O2
business

Mobile phone data provision

AECOM – Transport Scotland Origin Destination Demand

November 2019

Contents

Introduction & Project Scope

Introduction

O2 is a mobile network operator providing telephony services to over 22 million UK customers in both the public and private sectors. To ensure this service, O2 operates a network which supplies continuous nationwide coverage to each customer phone (device). The network and phone are in frequent communication to provide service. Intimate understanding of these networks allows O2 to build a contextual understanding of the movement of devices in space and time in the real world, with each phone creating events at specific points in time and space. These are chained into ‘breadcrumbs’, demonstrating whether each phone is moving or stationary at any point in time.

The result of O2’s processing creates a vast and valuable dataset which describes the movement and flow of O2 users across the UK. We track devices anonymously and associate each device with attributes derived from the user’s contract (age, gender, contract type and billing address) or their observed behaviour (affluence, lifestyle, home and work location and other points of interest). In aggregate, therefore, mobile phone data provides a robust insight into the movement patterns of the UK population.

Given the nature of mobile phone data, it can represent movements on a macro basis across larger areas effectively. The technology is generally better at identifying longer trips and those where the user dwells at their destination for a more extended period. For this reason, the data should not be used in isolation but combined with other data sources before application.

Customer privacy is of utmost importance to O2. All events processed are by-products of the core telephony network, and the process does not affect any user’s handset. We anonymise the records before storing them in the analysis platform, so all analysis of behaviour is done in a completely anonymous separate environment. We aggregate outputs from the analysis such that we do not provide any individual-level data to clients.

Scope

O2 was requested by AECOM to prepare origin-destination matrices for travel in the Scotland area. We included trips if they penetrated a cordon, as shown in Figure 1.

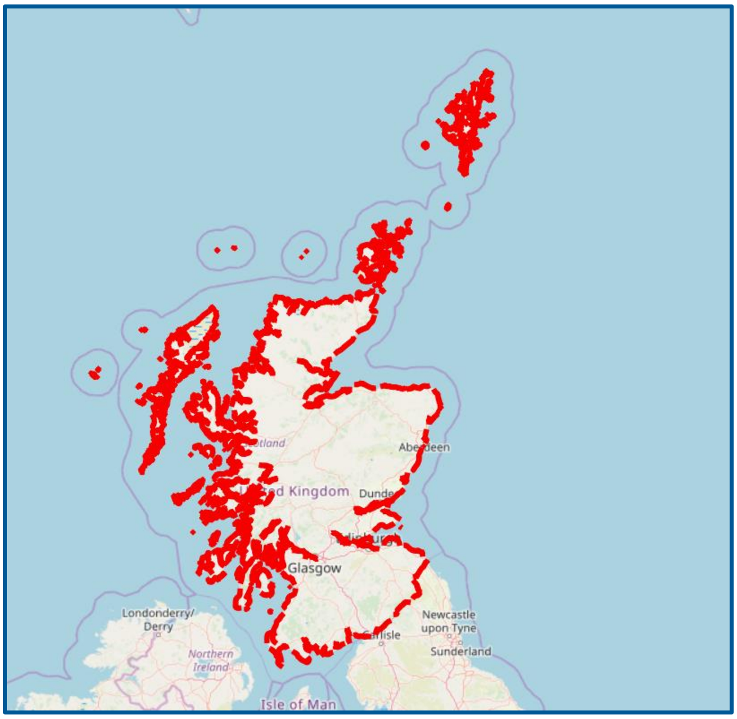


Figure 1: Image showing the extent of the model cordon.

We allocated trips to a start and end zone based on a zone system agreed with AECOM, featuring a total of 525 zones. AECOM provided the zoning system disaggregated in the following way:

- Cordon area (508 zones)
- Outer zones (17 zones)

Figure 2 shows the full zoning system, including the cordon area.

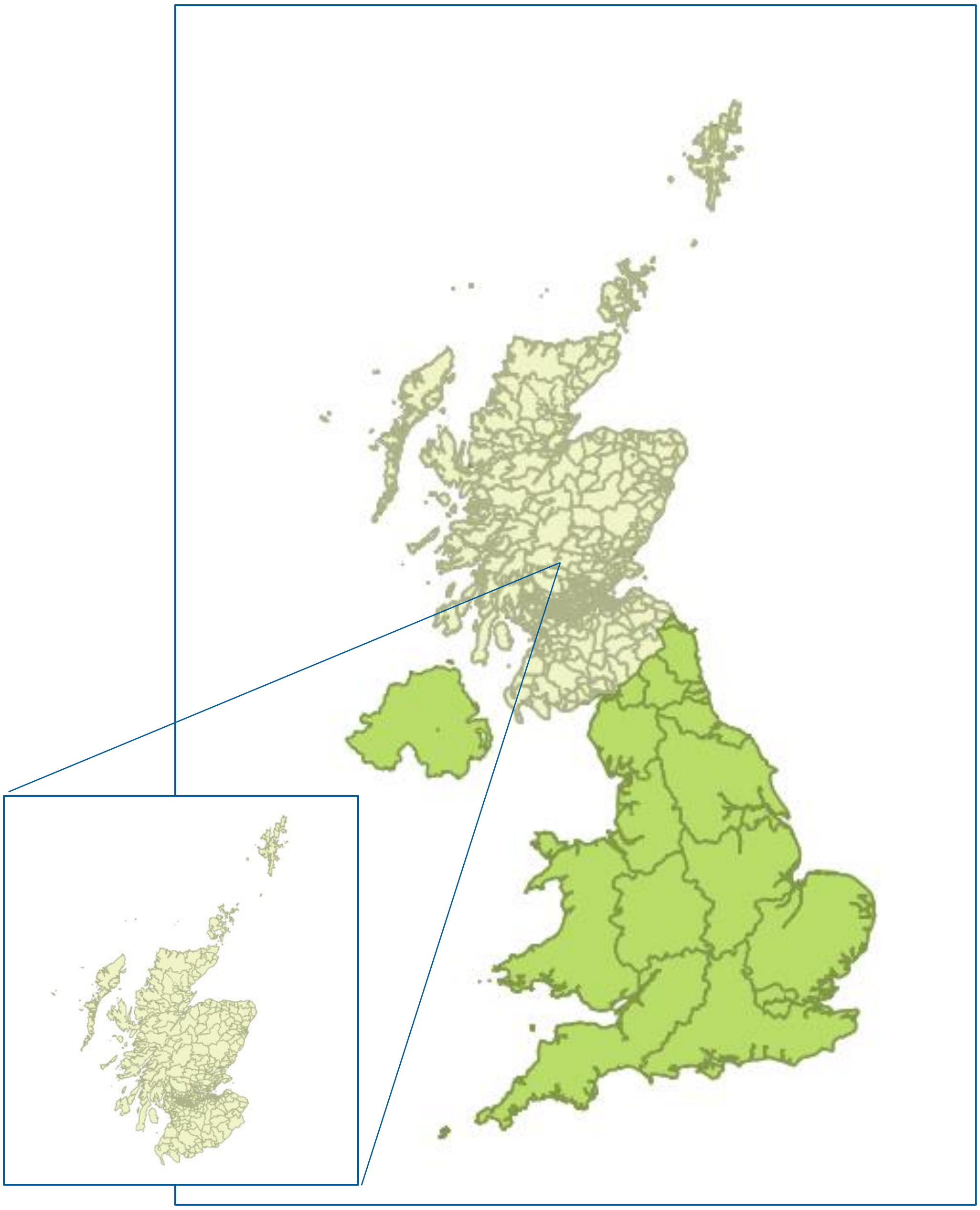


Figure 2: Image showing the zones used to identify the start and end zone of trips.

We segmented trips by different variables; the core segmentation variables are as follows:

- By travel mode:
  - Motorised road
  - Rail
  - Air
  - Ferry
- By travel purpose:
  - Outbound home-based work (OB_HBW)
  - Inbound home-based work (IB_HBW)
  - Outbound home-based other (OB_HBO)
  - Inbound home-based other (IB_HBO)
  - Non-home-based work (NHBW)
  - Non-home-based other (NHBO)
- By the time of day period into:

Morning peak period	(07:00-10:00)
Morning peak hour	(08:00-09:00)
Inter-peak period	(10:00-16:00)
Evening peak period	(16:00:19:00)
Evening peak hour	(17:00-18:00)
Off-peak	(19:00-00:00)

We segmented all trips into these brackets according to the time they entered the cordon.

Study Period

Trips were sampled using Tuesday to Thursdays between 1st March and 30th April 2019. We excluded the Easter holiday period and school Spring Half Term holidays from the 19th to the 22nd April 2019.

Mobile Phone Technology

Overview of the Cellular Network

A cellular or mobile network is a wireless network distributed over land areas called cells; each served by at least one fixed-location transceiver which is known as a cell site or base station. In a cellular network, each cell uses a different set of frequencies from neighbouring cells to avoid interference and provide guaranteed bandwidth within each cell. When joined together, these cells provide radio coverage over a wide geographic area, enabling a large number of portable transceivers to communicate with each other and with fixed transceivers and telephones anywhere in the network, via base stations, even if some of the transceivers are moving through more than one cell during transmission.

Adjacent cells form groups of cells. The names of these groups depend on the generation of the cells. For simplicity in this document, we use the 2G grouping, which is LAC (Location Area Code). LACs overlap and vary in size, depending on the area. Grouping cells into LACs is essential for the collection of event data.

Event Data

O2 mobiles phones generate “events” as they communicate with the national cell network. O2 collects these events on an anonymised basis for analysis. We link each event to a persistent yet anonymised user ID. Along with each event, O2 also stores a timestamp as well as the cell ID of the cell that recorded the event. In this manner, the spatial and temporal distribution of events can be analysed to determine users’ movement patterns. We classify events as either active or passive. It is the combination of both of these types of events that allow O2 to build a representative, stable dataset. Without the inclusion of passive events, the sample would bias toward more active users, and individual user profiles would be biased towards locations where they made calls.

Active Events

Connection events occur when a user turns their phone on or off, loses or regains connection

Call events occur when a user makes or receives a phone call, or moves between cells when on a call

Text events occur when a user makes or receives a text message

Passive Events

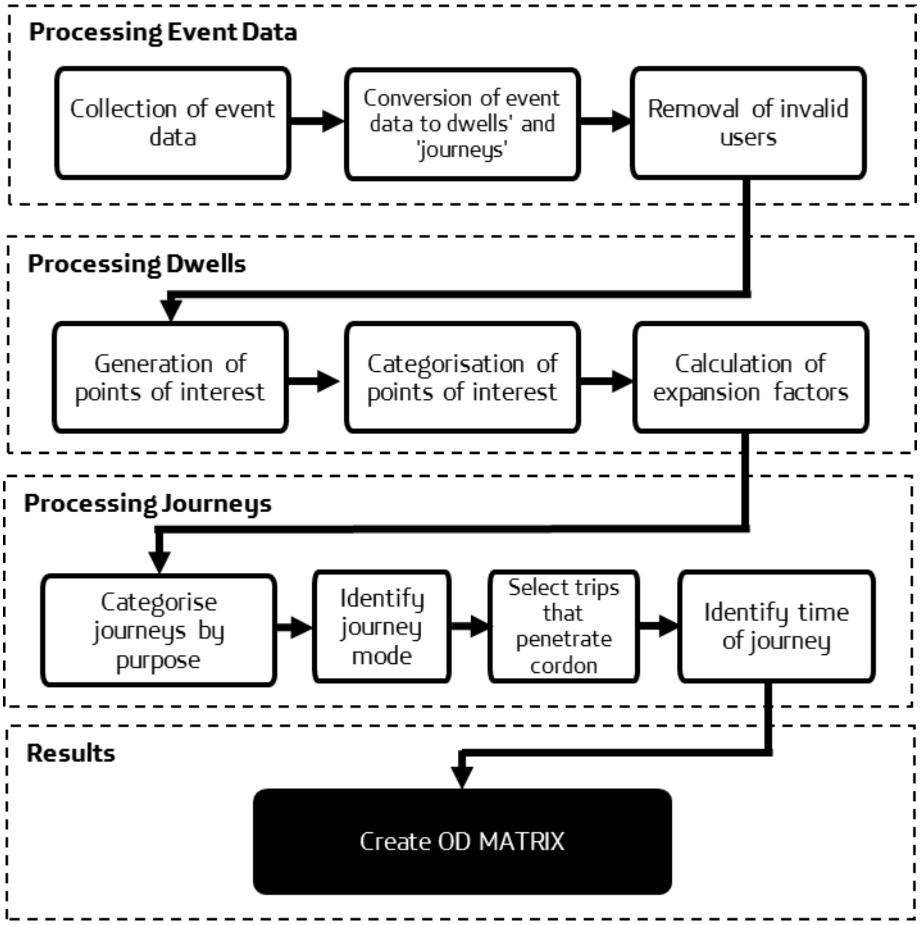
Movement events occur when a user moves from one LAC to another. LACs consist of nearby cells in the same band – so users also create passive events when they transition between 2G/3G/4G coverage. These events ensure that the analysis process records journeys that cover more than one LAC. The collection of these events is vital for accurately observing trips and allocating them to the correct mode.

Time-based events occur whenever a user does not create an event for a sustained period of 3 hours. We use these events to identify longer dwells even if they are in the same LAC as the previous dwell.

Methodology

Process Overview

Figure 3 summarises the process used to create the OD matrix deliverables. We have described each step in more detail in this chapter.



Processing Event Data
Collection of event data
Conversion of event data to dwells' and 'journeys'
Removal of invalid users
Processing Dwells
Generation of points of interest
Categorisation of points of interest
Calculation of expansion factors
Processing Journeys
Categorise journeys by purpose
Identify journey mode
Select trips that penetrate cordon
Identify time of journey
Results
Create OD MATRIX

Figure 3: Process diagram of the existing methodology.

Collection of event data

As described in section two, mobile phones regularly generate events. These are collected ('probed') by O2 for network management and billing purposes. The events are stored in a database to enable an analysis of travel patterns. O2 has access to data relating to the whole of the UK for the last five years, but for this project, we analysed data for 25 specific weekdays (Tuesday to Thursday). Although only these 25 days were used to create the OD matrix, we used the data from other days to define a number of the core segmentations (e.g. identifying valid users and home locations).

Conversion of Event Data to Dwells and Journeys

O2 converts the raw event data into 'dwells' (or settles) and 'journeys'. We take into account the geographic proximity of events, the propensity for phones to 'flicker' between cells without changing their location and the timing of each event. In general, we classify a dwell whenever a user is assumed to be stationary in one distinct place for at least 30 minutes. We classify the period between two dwells as a journey. We store the cells of the events which have been combined to make up each dwell and each journey as 'via points', these can be interrogated to understand the route of each journey or the location of each dwell. We represent journeys as person trips and not vehicle trips.

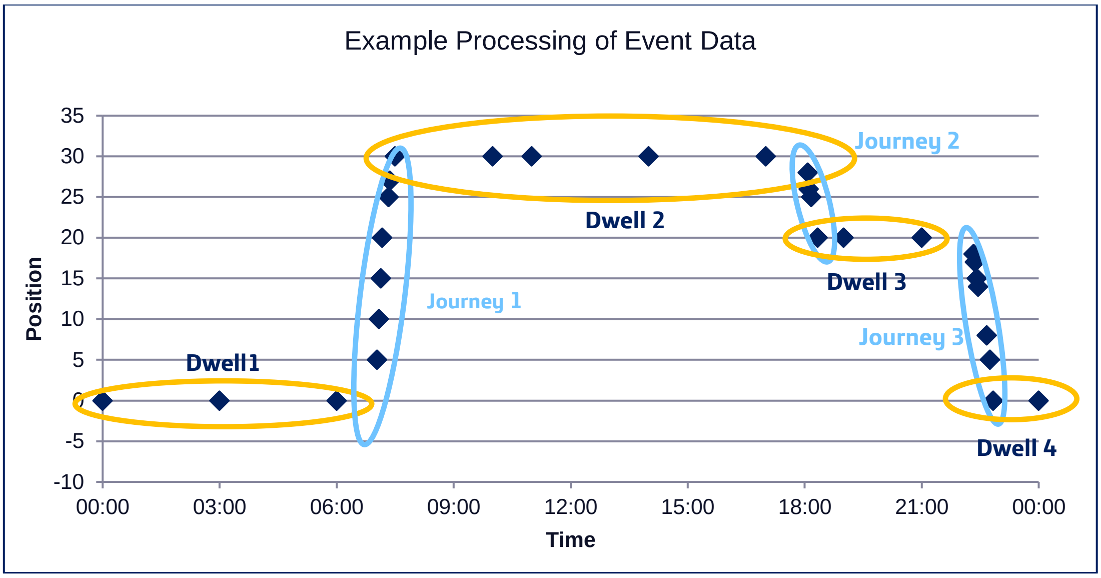


Figure 4: Processing event data: dwells and journeys.

Removal of invalid users

Events are created by all O2 users, corresponding to about 30% of the UK population or circa 22m connections. We allocate each user an anonymised user ID; this ensures that we cannot trace their records back to a particular person. The anonymous ID is set up to make sure that it is consistent even if a user changes their phone. If a user leaves O2 however, their records cease. We run a filtering process to identify and remove these inconsistent users.

Also at this stage, a filter is applied to ensure that we only include mobile devices; we exclude machine to machine (M2M) devices, tablets and GPS units, as they are less likely to be carried by users at all times. Large business contracts are also removed from the sample to reduce the risk of double-counting users who carry two phones.

Generation of Points of Interest

We define Points of Interest (POI) where a user has multiple dwells which overlap each other. By analysing all of the dwells associated with a particular POI the position of the POI can be identified with a higher degree of accuracy as we take the cell information from each of the dwells that contribute to the POI. We compare and analyse the relevant cell geographies associated with a POI and match them to the zone system agreed with Aecom. We associate each POI with a specific zone. Every time a user visits a cell associated with one of their POIs, we record this as a trip to the associated zone.

The categorisation of Points of Interest

The categorisation of POIs is based on the temporal patterns of a user’s dwells at each POI throughout the study period. We classify a POI where a user spends a significant amount of time overnight as their home POI. All users must have a home POI. We classify POIs where users spend a substantial period during the working day as their work POI. We classify all other POIs that are not ‘home’ or ‘work’ as ‘other’ POIs.

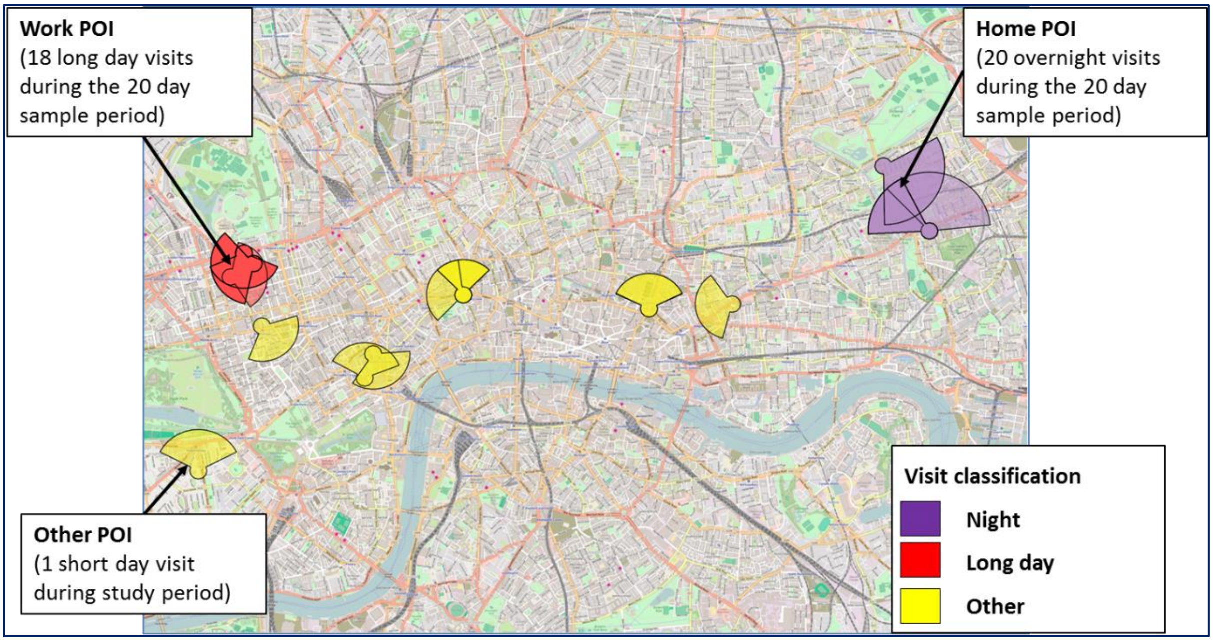


Figure 5: Example POI classification

The POI schematic used is designed to detect regular daytime commuters. As such, it may not correctly capture travel patterns for users who behave in unusual ways:

- **Working from home:** users who work from home have a home POI, but no work POI.
- **No fixed place of work:** users who have an inconsistent place of work (e.g. plumbers) will not usually have a work POI unless they spend most of the study period working at the same site. We include their trips to work will often in the home-based-other matrix.
- **Shift workers**: users who work unusual hours, e.g. night shifts, will not usually have a work POI – we include their trips in the home-based-other matrix.

Calculation of expansion factors

O2, Tesco Mobile and Giffgaff combined market share constitute a representative sample of the UK population with over 32% of the UK mobile network provider market-share. This market share varies across the country, under-representing some age, gender and socioeconomic segments due to technology penetration (phone devices ownership) and over-representing others. When expanding mobile data to represent the entire population, we need to take these bias' into account. The process for calculating the expansion is as follows:

- For every valid user (everyone with an ongoing contract with O2 during the entire duration of the study period), identify their home POI. We exclude under 12 years old as we understand that they are very poorly represented in the data, and are expected to be less independently mobile.
- Count the number of primary home POIs in each MSOA (Middle Super Output Area).
- For each MSOA, compare the number of primary home POIs with the mid-year ONS population estimates from 2017 (the most recent available) for the over 12 population, with a small adjustment for growth since then. We associate each MSOA with an expansion factor which is equivalent to the census population as described previously divided by the number of primary home POIs in that MSOA.
- For each region, compare the proportion of primary home POIs for users in each age/gender/socio-economic bracket with the proportion from the 2018 ONS population estimates. We associate each region with an age/gender/socio-economic reweighting factor which is equivalent to the proportion of the census population in that age/gender/socio-economic bracket divided by the proportion of primary home POIs of users in that age/gender/socio-economic bracket.

- We attach the expansion factor and the age/gender/socio-economic reweighting factor based on the user's primary home POI and the age/gender/socio-economic information of that user.
- All trips made by each user, regardless of origin or destination, are scaled up according to the weight of the user.

Categorising journeys by purpose

Journeys are assigned a travel purpose based on the categorisation of their start and end POI:

Origin POI	Destination POI	Purpose
Home	Work	Outbound Home-Based Work (OB_HBW)
Work	Home	Inbound Home-Based Work (IB_HBW)
Home	Other/Home	Outbound Home-Based Other (OB_HBO)
Other	Home	Inbound Home-Based Other (IB_HBO)
Work	Other/Work	Non-Home-Based Work (NHBW)
Other	Work	Non-Home-Based Work (NHBW)
Other	Other	Non-Home-Based Other (NHBO)

Table 1: Trip Purpose Categories

Education trips: Tertiary education trips made by users will usually be included in the home-based work trips because they are trips between home and a place where the user regularly spends long periods during the time. We recommend that Aecom apply alternative datasets to split out and supplement education trips from the matrices.

Note that education escort trips, where observed, will usually be included in home-based-other trips.

Identify journey mode

At this stage, we analyse the route and characteristics of each journey to allocate the journey to one of the following modes:

- **Rail** – we classify journeys which follow the rail network and which exhibit 'clustering' (see description below) as rail trips.
- **Air** - we classify journeys which have cell events at airport locations, but no cell hits during the flight as air trips. We also take into account travel speed in this classification. For this project, these trips are removed from the final matrices.
- **Ferry** – We classify ferry trips as those that travel between different landmasses that have potential ferry routes. We consider two types of ferry routes, these are:
  - Ferry routes between unconnected landmasses, and
  - Ferry routes between connected landmasses (i.e. to and from the Isle of Skye)

  For trips between unconnected landmasses, we consider all trips that we have not previously classified as air trips. For trips between connected landmasses, we consider the two options (either the user had taken the ferry, or the user had driven), to differentiate these we route the road option and consider the overlap of the journey via points with the connecting links between the landmasses (i.e. bridges). If there is a significant overlap between the journey via points of the journey and connecting links between the landmasses we classify the journey as road if there is a poor overlap we assume that the user had taken the ferry.

- **Road** – we allocate the remaining trips to the road matrix. Please note that this includes coach, bus, and LGV trips as well as car and motorcycle trips.

Clustering: we distinguish between road and rail journeys by identifying cell pairs that show characteristic travel time patterns for either of the two modes. When a train crosses the boundary of a LAC, the phone of every O2 customer onboard generates a passive event. These events occur in quick succession (depending on the length and speed of the train, as well as the device type and the current state of the mobile network), which results in clearly identifiable clustering patterns. We classify specific cell pairs as “rail” when these patterns become apparent. When we do not observe clustering, we classify them as a road pair. On the road, we usually observe a continuous flow of cars, and events (i.e., movements from one LAC to another) also occur continually. An algorithm examines the clustering patterns of all the journeys in the system to identify rail and road journeys.

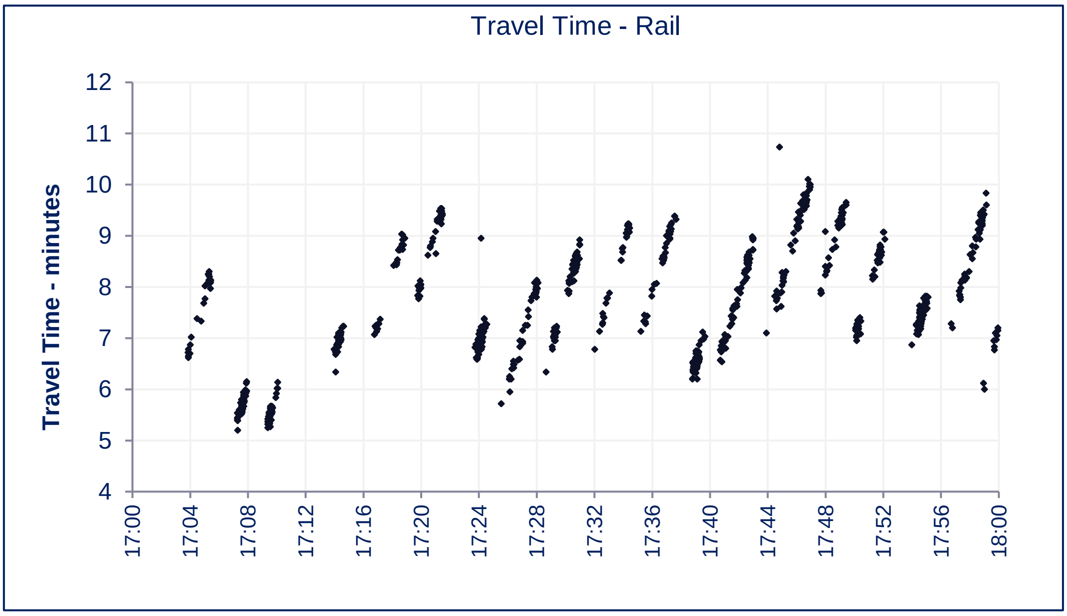


Figure 6: Characteristic clustering pattern of a rail cell pair

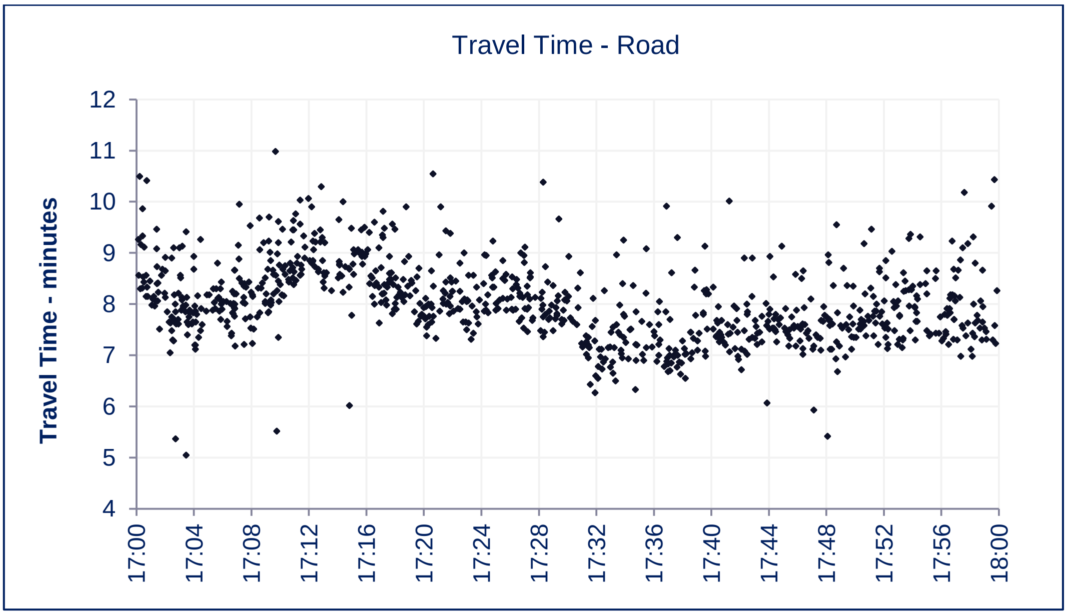


Figure 7: The lack of any identifiable cluster indicates a road cell pair

Select Trips that penetrate cordon

Once every journey is associated with a mode, we map it to a route based on the events (via points) generated during the journey. We compare these routes against the cordon agreed with AECOM and only included those trips which penetrate this cordon in the matrix.

Identify the time of the journey

We allocate journeys are to a time band based on their start time entering the cordon.

Create OD matrix split by mode

Once we allocate all journeys a time, purpose and mode, it is straightforward to create the OD matrix outputs. We allocate trips to a time-period, mode and purpose and include them in the relevant part of the matrix.

Stochastic rounding: to preserve personal data, O2 does not provide outputs relating to the movement of individuals. In the context of an origin-destination matrix, we achieve this by creating an average result representing multiple days of observations, and by rounding results to integer values.

Applying standard rounding methods would cause errors in the outputs because they would cause many cells in the matrix to be rounded to zero, reducing the volume of trips in the data. To avoid this, we use stochastic rounding whereby the probability of rounding a value up or down depends on a fractional part – so a value of 0.1 has a 90% probability of being rounded down to zero and a 10% probability of being rounded up to one. This method of rounding preserves the overall volumes of the matrix (and the size of any part of the matrix large enough for the rounding interval to be negligible) while also preventing the disclosure of individual-level data.

Validation

Before releasing the data, O2 carries out a range of validation checks to ensure internal consistency and check against relevant alternative data sources. Checks are usually limited to zones which are within the cordon because we only include trips if they penetrate the model cordon.

Comparison of Mobile Network Data (MND) home-based trip origins against Census zone home population

Figure 8 shows the number of outbound home-based trips starting in each zone within the cordon on an average day in the study period against that zone's home population, based on the 2011 Census (we extracted the population of each client zone from the Output Areas within each zone). As is to be expected, zones with a higher population tend to have more home trip ends per day. The correlation between the census population and outbound home-based trips has an R^2 of 0.89.

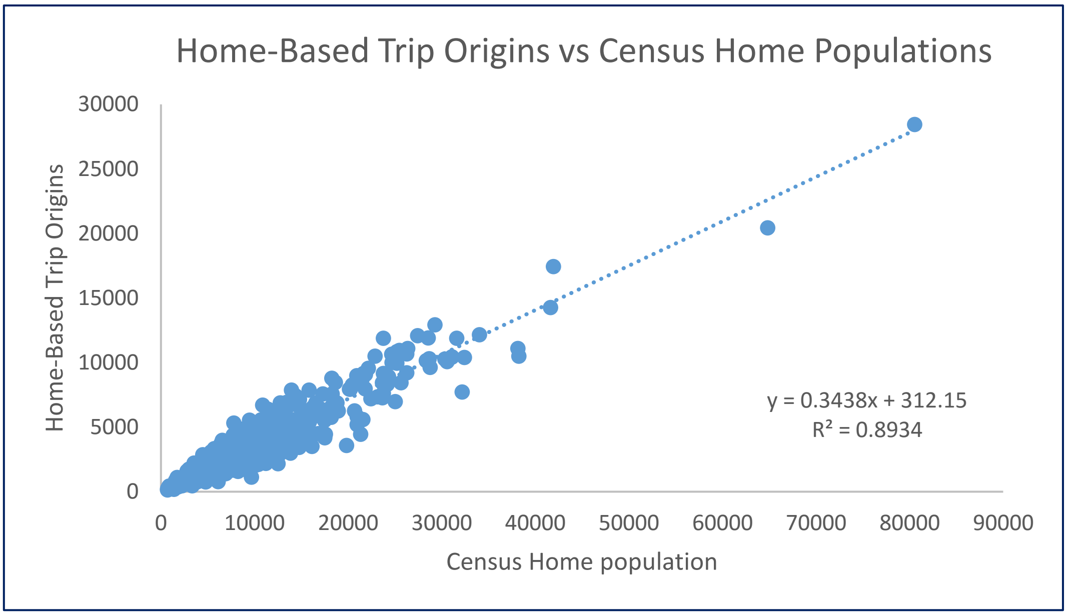


Figure 8: OB_HB Journeys Vs Census Home Population

Comparison of MND work-based trip destinations against Census zone workplace population

Figure 9 shows the number of outbound home-based work trips arriving at each zone within the cordon, during a typical day in the study period, against the work population of each zone (based on Census workplace statistics). We identified a strong correlation with an R^2 of 0.96.

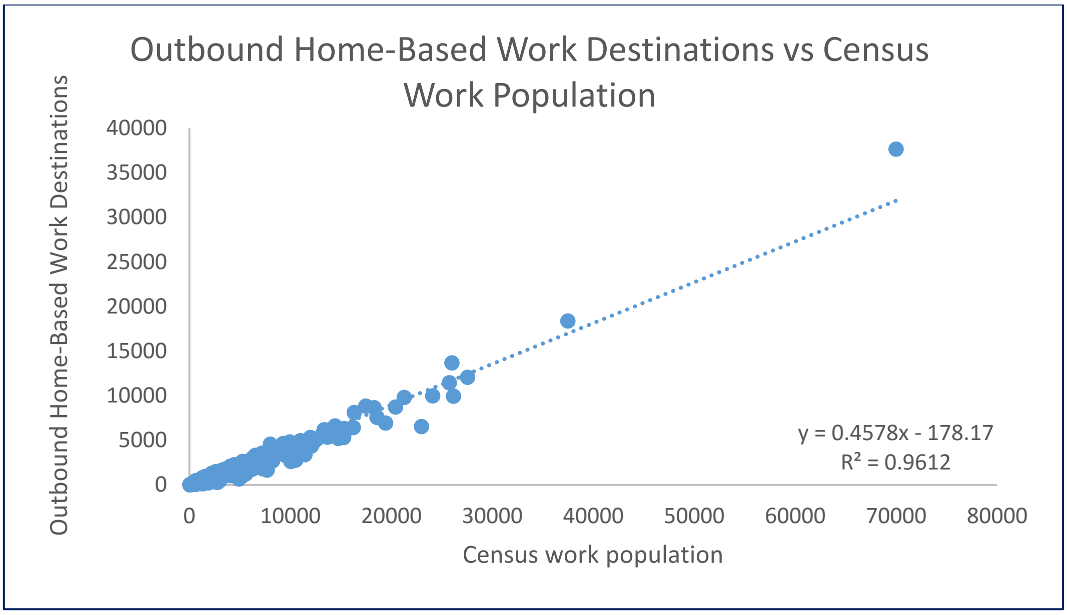


Figure 9: OB_HBW Destinations vs Census Work Population

Comparison of inbound trips and outbound trips per zone

Figure 10 shows a comparison of the number of trips starting (by all modes and purposes) with the number of trips ending in each zone. As expected, we see a strong correlation, with an R^2 of close to 1.

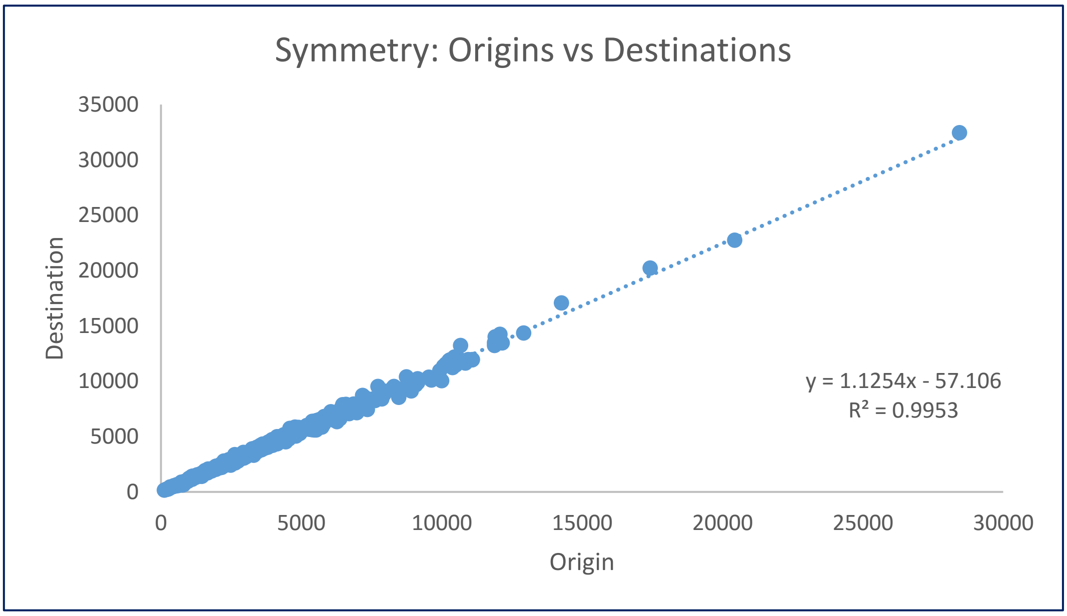


Figure 10: Trip Symmetry by Zones

Comparison of the MND trip length distribution for all trips against NTS trip length distribution

Figure 11 shows a comparison of the trip length distribution for trips starting in the cordon (by all modes and purposes) with the trip length distribution reported in the National Travel Survey for the North West region of England (NTS9911).

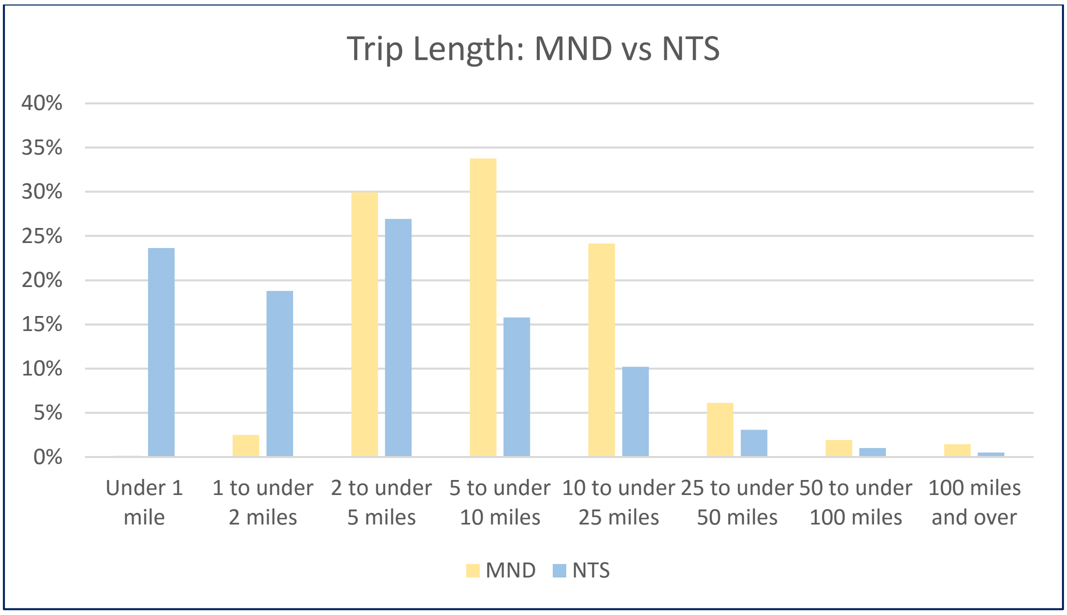


Figure 11: Trip Length Distribution (MND v NTS)

At first glance, the match between the two datasets is poor, with the NTS containing more trips below two miles and the mobile data containing more trips above two miles. However, we find a better match when we compare trips above two miles in length (Figure 12):

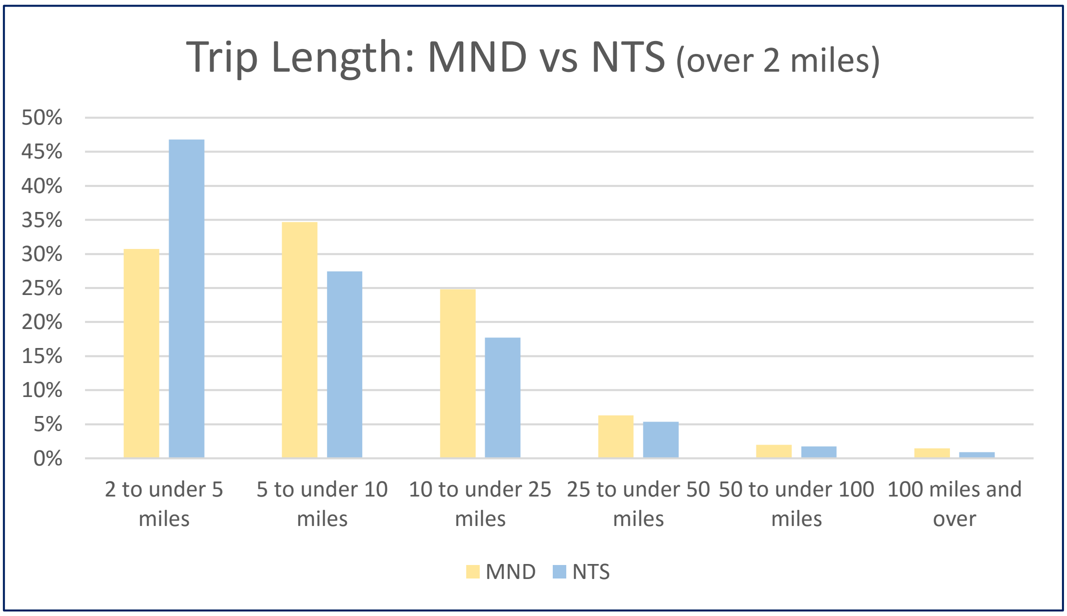


Figure 12: Trip Length Distribution for trips greater than 2 miles.

This result indicates that trips above two miles in length are well represented in the mobile phone data, while trips below two miles in length are only partially represented in the mobile phone data. Short trip under-representation is a well-known limitation of mobile phone data as we can only represent trips if the device moves between cells, which in rural areas, can be large. We recommend that secondary datasets are used to correct for this bias in the mobile phone data.

Comparison of the MND trip length distribution of commutes against Census journey to work data

We compare the trip length distribution from the Mobile Network Data against the Census journey to work (JTW) data. The graph below represents the distance profile of all journeys captured in the matrix of over 5,000 metres.

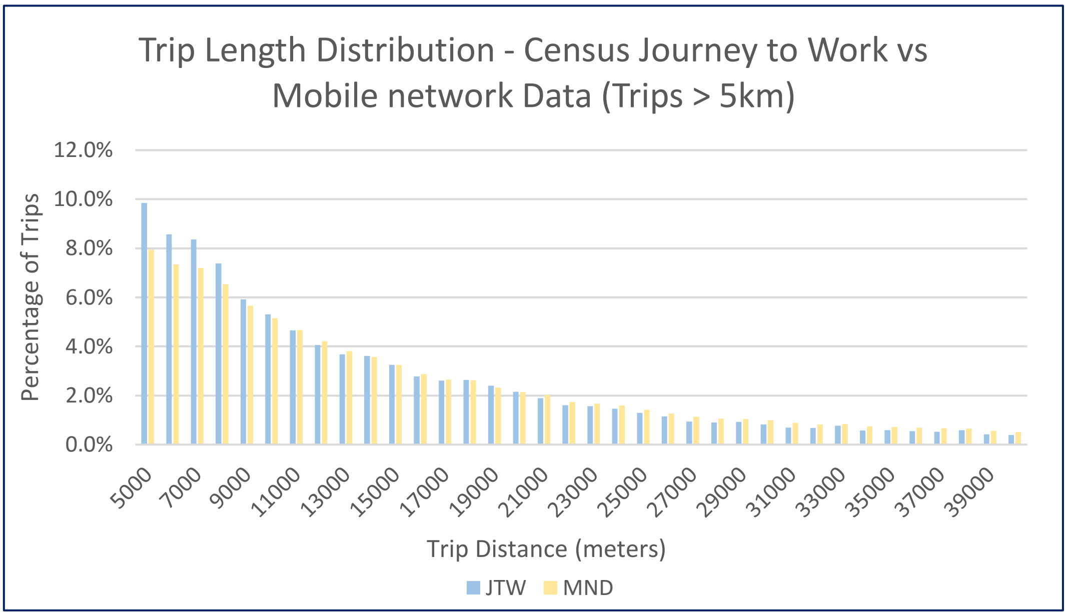


Figure 13: Trip Length Distribution for Commutes greater than 5km (MND v Census)

In Figure 13 above, we can see that the trip length distribution for commutes greater than 5,000 metres is quite closely aligned.

MND rail trip volumes against Office of Rail and Road station counts.

Figure 14 compares the proportion of trips, at Local Authority District Level for the Office of Rail and Road station entries and exits data collected at stations against the proportion of rail trips we observe in the mobile network data at Local Authority District (LAD) level.

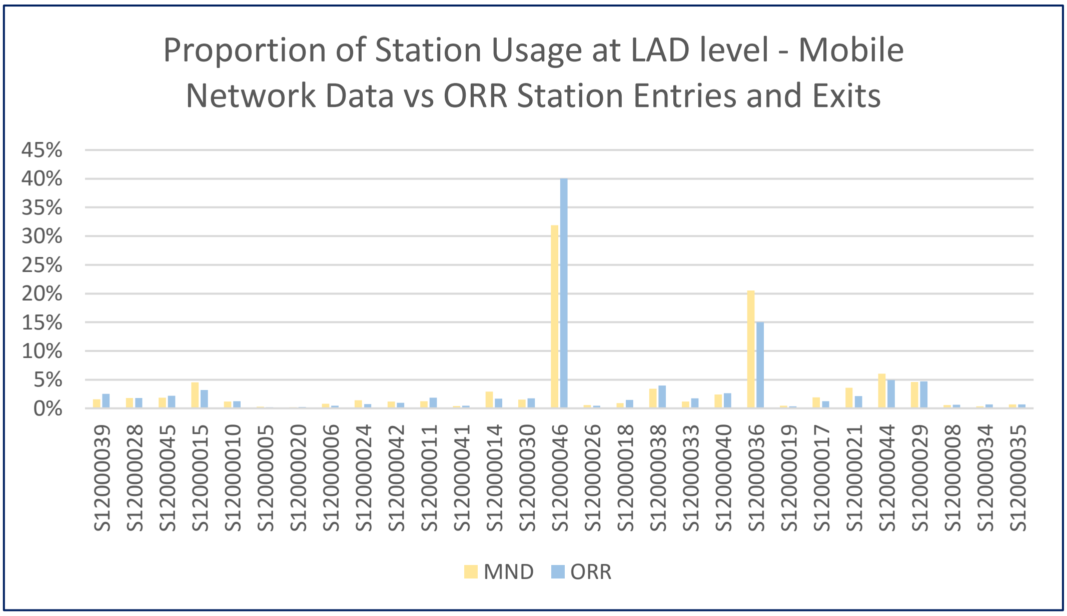


Figure 14: MND Rail Trip v ORR Station Count

Summary

The data for this project has been collected using established and proven methodologies for the application of mobile phone data to transport modelling. We have shown through validations that the mobile data provided is internally consistent and compares well to the secondary datasets. We performed checks that are limited to publicly available datasets and are not intended to be exhaustive. We, therefore, advise further comparisons with appropriate local datasets before applying the matrices to a transport model.

We have highlighted a few biases in the methodology and validation sections, all of which are recognised limitations of mobile phone data. The core limitations are as follows:

- We represent park and ride trips in the mobile data as rail trips from the initial origin to the final destination.
- Comparisons with trip length distributions from NTS indicate that trips below two miles are likely to be under-represented in the mobile phone data. However, this depends on the cell resolution – in urban areas short distance trips are more likely to be represented, while in rural areas the threshold may be slightly higher.

We recommended that secondary data sources are used to enhance the mobile phone data to correct for them.