DEVELOPMENT OF A TRIP INFORMATION SYSTEM FOR HIGHWAYS ENGLAND USING TELEFONICA / O2 MOBILE PHONE DATA

Nicolae Duduta
Telefonica Digital, London UK

Timothy McHugh
Telefonica Digital, London UK
Corresponding author: timothy.mchugh@o2.com

Nick Corby
Highways England

Stephen Rutherford
Jacobs, London UK

2016 European Transport Conference

Barcelona, Spain, October 5 – 7, 2016

1 INTRODUCTION AND LITERATURE REVIEW

In this paper, we provide an overview of the development of the Trip Information System (TIS), a transport database for Great Britain developed jointly by Highways England, Telefonica and Jacobs. We provide an overview of some of the key aspects of the methodology, lessons learned, as well as some guidance for transport modellers and public agencies that will be using the data in the future. The paper is mainly addressed at future TIS users, describing the data formats and the types of insights that can be extracted from the database. We also discuss some methodological aspects of the processing of mobile phone data in TIS that represent improvements over previous research and should be of interest to transport professionals and researchers interested in the topic of mobile phone data.

The Trip Information System (TIS) is a transport database for Great Britain, estimating road and rail trips over a 12-month period, and generated from a combination of data sources available to Telefonica, including primarily mobile phone cell ID data, but also other data such Wi-Fi connections. It will be accessible to transport consultancies working on Highways England projects as well as to Highways England and UK Department for Transport (DfT) staff for use in strategic network modelling.

Mobile phone data has emerged in recent years as a promising data source for an array of sectors, from insurance to finance, out of home media to tourism and more. It has perhaps been most successful the transport sector, particularly for creating origin-destination (OD) matrices. There is a growing body of research exploring methodologies for building OD matrices from different types of mobile phone data, including call records and cell ID data (e.g. Friedrich et al. 2014, Bahoken and Olteanu Raimond 2013, etc.)

Most of the earlier research has been theoretical, showing the feasibility of mobile phone data (MPD) as a data source, and comparing the outputs to those of traditional data sources, showing that MPD-based OD matrices can result in models that validate well against other data sources (Colak et al. 2015).

In the past three years, mobile network providers in the United Kingdom have partnered with transport consultancies and local governments to carry out proofs-of-concept on the use of MPD for real world projects. Tolouei et al. (2015) report on one such partnership between Telefonica and AECOM for the development of a transport model for Leicestershire. The authors show that MPD-based transport models can achieve the level of validation against other data sources indicated by earlier research.

The Highways England Trip Information System (TIS) builds on the experiences from previous research and previous applications of the technology, and represents the first time that MPD has been used at the national scale, making this the broadest and most ambitious application of MPD to transport to date.

2 OVERVIEW OF THE TIS

TIS will be a database including all motorized road and rail trips made by residents of Great Britain and will cover all of England, Wales, and Scotland. It will be accessible via a web interface, and data will be provided in the form of trip matrices for a user specified model zone system, which can be uploaded through the interface.

Data in TIS will be provided at the level of Census geographies. Model zones can be any Census geography from entire regions or countries, down to Middle Layer Super Output Areas (MSOAs). While mobile data can be applied at various geographic levels (driven by need and usecase), MSOAs were chosen as the most appropriate level at which data should be provided based upon a detailed review of previous projects and research.. MSOAs are designed to have a relatively constant Census home population, which means that their size varies with population density. This is a good match with cellular data, since the range and density of cells also varies with population density (O_2 2016).

Trips made by residents of Northern Ireland and non-UK residents will not be included in the data. Personal travel, business travel, and freight will all be included in the data, but there is no specific identifier for freight trips. Users will request data from TIS via a web interface, and will upload project zoning systems along with request for data to be processed according to various parameters, such as mode of transport (motorised road or rail), trip purpose (work and other), distinctions between home-based (HB) and non-home-based (NHB) trips.

TIS offers two possible definitions for trip origins and destinations. These can be defined in terms of the start and end point of the main mode of transport for the trip – this would be station to station in the case of a rail matrix, excluding the rail legs of domestic air trips. Alternatively, the user can opt to include station and airport legs. This option would give the actual origins and destinations of rail trips rather than the stations. A road trip matrix with this option would include road trips to stations and the road legs of domestic air trips as road trips in their own right.

In addition to a ‘full matrix’, TIS users will also be able to obtain a ‘route matrix’ for motorised road trips, which will feature all trips passing through a sequence of nodes on the road network as specified by the user.

3 TIS TRIP PROCESSING METHODOLOGY

3.1 DATA SOURCES

3.1.1 MOBILE PHONE CELL ID DATA

The main data source used in the development of TIS is mobile phone cell ID data. Telefonica has a market share of approximately 30% (Statistica 2015) among mobile providers in the United Kingdom, where it trades under the O_2 brand name and also has mobile virtual network operators (MVNO) Tesco Mobile and GiffGaff as partners, broadening its total user base to approximately 20 million users in total. Users generate data events whenever the phones connect to the cellular network and the data is probed by Telefonica in real time. Connections can be active, when the user is actively using the phone (i.e. calls, text messages) or passive, when the location of the phone is automatically checked or updated by the cellular network, or in a number of other situations, such as regaining signal or switching between 2G, 3G and 4G cells.

Each connection event includes the following information – an anonymized user ID, a cell ID, and a timestamp. The anonymized user ID disguises the actual user ID to protect customer privacy, but it is persistent over time, which allows for in-depth analysis of mobility patterns over days and months, helping to generate insights into the users' home and work locations. A cell is a fixed-location transceiver, mounted on a cell tower. A single tower will typically have multiple cells, to cover different areas around the tower and different technologies (2G, 3G, 4G). The cell ID is linked to the O_2 cell map, which provides the location of the cell tower and the geometry of the cell's coverage area.

It is important to note that a user ID in Telefonica's database is linked to a SIM card, and therefore identifies a mobile phone or a machine-to-machine (M2M) device, not a person. There were 91.5 million mobile subscriptions (including M2M devices) in the UK at the end of 2015 (Ofcom 2016), over 40% more than the resident population of the UK (ONS 2016). Therefore, one of the challenges in building OD matrices from mobile phone data is the conversion from user IDs to persons.

The TIS user base filters out any user ID associated with a device that is not a mobile phone, which solves the issue of M2M devices (or indeed tablets etc). Another important issue to address is multiple mobile phone ownership. This is addressed in TIS by removing user IDs associated with business phones from the sample, and keeping only personal phones. This should help address most of the problem of multiple phone ownership, since the most likely cause of this would be someone owning both a personal and a business mobile phone. There is another potential source of bias which cannot be addressed with the data available in TIS – persons owning multiple personal phones, all on O_2. Because user IDs relate to SIM cards not actual customers, and because they are anonymized to protect the privacy of O_2 customers, the extent to which this may be an issue in the sample is unknown. The TIS verification tests, discussed briefly at the end of this paper and in more detail in Corby et al. (2016) evaluated trip rates in TIS and did not find any bias resulting from potential multiple phone ownership.

3.1.2 WI-FI DATA

While cellular data provides an excellent picture of macro flow of population, one of the limitations relates to shorter trips, or trips which exist where no events occur. This is especially relevant in urban contexts, especially in London, with a complex mix of rail services including Underground and various national rail services. To overcome this challenge, TIS makes use of an additional data source – O_2 Wi-Fi. In addition to being a mobile network operator, O_2 also offers Wi-Fi connectivity to private and public sector clients, and there O_2 Wi-Fi access points in rail stations and retail locations across London. A connection to a Wi-Fi access point generates a similar type of data event to a connection to a macro cell tower (anonymized user ID, device ID, and time stamp), but the location accuracy is greatly improved compared to the macro network.

Whilst a macro cell tower might place an event within a radius of a few hundred meters from a rail station, a Wi-Fi access point will reliably place the event inside a rail station or retail venue. Not every O_2 user will connect to O_2 Wi-Fi or will be detected by a Wi-Fi device. Only those users who have previously registered for O_2 Wi-Fi and those with EAP-SIM enabled phones (typically newer smartphone models) will be detected. This means that although Wi-Fi provides greater locational accuracy, the O_2 Wi-Fi user base is a subset of the O_2 user base and this requires careful consideration as part of the expansion process.

3.1.3 CUSTOMER PRIVACY

Customer privacy is of utmost importance to Telefonica and O_2, and there is a very strict set of rules in place to ensure that the privacy of customers is fully protected throughout the data processing steps, and that no personally identifiable information is ever made public. The data from both the macro network and Wi-Fi is anonymized before it is processed for TIS and the outputs provided to third parties through TIS are always aggregated and averaged out across a minimum of 20 days, to ensure the data represents the movements of populations and that there is no information in the outputs regarding specific individuals.

3.2 IDENTIFICATION OF TRIP ORIGINS AND DESTINATIONS

In the first instance, the combinations of cell ID and timestamps are analysed in order to convert the data into *trips* and *dwells*. A dwell (or OD) is defined as any location where the user has stopped for at least 10 minutes, and it features a start and end time. A trip is defined as the movement connecting two consecutive dwells. Trips and dwells are created by an algorithm that analyses the difference in timestamps between the events and the IDs and positions of the different cells associated with those events.

The simplest example of a dwell is a user generating two consecutive events with the same cell, with a time difference between those events exceeding the acceptable travel time across that cell. Most cases will be considerably more complex, however. The coverage area of cells across the network overlap, partly to provide coverage for different technologies (2G, 3G and 4G) and partly to provide redundancy (i.e. to maintain coverage to an area even if there is an issue with a specific cell). As a result, a user that is not moving may connect to more than one cell. Telefonica implements a separate process to identify "flickering cells", i.e. cells where multiple handovers are observed over short time intervals. The algorithm that identifies trips is designed to discount any movement between cells that flicker. The advantage of this approach is an increased confidence that the trips which are detected do represent actual movement. The downside is that discounting flickers makes it more difficult to identify short trips. This is an important limitation of cell ID technology and the implications for transport modellers are discussed further below.

TIS also addresses the difference between the dwell / trip logic of mobile phone event data and the equivalent definitions used in transport modelling. In particular, not every dwell observed for a mobile phone can be considered as an origin or destination for transport modelling purposes. For instance, a user transferring from the London Underground to National Rail can generate a dwell at the National Rail station, but that dwell is in fact only a transfer as part of a longer trip. Similarly, a long distance road trip can include multiple stops at motorway service areas, which may or may generate dwells in the data, but which would never correspond to the true destination of the trip. Conversely, a user may stop briefly at a shop on the way home from work, and that stop may be too short to generate data events from which a dwell could be inferred. However, for a transport modeller, the visit to the shop is a separate trip which could potentially have involved a diversion from the route that would otherwise be taken to home.

TIS applies several additional rules to enhance the dwell / trip logic, and split or merge trips where necessary in order for the resulting dataset to better meet the needs of transport modellers. For instance:

- Merging of rail trips: whenever a user is observed to make two consecutive National Rail trips with a time of less than sixty minutes between them, the dwell between them is assumed to be a transfer, and the two trips are merged into a single one
- Inferring origins and destinations for motorized road trips: In some cases, a user can be observed to generate two consecutive events with timestamps indicating an exceedingly long travel time between those cells. It is likely that the user stopped somewhere in between, but did not generate any events to indicate a dwell. The TIS methodology identifies these cases, and then analyses other locations that the user typically visits in the area. If a nearby location is found that the user has visited before, and the travel times to and from that location, plus a minimum dwell time, would explain the difference between the observed difference in timestamps between the two cells, then the user's trip is broken at that location.
- Avoiding the creation of dwells at motorway service areas: When applying the above rule, the locations of dwells for certain long distance trips are cross-referenced against the locations of all motorway and A road service areas in Great Britain, and if there is a match, the dwell is discarded; this helps ensure that long distance trips are better represented in the database

3.3 IDENTIFICATION OF TRIP MODE

After a trip has been identified, along with an estimated start and end time, the cell ID and timestamp of data events that occurred during that trip can offer insights into the mode of transport used. The coverage areas of cells can provide valuable information here. Some cells do not cover any railway lines, and any trip that generates events only with these cells, can be allocated to road. Conversely, if a user has multiple connections to cells that cover railways lines within the same trip and also has a start and / or end location of the trip in close proximity to rail stations, this is indicative of rail as the primary mode of transport. Domestic air trips can also be detected because phones are typically switched off during flights. A sequence of events that has two consecutive connections at two airports is a good indicator of an air trip.

As discussed above, a methodology relying solely upon data from the macro cellular network will have difficulties identifying the correct mode of transport for very short rail trips. This can be particularly challenging in central London, where most trips will start and end in areas that are well served by the rail network (even more challenging when the underground is considered), and may be too short to generate enough macro events to allow for a thorough evaluation of transport mode. To overcome this challenge, TIS leverages both macro cellular events and O_2 Wi-Fi data. There are O_2 Wi-Fi access points at most rail stations within London, and our methodology cross-references events from the macro cell network with connections to Wi-Fi access points to identify rail trips. As O_2 users travel through the London rail network, those users with EAP-SIM enabled phones and registered O_2 Wi-Fi users will generate sequences of connections to access points in different rail stations along the way. These sequences of connections are then cross-referenced with the trip data from the macro network, and the corresponding macro network trip is assigned rail as the main transport mode.

TIS does not rely solely only on event data to determine mode of transport, but also uses insights from other data sources, such as the National Travel Survey (NTS). According to the NTS, the vast majority of rail trips are more than 5 miles long, and of those rail trips that are less than 5 miles, most are in London (NTS 2014). This type of independent information can be quite valuable when evaluating the mode of transport of a trip that is both under 5 miles and occurring outside of London, since there is a very strong probability that the trip is road.

3.4 IDENTIFICATION OF HOME AND WORK LOCATIONS

The use of anonymized user IDs over time makes it possible to analyse historical behaviour and extract more insights than would be possible from just analysing individual trips and dwells. In the context of TIS, the main insight derived from historical patterns is a list of points of interest (POIs) per user. A POI is defined as a unique location where a user has dwelled once or multiple times in the study period. A POI can have a single dwell, if it is a location that the user only visited once, or multiple dwells, if it is a very frequently visited location.

POIs are useful because they can be classified based on the types of dwells they include. For instance, it is typical when analysing mobile phone data to use the POI with the most overnight dwells to represent a user's home location. For instance, Colak et al. (2015) define home as the most frequently visited location between 8pm and 7am on weekdays and anytime on weekends. The challenge with this approach is that it may not be a correct representation of home location for people who work at night. To overcome this challenge, TIS uses a different definition of home, looking at the POI with the highest number of long dwells (defined as dwells longer than six hours). For over 90% of people, the two methods will give identical results, but for a small percentage, the TIS method suggests a different location as home. It is difficult to evaluate the relative accuracy of the two methods, since the user data is anonymized and there is no ground truth against which to test, such as billing addresses. However, tests carried out at an aggregate level suggest that the 6-hour definition results in a ratio of home-based (HB) to non-home-based (NHB) trips that better reflects those found in other data sources (e.g. NTS), compared to the overnight definition.

A similar logic can be applied to identify work locations. However, the ability to accurately detect workplaces from mobile phone data is slightly lower than the ability to detect home. The tools available for POI classification are the start and end times of dwells, the count of dwells, and the types of days on which they most frequently occur. Whatever rule is created based on these parameters (e.g. daily dwells between 9am to 5pm on business days), there will always be some users with locations that meet the criteria without actually having a workplace, and also users with different work arrangements whose locations will not qualify. TIS aims to address these limitations based on the premise that anyone spending more than three hours a day in a single location (other than home) on three or more days per week over a six-week period is likely to be working at that location, or otherwise is in full-time education. Education can be filtered out by looking at which users have, for example, the same work POI in May and October in a given year, but a different work POI in August of that same year.

Overall, the TIS home and work location rules represent a refinement over previous definitions used in the literature, as they are able to handle more complexity in the users' work arrangements, including night-shift and weekend workers. Preliminary test results using data for March 2015 (a month that will not be featured in the actual TIS database) show promising results. As Figures 1 and 2 show, home-based trips and home-based work trips correlate well against Census home and work populations.

3.5 IDENTIFICATION OF TRIP PURPOSE

TIS users will have the opportunity to select OD matrices segmented by trip purpose. Purpose is inferred from the types of POIs that each trip connects. Each user included in the database will have a home POI, and some users will also have a work POI, if they have a location that meets the criteria. All remaining POIs for each user will be labelled "other" POIs. Using these three POI types, the following trip purpose and directions can be identified in the data:

- Home-based work (HBW) outbound: a trip from the home to the work POI
- Home-based work (HBW) inbound: a trip from the work to the home POI
- Home-based other (HBO) outbound: a trip from the home POI to another POI that is not the work POI
- Home-based other (HBO) inbound: a trip from a POI that is neither home nor work to the home POI
- Non-home-based (NHB): a trip that connects two POIs that are not the home POI

3.6 TREATMENT OF MULTI-MODAL TRIPS

The mode of a multi-modal trip in TIS is defined as that mode which accounts for the greatest proportion of total distance travelled (i.e. the ‘main mode’). A matrix of motorized road trips will therefore include only those multi-modal trips which have a main mode of motorized road. Typically therefore, a motorized road trip involving walk would be included, whereas a motorized road trip including rail and air would not.

There is also the option of including all motorized road or rail trips in a matrix, irrespective of whether or not this is the main mode of travel, For example, in the case of a motorized road matrix, the user could include the motorized road stages or ‘legs’ of rail and domestic air trips in the matrix. In relation to airport legs, it should be noted that trips to and from airports for international travel are not considered multi-modal trips since the whole of the trip is not within Great Britain and TIS relates to trips within Great Britain only. A trip from home to the airport by motorized road or rail for purposes of international travel is therefore considered as a single mode trip and would appear in the matrix.

If selecting rail trips by main mode only, users will receive an OD matrix where the origins and destinations correspond to the start and end stations of the rail trips (excluding the rail legs of domestic air trips). If the user also selects the ‘include station and airport legs’ option, then the origins and destinations will instead correspond to the actual trip ends. Whilst this changes the ODs rather than the number of trips in the matrix, there will be a small increase in trip numbers resulting from the inclusion of the rail legs of domestic air trips.

The choice of rail station is determined by an algorithm that considers the proximity of different rail stations to the start and end of the trip, but also which stations serve which routes. For longer distance rail trips, the algorithm will consider which rail station serves the destination from the origin city, and will consider that as part of the allocation process.

In the case of road trips, the selection of station and airport legs in the OD matrix will add the motorised road legs of rail and domestic air trips. These will appear as additional road trips when compared to a main mode only matrix. Some road legs may not be included if the TIS mode detection algorithm has determined that those legs are more likely to have been made via a non-motorised mode.

3.7 USER BASE AND EXPANSION

One of the challenges in using the customer base of a mobile network operator as a population sample is that the customer bases changes constantly as uses join or leave the network. From the perspective of TIS, this poses the problem that if a user is not present for a sufficient amount of time in the customer base, there may not be sufficient observations of that user's dwells over time in order to determine their home location. Home detection is crucial because the user base is expanded at MSOA level based on the place of residence.

Therefore, TIS defines, for each month in the study period, a list of valid users, defined as those users that have been generating events with the cellular network on at least ten distinct days during the study period. A further refinement is then applied, requiring that there be at least sixteen days between the timestamp of the first and last connection to the cellular network on that month, which should help to some extent eliminate tourists using a UK phone number.

Expansion of the TIS user base is undertaken on a daily basis. This involves allocating an expansion factor to each user which is applied to all of the trips made by that user on the day in question. As discussed above, the use of Wi-Fi data complicates the expansion process, because the O_2 Wi-Fi user base is a subset of the O_2 customer base. TIS addresses this by using a separate expansion process for Wi-Fi and non-Wi-Fi users. It should also be noted that expansion is undertaken separately for different age groups since it is known that mobile phone ownership rates differ by age group.

Another concept included in the TIS methodology is that of a 'reliability grading'. The movements of every user are analysed every day with the goal of determining all of their dwells and the mode of travel between them. Where this is possible, the user receives a reliability grading of 1 for that day and is included in the daily expansion base. Otherwise, the user is graded 2 for that day and excluded. A user may therefore be graded 1 on one particular day and 2 on another.

4 COMPARISON OF TIS DATA WITH INDEPENDENT DATA

TIS data was subjected to a rigorous assessment through comparing aggregations of TIS trip data with equivalent numbers from other transport data sources, including the Census, Census Journey to Work (JTW), the National Travel Survey (NTS), the National Trip End Model (NTEM), Transport for London (TfL), and Office of Rail and Road (ORR) statistics. The full results of these tests and a discussion of the findings can be found in Corby et al. (2016). Tolouei et al. (2015) offer another example of in-depth validation of MPD-based OD matrices, and whilst the data source is the same (O_2 customer cell ID data), their methodology differs slightly from that of TIS.

In this section, we will present a brief overview of preliminary findings from the TIS testing process, using data from March 2015. These results should be interpreted with caution, since the actual TIS data will feature calendar year 2016 and actual test results based on the final dataset may vary. Nevertheless, the results in this section should provide a good overview of how TIS data should generally be expected to compare against other data sources, and which are the areas that may require additional work from transport modellers.

In general, TIS data was found to validate well against other data sources for motorized trip rates and also for the correlation of home-based outbound trips against Census home populations, where R^2 values of 0.7 or higher were observed for groups of similar MSOAs (Figure 1).

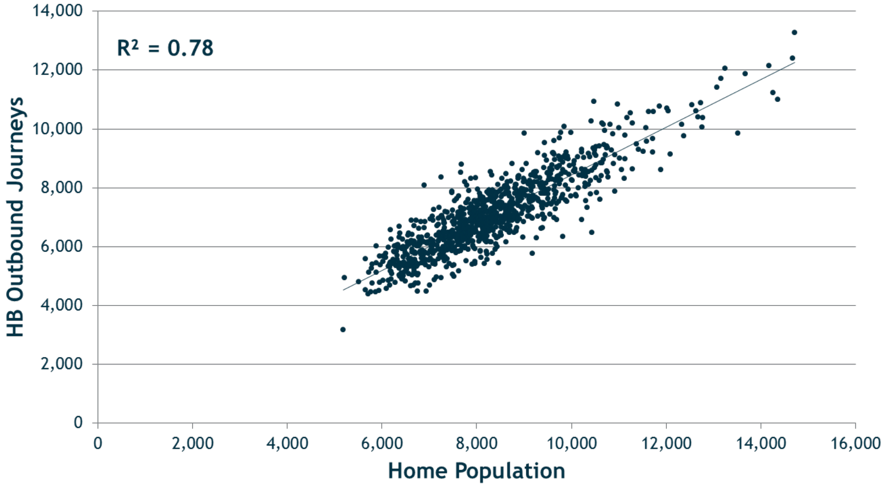

Figure 1: Correlation between Census home population and home-based outbound trips, by MSOA (Greater London only).

Despite the limitations in the ability to detect work locations, home-based work (HBW) trips were found to correlate well against Census Journey to Work (JTW) data. As Figure 2 shows, the correlation between motorized HBW trips in TIS and Census JTW was high, with an R^2 value of 0.95. The high correlation is driven by the major outlier that is the City of London, but even when excluding MSOAs with Census work populations over 50,000, the R^2 value remains above 0.8.

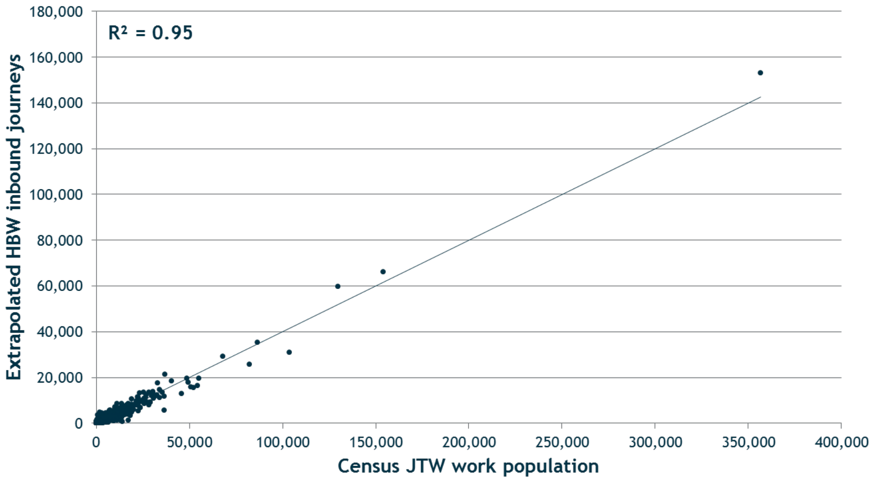

Figure 2: Correlation between Census work populations derived from Census Journey to Work (JTW) data, and origins of home-based work (HBW) inbound trips, by destination MSOA.

The data was also found to be internally consistent and symmetric, with a difference of less than 3 – 5% between home-based outbound and home-based inbound trips over a four-week study period (Figure 3).

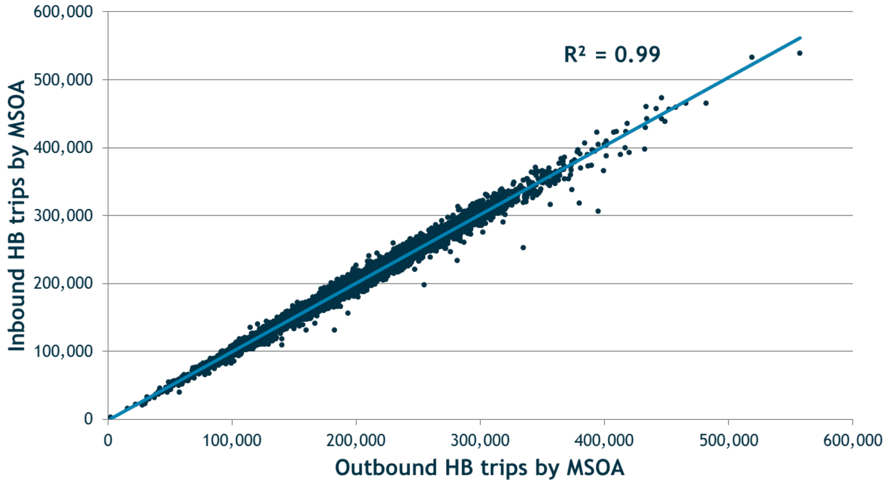

Figure 3: Outbound versus inbound home-based trips by MSOA

The percentage of non-home-based (NHB) trips in TIS was typically above 25% across the UK, slightly higher than what would be expected from other data sources (such as NTS). This is partly due to the fact that TIS, unlike NTS, includes freight trips and other employer's business trips in commercial or specially adapted vehicles. Another factor is that TIS includes all trips undertaken for the same purpose in a trip chain as separate trips, whereas NTS does not. For example, a home based leisure or shopping trip involving visits to various locations (before returning home) would be considered as an outbound and return home based trip in NTS. Within

TIS, each detected shopping or leisure location in the trip chain would generate an additional NHB trip. Home-based work (HBW) trips, on the other hand, validate relatively well against other data sources, despite the limitations described above, both in terms of total volume of trips (HBW represent around 20% of total trips in TIS) and MSOA-level correlation against Census Journey to Work (JTW) data (Figure 2).

5 CONVERTING TIS MATRICES TO MODEL MATRICES

The trip data in TIS is derived from mobile phone cell id data. This data is limited in terms of the information it directly provides and what can reliably be inferred from that information. Whilst the TIS methodology leverages as much information as possible, this still leaves a deficit in terms of what is required for transport modelling purposes. In particular, the mobile matrices extracted from TIS will require further work in order to convert them to model matrices. The following paragraphs describe the limitations in cell id data and the issues which will need to be addressed as a consequence of these limitations when converting from mobile to model matrices.

As mentioned earlier, shorter distance trips are more difficult to identify from cell id data. This is related to the fact that user locations are identified through the ID of the cell that a phone connects to. Each cell has a range and if the trip is shorter than the range of the cell that the phone is connected to i.e. an intra-cellular trip, then the trip will be invisible to the mobile network (the phone will not appear to have moved). In an urban area, where the cell density is higher and cell ranges lower, short walking trips will generally not be picked up (e.g. a 50-meter walk from an office building to a local coffee shop. In rural areas, where there are considerably fewer cells with longer ranges (O_2 2016), the minimum distance that a user must travel before being picked up by the technology will be correspondingly longer.

Unlike intra-cellular trips, short distance inter-cellular trips are potentially detectable. They are however still less likely to be detected than longer distance trips. The reason for this is that cell events are less likely to be generated because of the following characteristics of mobile networks and user behaviour:

- Shorter distance trips are of short duration and often involve travel to locations where the duration of stay is also short. In such cases, the user is less likely to generate events such as calls or texts and there is little likelihood of a periodic location update occurring. Even though a trip may therefore be inter-cellular, it could appear from the absence of events that the user has not moved. An example of this could be a parent leaving home to escort a child to school and then returning home.
- A short distance trip is confined to a smaller spatial area. This makes it less likely that a passive event will be generated when crossing a network location or routing area boundary.

Even where sufficient events are generated to indicate that a user has in fact moved location, there will not necessarily be sufficient events occurring at the destination to indicate that a dwell has occurred at that location. For example, a trip from home to shop and home could generate an event en-route only, or just a single event at the shop. Or a trip chain starting/ending at home with several destinations may only generate sufficient events at one of those destinations. This limitation is addressed within TIS to a large extent because algorithms are included which seek to identify the unobserved trip ends. Where trip ends cannot be identified, a user will be excluded from the expansion base for the day in question on the basis that their trip records are incomplete. However, the act of removing trip chains which should be broken but cannot be will potentially leave some bias in the trip length distribution.

Assuming the data is correct in all other respects, the limitations described above will mean that a trip length distribution from TIS will differ from that observed in surveys such as the NTS. The TIS distribution will have a higher proportion of longer distance trips and a correspondingly lower proportion of short distance trips, even after accounting for the absence of freight from NTS. There are several other limitations of TIS data that modellers should be aware of and will need to address when converting TIS matrices to model matrices:

- Whilst it is possible, to a large extent, to split journeys into rail and motorized road, the TIS methodology cannot distinguish between different vehicle types for motorized road. It will therefore be necessary for modellers using TIS data to convert motorized road to different types of vehicles.
- Although TIS can identify London Underground trips explicitly, (approximately 50% of rail trips in Great Britain), it cannot distinguish light rail and tram trips (4% of rail trips) from motorized road trips. Light rail and tram trips will therefore be included as motorized road trips. Modellers with an interest in urban areas outside Greater London should be aware of this limitation and adjust the data accordingly.
- Finally, since trip purpose is derived from the types of POIs connected by a trip, it is not possible to identify employer's business as a separate category. Employer's business trips will likely be split among the home-based other and non-home based categories, and modellers will need to use other data sources to identify this trip purpose.

Whilst it is important for TIS users to understand and address the limitations listed above, TIS data will offer a significant improvement over traditional data sources used in transport modelling. The large sample of users and the reliance on observed population movements, filtered through a robust processing algorithm will help ensure that transport modellers have access to a more reliable and consistent data sources in the development of transport models. Furthermore, the lessons learned from the development of TIS and the use of its data in the transport community will contribute to a much better understanding of mobile phone data and how it can help meet the needs of transport modellers.

ACKNOWLEDGMENTS

The authors would like to thank the following individuals for their valuable contributions to the development of the Trip Information System and for their valuable feedback and review: Vipin Chadha, Vincent Quantmeyer, Anne Griffith, Dave Sweeney, Kin Pang, Tom Brealey, Shaleen Srivastava, Boon Sey, Antonio Nunes, Graeme Dodds, Susanne Afra, Adam Czopp, Mohammad Mamajiwala, Daniel Ramirez Cano, Alex Petrie, Liam Birkett, Linda Lilly Salvesen, Gary Pang, Joaquin Guanter Gonzalbez, Csaba Kellen, Simon Knepper.

REFERENCES

Bahoken, F., A-M. Olteanu Raimond. 2013. *Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths. The effect of spatiotemporal filtering on flow measurement*, Selected proceedings of the 26th International Cartographic Conference.

Colak, S., L.P. Alexander, B. Alvim, S. Mehndiratta, M. C Gonzalez. 2015. *Analyzing Cell Phone Location Data for Urban Travel: current methods, limitations, and opportunities.* Transportation Research Record 2526, Journal of the Transportation Research Board.

Corby, N., S. Rutherford, T. McHugh. 2016. *Highways England Trip Information System Verification,* paper submitted to the 2016 European Transport Conference, Barcelona, Spain.

Friedrich, M., K. Immisch, P. Jehlicka, T. Otterstatter, J. Schlaich. 2014. *Generating Origin – Destination Matrices from Mobile Phone Trajectories.* Transportation Research Record 2196, Journal of the Transportation Research Board.

Iqbal, Md. S., C. Choudhury, P. Wang, M. Gonzalez. 2014. *Development of Origin-Destination Matrices Using Mobile Phone Call Data,* Transportation Research Part C, 40 (2014).

National Travel Survey (NTS). 2014. *Information and Statistics on the National Travel Survey.* Retrieved online from: https://www.gov.uk/government/collections/national-travel-survey-statistics

O_2. 2016. *Check coverage and network status.* Retrieved online from: http://www.o2.co.uk/coveragechecker

Office for National Statistics (ONS). 2016. *Population Estimates.* Retrieved online from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

Office of Communications (Ofcom). 2016. *Facts and figures. Landline / Mobile. Proportion of adults who personally own / use a mobile phone.* Retrieved online from: http://media.ofcom.org.uk/facts/

Statistica. 2015. *Market Share Held by Mobile Phone Operators in the United Kingdom*. Retrieved online from: http://www.statista.com/statistics/375986/market-share-held-by-mobile-phone-operators-united-kingdom-uk/

Tolouei, R., P. Alvarez-Indave, N. Duduta. 2015. *Developing and Verifying Origin-Destination Matrices Using Mobile Phone Data. The LLITM Case,* Proceedings of the 2015 European Transport Conference, Frankfurt.