Journal of Urban Technology, 2014
Vol. 21, No. 2, 109–132, http://dx.doi.org/10.1080/10630732.2014.888904

Population Mobility Dynamics Estimated from Mobile Telephony Data

John Doyle, Peter Hung, Ronan Farrell, and Seán McLoone

ABSTRACT *In the last decade, mobile phones and mobile devices using mobile cellular telecommunication network connections have become ubiquitous. In several developed countries, the penetration of such devices has surpassed 100 percent. They facilitate communication and access to large quantities of data without the requirement of a fixed location or connection. Assuming mobile phones usually are in close proximity with the user, their cellular activities and locations are indicative of the user's activities and movements. As such, those cellular devices may be considered as a large scale distributed human activity sensing platform. This paper uses mobile operator telephony data to visualize the regional flows of people across the Republic of Ireland. In addition, the use of modified Markov chains for the ranking of significant regions of interest to mobile subscribers is investigated. Methodology is then presented which demonstrates how the ranking of significant regions of interest may be used to estimate national population, results of which are found to have strong correlation with census data.*

KEYWORDS *call detail records; imperfect trajectories; Markov chain; stationary distribution; population estimation*

Introduction

A mobile cellular telecommunication network is a geographically distributed radio network that enables communication via voice, text, or data between two or more devices (Theodore, 2001; Mishra, 2004; Korhonen, 2003; Olsson et al., 2009; Ghosh et al., 2010). It routinely collects a wealth of information related to customer interactions in the context of its normal service operations. Functions such as connecting calls, delivering text messages via SMS, and providing Internet access generate a huge amount of data which mobile network operators use for customer billing and service delivery.

The operator-based data includes, among others, network bandwidth usage measurement logs, which are typically measured in Erlang (units of person phone use per hour), handover records, locating area logs, and call detail records (CDR). Handover records are recordings of migrations of a user from one servicing cell to another while in the process of an active call. Location area logs consist of periodic location updates relating to the set of cell towers, which are prepared to serve a particular mobile device at any given time. CDR contains

Correspondence Address: John Doyle, The Callan Institute for applied ICT, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland. E-mail: jcdoyle@eeng.nuim.ie

information about all interactions between a mobile network and its customers that are required for billing purposes. It may consist of user information relating to people in connection with the network operators, the nature of the communication activity (voice, SMS, data, etc.), duration of the activity, starting time of the activity, and servicing cell identification numbers of both the sender and the receiver when available.

This research uses anonymized CDR from Meteor, a mobile network operator in the Republic of Ireland, to visualize the regional flows of people across Ireland, and also investigates the use of Markov chain fixed row vectors for the identification of significant mobile subscriber regions of interest as well as population density prediction. The Meteor network under investigation has just over one million customers, which represents approximately a quarter of the country's 4.6 million inhabitants in 2012 (CSO, 2012), and operates using both 2G and 3G telephone technologies. The CDR collected from the operator's core network includes records related to voice calls, short message service (SMS), and data transfer. Cell tower information, which contains geo-spatial coordinates in the Irish Grid Coordinate Reference System (The Irish Grid, 2011), network type, and transmitter azimuth, was also provided.

The voice calls and SMS records are split into originating and terminating files, while data logs contain information on mobile Internet sessions. The voice originating and terminating logs contain information on the time of each call, the caller and the called subscriber's anonymized unique identifier, the duration of each call, and the servicing cell towers of both caller and called subscribers at the start and end of each call when available. Similar information related to SMS activity is contained in the SMS originating and terminating logs. For each Internet session recorded in the data logs, information on the anonymized unique identifier, access point name (APN), session start time, duration of the session, servicing cell at the start of the session, quantities of data uploaded and downloaded, and servicing network node is collected. Note cell information is only available for Meteor subscribed mobile devices. The available dataset consists of approximately three months of voice and SMS records from 09/11/2010 to 27/02/2011 and approximately two weeks of data records from 08/02/2011 to 27/02/2011, and contains tens of millions of CDR activities per day.

Background

In recent years, the wider availability of human mobility data gives rise to an increase of research activities involving human movement and spatial-temporal behavior at both urban and national scales. This is probably due to the variety of potential applications resulting from the classification and prediction of human mobility (Song et al., 2010a, 2010b), such as resource planning (Deruyck et al., 2012), dynamic transportation services (Steenbruggen et al., 2011), and understanding of human mobility behavior (Um et al., 2009; Park et al., 2010).

Ratti et al. (2006, 2007), Calabrese and Ratti (2006), Calabrese et al. (2011a), and Horanont and Shibasaki (2008) each focused on human activity mapping. Ahas et al. (2007a, 2007b, 2008, 2010a) demonstrated that suburban commuter movements and tourist movement dynamics could be extracted from sources of mobile telephony data. Tourist movements have also been examined by Kuusik et al. (2010, 2011), while methods for home and work location estimation and

population movement dynamics have been proposed by Ahas et al. (2010b), Silm and Ahas (2010), Calabrese et al. (2011b), Isaacman et al. (2011), Kelly et al. (2011), and Järv et al. (2012).

Areas associated with mass urban activity may be readily sourced from cell activity counts, as demonstrated by Reades et al. (2007, 2009), Andrienko et al. (2010a, 2010b), Becker et al. (2011a), Isaacman et al. (2010), Vieira et al. (2010) and Caceres et al. (2012). This type of work has generally focused on clustering areas of similar activity profiles with applications such as group movement patterns (Becker et al., 2011b) and marketing (Lin and Wan, 2009).

Other applications include examinations on the differences between rural and urban societies (Eagle et al., 2009b), the production of agent-based models of epidemic spread (Frias et al., 2011), the construction of social network graphs (Onnela et al., 2007a, 2007b; Nanavati et al., 2008; Kamola et al., 2011), studies on the characteristics of communication flow (Lambiotte et al., 2008; Krings et al., 2009; Kelly et al., 2011), and the identification of community structure using communication flow (Walsh and Pozdnoukhov, 2011; Ratti et al., 2010).

Movement and mobility predictability by mobile telephony data has also been a topic of discussion; most notably the works of González et al. (2008) and Song et al. (2010a, 2010b) have provided insights into the basic laws governing human motion and limiting thresholds on the predictability of human movement. The range of human motion was quantified in González et al. (2008) using the radius of gyration, which measures the overall range of an individual trajectory. This research demonstrated a stark contrast between actual human motion and classical random walk models. By measuring the entropy of individual trajectories, Song et al. (2010a, 2010b) showed that there was a potential predictability of 93 percent in user mobility across a mobile network operator's subscriber base, despite the significant differences in travel patterns.

Eagle et al. (2009a, 2009c) demonstrated the application and design of community structure algorithms that are appropriate for the identification of location clusters relevant to a mobile user's life. Bluetooth beacons located at subscriber homes supplemented validation of the proposed techniques. Mobility modeling algorithms were also developed using discrete Markov chains, for example by Park et al. (2010) in which it was demonstrated that the approximation of user mobility through Markov chains reproduces the slow, sub-polynomial growth predicted by the evolution of the radii of gyration. They also proposed how the eigenvalues and eigenvectors of a Markov chain were related to subscriber mobility. Järv et al. (2014), Steenbruggen et al. (2011), Caceres et al. (2008), Yim (2003), and Rose (2006) discussed how the availability of large quantities of human movement data had also been of interest in the transportation sciences.

Cell Coverage Regions

A mobile network topology is governed by coverage and capacity requirements. While cell coverage is generally constrained by geographical factors, capacity is generally influenced by traffic demand (Theodore, 2001; Korhonen, 2003; Mishra, 2004; Olsson et al., 2009; Ghosh et al., 2010). As traffic demand is strongly linked to population density, parameters such as cell size and density vary with mobile user density. Typically, mobile network topology for 2G, 3G, and 4G are designed separately. This results in the possibility of several cell towers of

varying telecommunication standards covering a single geographical area in different network configurations.

Using the collective cell tower data, namely the geo-spatial coordinates and network type of each cell, it is possible to approximate idealized cell site coverage areas. This is achieved via Voronoi tessellation (Okabe et al., 1992), of each mobile network of interest, where each center represents a cell base station site location. Note the accuracy of the tessellation in approximating cell coverage areas is affected by channel characteristics, topography of the area, and physical layer parameters that include among others, transmitter frequency, tilt, height, and transmission power (Mishra, 2004). These have not been factored into this analysis and hence may introduce some approximation error.

The coverage regions produced in this way are a reasonable approximation for cell site locations that lie within central locations. However, the absence of limiting threshold on coverage regions size means that cells sites along coastal regions are poorly approximated. As a result, the introduction of a maximum cell site radius of 20 km and 15 km is introduced for 2G and 3G networks, respectively. The choice of each threshold reflects the realistic limit for communication with each standard given our network topology. Each site radius S_r is calculated via

$$S_r = \sqrt{\frac{S_a}{\pi}} \tag{1}$$

where S_a denotes the cell site coverage area and is given by

$$S_a = \frac{1}{2}\left|\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i)\right| \tag{2}$$

where n is the number of vertices in the coverage polygon and (x,y) are the spatial coordinates of each point. To establish more specific cell sectored regions for analysis, each tessellation may be subdivided using cell transmitter azimuth angle information. Note multiple cells on one site with the same transmitter azimuth angles share the same cell coverage polygon. Cell radius (C_r) and area (C_a) can thus be estimated using Equations 1 and 2, while individual cell centroid Easting and Northing locations (C_x, C_y) are given by;

$$C_x = \frac{1}{6C_a}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \tag{3}$$

$$C_y = \frac{1}{6C_a}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \tag{4}$$

An illustrative example of cell boundary polygons for Meteor's 2G and 3G networks is given in Figure 1.

Movement Transition Flows

With the ever-increasing availability of trajectory data, observing the aggregated flows of people or animals between regions of interest has been a growing area of research. The work of Andrienko and Andrienko (2002, 2011), Andrienko et al. (2012), Buchin et al. (2011), and Doyle et al. (2011) has explored varying techniques to visualize and group similar movement patterns. Similarly, CDR subscriber trajectories may be exploited in this regard.

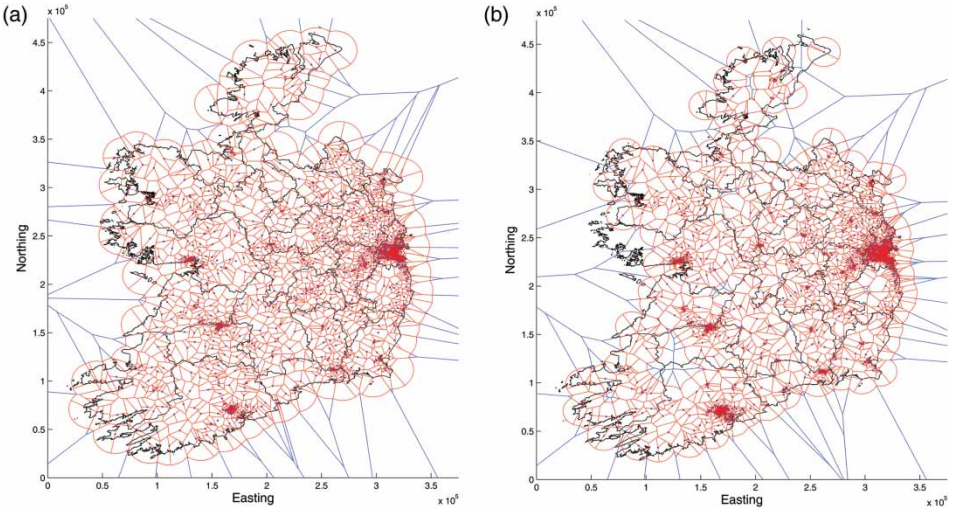

Figure 1: Sectored cell coverage of Meteor (a) 2G (4042 cells); and (b) 3G (6679 cells) networks in the Republic of Ireland. Blue lines indicate original Voronoi cell boundary edges, while red lines correspond to the cell boundary edges after the cell radius limiting threshold was applied.

A mobile device CDR trajectory is a path observed from CDR that a subscriber follows through a cell network as a function of time. Such trajectories are readily extracted from CDR by selecting device-specific temporally sorted cell tower connections. The trajectory may be spatially correlated by relating the spatial information of a servicing cell to each trajectory point. By counting the number of subscriber transitions between servicing cell towers in a given time frame, we can construct an aggregated transition matrix, T_a (k),

$$T_a(k) = \begin{pmatrix} t_{1,1}(k) & t_{1,2}(k) & \cdots & t_{1,R}(k) \\ t_{2,1}(k) & t_{2,2}(k) & \cdots & t_{2,R}(k) \\ \vdots & \vdots & \ddots & \vdots \\ t_{R,1}(k) & t_{R,2}(k) & \cdots & t_{RR}(k) \end{pmatrix} \tag{5}$$

where R is the number of regions of interest, and $t_{i,j}$ (k) is the transition intensity from region i to j at time k. For the mobile network considered in this research, T_a is a large matrix containing approximately 115 million elements ($R = 10{,}721$). For transition flow analysis, the matrix size needs to be reduced in order to lower both the computational complexity and memory requirements.

Several existing clustering methods can be used to combine adjacent cell coverage polygons of varying standards into a larger polygon representative of a symbolic location such as a population center. To this end, an agglomerative hierarchical based clustering (Xu and Wunsch, 2005) was chosen, using the Euclidean distance between each base station site location as a similarity measure. Initially, as a compromise between spatial accuracy and computational intensity, 500 clustered cell regions were selected. By performing a spatial union on the coverage polygons of individual cells within each cluster, the coverage region of each cluster can be generated and visualized as in Figure 2(a). This reduces T_a to T ($R = 500$). The flow of people between clustered regions and the geographical areas covered represents a proxy for the flow of people between individual

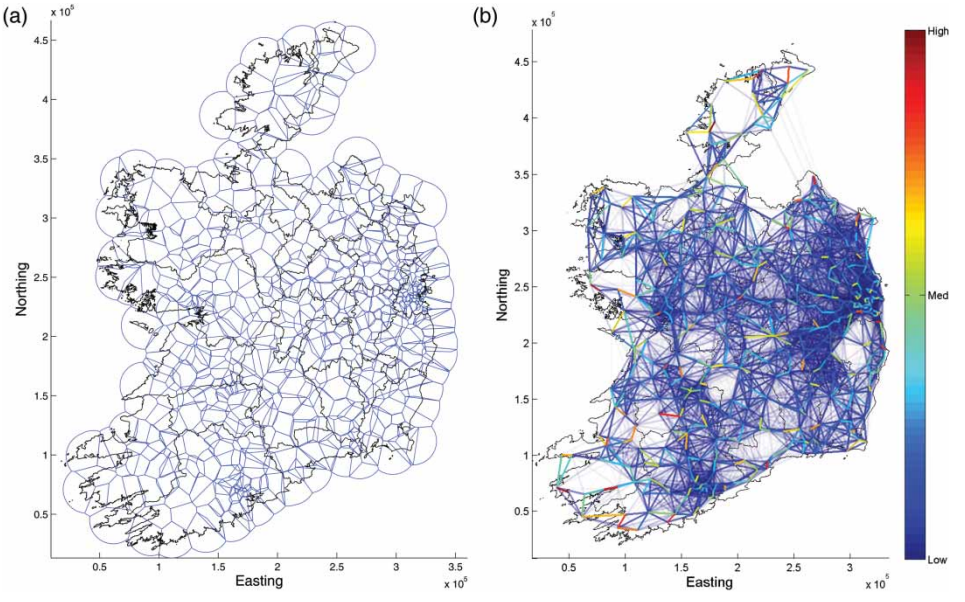

Figure 2: The proportional link strengths demonstrating observed transitions between clustered cell regions: (a) Clustered cell coverage regions; (b) Proportional strengths of transitions between clustered cell regions

population centers. The proportional link strengths demonstrating observed transitions among regions are illustrated in Figure 2(b).

The transition intensity or strength may also be observed temporally between two individual regions. A comparison of average daily activity volumes between Maynooth and Leixlip, towns in the north-east corner of County Kildare, Ireland, is displayed in Figure 3. With populations of 12,510 and 15,452, respectively, both are served by a commuter train service to Dublin and are close to the M4 motorway. From Figure 3, as expected Friday evening (4–8pm) has the highest transition volume in comparison to other days of the week. The commuting behavior that exists between Maynooth and Leixlip is also evident, as early spikes in transition intensity from Maynooth to Leixlip is recorded on week days, with the expected lull on weekends.

The temporal directionality of the transition intensity is an important feature when observing the flow of individuals. A sample of regional transition flows constructed using a customized interface between MATLAB® and Google Earth® is depicted in Figure 4. It illustrates Dublin city regional transition flows in time periods of low and high intensities, where the width of a connecting arrow corresponds to directional transition intensity. The anchor point for each connection arrow is positioned at the centroid position of each cluster, which is the center of the mass of cell base station sites independent of the number of cells in each site.

Population Density Estimation

A Markov chain is a mathematical representation of a stochastic process that undergoes step transitions from one state to another within a finite or countable state space. They have been extensively used in many domain areas including mobility modeling (Eagle et al., 2009a; Park et al., 2010), biomedical data analysis

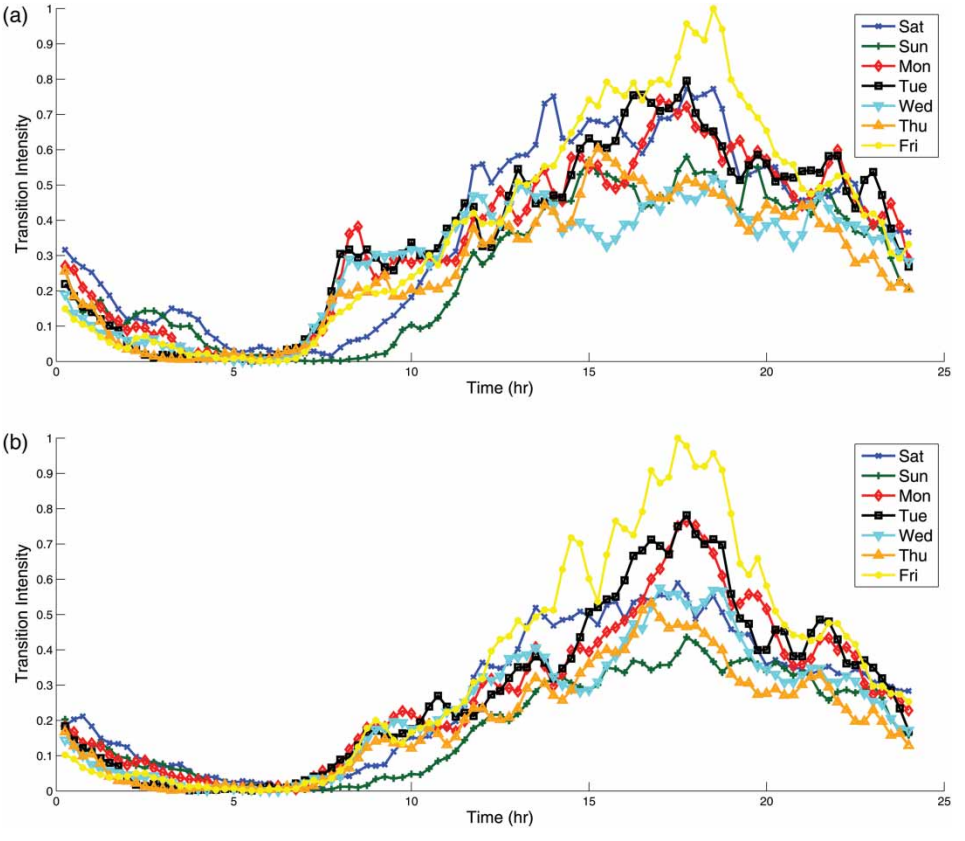

Figure 3: Average daily activity volumes of subscribers moving between clustered regions covering the towns of Maynooth and Leixlip.

(Ocan, 2005), and speech recognition (Rabiner, 1989; Ostendorf et al., 1996). A first-order, discrete-time Markov chain is used to mathematically represent a process, $S(k)$, $k = 0, 1,2, \ldots$, that undergoes random step transitions such that

$$P[S(k) = j|S(k-1) = i] = p_{ij}(k) \tag{6}$$

for all i, j and k (Ibe, 2008). Here $p_{i,j}$ (k) is the conditional probability that the process will transition from state i at time $k-1$ to state j at time k. A Markov chain which does not depend on the time unit, is known as a homogeneous Markov chain and implies

$$P[S(k) = j|S(k-1) = i] = p_{ij}. \tag{7}$$

From this, it is inferred that the state transition probability $p_{i,j}$ only depends on the current state and not on the sequence of previous states. This specific kind of memorylessness is called the Markov property.

Homogeneous Markov chains are useful when the state sequence, $S(k)$, $k = 0$, $1,2, \ldots$, is directly observable. By extracting a subscriber CDR trajectory, it is possible to directly observe an individual subscriber's cell tower state sequence. As previously discussed, cells may be linked to symbolic locations defined by their coverage regions, thus Markov chains may be used to model a mobile subscribers

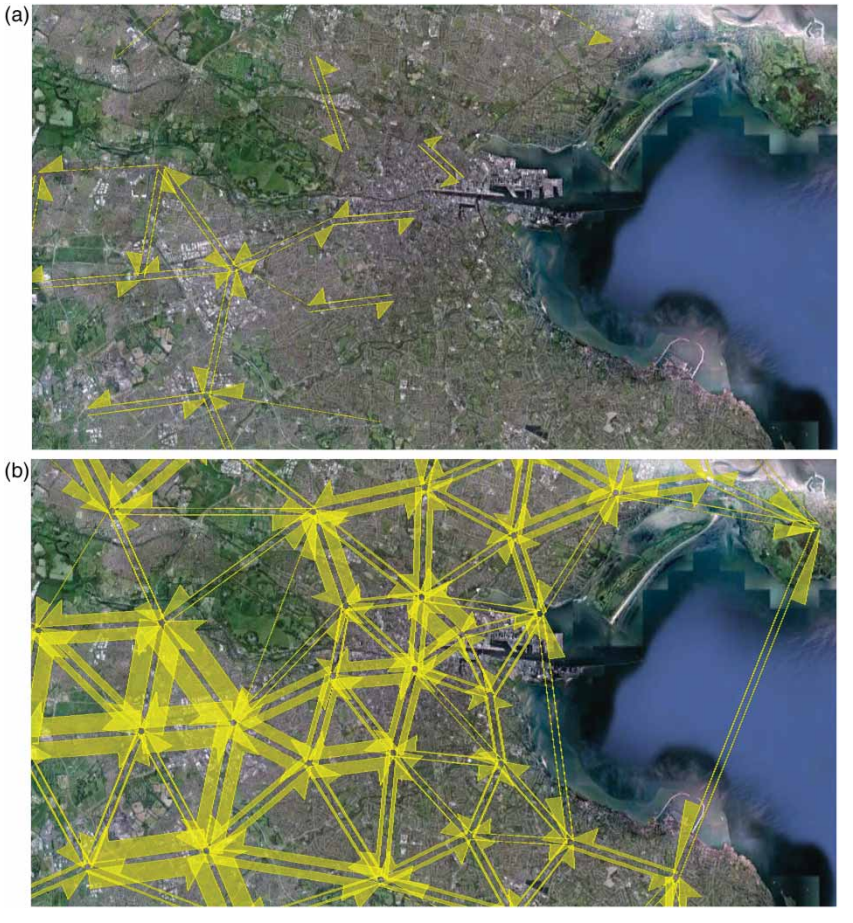

Figure 4: Dublin city regional transition flows in time periods of (a) low, and (b) high intensity, where the width of the connecting arrow corresponds to directional transition intensity. Note connecting arrows with very low intensity have been removed for visual clarity.

transient movement between these symbolic locations. Similarly, national mobility may be modeled when subscriber movements are combined into a single mobility model characterizing flow throughout the country.

$$T = \sum_{u=1}^{N} T_u \tag{8}$$

where N is the number of subscribers.

By counting the transitions between clustered regions from concurring activities, an aggregated transition matrix, T_u, can be constructed which summarizes the movement of the uth subscriber. To reduce high frequency transitions and to ensure uniformly sampled trajectories, each subscriber trajectory was sampled at a regular interval every 15 minutes from the start of the observation period. The procedure is illustrated in Figure 5. Within each 15-minute temporal window, the estimate of location is based on the last recorded servicing cell

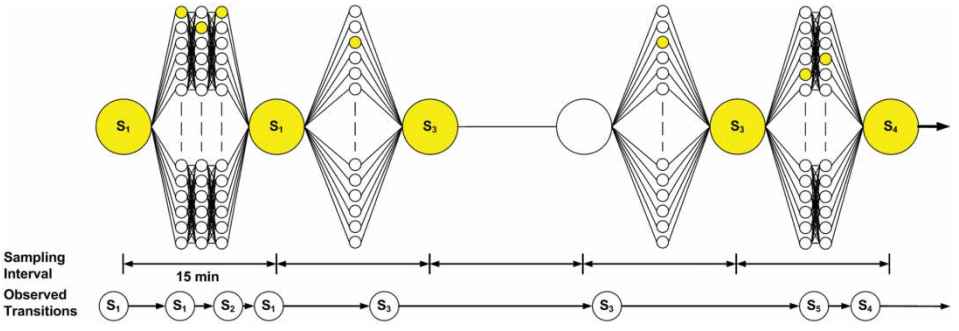

Figure 5: CDR trajectory state sequence sampling, where the output sequence $S = \{S_1, S_1, S_3, S_3, S_4\}$. Smaller yellow circles represent actual regional transitions within a sample period and larger yellow circles represent the observed output transition sequence before resampling. The larger white circle represents missing information.

tower recorded for that subscriber during that period. When no CDR activity occurs during a particular temporal window, no sample would be taken.

CDR trajectory sampling distributions are non-uniform and dictated by the activity profiles of individual subscribers (González et al., 2008; Ranjan et al., 2012). As a result, the number of observable subscribers fluctuates with normal cyclic national activity patterns. The net effect is that during periods of normal social activity (i.e., excluding public holidays such as New Year's Eve and large events such as St Patrick's Day) only a portion of the total subscriber population may be observed at any one instance. The percentage of active subscribers over a seven-day period observed using the aforementioned sampling technique is illustrated in Figure 6.

The transition matrix of each subscriber, T_u, may be translated into a probability transition matrix, P_u, by scaling each row such that

$$P_u = [p_{ij}]_{R\times R} \rightarrow \sum_{j=1}^{R} p_{ij} = 1, \quad \forall i \tag{9}$$

The resulting probability transition matrix, P_u, characterizes the movement of an individual subscriber. Using this framework, a Markov chain mobility model for each subscriber registered to the cellular network was obtained using a month

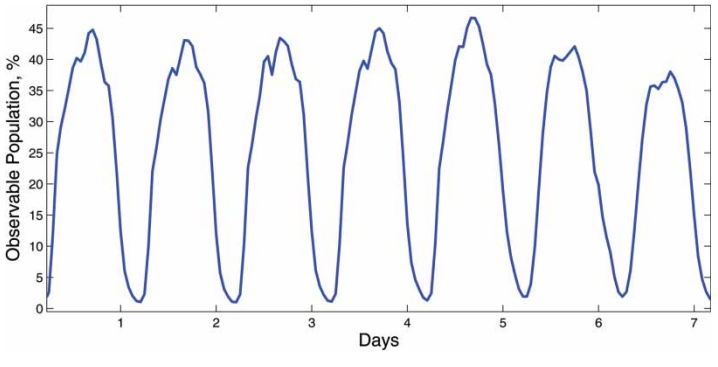

Figure 6: Typical percentage of subscribers which may be observed over a seven-day period.

long period of observations, where the number of observable state equals the number of regions of interest R.

If a Markov chain is irreducible, or ergodic, it is possible to go from every state to every other state in one or more steps (Grinstead and Snell, 1997). In such cases, the following holds

$$W = \lim_{j \to \infty} P^j \tag{10}$$

where W is a matrix with identical rows w, and all components of w sum to 1. Then $wP = w$, and any row vector v such that $vP = v$ is a constant multiple of w. A row vector w with the property $wP = w$ is called a fixed row vector for P and may be calculated by various methods, as outlined in by Grinstead and Snell (1997).

A fixed row vector characterizes the long term probability of a system being in a given state when the state transitions are governed by an underlining Markov chain. The fixed row vector of a mobile subscriber's mobility Markov chain, w_u, conveys the probability of observing that subscriber at a region in space over a long period of time. In order to extract national population counts using fixed row vectors, the home location of each subscriber needs to be segregated from these regions of interest. Here, the maximum weighting approach used involves assigning a subscriber's home location to the region that has the maximum fixed row vector weight. The population count of any region may then be calculated by counting the number of subscribers who are estimated to live in that region.

Alternatively, it is also hypothesized that the fixed row vector for the aggregated Markov chain mobility model, w_a, will convey the likelihood of observing the mobile operators active subscriber base at a particular region in space over a long period of time, which in turn provides an estimate for national population density. The model has the advantage that the calculation is based on the overall subscriber data rather than individual subscriber regions of interest. Hence, it is totally privacy preserving in the sense that none of the subscribers are individually tracked. Also, only a single calculation is required to form the aggregated fixed row vector, which is less computationally intensive than the maximum weighting approach, where the number of calculation is proportional to the number of subscribers. Each of the proposed techniques is summarized in Figure 7.

However, mobility Markov chains are not necessarily ergodic. Instead, they are typically sparse and may contain absorbing states (i.e. $p_{ii} = 1$): for example, if a subscriber is only ever serviced by a single cell tower or if its last trajectory sample was to a previously unvisited tower. For an aggregated mobility Markov chain an absorbing state may occur if subscribers from a particular region of interest never left that area during the time concerned. Likewise, a non-ergodic chain may form if every region of interest was not visited during the observation period.

To ensure each mobility Markov chain is ergodic and thereby non-absorbing, a regularization process similar to that used by Google® PageRank algorithm (Brin and Page, 1998) is introduced. It consists of applying a small transition weight to all state transitions before the fixed row vector is calculated and is given by

$$Q = \alpha P + (1 - \alpha)\frac{J}{R} \tag{11}$$

where Q is a modified Markov chain, J is a $R \times R$ matrix of ones, and α balances the learnt mobility patterns summarized by P with the influence of random

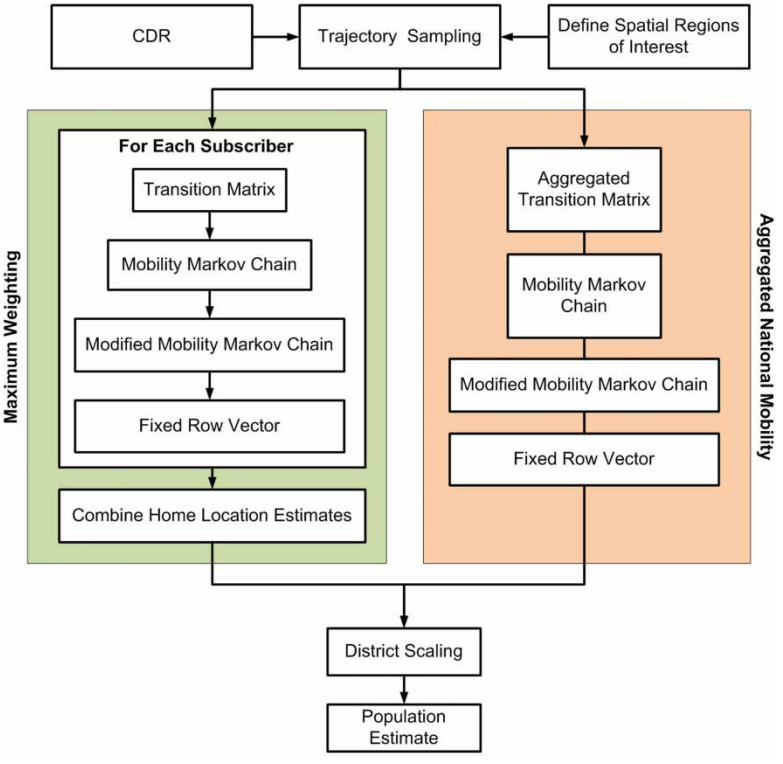

Figure 7: Overview of each population estimation technique.

transition probabilities introduced by the term J/R. To this end α is estimated as $(1-1/R)$. Note Q should satisfy to the following conditions:

(1) $p_{ii} < 1, \quad \forall i$

(2) $0 \leq p_{ij} \leq 1, \quad \forall i \forall j$

(3) $Q = [q_{ij}]_{R\times R} \rightarrow \sum_{j=1}^{R} q_{ij} = 1, \quad \forall i$

The incorporation of uniformly regularized weighting has the added benefit of accounting for the likelihood of observing transitions which relate to all plausible but unobserved journeys. Using the previously mentioned clustered cell regions ($R = 500$) as a proxy for spatial regions of interest, a visualization of Q for a randomly selected subscriber is visualized in Figure 8. Note all weights which are asymptotically zero are removed for visual clarity. Figure 8(a) depicts the weighted transition probabilities, with arch height corresponding to the probability of transitioning from one region to another. Figure 8(b) depicts the region rank estimated from Q given its fixed row vector w. The observed regional ranking suggests that the subscriber tends to travel in County Meath, with occasional trips into Dublin City.

Using the same clustered cell regions, an estimate of population density was calculated using both proposed approaches. The results are visualized in Figure 9, with density normalized between 0 and 1 for visual clarity. While maximum weighting relies only on information collected from individual subscribers, it is prone to noise in CDR data as it relies on the assumption that both a significant

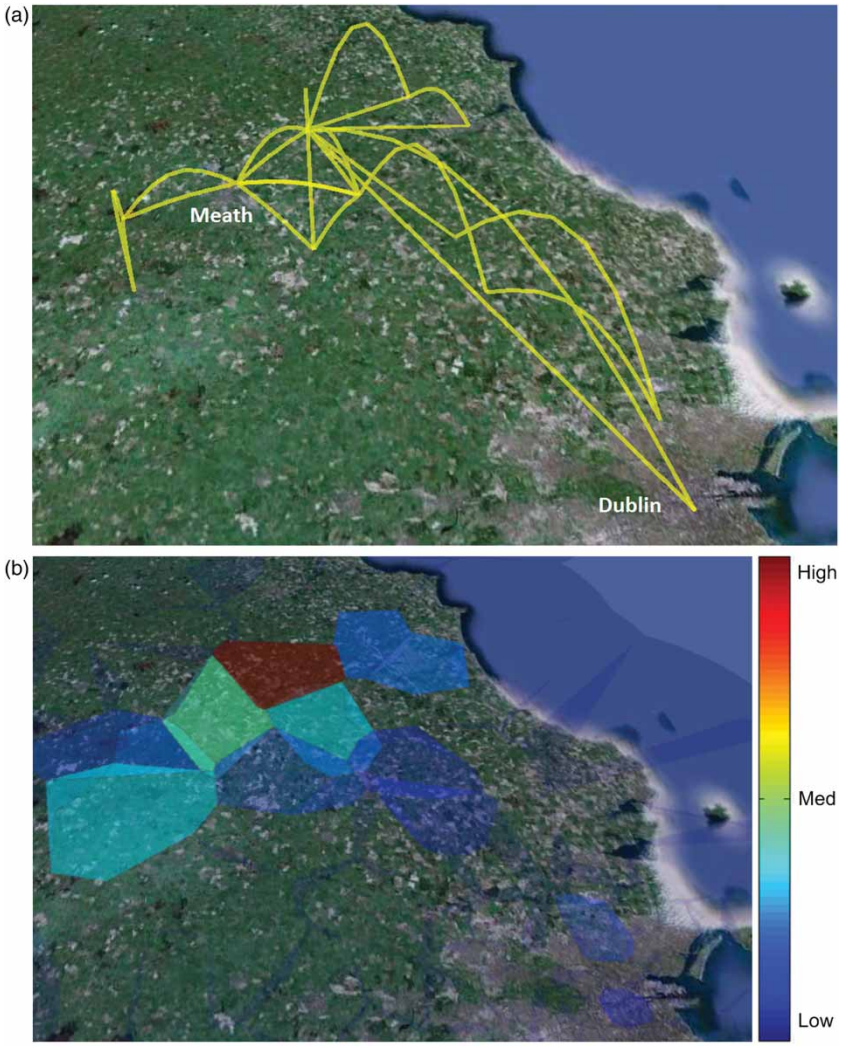

Figure 8: Visualization of an example: (a) subscriber transition probability matrix, Q_u, where heights correspond to transition intensity; (b) corresponding fixed row vector, w_u, where polygon color corresponds to state weight.

amount of time is spent and a significant amount of CDR activities are carried out at home locations by each subscriber, which may not be true. The mobility Markov chain is constructed such that it takes account of both aspects of user behavior and reflects that in the form of individual fixed row vectors. Comparing Figure 9 with the locations of towns and urban districts in the Republic of Ireland as presented in Figure 10, it can be seen that each area of high proportional population density corresponds well to urban centers and large towns.

District Scaling

Problems arise with the estimation of population density through CDR as both clustered cell regions and cell coverage areas do not naturally correspond to the

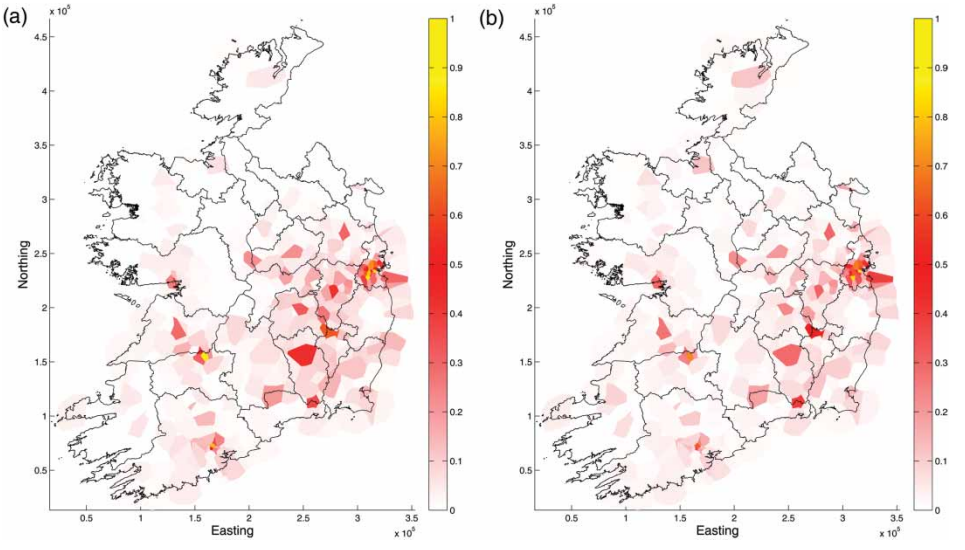

Figure 9: Population density estimates based on (a) individual home locations as sourced from subscribers; (b) aggregated mobility model.

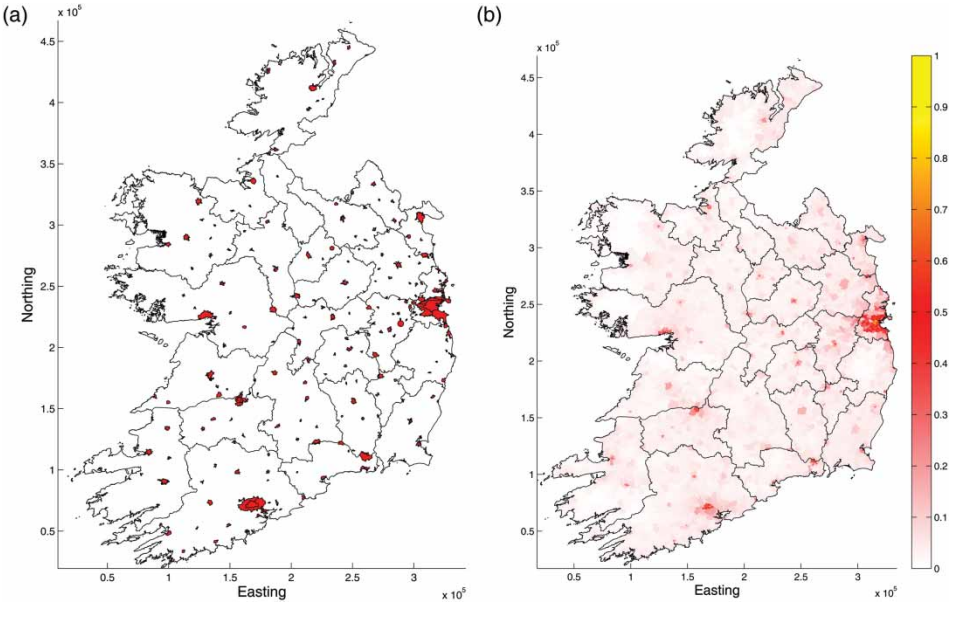

Figure 10: Sourced from the Central Statistic Office (CSO) Ireland: (a) town locations across the Republic of Ireland, with (b) corresponding normalized population density.

boundaries of districts or municipalities used by governments in the calculation of regional or local population. To allow direct comparisons between estimated populating density and census ground truth, where census data is supplied from the Central Statistics Office (CSO) Ireland, measurements of population observed at each region need to be redistributed to regions within officially

defined district boundaries. In Ireland, a common local area district used in the calculation of population is known as an Electoral Division (ED). There are approximately 3400 ED in Ireland ranging in size from several hundred squared meters in urban areas to several squared kilometers in rural regions.

A sample of the spatial distribution of buildings is displayed in Figure 11. The property use and location of each building is sourced from Geodirectory (2012). Established and maintained by An Post and Ordnance Survey Ireland (OSI), it is one of the most comprehensive building address databases available in the Republic of Ireland.

The procedure used to distribute estimated population onto EDs consists of several steps. First, assign each identified occupied home to an ED. Next, allocate each home in an individual ED to a cell region of interest that covers the dwelling concerned in its spatial coverage polygon. If multiple regions of interest cover a particular building, due to instances of overlapping 2G and 3G coverage, randomly assign a covering region from that list. Once all dwellings have been assigned, group them into a matrix H,

$$H = \begin{pmatrix} h_{1,1}(k) & h_{1,2}(k) & \cdots & h_{1,R}(k) \\ h_{2,1}(k) & h_{2,2}(k) & \cdots & h_{2,R}(k) \\ \vdots & \vdots & \ddots & \vdots \\ h_{M,1}(k) & h_{M,2}(k) & \cdots & h_{MR}(k) \end{pmatrix} \tag{12}$$

where $h_{i,j}$ is the number of homes from ith ED assigned to region of interest j, while M and R are the numbers of EDs and clustered regions, respectively. H is then ED

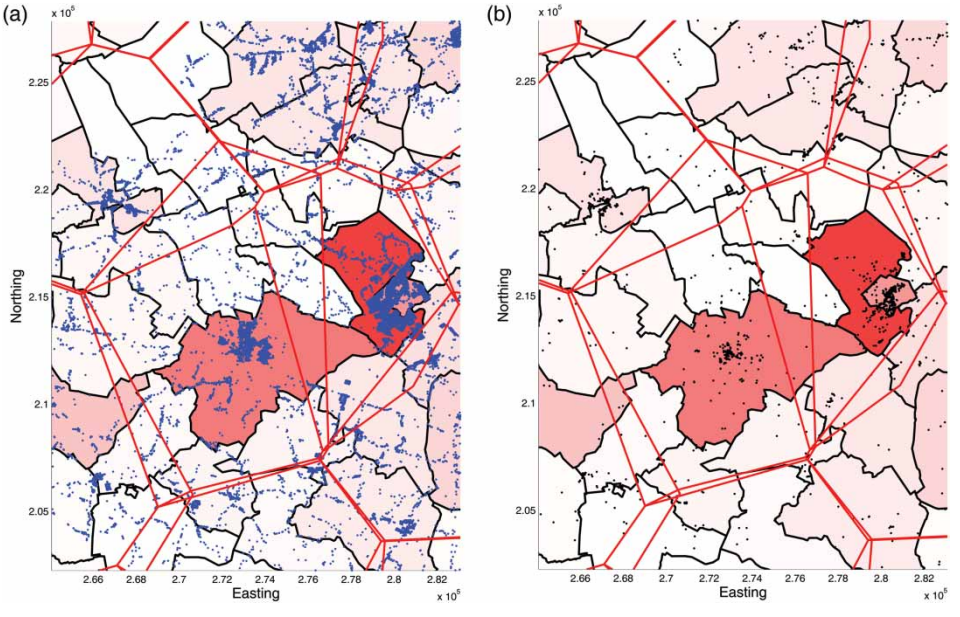

Figure 11: A sample of the spatial distribution of buildings across the Republic of Ireland. Also included is the cell coverage regions in the area (indicated by red lines), and ED boundaries (black lines); (a) Residential locations (blue dots), and (b) Commercial buildings (black dots).

normalized such that

$$\overline{H} = [\overline{h}_{ij}]_{M\times R} \rightarrow \sum_{i=1}^{M} \overline{h}_{ij} = 1, \quad \forall j \tag{13}$$

The number of subscribers living within an individual ED, N_i, is then estimated by,

$$\overline{N_i} = \sum_{j=1}^{R} N_j \overline{h_{ij}} \tag{14}$$

where N_j denotes the number of estimated subscribers living in a region of interest j. Using this method of distribution, Figure 12 depicts the population for each ED as estimated using the aforementioned fixed row vectors techniques for $R = 500$. In particular, proportional population estimated for the Dublin region is displayed in Figure 13. It can be observed that the ED segregated spatial distribution of subscribers between census data and both estimation techniques are strongly correlated. The discrepancies, such as in the city center region, could be attributed to the differences in the nature of census, where only residential addresses are recorded, and the activity observed by mobile networks.

Finally, to obtain the population density, perhaps a more important measure, H is transformed to D where

$$D = HA \tag{15}$$

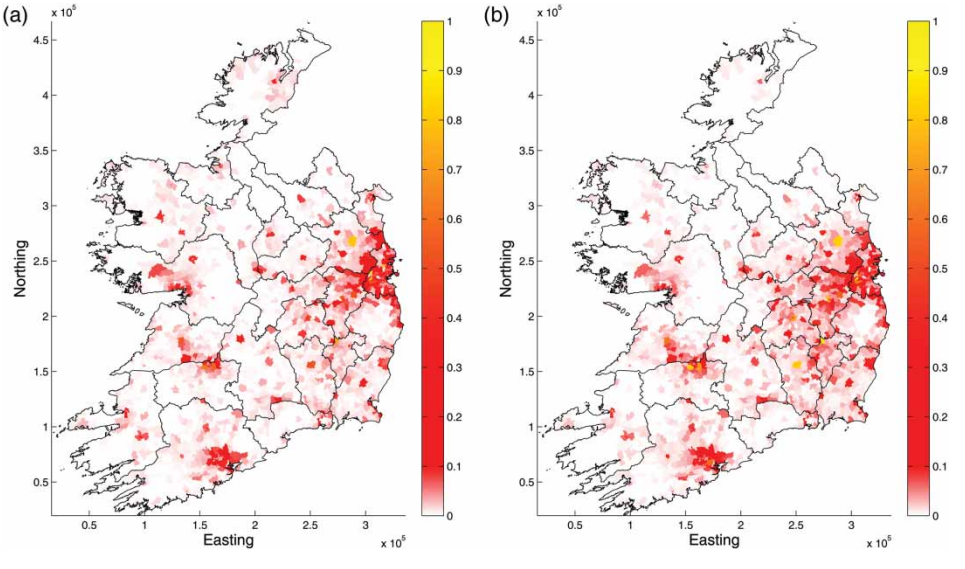

Figure 12: Electoral division population estimates across the Republic of Ireland from (a) maximum weighting; (b) aggregated vector.

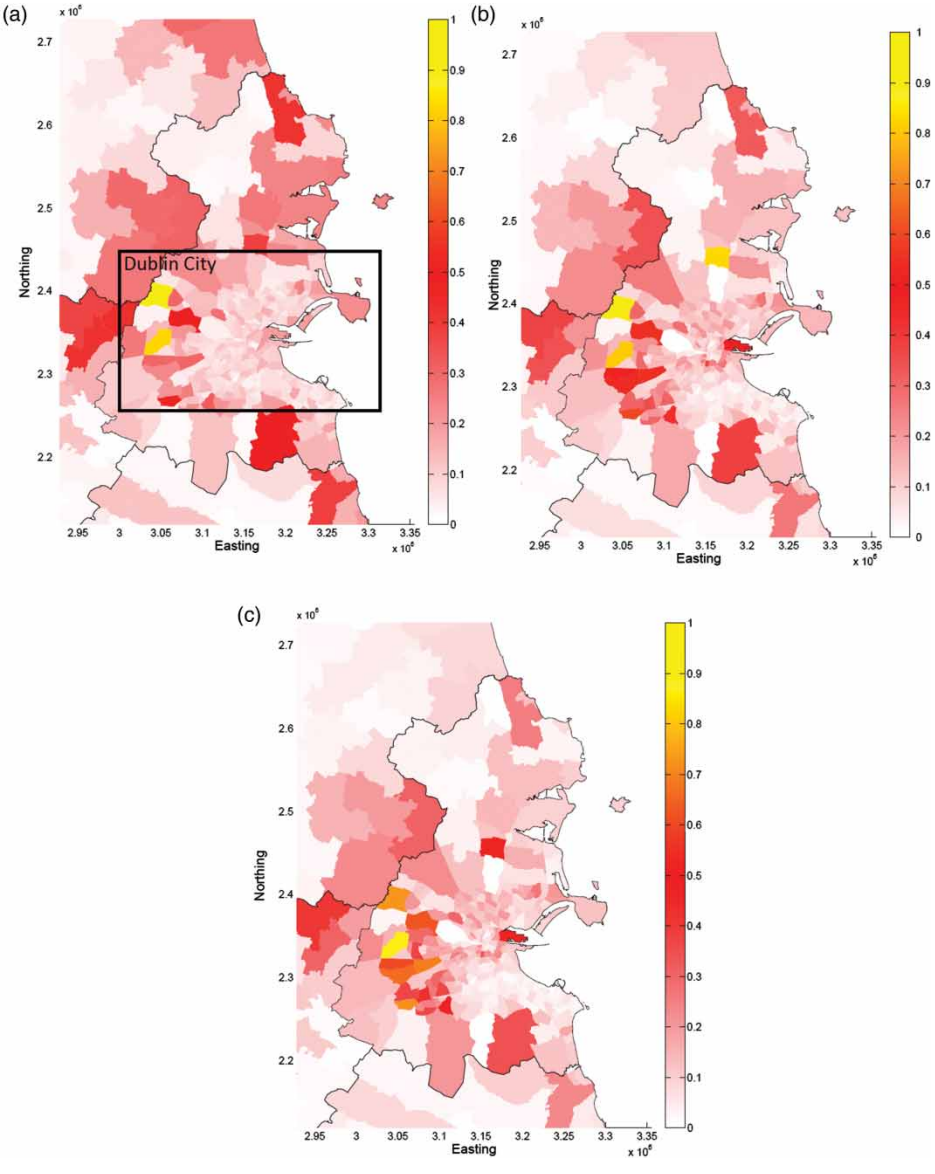

Figure 13: Proportional population estimated for the Dublin region from (a) census counts; (b) maximum weighting, and (c) aggregated vector.

and A is a diagonal matrix

$$A = \begin{pmatrix} \frac{1}{a_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{a_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & \frac{1}{a_M} \end{pmatrix} \tag{16}$$

where a_i is the spatial area of the ith ED. Then, D is normalized to $\bar{D}$ using ED columns similar to the normalized $\bar{H}$.

Census Validation

To validate the population estimate given by each fixed row vector from modified Markov chain mobility models, a direct comparison is drawn between the estimated populations and the Irish 2011 census (CSO, 2012). The correlation of census population counts with the population estimates based on maximum weighting was found to be 0.8645 while that with aggregated vector was 0.8088. The results indicate that both approaches have a strong spatial relationship to census count measurements.

On a national level, the spatial variance of percentage error between census data and estimated population is shown in Figure 14. In general, the mean of the percentage error between census data and estimates from aggregated vector is 0.64037 percent with a standard deviation of 0.51335 percent while the corresponding values for the population estimates based on maximum weighting are 0.54007 percent and 0.42464 percent, respectively. As a result, the maximum weighting approach appears to provide population estimates which match more closely with the census data. Note the percentage error is calculated based on normalized population count. From Figure 14, there is no clear pattern associated with the spatial distribution of error. In the absence of accurate Meteor subscriber demographics, it is hypothesized that estimation error fluctuates with the spatial density of meteor's subscriber population. If the age of each subscriber were known, this hypothesis could be tested by proportionally scaling each population estimate by its corresponding ED age profile.

Comparing the population density between census data and both techniques, correlations of 0.8661 and 0.8438 are obtained from maximum weighting and aggregated vector approaches, respectively. While the census correlations from $\bar{H}$ are similar to those from $\bar{D}$ using both techniques, it appears that the maximum weighting approach provides better estimates compared to aggregated

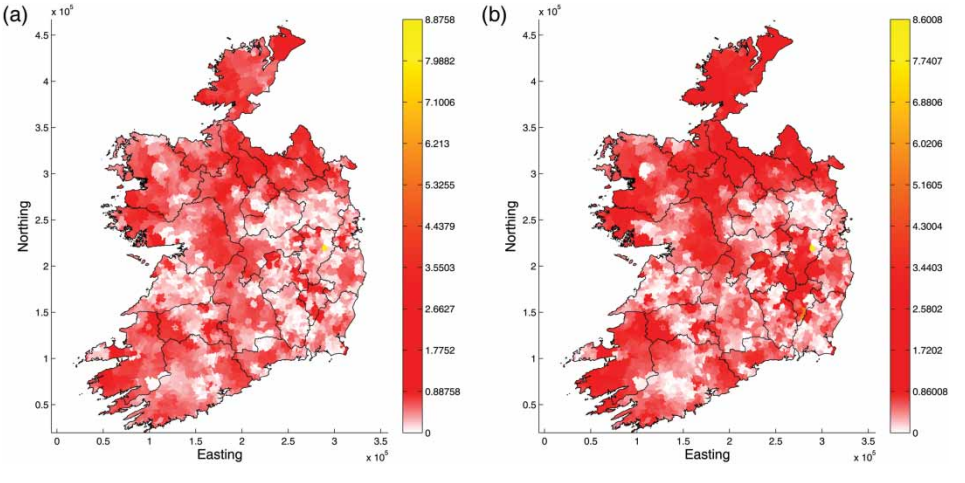

Figure 14: The spatial variance of percentage error between census data and estimated population from (a) maximum weighting; (b) aggregated vector.

vector. However, it is noted that maximum weighting is much more computationally intensive and privacy non-preserving as the assumed home location of each individual subscriber is tracked anonymously during the process.

On the regional scope, a comparison of census data and estimated population count for each county in the Republic of Ireland is summarized in Table 1. In each case, the total county population count is the sum of all ED counts which are located in a particular county. From this table, the maximum weighting approach and aggregated vector approach had a percentage mean squared error (MSE) with CSO census of 6.2288 and 4.0415, respectively. When the measurement for Dublin county is omitted the percentage MSE was 1.0491 and 2.0832, respectively.

In the study of mobility, correlation is an important measure, as it captures the relationship between measured population count and population estimates. By observing changes in correlation, we can capture relative displacements of population over time. Here, the maximum weighting approach and aggregated vector approach had correlations with CSO census of 0.98408 and 0.97731, respectively. When the measurement for Dublin county is omitted, the correlations were 0.9124 and 0.8515, respectively.

Table 1: A comparison of Central Statistics Office Ireland census data and estimated population density for each county in the Republic of Ireland, where measurements are the percentage of total population.

County	Central Statistics Office Ireland (%)	Maximum Weighting (%)	Aggregated Vector (%)
Carlow	1.19	2.22	2.93
Cavan	1.60	0.53	0.39
Clare	2.55	2.81	3.16
Cork	11.31	10.24	11.98
Donegal	3.51	1.14	0.48
Dublin	27.75	39.40	35.03
Galway	5.46	5.27	4.17
Kerry	3.17	1.36	0.92
Kildare	4.58	6.09	7.24
Kilkenny	2.08	2.24	2.98
Laois	1.76	2.35	3.33
Leitrim	0.69	0.14	0.07
Limerick	4.18	5.36	6.82
Longford	0.85	0.50	0.44
Louth	2.68	1.34	0.70
Mayo	2.85	1.35	0.99
Meath	4.01	3.74	3.55
Monaghan	1.32	0.27	0.15
Offaly	1.67	1.43	1.83
Roscommon	1.40	0.66	0.52
Sligo	1.43	0.65	0.34
Tipperary	3.46	2.02	2.37
Waterford	2.48	1.97	1.69
Westmeath	1.88	1.89	2.19
Wexford	3.17	2.63	3.14
Wicklow	2.98	2.39	2.62
MSE	0	6.2288	4.0415
MEE Excluding Dublin	0	1.0491	2.0832

The strong correlations (>0.95) of estimated counts demonstrates the effectiveness of Markov chain fixed row vector analysis for approximating population density based on CDR at county level. However, while similar correlations are also observed at the Electoral Division level, there is an overall reduction in the measured correlations at ED relative to county level. This may be a result of greater fluctuations in proportional population representation over smaller geographical regions, caused by the spatial variations of mobile operator penetration.

Moreover, the consistently high levels of errors found in Dublin County (city center in particular; refer to Figure 13) might illustrate one possible limitation of a census, i.e., that it only contains records of residential address. Indeed, subscribers may tend to carry out a substantial amount of CDR activities in densely populated areas. Alternatively, a substantial number of subscribers might be staying in the city center temporally for an extended period of time (shopping, night-out, etc.), and may not be recorded as residing there through a census.

Evaluating the impact of spatial resolution has on accuracy and correlation with respect to census measurements of ED population density and population count, it was observed that the aggregated vector approach had an optimal region of spatial resolution. Figure 15 illustrates the observed measurements of

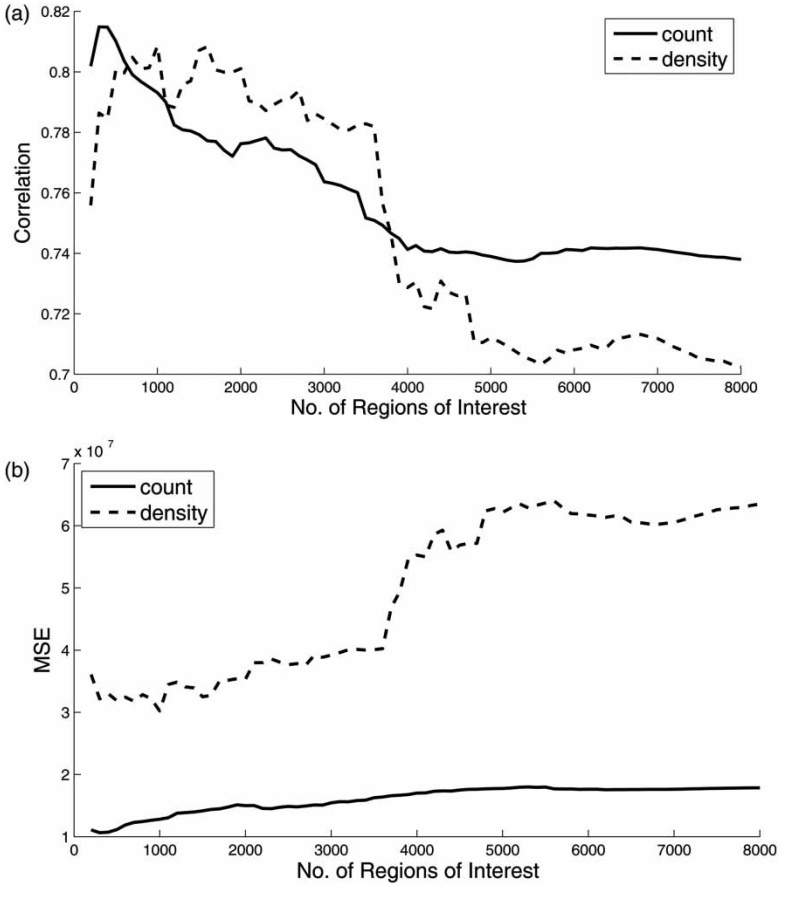

Figure 15: Impact of spatial resolution on (a) correlation; (b) MSE between census data and population density estimates from the aggregated vector approach.

correlation and MSE over a range of *R*, a parameter that generally corresponds to spatial resolution.

Results indicate that if spatial regions of interest are too small, the approximation error due to transition noise, such as that introduced by localization uncertainty and network penetration, will increase. This suggests that spatial regions of interest should mirror national population centers, where towns and other urban district boundaries are maintained. Note, the effects of spatial resolution on subscriber based maximum weighting could not be evaluated in this extent due to computational limitations. Nonetheless, results presented by Eagle et al. (2009a, 2009c) show that when locating the home of a subscriber, a set of home cell towers is typically found. Thus, similar to the results obtained for the aggregated vector approach, it is reasonable to assume that the subscriber maximum weighting approach will also have a threshold of R above which transition noise will affect the accuracy of home location estimation.

Another important factor when evaluating performance is temporal homogeneity. If each mobility Markov chain used in the calculation of population is temporally heterogeneous, it is statistically different when computed from different observation periods, meaning each estimate of population will be different. Thus, the accuracy of the proposed technique will be a function of the observation period. The study of temporal homogeneity is the subject of future work.

Conclusion

This paper used call detail records (CDR) from Meteor, a mobile network operator in the Republic of Ireland, to visualize the regional flows of people across Ireland. The use of CDR location estimates for the identification of significant mobile subscriber regions of interest and the estimation of population density was also investigated. Potential practical applications include utility load forecasting and dynamic transportation services as they require a human activity feed/source, whether live or historic, to help predict human-related service demands.

The high correlation of estimated results with the Central Statistics Office Ireland census data demonstrates the effectiveness of Markov chain fixed row vector analysis for approximating population density proportions. While population estimation via such techniques does not result in the fine-grained measurements achieved through a census, population fluctuations can be monitored at much finer temporal resolutions and lower cost.

Each approach to the estimation of population density discussed has its own advantages and disadvantages. Although the estimates derived from the subscriber-based maximum weighting approach are more accurate, the calculation of each individual vector is more computationally intensive compared to the single calculation required for the aggregated vector approach. Moreover, the aggregated approach is totally privacy preserving, as calculations are based on the overall subscriber data instead of individual subscriber regions of interest.

Acknowledgments

The 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 under their Embark Initiative in part-

nership 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).

Notes on Contributors

John Doyle is a post-doctoral research fellow at the National University of Ireland Maynooth, where his research is focused on the development of data mining and machine learning applications related to semiconductor manufacturing.

Peter Hung is a post-doctoral research fellow at the National University of Ireland Maynooth. His research interests include system identification, online and offline signal processing, pattern classification, and machine learning, as well as geocomputation.

Ronan Farrell is the director of the Callan Institute for applied ICT at National University of Ireland Maynooth. He has published over a hundred peer-reviewed papers and holds three patents; his research interests include radio and wireless systems, with particular interest in their use in non-communication domains.

Seán McLoone is a professor and director of the Energy Power and Intelligent Control Research Cluster in the School of Electronics, Electrical Engineering, and Computer Science at Queen's University Belfast, Northern Ireland. He also holds a joint appointment in the department of electronic engineering, National University of Ireland Maynooth, where he was head of department from 2009 to 2012.

Bibliography

R. Ahas, A. Aasa, U. Mark, T. Pae, and A. Kull, "Seasonal Tourism Spaces in Estonia: Case Study with Mobile Positioning Data," *Tourism Management* 28: 3 (2007a) 898–910.

R. Ahas, A. Aasa, S. Silm, and M. Tiru, "Mobile Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia," *Information and Communication Technologies in Tourism* 2007 (2007b) 119–128.

R. Ahas, A. Aasa, A. Roose, U. Mark, and S. Silm, "Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study," *Tourism Management* 29: 3 (2008) 469–486.

R. Ahas, A. Aasa, S. Silm, and M. Tiru, "Daily Rhythms of Suburban Commuter Movements in the Tallinn Metropolitan Area: Case Study with Mobile Positioning Data," *Transportation Research Part C: Emerging Technologies* 18: 1 (2010a) 45–54.

R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," *Journal of Urban Technology* 17: 1 (2010b) 3–27.

G. Andrienko and N. Andrienko, "A General Framework for Using Aggregation in Visual Exploration of Movement Data," *Cartographic Journal* 47: 1 (2002) 22–40.

G. Andrienko, N. Andrienko, P. Bak, S. Bremm, D. Keim, T. von Landesberger, C. Politz, and T. Schreck, "A Framework for Using Self-Organising Maps to Analyse Spatio-Temporal Patterns, Exemplified by Analysis of Mobile Phone Usage," *Journal of Location Based Services* 4: 3–4 (2010a) 200–221.

G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Plitz, "Discovering Bits of Place Histories from People's Activity Traces," in Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium On (2010b) 59–66.

N. Andrienko and G. Andrienko, "Spatial Generalization and Aggregation of Massive Movement Data. Visualization and Computer Graphics," *IEEE Transactions On* 17: 2 (2011) 205–219. ISSN 1077–2626. doi:10.1109/TVCG.2010.44.

N. Andrienko, G. Andrienko, H. Stange, T. Liebig, and D. Hecker, "Visual Analytics for Understanding Spatial Situations from Episodic Movement Data," *Kunstl Intelligenz* 74 (2012) 1–11. ISSN 0933–1875. 10.1007/s13218–012-0177-4.

R.A. Becker, R. Caceres, K. Hanson, J.M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A Tale of One City: Using Cellular Network Data for Urban Planning," *Pervasive Computing, IEEE* 10: 4 (2011a) 18–26. ISSN 1536-1268.

R.A. Becker, R. Cáceres, K. Hanson, J.M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Clustering Anonymized Mobile Call Detail Records to Find Usage Groups," 1st Workshop on Pervasive Urban Applications (PURBA) (2011b).

S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems* 30 (1998) 107–117. ISSN 0169-7552.

K. Buchin, B. Speckmann, and K. Verbeek, "Flow Map Layout Via Spiral Trees. Visualization and Computer Graphics," *IEEE Transactions On 17:* 12 (2011) 2536–2544. ISSN 1077-2626. doi:10.1109/TVCG.2011.202

N. Caceres, J.P. Wideberg, and F.G. Benitez, "Review of Tra_c Data Estimations Extracted from Cellular Networks," *Intelligent Transport Systems, IET* 2: 3 (2008) 179–192. ISSN 1751-956X.

R. Caceres, J. Rowland, C. Small, and S. Urbanek, "Exploring the Use of Urban Greenspace Through Cellular Network Activity," *Proceedings of 2nd Workshop on Pervasive Urban Applications* (Newcastle, June 18–22, 2012).

F. Calabrese and C. Ratti, "Real Time Rome," *Networks and Communication Studies* 20: 3–4 (2006) 247–258.

F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. Intelligent Transportation Systems," *IEEE Transactions On* 12: 1 (2011a) 141–151. ISSN 1524-9050.

F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating Origin-Destination Flows Using Mobile Phone Location Data," *Pervasive Computing, IEEE* 10: 4 (2011b) 36–44, ISSN 1536-1268.

CSO, Population Classified by Area, vol. 1, <http://www.cso.ie/en/census/census2011reports/census2011populationclassifiedbyareaformerlyvolumeone/> Accessed August, 2012.

M. Deruyck, E. Tanghe, W. Joseph, and L. Martens, "Characterization and Optimization of the Power Consumption in Wireless Access Networks by Taking Daily Traffic Variations into Account," *EURASIP Journal on Wireless Communications and Networking* (2012) 1–12.

J. Doyle, P. Hung, D. Kelly, S. McLoone, and R. Farrell, "Utilizing Mobile Phone Billing Records for Travel Mode Discovery," In IET Irish Signals and Systems Conference, 2011.

N. Eagle, A. Clauset, and J.A. Quinn, "Location Segmentation, Inference and Prediction for Anticipatory Computing," in Proceedings of AAAI Spring Symposium on Technosocial Predictive Analytics (Stanford, CA, 2009a).

N. Eagle, Y.-A. de Montjoye, and L.M.A. Bettencourt, "Community Computing: Comparisons between Rural and Urban Societies Using Mobile Phone Data," in Computational Science and Engineering, 2009. CSE '09. International Conference on, volume 4 (2009b) 144–150.

N. Eagle, J.A. Quinn, and A. Clauset, "Methodologies for Continuous Cellular Tower Data Analysis," *In Pervasive* 63 (2009c) 342–353.

E. Frias Martinez, G. Williamson, and V. Frias-Martinez, "An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information," in Privacy, Security, Risk and Trust (passat), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (socialcom) (2011) 57–64.

Geodirectory, http://www.geodirectory.ie/, accessed Aug. 2012.

A. Ghosh, J. Zhang, J.G. Andrews, and R. Muhamed, *Fundamentals of LTE*, 1st edition (Upper Saddle River, NJ, USA: Prentice Hall Press, 2010). ISBN 0137033117, 9780137033119.

M.C. González, C.A. Hidalgo, and A.L. Barabási, "Understanding Individual Human Mobility Patterns," *Nature* 453: 7196 (2008) 779–782.

C.M. Grinstead and J.L. Snell, *Introduction to Probability: Second Revised Edition* (Providence: American Mathematical Society, 1997).

T. Horanont and R. Shibasaki, "An Implementation of Mobile Sensing for Large-Scale Urban Monitoring," In the Proceedings of Urbansense08 (2008).

O. Ibe, *Markov Processes for Stochastic Modeling* (New York: Elsevier, 2008).

S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky, "A Tale of Two Cities," In Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications (ACM, 2010) 19–24.

S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in Peoples Lives from Cellular Network Data," *Pervasive Computing* 10 (2011) 133–151.

O. Järv, R. Ahas, E. Saluveer, B. Derudder, and F. Witlox, "Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records," *PLoS One* 7: 11 (2012) e49171.

O. Järv, R. Ahas, and F. Witlox, "Understanding Monthly Variability in Human Activity Spaces: A Twelve-Month Study Using Mobile Phone Call Detail Records," *Transportation Research C: Emerging Technologies* 38 (2014) 122–135.

M. Kamola, E. Niewiadomska-Szynkiewicz, and B.C. Piech, "Reconstruction of a Social Network Graph from Incomplete Call Detail Records," in Computational Aspects of Social Networks (CASoN), 2011 International Conference on (2011) 136–140.

D. Kelly, J. Doyle, and R. Farrell, "Analyzing Ireland's Social and Transport Networks Using Sparse Cellular Network Data," In IET Irish Signals and Systems Conference (2011).

J. Korhonen, *Introduction to 3G Mobile Communications*, Second Edition (London: Artech House, 2003).

G. Krings, F. Calabrese, C. Ratti, and V.D. Blondel, "Urban Gravity: A Model for Inter-City Telecommunication Flows," *Journal of Statistical Mechanics: Theory and Experiment* 2009: 07 (2009) 936–939.

A. Kuusik, R. Ahas, and M. Tiru, "The Ability of Tourism Events to Generate Destination Loyalty Towards the Country: An Estonian Case Study," in S. Meltsamees and J. Reiljan, ed., *Discussions of Estonian Economic Policy XVIII* (Berlin: Berliner Wissenchafts-Verlag, 2010) 140–155.

A. Kuusik, M. Tiru, and R. Ahas, "Innovation in Destination Marketing: The Use of Passive Mobile Positioning for the Segmentation of Repeat Visitors in Estonia," *Baltic Journal of Management* 6 (2011) 378–399.

R. Lambiotte, V.D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical Dispersal of Mobile Communication Networks," *Physica A: Statistical Mechanics and its Applications* 387: 21 (2008) 5317–5325.

Q. Lin and Y. Wan, "Mobile Customer Clustering Based on Call Detail Records for Marketing Campaigns," in Management and Service Science, 2009. MASS '09. International Conference on (2009) 1–4.

A.R. Mishra, *Fundamentals of Cellular Network Planning and Optimization: 2 g/2.5 g/3 g... Evolution to 4 g* (Hoboken: John Wiley & Sons, 2004).

A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyzing the Structure and Evolution of Massive Telecom Graphs. Knowledge and Data Engineering," *IEEE Transactions On* 20: 5 (2008) 703–718. ISSN 1041-4347.

R. Ocan-Riola, "Non-homogeneous Markov Processes for Biomedical Data analysis," *Biometrical Journal* 47: 3 (2005) 369–376.

A. Okabe, B.N. Boots, K. Sugihara, and S.N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Chichester: Wiley & Sons, 1992).

M. Olsson, S. Sultana, S. Rommer, L. Frid, and C. Mulligan, *SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution* (Waltham: Academic Press, 2009).

J.P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A.L. Barabási, and J. Kertész, "Analysis of a Large-Scale Weighted Network of One-To-One Human Communication," *New Journal of Physics* 9 (2007a) 179.

J.P. Onnela, J. Saramaki, J. Hyvonen, G. Szabó, D. Lazer, K. Kaski, J. Kertesz, and A.L. Barabasi, "Structure and Tie Strengths in Mobile Communication Networks," *Proceedings of the National Academy of Sciences* 104: 18 (2007b) 7332–7336.

M. Ostendorf, V.V. Digalakis, and O.A. Kimball, "From Hmm's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *Speech and Audio Processing, IEEE Transactions On* 4: 5 (1996) 360–378. ISSN 1063-6676.

J. Park, D.S. Lee, and M.C. González, "The Eigenmode Analysis of Human Motion," *Journal of Statistical Mechanics: Theory and Experiment* 2010: 11 (2010) 11021–11036.

L.R. Rabiner, "A Tutorial on Hidden MARKOV Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE* 77: 2 (1989) 257–286. ISSN 0018-9219.

G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are Call Detail Records Biased for Sampling Human Mobility?," *SIGMOBILE Mob. Comput. Commun. Rev.* 16: 3 (2012) 33–44. [Online]. Available: http://doi.acm.org/10.1145/2412096.2412101

C. Ratti, S. Williams, D. Frenchman, and R.M. Pulselli, "Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis," *Environment and Planning B Planning and Design* 33: 5 (2006) 727–748.

C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer, "Mobile Landscapes: Graz in Real Time," G. Gartner, W. Cartwright, and M. Petersons, eds., *Location Based Services and Telecartography*, Berlin: Springer, 2007.

C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the Map of Great Britain from a Network of Human Interactions," *PLoS ONE* 5: 12 (2010) 14248.

J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular Census: Explorations in Urban Data Collection," *IEEE Pervasive Computing* 6: 3 (2007) 30–38.

J. Reades, F. Calabrese, and C. Ratti, "Eigenplaces: Analyzing Cities Using the Space- Time Structure of the Mobile Phone Network," *Environment and Planning B: Planning and Design* 36: 5 (2009) 824–836.

G. Rose, "Mobile Phones as Traffic Probes: Practices, Prospects and Issues," *Transport Reviews* 26: 3 (2006) 275–291.

S. Silm and R. Ahas, "The Seasonal Variability of Population in Estonian Municipalities," *Environment and Planning. A* 42: 10 (2010) 2527–2546.

C. Song, T. Koren, P. Wang, and A.L. Barabási, "Modelling the Scaling Properties of Human Mobility," *Nature Physics* 6: 10 (2010a) 818–823.

C. Song, Z. Qu, N. Blumm, and A.L. Barabási, "Limits of Predictability in Human Mobility," *Science* 327: 5968 (2010b) 1018–1021.

J. Steenbruggen, M. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile Phone Data from GSM Networks for Traffic Parameter and Urban Spatial Pattern Assessment: A Review of Applications and Opportunities," *GeoJournal* 78 (2011) 1–21.

The Irish Grid, The Irish Grid: A Description of the Coordinate Reference System Used in Ireland, <http://www.osi.ie/Services/GPS-Services/Reference-Information/Irish-Grid-Reference-System.aspx> Accessed September, 2011.

C. Theodore, *Wireless Communications: Principles and Practice*, 2nd edition (Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001). ISBN 0130422320.

J. Um, S.W. Son, S.I. Lee, H. Jeong, and B.J. Kim, "Scaling Laws Between Population and Facility Densities," *Proceedings of the National Academy of Sciences* 106: 34 (2009) 14236–14240.

M.R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Mandnez, "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics," in IEEE Second International Conference on Social Computing (SocialCom) (2010) 241–248.

F. Walsh and A. Pozdnoukhov, "Spatial Structure and Dynamics of Urban Communities," *Proceedings of the 1st Workshop on Pervasive Urban Applications* (San Francisco, June 12–15, 2011).

R. Xu and D. Wunsch, "Survey of Clustering Algorithms," *Neural Networks, IEEE Transactions On* 16: 3 (2005) 645–678. ISSN 1045-9227.

Y. Yim, "The State of Cellular Probes. Institute of Transportation Studies, Research Reports, Working Papers," Proceedings, Institute of Transportation Studies (UC Berkeley, 2003).