Deutsche Bank
COO Chief Data Office

Estimating Large Scale Population Movement
ML Dublin Meetup

John Doyle PhD
Assistant Vice President
CDO Research & Development
Science & Innovation

john.doyle@db.com
https://www.db.com/ireland/

Estimating Large Scale Population Movement
Presentation Outline

- Introduction: Research Motivation & Data
- Mobility: Trajectories & Large Scale Movement
- Population: Density Estimates
- Application: How to Use the Data
- Conclusions: Summary of the Research

StratAG
Strategic Research in Advanced Geotechnologies
Taighde Straitéiseach sna hArd-Gheoteicneolaíochtaí

sfi
science foundation ireland
fondúireacht eolaíochta éireann

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

esri Ireland Understanding our world.

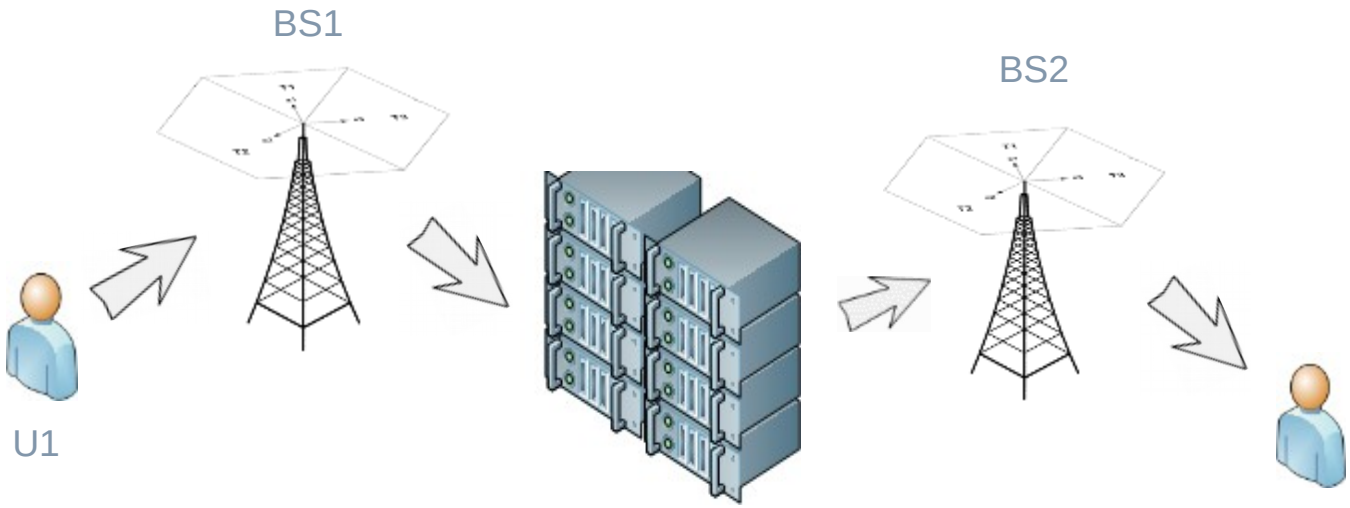
meteor

Research Motivation

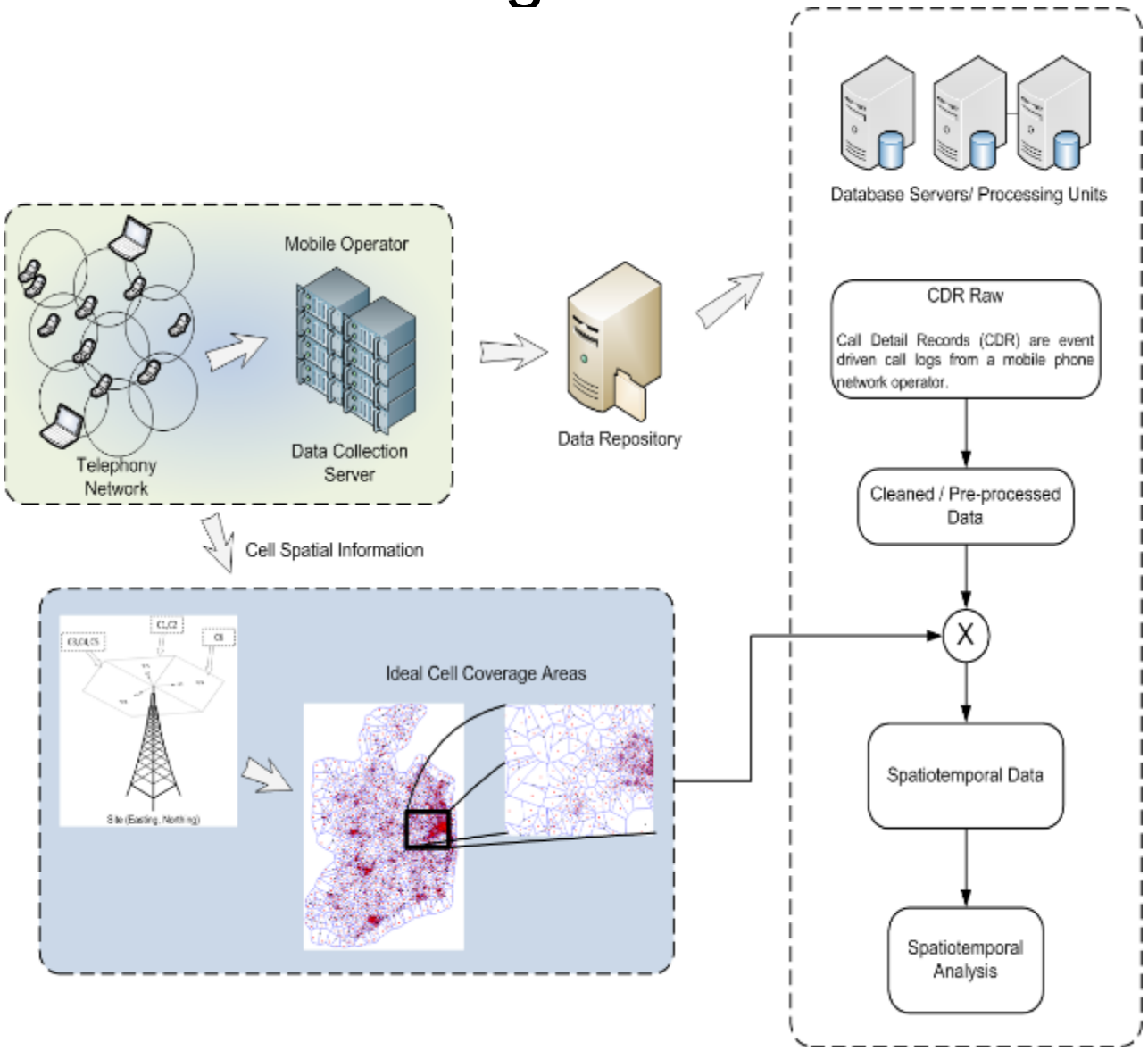
- Measuring the movement of people is a fundamental activity in modern society
- Movement data is used by:
  - Transportation services
  - Planning authorities
  - Governmental departments
- It is also the primary data source used in the delivery of mobile communications and location based services
- This research documents novel algorithms and techniques for the estimation of movement from mobile telephony data addressing practical issues related to sampling, privacy and spatial uncertainty.

Mobile Telephony Data

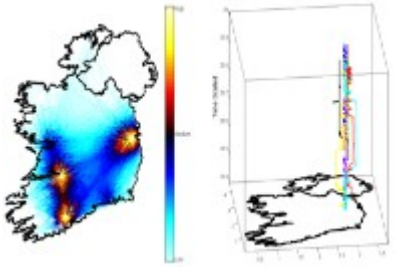
- Call Detail Records (CDR)
  - CDR is a data log of recorded Call, SMS and data activities which occur on a mobile operator's telephony network.
  - Approximately 1 million customers generating over 1.5 billion records



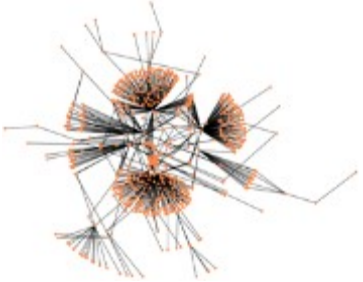
CDR Data Mining



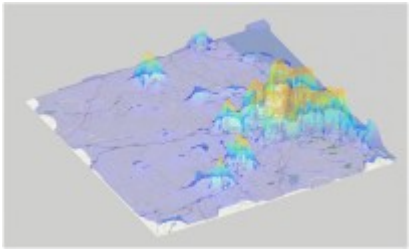
CDR Spatiotemporal Data Types



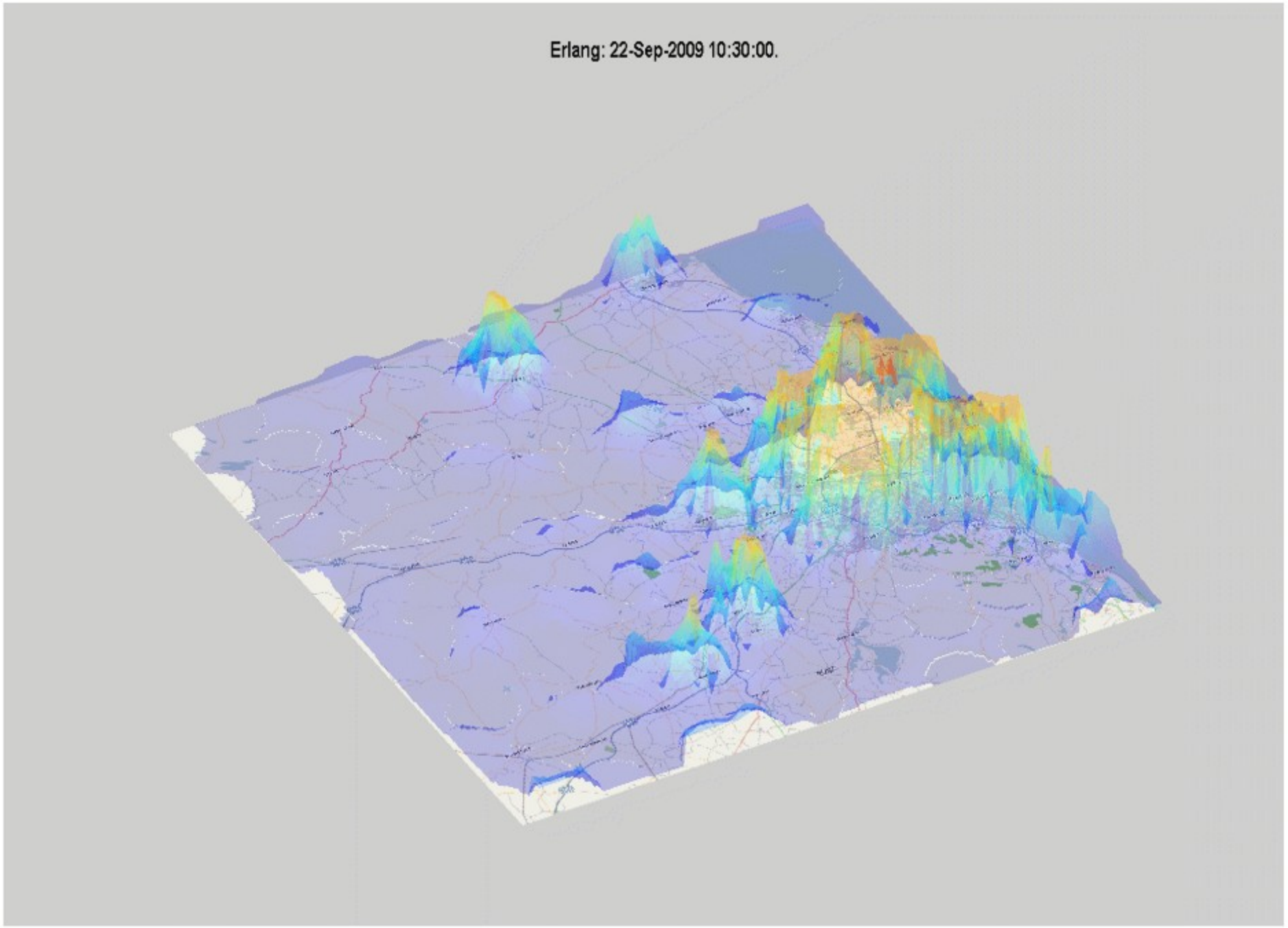
Trajectory Information



User Social / Cell Network



Cell Activities


Erlang: 22-Sep-2009 10:30:00.

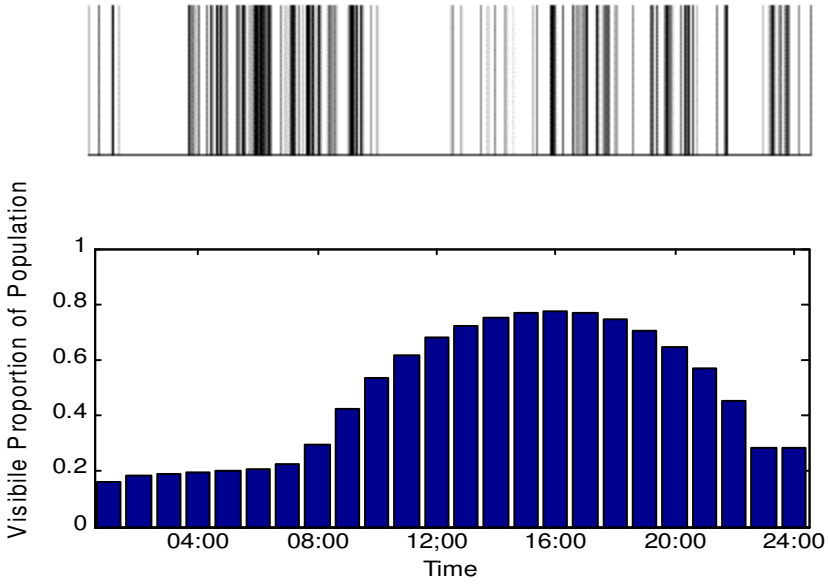
Subscriber Trajectories

Trajectories from CDR only capture cell locations of individuals when they record mobile phone activity

Trajectory Issues

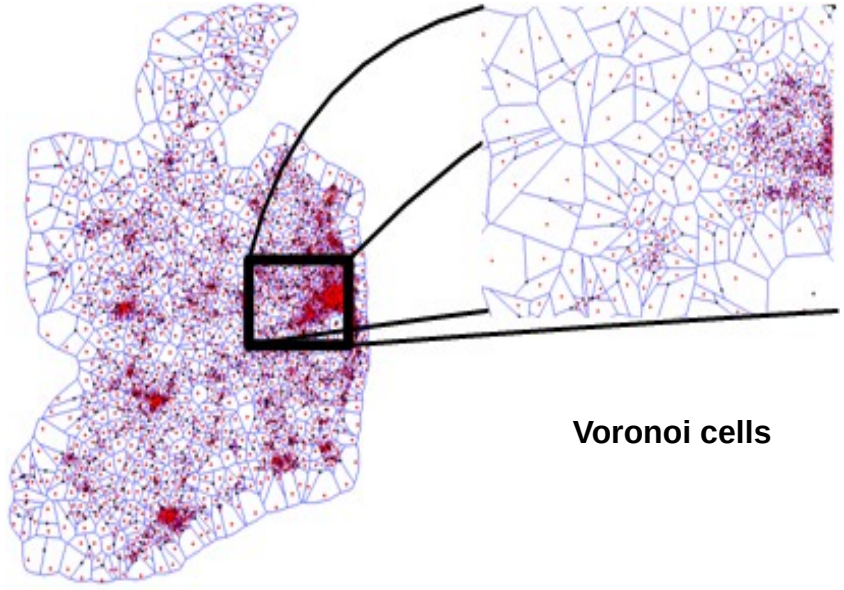
Sampling rate

- User activity follow a burst mentality



Spatial Resolution

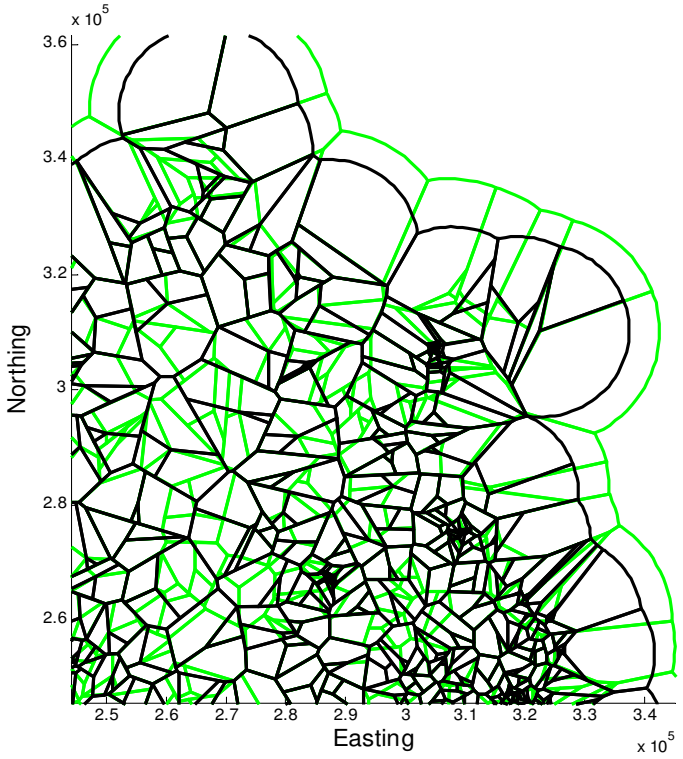
- Location estimates are fixed to cell tower coverage areas



Voronoi cells

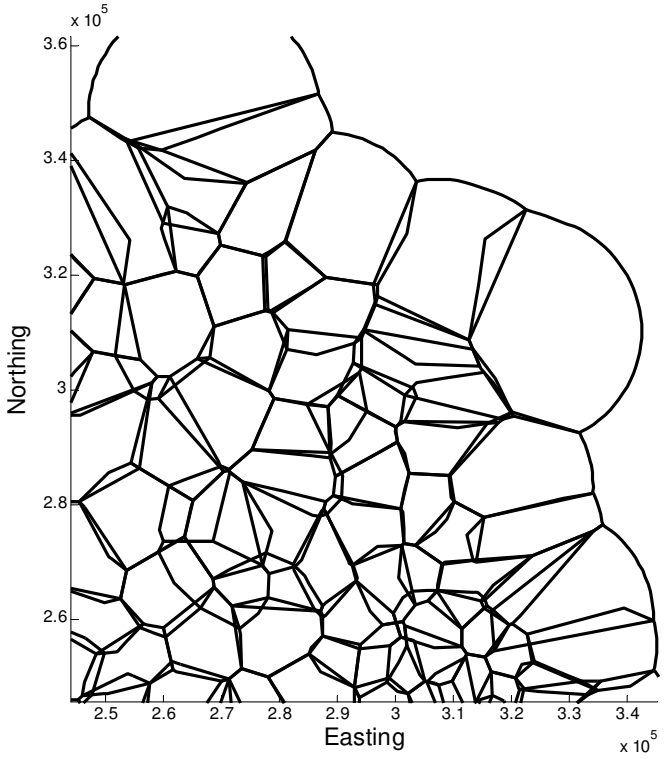
Scaling Cells to Regions

Cell Coverage



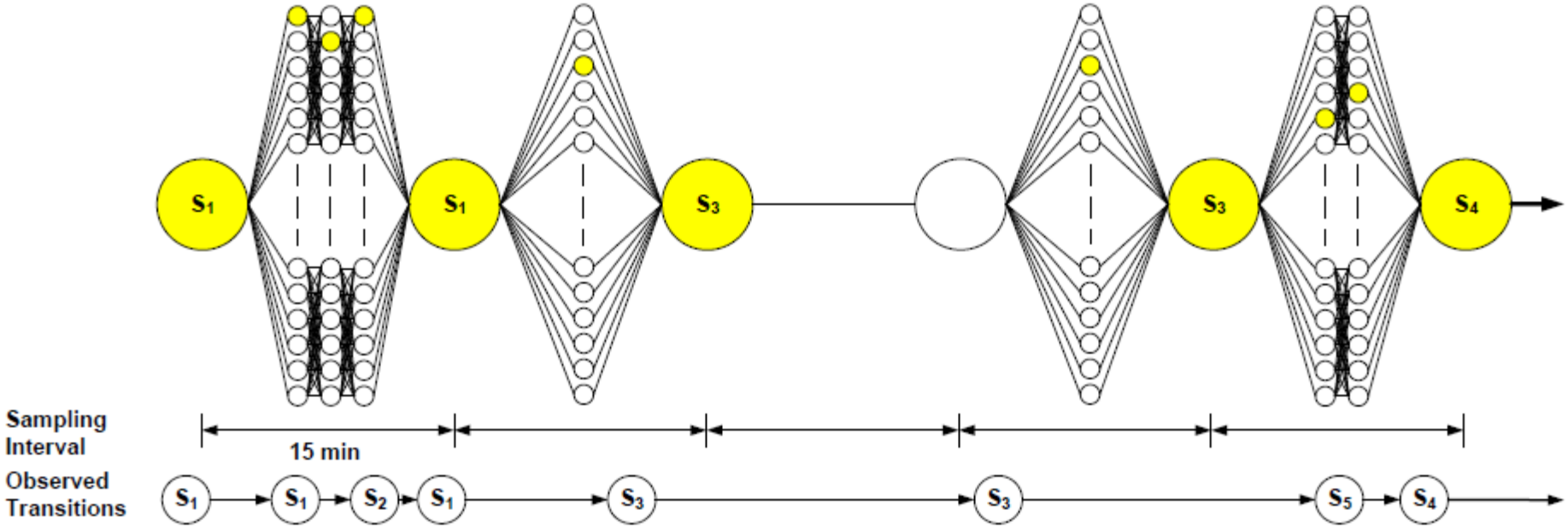
10721 cells

Spatial Regions of



500 regions

Uniform Sampling



Within each 15-minute temporal window, the estimate of location is based on the last recorded servicing cell tower recorded for that subscriber during that period.

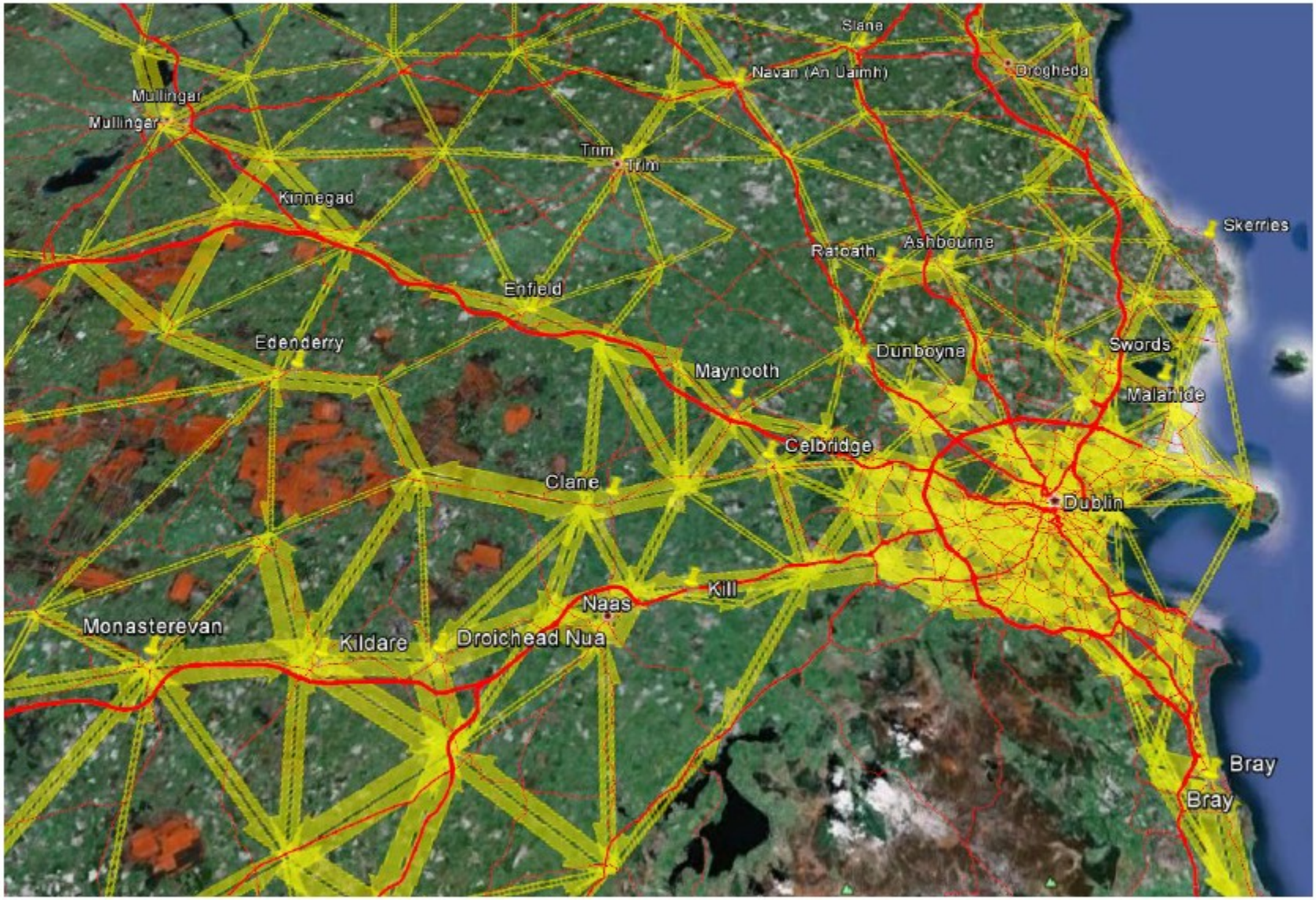
CDR trajectory state sequence sampling of the output sequence S = {S_1, S_1, S_3, S_3, S_4}. Smaller yellow circles represent actual regional transitions within a sample period and larger yellow circles represent the observed output transition sequence before resampling.

Regional Flows of Subscribers

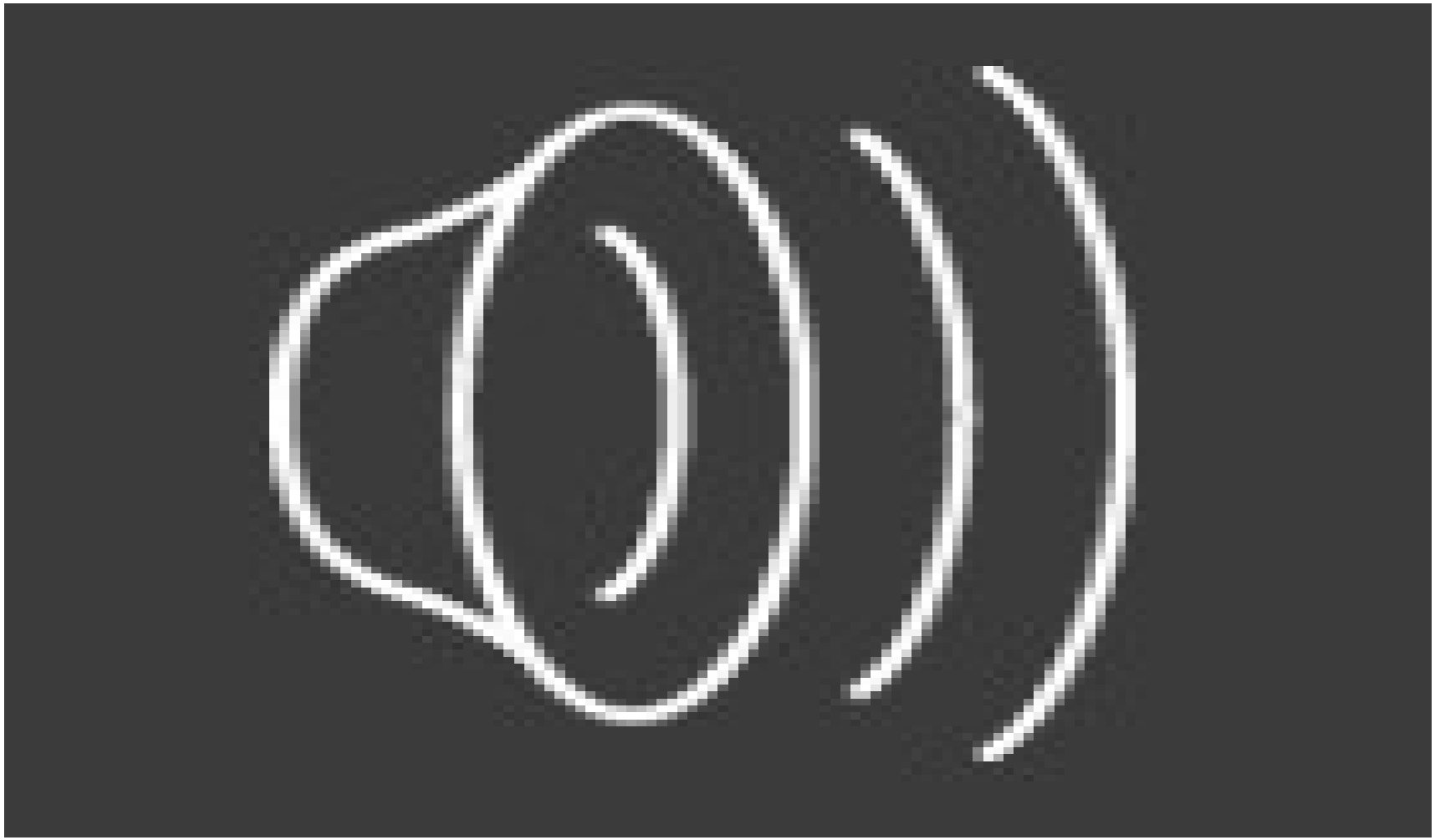
- By observing the flow of people between clustered regions and the geographical areas covered, a proxy for the flow of people between individual population centres can be established. These results can summarised in an aggregated transition matrix T(k),

$$T(k) = \begin{pmatrix} n_{1,1}(k) & n_{1,2}(k) & \cdots & n_{1,N}(k) \\ n_{2,1}(k) & n_{2,2}(k) & \cdots & n_{2,N}(k) \\ \vdots & \vdots & \ddots & \vdots \\ n_{N,1}(k) & n_{N,2}(k) & \cdots & n_{N,N}(k) \end{pmatrix}$$

Average Intensity of Subscribers Between Regions



Temporal Flow of Subscribers



Population Estimation

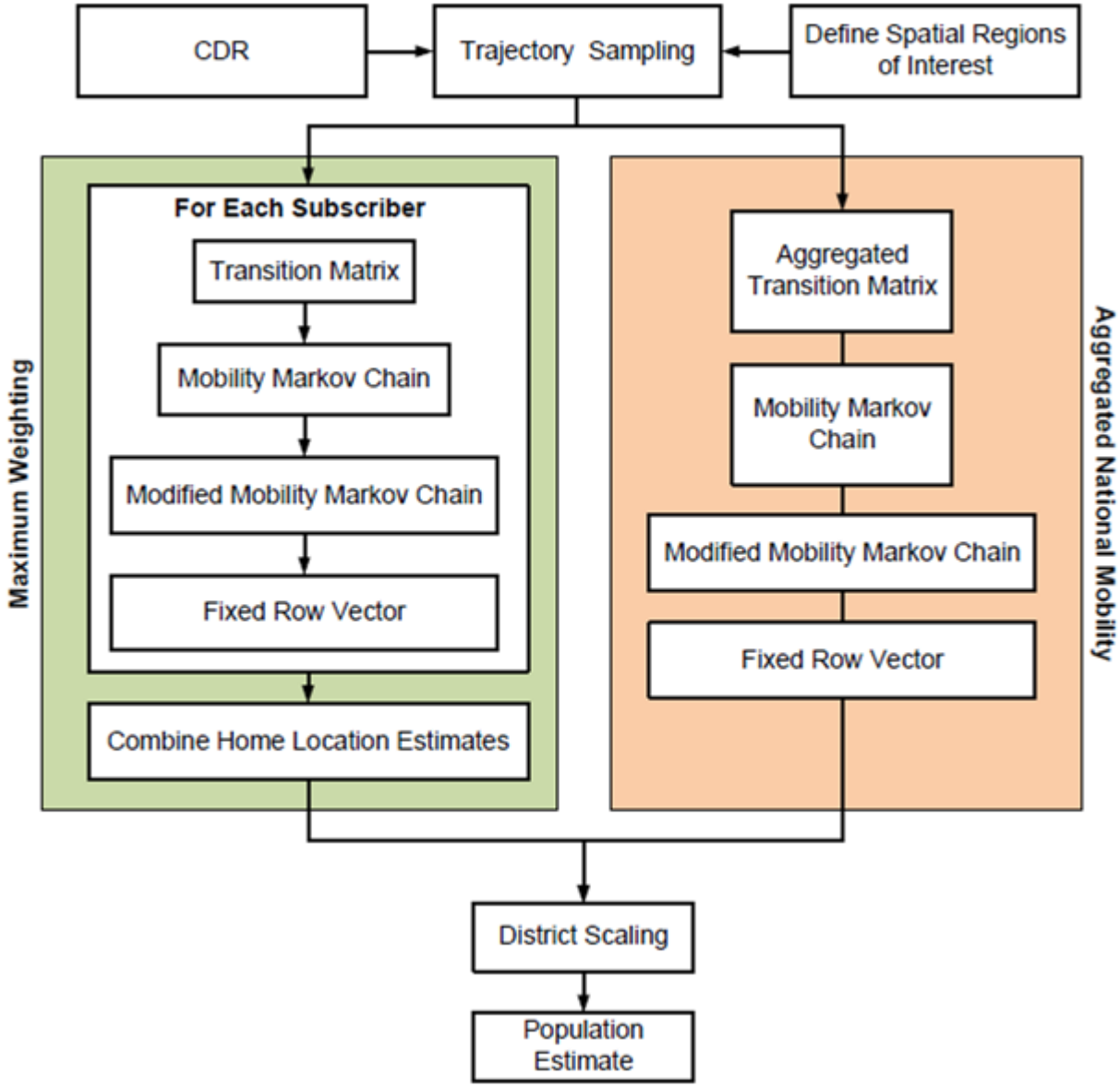
- A census is the primary tool used by national governments to gather information on population metrics, which includes among others population count, religious status, material status and household occupancy.

- The knowledge obtained dictates future policy on decisions related to the planning of future infrastructure and public services.

- While the information gathered is extremely important for the delivery of such services, the cost of carrying out a census is prohibitively expensive. As a result a census may be only carried out every 5-10 years.

- Consequently, they provide poor temporal resolution and are incapable of providing information on the current status of a population.

- This motivates the requirement for low cost alternatives.


CDR
Trajectory Sampling
Define Spatial Regions of Interest
Maximum Weighting
For Each Subscriber
Transition Matrix
Mobility Markov Chain
Modified Mobility Markov Chain
Fixed Row Vector
Combine Home Location Estimates
Aggregated National Mobility
Aggregated Transition Matrix
Mobility Markov Chain
Modified Mobility Markov Chain
Fixed Row Vector
District Scaling
Population Estimate

Modelling User Movement

- We can model individual user movement with Markov chains.
- Homogeneous Markov chains are useful when the state sequence, S(k), k = 0; 1; 2; . . . , is directly observable.
- By extracting a subscriber CDR trajectory, it is possible to directly observe an individual subscriber’s cell tower state sequence.
- Markov chains may be used to model a mobile subscribers transient movement between the symbolic locations represented by the clustered cell regions.

Subscriber Regions of Interest

- If a Markov chains is ergodic

$$W = \lim_{n \to \infty} P^n$$

where W is a matrix with identical rows *w*, and all components of w sum to 1.

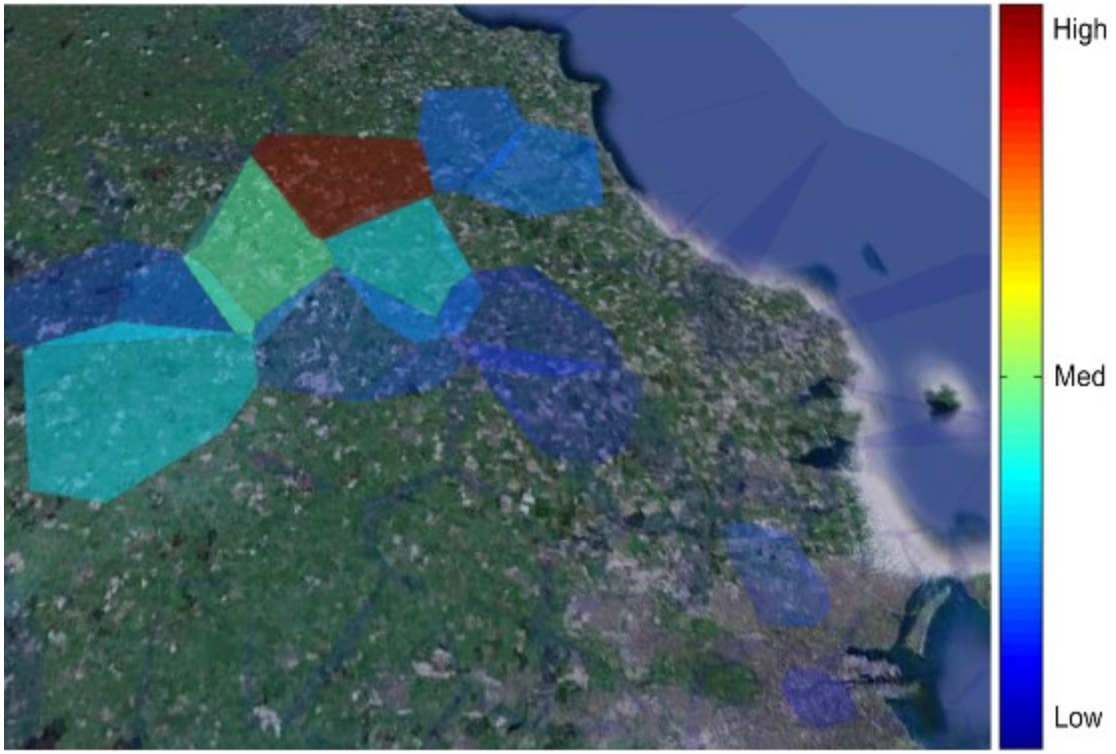
- The fixed row vector, *w*, of a mobile subscriber's mobility Markov chain conveys the probability of observing that subscriber at a region in space over a long period of time.

- As not all mobility Markov chains are ergodic, introduce a regularisation weight

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

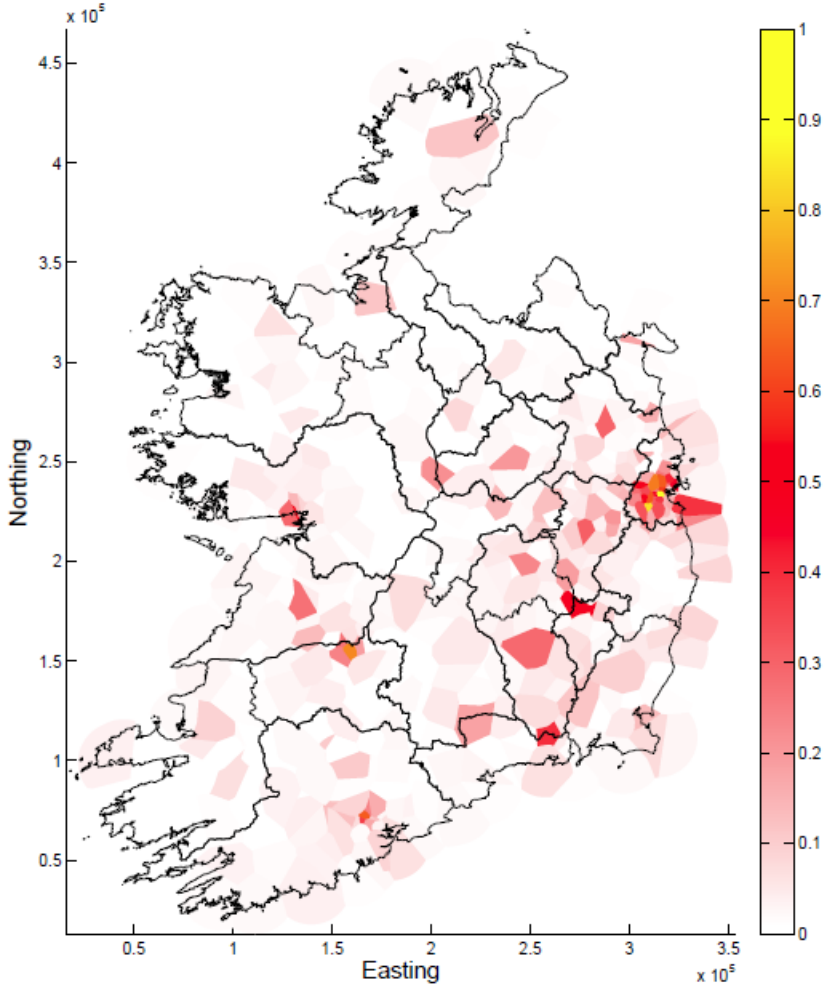
where *Q* is a modified Markov chain, *R* is the number of states, *J* is a *R x R* matrix of ones and α balances the learnt mobility patterns summarised by P with the influence of random transition probabilities introduced by the term J/R

- The *Q* of a randomly select subscriber
- Low transition probabilities are not illustrated for visual clarity

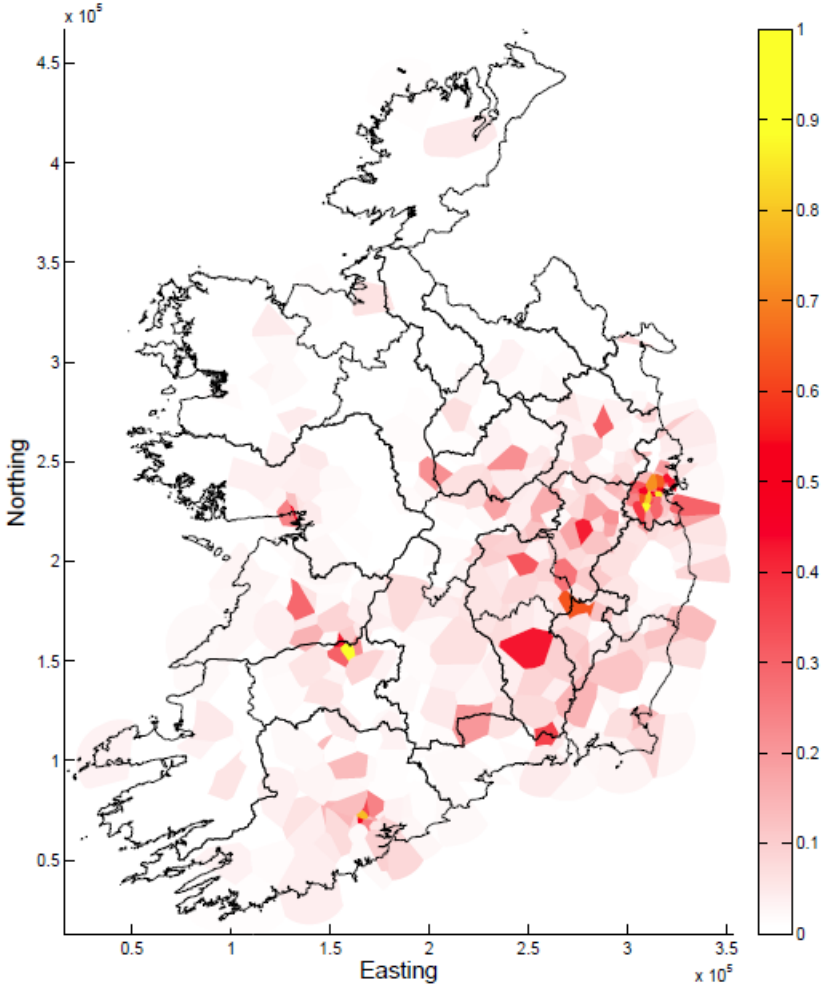


- The observed regional ranking suggests that the subscriber tends to travel in County Meath, with occasional trips into Dublin City

Population Density

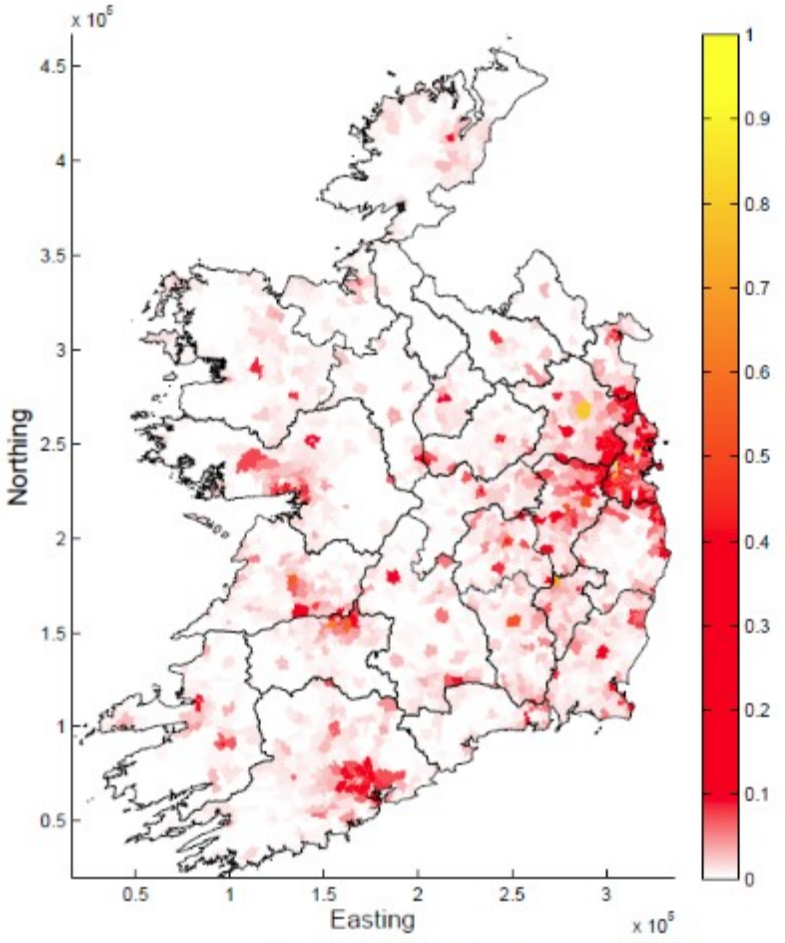


(a) Maximum weighting



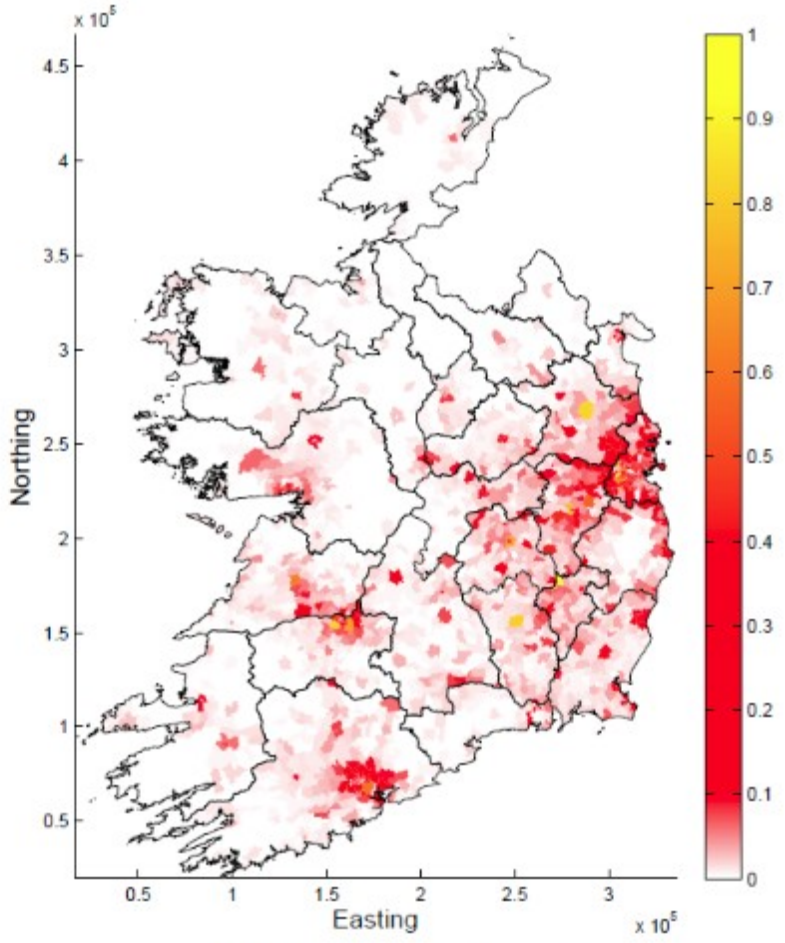
(b) National mobility

ED Population



(a) Maximum weighting

Corr – 86.61%



(b) National mobility model

Corr – 84.38%

Population Estimation

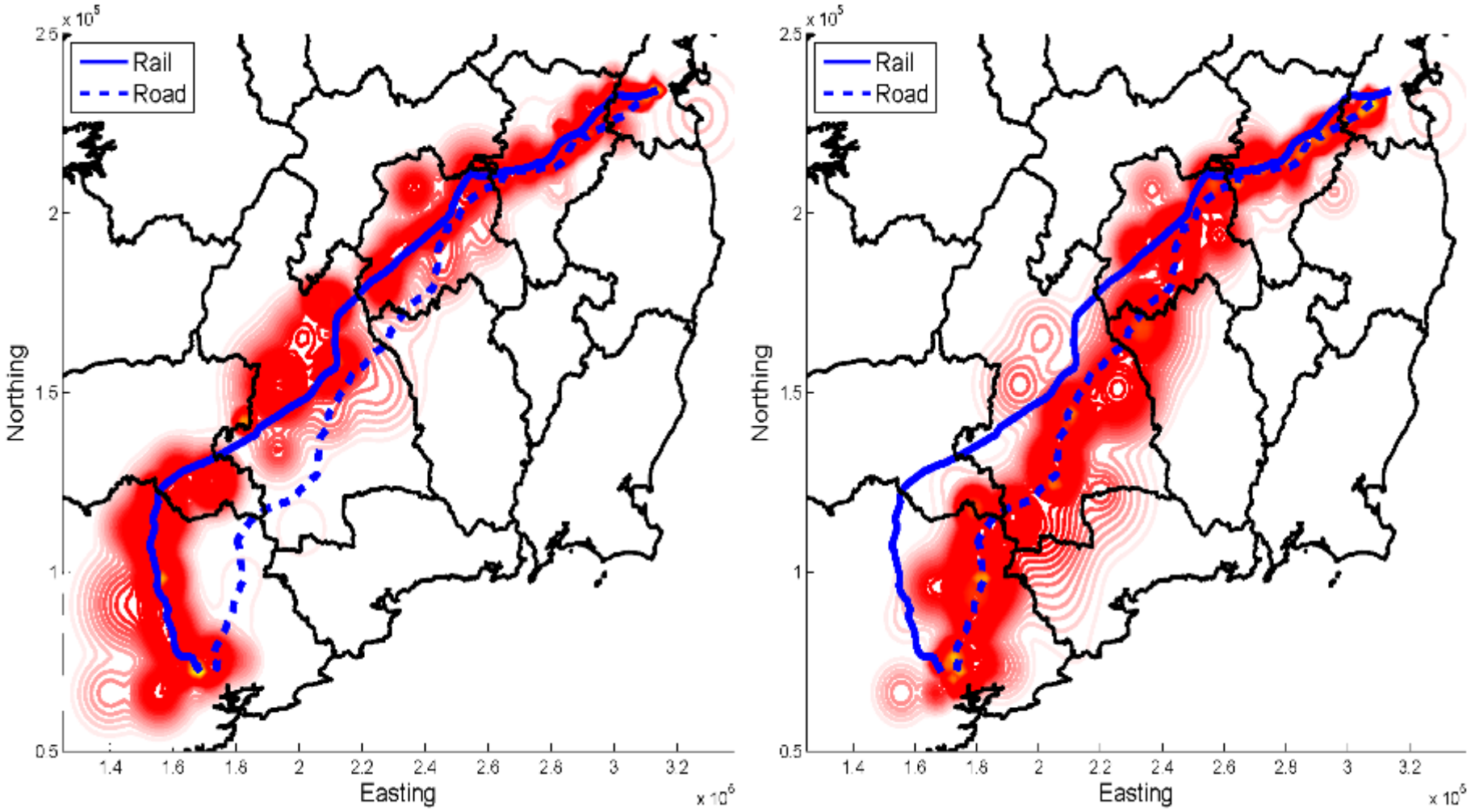
- The correlation between census data and maximum weighting approach is approximately 98.4%.
- The correlation between census data and aggregated approach is approximately 97.7%.
- However, as performance is restricted by its ability to measure population proportions in different areas, but not the ability to estimate counts, the effectiveness of such techniques for inferring census type data needs further research and is the subject of future work.

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**
MSE Excluding Dublin	**0**	**1.0491**	**2.0832**

Application Areas

- Mobile network operators are beginning to see profit margins fall due to
  - tighter regulation
  - increasing demand for data services
  - falling revenues generated from call and SMS traffic
- In this context, network operators are increasingly focusing their efforts on
  - new revenues generation schemes
  - lower subscriber churn
  - increasing customer satisfaction rates
- However, this shift in focus has unearthed significant gaps in their knowledge of how subscribers use and perceive the mobile services on offer to them.

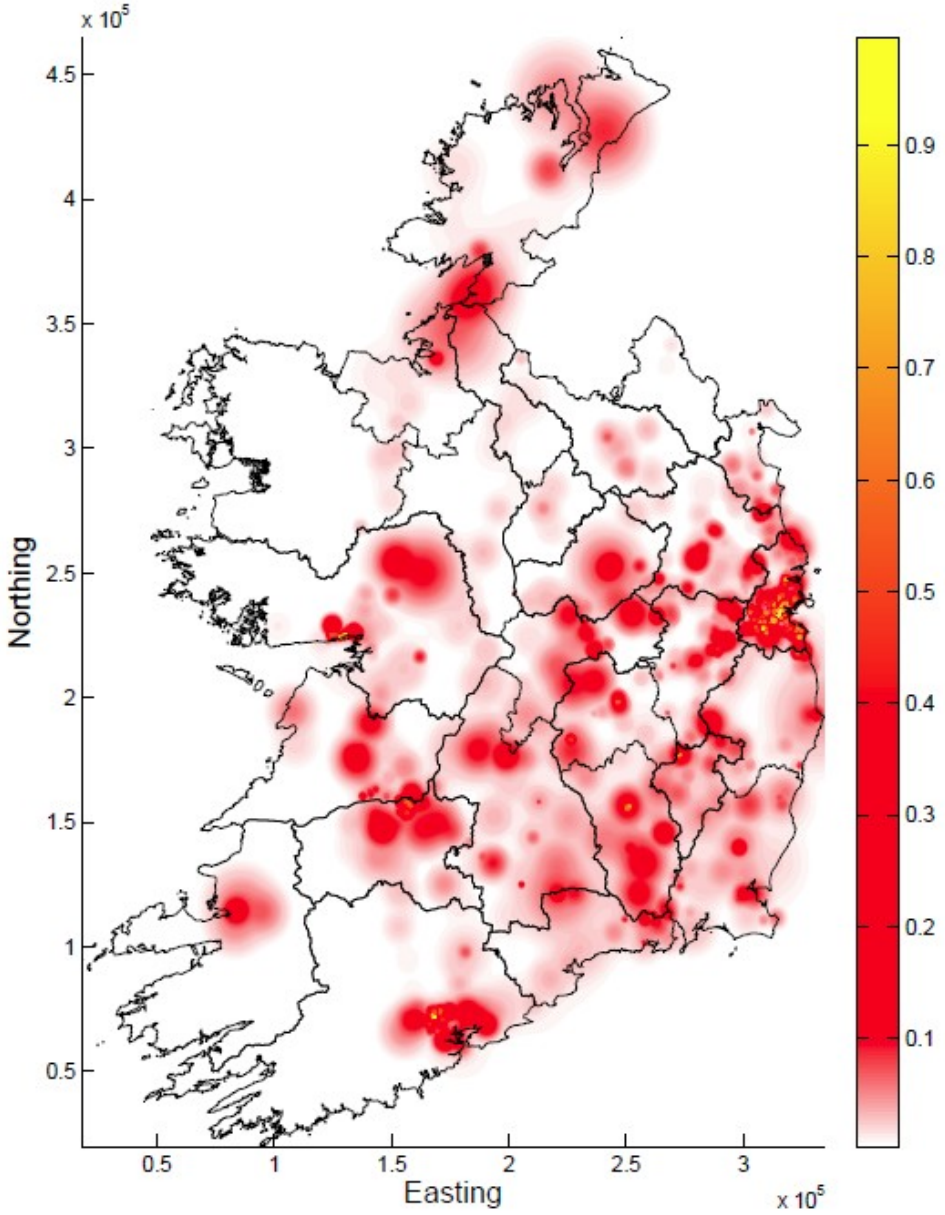
Transportation Planning



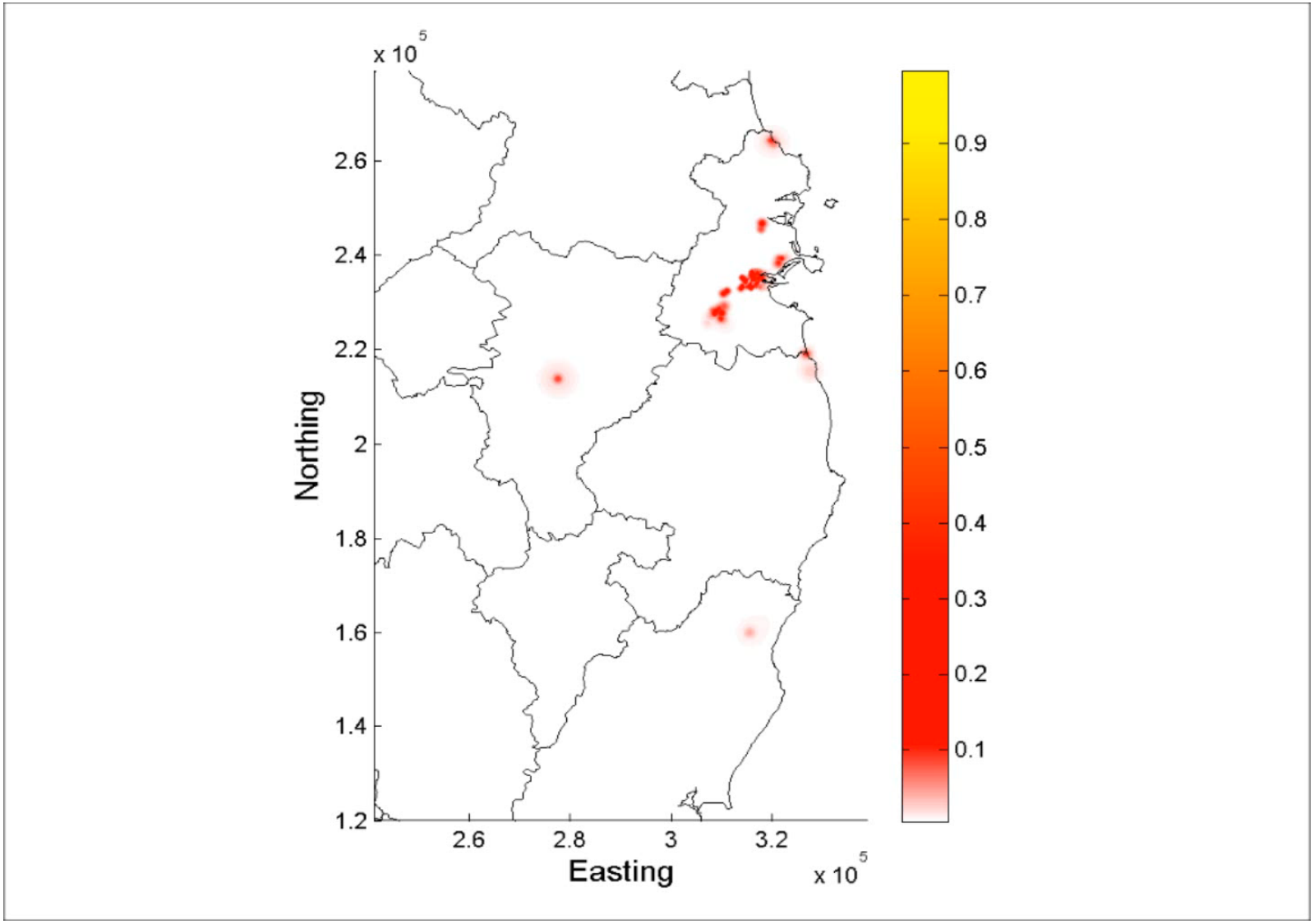
Kernel density estimate of journey trajectories identified as travelling along (a) road and (b) rail travel paths.

High Mobile Traffic Regions Of Interest

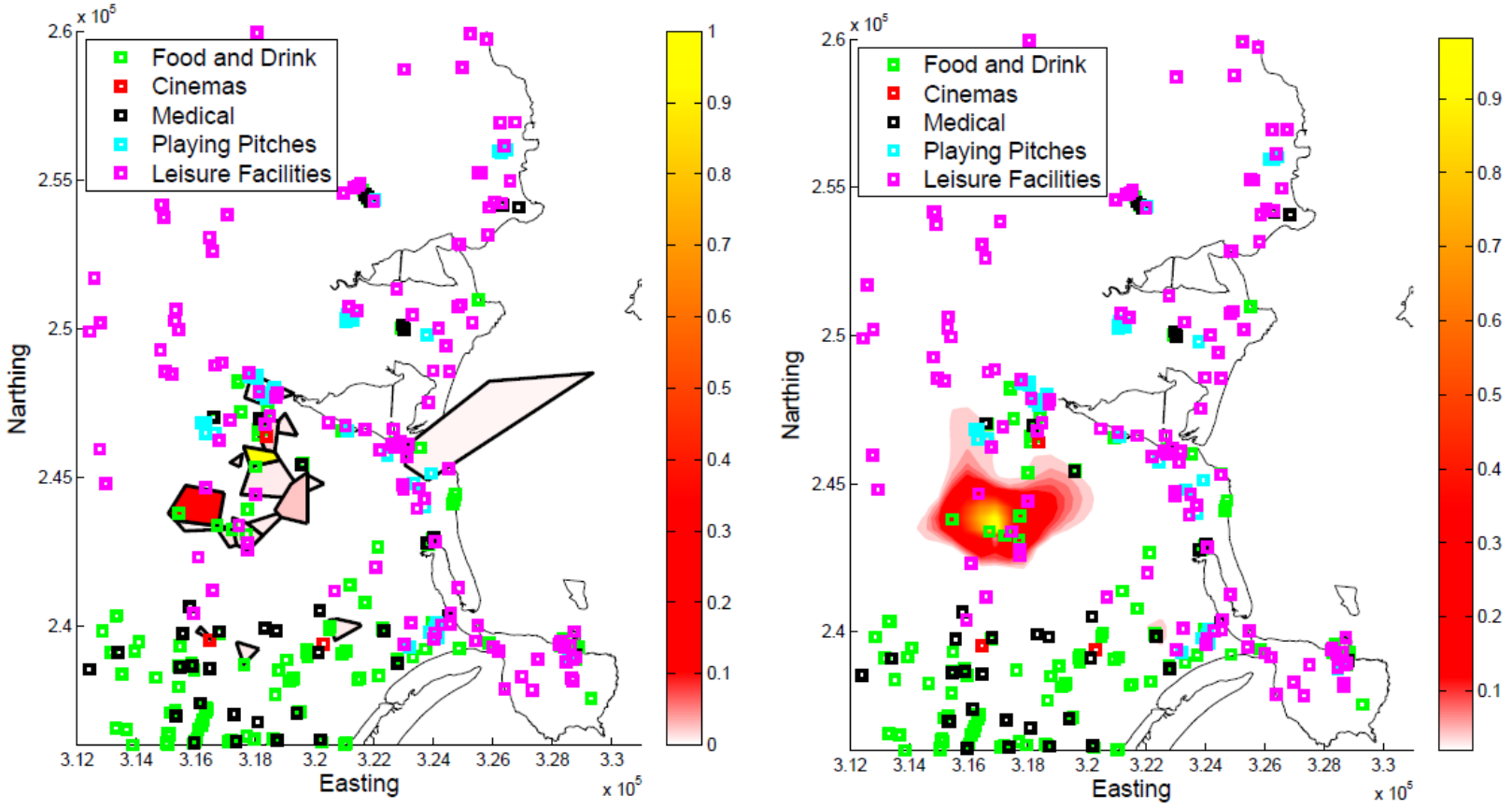
- Combine the vector weights of high data usage subscribers
- Better understanding of the areas they occupy on a daily basis
- Design more efficient networks
- Identify coverage black spots
- Better data for marketing



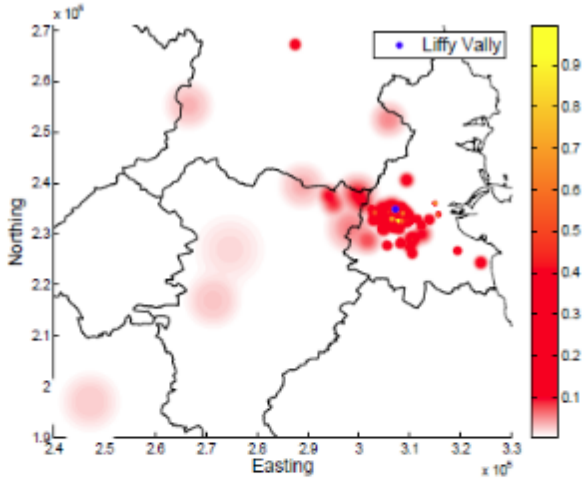
Identify Event Mobility



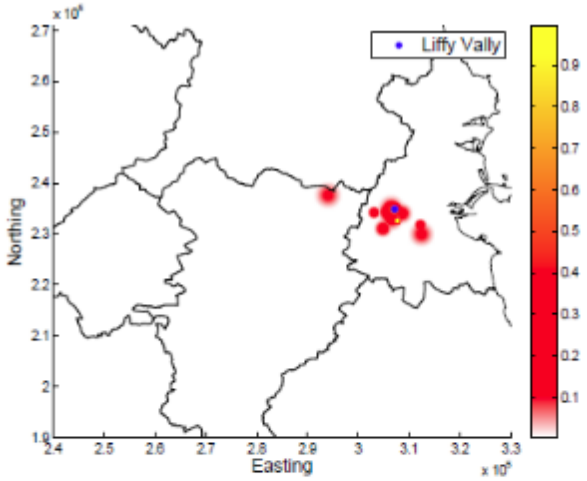
Geographically Weighted Amenities



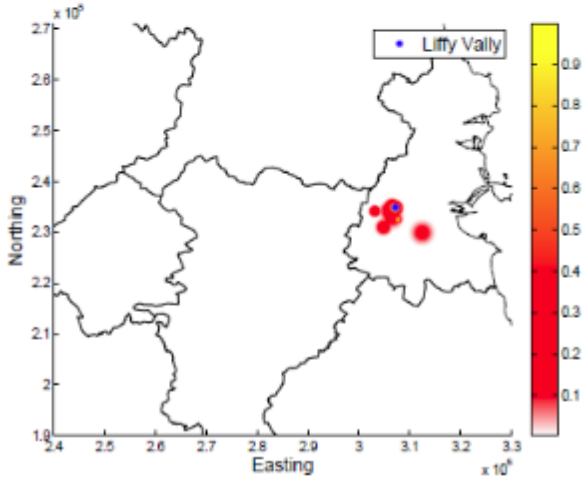
Catchment Area



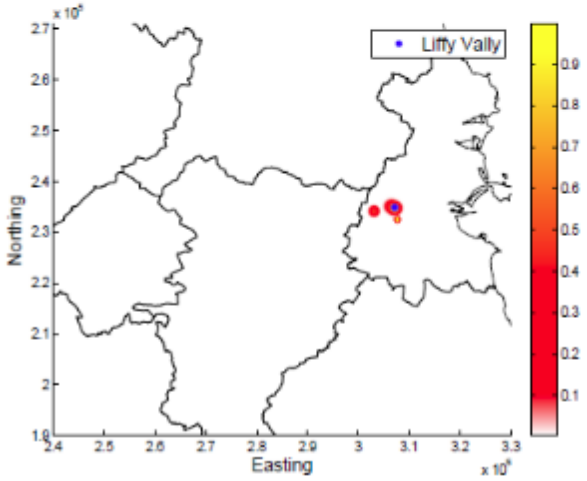
(a) ≥ 1%



(b) ≥ 5%



(c) ≥ 10%



(d) ≥ 20%

Acknowledgements

This research 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 partnership with ESRI Ireland.

I would also like to gratefully acknowledge the support of Meteor for providing the data used in this research, in particular John Bathe and Adrian Whitwham.

Questions?