Utilising Mobile Phone Billing Records for Travel Mode Discovery

John Doyle†, Peter Hung, Damian Kelly, Seán McLoone and Ronan Farrell

Strategic Cluster for Advanced Geotechnologies,
Callan Institute,
National University of Ireland Maynooth,
IRELAND

E-mail: †jcdoyle@eeng.nuim.ie

Abstract — **A novel methodology to infer transportation mode taken by mobile device users between regions of interest is introduced. It relies on analysing anonymised billing data, namely call detail records, supplied by mobile network operators as the primary source of user-created data. Coupled with the spatial coverage and distribution of mobile network cells and geographical route map information of major transportation modes, assumed to be partially non-overlapping, user travel paths can be predicted. Journey specific trajectories are constructed and analysed using the concept of virtual cell path for each qualified pre-processed list of activities from each unique user. After classification, kernel density paths for each route were generated both for illustration and validation purposes. Differentiation between rail and road users travelling between Dublin and Cork in the Republic of Ireland is shown as an example application case study.**

Keywords — **Call detail records, Transportation mode inference, Kernel density estimation, Estimated travel path.**

I Introduction

Transportation surveys are one of the most important instruments for gathering sample data used by government agencies and transportation scientists for strategic infrastructural planning and service provision. Such surveys, while informative, are both time consuming and expensive to undertake.

The mobility patterns of mobile phones are intrinsically linked to human travel behaviour. It can be used to reveal some of the space-time behaviour patterns relating to human mobility [1, 2], social structure [3] and land use [4]. The data individual working groups have at their disposal strongly influences the type of possible analysis, and information that can be extracted.

In this context, mobile phone networks may provide sources that convey transportation survey parameters. Rose [5] critics the use of mobile phones as traffic probes, noting several issues and opportunities in the extraction of transport related information. Caceres *et al.* [6] reviews traffic data estimations using a variety of metrics extracted from mobile phone networks. The development of origin-destination matrices traffic flow estimates, user speed, travel time calculations and traffic volumes is also discussed.

However, due to spatial and temporal sampling issues, it has proven difficult to segregate mobility patterns observed through the mobile telephony data to specific transportation modes. For example, Wang *et al.* [7] examined transportation mode inference from mobile billing records but their technique fails to account for transportation modes that have similar travel times. The contribution of this paper is to provide the means to infer transportation modes or major routes taken by mobile telephone users between regions of interest via their anonymised billing records. By focusing on the likelihood of a user travelling along a spatial transportation route, given their recorded travel path through the mobile phone network, users with similar travel times but different travel modes can readily be segregated.

The data set used in this publication is a continuous one week sample of call detail records (CDR) provided by Meteor, a mobile network operator in the Republic of Ireland, collected between the 26th of November and the 2nd of December 2009. This data contains in excess of 300 million entries and provides temporal and cell tower connecting information relating to over a million users across the whole of the Republic of Ireland. The records are comprised of billable interactions between a mobile phone network and their customers. This consist of anonymised information relating to the SIM cards inserted into mobile devices that are in connection with the network, consisting of those who initiated activities and their intended recipients. CDR also contain the nature of the communication (voice, SMS, data, etc.), duration of the activity, starting time of the activity and cell identification numbers of both the sender and receiver where available.

This paper is organised as follows. Section II describes the process of constructing CDR journey trajectories while section III summarises transportation mode inference using virtual cell paths. Section IV discusses the procedure for estimating user travel paths based on the spatial distribution of their activities. Section V shows the results of the transportation mode inference processes while the conclusions are discussed in section VI.

II CDR Journey Trajectories

To infer transportation mode from CDR, it is first required to extract mobility information for each unique user in the form of journey trajectories. To transform a unique user's list of CDR activities into journey trajectories, spatial information must be first derived from the serving cells associated to the logged events. Common cell parameters include network type, site location, allocated transmitter reference and transmitter azimuth angle. This information can collectively be converted to cell site coordinates or coverage areas of the serving cells via Voronoi tessellation [8] of the site locations. A section of the approximate coverage map is depicted in Fig. 1. The accuracy of the tessellation is affected by physical layer parameters such as the channel characteristics, transmitter frequency, tilt, height, transmission power etc.

User event trajectories are formed by associating the centroid location of the connecting cell Voronoi polygons to the temporally sequenced events. Location information of trajectories generated in this way will display spatial heteroskedasticity, as the variance in estimation accuracy will be influenced by the physical topology of the mobile network (i.e. the size and density of the cells). For example, it is less likely for a user 20 km away from a cell tower to be associated to that cell in a city compared to one located in a rural area.

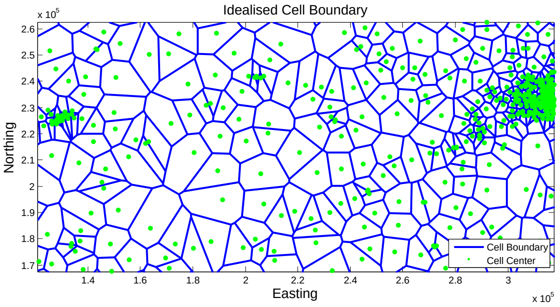


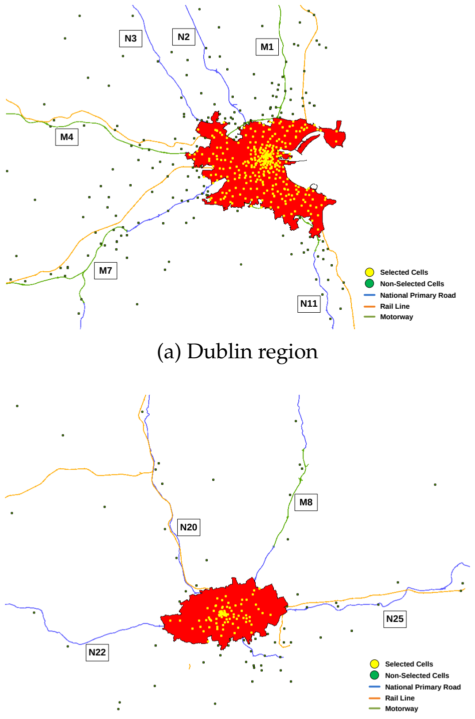
Fig. 1: Sample of idealised Voronoi tessellation used to calculate cell network coverage map.

To develop journey specific trajectories (J_i), regions of interest R_A and R_B are defined. Cells located within each region boundary are assigned to belong to that region. This provides a means of capturing each region in terms of the cell network coverage as well as restricting CDR to samples related directly to the regions of interest. The next step is to find the users who were deemed to travel directly between the cells in R_A and R_B. For the propose of this investigation journeys between the Republic of Ireland's two largest cities, Dublin and Cork, were chosen. Both regions of interest are depicted in Fig. 2a and Fig. 2b, respectively. Hereafter, only trajectories between R_A and R_B are considered.

Ensuring anonymity of individual users whose trajectories are being tracked is important. Researchers have shown that the combination of anonymised user data with external information sources can reveal the identity of the previously unidentifiable users [9]. To address this issue we have chosen to remove anonymised user IDs hereafter. While this decimation may not guarantee the anonymised user associated with a particular journey may never be identified. The omission of anonymised user ID breaks the linkage between blocks of travel information at different time intervals, meaning no individual may be tracked for prolonged periods of time.

III Transportation Mode Inference based on VCP

The travel mode discovery technique focuses on identifying journeys taken by the major modes of transport, namely rail-line and primary road (the M7-M8 motorway) between Dublin and Cork. The first step in the process is to select those users whose journey travel times are related to the achievable travel time along the particular primary routes. The non-flight travel time between Dublin and Cork is approximately 3 hours, thus



(a) Dublin region

(b) Cork region

Fig. 2: Regions of interest (a) cells assigned Dublin region; (b) cells assigned Cork region.

by extracting those users who had journey travel times of less than 3.5 hours, we can focus on users who travel directly between the two regions on primary routes. This reduces the number of individual journeys from 7500 to 2537.

Proximity measurements may fail to account for the topology of the network given that each J_i is artificially fixed to the connecting cell towers. This is due to the fact that cell towers are unlikely to be aligned in parallel routes to the rail and roads. In other words, 'zig-zag' shaped J_i would normally be expected, contributing noise during subsequent analyses.

Instead, we measure similarity based on the proportion of locating events that occur at cells that are deemed to represent a route of interest (T_i). We call such a collection of cells a virtual cell path (VCP), defined as a representation of the path through a mobile telephony network along which a user may travel while on T_i. Cells in the VCP are selected if their area of coverage overlaps with T_i. Additional cells associated with T_i as manually identified from training data (c.f. 'zig-zag' J_i) are collected to improve the spatial coverage of the resultant VCP. Training data consists of randomly select J_i whose transportation mode is manually identified. Once completed, each VCP consists of a list of cells whose spatial coverage either coincides with part of T_i or serviced as a connecting cell to a user while they travelled along T_i.

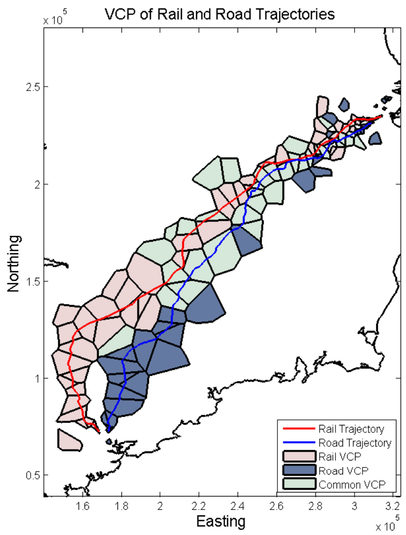


Fig. 3: Virtual cell path after training of rail and road (M7-M8) route corridors between Dublin and Cork.

Inference of the transportation mode taken by the J_i based on VCP is accomplished by labelling J_i as a particular T_i if the probability $P(T_i|J_i)$ is deemed sufficient in comparison to the probability $P(T_j|J_i)$, where T_j are other plausible transportation routes. Equation 1 shows the simplified transport inference decision between T_{rail} and T_{road} trajectories.

$$Decision = \begin{cases} Road & \text{if } [P(road) - P(rail)] > \epsilon \\ Rail & \text{if } [P(rail) - P(road)] > \epsilon \\ Unknown & \text{if } |P(rail) - P(road)| \leq \epsilon \end{cases} \tag{1}$$

where $P(road)$ and $P(rail)$ are the likelihood of being on the trajectory T_{road} and T_{rail}, respectively.

However the variances in VCP measures alone is not necessarily enough to classify travel path if those VCP have large regions of overlap. As depicted in Fig. 3 there are several regions of overlap between T_{rail} and T_{road} VCP in this case study. As a result some necessary conditions are required for VCP based travel path identification to be feasible. These conditions are

1. a minimum number of cell connections in the areas of spatial divergence between T_i; and
2. a minimum weight ϵ of difference in measures of similarity among all T_i.

IV Estimated Path Generation

The estimated travel path (ETP) represents the likely path travelled by a group of similar CDR trajectories migrating between regions of interest.

The ability to observe this path, with no transportation mode assumptions, enables the comparison of the CDR trajectory groups solely based on their on-route travel characteristics with transportation travel paths.

The process used to develop the ETP consists of developing a path of least cost between a chosen start and end location. Firstly, bin activity counts from users trajectories between the start and end location onto their servicing cell towers. Once completed, weight each spatial sample point in a dispersed grid with annealed distributed cell tower counts. Using the spatial sample point weights, derive a transition cost matrix for each sample point to its neighbouring points. This is effectively the cost of moving from one sample point to its neighbour. Sample points can then be selected to form the least cost path, if the cost of moving from the start location to end location while visiting those points is minimal in comparison to alternative routes, given that you may only transverse to a neighbouring cell in any one hop.

The spatial points chosen consist of centre of gravity locations from an evenly dispersed hexagonal grid over the study area. The size of each hexagon (5 km in diameter) is chosen to be much smaller than the rural area cells typically (10 km to 20 km). The reason for this choice is to balance the compromise between speed and spatial accuracy. In areas such as the city centre, where cell diameters are sometimes smaller than that of the hexagon grid, the error margin will be $max\{$diameter of hexagon, diameter of cell$\}$.

The spatial sample points weights are calculated based a kernel density smoothing of user activity counts on servicing cell towers. The non-negative spatial kernel density weight $W(x, y)$ at each grid point is calculated as in equation 2,

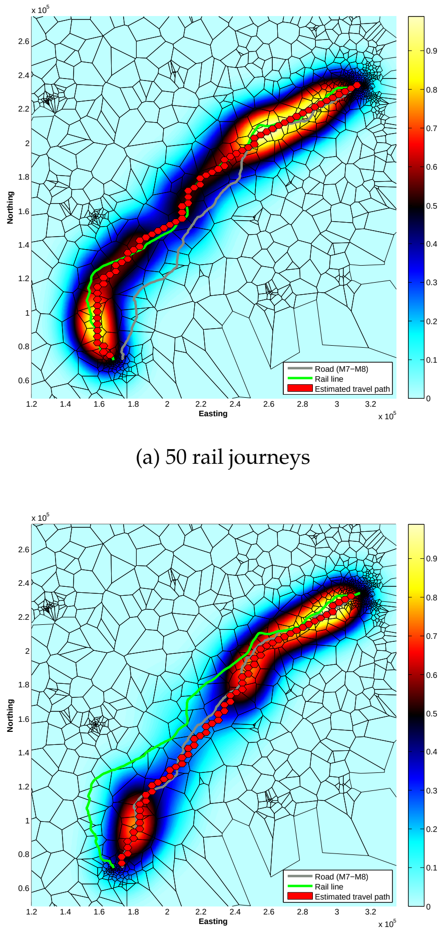
$$W(x,y) = \frac{n_c}{2\pi r^2 k} \sum_{i=1}^{M} exp^{\left(\frac{(x-X_i)^2+(y-Y_i)^2}{2r^2}\right)}, \tag{2}$$

where x, y are the spatial points Easting and Northing coordinates respectively, k is a normalisation factor such that the largest value of $W(x, y) = 1$ and r is the Gaussian kernel width, in this instance $r = 10000$. X_i and Y_i are the spatial coordinates of the cell towers associated to each activity, while n_c represent the number of CDR activities at that cell.

However, kernel density weights cannot be directly employed in a transition cost matrix between neighbouring spatial sample points. The following transition matrix M_n relates $W(x, y)$ at each sample point to the cost of moving to a neighbouring sample point.

$$M_n = \frac{W(x_n, y_n)}{W(x, y)}[(2 - W(x_n, y_n))] \,, \tag{3}$$

where $W(x_n, y_n)$ is the kernel weight of a neighbouring grid section of $W(x, y)$. The ratio $W(x_n, y_n)/W(x, y)$ scales the transition cost such that the cost of moving from a high $W(x, y)$ to low $W(x_n, y_n)$ is large, while the cost of moving from a low $W(x, y)$ to high $W(x_n, y_n)$ is small. The transition cost of moving from similar weights is also small, so to penalise the transition of low $W(x, y)$ to low $W(x_n, y_n)$, $(2 - W(x_n, y_n))$ is introduced. We then utilise Dijkstra's algorithm [10] to select the sample points which form the path of least cost between the chosen start and end locations.

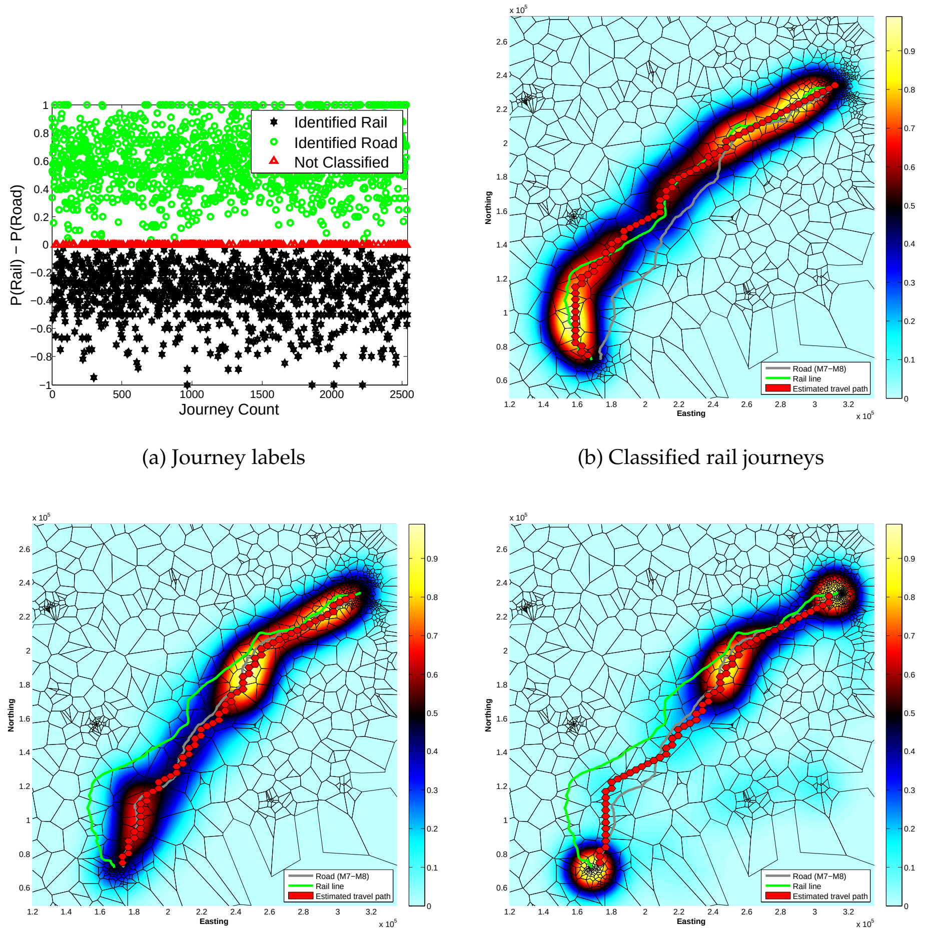


(a) 50 rail journeys

(b) 50 road journeys

Fig. 4: Estimated travel paths derived from empirical CDR trajectories between Dublin and Cork, with annealed distributed cell tower counts displayed (a) estimated travel paths associated with 50 empirical rail journeys; (b) estimated travel paths associated with 50 empirical road journeys.

The paths depicted in Figs. 4a and 4b visually demonstrate the relationship between the ETP and respective rail and road trajectories. The paths are constructed using 50 journeys manually identified as rail and road journeys. The results highlights that it may be possible to use the estimated travel



(a) Journey labels

(b) Classified rail journeys

(c) Classified road journeys

(d) Non-classified journeys

Fig. 5: Results from the VCP classification of the mode of transport used by persons travelling between Dublin and Cork (a) the classification result for each individual journey; (b) ETP with annealed distributed cell tower counts displayed of identified rail journeys; (c) ETP with annealed distributed cell tower counts displayed of identified road journeys; and (d) ETP with annealed distributed cell tower counts displayed of non-classified journeys.

paths to help verify the accuracy of the classification process.

V Results

Table 1 contains the number of road, rail and non-classifiable journeys inferred through VCP classification from our test data set. Fig. 5a depicts the weights corresponding to each T_i.

Rail	Road	Unknown	Total
1331	960	246	2537

Table 1: Classification of transportation mode

Depicted in Figs. 5b, 5c and 5d is the spatial relationship between classified J_i and T_i as described by a kernel density map for each of the classes (rail, road, unknown). As expected, the kernel density map for non-classified journeys clearly demonstrates a lack of CDR activities in spatial regions where T_{rail} and T_{road} do not overlap. The kernel density map of both rail and road journeys highlights the ability of transport mode identification using VCP on CDR data when necessary conditions are met. From Figs. 5b, 5c, it is noted that the estimated paths do not necessarily follow the true transport trajectories because of the Dijkstra algorithm attempts to follow a least cost weighted shortest path through the spatially overlain hexagonal grid.

VI Conclusions

In this paper, a novel methodology is described for inferring transportation mode taken by individuals between regions of interest via mobile phone network billing records. User trajectories are constructed from a qualified pre-processed list of activities from each unique user and the spatial parameters of the servicing cells. The development of virtual cell paths is described, for the purpose of mapping user journey trajectories to individual modes of transport between Dublin and Cork. The technique may be expanded to other routes, by developing route specific virtual cell paths.

Kernel density path generation can be used both for illustration and validation of transport mode predictions. The computation complexity associated with the generation of these paths is a compromise between grid diameter and spatial resolution.

Inferring the transportation mode of users allows a more accurate association of anonymised mobile billing data to transportation survey parameters. Thus mobile phone CDR could be proposed as an alternative to conventional regional transportation surveys with the advantages of less costly execution and a reduction in completion time.

Rose [5] suggests that the use of mobile phone sourced data for traffic monitoring may be more suited to an interurban motorway context rather than an urban setting, due to spatial and temporal sampling issues. However, as mobile device usage increases, sampling issues become less significant, and may lead to more reliable traffic monitoring for urban environments.

There are other issues to be considered that may affect the accuracy of transport mode prediction [5, 6]. These include market penetration, customer profile and network infrastructure of the mobile network operator supplying the billing information, citizens who do not carry or have their mobile devices turned off during travelling. From the technical point of view, analysis may potentially be complicated by the fact that some citizens carry more than one mobile device.

Acknowledgment

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan and by the Irish Research Council for Science, Engineering and Technology under their Embark Initiative in partnership with ESRI Ireland. The authors gratefully acknowledge this support. The authors would like to gratefully acknowledge the support of Meteor for providing the data used in this paper, in particular John Bathe and Adrian Whitwham. Ordinance Survey Ireland are also gratefully acknowledged for the provision of geographical region polygons.

References

[1] M. González, C. Hidalgo, and A. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[3] J. Onnela, J. Saramaki, J. Hyvonen, G. Szab'o, D. Lazer, K. Kaski, J. Kertesz, and A. Barabasi, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, p. 7332, 2007.

[4] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, 2007.

[5] G. Rose, "Mobile phones as traffic probes: practices, prospects and issues," *Transport Reviews*, vol. 26, no. 3, pp. 275–291, 2006.

[6] N. Caceres, J. Wideberg, and F. Benitez, "Review of traffic data estimations extracted from cellular networks," *Intelligent Transport Systems, IET*, vol. 2, no. 3, pp. 179 –192, Sept 2008.

[7] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records," pp. 318 –323, sept 2010.

[8] F. Aurenhammer, "Voronoi diagrams a survey of a fundamental geometric data structure," *ACM Comput. Surv.*, vol. 23, pp. 345–405, September 1991.

[9] M. Barbaro and T. Zeller, "A face is exposed for aol searcher no. 4417749," *New York Times*, vol. 9, p. 2008, 2006.

[10] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.