage.com/.

4 https://www.teralytics.net/.

5 https://www.nommon.es.

6 https://www.unacast.com/.

7 https://www.cuebiq.com/.

8 https://carto.com/.

9 https://www.streetlightdata.com/.

10 See for example https://wiki.openstreetmap.org/wiki/QuadTilesv.

11 https://www.itf-oecd.org/itf-work-shared-mobility.

12 Weblog is used in this context as a list of internet sites recently visited using the mobile phone; that may be used to characterise the user and provide better segmentation of the sample.

13 Picocells are small local cells, usually located in places with large number of people like shopping centres; as they are smaller than a conventional cell they provide a more accurate location of the device, especially indoors.

14 https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data.

References

Bassolas, A. et al. (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, March 2019, pp. 56-74, https://doi.org/10.1016/j.tra.2018.12.024.

BID (2019), "Como aplicar big data en la planificación del transporte urbano" [How to apply big data in urban transport planning], Nota Técnica No. IDB-TN-1773, Banco Interamericano de Desarrollo [Inter-American Development Bank], http://dx.doi.org/10.18235/0002009.

Bonnel, P., M. Fekih and Z. Smoreda (2018), "Origin-Destination estimation using mobile network probe data", *Transportation Research Procedia*, Vol. 32, pp. 69-81, https://doi.org/10.1016/j.trpro.2018.10.013.

Bonnel, P. and M. Munizaga (2018), "Transport survey methods – in the era of big data facing new and old challenges", *Transportation Research Procedia*, Vol. 32, pp. 1-15, https://doi.org/10.1016/j.trpro.2018.10.001.

Burrieza, J. et al. (2019), "Enhanced passenger characterisation through the fusion of mobile phone records and airport surveys", proceedings of the 9th SESAR Innovation Days, EUROCONTROL.

Cherrett, T., and M. McDonald (2002), "Traffic composition during the morning peak period: Implications for urban traffic management systems ", *European Journal of Transport and Infrastructure Research*, Vol. 2, No. 1, pp. 41 – 55, https://doi.org/10.18757/ejtir.2002.2.1.3675.

Del Mistro, R. and R. Behrens (2008), "How variable is variability in traffic? How can TDM succeed?", proceedings of the 27th South African Transport Conference (SATC 2008), 7-11 July 2008, Pretoria, https://www.gtkp.com/assets/uploads/20091129-122725-7589-Del%20Mistro.pdf.

FHWA (2016), *Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data*, Federal Highway Administration, Report FHWA-HEP—16-083, https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/origin-destination/fhwahep16083.pdf.

ITF (2020), "Shared Mobility Simulations for Lyon", *International Transport Forum Policy Papers*, No. 74, OECD Publishing, Paris, https://doi.org/10.1787/031951c3-en.

ITF (2015), "Big Data and Transport: Understanding and Assessing Options", *International Transport Forum Policy Papers*, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/5jlwvzdb6r47-en.

Montero Mercadé, L. et al. (2019), "Fusing mobile phone data with other data sources to generate input OD matrices for transport models", *Transportation Research Procedia*, Vol. 37, pp. 417-424, https://doi.org/10.1016/j.trpro.2018.12.211.

Picornell, M. et al. (2015), "Exploring the potential of phone call data to characterize the relationship between social network and travel behavior", *Transportation*, 42(4), pp. 647-668, https://doi.org/10.1007/s11116-015-9594-1.

Tolouei, R., S. Psarras and R. Prince (2016), "Origin-Destination trips matrix development: Conventional methods versus mobile phone data", *Transportation Research Procedia*, 26 pp. 39-52, https://doi.org/10.1016/j.trpro.2017.07.007.

Transport Systems Catapult (2016), "Utilising mobile network data for transport modelling: Recommendations paper", document reference: MV7/RPT001, available at: https://www.gov.uk/government/publications/mobile-phone-data-in-transport-modelling (accessed 10 November 2020).

International Transport Forum

Use of Big Data in Transport Modelling

This paper guides transport planners in making the best use of mobile phone traces, derived either from mobile network data or from smartphone app data. It suggests combining such new data sources with conventional travel surveys whose sample size and cost could ultimately be reduced. In the context of a rapidly evolving mobility landscape, with new modes and new services available, big data can help monitor behaviour change, learn from quasi-experiments and develop next-generation travel demand modelling tools.

All resources from the Roundtable on Use of Big Data in Transport Models are available at:
www.itf-oecd.org/big-data-transport-models-roundtable

International Transport Forum
2 rue André Pascal
F-75775 Paris Cedex 16
+33 (0)1 73 31 25 00
contact@itf-oecd.org
www.itf-oecd.org

OECD

Published: 01/2021 | Photo credit: Maria Savenko/Shutterstock