FUSING MOBILE NETWORK DATA WITH MOBILITY DATA FOR TRANSPORT FOR LONDON'S PROJECT EDMOND

Chris Wroe
Telefonica UK
Stephen Rutherford
Jacobs
Paul Hanson, Reza Tolouei
AECOM
Aliasgar Inayathusein
Transport for London

1. THE MODELLING CONTEXT

London is a dynamic and dense city which has the challenge of accommodating a population of 8.7 million people and 26.7 million daily trips made as people travel around the city. This is expected to grow further by 2041 with an estimated population of 10.5 million and 32 million trips being made by this time. The Mayor's Transport Strategy[1], currently published in draft, sets out an ambitious plan to accommodate sustainable growth and create a city for all Londoners.

In order to inform policy and appraise the impacts of infrastructure investment, Transport for London (TfL) has developed and maintains a set of strategic transport model which represent the behaviour of drivers, passengers, cyclists and pedestrians as they travel on London's transport network. They cover all the main modes of travel and help TfL and others to plan London's future transport needs and identify which transport schemes and policies should be implemented to meet the goals set in the Mayor's Transport Strategy. Further information can be found on TfL's Strategic Transport Models[2] webpage (Search *TfL Strategic Transport Models*).

[1](https://www.london.gov.uk/what-we-do/transport/our-vision-transport/draft-mayors-transport-strategy-2017

[2] https://tfl.gov.uk/corporate/publications-and-reports/strategic-transport-and-land-use-models

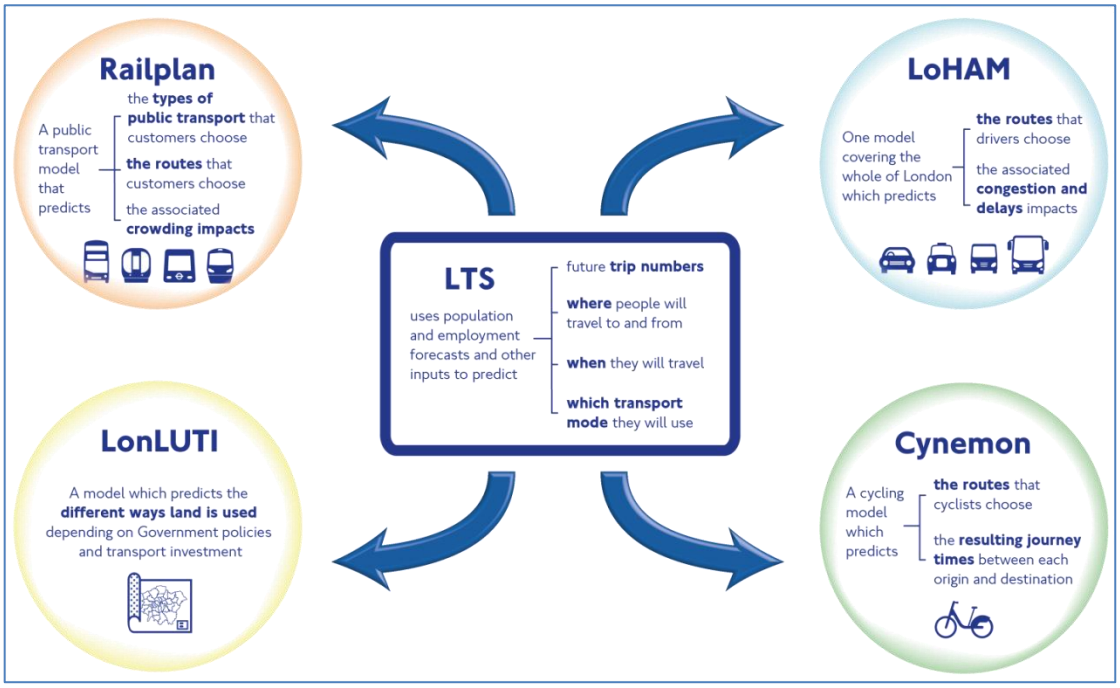

Figure 1: Overview of TfL Strategic Models

These models are based on data collected from a variety of datasets including data collected from TfL's Oyster Smartcard system and from the London Travel Demand Survey, a rolling household interview survey. However, data on mainline rail usage and on road movements is more limited, with the last series of road side interviews carried out in 2009.

To collect further data for these models, TfL has commissioned project EDMOND ('Estimating Demand from Mobile Network Data'), which is being delivered collaboratively with TfL by Jacobs, Telefonica and AECOM. The project is focused on collecting anonymous, aggregated mobile network data and fusing it with other datasets to develop trip matrices, providing the levels of mode and purpose segmentation required for the project. The project is ongoing, and this paper describes some of the project team's proposed methodology for data collection and fusion.

The use of mobile network data to provide origin-destination information for transport models is becoming increasingly common in the UK. Data from mobile network operator Telefonica has been used in collaboration with Jacobs to create a 'Trip Information System' for Highways England, who manage the strategic road network in England (Duduta et al, 2016). Telefonica data has also been used to provide origin-destination data for various local authority models, including for the Leicester and Leicestershire Integrated Transport Model (LLITM) delivered by AECOM (Tolouei et al, 2015).

2. MOBILE DATA COLLECTION

Mobiles phones generate “events” as they communicate with the national cell network. Event data is collected from Telefonica’s infrastructure through probes that are used to monitor the network performance and customer experience. The probes collect event data in real time, on an ongoing basis, across the whole of the UK. The output is anonymised to ensure that individuals cannot be identified, and then stored in a data ‘warehouse’ for the purpose of analysis of travel patterns. All of the outputs used for the project are also aggregated, to prevent the disclosure of individual trip patterns.

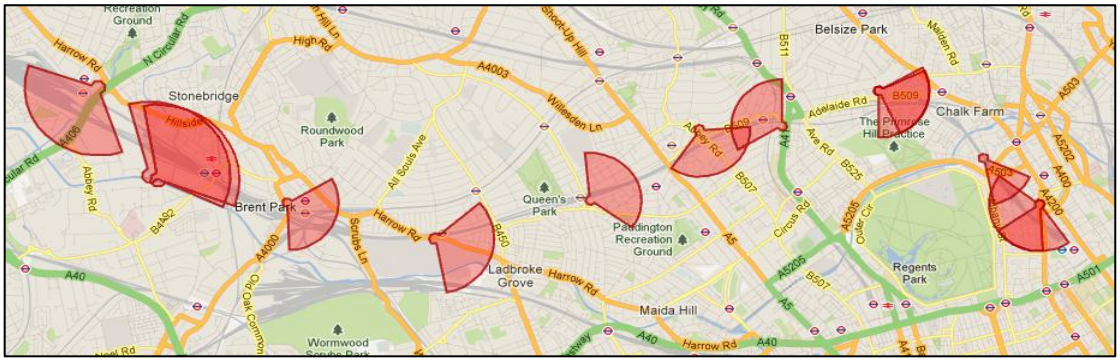

Figure 2: As handsets move around the UK they leave a trail of anonymous events

The source of the dataset is from Telefonica UK and the feed includes both contract and pay as you go (PAYG) customers, including virtual mobile network operators that use the Telefonica infrastructure (Tesco Mobile and Giffgaff).

The stream covers the whole of the UK and contains active as well as passive events. Active events occur when a handset makes or receives a phone call, or sends or receives a text message. Passive events occur at set regular intervals, or when a handset moves from one group of cells to another. Events from the 2G, 3G and 4G networks were collected and both smartphones and non-smartphone devices contribute to the dataset. About six billion events per day were collected, relating to about 26 million devices. This was later filtered down to a ‘core sample’ of approximately 15 million devices, after removing handsets that were not regularly used during the study period.

For the project, events were processed relating to September, October and November 2016 – three months chosen because of their timeliness at project inception and general neutrality.

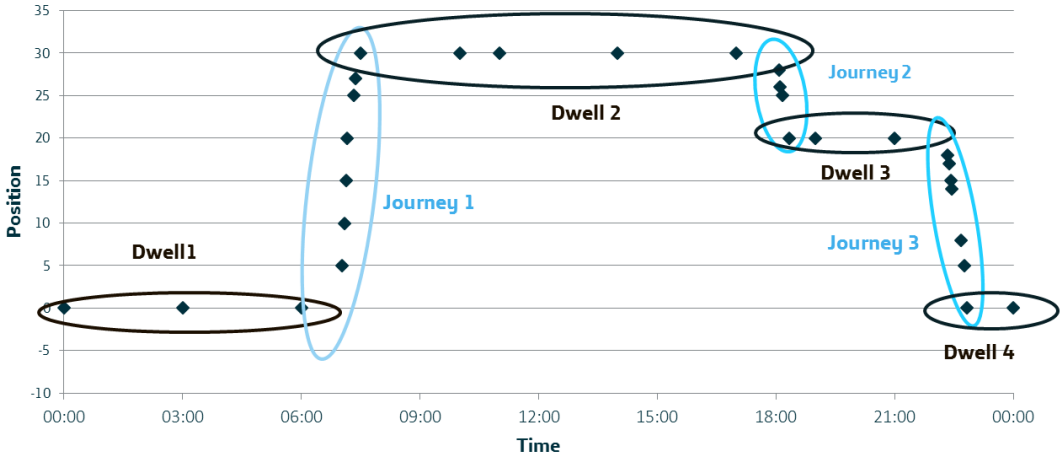

Figure 3: A pre-existing process was used to convert the mobile network events into dwells and journeys

After the events were collected they were converted using a pre-existing algorithm into 'dwells', when handsets were deemed to be stationary, and journeys', when handsets were moving. Approximately 30m journeys per day were observed by the core sample across the study period, as illustrated below:

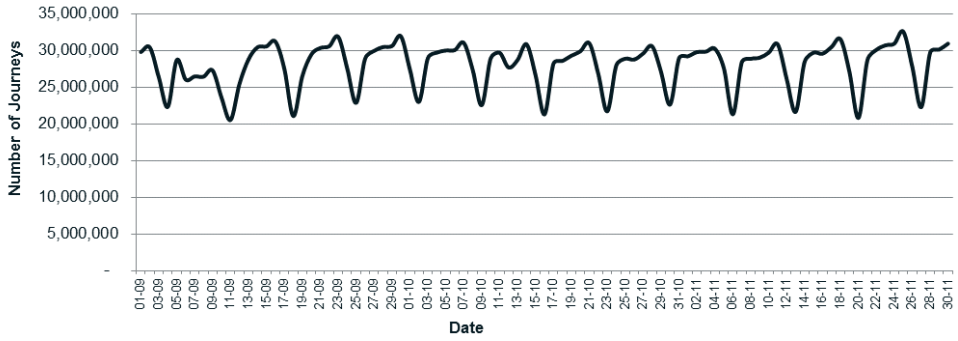

Figure 4: Number of journeys per day in the processed mobile network data

After the journeys descried above were identified, they were filtered to create a dataset of journeys 'that did, or could have' gone through London. This was done through the use of a routing engine which indicated the possible routes for each journey. Including journeys that 'could have' gone through the study area ensures that the effect of modal shift or infrastructure improvements on travel in London can be properly understood.

Since the mobile network data was based on trips by a sub-set of the UK population, the trips were expanded based on demographic analysis of the core sample. The home location of each mobile phone user was identified through analysis of their regular overnight location, and the number of residents counted

in each model zone was compared to recent population estimates for those zones provided by TfL. The ratio of these results was used as an 'expansion factor' and applied to all trips by those handsets.

Statistics provided by OFCOM, the UK's communications regulator, as well as analysis of London Travel Demand Survey (LTDS) has shown that variation in trip patterns is correlated with variation in mobile phone ownership; that is, individual groups who tend to have more access to mobile phones (irrespective of the operator), are also more likely to drive and to travel longer distances. Two key factors identified were age and level of income. Adjustment factors are currently being developed to account for these variations during the expansion process.

A considerable number of trips in London are made by international visitors. Some of these 'inbound roamers' were included in the mobile network data, and these were used to generate journeys and dwells by overseas visitors. The port or airport of entry and exit for these users was identified, and an expansion factor applied based on comparisons of the number of international visitors seen arriving at those ports with information provided by the UK Civil Aviation Authority (CAA) and sourced from the International Passenger Survey (IPS).

Trips observed in the mobile network data were analysed to identify their trip purpose. The home and work location of handsets was analysed based on regular overnight and weekday dwells, and trips between these points have been tagged as commutes. This analysis also enables a distinction between non-home based and home-based trips. Analysis is currently underway to identify education trips from commutes, based on an analysis of the seasonality of these trips.

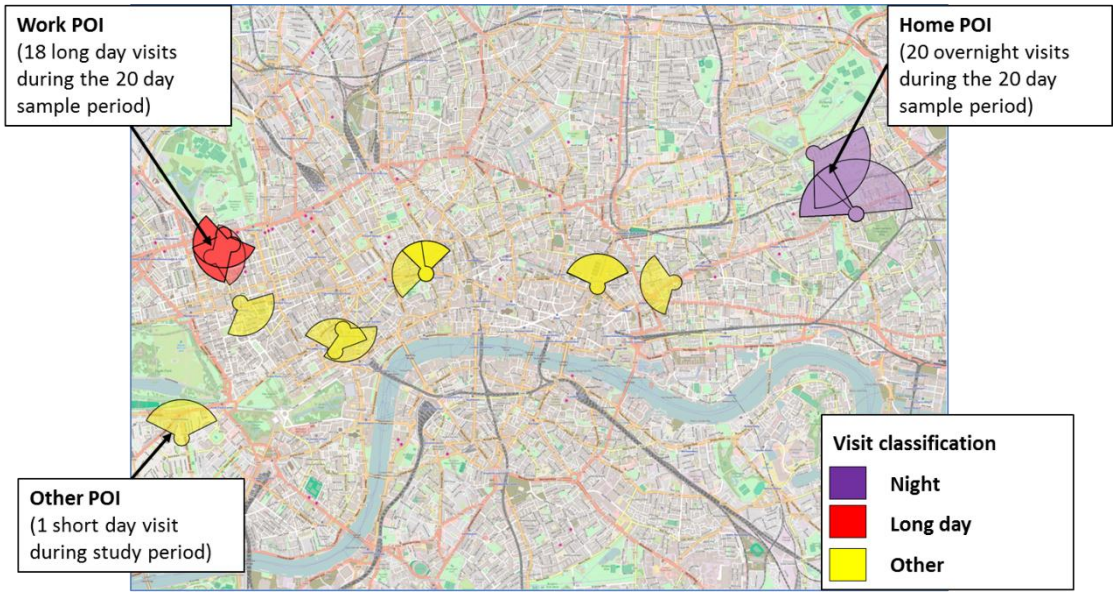

Figure 5: Dwells in the mobile network data can be analysed to identify the home and work location of users

Although the primary process of mode disambiguation for the project will be carried out based on data fusion, some initial analysis has been carried out for trips observed in the mobile network data to provide some indicators of mode. This included:

- Analysis of the spatial events created during the observed journeys to see if they match a particular rail or road route
- Analysis of the temporal pattern of events generated during the observed journeys to establish the maximum, and average speed
- Analysis of clusters of events seen by multiple handsets at the same time and place – these clusters are generally indicative of rail travel, as multiple handsets travelling in the same train create clusters of events.

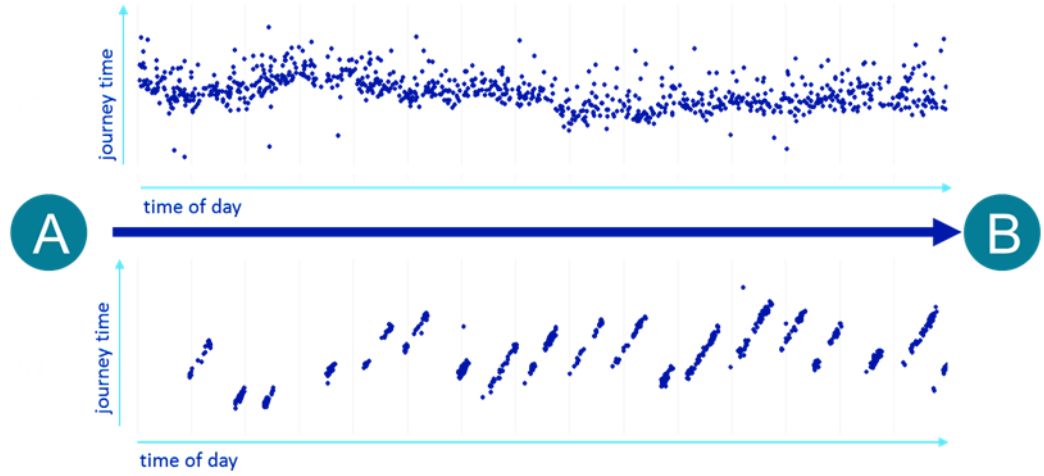

Figure 6: This image compares the pattern of events for road users (top) with rail users (bottom) – clusters of events are indicative of rail travel

3. SECONDARY DATA COLLECTION

Transport for London already collects a wide variety of data across London, and much of this will be used in the project for the purpose of disambiguation of mode in the mobile network data. Some of the key datasets proposed to be used for this purpose are described below.

London Travel Demand Survey (LTDS)

This is a household survey that has been run since 2005. Surveys are filled in by all members of a chosen household aged five or over. On a designated day, the respondents will list all of their trips including trip purpose, modes used, start and end time and origin, interchanges and destination. Approximately 0.5% of London's population are sampled per year. The trips sampled in the LTDS will form part of a 'training dataset' which will be used to identify the mode of travel of trips seen in the mobile network data.

National Travel Survey (NTS)

Similar to LTDS, this is a household survey that is run on a rolling basis across the UK. Since LTDS only includes London residents, NTS results will be used in the training dataset to represent the behaviour of visitors to London.

Cycle Hire Docking Station Data

Santander Cycle hire is London's self-service, bike-sharing scheme for short journeys. Bikes are located at docking stations throughout central and inner London and each trip begins and ends at a docking station. The data includes

100% of Cycle hire users with around 2.3 million hires in Sep-Nov 2016. This cycle hire data will be used to enhance the training dataset, replacing the small number of cycle hire trips recorded in LTDS with a much larger number of observed records, and increasing the resolution of the data while controlling for overall mode share.

Taxi and Private Hire Vehicle Supply and Demand Survey

This survey is focussed on private hire drivers who work in the minicab or chauffeur sectors as well as taxi drivers. Taxi and PHV drivers were recruited on the telephone using the respective driver databases, held by Transport for London - Taxi & Private Hire, as the sample frame. Taxi and PHV drivers completed a diary for two days, recording details of all journeys undertaken during their driving shifts. This dataset currently contains approximately 10,000 taxi and PHV journeys, which will be used to enhance the training dataset used in the project, increasing the resolution of observed taxi and PHV journeys.

HGV and LGV Usage Surveys

The Continuing Survey of Road Goods Transport (CSRGT) provides information on the UK activity of GB-registered heavy goods vehicles (HGVs). The survey is based on a weekly sample of around 350 vehicles which are selected from the DfT's vehicle database. The operator of the vehicle is asked to record information about their activity for one week. Despite a relatively small sample size, the CSRGT survey has key advantages of being detailed, comprehensive, and consistent through the years. It will be used to identify the typical behaviour of HGVs, which will then be searched for in the mobile network data.

The Department for Transport has carried out two van-related surveys in the past: Survey of Company Owned Vans (2003-2005), and Survey of Privately Owned Vans (2002-2003). Despite being old, these two surveys provide valuable insights into travel patterns of various van users. Similarly, these are being analysed to identify the typical behaviour of HGVs

Trafficmaster OD Data

The Trafficmaster database holds anonymous data collected from in-vehicle GPS tracking devices. These data can be used to derive Origin-Destination matrices, as well as other statistics (e.g. average speed, journey times, journey time variability, and journey time reliability). Each record in the OD data is associated to a single trip, which is registered as the point from when the vehicle

ignition is turned on to the point the ignition of the vehicle is turned off. The database is created and owned by Department for Transport. For LGVs, data is believed to include a sample of around 75,000 LGVs, which constitutes approximately 2% of the national fleet. This will be used to identify typical journey patterns for LGVs, which will then be searched for in the mobile network data.

Oyster Smartcard and Contactless Payment Data

TfL's ticketing data allows them to build a detailed picture of travel patterns across the rail, tube and bus networks. The use of Oyster and contactless payments through bank cards, Apple Pay and Android Pay, has given TfL tube and rail station entry and exit data as customers (using these payment mechanisms) must touch in and out for their journeys. Bus journeys, on the other hand, are more problematic to monitor, as customers are only required to tap in when they get on, but not when they exit. However, through work with the Massachusetts Institute of Technology, TfL have developed a process to infer when customers are leaving a bus using a Big Data tool which looks at origin and bus interchange information – which they call ODX. It combines bus location and ticketing data to try and match up origin and destination pairs to create a multi-modal travel dataset.

The ODX data contains approximately 46m weekly bus journeys and 25m weekly rail journeys. Compared to ad-hoc surveys the sample size is very large and temporal, and geo-spatial resolution is very detailed. The data from the EDMOND study period will be incorporated into the training dataset, greatly enhancing the resolution of public transport trips.

4. MODE ALLOCATION MODEL

Data fusion

As described above, various datasets will be combined into a training dataset to inform the calibration of a mode allocation model. Each dataset will have different characteristics, sample rates and data points so as they are combined, the overall mode share from LTDS and NTS will need to be retained and some data points not present in all data (e.g. age and gender) will need to be infilled from the LTDS and NTS data.

Model formulation

Once the training dataset has been built, a mode choice model will be calibrated using the training dataset (see, for example, Koppelman and Bhat, 2006 for theoretical background on mode choice modelling). The process is complicated by the fact that some features in the mobile network data, such as the extent to which trips match a certain road or public transport route, are not represented in the training dataset. To accommodate this, each trip record between a given origin destination pair in the training dataset will be allocated a distribution of route match scores based on several observed route match scores in the mobile network dataset for trips made between the same origin destination pair. Using a distribution and calculating individual probabilities of mode for small intervals of the distribution (i.e. integration) ensures that, unlike using a simple average for example, consistent and unbiased parameters can be estimated during the calibration process, and that they can then be directly applied to individual records within the MND (see Tolouei, et al., 2013 for more theoretical details of this modelling approach).

The mode choice model will be developed progressively using a step-wise approach, incorporating different variables found in both the mobile network data and the training dataset to improve the model. It is expected that different trip purposes may exhibit different mode preferences, so separate models may be calibrated to reflect this.

The variables that will be investigated for inclusion in the model are listed below:

Default attributes	Description
Mode	Transportation mode of the trip
Origin	Origin of the trip
Destination	Destination of the trip
Timestamp	Reported time of the start of the trip
Distance	Reported distance of the trip
Purpose	Reported purpose of the trip
Age and Gender	Age and gender of the responder
Affluence	Household income
Working status	Employed, student, retired
Residency	Home location of user
User trip rate	Trips made by the user that day
User distance	Distance travelled by the user that day
Speed	Average speed of the journey
Journey time	Approximate journey time.

Table 1: Attributes used in the mode allocation process

Model application

Once the parameters are estimated for all mode alternatives, the mode allocation model, in the form of a multinomial logit model, will be directly applied to the mobile network data to estimate mode probabilities. Rather than allocating trips to the highest probability mode, the individual probabilities are proposed to be retained and used in the resulting origin-destination matrices, thus retaining the volumes for 'minority modes' such as cycling. After applying the model, the results will be validated against datasets not used in the calibration process. It is expected that this process will be iterative, with different model forms and parameters used to improve the results.

5. MATRIX ADJUSTMENTS AND VALIDATION

The next step in the project will be the application of the mode allocation process described above. Adjustments will be made to the data, using data fusion techniques, to reflect the limitations of the mobile network data, including the synthesis of short trips, further segmentation of trip purpose, and spatial disaggregation to meet the requirements of TfL's models. A validation process will then be carried out based on a number of independent datasets, prior to the trip matrices being delivered to Transport for London for use in their model development.

A number of other applications of the data are currently being explored, such as the use of the data in real time operational contexts, and in policy making areas such as air quality impact analysis. Many other cities in the UK are also currently exploring opportunities to use mobile network data, together with other

emerging datasets, to further improve the quality of their transport modelling data.

BIBLIOGRAPHY

Duduta, N., Corby, N., Rutherford, S., (2016) “Development of a Trip Information System for Highways England Using Telefonica / O2 Mobile Phone Cell ID Data”, proceedings of the European Transport Conference, Barcelona.

Tolouei, R., Álvarez, P., Duduta, N., (2015) “Developing and Verifying Origin-Destination Matrices using Mobile Phone Data: The LLITM Case”, proceedings of the European Transport Conference, Frankfurt, 2015.

Tolouei, R., Maher, M., Titheridge, H., (2013), “Vehicle mass and injury risk in two-car crashes: a novel methodology”, Accident Analysis and Prevention, 50, 155-166.

Tolouei, R., Maher, M., Titheridge, H., (2013) “Vehicle mass and injury risk in two-car crashes: a novel methodology”, Accident Analysis and Prevention, 50, 155-166.