Mobile Phone Data Validation – A Validation Framework for Mobile Phone Data for Transport Planning

TIL5060 Thesis

Author:
Simon Cornelis Jacobus Knepper

Supervisors:
Erik de Romph – TU Delft
Gonçalo Homem de Almeida Correia – TU Delft
Mark de Reuver – TU Delft
Boon Sey – Jacobs

30/06/2016
Contents

1 Introduction .................................................................................................................. 4
  1.1 Scope ..................................................................................................................... 5
  1.2 Research questions ............................................................................................... 5
2 Literature review .......................................................................................................... 7
  2.1 Studies on mobile phone based matrix building .................................................... 7
    2.1.1 Summary of literature gaps on the topic of mobile phone based matrix building ....... 8
  2.2 Validation of mobile phone data in literature ....................................................... 9
    2.2.1 Summary of matrix validation in literature .................................................. 10
3 Reflection on the matrix creation algorithms ............................................................ 11
  3.1 Event generation .................................................................................................... 11
  3.2 Creation of ‘dwell’ and ‘journey’ ........................................................................... 11
  3.3 Points of interest ................................................................................................... 12
  3.4 Estimation of start time ......................................................................................... 12
  3.5 Expansion ............................................................................................................. 12
  3.6 Classification of ‘road’ and ‘rail’ users ................................................................... 13
  3.7 Creation of routes ................................................................................................ 13
  3.8 Privacy .................................................................................................................. 13
  3.9 Hypotheses .......................................................................................................... 14
4 Method of validation .................................................................................................... 15
  4.1 Validation data sources ........................................................................................ 15
    4.1.1 Census Journey to Work Matrix ................................................................... 15
    4.1.2 National Trip End Model (NTEM) ................................................................. 16
    4.1.3 Census data ................................................................................................... 16
    4.1.4 National Travel Survey ............................................................................... 16
  4.2 Hypotheses and datasets ...................................................................................... 16
  4.3 Zone systems ........................................................................................................ 18
  4.4 Framework structure ............................................................................................. 19
    4.4.1 Statistical tests and criteria .......................................................................... 20
5 Validation & results ..................................................................................................... 24
  5.1 Mode share and journey purpose share .............................................................. 24
    5.1.1 Mode share .................................................................................................... 24
    5.1.2 Journey purpose share .................................................................................. 25
  5.2 Daily profile .......................................................................................................... 27
    5.2.1 Reflection on National Travel Survey ............................................................ 28
5.3 National Trip End Model ........................................................................................................... 29
   5.3.1 Trip End validation, all day, all purposes ............................................................................. 29
   5.3.2 Trip End validation, AM Peak, home based work outbound ............................................. 32
   5.3.3 Trip End validation, PM Peak, home based work inbound ............................................... 32
   5.3.4 Final note on NTEM .......................................................................................................... 33
5.4 Census population and workplace data .................................................................................. 34
   5.4.1 United Kingdom level ........................................................................................................ 34
   5.4.2 Final remarks on census .................................................................................................. 34
5.5 Journey to Work validation ..................................................................................................... 35
   5.5.1 Matrix comparison ......................................................................................................... 35
   5.6 Reflection on validation methodology .................................................................................. 36
   5.6.1 Discussion on the hypotheses ......................................................................................... 36
   5.6.2 Conclusions on the data ................................................................................................ 38
   5.6.3 Future work on matrix generation methodology ............................................................ 38
6 Conclusions, discussion and further research .......................................................................... 40
   6.1 Discussion ............................................................................................................................ 41
   6.2 Contribution to existing literature ....................................................................................... 41
   6.3 Further research ................................................................................................................ 42
Bibliography ..................................................................................................................................... 43
i. NTEM validation charts, all day, all purposes ......................................................................... 45
ii. NTEM validation charts, AM Peak, home based work outbound ......................................... 48
iii. NTEM validation charts, PM Peak, home based work inbound .......................................... 51
iv. Data quality maps ................................................................................................................ 54
v. Census validation results ....................................................................................................... 56
vi. Census Journey to Work matrix comparison ..................................................................... 57
vii. Glossary ................................................................................................................................... 58
1 Introduction

This thesis deals with the topic of using mobile phone data for transport and traffic assessment. Mobile phone data, used for this purpose, is a new, innovative technology that can help improve transport modelling by providing more detailed data on movement patterns. At the time this thesis is written, mobile phone data are on the verge of being applied for transport assessments on a large scale in the United Kingdom. This provides opportunity for more research on the topic. This thesis will focus on the validation of this new technology. The aim is to provide a framework for validating the accuracy of the movement patterns inferred by the mobile phone data.

Transport models have been used for multiple decades to aid in policy making. Traditionally, assessments in the field of transportation were done using a multi-stage process including trip generation, trip distribution, mode choice and trip assignment models. A trip generation model would aim to predict the amount of trips generated in a certain zone, while trip distribution splits these trips over the destination area. The total trips generated in these models depend on the number of households and household characteristics such as car ownership, income and family size. Well known trip distribution models are the gravity model and the Logit Model, factoring in personal/household characteristics or zonal characteristics such as number of jobs in a destination zones and travel distance between two zones. This results in an Origin-Destination (OD) Matrix, containing the number of trips from an origin to a destination.

As of today, mobile phones have almost fully penetrated the market and have become part of our everyday lives. Whenever we make a call, send a text or connect to the internet, data are stored at the mobile phone provider the phone user is connected to, which results in large quantities of data being stored in the databases of mobile phone providers. This data remains in the databases for some time for billing purposes. This provides a lot of opportunities for applying mobile phone data in all kinds of research fields. One of these fields is transportation modelling, where mobile phone data can be used to track the phone users, determining their origins, destinations and the trips the users make. There is a lot of potential for mobile phone data to replace the traditional trip generation and trip distribution models for the base year. In fact, this has already been applied in multiple contemporary transportation assessments (examples include Demissie, 2014, Wang et al., 2015, Calabrese et al., 2011 and Alexander et al., 2015). With more and more data becoming more easily available to the researcher, it seems likely that mobile phone data are a very helpful addition to transportation modelling in the future.

Even though there is great potential for the use of mobile phone data on this particular application, there are still many drawbacks and unresolved issues. First of all, there are serious privacy concerns, and the data will have to be anonymized and handled cautiously every time it is being used. There may also be a bias due to the data sample being nonrandom with certain groups being over- or underrepresented (Calabrese, 2013). Another issue is the precision of mobile phone data. The mobile phone provider usually has data on the cell location the mobile phone is in. The problem with this is that the area of the cell can vary a lot between being small in city centres and large in rural areas (Cacares, Wideberg, & Benitez, 2008). This leads to positioning accuracy outside of urban areas being poor. Positioning accuracy also ranges a lot between the different data sources 2G, 3G and 4G.

Besides this, mode identification and trip purpose identification prove to be difficult. No defined methods for identifying transport modes of road users have been developed in literature. The problem is often lack of socio-economic or demographic data on the mobile phone users (Calabrese, 2013). Besides cars, there are buses, Light Goods Vehicles and Heavy Goods Vehicles on the road, and train users may also generate events that cannot be distinguished from road users. This is a problem if the researcher is only interested in trips made using a particular mode, e.g. motorized road vehicles. Similar to mode
identification, trip purpose is hard to determine because just being able to track mobile phone users does not help determining trip purpose. Further steps are required for trip purpose determination.

There is no defined framework for validating Origin-Destination matrices yet. This often makes comparing mobile phone matrices to other (mobile phone or synthetic) matrices and external sources of data unstructured and inconsistent with other studies on mobile phone data. This has been recognized as an issue in literature (Demissie, 2014). As research goes on on the outstanding issues mentioned here such as inferring mode and journey purpose and applying expansion factors, such a validation framework is required. Being able to test mobile phone matrices in a more consistent manner will provide with a track record clearly showing improvements on mobile phone based matrix creation methodologies over time, as research progresses and better methodologies are developed.

These issues can seriously hamper the application and validation of mobile phone as a valid and useful technology for traffic and transport assessments. At present, transport modellers especially need to be aware of these issues and work around them. In the future, as more granular data and improved algorithms become available, mobile phone OD matrices will likely improve in quality. Developments will need to be continuously monitored, with each new matrix created using mobile phone data validated and tested for quality.

1.1 Scope
This thesis is being developed at Jacobs in London, United Kingdom. All mobile phone data used in this thesis are processed by a mobile phone provider and contains location data (mobile phone traces) from users of one mobile phone provider in the United Kingdom. This includes data from active events (calls, texts, connecting to internet), passive events and handover data. These data are processed by the mobile phone provider to create mobile phone matrices. Both the data and the processing algorithms are confidential and only the output, being aggregated mobile phone matrices, can be analysed. The purpose is to create a database which contains origin-destination data by time, mode and journey purpose of the mainland United Kingdom. This database is to be used by transport modellers to use on highway related projects. This research is carried out within this project and has the goal of developing a validation framework to be used on the matrices. The outcome of this validation will lead to further steps to improve the matrix creation methodology. This framework will also help with validating matrices being released to consultants in the future.

Developing a validation framework is central in this thesis. The framework will not be specific to the particular matrices evaluated within this scope, but can be universally applied to all mobile phone matrices in the United Kingdom. This will also be true for other (developed) countries, provided the same or similar sources of data are available to test against. A validation framework is needed, because the matrix creation process is iteratively being fine-tuned by mobile phone providers and will require some sort of benchmark to see if a new iteration is an improvement over the last. Within the scope of this thesis, a prior matrix validation will be carried out, but work on the database will proceed beyond the scope and will probably take multiple years.

1.2 Research questions
A main research question can be formulated. This research question will directly relate to the goal of developing a universally applicable validation framework.

"Which validation steps have to be undertaken using which sources of data, to be used to test the quality of a trip matrix derived from mobile phone data?"
This will lead to a series of steps needed to be undertaken to validate a mobile phone based matrix on its most important aspects. These aspects will have to be quantifiable to include output such as number of trips by zone, trip length and journey purpose. All of these aspects will require a different source of data to validate against. There is a large quantity of data sources available in the United Kingdom, as well as in other developed countries, with varying degrees of quality and relevancy. This makes careful selection of data sources an important part of the validation methodology.

Applying the framework will already provide insight in quality and shortcomings of the prior mobile phone based matrix being analysed in this thesis. The following secondary research question follows logically from applying this process:

“Is the tested trip matrix derived from mobile phone data valid and applicable for use, and if not, which aspects will need to be considered for improvement?”

With this outline, there will be two elements to this thesis. The first element is developing a framework that can be applied in the following years to track improvements of mobile phone data and matrix creation algorithms. The second element deals with mapping the matrix quality already achieved and can provide a benchmark for future improvements.

It is not possible to develop a sound validation framework without prior knowledge of how the matrices are built, what the shortcomings in data collection and processing algorithms are and which assumptions have been made. A literature review is conducted on the topic of mobile phone data, matrix building and journey purposes, which can be found in chapter 2. The matrix creation process is confidential, and therefore has not been included in this version of the thesis. A critical reflection of the matrix creation processes, discussing and summing up its shortcomings and assumptions, is found in chapter 3. Hypotheses are formulated around these shortcomings and assumptions. This is used to create the validation framework in chapter 4 containing a series of steps of the whole validation process and a detailed description of the sources of data used for these steps. The results are discussed in chapter 5 and can be found in appendix i through appendix vi. The conclusion and discussion can be read in chapter 6.
2 Literature review

This chapter gives an overview of literature about applications of mobile phone data in transportation and urban modelling. This application is not new. In fact, very early attempts to use mobile phone data date back to the 1990’s. The general idea has been to use call data records generated by use of mobile phones within the cellular network. The time- and location data of these records can be used to derive where the users started and ended trips, as well as where they stayed in between these trips. Since the penetration rate of mobile phones was quite small in the 1990’s, these attempts have not always been very successful. However, the interest in mobile phone data from researchers remained to exist throughout the years. Paragraph 2.1 provides an overview of some literature on attempts that have been undertaken to use mobile phone data for transport- and spatial planning and modelling related purposes throughout history. Paragraph 2.2 presents the validation part of the mobile phone data in more detail.

2.1 Studies on mobile phone based matrix building

The penetration rate of mobile phones in most western countries got close to a 100% in most developed countries around 2006, which shows in the number of papers published at that time. In Cacares et al. (2007) one of the earlier attempts to develop a methodology for building Origin-Destination matrices using mobile phone data has been made. This was possible since at that time, 89.4% of the Spanish population had a mobile phone. The researchers made use of handover data, being the time and location the mobile phone user switches between two masts with a corresponding network cell. If the phone is on at the origin and remains on until the destination is reached, it is possible to track the route through all the network cells. Becker et al. (2011) explores handover data further, using this data source to determine these routes. A problem with the route tracking application is that the network cells may be large or varying in size throughout the network. Cell size is based on the number of masts that are in the area, the more masts there are, the denser the network is and the smaller the cells are, allowing for greater accuracy while pinpointing location. It is possible for different routes to go through the same cells, making it impossible to track the exact route. Becker et al. used a Nearest Neighbours algorithm and signal strength data to determine the likelihood that a route was taken.

A good review of different contemporary types of data-collection, including mobile phone data, can be found in Cacares et al. (2008). Sources of data discussed here are GSM (from mobile phones) and GPS (from satellites). GSM networks generate more data but have an accuracy ranging from 100 m to 150 m with lower accuracy in rural areas, while the GPS is more accurate but only has a small and non-random sample. GSM data also has a wider coverage. A significant drawback of using GSM data is that it is hard to measure intra-area trips because the cells may be too large, while GPS would not have this kind of problem because of its high accuracy. A more in-depth discussion of applying GPS data for traffic purposes can be found in Herrera et al. (2010). GPS is very accurate in determining speed and location but suffers from a low sample size and a sample bias was found during the research, while mobile phone data has a much higher sample size and a smaller sample bias. The authors only make a comparison between these data sources and do not discuss applying a combination of both sources.

The higher position accuracy of GPS data may make it more suitable for route tracking applications than GSM (mobile phone) data. This limits or removes the need for further classifying algorithms such as used by Becker et al. (2011). There is great potential in using GPS data to support the GSM based matrix estimations, making use of the advantage of higher positioning accuracy. GPS and other sources such as radio frequency identification (RFID) and license plate recognition (LPR) also provide journey time data that can be used to calibrate and validate OD matrices built using mobile phone data. More research should be done on the combination of mobile phone, GPS, RFID and LPR as there is a lot of potential for fusion of these different data sources.
An important limitation is that the sample only consists of mobile phone users of a specific operator. Cacares et al. (2007) solve this by applying an expansion factor to expand the mobile phone sample to the number of vehicles. This expansion factor is based on the mobile phone provider’s market share in Spain and the number of cell phones that are on and in the vehicle. While this gives a decent approximation, it does not take into account that market shares differ per region and also does not deal with potential non-randomness of the sample. There was also no segregation of the matrix by mode or trip purpose, either because this could not be done or because this was outside of the scope of the research. Despite these limitations, the results were promising and clearly showed the advantages of the large sample size and the shorter time between data collection and making the OD-matrix, which is close to real-time.

Around the same time, attempts have been made to research urban dynamics or population density in urban areas. Although this is another interesting application, it is not directly related to transport and therefore will not be further covered in this review. For examples, see Pulselli et al. (2008), Calabrese et al. (2010), Reades et al. (2007) and Reades et al. (2009). More interesting from a transportation perspective is Ahas et al. (2010). Even though this research was not conducted with the goal of understanding mobility patterns, its methodology allows for determining ‘anchor points’ or Points of Interest (POI’s), being places where users make a lot of calls. This method is used by Novak et al. (2013) to build OD-matrices from these anchor points. Taking this method further would allow for activity based modelling and possibly being able to determine trip purpose if users are travelling between these POI’s.

Other studies that may help determining activity patterns are Sevtsuk & Ratti (2010) and Csáji et al. (2012). These activity patterns have a lot of potential for determining trip purposes in OD-matrices. Looking at more recent attempts to develop OD-matrices from mobile phone data, we see that, although these attempts were relatively successful, segregation on mode and trip purpose was usually outside of the scope (Calabrese et al., 2011; Nanni et al. 2014; Wang et al., 2015). Järv et al. (2012) also shows the possibility to build OD-matrices from mobile phone data yet found no solution to problems such as determining trip purpose, modal split and trip chaining. The studies by Nanni et al. and Wang et al. (2015) used OD-matrices based on only the sample, which was not expanded to the population. Calabrese et al. (2011) use a scaling factor to extrapolate the ODs from mobile phone data to the population, which is based on the monitored trips and census data. Determining purposes, modes and developing a more sophisticated model for sample expansion would be going one step further. A recent attempt to identify journey purpose can be found in Alexander et al. (2015). The algorithm proposed in this study makes use of the number of times a location is visited. The most visited location between 19.00 and 20.00 is identified as the home location for a particular user. The work location is then determined by calculating the location with the largest travel distance multiplied by the amount of times the location is visited. All other locations are flagged as ‘Other’. This allows for trips to be separated between home based work, home based other and non-home based work. In these papers, a validation using household surveys and census data shows that the method to flag journeys by their starting and ending POI is accurate. Tracking users over a period of time (multiple months) to determine their frequency of visit of locations is a very sensible approach indeed, and has now proven to give valid results.

2.1.1 Summary of literature gaps on the topic of mobile phone based matrix building

Even though a lot of research has been published on mobile phone data and OD matrix building, some issues are still unresolved. The most notable ones are described below.

- **Sample expansion.** Mobile phone data usually only comes from one source: the mobile phone provider. This means that only a sample of the data is available, which may be biased. There is also the problem of a mobile phone provider having different market shares in different regions,
meaning some regions are under- or overrepresented in the sample. A sophisticated model for sample expansion is required to handle this problem.

- **Mode choice identification.** A road is shared by users of bikes, cars, buses, LGVs and HGVs. The mode of transportation cannot be directly derived from the data sample. Characteristics of the trip, such as speed and travel distance, have to be used to classify the trips into trip modes.

- **Journey purpose identification.** Journey purpose faces a similar problem as mode choice: it cannot be directly derived from the data. It is possible to do this indirectly by using characteristics of the stay at the location, such as frequency of trips to the location length of stay.

- **Validation.** Demissie (2014) mentions that most studies regarding mobile phone data haven’t validated their results. He also mentions a lack of ground truth data to validate against. It is not clear how these matrices should be validated before they are used for modelling. Synthetic matrices from traditional methods may contain errors themselves, and calibration is needed to further refine matrices. Mobile phone matrices would add value to modelling if the quality of the data is better than that of a synthetic matrix. A method of validation is required to test the quality of the mobile phone matrices and prove their added value.

As mobile phone providers are currently attempting to resolve these issues, a solid framework examining this is required to test the quality of mobile phone matrices and find out if and on what degree errors in the matrices are caused by the issues mentioned here. This framework is not currently defined, even though work has already been done to validate mobile phone matrices by those who have already undertaken studies.

### 2.2 Validation of mobile phone data in literature

Most studies dealing with mobile phone data have dealt with validating the output origin-destination matrix. Every study defines its own validation methodology which is based on datasets available or purpose of the research. There is merit in looking into some individual validation methods, and learn lessons from validation work done before. The overview of validation work is described in this paragraph.

Validation is not clearly defined in most papers using mobile phone data and sometimes not even carried out (Demissie, 2014). This is mostly because the nature of the papers is to explore OD matrix creation using mobile phone data, and not so much define a validation framework as is the goal of this thesis. Alexander et al. seem to have defined their own validation steps more extensively (Alexander, Jiang, Murga, & González, 2015). The first part in their validation process deals with the allocation of home locations. The number of home locations inferred from the CDR (call data records) correlates with the number of home locations from census population data. This would inspire confidence in the expansion process being accurate, as the inferred home locations are used to compute these expansion factors. Alexander et al. have used household surveys to compare the journey purpose segmentation and the trip length distribution. This quickly provides with some high-level figures for comparison that can be easily interpreted. However, no metric has been defined for the mobile phone data to ‘pass’ or ‘fail’ a test. The researchers also test the trips for each OD pair against a dataset called CTPP, a census-based matrix for home based work origins and destination in the United States. The statistical method used for this part of the validation is a correlation analysis. Finally, a trip length distribution derived from an assignment model using the mobile phone matrices has been compared to the CTPP.

Bonnel et al. (2015) also validated a matrix created from mobile phone data. Similar to Alexander et al., Bonnel et al. have used correlation analysis by plotting census commuting data against their home based work mobile phone trips. The $R^2$ of the analysis was used as an indicator of quality. A second test involved a survey (Enquête Globale Transport de l’Ille-de-France) containing the origins and destinations of trips. Again, correlation analysis was used with $R^2$ as an indicator of quality. Bonnel et al. did not do a
validation based on journey purpose split, distribution over time of day or trip length distribution. This still calls for metrics for these types of validation. As for testing matrix to matrix or matrix to residential population, correlation analysis seems to be a popular method.

A similar validation method is used by Wang et al. (2015). A regression analysis is used for comparing a mobile phone based matrix against a gravity model-based matrix. This is also done for the relation between inferred home locations from mobile phone data and census population data. Comparing mobile phone home locations and census data is proposed by Calabrese et al. (2011). It is notable that both Wang et al. (2015) and Calabrese et al. (2011) use a logarithmic scale for these regression analyses. Applying a logarithmic scale helps dealing with points of data over a large range, and makes it easier to interpret the graphs.

2.2.1 Summary of matrix validation in literature
In summary, regression is very often used for comparing a mobile phone based matrix to either a synthetic matrix or census data. $R^2$ is the chosen indicator of quality, with a high $R^2$ implying a good fit of the regression model. If the range is so large that interpretation of the regression analysis becomes problematic, a logarithmic scale is used. This is often needed when comparing a matrix to census data, as the values (trips or population) can vary widely among the zones. For comparing matrix characteristics, such as trip length, trip purpose, mode share and trip start time, statistics are required. For the United Kingdom, such a statistic is prescribed for validating modelled flows against traffic count data (Department for Transport, 2014). Outside of the United Kingdom, other statistics may either be prescribed to modellers, or may be common practice. In some cases, such a statistical test is not as clearly defined. Defining these tests will be useful to researchers, consultants and government authorities attempting to create matrices from mobile phone data, and will allow for unambiguous interpretation of the quality of mobile phone matrices.
3 Reflection on the matrix creation algorithms

To write this chapter, a matrix creation methodology from a mobile phone provider has been analysed. Due to confidentiality reasons, this methodology cannot be described in this thesis. However, to better understand the shortcomings of mobile phone based matrices, this methodology will be reflected upon. Many of the algorithms in this methodology are based on assumptions, which may prove to be oversimplified or false. Furthermore, the data may lack detail to sufficiently find all the trips with sufficient accuracy. For each step, one or several potential issues will be raised. These issues are translated into hypotheses meant for further investigation of the quality of the data. If there is evidence that one or multiple of these hypotheses are true, a case can be made for changing and improving the current methodology of matrix creation.

3.1 Event generation

The quality of the OD matrix relies largely upon the quality of data generated by events in the network. Well known problems with mobile phone data also apply here, being varying levels of accuracy because of varying mast coverages, and not having data on a continuous basis. The first is mostly a problem in rural areas, where mobile phone masts cover large areas and it is not possible to track the exact location of a user. This may especially have implications for determining ‘home’ and ‘work’ locations later on in the process, as these locations may be anywhere in an entire village covered by only one mast. An additional problem in this scenario is that trips within the same cell area may not get detected, since no handover takes place if the user does not leave the LAC. In urban areas, it is suspected that this is less of a problem.

Not generating events on a continuous basis, and in some cases, events being far apart, can also be a potential issue. Some networks have the benefit of sending out a ‘ping’ to a phone at certain time intervals, so a passive event will take place at least some times during long stays at location. For this reason, it can be argued that the ‘home’ and ‘work’ location can reliably get detected, because a user is likely to spend more than the ping time interval at a certain location. A problem arises with short stays such as shopping or doctor visits. For light mobile phone users, no event may be generated at these locations at all, causing a trip to be completely missed.

In summary, the following three issues can be raised from the way data is generated by the network.

1. Lack of spatial granularity makes it difficult to find user ‘home’ and ‘work’ location.
2. Lack of spatial granularity causes trips within the same cell to be missed.
3. Short stays can be ‘hidden’ in between two events.

It should be noted that these issues will eventually be resolved with higher market penetration of 4G technology. Most of the time, 4G data has significantly less problems with spatial granularity and event generation, as the 4G network leaves more ‘breadcrumbs’ with finer spatial accuracy. This is due to 4G data leaving a passive event when a user connects to a new cell tower, rather than crossing a LAC boundary as is the case with 2G and 3G data.

3.2 Creation of ‘dwells’ and ‘journeys’

Creating dwells and journeys is the first step in the processing of the event data. A threshold is applied for classifying ‘dwells’, a synonym for a stay at a location, which is currently set to a certain time threshold. This means that short stops on the way to another destination will not get a separate OD pair in the matrix, and a trip chain with several stops will become a single trip. In order to catch these ‘trip chains’, the threshold will have to be lowered. However, this will create the additional problems of journeys being misclassified as dwells. It is probable that a user will generate two events more than the dwell threshold apart in the same cell during their journey, and there is no telling if a user stopped briefly or is delayed by
traffic, in which case it would not be a trip chain. This provides another inherent limitation for the detection of short stay trips. The single issue raised in this step is:

4. Short stays are not classified as such by the threshold being too high.

This can only be resolved with the availability of data with finer spatial accuracy, as the threshold can then be lowered to catch short stays. A smaller cell size would mean the exact location in which events are generated is better known which provides better tracking of journeys and dwells.

3.3 Points of interest
The search for the points of interest ‘home’ and ‘work’ for the users in the network is based on several assumptions.

- Each user has a home and work location.
- Each user is at home most of the time at night.
- Each user is at work most of the time during the day.

These assumptions will hold true in most cases, but certainly not in all cases. Unemployed users also get assigned a ‘work’ point of interest, and children carrying a mobile phone to school will generate work trips. Furthermore, not everybody follows the typical pattern of working over the day and sleeping at home over the night, the obvious example being nightshift workers. No attempt has been made to find these nightshift workers or filter out those unemployed or going to school, and will likely result in errors in the amount in home based work trips. In summary, the following issue has been identified.

5. Users not actually going to work are misclassified as going to work.
6. Nightshift workers and other similar exceptions have misclassified ‘home’ and ‘work’ POI’s.

3.4 Estimation of start time
The estimation of start time follows these two assumptions:

- The observed distance between origin and destination is correct.
- Vehicles travel with a constant speed.

While simplified, these assumptions may not prove to be much of a problem since this process is only in place to split the trips by hourly matrices according to starting hour. This is only relevant when modelling peak hours. Trips can be wrongly allocated inside or outside of a peak period, causing the modelled peak period to be more or less intensive than in reality. Only one issue is raised here:

7. Trip starts may be allocated to the wrong hour.

3.5 Expansion
The expansion step is an important step in the matrix creation process. If this is done incorrectly, the matrices will not reflect reality at all. If the final numbers of trips in the matrices are far off, there may be evidence that there is a problem with the current method of validation.

The expansion is based on two types of information, one being residential population from census, and the being the number of classified ‘homes’ in every area. In a sense, the factor for expansion is the result of a function of census Output Area residents and users with their ‘home’ in this Output Area. A valid expansion factor would require both the census data to be reliable and the number of ‘home’ Point of Interest, or POIs, to be reliable. If this is not the case, the trips in the matrix get multiplied by the wrong factor and, as a result, misallocated across Output Area zones.
The census population data can be argued to be correct. The final census was conducted in March 2011, and factors for population growth have been applied to find the number of United Kingdom residents in 2015. While these factors of population growth may produce errors, this data should be valid enough for this specific purpose.

The location of the ‘home’ POIs are likely to be of higher concern than the accuracy of the census data. Two issues have already been raised while reflecting on the ‘event generation’ step and the ‘points of interest’ step regarding the assignment of home POIs to a user. If the data is indeed of insufficient quality and if the assumptions prove to be invalid, homes may be assigned the wrong location. The issue with expansion is summarised as follows:

8. A misclassification of ‘home’ POIs leads to wrong factors for expansion, quickly resulting in the matrix being far off reality.

If this is to remain the method for expanding data from sample to population, the step for finding ‘home’ is probably the most important step to get right.

3.6 Classification of ‘road’ and ‘rail’ users
If the method of finding rail users is done incorrectly, road users may incorrectly get classified as rail users, and rail users may incorrectly be classified as road users.

The method is much more difficult to do where there is a railroad next to a motorway. This scenario will result in many ambiguous trips. These trips will then be classified as road. Issue 9 relates to non-obvious users being classified as ambiguous and put in the matrix as road trips. It is expected that the rail mode share is lower than it should be. The main issue raised here is:

9. Trips that are not evidently classified as rail get classified as road trips, potentially missing out on rail trips.

3.7 Creation of routes
A danger of applying this routing engine for removing external area to external area is that a journey may not actually start in the external area, even if it is classified as such. There are cells covering both a part of an external area and a part of a zone within the model simulation area. In this case, it is not possible to know if a trip starting or ending in this cell should be included in the matrix. In addition, external to external area trips may be falsely put in or left out of the matrix by predicting the wrong route. In particular, users with their mobile phone turned off during the entire journey may have the wrong route assigned to the journey.

This step has not been applied on the matrix that is validated in this thesis. As there is no way to test the potential issues with the matrix at hand, the creation of routes will not be covered any further.

3.8 Privacy
After the creation of the trip matrices, the OD pairs containing less than 20 trips observed over the study period of 4 weeks need to be removed from the matrix. This could lead to a higher amount of OD pairs with 0 trips.

10. For privacy reasons, some OD pairs will have 0 trips while there are actual trips observed.

The only ways to bypass this issue is to use larger zones as Output Areas or not split the matrix into hourly matrices. This way, there will be more trips for each zone, reducing the chance that trips have to be removed.
3.9 Hypotheses

Throughout the reflection of the matrix creation methodology, a list of 10 issues has been created. Out of these issues, it is possible to create a list of hypotheses. The 10 identified potential issues with mobile phone data are summarised below, labelled 1 to 10.

Issue 1 - Lack of spatial granularity makes it difficult to find user ‘home’ and ‘work’ location.
Issue 2 - Lack of spatial granularity causes trips within the same cell to be missed.
Issue 3 - Short stays can be ‘hidden’ in between two events.
Issue 4 - Short stays are not classified as such by the threshold of 30 minutes being too high.
Issue 5 - Users not actually going to work are misclassified as going to work.
Issue 6 - Nightshift workers and other similar exceptions have misclassified ‘home’ and ‘work’ POI’s.
Issue 7 - Trip starts may be allocated to the wrong hour.
Issue 8 - A misclassification of ‘home’ POIs leads to wrong factors for expansion, resulting in the matrix being far off reality.
Issue 9 - Trips that are not evidently classified as rail get classified as road trips, potentially missing out on rail trips.
Issue 10 - For privacy reasons, some OD pairs will have 0 trips while there are actual trips observed.

Any of these issues may cause trips to be missing, be in the wrong zone, wrong time period or wrong journey purpose. In order to investigate if this is indeed the case and what the impact is of certain shortcomings of the data and decisions in matrix creation methodology, the following 6 hypotheses have been formulated, labelled A to F.

A. Home locations do not have a correlation with census residential population.
B. The proportion of ‘Home based other’ and ‘Non-home based’ trips in the matrix is too small if compared to the National Trip End Model.
C. The ‘Home based work’ work trip ends have no correlation with census workplace data.
D. The trip origins and destinations have no correlation with the trip origins and destinations in the National Trip End Model.
E. There is no correlation between trip distribution of the mobile phone home based work matrix and the census journey-to-work matrix.
F. Not enough trips have been classified as ‘rail’ if compared to the National Travel Survey. The proportion of rail trips is too low.

In chapter 4, a method of validating the matrices will be discussed. The matrices will be tested to see if the hypotheses are true or false. It is possible to find evidence for the issues listed above using this method of testing.
4 Method of validation

This chapter aims to answer the research question by developing a validation framework and describing the sources of data available to test mobile phone matrices against. A list of shortcomings and assumptions has been made in the previous chapters. The validation framework described in this chapter will be able to provide insight in the quality of the data and the related assumptions and shortcomings. A description of the sources of data used in this framework can be found in 4.1. These datasets are linked to the hypotheses in 4.2. The various zone systems used for this data will be covered in 4.3. The stepwise structure of the framework is discussed in 4.4.

4.1 Validation data sources

For validation, the following sources of data are used:

- Census Journey to Work matrix
- National Trip End Model (NTEM)
- Census residential population
- Census workplace population
- National Travel Survey

These five data sources will now be discussed.

4.1.1 Census Journey to Work Matrix

The Census Journey to Work Matrix is made by the UK government and based on census information of residences and workplaces. This is collected at the level of Census output areas (OA), areas with an average population of 125 households and mostly based on postcode blocks (Office for National Statistics, 2015a). An example of a part of London split in Census output areas is shown below in Figure 1 as an example. The latest available matrix is based on 2011 data, which is the year of the census. There are various levels of output areas. The one used in this analysis is in Middle Layer Super Output Area (MSOA) level, which consists of a combination of Output Areas, because the mobile phone origin-destination matrix is based on this zone system. MSOA zones have roughly 6000 households.

Figure 1: Example of the OA zone system
4.1.2 National Trip End Model (NTEM)
The second data source for validation is the National Trip End Model. The NTEM provides with predictions of the number of trips, split by purpose (Department for Transport, 2013). NTEM utilizes an underlying car ownership model and data from population, household, employment and GDP to predict trip ends by zone on a unique zone system. Trips are split by walk, cycle, car driver, car passenger, bus/coach and rail/underground modes. Data is available for the time periods weekday AM peak (7:00-9:59), weekday inter peak (10:00-15:59), weekday PM peak (16:00-18:59), weekday off peak (0:00 to 6:59), Saturday all day, Sunday all day and the average weekday. Since the mobile phone based matrix is based on the average weekday, only the average weekday and the weekday AM, inter peak and PM matrices are compared to the mobile phone data. Because walk and cycle trips are filtered out of the mobile phone based matrix, only car, car passenger, rail/underground and bus/coach trips are selected from the NTEM database as a validation source.

Because the NTEM database uses its own zone system, data will need to be converted to the MSOA zone system used for the mobile phone matrices. This procedure will be described in 4.3.

4.1.3 Census data
Census is a national count of population performed every 10 years in the United Kingdom. The last census took place in 2011 (Office for National Statistics, 2015b). This data is aggregated on the census output areas (OA) discussed in 4.1.1 and shown in Figure 1. During the 2011 census, data on the amount of workplaces in the United Kingdom was also collected and output was generated on a unique zone system.

4.1.4 National Travel Survey
The National Travel Survey (NTS) is a survey conducted by the United Kingdom government to monitor trends in personal travel (Department for Transport, 2015). The survey is held every year, with the most recent available survey being from September 2014. This includes data on mode choice, journey purpose and travel distance. The survey is used to calculate the split between the various journey purposes (commuting and home based/non-home based trips) and the road and rail modes.

4.2 Hypotheses and datasets
Six hypotheses, corresponding with data shortcomings and assumptions, are discussed in chapter 0. These hypotheses can be tested against the following sources of data, available in the United Kingdom:
Table 1: Validation sources suitable for each hypothesis

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Validation source</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Home locations do not have a correlation with census residential population.</td>
<td>Census residential population</td>
</tr>
<tr>
<td>B. The proportion of ‘Home based other’ and ‘Non-home based’ trips in the matrix is too small, caused by an underestimation of trips with a short stay at the trip end.</td>
<td>National Travel Survey</td>
</tr>
<tr>
<td>C. The ‘Home based work’ work trip ends do not correlate well against census data.</td>
<td>Census workplace population</td>
</tr>
<tr>
<td>D. The trip origins and destinations have no correlation with the trip origins and destinations in the National Trip End Model.</td>
<td>National Trip End Model</td>
</tr>
<tr>
<td>E. There is no correlation between trip distribution of the mobile phone home based work matrix and the census journey-to-work matrix.</td>
<td>Census Journey to Work matrix</td>
</tr>
<tr>
<td>F. Not enough trips have been classified as ‘rail’ if compared to the National Travel Survey. The proportion of rail trips is too low.</td>
<td>National Travel Survey</td>
</tr>
</tbody>
</table>

Hypothesis A is formulated to test if the distribution of ‘home’ locations is correct. A test is test is against the census residential population in each zone. There should be a correlation between ‘home based’ trips and population per zone. If this correlation is not found, there is reason to believe home locations are misclassified, skewing expansion and causing large errors in the matrix.

Hypothesis B is tested against a survey. A proportion of trips of all the journey purposes can be calculated out of the trip records of the National Travel Survey. The proportions will be on the level of the UK mainland. This is on the same spatial scale as the mobile phone based matrix. The sample of the survey is presumed to be representative of the United Kingdom, so the proportions should be comparable. From the way the hypothesis is formulated, the proportion of home based other trips is expected to be lower because these will typically be harder to observe than commute trips having a long dwell at the trip end.

Hypothesis C can be tested against census data regarding workplaces. Trips with their origin or destination in a specific zone must have a correlation with the amount of observed workplaces. This is because workplaces are the attractor for commute (or home based work) trips. If this is not the case, this is evidence hypothesis C is correct. The p-value of the regression can be used to determine if there is a correlation or not. If the p-value is higher than 0.05, hypothesis C will be accepted. If below 0.05, there is a correlation and this hypothesis can be rejected.

The National Trip End model used to test the ‘trip generation’ aspect of the matrix, which is represented by origins and destinations of the zones in the mobile phone data. If there is no correlation (p-value > 0.05) between mobile phone origins and destinations and National Trip End Model generated origins and destinations, there is evidence that hypotheses D may be true. However, another note should be made that a mismatch between the mobile phone trips and the amount of trips estimated by the National Trip End Model may also be due to errors in the National Trip End Model itself. A model is wrong by definition, so it is difficult to accept or reject a hypothesis based on a synthetic model rather than observed data, such as census.

Hypothesis E will also be tested using regression. This hypothesis is concerned with the ‘trip distribution’ aspect. It asks the question of how trips generated in all zones are distributed among the other zones.
Some zones will have a strong production-attraction relationship, e.g. a zone with a large residential population and another zone with many workplaces that is well connected by the road network and by public transport links. These zone pairs should have a high amount of trips in the census journey-to-work matrix as many people have been observed to commute between the zones. There should be a large amount of home based work trips between the same zones in the mobile phone based matrix. If there is no correlation there (p-value > 0.05), hypothesis E will be accepted.

The final hypothesis, hypothesis F, is tested against the National Travel Survey. A proportion of road and rail trips can be calculated out of the trip records of the National Travel Survey. The proportions will be on the level of the UK mainland. This is on the same spatial scale as the mobile phone based matrix. The sample of the survey is presumed to be representative of the United Kingdom, so the proportions should be comparable. From the way the hypothesis is formulated, the proportion of rail trips is expected to be lower.

In conclusion, each of these sources of validation will provide with another insight into an aspect of the mobile phone data. Any aspect of the mobile phone data relates to an algorithm which may or may not need to be revised, depending on the outcome of the validation. This would provide recommendations to changes or improvements in the methodology for the mobile phone providers.

### 4.3 Zone systems

Because of the different sources of data, there are many different zone systems and levels of detail in which the data exists. This section describes how the zone systems were converted. All data sources use different zone systems for their OD matrices. Census data is available at OA level, NTEM comes with its own zone system as discussed in 4.1.2 and the mobile phone data matrix is built for the Middle Layer Super Output Area (MSOA) system, which are output areas that consist of multiple OAs.

Converting OA and workspace zones from census to MSOA zones is quite easy. MSOA zones are larger than OA and LSOA and completely overlap, so simple aggregation by adding up the trip ends in the OA and LSOA zones is sufficient for converting the Census OD matrix to be in the same format as the mobile phone OD matrix. The NTEM trip end data will be harder to convert, since NTEM has its own zone system. The area of overlap between NTEM zones and MSOA zones can be calculated in order to do a conversion from one zone system to another, where the MSOA zones can be assigned a proportion of the NTEM trip ends according to the proportion of overlap between the zones. Figure 2 shows an example of the two zone systems and the zones only partially overlapping. The NTEM zones are displayed in red and the MSOA zones are displayed in blue.
The problem with converting NTEM data to MSOA level using area overlap is that this method assumes equal population density throughout the entire zone system. This will produce errors in highly and sparsely populated MSOA areas, which will get assigned too few or too many trips respectively. To avoid these errors, all zones have been aggregated to region, county and authority levels for the NTEM analysis.

4.4 Framework structure

A framework structure is developed with the data sources discussed in this chapter. This consists of a series of 6 validation steps. The 6 steps are presented in the flowchart in Figure 3. In step 1, high level statistics will be checked against the National Travel Survey. In this case, the percentage of rail trips will be compared, as well as the percentage of trips by journey purpose. Step 2 will compare the daily profile of trips generated during the hours of day against the daily profile of the National Travel Survey. During step 3, trips are examined on the trip end totals according to NTEM on regional, county, authority level. This will give a general overview about the quality of the data without going into much detail. The quality of classification of ‘home’ and ‘work’ locations are then tested against census data in steps 4 and 5. Afterwards, the distribution of home to work trips can be validated in step 6 against a matrix. It should be noted that step 5 can only be performed if steps 3, 4 and 5 prove that the trip totals, home locations and work locations are of sufficient quality. If the trip ends are too low or too high when compared to trip end models, then trip distribution will automatically be off, because the trip generation step is inaccurate.
1. Test mode share and journey purpose split against NTS.

2. Compare against NTS daily profile.

3. Test trip end origin and destinations against the NTEM model.

4. Test home based trip origins against census population data.

5. Test home based work trip destinations against census workplace data.

6. Compare Home based Work trips with the census Journey to Work matrix.

**Figure 3: Validation framework**

This framework can be used for every mobile phone based matrix generated in the United Kingdom. This is provided that an attempt is made to classify home and work locations and create separate matrices for the home based work and home based other journey purposes. The framework allows for all hypotheses in 3.9 to be tested, covering a wide range of aspects of mobile phone data and matrices in general, including mode share, journey purpose share, trip generation and trip distribution. The details of the analyses will be further described in 4.4.1.

**4.4.1 Statistical tests and criteria**

Several statistical tests are formulated to carry out the 6 steps in Figure 3. Based on these tests, the hypothesis in 3.9 can be accepted or rejected. The criteria will be defined in this paragraph.

**4.4.1.1 One-Sample z-test for proportions**

The difficulty of testing two proportions is that one proportion is based on an expanded matrix rather than a statistical dataset. In this case, it is unknown to the researcher what the observed amount of trips is exactly. A solution is to treat the mobile phone data as the population rather than as a statistical sample. This bypasses the need for the sample size to be known. This allows for a One-Sample z-test for proportions to be applied. The sample can be any survey. In this case, the National Travel Survey is used. This test will be undertaken during step 1 of the validation.

The following six steps will be undertaken during the process:

1. Define a null and alternative hypothesis. The null hypothesis is that there is no difference between the mobile phone proportion (for mode or journey purpose, in this case) and the National Travel Survey proportion. The alternative hypothesis is that there is a difference between the two proportions.
2. Define alpha. An alpha of 0.05 will be used in this case.
3. Stating the decision rule. The test will be a two-tailed test, as there is no prior indication of the NTS proportions being over or under the mobile phone proportions. For a two-tailed test with an alpha of 0.05, the z-score needs to be between -1.96 and 1.96. If it is not, the null hypothesis can be rejected.
4. Calculating the z-score. This is done by applying the following formula:

\[ z = \frac{p_s - p_m}{\sqrt{p_m(1 - p_m)/n}} \]

Where:
- \( z \) = z-score
- \( p_s \) = Proportion in survey
- \( p_m \) = Proportion in mobile phone based matrix
- \( n \) = survey sample size

5. Accept or reject the null hypothesis based on the z-score.
6. State the conclusion.

This test will be carried out for the journey purpose for home based work, home based others and non-home based trips as a proportion of the total trips, for every individual English region. It will also be carried out for the proportion of rail trips over the combined motorized road and rail trips.

4.4.1.2 Analysis of peak periods
The purpose of this analysis is to investigate the overall trend of trips made throughout the day. A clear AM peak and PM peak should be observed in the mobile phone data, which should be in the same hour of day as the National Travel Survey data. This applies to test 2 in the matrix validation framework in Figure 3. Three criteria are formulated so that the trend lines can be compared:

1. The AM peak of the mobile phone based matrix should be in the same hour as the AM peak of the National Travel Survey.
2. The PM peak of the mobile phone based matrix should be in the same hour as the PM peak of the National Travel Survey.
3. The inter-peak of the mobile phone based matrix should be minimal in the same hour as the inter-peak of the National Travel Survey is minimal.

It is not possible to distinguish weekend days and weekdays from the NTS data. Because it is a household survey, children under the age of 16 are included in the NTS data, while this is filtered out in the mobile phone data due to privacy reasons. To make the NTS consistent with the mobile phone data, only commute trips will be plotted in this profile.

4.4.1.3 Regression analysis for matrix comparison
Regression analysis is an often used method for testing matrices against other matrices, or matrices against data such as census (see the literature review in 2.2). These tests will be carried out on different spatial levels, being Region, County and Unitary Authority in the United Kingdom. This can translate to different spatial levels in the country or geographical region the framework is being applied on, if applied outside of the United Kingdom. It may be that the mobile phone matrices are only applicable on some spatial levels. This analysis will provide this insight.

An indicator for quality of model fit is the R² coefficient. The problem with formulating a criterion based on this statistic is that it is subjective when the R² can be considered ‘good’. This is dependent on the application. For example, if a matrix is to be compared to a similar matrix, a higher R² is expected than if the matrix is compared to for example census data. In the latter case, just observing a correlation without a perfect fit is sufficient.

A solution to this is to use the p-value. The null hypothesis is that there is no correlation between the variables. This hypothesis can be rejected if a p-value of 0.05 means that. If the p-value is below 0.05, it
means that there is a correlation between the variables. The criterion formulated for tests 3 through 6 in Figure 3 is that the p-value is below 0.05. Further interpretation is possible by looking at the $R^2$. A p-value below 0.05, but having a low $R^2$, would mean that there is a correlation between the variables, but not a strong one.

The zones being tested and plotted in the regression can vary widely by size. Typically, a zone in the study area can be as small as a town centre, while a zone far away from the study area can be as large as Scotland. This will cause the $R^2$ to be artificially high, because the residual error of the smaller data points becomes small due to the large scale. See Figure 4 for an example.

![Figure 4: Mobile phone road and rail trips vs. census residential population](image)

It is hard to interpret such a figure and the corresponding $R^2$. A solution is found in a paper by Calabrese et al., where a logarithmic scale was used (Calabrese, Di Lorenzo, Lui, & Ratti, 2011). A logarithmic scale brings the data points closer to each other and makes meaningful analysis possible. The data points will be transformed by the natural logarithm before plotting the data in a regression chart. Applying the natural logarithm will transform the chart as following (Figure 5):
Figure 5: Mobile phone road and rail trips vs. census residential population after logarithmic transformation

A difference between Figure 4 and Figure 5 is that the latter has a lower $R^2$. This is because the range is smaller, causing the error in the data points with a lower value to account for more. It is therefore more meaningful to interpret this $R^2$. It also makes it easier to observe the single outlier on the lower end of the chart. For reasons mentioned in this section, all regression charts will be logarithmically transformed before analysis is undertaken.
5 Validation & results
The framework described in chapter 4 is carried on the South East Regional Transport Model (SERTM) matrix data. This matrix is produced by the methodology described in chapter 2. Two sets of matrices are produced, a car only matrix and a matrix with rail added to the car trips. Because the ‘clustering’ methodology for rail is insufficiently tested, it is unclear what the quality of the rail matrix is. Therefore, these two sets of matrices have both been tested against all data sources. These analyses, as well as the results, are described in this chapter. The results are presented in the order of the framework displayed in Figure 3.

Paragraph 5.1 reports some high-level statistics of the mobile phone data and the National Travel Survey. Mode share and journey purpose split is included here. This corresponds to step 1 in the framework. Paragraph 5.2 discusses the trip profile of the matrix throughout the day, as well as the profiles of data from the National Travel Survey and traffic counts throughout the United Kingdom. In the framework, this is step 2. Paragraph 5.3 gives an overview on how exactly a comparison against available data from the National Trip End Model (NTEM) was carried out and presents the results on the trip end comparison, which is step 3 in the framework. In 5.4, the South East Regional Model matrix data is matched against residential population and workplaces. This is step 4 and 5. The final part of the validation using census Journey to Work data, presented in paragraph 5.5 and corresponds to step 6. The method will be reflected upon in paragraph 5.6.

5.1 Mode share and journey purpose share
The first test reports some high level statistics from NTS and the mobile phone data, concerning proportions of trip classifications into mode and purpose. Paragraph 5.1.1 contains a description of the statistics on mode share. The statistics on journey purpose share can be found in 5.1.2.

5.1.1 Mode share
All trips in the mobile phone based matrix without rail were summed up by region. The same was done for the mobile phone based matrix with rail, so that the percentage of road and rail trips can be calculated based on the difference. The percentage of rail trips from the NTS is also based on the sum of all road modes and all rail modes (including London Underground). Only 2014 data is selected from the NTS, which is the latest available, to better correspond to the mobile phone trips. The NTS includes English regions only, so Scotland and Wales are not considered here. See Table 2 for the test results.
Table 2: Modal split tests

<table>
<thead>
<tr>
<th>Region</th>
<th>NTS Rail Share</th>
<th>Mobile Phone Rail Share</th>
<th>NTS Sample</th>
<th>z-score</th>
<th>Null hypothesis accepted/rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>North East</td>
<td>0.772%</td>
<td>0.432%</td>
<td>11399</td>
<td>5.526048</td>
<td>Rejected</td>
</tr>
<tr>
<td>North West</td>
<td>1.521%</td>
<td>1.321%</td>
<td>29052</td>
<td>2.994534</td>
<td>Rejected</td>
</tr>
<tr>
<td>Yorks &amp; Humber</td>
<td>1.445%</td>
<td>0.711%</td>
<td>22631</td>
<td>13.14039</td>
<td>Rejected</td>
</tr>
<tr>
<td>East Midlands</td>
<td>0.772%</td>
<td>0.674%</td>
<td>18257</td>
<td>1.615552</td>
<td>Not rejected</td>
</tr>
<tr>
<td>West Midlands</td>
<td>1.803%</td>
<td>1.461%</td>
<td>22021</td>
<td>4.224963</td>
<td>Rejected</td>
</tr>
<tr>
<td>East of England</td>
<td>1.855%</td>
<td>2.784%</td>
<td>22746</td>
<td>-8.5159</td>
<td>Rejected</td>
</tr>
<tr>
<td>London</td>
<td>19.479%</td>
<td>8.977%</td>
<td>27594</td>
<td>61.03272</td>
<td>Rejected</td>
</tr>
<tr>
<td>South East</td>
<td>2.558%</td>
<td>2.619%</td>
<td>34758</td>
<td>-0.71862</td>
<td>Not rejected</td>
</tr>
<tr>
<td>South West</td>
<td>0.621%</td>
<td>0.706%</td>
<td>23031</td>
<td>-1.54927</td>
<td>Not rejected</td>
</tr>
</tbody>
</table>

The null hypothesis is that the NTS rail share and the mobile phone rail share is the same. This is only the case in the East Midlands, South East and South West regions, but is different in the 6 other regions being investigated here. The difference is greatest in London (more than 10 percent point), possibly because of the underestimation of deep Underground trips. Where the null hypothesis is rejected, there is an underestimation of rail trips in all regions besides the East of England region. An explanation for this systematic underestimation is that all trips that cannot be classified as either road or rail by the clustering method (the ambiguous trips) are automatically classified as road.

5.1.2 Journey purpose share
The trips for each individual journey purpose were summed up by region from the mobile phone road and rail matrix. The same is done for NTS trips using the purpose to and purpose from categories. This allowed for creating home based work trips, home based other trips and non-home based trips. Only 2014 data is selected from the NTS, which is the latest available, to better correspond to the mobile phone trips. The NTS includes English regions only, so Scotland and Wales are not considered here. See Table 3 for the test results.
### Table 3: Home based work proportion tests

<table>
<thead>
<tr>
<th>Region</th>
<th>NTS Home based work proportion</th>
<th>Mobile Phone Home based work proportion</th>
<th>NTS Sample</th>
<th>z-score</th>
<th>Null hypothesis accepted/rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>North East</td>
<td>0.160</td>
<td>0.242</td>
<td>13149</td>
<td>-21.849</td>
<td>Rejected</td>
</tr>
<tr>
<td>North West</td>
<td>0.163</td>
<td>0.241</td>
<td>33178</td>
<td>-33.377</td>
<td>Rejected</td>
</tr>
<tr>
<td>Yorks &amp; Humber</td>
<td>0.153</td>
<td>0.239</td>
<td>25936</td>
<td>-32.334</td>
<td>Rejected</td>
</tr>
<tr>
<td>East Midlands</td>
<td>0.148</td>
<td>0.256</td>
<td>20357</td>
<td>-35.272</td>
<td>Rejected</td>
</tr>
<tr>
<td>West Midlands</td>
<td>0.160</td>
<td>0.250</td>
<td>24529</td>
<td>-32.544</td>
<td>Rejected</td>
</tr>
<tr>
<td>East of England</td>
<td>0.150</td>
<td>0.264</td>
<td>25925</td>
<td>-41.885</td>
<td>Rejected</td>
</tr>
<tr>
<td>London</td>
<td>0.204</td>
<td>0.278</td>
<td>32388</td>
<td>-29.741</td>
<td>Rejected</td>
</tr>
<tr>
<td>South East</td>
<td>0.154</td>
<td>0.251</td>
<td>40003</td>
<td>-45.089</td>
<td>Rejected</td>
</tr>
<tr>
<td>South West</td>
<td>0.133</td>
<td>0.249</td>
<td>26498</td>
<td>-43.592</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

The home based work proportions do not match between the NTS and the mobile phone data. This is implied by all z-scores being under -1.96. There are too many home based work trips in the matrix. All null hypotheses are rejected. See Table 4 for the test results.

### Table 4: Home based other proportion tests

<table>
<thead>
<tr>
<th>Region</th>
<th>NTS Home based other proportion</th>
<th>Mobile Phone Home based other proportion</th>
<th>NTS Sample</th>
<th>z-score</th>
<th>Null hypothesis accepted/rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>North East</td>
<td>0.717849</td>
<td>0.566844</td>
<td>13149</td>
<td>34.94501</td>
<td>Rejected</td>
</tr>
<tr>
<td>North West</td>
<td>0.705467</td>
<td>0.569434</td>
<td>33178</td>
<td>50.04124</td>
<td>Rejected</td>
</tr>
<tr>
<td>Yorks &amp; Humber</td>
<td>0.710248</td>
<td>0.575518</td>
<td>25936</td>
<td>43.89921</td>
<td>Rejected</td>
</tr>
<tr>
<td>East Midlands</td>
<td>0.710665</td>
<td>0.575866</td>
<td>20357</td>
<td>38.91614</td>
<td>Rejected</td>
</tr>
<tr>
<td>West Midlands</td>
<td>0.696441</td>
<td>0.580918</td>
<td>24529</td>
<td>36.66932</td>
<td>Rejected</td>
</tr>
<tr>
<td>East of England</td>
<td>0.707155</td>
<td>0.569024</td>
<td>25925</td>
<td>44.91175</td>
<td>Rejected</td>
</tr>
<tr>
<td>London</td>
<td>0.684111</td>
<td>0.414575</td>
<td>32388</td>
<td>98.46269</td>
<td>Rejected</td>
</tr>
<tr>
<td>South East</td>
<td>0.709172</td>
<td>0.560309</td>
<td>40003</td>
<td>59.98529</td>
<td>Rejected</td>
</tr>
<tr>
<td>South West</td>
<td>0.703902</td>
<td>0.588723</td>
<td>26498</td>
<td>38.10273</td>
<td>Rejected</td>
</tr>
</tbody>
</table>
The home based other proportions in the mobile phone based matrix are underestimated if compared to the National Travel Survey. This is implied by all z-scores being higher than 1.96. All null hypotheses are rejected. See Table 5 for the test results.

Table 5: Non-home based proportion test

<table>
<thead>
<tr>
<th>Region</th>
<th>NTS non-home based proportion</th>
<th>Mobile Phone non-home based proportion</th>
<th>NTS Sample</th>
<th>z-score</th>
<th>Null hypothesis accepted/rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>North East</td>
<td>0.12191</td>
<td>0.191331</td>
<td>13149</td>
<td>-20.2375</td>
<td>Rejected</td>
</tr>
<tr>
<td>North West</td>
<td>0.131593</td>
<td>0.189218</td>
<td>33178</td>
<td>-26.7981</td>
<td>Rejected</td>
</tr>
<tr>
<td>Yorks &amp; Humber</td>
<td>0.136528</td>
<td>0.185656</td>
<td>25936</td>
<td>-20.3477</td>
<td>Rejected</td>
</tr>
<tr>
<td>East Midlands</td>
<td>0.141033</td>
<td>0.16791</td>
<td>20357</td>
<td>-10.2595</td>
<td>Rejected</td>
</tr>
<tr>
<td>West Midlands</td>
<td>0.143096</td>
<td>0.168584</td>
<td>24529</td>
<td>-10.6624</td>
<td>Rejected</td>
</tr>
<tr>
<td>East of England</td>
<td>0.143259</td>
<td>0.166682</td>
<td>25925</td>
<td>-10.119</td>
<td>Rejected</td>
</tr>
<tr>
<td>London</td>
<td>0.111986</td>
<td>0.307489</td>
<td>32388</td>
<td>-76.2461</td>
<td>Rejected</td>
</tr>
<tr>
<td>South East</td>
<td>0.13704</td>
<td>0.188076</td>
<td>40003</td>
<td>-26.1216</td>
<td>Rejected</td>
</tr>
<tr>
<td>South West</td>
<td>0.162616</td>
<td>0.161941</td>
<td>26498</td>
<td>0.298426</td>
<td>Not rejected</td>
</tr>
</tbody>
</table>

There is no match between the NTS non home based proportion and mobile phone non-home based proportion except for in the South West region. Where the null hypothesis is rejected, there is an overestimation of non-home based trips.

5.2 Daily profile

The mobile phone data from the South East Regional Transport Model (SERTM) was summed up by starting hour and plotted in a graph. This creates a profile, where traffic intensity throughout the day, including peak periods, can be observed. The profile is plotted in Figure 6, along with the profiles created in the same way using the National Travel Survey (NTS). The horizontal is the time scale with hours of the day. Both the SERTM and NTS data have been summed up by hour starting. This means that if a trip in SERTM or NTS starts at any time between 8:00 and 9:00, the trip is put in the ‘8’ band. The vertical is the proportion of trips starting, in case of SERTM or NTS, or observed, in case of the traffic counts. The sum of the proportions between 7:00 and 18:59 is 1, so these are the only trips being taken into account. The reason for this limitation is that no hourly mobile phone is in the matrix outside of this time frame.

Two graphs have been plotted in Figure 6. The SERTM mobile phone data for car and rail is plotted in red. The same is done for the NTS data, plotted in blue. Only home based work trips were selected from the mobile phone data and only commute trips were selected from the NTS.
The NTS data and mobile phone based matrix from the SERTM have the same PM peak between 17:00 and 18:00, so the mobile phone data passes on this criterion. The PM peak in the NTS is sharper than it is in the mobile phone data. The AM peak happens between 7:00 and 8:00 in the NTS while the AM peak happens between 8:00 and 9:00 in the mobile phone data. The mobile phone data does not pass on this criterion. The inter-peak proportion of trips is minimal between 11:00 and 12:00 in both the mobile phone based matrix and the NTS. The mobile phone data passes on this criterion as well. In total, the mobile phone data passes on 2 out of 3 criteria. The NTS has sharper peak periods, where the mobile phone data has its trips more spread out throughout the day. Also, the AM peak seems to be broader in the mobile phone data.

5.2.1 Reflection on National Travel Survey

In practice, there is a list of problems which may make the NTS (or surveys in general) a less suitable source of data to use for mobile phone validation as a source for ground truth.

- A survey usually contains errors such as a sampling bias and has a relatively low sample size. The survey contains about 15000 respondents every year, with a response rate of about 63% (NatCen, 2014). The participants for NTS are randomly selected from an address base, so the sample should be relatively random, although this does not eliminate the possibility of response bias. Only 2014 data has been selected, making the sample size relatively small. For the daily profile, around 32000 trips were disaggregated into hourly bands, resulting in some hours having less than a 1000 trips in the sample. It can be argued that this sample is too small.

- A trip in NTS can be broken up into two or more trips by the mobile phone data processing algorithms. This happens if a user makes a stop somewhere along the journey for more than 30 minutes, and generates two events. This causes the mobile phone data to have a higher proportion of trips in certain time periods than there would be according to NTS.
This leads to a problem in that a mismatch between mobile phone data and NTS, as found in tests in 5.1 and 5.2, does not necessarily mean that the mobile phone data is wrong. A survey cannot be considered ‘ground truth’ for reasons mentioned above, although it does provide some evidence that there are some problems with the mobile phone data.

5.3 National Trip End Model

The next step in the validation process is the comparison against the National Trip End Model (NTEM). NTEM contains data on trip ends for each purpose and mode of travel. It uses its own zoning system, which is different from the MSOA based zone system used for the SERTM matrix and therefore incompatible at the most detailed level. This is why all data has been aggregated to United Kingdom regions, counties and authorities, so that a valid comparison can be made on these levels. The starting point of the analysis has been the regional comparison, so that it is possible to see if the data is useable on the regional level. If it is not, the analysis can stop here, but if it is, the analysis can move on to county level, and possibly on to authority level.

Besides spatial granularity, time of day and trip purpose can also be varied throughout the analysis. Three selections have been made for time period and journey purpose:

- All times of day, all purposes
- Morning (AM) Peak period, Home based work outbound trips only
- Afternoon (PM) Peak period, Home based work inbound trips only

The data was selected from the Trip End Model Program (TEMPro), which defines the AM peak as the period between 7:00 and 9:59 and the PM peak as the period between 16:00 and 18:59. TEMPro does not separate Home based work outbound and inbound trips, both fall under the category 'commuting'. In order to be able to make a comparison, commuting trips in the AM peak period are all assumed to be Home based work outbound trips, while commuting trips in the PM peak are Home based work inbound trips. All motorized road and rail modes have been included in both TEMPro and the mobile phone matrices.

5.3.1 Trip End validation, all day, all purposes

Firstly, the analysis is done for the trip ends throughout the entire day with all journey purposes combined. Afterwards, more specific times of day and purpose are analysed in isolation.

5.3.1.1 Region analysis

The trip ends for both the origin and destination side have been aggregated over the 11 regions in the United Kingdom (Northern Ireland excluded) and summed up. The charts can be found in appendix i, Figure 7 and Figure 8.

There are generally more trips predicted by the NTEM model than there are observed using the mobile phone data. This is partly because of the short trips and short stays missing from the mobile phone data, and partly because of different definitions of a trip. In NTEM, it is assumed that a person makes a single trip to perform a single activity. For example, a trip undertaken to perform the activity ‘work’ is a commuting trip. In mobile phone data classification algorithms, activity is unknown, so location is used to classify trip purpose. This means that commuting (home based work outbound and inbound) trips are only the trip starting in the home location and ending in the work location. If a trip is broken in two because a user stopped in a location that is not work for more than 30 minutes in between home and work, the user will have made two trips (both not commuting), rather than a single commuting trip.
Because of these fundamental differences, a regression analysis is chosen as a method. The hypothesis is that the more trip ends are observed from the mobile phone data, the more trip ends are predicted by NTEM, and vice versa. Rather than looking at absolute numbers and absolute differences, a ‘trend’ is being analysed across the zones with $R^2$ as an indicator for data quality, as proposed in 4.4. The logarithmic scale is applied on all regression analyses from this point on. Applying the logarithmic scale for comparing data points in a wide range is common practice, examples being Wang et al. (2015) and Calabrese et al. (2011). This method leads to the following set of $R^2$ values.

<table>
<thead>
<tr>
<th>Region regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.885</td>
<td>0.887</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>69</td>
<td>71</td>
</tr>
</tbody>
</table>

The p values are close to zero, which means that there is a correlation between the variables. ($p<0.05$)

The $R^2$ values indicate a good match between NTEM and the mobile phone data, not in terms of absolute values, but more trips are observed where more trips are predicted by the model. The origin and destination ends are similar because origins and destinations on regional level have mostly the same values – trips starting in a UK region usually end in the same region.

5.3.1.2 County analysis

The same analysis was carried out with data on the county level. This introduces the problem that some of the bigger zones from SERTM consist of two counties, making it impossible to compare with NTEM data without making more assumptions. The counties that are larger than the SERTM zones are removed from the analysis. 37 out of the 93 UK counties remain. The results are listed below. The charts can be found in appendix i, Figure 9 and Figure 10.

<table>
<thead>
<tr>
<th>County regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>572</td>
<td>599</td>
</tr>
</tbody>
</table>

The p values are close to zero, which means that there is a correlation between the variables ($p<0.05$). The analysis on county level still yields high $R^2$ values, indicating a good match with the NTEM model. This makes mobile phone data useable on the county level.

5.3.1.3 Authority analysis

The analysis is repeated for the authority level. Again, SERTM zones that are made up of multiple authorities are removed from the analysis, because the available data is not detailed enough. This leaves 235 out of the 411 United Kingdom authorities. The results are listed below. The charts can be found in appendix i, Figure 11 and Figure 12.

<table>
<thead>
<tr>
<th>Local authority regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>46</td>
<td>47</td>
</tr>
</tbody>
</table>
On this level of detail, the $R^2$ becomes lower. This implies a larger scatter, with several authorities having a much higher or much lower amount of trip ends than predicted by the NTEM, both on the origin and destination end. The $p$ value of the regression is still below 0.05, so the hypothesis that there is no correlation between NTEM and the mobile phone data can be rejected.

It is interesting to see which authorities in particular cause these $R^2$ to be lower. An effort has been made to remove authorities from the analysis in order to increase the $R^2$ values. This is done by identifying authorities with a bad match by eye, then removing them until the $R^2$ values got up to acceptable levels. The following list of authorities has been removed in this process, in addition to the Inner London authorities, where the data is assumed to be of lesser quality:

- Charnwood
- Bath & NE Somerset
- City of Bristol
- Bournemouth
- East Dorset
- North Dorset
- Purbeck
- Gloucester
- North Wiltshire
- Nuneaton and Bedworth
- Warwick
- Coventry
- Solihull
- Stratford-on-Avon
- Christchurch
- Mendip
- South Somerset
- West Wiltshire

Having identified these ‘problematic’ authorities, a map can be created showing the geographic location of these problematic areas. These charts can be found in appendix iv, Figure 25, Figure 26, Figure 27 and Figure 28. The areas in grey are omitted because they are aggregated beyond authority level for the SERTM. The areas in red are the authorities listed above, with the addition of Inner London. Of the remaining zones, the half with the least good match with NTEM is coloured orange while the half with the best match with NTEM is coloured green. This is based on the distance from the regression line, where the data point is predicted to be.

It is difficult to tell by geographical location why the authorities coloured in red or orange have a bad fit. Most of the better scoring areas are in the southeast of the United Kingdom, with the exception of London and surrounding towns. This may be caused by a higher market share of the mobile phone provider in this region. A higher market share will lead to a higher sample size and will cause the expansion factor to be lower, painting a more accurate picture of the trips made in the area. This will need to be confirmed with external sources for regional market share of the mobile phone provider, which are not currently available to this research.
5.3.2 Trip End validation, AM Peak, home based work outbound
This paragraph describes the results of the analysis for the AM peak (7:00 to 9:59) for home based work outbound trips only.

5.3.2.1 Region analysis
The $R^2$ value table for the United Kingdom regions is shown below. The charts can be found in appendix ii, Figure 13 and Figure 14.

<table>
<thead>
<tr>
<th>Region regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.953</td>
<td>0.93</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>182</td>
<td>120</td>
</tr>
</tbody>
</table>

This shows a good $R^2$ for the AM peak considering commuting trips starting at home and ending at work. The p value of the regression is below 0.05, so the hypothesis that there is no correlation between NTEM and the mobile phone data can be rejected.

5.3.2.2 County analysis
The $R^2$ value table for the United Kingdom counties is shown below. The charts can be found in appendix ii, Figure 15 and Figure 16.

<table>
<thead>
<tr>
<th>County regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.924</td>
<td>0.896</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>427</td>
<td>299</td>
</tr>
</tbody>
</table>

The $R^2$ drops slightly if compared to the region level. The p value is below 0.05, so there is a correlation between NTEM and the mobile phone data on the county level.

5.3.2.3 Authority analysis
The $R^2$ value table for the United Kingdom authorities is shown below. The charts can be found in appendix ii, Figure 17 and Figure 18.

<table>
<thead>
<tr>
<th>Local authority regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.871</td>
<td>0.79</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>1567</td>
<td>880</td>
</tr>
</tbody>
</table>

The $R^2$ for commuting trips in the AM peak scores much better than the all day trips. A reason for this could be that the AM peak has a high proportion of trips that are strictly commuting, and these types of trips can be more easily observed by mobile phone data than other types of trips that have shorter times of stay, such as shopping trips. The $R^2$ on the destination end is lower than on the origin end. This implies that the inferred work location is less accurate than the inferred home location.

5.3.3 Trip End validation, PM Peak, home based work inbound
This paragraph describes the results of the analysis for the AM peak (16:00 to 18:59) for home based work outbound trips only. These are trips starting at the work location and ending at the home location.
5.3.3.1 Region analysis
The $R^2$ value table for the United Kingdom regions is shown below. The charts can be found in appendix iii, Figure 19 and Figure 20.

<table>
<thead>
<tr>
<th>Region regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.937</td>
<td>0.953</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>134</td>
<td>182</td>
</tr>
</tbody>
</table>

The $R^2$ on region level is high on all matrices. The p value of the regression is below 0.05, so the hypothesis that there is no correlation between NTEM and the mobile phone data can be rejected.

5.3.3.2 County analysis
The $R^2$ value table for the United Kingdom counties is shown below. The charts can be found in appendix iii, Figure 21 and Figure 22.

<table>
<thead>
<tr>
<th>County regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.922</td>
<td>0.939</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>417</td>
<td>542</td>
</tr>
</tbody>
</table>

On the county level, the $R^2$ remains high for all matrices. The origin end scores worse than the destination end. Since the origin end is work and the destination end is home, this would imply that the work end is estimated less well than the home end. The p value of the regression is below 0.05, so the hypothesis that there is no correlation between NTEM and the mobile phone data can be rejected.

5.3.3.3 Authority analysis
The $R^2$ value table for the United Kingdom authorities is shown below. The charts can be found in appendix iii, Figure 23 and Figure 24.

<table>
<thead>
<tr>
<th>Local authority regression</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ value</td>
<td>0.218</td>
<td>0.196</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>65</td>
<td>56</td>
</tr>
</tbody>
</table>

Similar to the ‘all day’ analysis, the PM peak $R^2$ values are far worse on the authority level than they are on the county and region levels. It is interesting to compare this to the AM peak analysis and conclude that the PM peak is a lot more ‘noisy’ than the AM peak. This is possibly because more short trips take place towards the end of the day rather than the beginning, where activities such as grocery shopping and visiting friends and relatives are generally performed. Removing the same problematic zones that were removed in 5.3.1.3 brings the $R^2$ up to a more acceptable level. It is suspected that there is a problem with expansion in authorities where the mobile phone providers’ market share is low. The p value of the regression is below 0.05, so the hypothesis that there is no correlation between NTEM and the mobile phone data can be rejected.

5.3.4 Final note on NTEM
As was mentioned before, the NTEM is based on several models that are mostly being used for predicting growth factors. Like any model, these models are based on assumptions which may be false or overly simplified. Furthermore, traditional transport models have more difficulties dealing with
phenomena such as trip chaining. This gives another definition to a commuting trip, which, in NTEM, may be a chain of trips leading from home to work. It is therefore wrong to take these models as ‘ground truth’ and conclude that the mobile phone data is erroneous on the authority level just because it does not match up well with NTEM. What this analysis was useful for is finding that the PM period is harder to observe than the AM period, providing some evidence that short stays may be more problematic to observe with mobile phone data than, for instance, the conventional commuting trip.

5.4 Census population and workplace data

Whereas the previous part of the analysis compares the mobile phone data to a model, this step aims to do the comparison with something much closer to what can be considered ‘ground truth’. A correlation between census and the mobile phone data is expected because of the simple assumption that residents produce trips and workplaces attract trips. Of course, population and workplaces are not the only factors that cause trips, so correlation is not expected to be perfect. The more homogenous the analysed zones are in terms of factors such as wealth, land type and employment rate, the higher the correlation between population/workplaces and mobile phone trips are.

The comparison takes place on MSOA level, which can be done because it is consistent with the SERTM zone system. The SERTM zones consist of one or multiple MSOA zones. First, the entire SERTM, containing all of mainland UK, will be tested. Afterwards, only zones of the SERTM that consist of 1 MSOA will be compared. This analysis will be concluded with some more specific tests for Inner London and Outer London zones. For consistency purposes, only trips with the home based work outbound purpose will be used. The home end of the trip will be tested against population while the work end of the trip will be tested against the number of workplaces of the zone. Besides testing data quality, the journey purpose classification can also be tested in this manner. The number of trips is the sum of the home based work outbound trips throughout the entire day.

5.4.1 United Kingdom level

For all SERTM zones that consist of a multitude of MSOA’s, the corresponding MSOA zones have been aggregated. The scatterplots on which these R² values are based can be found in appendix v, Figure 29 and Figure 30.

To reduce the problem of upscaling the chart axes, a logarithmic transformation has been applied here, as was already done in 5.3. If the natural logarithm (ln) of all the data points is plotted into a regression analysis, the R² values are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Home end</th>
<th>Work end</th>
</tr>
</thead>
<tbody>
<tr>
<td>R² value</td>
<td>0.889</td>
<td>0.643</td>
</tr>
<tr>
<td>p value</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F value</td>
<td>25537</td>
<td>5761</td>
</tr>
</tbody>
</table>

Applying the natural logarithm allows to more clearly see the scatter. There is still a high correlation for the home end, and a slightly lower correlation on the work end. The p value is 0 for both ends, so there is a statistical correlation between census data and mobile phone data.

5.4.2 Final remarks on census

Census can be considered to be close to ground truth, making it a very valid source for comparison. However, as can be seen in this paragraph, there are some problems.

- No perfect correlation will ever be observed, even if the mobile phone data quality is perfect. This is because one resident does not directly correspond to a constant amount of trips. The
amount of trips one person generates is dependent on a large amount of factors. This makes setting hard criteria a mobile phone based matrix has to pass to be valid impossible. It may be more useful to look at the shape of the graph rather than the $R^2$ and see if a correlation between population and number of trips can be spotted.

- Zones with significantly higher population or workplaces scale the $R^2$ values towards high numbers. Even a very poor dataset can reach a high $R^2$ value in this manner. Every graph should be looked at, and a certain amount of zones will most likely have to be removed in order to do a meaningful analysis. Besides problems with this process being quite arbitrary, somebody performing the analysis this way will not get to test the high population or workplace zones for quality, as these are simply omitted from the regression. Applying the logarithm to the data helps solve this problem without removing the high population or workplace zones by bringing faraway data points close together. This is a recommended method for dealing with this type of problem.

This analysis could not provide hard evidence that the mobile phone based matrix quality for the SERTM is good or bad. However, it implies that the home-end of the trips corresponds well to the population, meaning that the home of a user is placed in the right MSOA in a good amount of cases. The work-end of the trip has more problems, with too many or too few users travelling to a MSOA for ‘work’ in a larger amount of cases.

So far, only tests against trip ends have been done. This tests the quality of the production and attraction of zones, according to mobile phone observations. The next paragraph will proceed with an analysis on actual Origin-Destination pairs and will therefore test the ‘trip distribution’ part of the matrix.

### 5.5 Journey to Work validation

The census Journey to Work matrix provides a number of trips on MSOA level. This is similar to the SERTM matrix, making the comparison straightforward. Just like with census population and workplaces, only zones consisting of a single MSOA have been selected.

#### 5.5.1 Matrix comparison

The trips of OD pairs of a single MSOA zone to a single MSOA zone have been analysed here. There are 2914 single MSOA zones, resulting in 478490 pairs. This excludes the pairs without any trips in either the Journey to Work or the SERTM matrix, because it is not possible to do a logarithmic transformation of 0. The $R^2$ values are 0.415 for the car and rail matrix versus the Journey to Work matrix ($p$ value = 0, $F$ value = 339564). While the graph in Figure 31 could be considered noisy, a $p$ value under 0.05 means that there is a statistical correlation between the two matrices. Two cases stand out from the graph.

I. The number of Home based work trips in the SERTM matrix is zero or close to zero, while the number of trips in the Journey to Work matrix is relatively high.

II. The number of Home based work trips in the SERTM matrix is very high, while the number of trips in the Journey to Work matrix is zero or close to zero.

These cases have been further investigated. Most of the times when the number of trips in either matrix is high, the trip is intrazonal or the corresponding origin and destination zones are relatively close. This is expected, as the cost of travelling between zones together is low. Considering case 1, it would make little sense for an intrazonal OD pair to have zero or close to zero trips. Looking at the journey purpose split for these specific OD pairs, it turns out that all or almost all trips fall under the ‘home based other’ and ‘non-home based’ categories. Considering case 2, the opposite occurs. There are close to zero ‘home based other’ and ‘non-home based’ trips, while there are relatively many ‘home based work’ trips. This is
strong evidence that journey purpose classification is still problematic. This is in line with the observations made in 5.1.2.

5.6 Reflection on validation methodology

Performing a validation in this manner comes with many problems already discussed. Most are caused by there being no actual ‘ground truth’; making it impossible to know if the data being validated is actually a match with reality, defeating the purpose of validation.

The method chosen in this analysis was regression. Regression is viable for comparing two datasets with different units, as was the case with the census analysis, and varying definitions of the same unit, as was the case with NTEM trip ends. Regression analysis also comes with its own problems. A single point of data, being much larger than all the other data points, can increase the scale of the axes in such a way that $R^2$ becomes much higher without the correlation being significantly better. This issue has been addressed by applying a logarithmic transformation of the data. Caution is therefore advised while using $R^2$ as a criterion, as there is no threshold for $R^2$ values implying that the match is ‘good’ or ‘bad’. Looking at the regression chart does however provide insight in how good the data matches the regression model. All of the regression models plotted in this chapter have a statistical fit, even though $R^2$ is sometimes low. It can therefore be said that there is a correlation between the mobile phone data and the external data sources, even though there is a lot of noise in the data. It is interesting to look into this noise into more detail and find the underlying problem.

Another problem with the linear regression performed in this chapter is that it only compares one variable to another. In reality, there are many variables causing a zone to generate or attract trips, with population and number of workplaces being only two of them. A low $R^2$ does not necessarily imply that the validated data is bad, but could also mean that there is a lot of variance of the other explanatory variables among the zones. This can be observed in 5.4, where eliminating some of the variance among the explanatory variables by selecting a subset of the data causes $R^2$ to increase.

Despite all these problems, insight is gained in shortcomings of the data and matrix creation methodology. This will be discussed in detail in chapter 6. A discussion on potential improvement of the matrix creation methodology can also be found in chapter 6.

In the chapters up to now, a framework for the validation of matrices created using mobile phone event data has been proposes and used for validating a set of matrices on various aspects. As the set of matrices being validated in this thesis are a prior matrices, there is much room for improvement, which is indeed happening at the stage of which this is written. The thesis will end with recommendations directed at mobile phone providers and researchers attempting to develop their own algorithms for creating mobile phone matrices.

A discussion on the 6 hypotheses will take place in paragraph 5.6.1. This will lead to the conclusion, which can be found in paragraph 5.6.2. Finally, paragraph 5.6.3 contains the discussion on the usefulness of this thesis and future work that can be done.

5.6.1 Discussion on the hypotheses

The 6 hypothesis formulated in chapter 0 are listed here again. An attempt will be made to reject the hypotheses if evidence is found during the validation in this chapter.

A. Home locations do not have a correlation with census residential population.
The comparison of trip ends against census population did not show to be distorted on MSOA level. A correlation was found between origin trip ends of home based trips and census population, especially in homogenous zones. This hypothesis can be rejected based on the census validation results.

_Hypothesis A is rejected._

**B. The proportion of 'Home based other' and 'Non-home based' trips in the matrix is too small.**

From comparing the home based other trips to the National Travel Survey, it turns out that the proportion of home based other trips is too low. Looking at NTS data, a lot of home based other trips are 'escort education' trips, which is bringing somebody else (children, in most cases) to a school. This is hard to observe by mobile phone technology as the user is unlikely to generate an event during this short dwell, and even if an event is generated, it will be considered part of the journey because a user is not observed there for 30 minutes or more. Furthermore, children under 16 cannot legally be in the matrix because of United Kingdom law.

_Hypothesis B is not rejected._

**C. The ‘Home based work’ work trip ends have no correlation with census workplace data.**

The home based work trip ends from the mobile phone data have been plotted against workplace data. Because the p value of this regression is below 0.05, it can be concluded that there is a correlation and that this hypothesis will have to be rejected. However, the quality of the correlation may not be desirable and will have room for improvement. The $R^2$ for the correlation between trip ends and work population is lower than the $R^2$ for the correlation between home ends and residential population. This means that the 'work' end of the trip is harder to predict than he 'home' end.

_Hypothesis C is rejected._

**D. The trip origins and destinations have no correlation with the trip origins and destinations in the National Trip End Model.**

The mobile phone trip ends have been compared against the National Trip End model on the regional, county and authority level. On all levels, a correlation was observed (p value < 0.05). However, the match of the mobile phone data with NTEM decreased as the data was tested on a spatially more detailed level. In conclusion, the hypothesis is rejected because a correlation was observed, but there is still room for improvement.

_Hypothesis D is rejected._

**E. There is no correlation between trip distribution of the mobile phone home based work matrix and the census journey-to-work matrix.**

Census journey-to-work data was used to test the trip distribution aspect. It showed that there was a correlation between trips in the mobile phone data and trips in the census journey-to-work matrix (p value < 0.05). The hypothesis is rejected. However, there is noise in the mobile phone data. Many OD pairs have zero trips in the mobile phone data, but have some trips in the Census journey-to-work matrix (see appendix vi, Figure 31). The quality of the trip distribution part of the mobile phone data can be improved by looking into these 0 trip OD pairs and finding out the cause.

_Hypothesis E is rejected._
F. Not enough trips have been classified as ‘rail’ if compared to the National Travel Survey. The proportion of rail trips is too low.

The proportion of rail trips matches the National Travel Survey in only 3 of the 9 regions. In 5 out of the 6 regions where there is a mismatch, the proportion of rail trips in the mobile phone data is too low. This is evidence that the number of rail trips is underestimated. This is expected, because some rail clusters are ambiguous if many road events are generated at the same time, which tends to happen at a LAC boundary with both a road and a railway close to each other. This is known by the mobile phone providers and extra steps to determine mode are currently being developed.

Hypothesis F is not rejected.

5.6.2 Conclusions on the data

The findings in this analysis will be summed up in this section. A discussion will follow in 5.6.3 based on these findings. This relates to the secondary research question, formulated in chapter 1.

“Is the tested trip matrix derived from mobile phone data valid and applicable for use, and if not, which aspects will need to be considered for improvement?”

Hypothesis A was rejected, meaning that home locations of users were inferred correctly. This is an important part of the output as this determines the expansion factors for the mobile phone data. Hypothesis B was accepted, implying a problem with journey purpose classification. Hypothesis C being rejected means that work trips often ended in zones with a higher amount of workplaces, as one would expect. However, the match was not as good as the home end was against residential population. This implies that it is harder to correctly infer the work end than it is to infer the home end. Trip generation (hypothesis D) and trip distribution (hypothesis E) are validated using the National Trip End Model and the census journey-to-work matrix, although more noise occurs at spatially more detailed levels. Hypothesis F is accepted. Rail trips have been misclassified as road trips. There are still improvements to be made on the classification of the rail mode.

The mobile phone technology and this specific methodology can correctly predict trip patterns, but it is hard to correctly infer journey purpose and mode. The quality of the mobile phone data at this stage is insufficient to make a proper distinction between ‘home based work’ and ‘home based other’ trips. A lot of ‘home based other’ trips are probably not observed because the dwells are generally shorter at the ‘other’ trip end. This was already found in literature and discussed in 2.2. Improved methodologies will need to deal with journey purpose and mode more extensively. This can again be tested using the steps defined in the framework in 4.4, and improvements on the journey purpose and mode detection can be tracked. This will also allow the researcher to see if R^2 values go up if compared to the regression models plotted for this analysis. This would imply that the quality of the new matrices is improved over the quality of the old matrices.

The PM peak causes more errors than the AM peak. This could be caused by the higher amount of non-commuting trips in the PM peak. Non-commuting trips tend to be shorter, with a shorter dwell. Some zones provide a significantly worse match than other zones (see appendix iv). This could be due to the sample of valid users being insufficient to achieve a matrix of good quality.

5.6.3 Future work on matrix generation methodology

The validation framework presented in this thesis already provided some insight on the value and shortcomings of mobile phone matrices today. To improve future mobile phone matrices, methodology and algorithms will need to be changed or added to the existing process.
One of the biggest problems so far has been the short trips and short stays. This can be resolved by changing the minimum time between two events for a location to be classified as a ‘dwell’. This will cause more dwells to occur and more trips to occur in a chain between origin and destination. Caution is advised when lowering the dwell threshold, as too low a threshold can provide dwells when there actually are none, for example during traffic delays, incidents or simply slow drivers. New methodology, subject to the same set of validation steps, can be tested for improvements in this area.

Inner London calls for a separate methodology dealing with underground trips. To properly detect users entering and leaving an underground station, additional equipment such as femtocells, cells with a very small cell area, can be installed. There is also potential for data fusion with other sources. Oyster cards are often used to enter and leave the London Underground system, providing an additional set of data that can be used to build matrices. GPS data can perhaps be fused with mobile phone data help find the shorter trips. Whatever the solution, this issue will need to be accepted and considered as a shortcoming in mobile phone technology.

The ‘home based work’ classification is of insufficient quality to be used for modelling. A modeller will need to look for alternative ways to split trips between ‘home based work’ and ‘home based other’. Shift workers or users working irregular hours will also have to be taken into account. Another issue is that education trips have similar time patterns as work trips. Students may be in universities during similar hours as an office worker spends in the office. Without additional methodology, this will cause education based trips to be classified as work based trips. In conclusion, finding an alternative way to find ‘work’ locations of users is required.

Issues concerning spatial granularity and lower sample discussed in this chapter are not as easily resolved. One clear solution to achieve higher spatial granularity is the inclusion of 4G users when the market penetration rate for 4G becomes higher. This will improve both spatial and temporal granularity, as 4G cells are smaller than 2G and 3G cells and events generated in the 4G network happen on a more continuous basis.

Much work has already been done and is still in progress on developing this new technology. For highway models, the short trips and trips with a short stay are not as relevant, as they do not require that level of detail in urban areas. Up to this stage, regional models are being updated by these matrices and the matrix will probably be of sufficient quality, provided the journey purpose split used is only ‘home based’ and ‘non-home based’. For modelling towns and cities, much additional work is still to be done and higher spatial and temporal granularity needs to be achieved.
6 Conclusions, discussion and further research

A framework for validation is developed in this thesis and applied on a mobile phone based matrix. This provided some useful insight, providing an answer to the research question formulated in chapter 1.

“How validation steps have to be undertaken using which sources of data, to be used to test the quality of a trip matrix derived from mobile phone data?”

These steps have been presented below. The framework follows a simple 6 step process.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Test mode share and journey purpose split against NTS.</td>
</tr>
<tr>
<td>2.</td>
<td>Compare against NTS daily profile.</td>
</tr>
<tr>
<td>3.</td>
<td>Test trip end origin and destinations against the NTEM model.</td>
</tr>
<tr>
<td>4.</td>
<td>Test home based trip origins against census population data.</td>
</tr>
<tr>
<td>5.</td>
<td>Test home based work trip destinations against census workplace data.</td>
</tr>
<tr>
<td>6.</td>
<td>Compare Home based Work trips with the census Journey to Work matrix.</td>
</tr>
</tbody>
</table>

These 6 tests provide insight in a multitude of characteristics of a matrix. It includes mode split, journey purpose split, trip generation, trip distribution and the daily profile. It is possible to carry out this validation process with just United Kingdom census data, the National Travel Survey, the National Trip End Model. Most developed countries would have surveys and models equivalent to the United Kingdom data, making this framework applicable to mobile phone matrices in most developed countries.

The step 1 compares percentages of mode split and journey purpose split on a very high level. This allows the analyst to quickly see if there are major problems with the way mode and journey purpose are determined. A one sample z-test for proportions was proposed as a statistical test to see if the mode proportions and journey purpose proportions differ between the mobile phone data matrix and a survey. It is important that this aspect can now be properly tested, as journey purpose and mode classification have been identified as difficult, with only few successful attempts in literature.

Step 2 is another high level analysis to test the temporal dimension of the matrix. The profile of trips throughout the day can be plotted from the matrix, considered the analyst possessed hourly matrices. A profile from a survey or model can be plotted in the same chart. An indicator is the time of the peak periods, and the mobile phone based matrix can be checked to see if the peak period is in the right place. After step 2, the analyst will already know if trips are assigned the right time of day, mode and purpose.

Regression analysis was selected as the method for carrying out steps 3 through 6. A correlation between the mobile phone data and the external data was assumed for the mobile trip ends, allowing for a regression model to be plotted. The benefit of applying regression is that the units do not have to be
defined in the exact same way. For example, commute trips in the National Trip End Model are not the exact same, but would have a correlation. This is even more true for comparing residential population and the home end of a trip in step 4. In this case, two different units have been compared. It is recommended to apply a logarithmic transformation on the data before undertaking the regression analysis. This allows for the regression charts to be more easily interpreted, especially when comparing smaller zones to the regression line.

A drawback of regression analysis is that it can be hard to interpret. In the validation of the SERTM mobile phone based matrix in chapter 5, a correlation was observed in every single case because the p value was always below 0.05. However, this does not imply that there is a strong correlation, just that there is one. The $R^2$ value is an indicator of quality. It is hard to conclude that data quality is ‘good’ or ‘bad’ based on this $R^2$, as sometimes a high $R^2$ is expected (if comparing trips against trips), while sometimes $R^2$ can be lower because there are other explanatory variables in play (if comparing trips against residential population). A solution would be to do a multivariate regression, taking into account these other explanatory variables. In step 4, for example, income and car ownership together with residential population would be explanatory variables for the number of trips. This would, however, require more data and would introduce possible problems if the data is not ground truth.

6.1 Discussion

A potential drawback of regression that was found during the analysis in chapter 5 is that very high numbers in the data can skew the axes in such a way that the residuals of the lower values become lower, resulting in a high $R^2$ just because a large data point is included. Removing these zones revealed that the $R^2$ will in fact be much lower without the existence of these data points. This means that it is possible to ‘cheat’ the $R^2$ value to a high value by including a very large, aggregated zone. To get any insight in the scatter, or ‘noise’ in the data, the axes need to be scaled down again. A logarithmic transformation is recommended in these scenarios. Applying the natural logarithm on all the values in the data will produce a more insightful regression chart. Using this method, it was revealed that the ‘work’ end of a home based work trip is of lower quality than the ‘home’ end of the trip, even though the $R^2$ was higher for the work end before applying the logarithmic transformation.

Besides regression, other tests in the framework include comparing percentage to percentage, or proportion to proportion. To arrive at a statistically sound comparison, the one sample z-test for proportions was proposed. This allows for the analyst to objectively reject a hypothesis that there is a difference between mobile phone data journey purpose and mode proportions and survey or model journey purpose and mode proportions.

The daily trip profile, showing the amount of trips starting throughout the hours of the day, is often reported to help gain insight in travel patterns throughout the day. It is not common that trip profiles from two different data sources are plotted against each other, let alone compared to each other. As the inferred start time of a trip is part of the methodology, based on an assumption and a potential issue, this needs to be tested. It has proven difficult to define an objective measure to this test, rather than just doing a visual interpretation of the trend. The criteria defined in this thesis only deal with the peak periods in the morning (AM) and afternoon (PM), as well as the local minimum of the graph in between the two peaks. This does not compare the height and sharpness of the peaks and overall trend line.

6.2 Contribution to existing literature

In this thesis, a contribution has been made to existing literature by providing a framework where researchers and transport modellers can test a matrix against. This answers a problem raised by Demissie (2014), where no proper validation framework has previously been defined. The framework proposed in this thesis covers trip generation, trip distribution, mode choice and journey purpose, providing the
researcher with insight on this with relatively little effort. No modelling is required to do any of these tests. The series of tests defined in this framework can be repeated should methodology be improved, and it is easy to see if the methodology improves any of the mentioned aspects of a mobile phone based matrix.

6.3 Further research
If, by applying this framework, the analyst gains confidence in the mobile phone data, the next step would be to apply the matrix to a model. This would allow for different statistics to be validated, such as a Trip Length Distribution, and it would be possible to test the assignment step against traffic counts. It is recommended to start by applying the framework presented in this thesis, as it will quickly allow the user of a matrix to gain confidence in (parts of) the mobile phone based matrix or identify problems on a high level. In the United Kingdom, Origin-Destination matrices are often derived from road side interviews (RSIs). These interviews are relatively expensive to conduct and contain only a small sample size. Matrices based on mobile phone data can be expected to be better than the RSIs because of the much larger sample. A possible test is to run model assignment using a mobile phone based matrix and comparing it to a fully calibrated and validated transport model. The same can be done using the initial matrix derived from the RSIs. The hypothesis is that the modelled flows from the mobile phone based matrix are closer to that of the fully calibrated and validated model than the modelled flows from the initial RSI matrix.
Bibliography


i. NTEM validation charts, all day, all purposes

Figure 7: Region origin comparison, all day.

Figure 8: Region destination comparison, all day.
Figure 9: County origin comparison, all day.

Figure 10: County destination comparison, all day.
Figure 11: Authority origin comparison, all day.

Figure 12: Authority destination comparison, all day.
ii. NTEM validation charts, AM Peak, home based work outbound

![Graph 1: Region origin comparison, AM peak.](image)

R² = 0.9526

![Graph 2: Region destination comparison, AM peak.](image)

R² = 0.9281
Figure 15: County origin comparison, AM peak

Figure 16: County destination comparison, AM peak.
Figure 17: Authority origin comparison, AM peak.

Figure 18: Authority destination comparison, AM peak.
iii. NTEM validation charts, PM Peak, home based work inbound

Figure 19: Region origin comparison, PM peak.

Figure 20: Region destination comparison, PM peak.
Figure 21: County origin comparison, PM peak.

Figure 22: County destination comparison, PM peak.
Figure 23: Authority origin comparison, PM peak.

Figure 24: Authority destination comparison, PM peak.
iv. Data quality maps

Figure 25: Authority goodness of match, origin trip ends

Figure 26: Authority goodness of match, destination trip ends
Figure 27: Authority goodness of match, origin trip ends zoomed in

Figure 28: Authority goodness of match, destination trip ends zoomed in
v. Census validation results

Figure 29: Home based work outbound origins vs. residential population

Figure 30: Home based work outbound destinations vs. workplace population
vi. Census Journey to Work matrix comparison

Figure 31: Journey to work comparison by origin-destination pair
vii. Glossary

Active event: Mobile phone ‘trace’ which is generated when a user makes a call, sends a text or connects to the internet.

BSC: Base Station Controller, a controller that supervises a number of Base Transceiver Stations.

BTS: Base Transceiver Station. A station, usually on top of a mast or tower, that connects to a mobile phone when an active event is produced.

Cell ID: A unique identifier given to each cell.

Dwell: A location where a user is observed for more than 30 minutes.

Journey: A trip between two dwell locations.

LAC: Location Area Code, the location covered by multiple cells grouped together.

MSC: Mobile Switching Centre. Performs the task of switching a phone call to the right BSC and arranges handover events.

MSOA: Middle Layer Super Output Area. An aggregation of the census output areas.

NTEM: National Trip End Model. A model used for predicting Production and Attraction of zones in mainland UK using growth factors.

NTS: National Travel Survey.

Passive event: Mobile phone ‘trace’ automatically generated after 180 minutes and when the user crosses a LAC boundary (handover event).

POI: Point of Interest, a location visited by a user classified as ‘home’, ‘work’ or ‘other’.

SERTM: South East Regional Model, a highways model of the South East of the United Kingdom.

TEMPRO: Trip End Model Program. This program is used to access the data from NTEM.