Big Data Analytics

FOSCA GIANNOTTI AND LUCA PAPPALARDO

HTTP://DIDAWIKI.DI.UNIPI.IT/DOKU.PHP/BIGDATAANALYTICS/BDA/

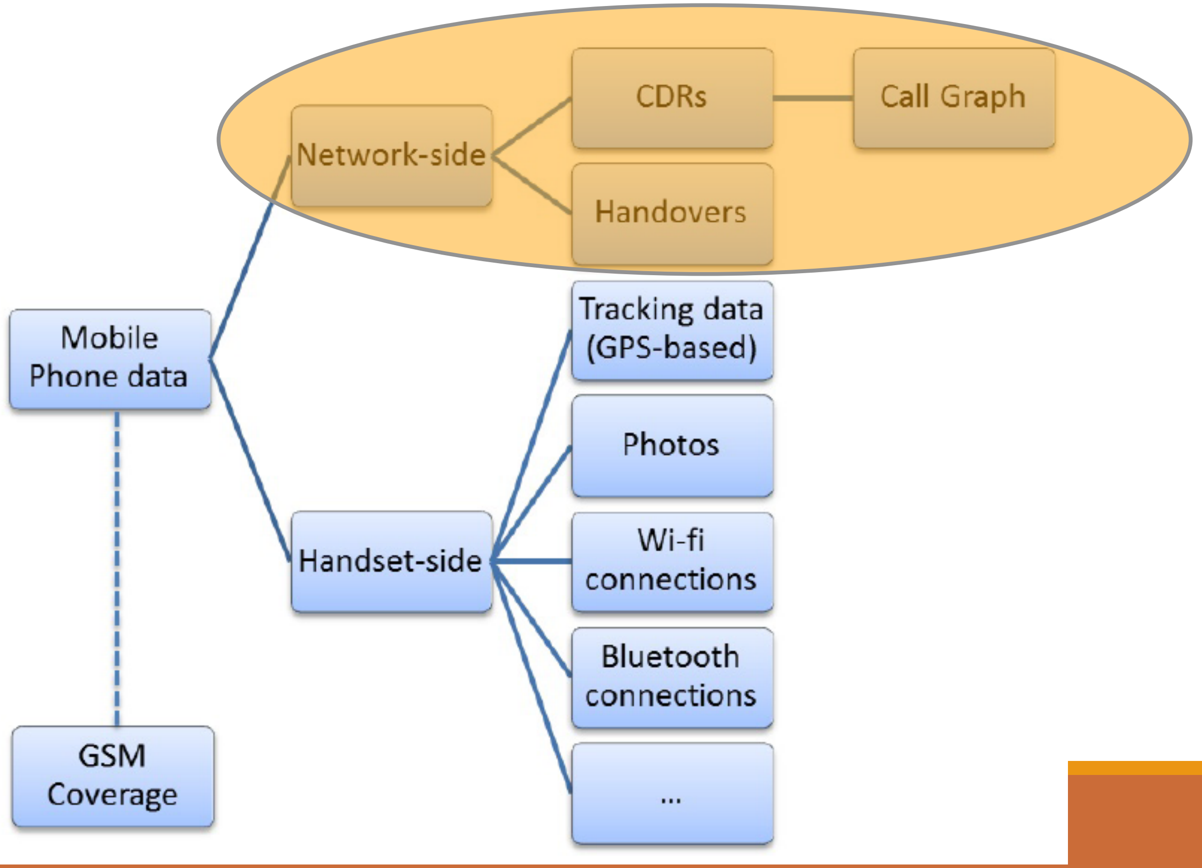
DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2018/2019

Mobility Data Mining

CITY DYNAMICS WITH GSM DATA

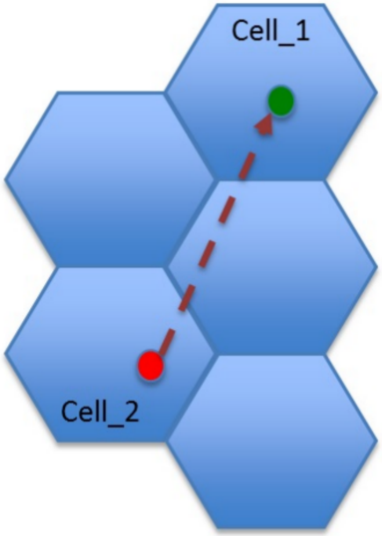
What are GSM data

- Most popular resource for mobile phone data
- In principle, several kinds of data



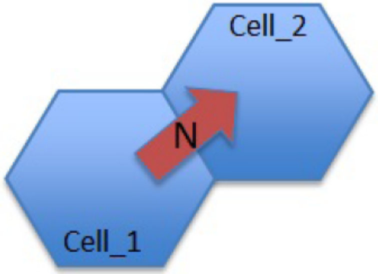
GSM data types

CDR
Who calls, **where** and **when**



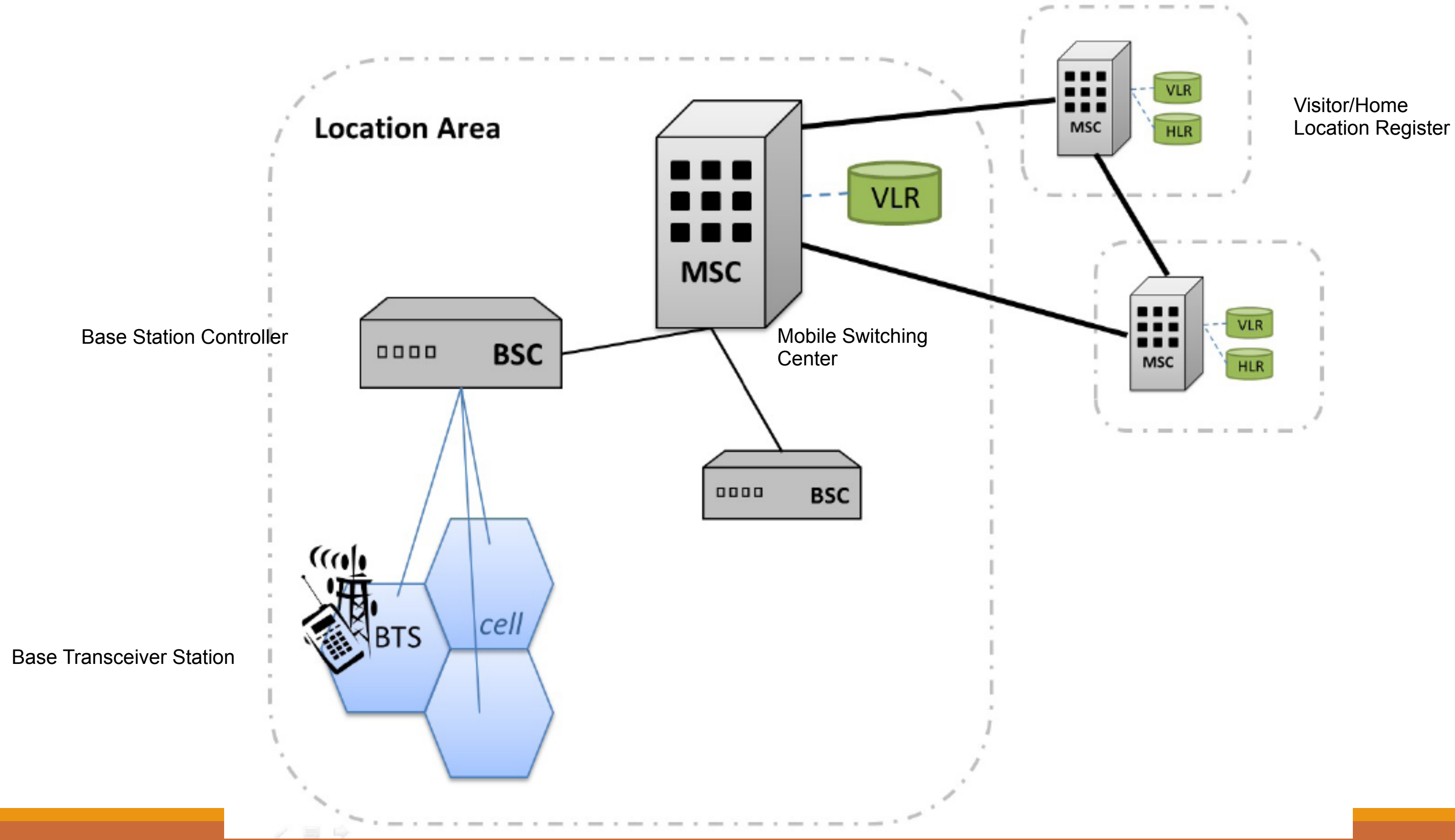
Call Graph
Who calls **whom** and **when**

Hand over
Inter-cell flow **counts**



GSM infrastructure

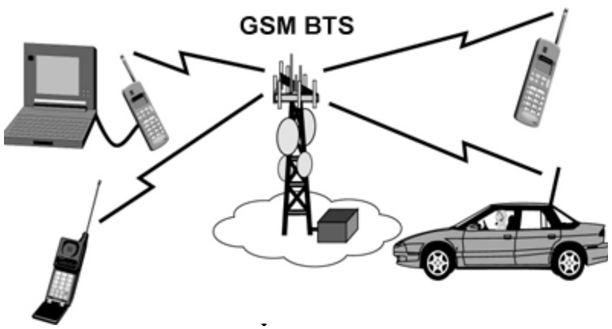
- Aimed at providing voice/data telecom.

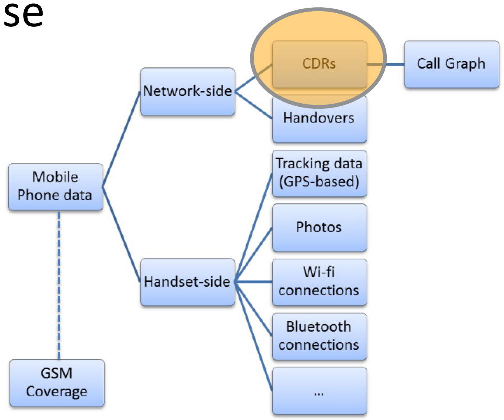


GSM data - Description

Call Data Record (CDR)

Data gathered from mobile phone operator for billing purpose



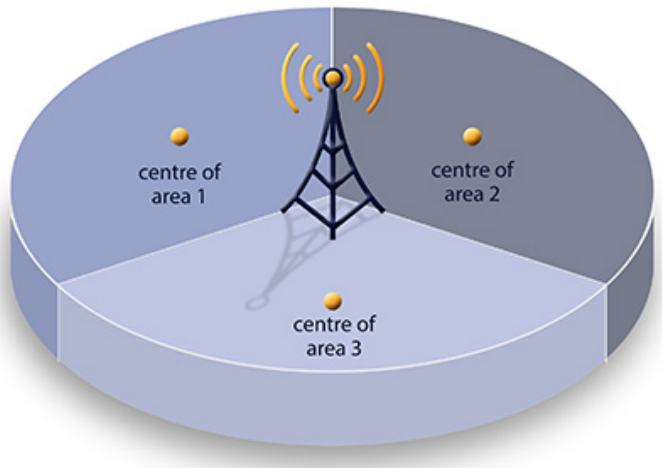


User id	Time start	Cell start	Cell end	Duration
10294595	"2014-02-20 14:24:58"	"PI010U2"	"PI010U1"	48
10294595	"2014-02-20 18:50:22"	"PI002G1"	"PI010U2"	78
10294595	"2014-02-21 09:19:51"	"PI080G1"	"PI016G1"	357

GSM data - Description

- Distinction between antenna and tower
  - Usually one “tower” carries 3 directional antennas
- Which one is in the data depends...

cell tower with 3 cells, each with 120° angle



Pros and cons of using GSM data

Pros

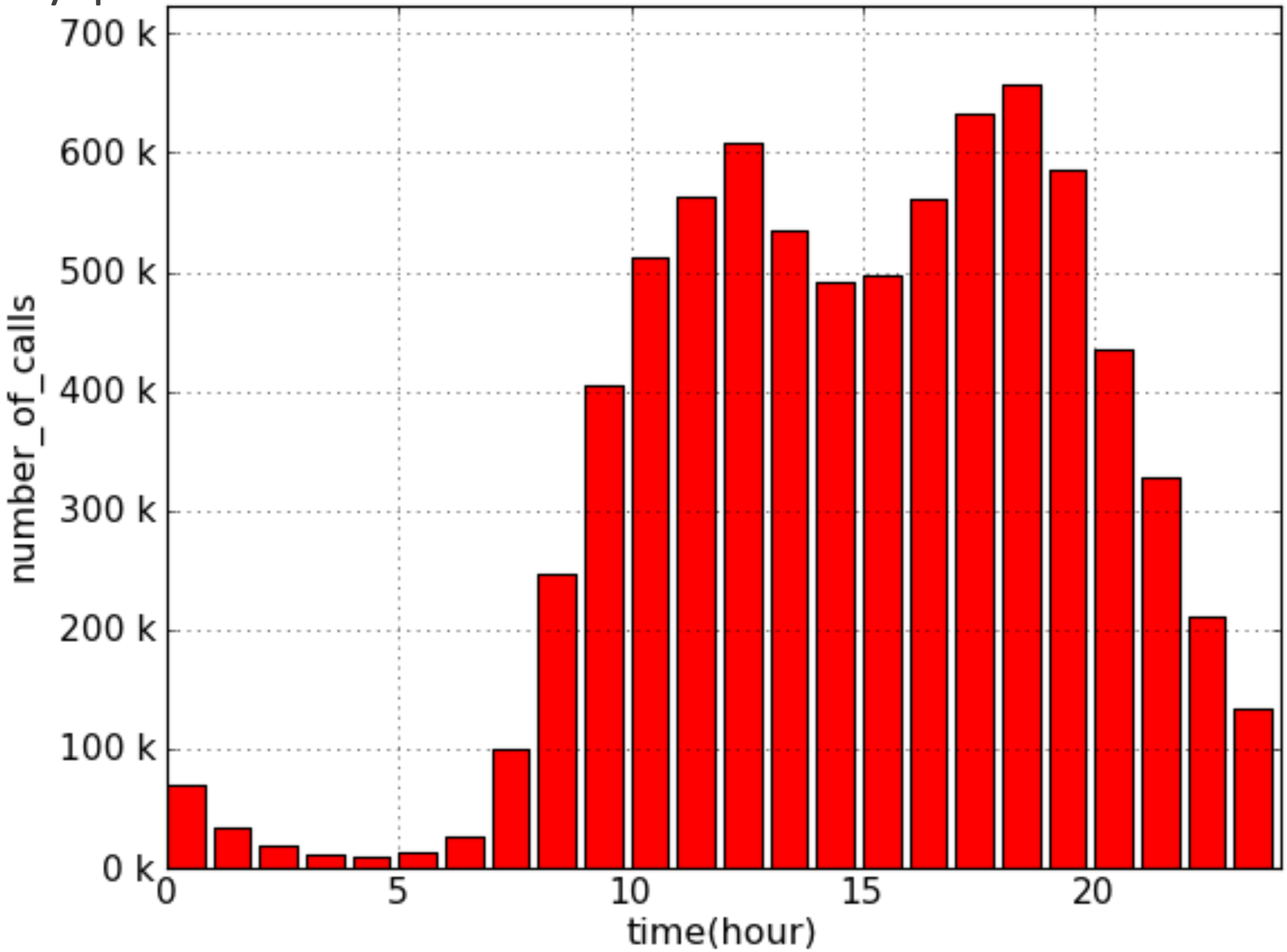
- Passive sensing: does not require an active contribution of the users
- Contains huge amount of information of how, when, with whom we communicate
- Same data format in all the world

Cons

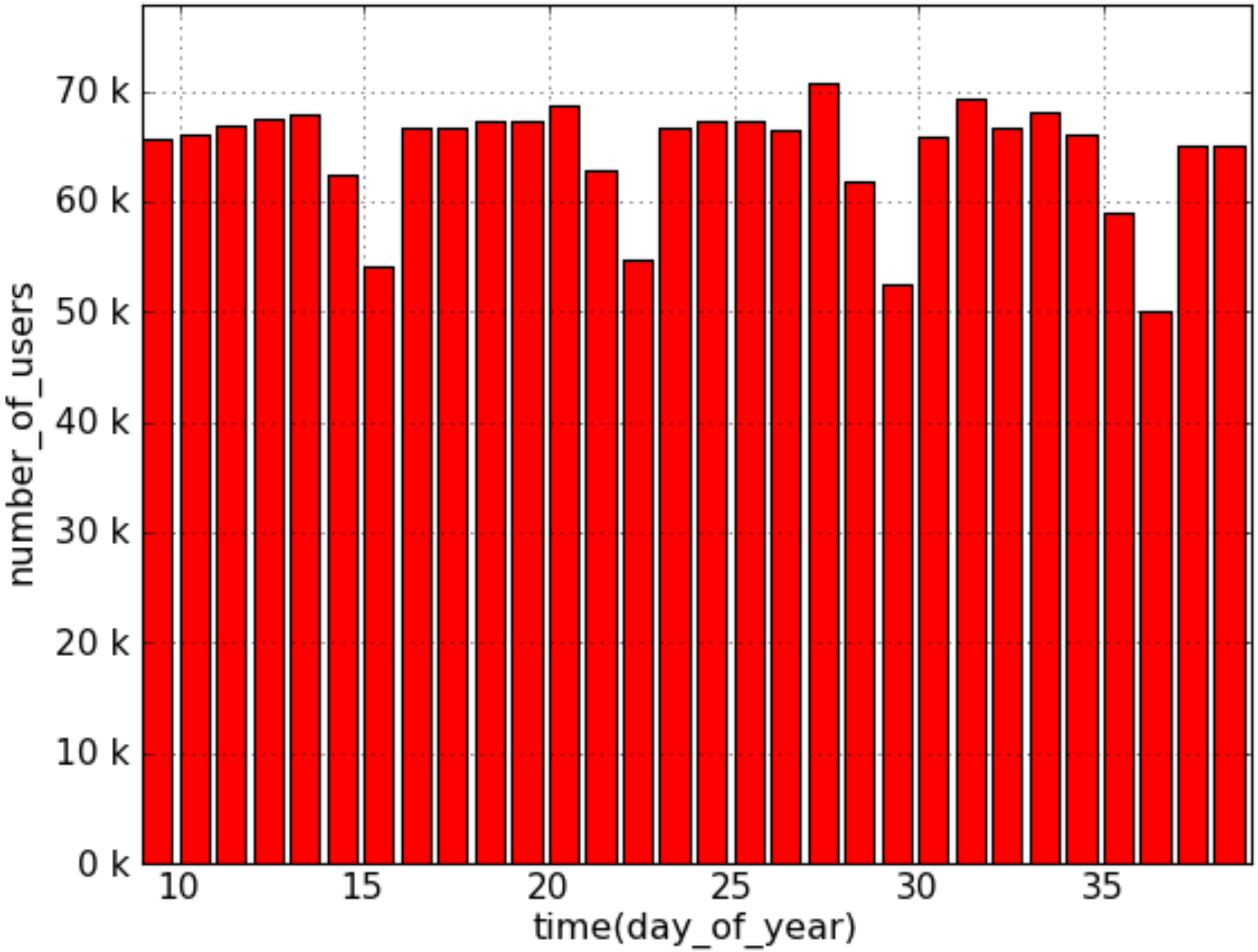
- Poor demographic and economic data
- Privacy concern: different legislations for different countries
- Low sampling: few events of calls for a considerable amount of users

Simple CDR-based statistics

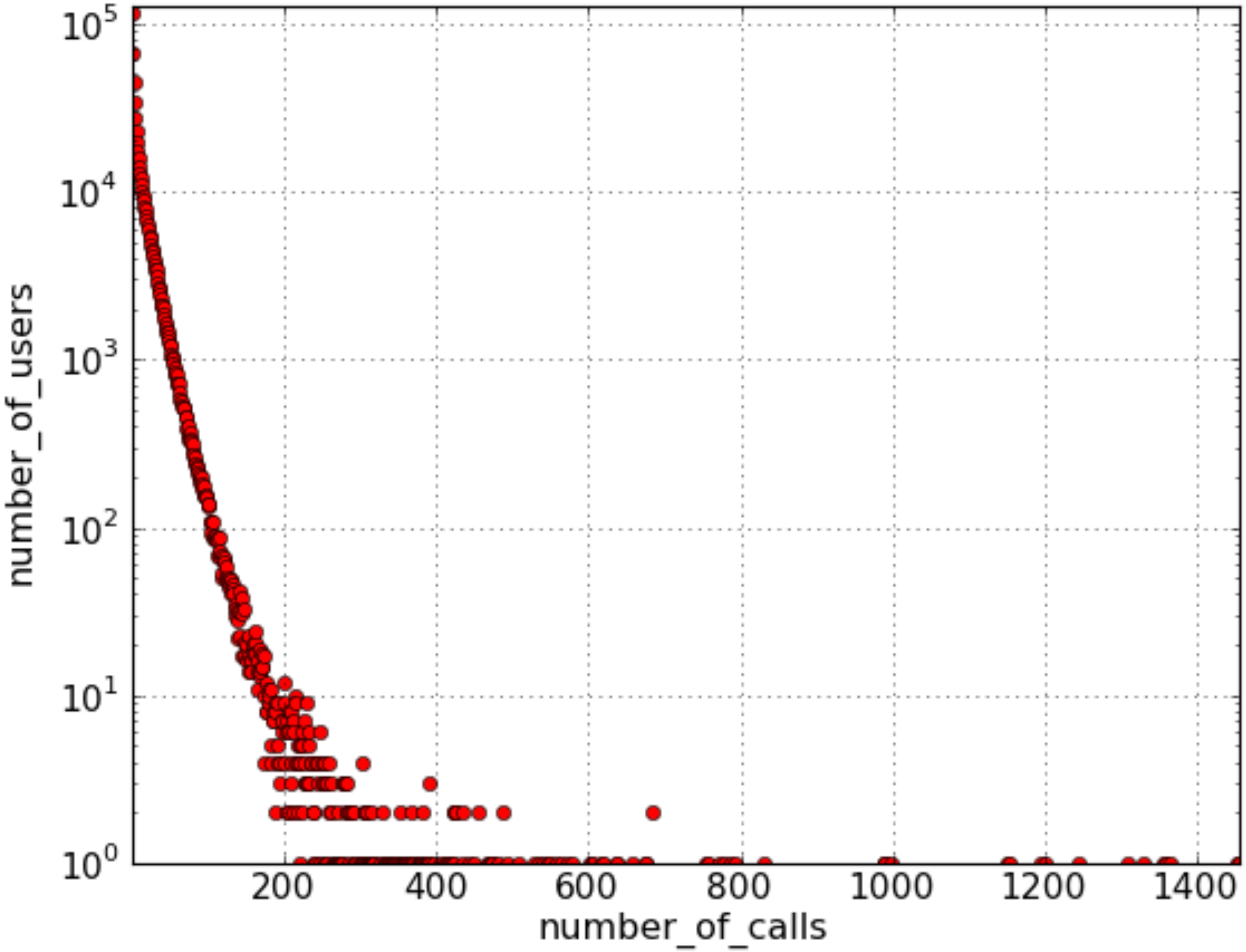
Daily pattern behavior



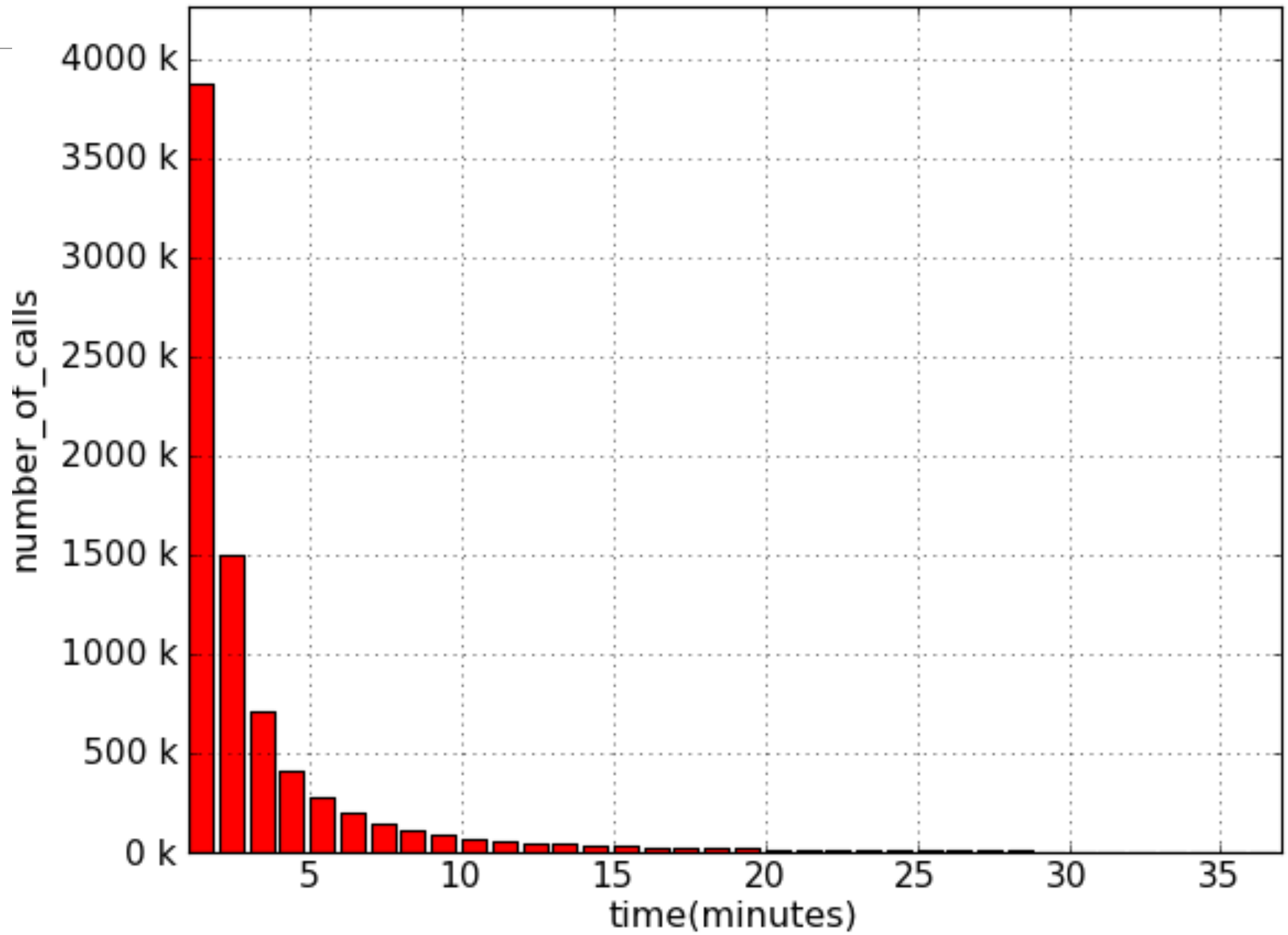
Weekly pattern behavior



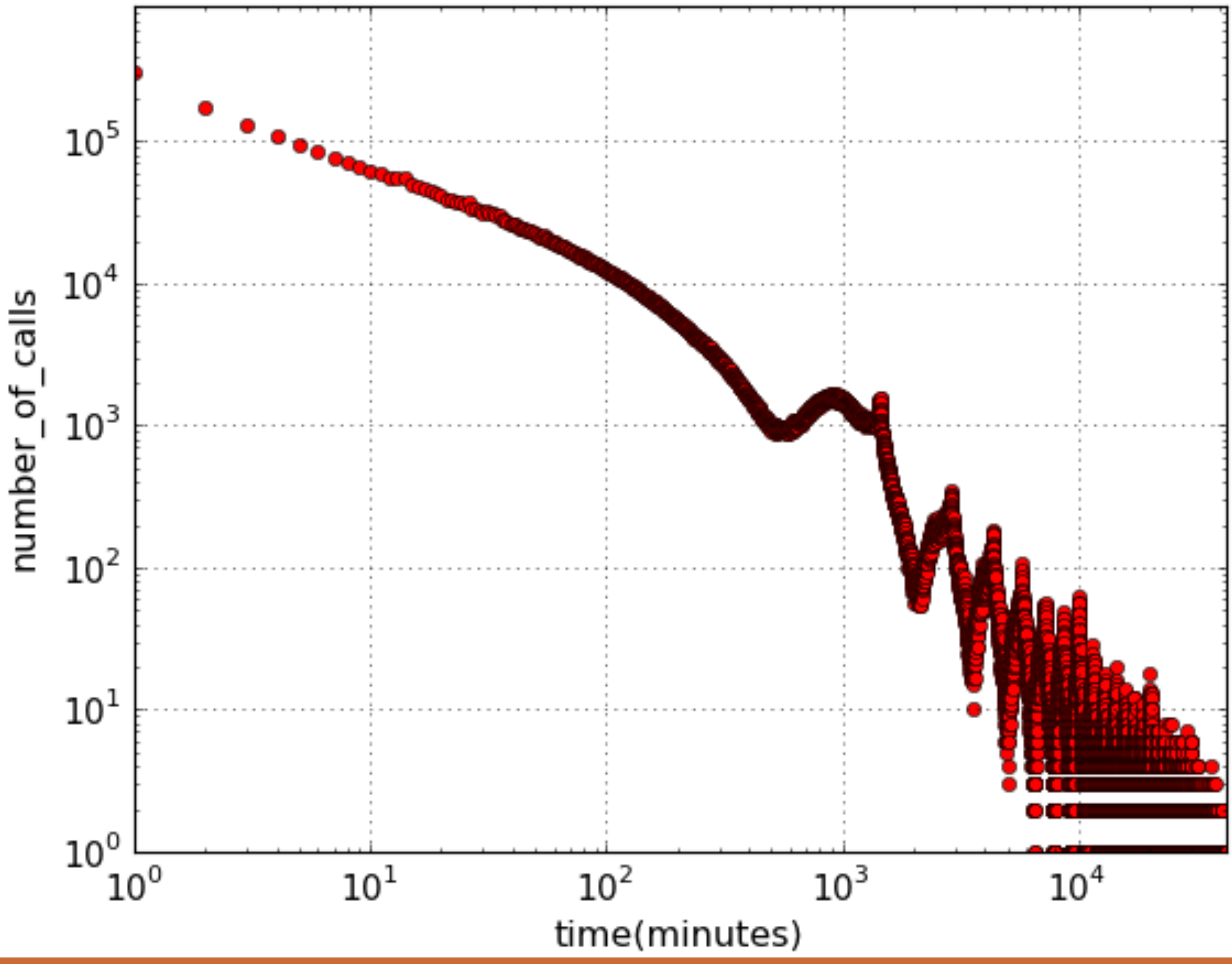
How many times we call?



How long we talk on the phone?



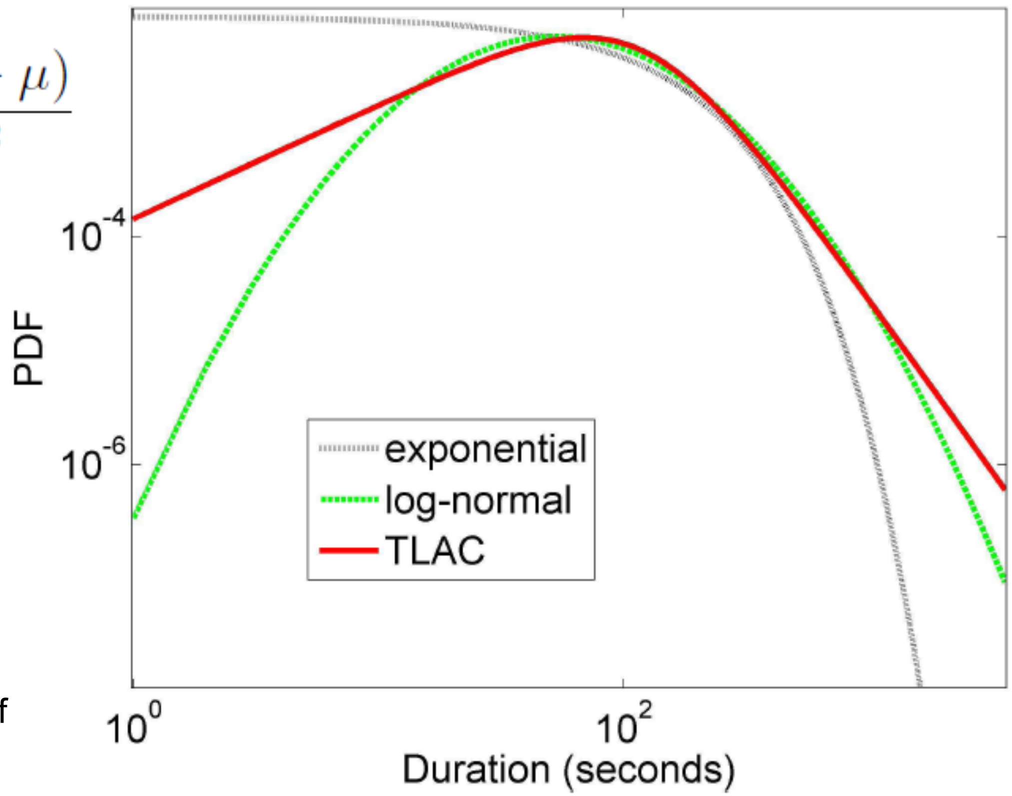
How many minutes goes by a call to the next?



Theoretical model of call durations

- Truncated Lazy Contractor (TLAC)

$$PDF_{TLAC}(x) = \frac{\exp(z(1+\sigma)-\mu)}{(\sigma(1+e^z))^2}$$

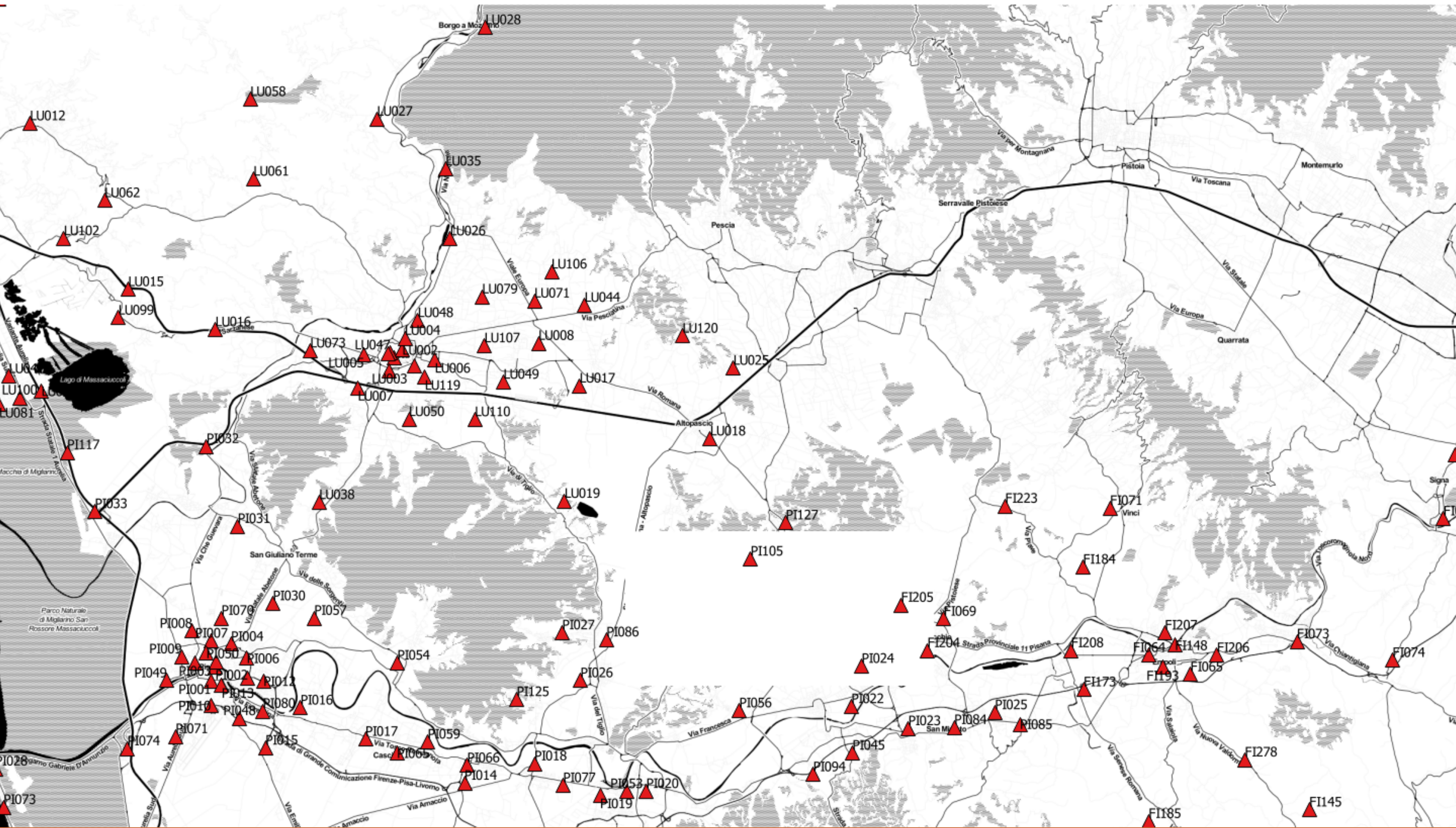


de Melo-Akoglu-Faloutsos-Loureiro.
Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. ECML PKDD 2010.

Join the **spatial** part of the mobile phone data

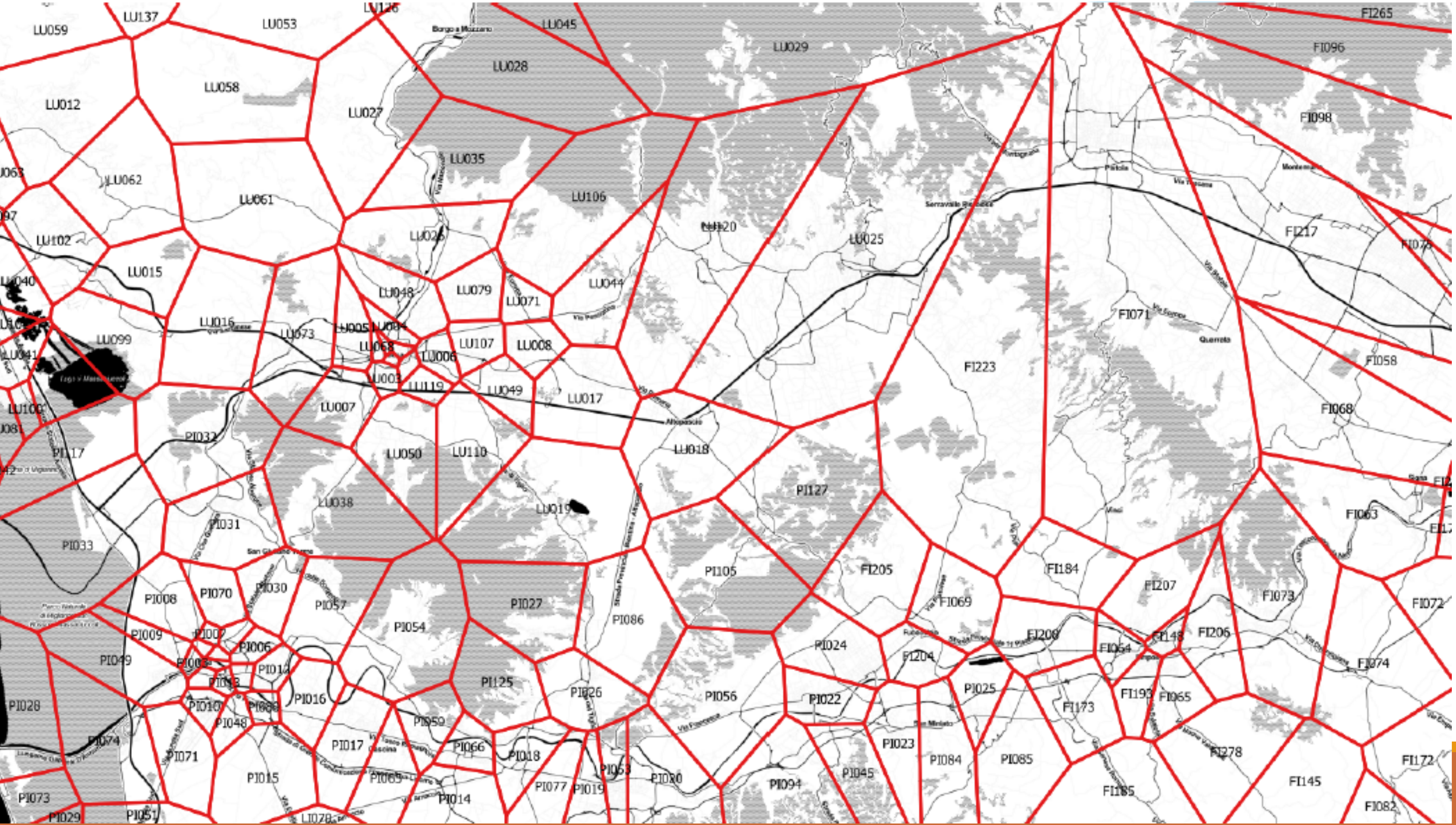
From CDR to Geography:
CDRs describe where the calls started

Antennas

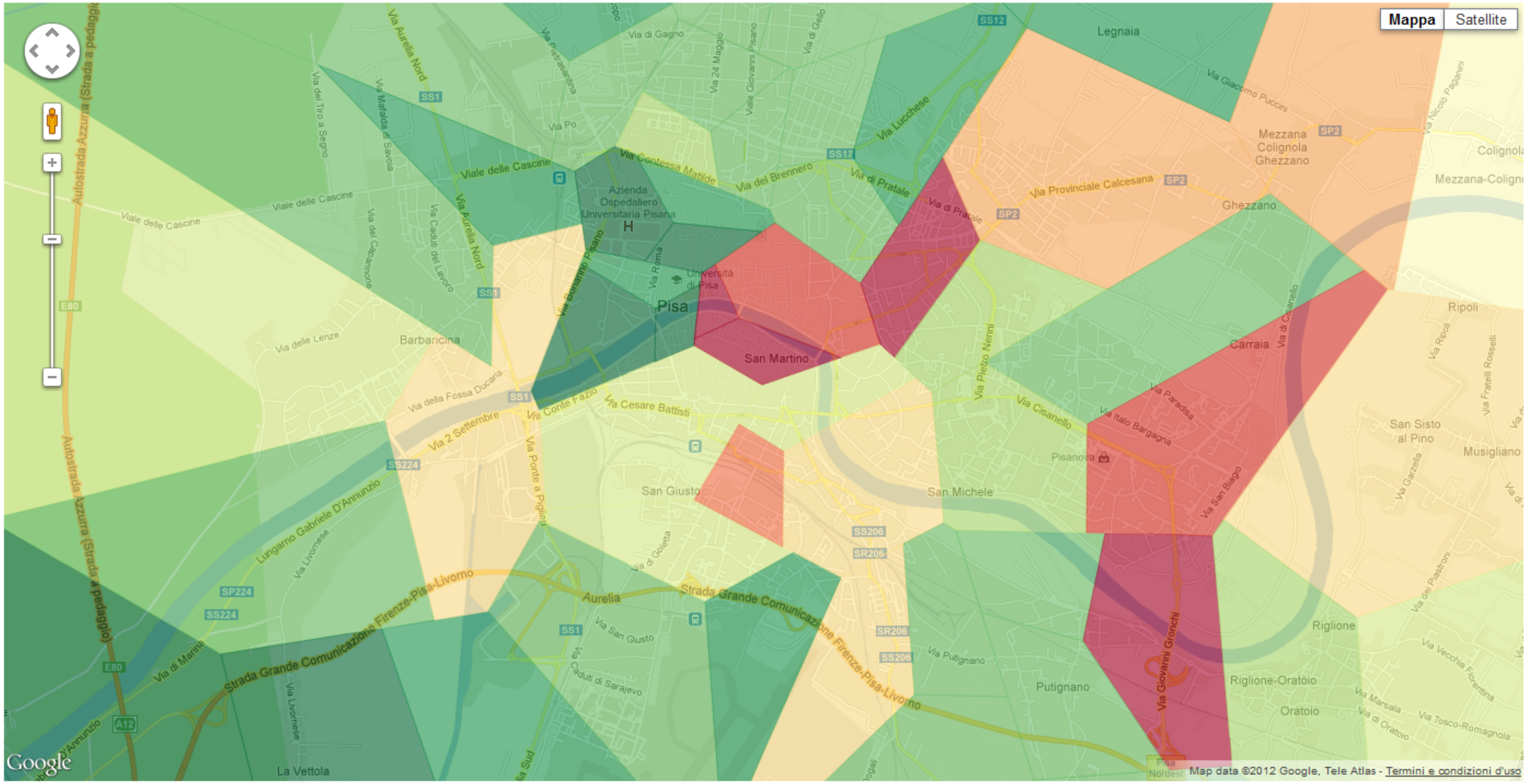


From CDR to Geography:
CDRs describe where the calls started

Voronoi tesselation



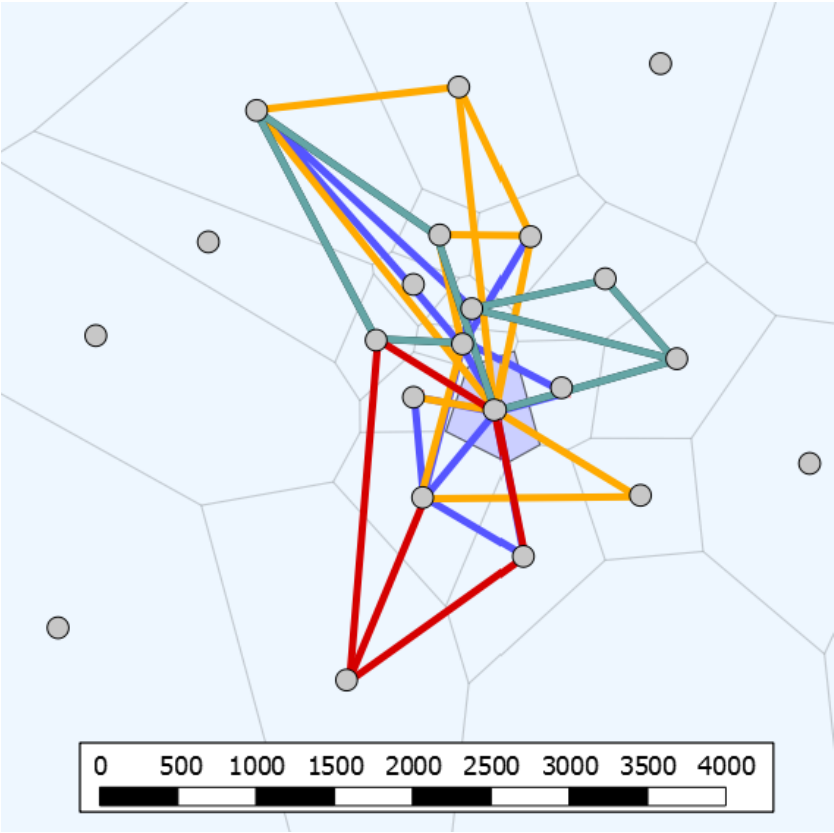
Spatial distribution of calls



ple within the working area of Pisa

Observing the **mobility** of individuals

Mobility Behaviours



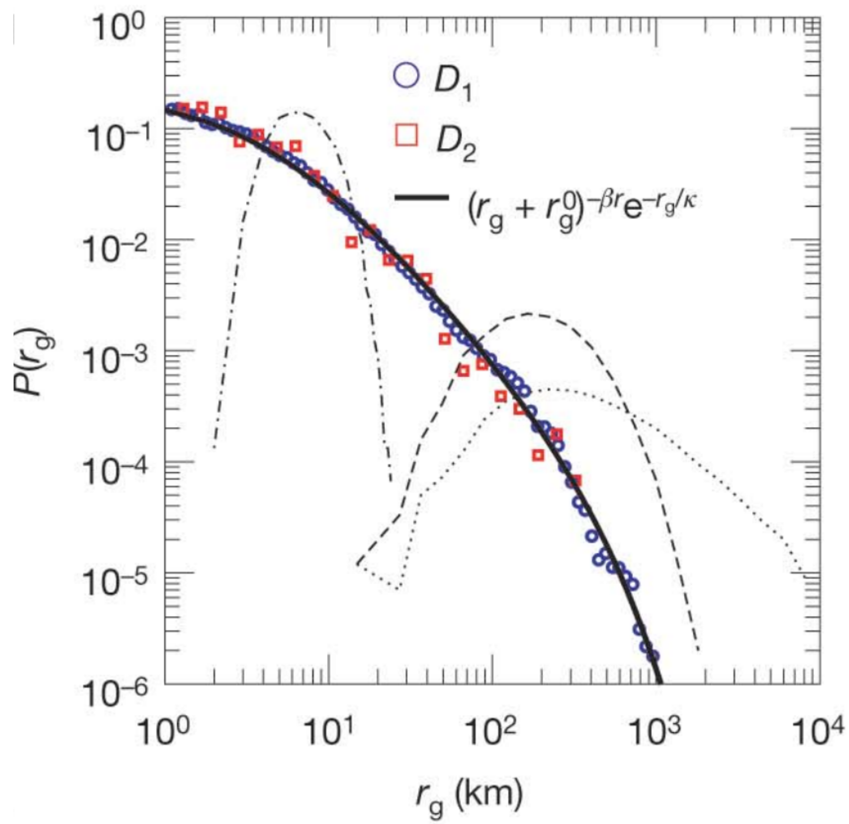
- The phone towers are shown as grey dots
- The trajectory describes the user's movements during 4 days (each day in a different color).

users move within a territory

Characteristic distance traveled by an individual

s of gyration
es heavy tails

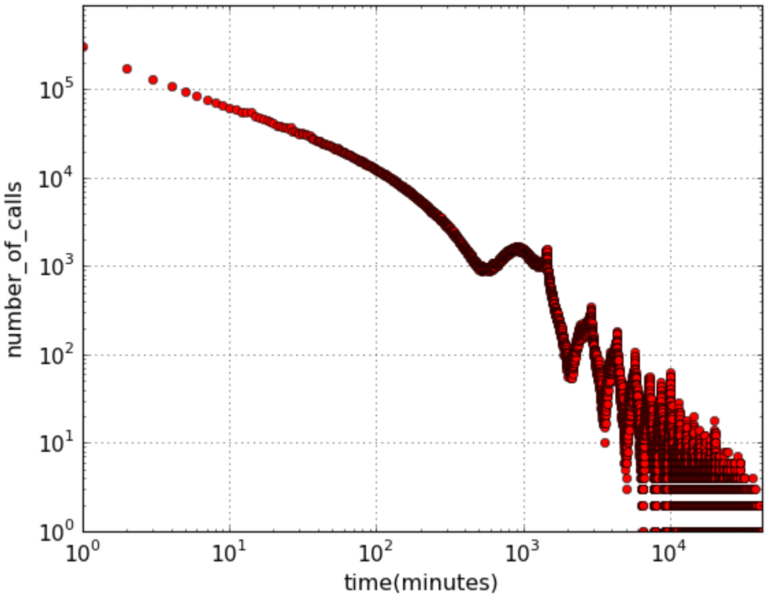
$$r_g = \sqrt{\frac{1}{N}\sum_{i\in L} n_i(\vec{r}_i - \vec{r}_{cm})^2},$$



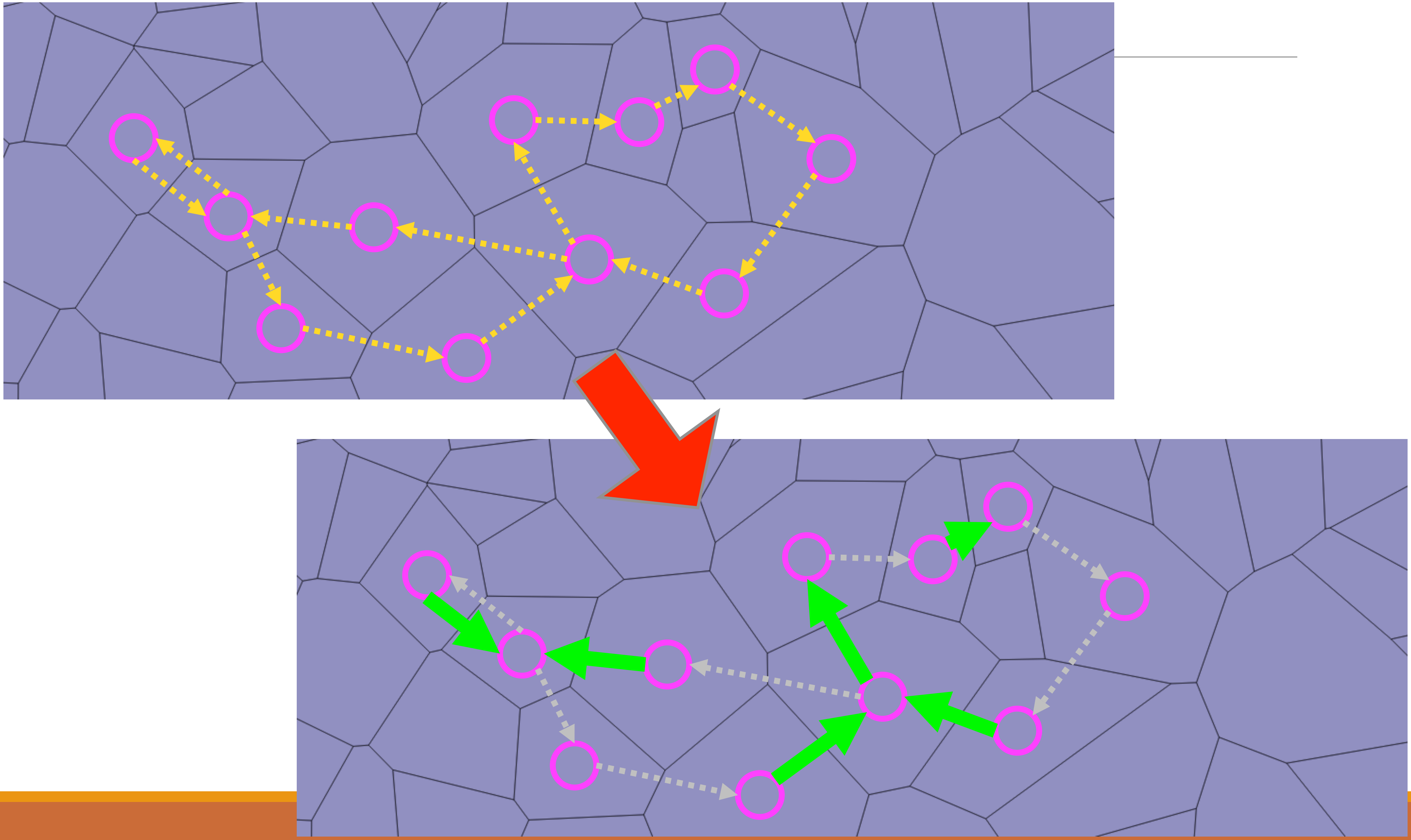
ez, Hidalgo, Barabási. *Nature 453(7196):779--782* (*June 2008*)

Estimating movements

- Reconstruct individual mobility through consecutive locations (individual flows)
- If **| time(Call_1) – time(Call_2) | < ΔT**
  then consider movement **Call_1 → Call_2**
- Issue: how to choose threshold?
  - Large ΔT => spurious data
  - Small ΔT => miss data

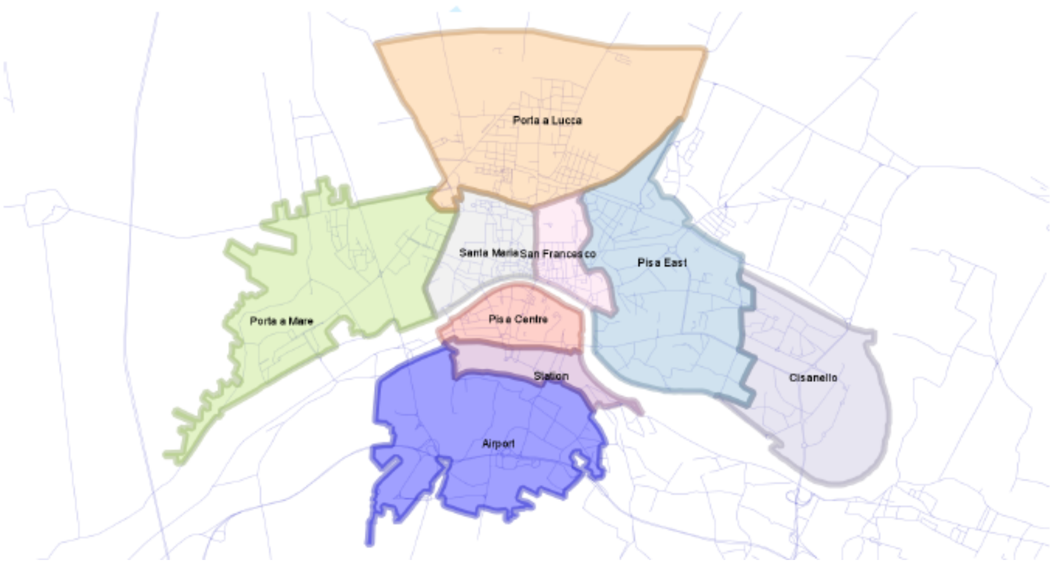


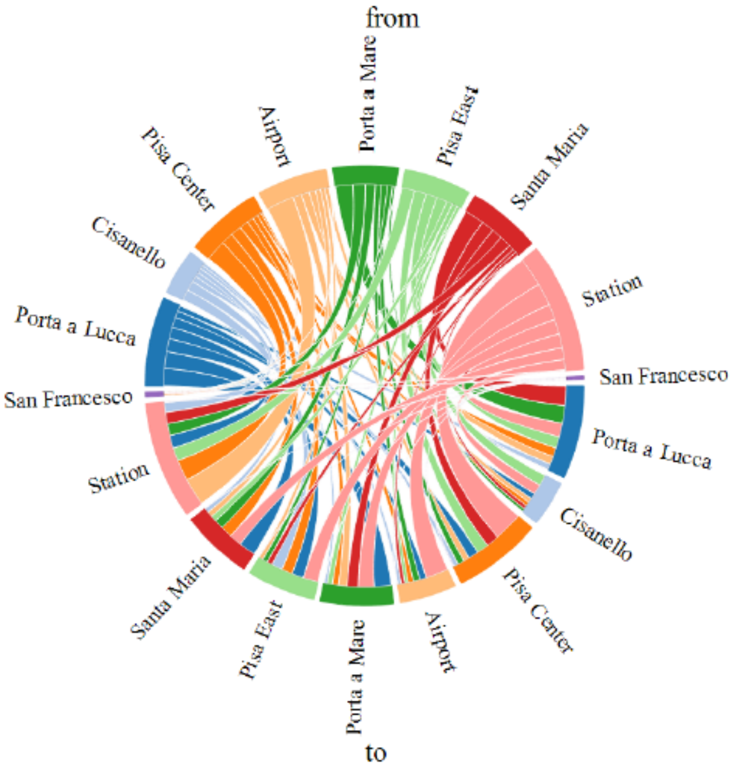
Estimating movements



Estimating movements

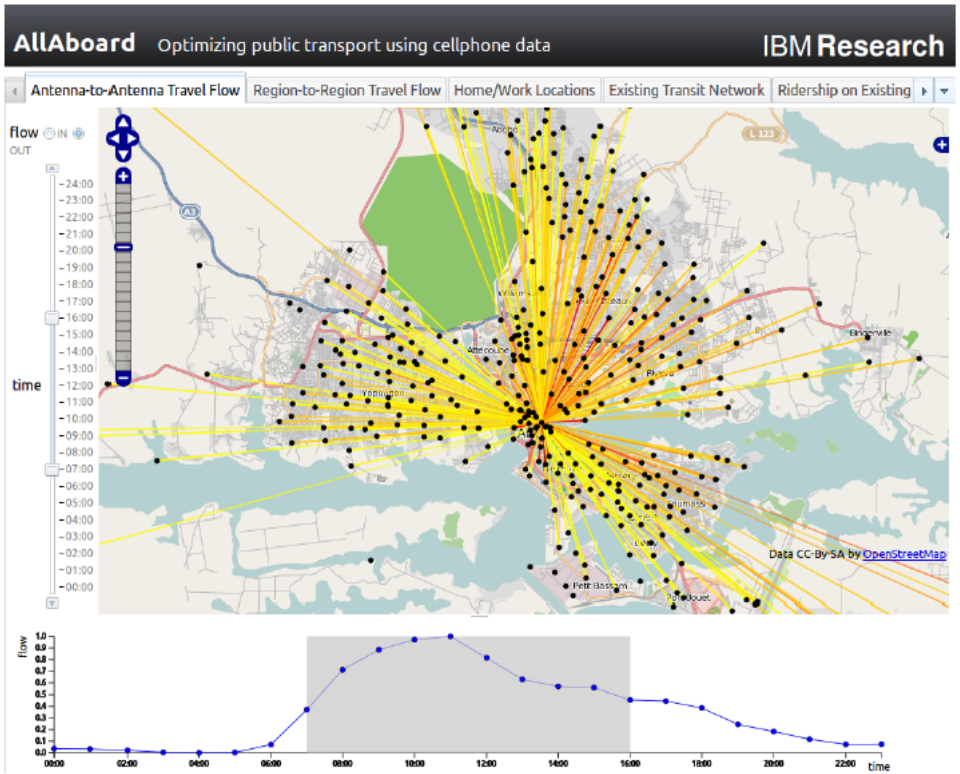
- Example on Pisa city

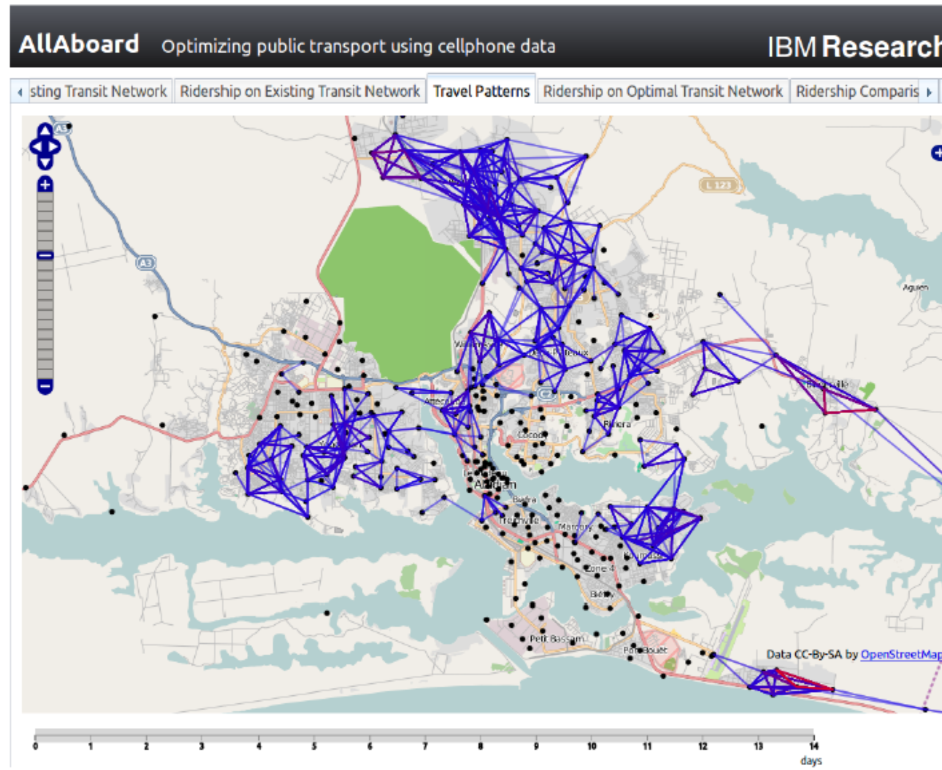




Estimating movements

- Example on Abidjan (Ivory Coast)





Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, Marco Luca Sbodio.

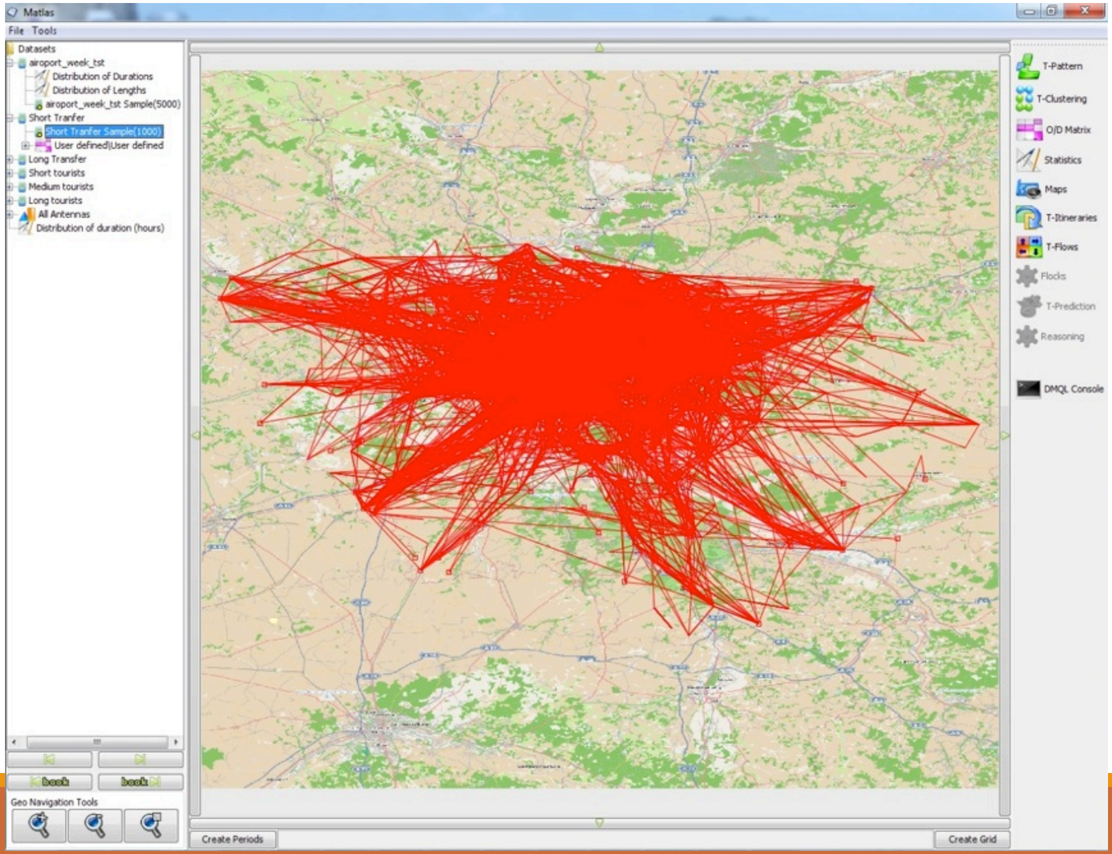
AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data.

http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=4746

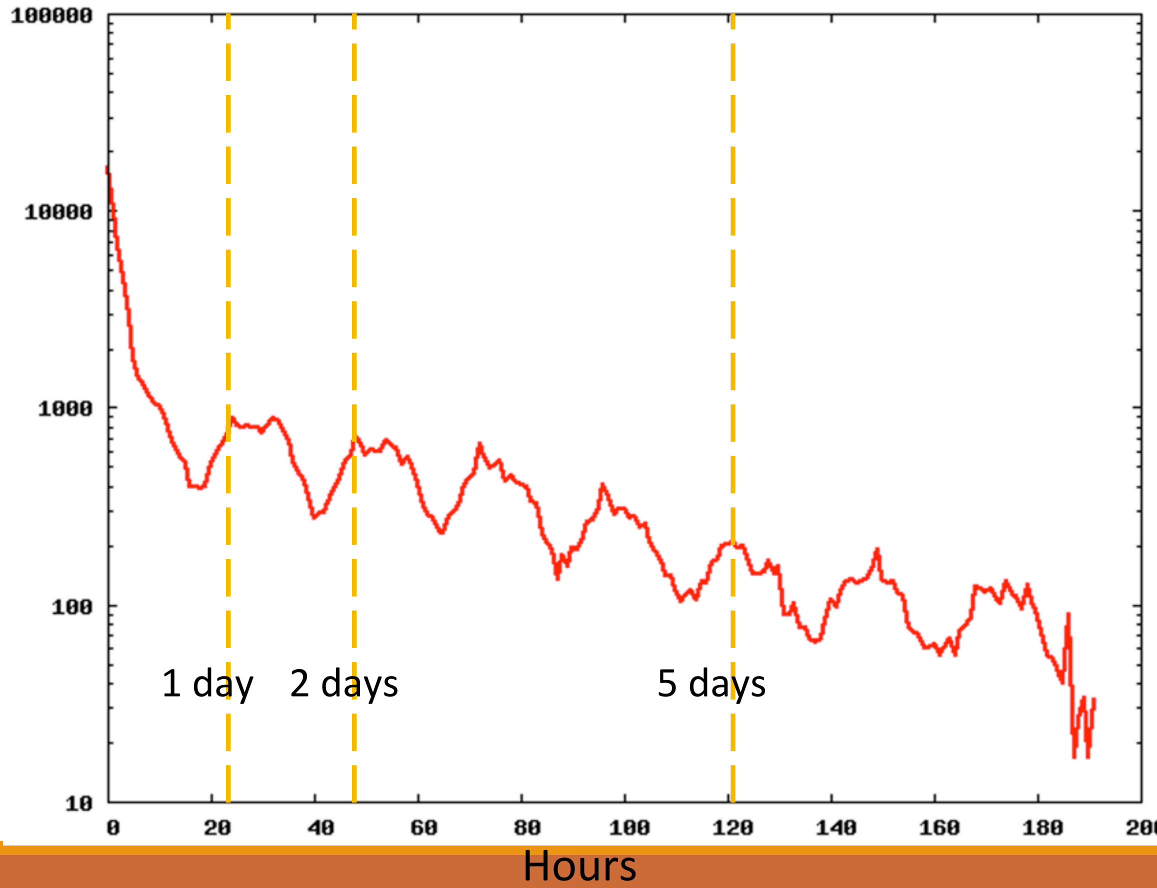
Sample application: Analyzing tourist data

- Case study of foreign (roaming) visitors of Paris area
- Users arriving and leaving at CDG airport

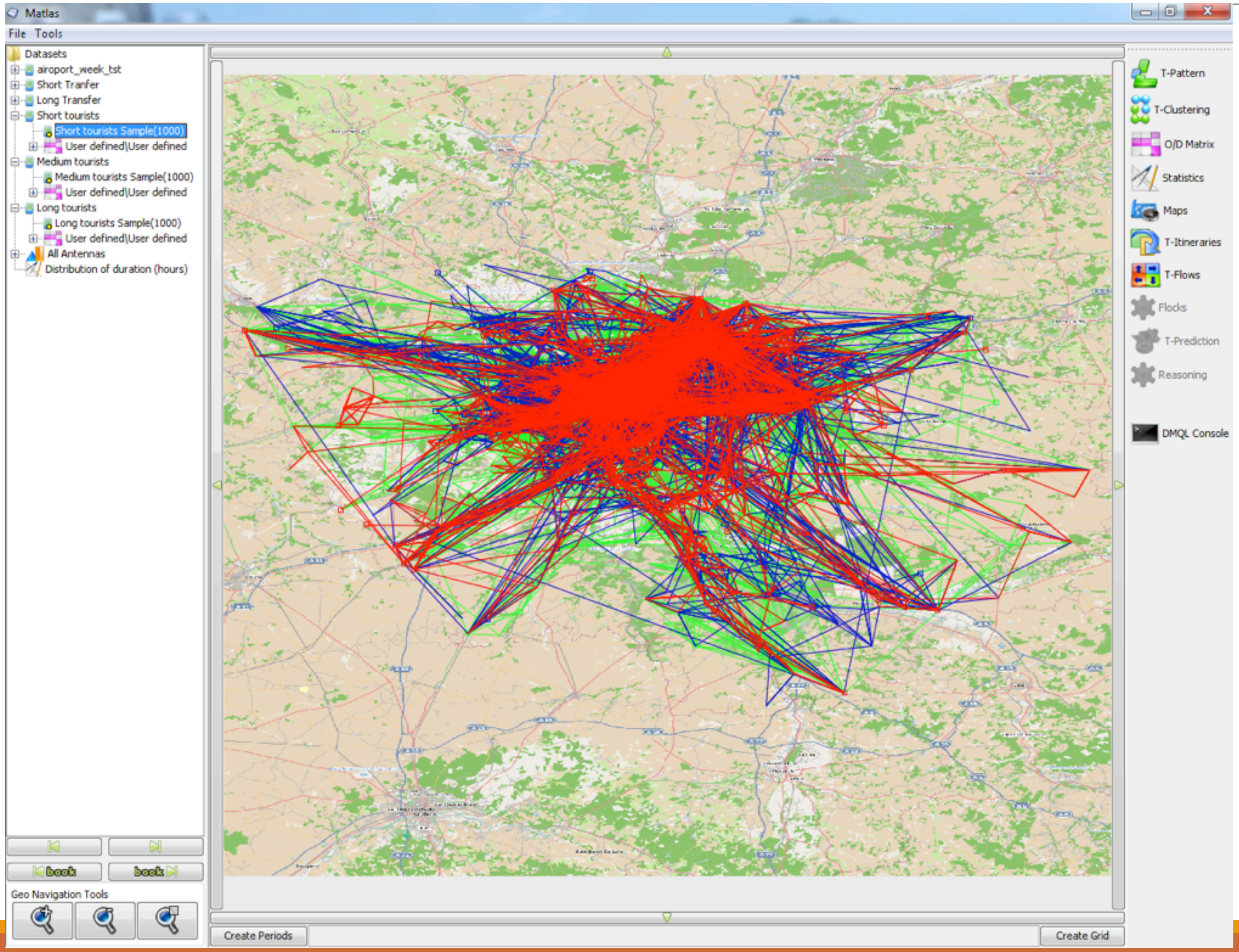
106 000 Users

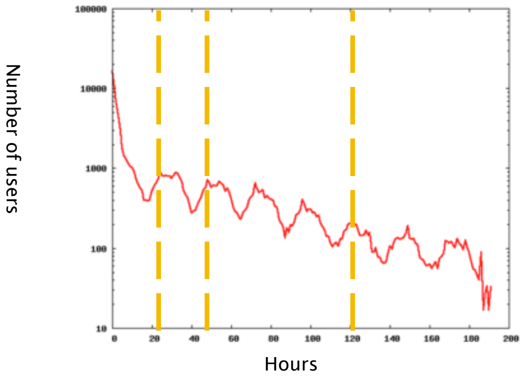


Distribution of visiting time



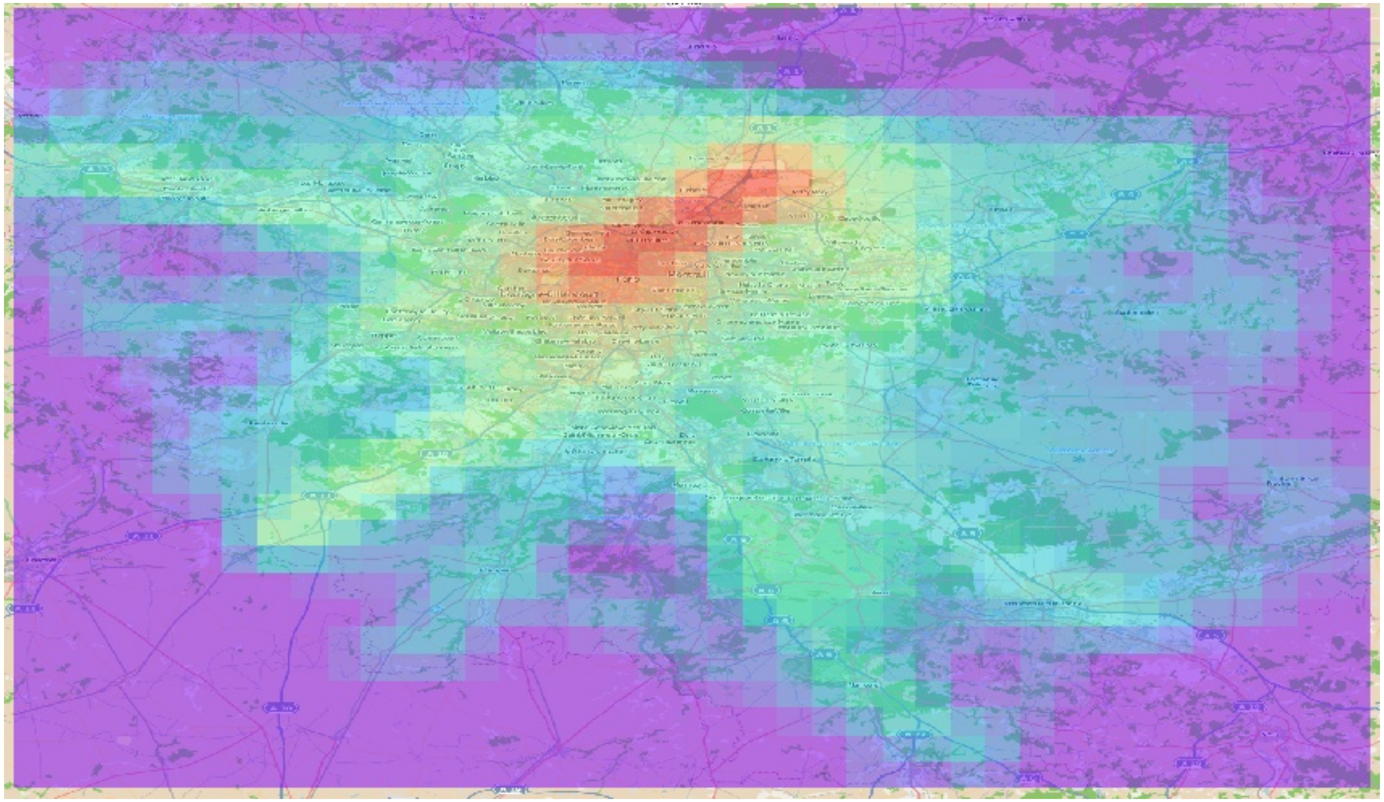
Categorization of tourists





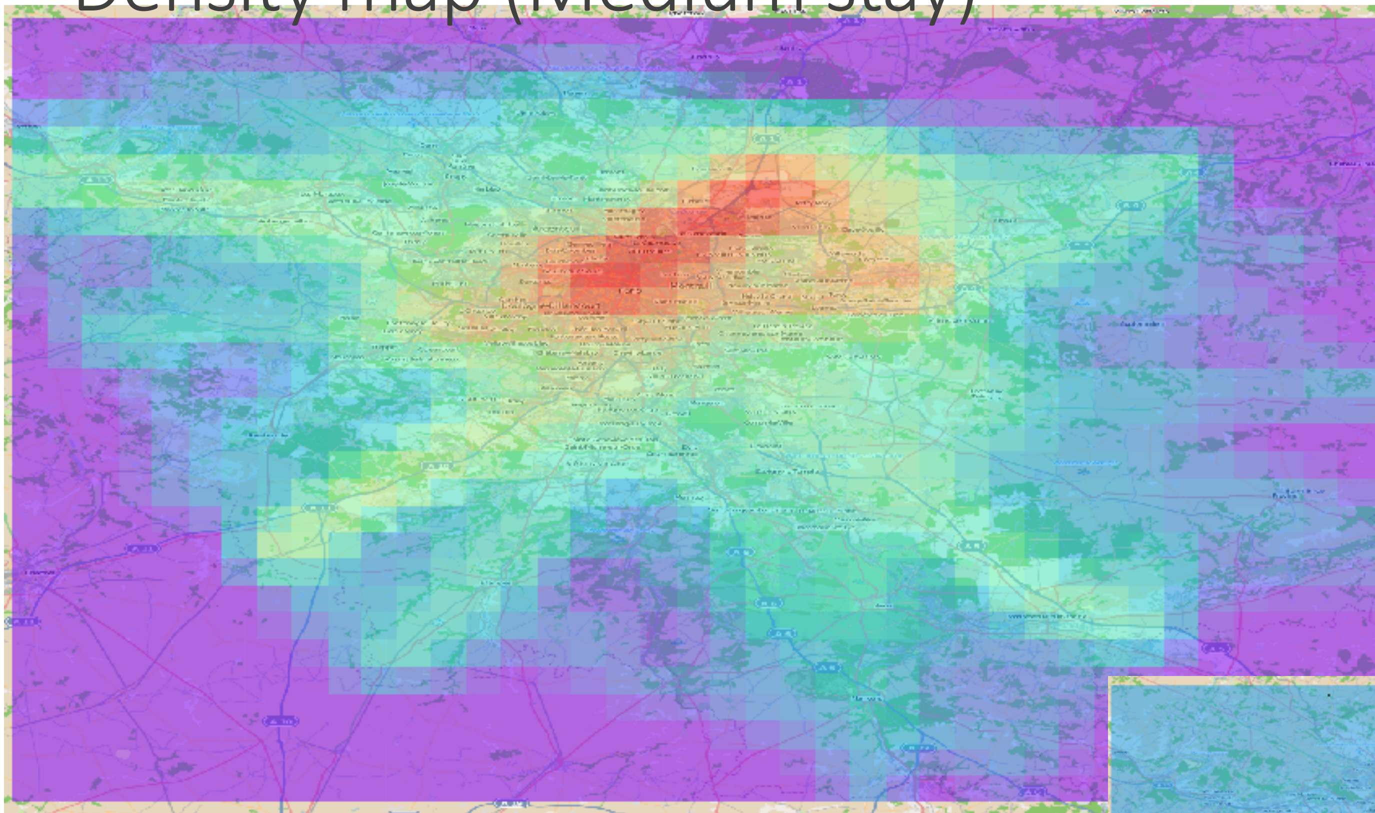
Short period stay Tourist (1 day → 2 days)
Medium period stay Tourist (2 day → 5 days)
Long period stay Tourist (5 day → 7 days)

Density map (Short stay)



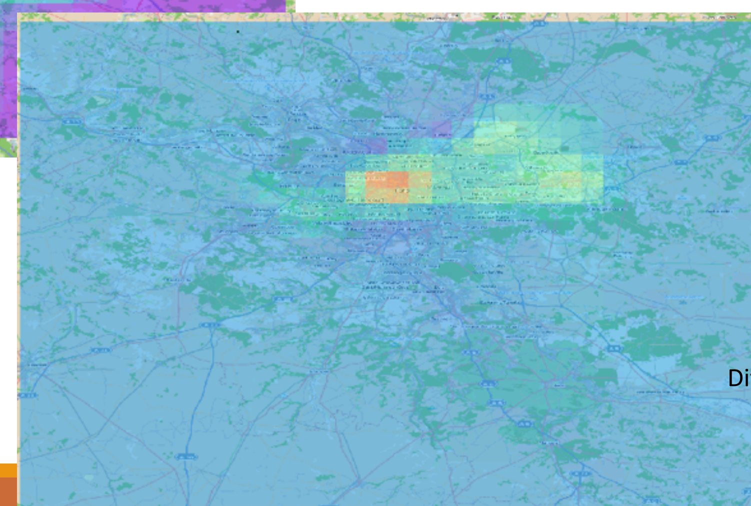
Short stay tourists visi

Density map (Medium stay)



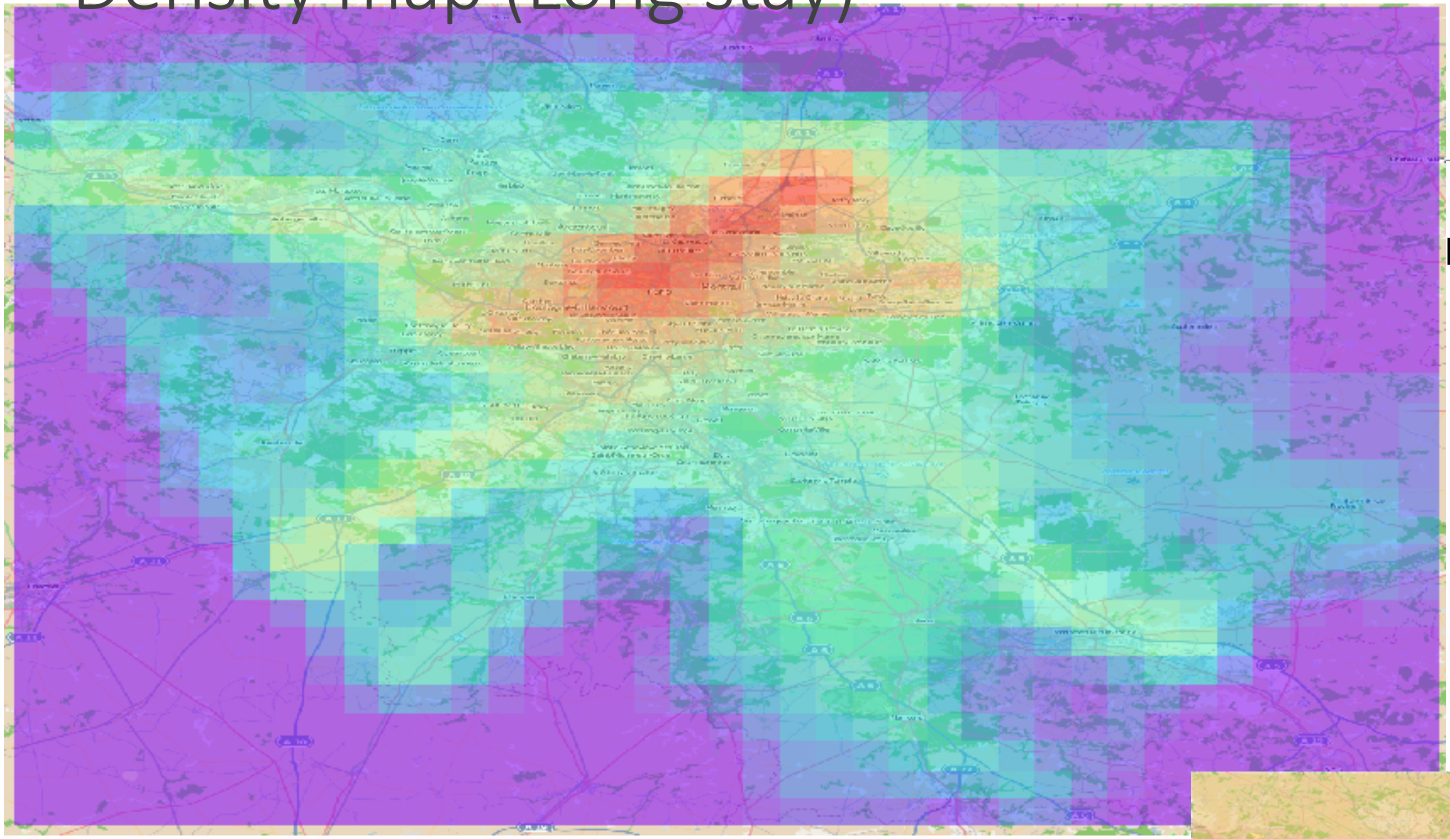
Medium stay tourists visit t

Green = Disneyland Paris
Red = Versailles



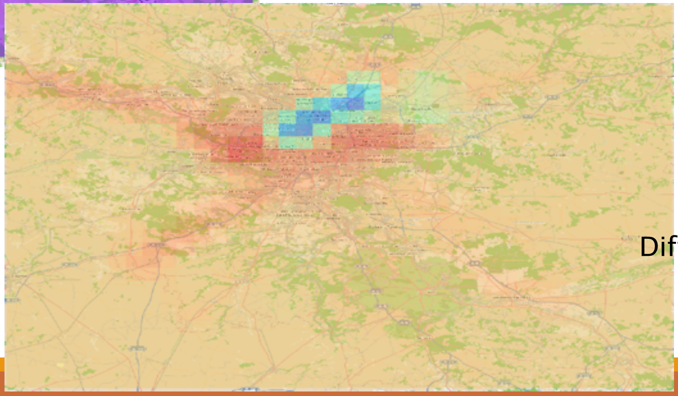
Differ

Density map (Long stay)



Long stay tourists visit t

Green = Disneyland Paris
Red = Versailles
Blue = Highway/Train to Mante la jolie
Black = Highway to South-West



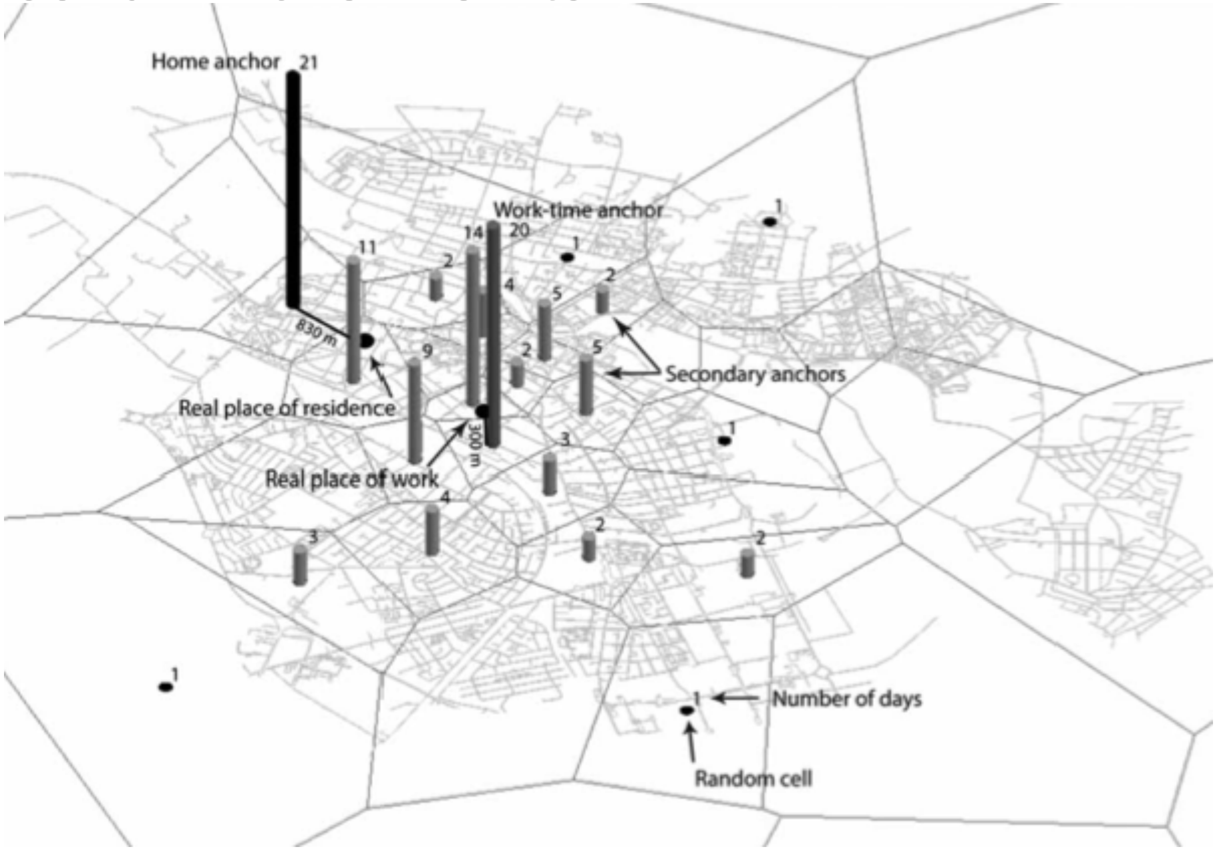
Diff

Identifying important locations

- Home (residence) and Work play an important role in understanding urban mobility
- “**Personal Anchor Points**”: high-frequency visited places of a user
  - Select top 2 cells with max number of days with calls
  - Determine home and work through time constraints:
    - average start time of calls and its deviation

Identifying important locations

- **"Personal Anchor Points"**



AHAS, R., SILM, S., JARV, O., SALUVEER, E., AND TIRU, M. 2010. *Using mobile positioning data to model locations meaningful to users of mobile phones*. Journal of Urban Technology 17, 1, 3–27.

Identifying important locations

- Estimating users' **residence through night activity**
  - Home = region with highest frequency of calls during nighttime
- First issue: cells might not correspond perfectly to the regions to measure
- Second issue: cells might not have uniform density of population

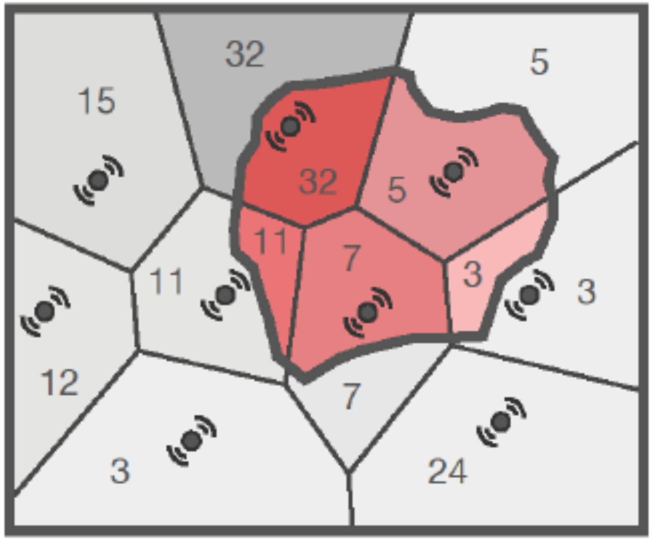
Pierre Deville et al.

Dynamic population mapping using mobile phone data.

PNAS vol. 111 no. 45, pp. 15888–15893, doi: 10.1073/pnas.1408439111

Identifying important locations

- First issue: cells might not correspond perfectly to the regions to measure

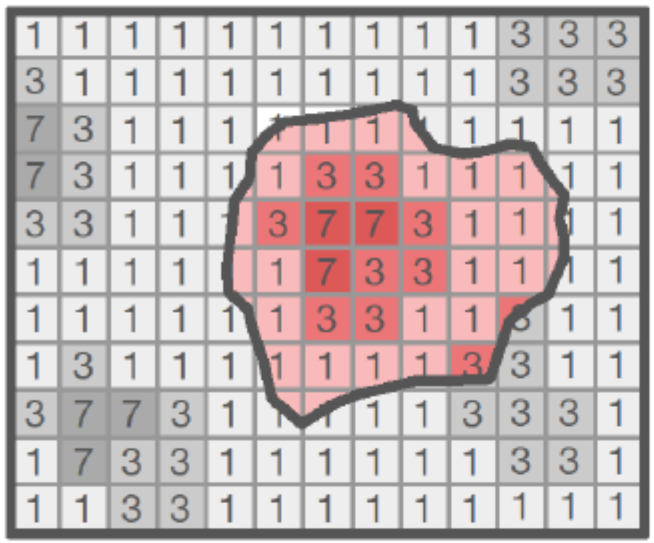


$$\sigma_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} \sigma_{v_j} A_{(c_i \cap v_j)}$$

- Approach: each cell contributes proportionally to its overlap with the region

Identifying important locations

- Second issue: cells might not have uniform density of population



$$\rho_i^{RS} = \frac{w_i}{\sum_j w_j} P$$

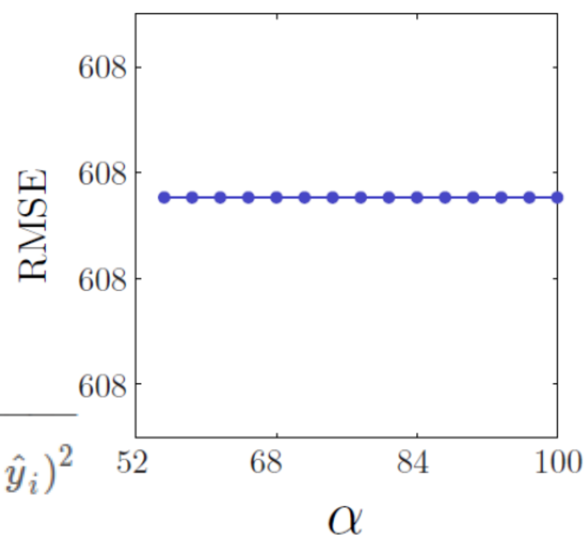
- Approach: integrate external indicators of relative density – e.g. from environment and infrastructures – to distribute cells' contrib.

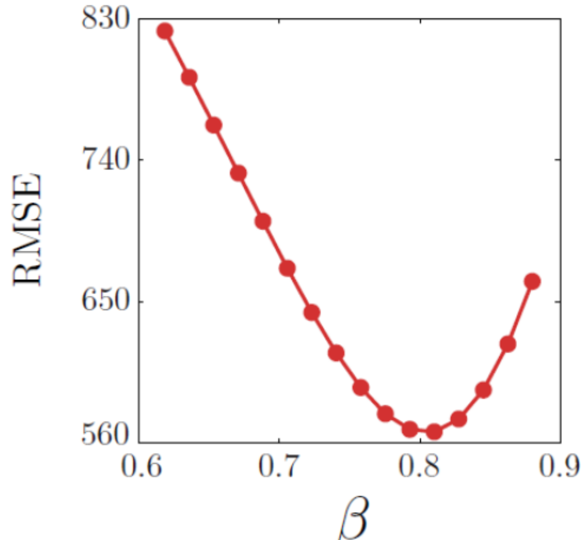
Identifying important locations

- Linear or superlinear relation?

$$\rho_c = \frac{P}{\hat{P}}\,\alpha\sigma_c^{\beta}$$

  - ρ_C = population density
  - σ_C = mobile phone residents
  - P = national population (real vs. estimated)

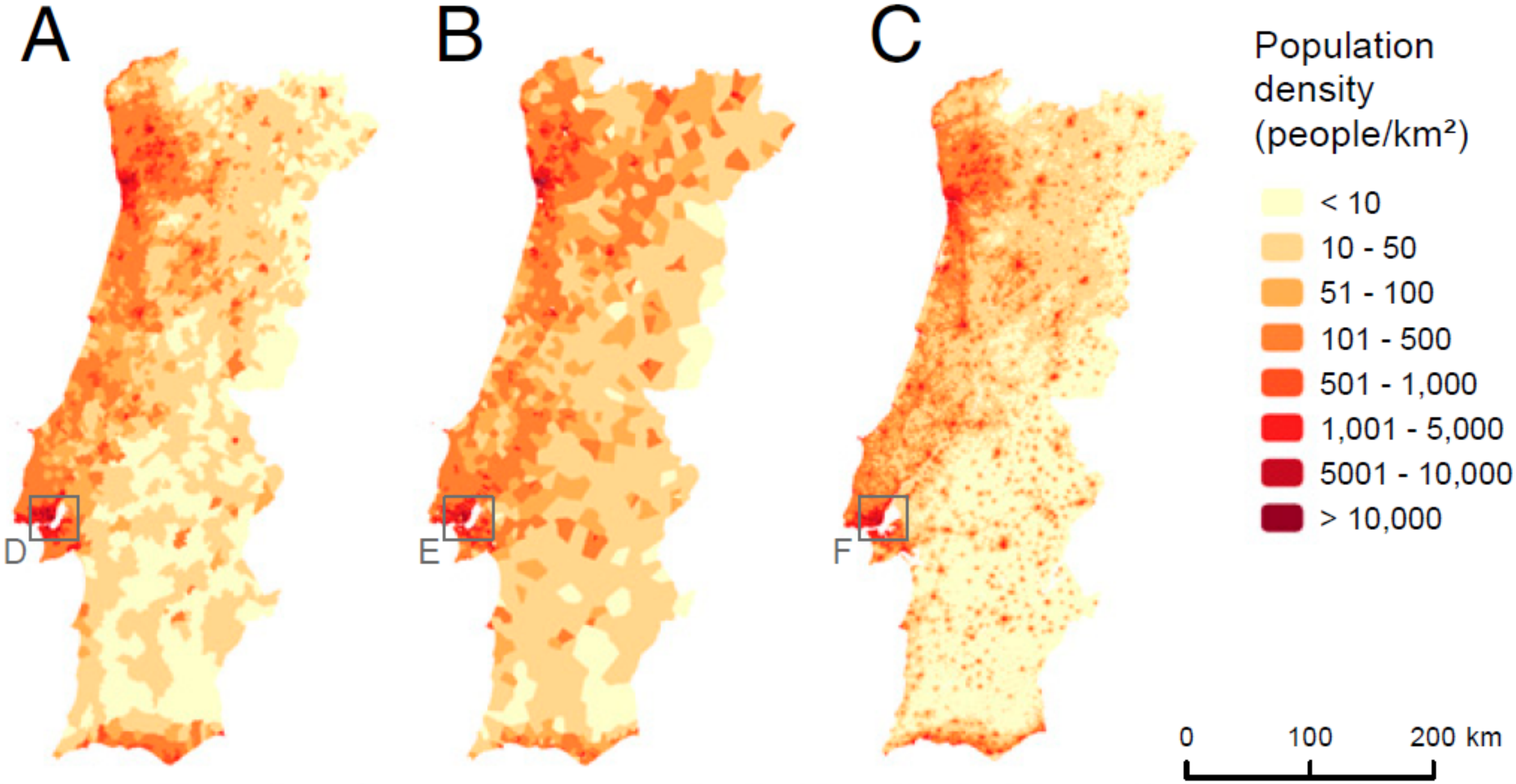




$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Identifying important locations

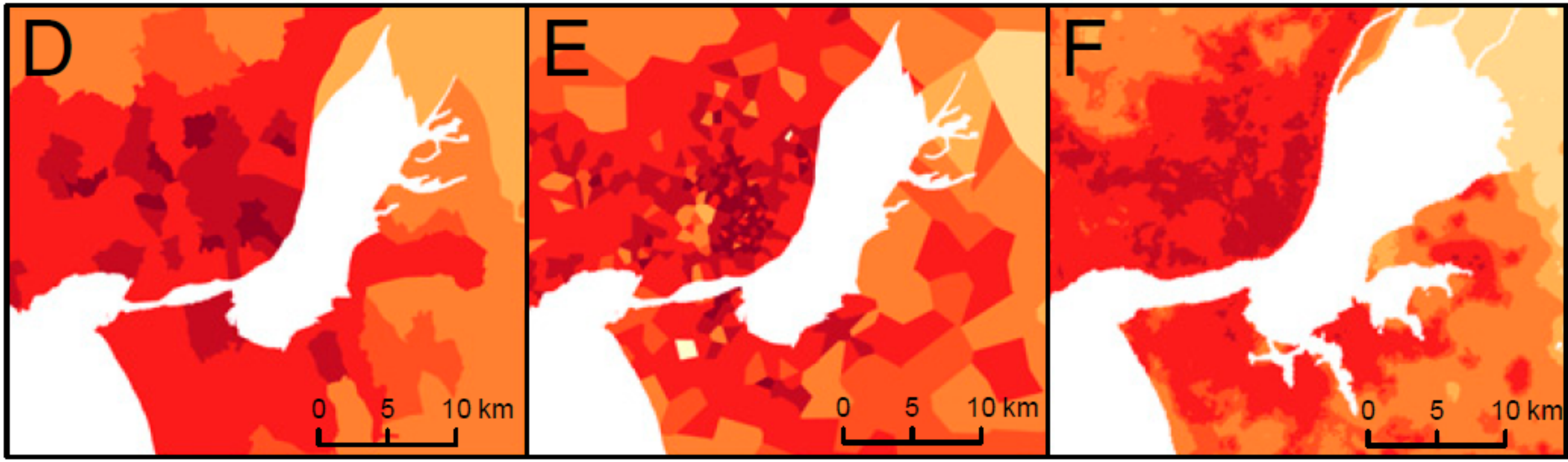
- Sample results on Portugal



A = Census B = GSM data C = Environment/Infrastructures-based

Identifying important locations

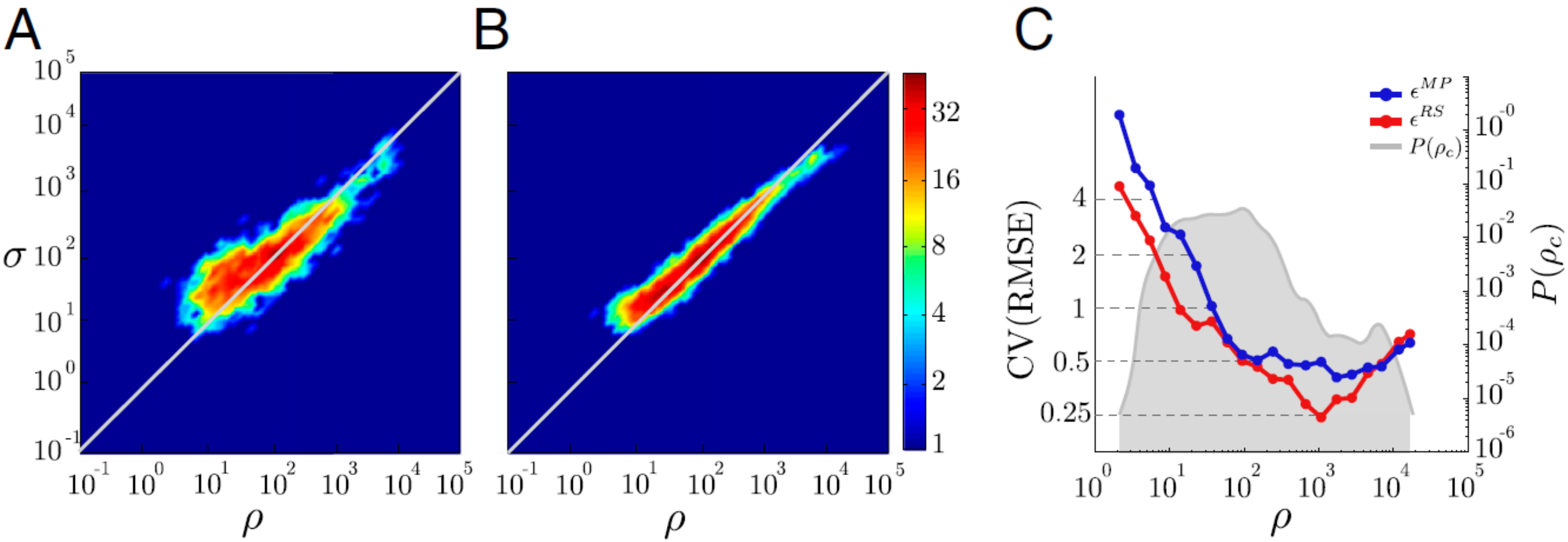
- Sample results on Portugal (close-up)



D = Census E = GSM data F = Environment/Infrastructures-based

Identifying important locations

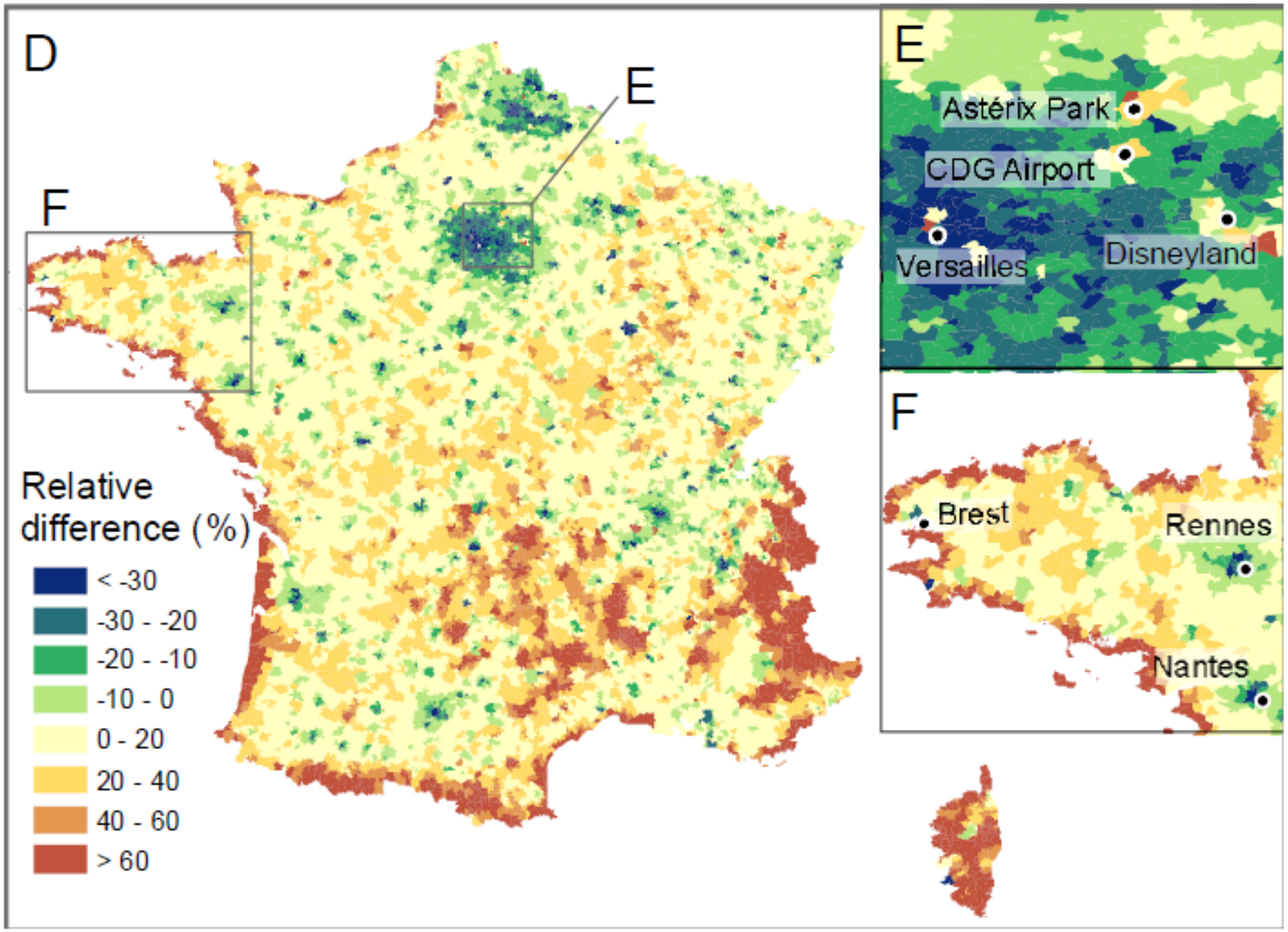
- Sample results



A = GSM data   B = Environment/Infrastructures-based

Identifying important locations

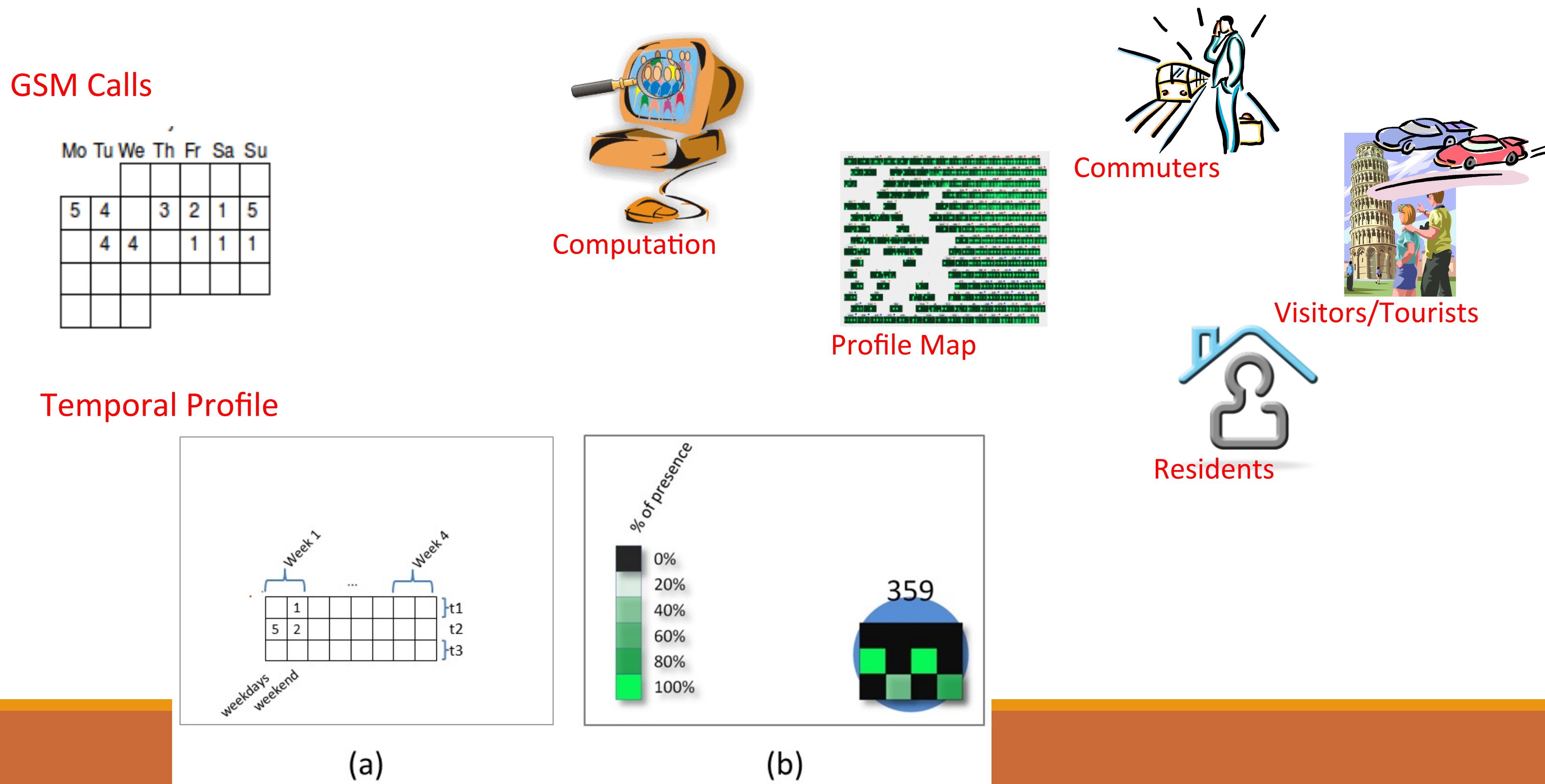
- Sample usage: evaluate seasonal changes
  - Summer variations vs. Winter period



Classifying into **city users** categories

Basic methodology: Sociometer

- GSM calls used as proxy of users' presence in a specific area
- 3 categories used: Residents, Commuters, Visitors

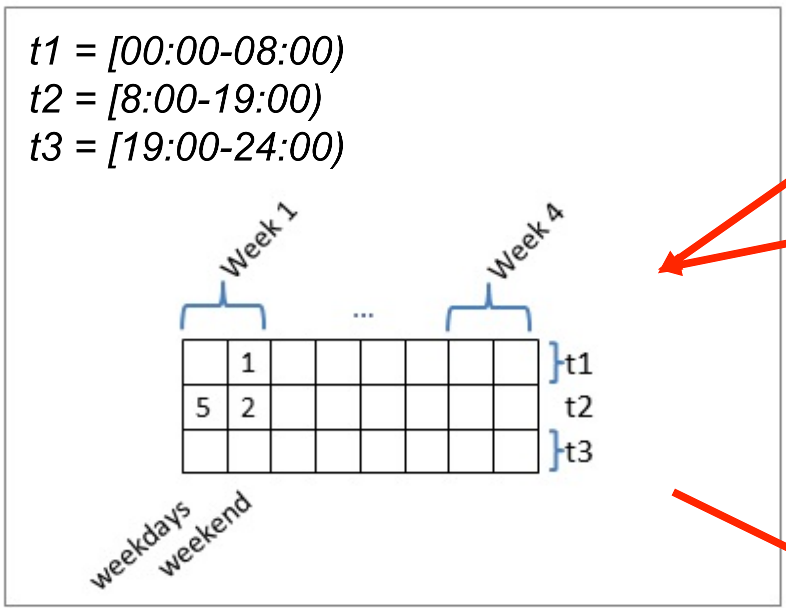


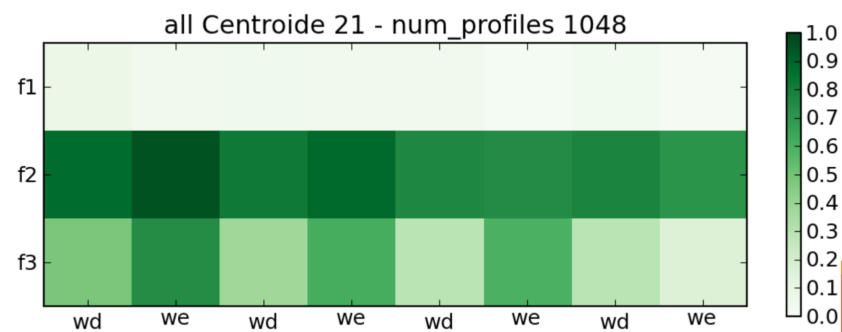
Sociometer

Step 1: build individual profiles

- Derive presence distribution for each < user, municipality >

123643 Cell12 24/06/2012 14:05
123643 Cell12 24/06/2012 18:13
123643 Cell15 25/06/2012 11:05
123643 Cell15 25/06/2012 20:42
123643 Cell11 25/06/2012 21:05
123643 Cell12 26/06/2012 10:01
….

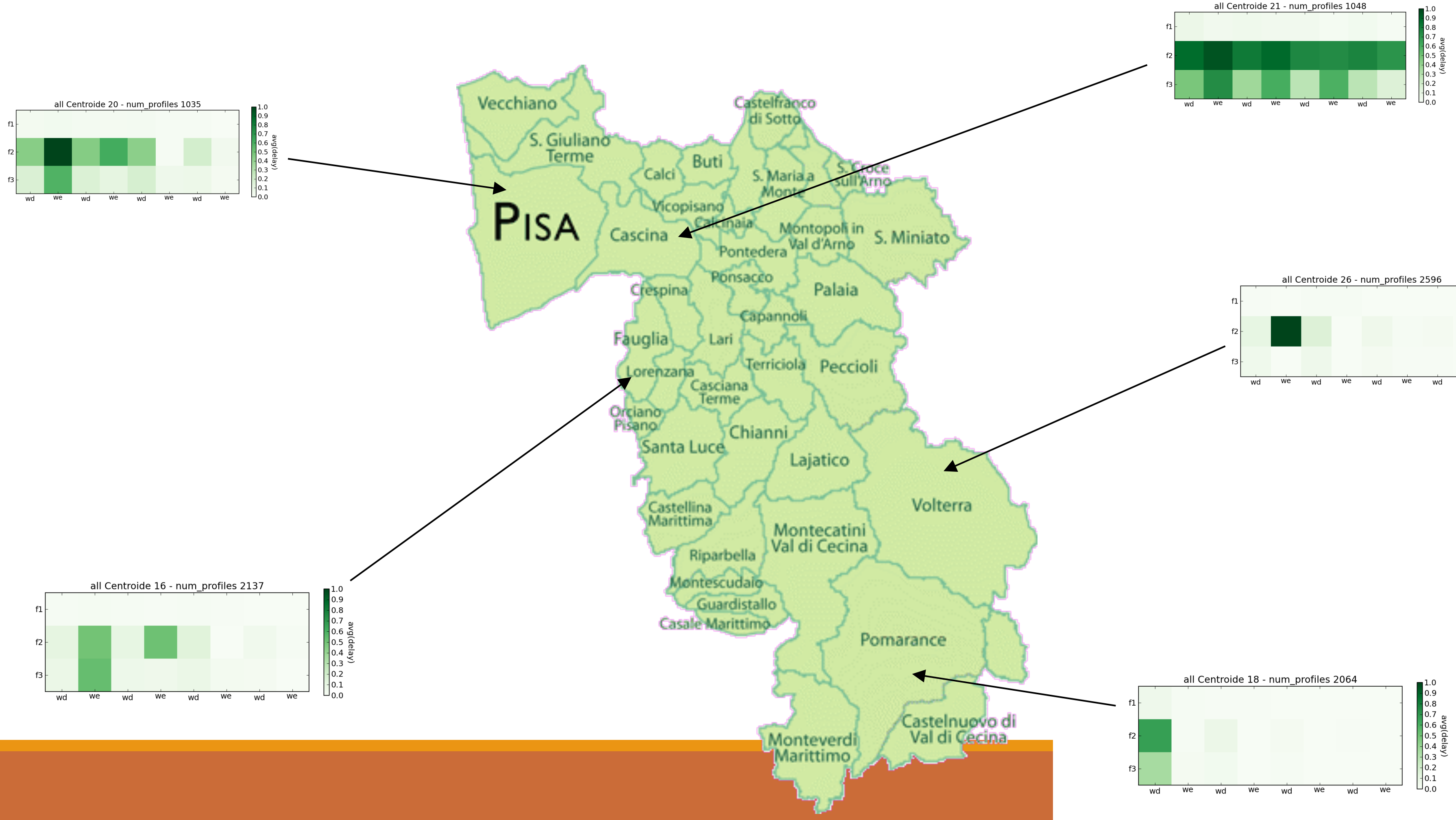




Sociometer 2.0

Step 1: build individual profiles

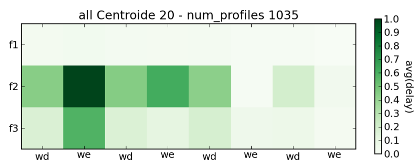
- Result for each user: set of individual profiles:

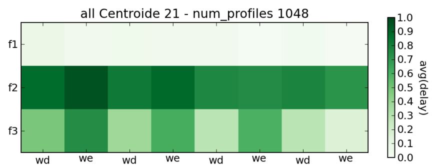


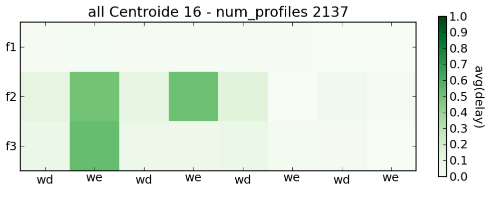
Sociometer 2.0

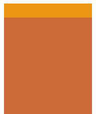
Step 2: find representative profiles across all dataset

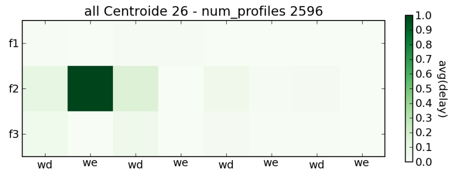
- Based on clustering
  - simple k-means: start with K random representatives, and iteratively refine them
  - in our experiments, k=100
- Output: set of reference (unlabelled) profiles

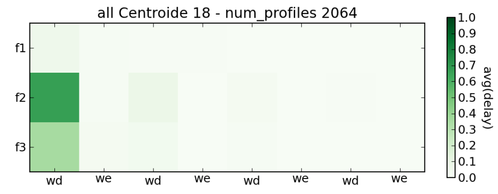








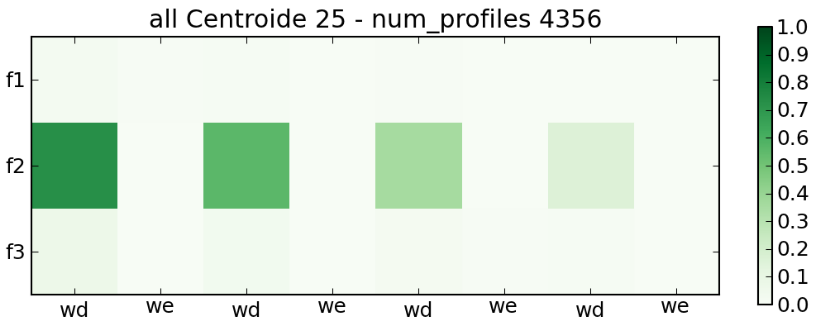




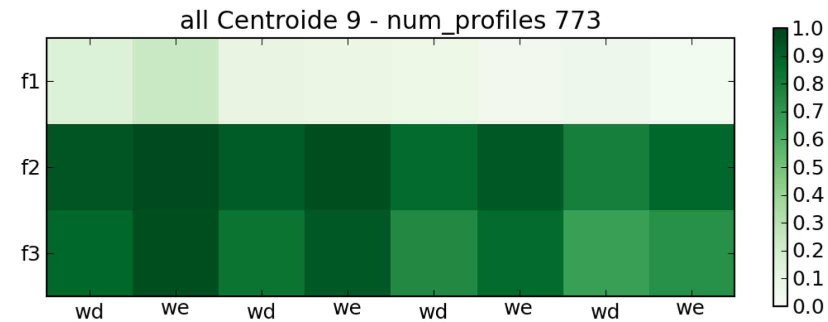
Sociometer 2.0

Step 3: associate representative profiles to categories

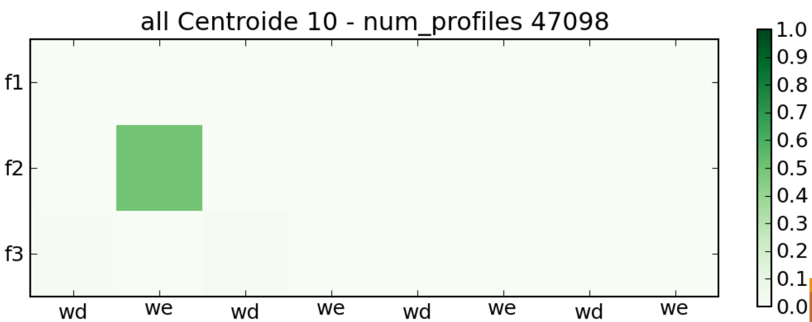
- Manual labelling
  - Use fuzzy rules, difficult to formalize
  - Crisp classification, no weights (reliability of labels)



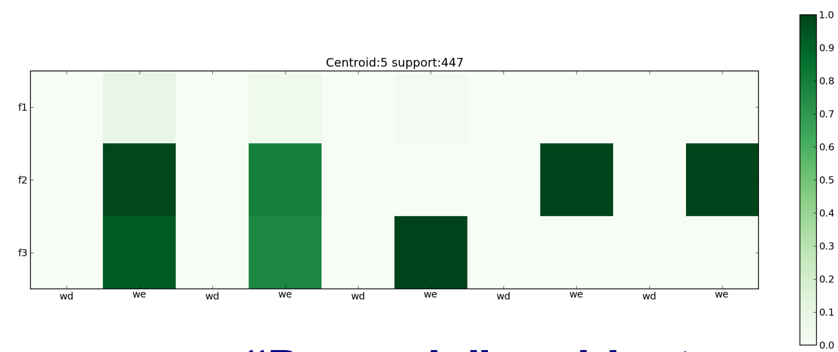
Commuter



"Static" resident



Occasional

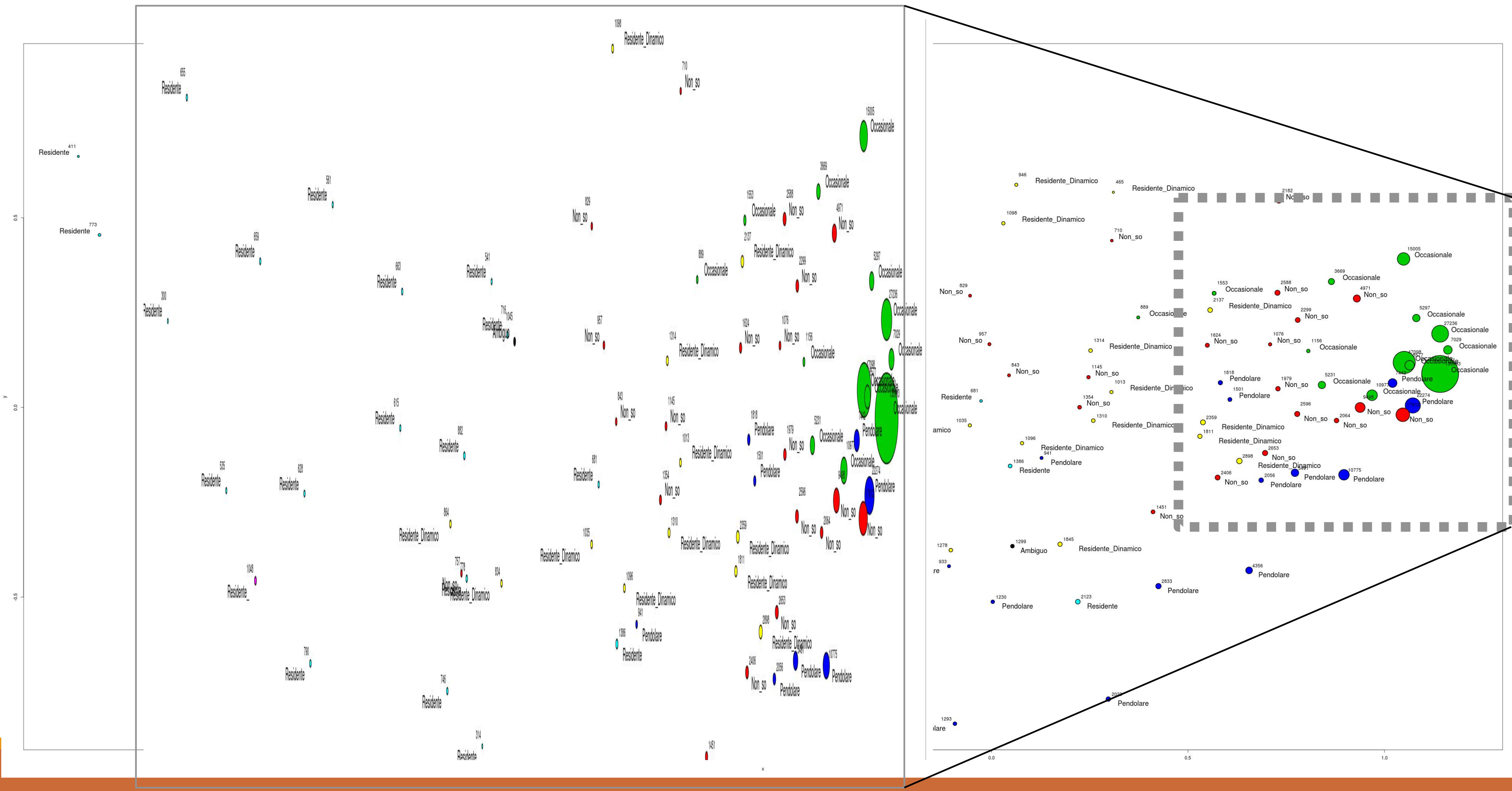


"Dynamic" resident

Sociometer 2.0

Step 3bis: consistency check / labels distribution / fix bugs

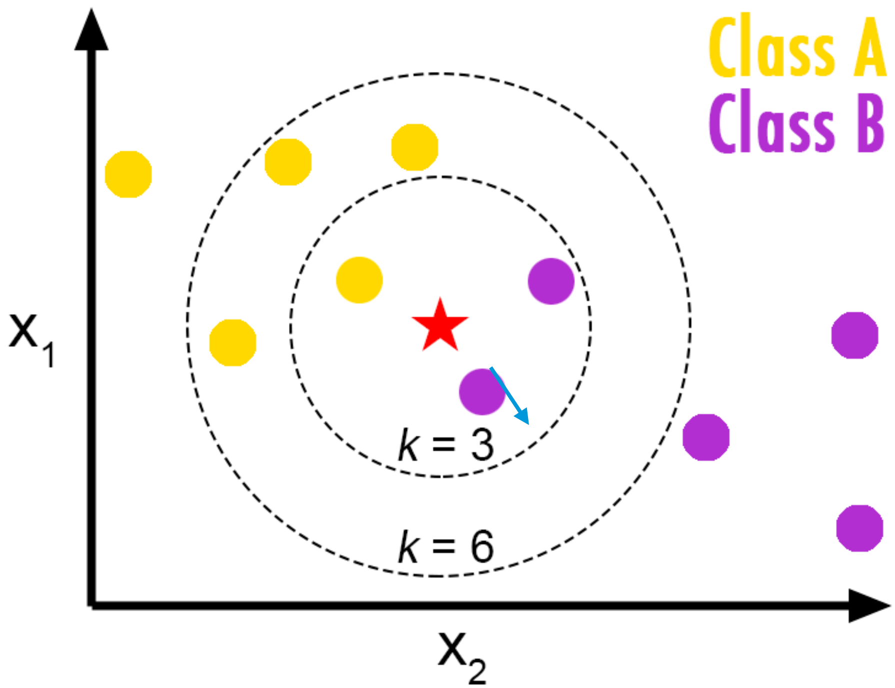
- Profiles (individual and representative) are 24-dimensional
- MDS (24 → 2) to visualize them



Sociometer 2.0

Step 4: label propagation

- Simple k-NN classification, k=1
  - Associates each individual profile to the closest representative profile
- So far, no voting schema (k>1) was used



Sociometer 2.0

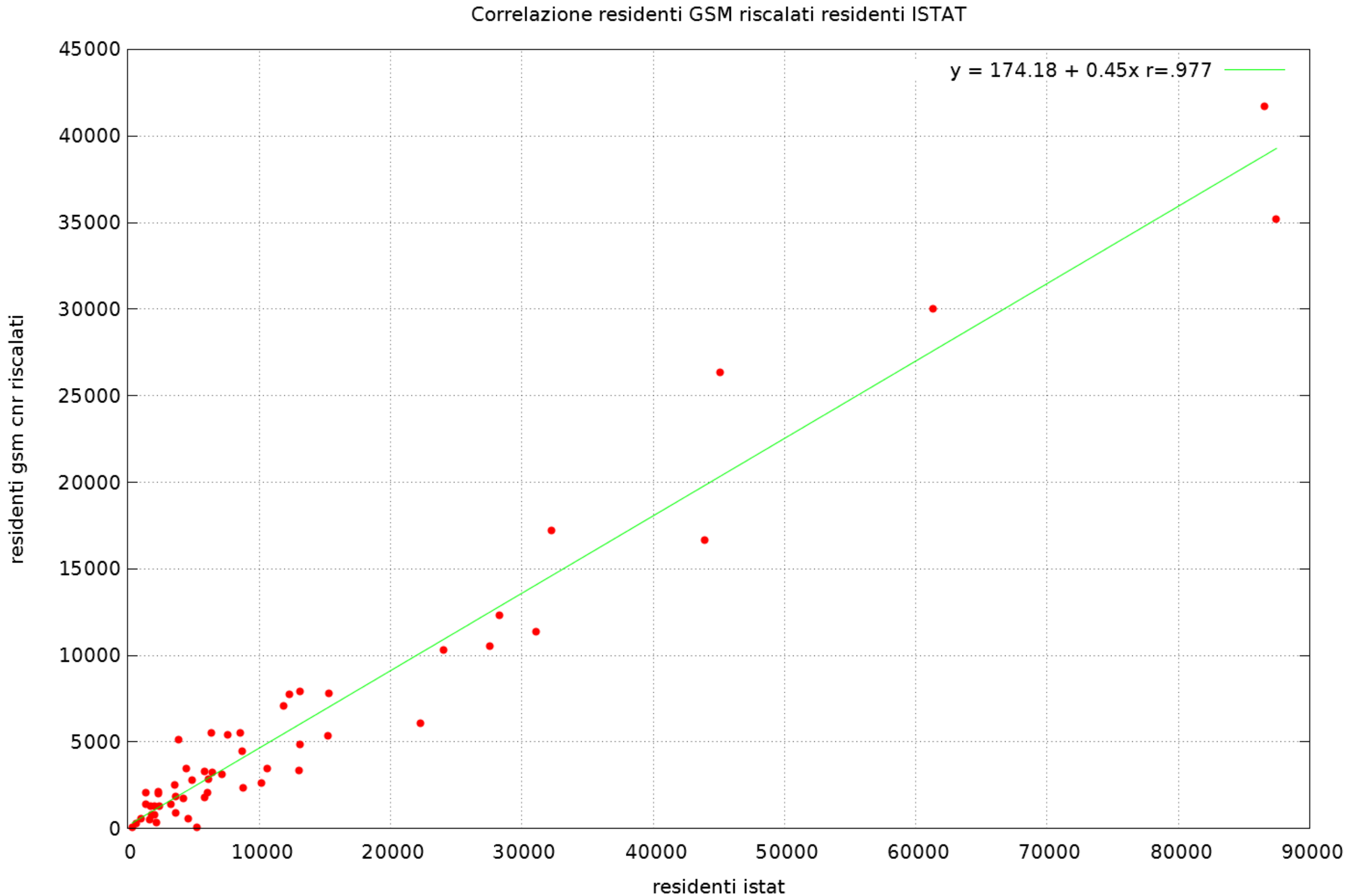
Step 5: aggregate into presence stats and O/D flows

- Presence aggregates
  - Residents = Static + Dynamic residents
- Kind of flows represented:
  - Dynamic residence → sites of commuting
  - Dynamic residence → sites of occasional visits

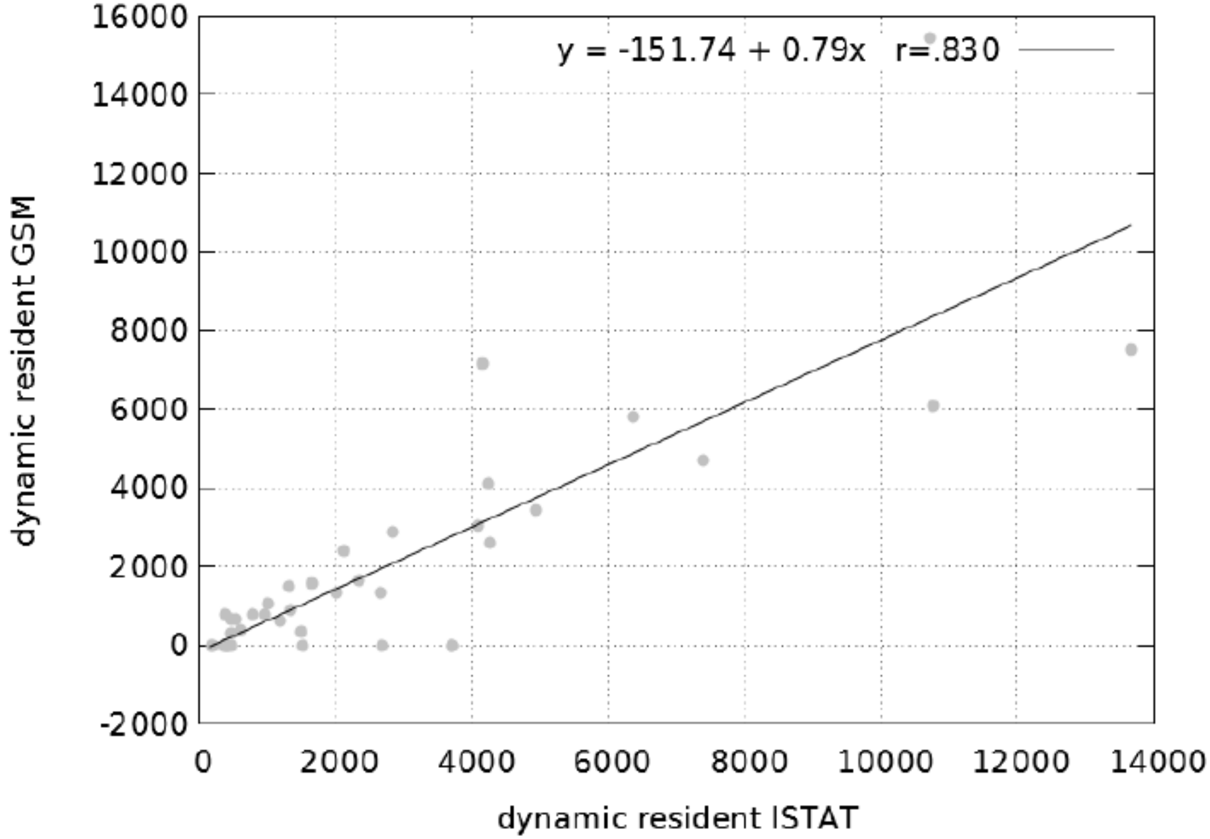
ISTAT Persons & Places project

- Ultimate goal: Use Big GSM data to
  - Estimate user categories on a given territory
  - Infer O/D matrix across municipalities
- Goal of this project:
  - Apply/adapt GSM-based user categorization (Sociometer) on municipalities of a large territory
  - Infer partial O/D matrix
  - Direct/Indirect comparison against official data
- GSM 4-weeks Dataset on Pisa and Lucca provinces

Static residents GSM

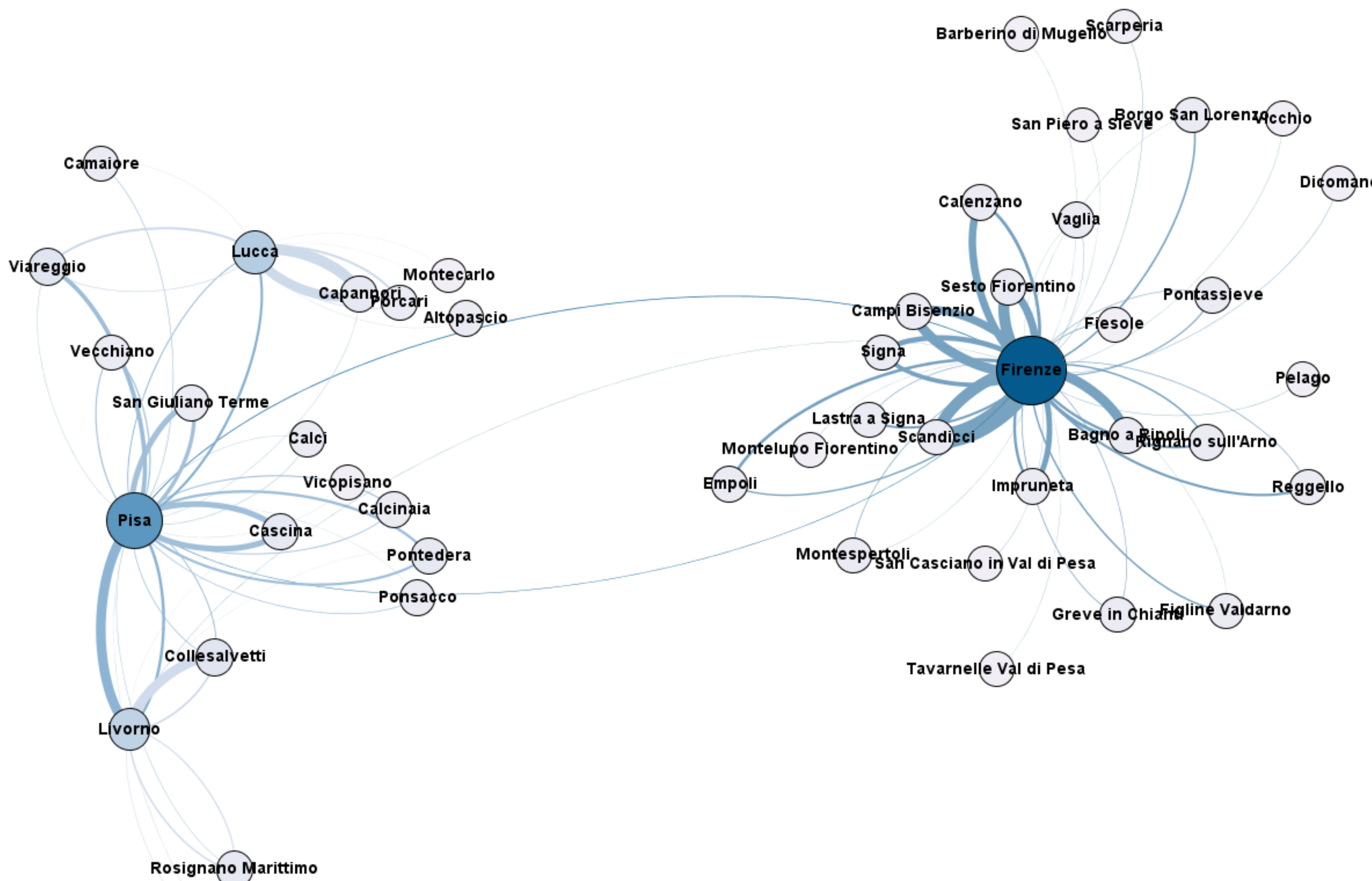


Dynamic residents (outgoing)



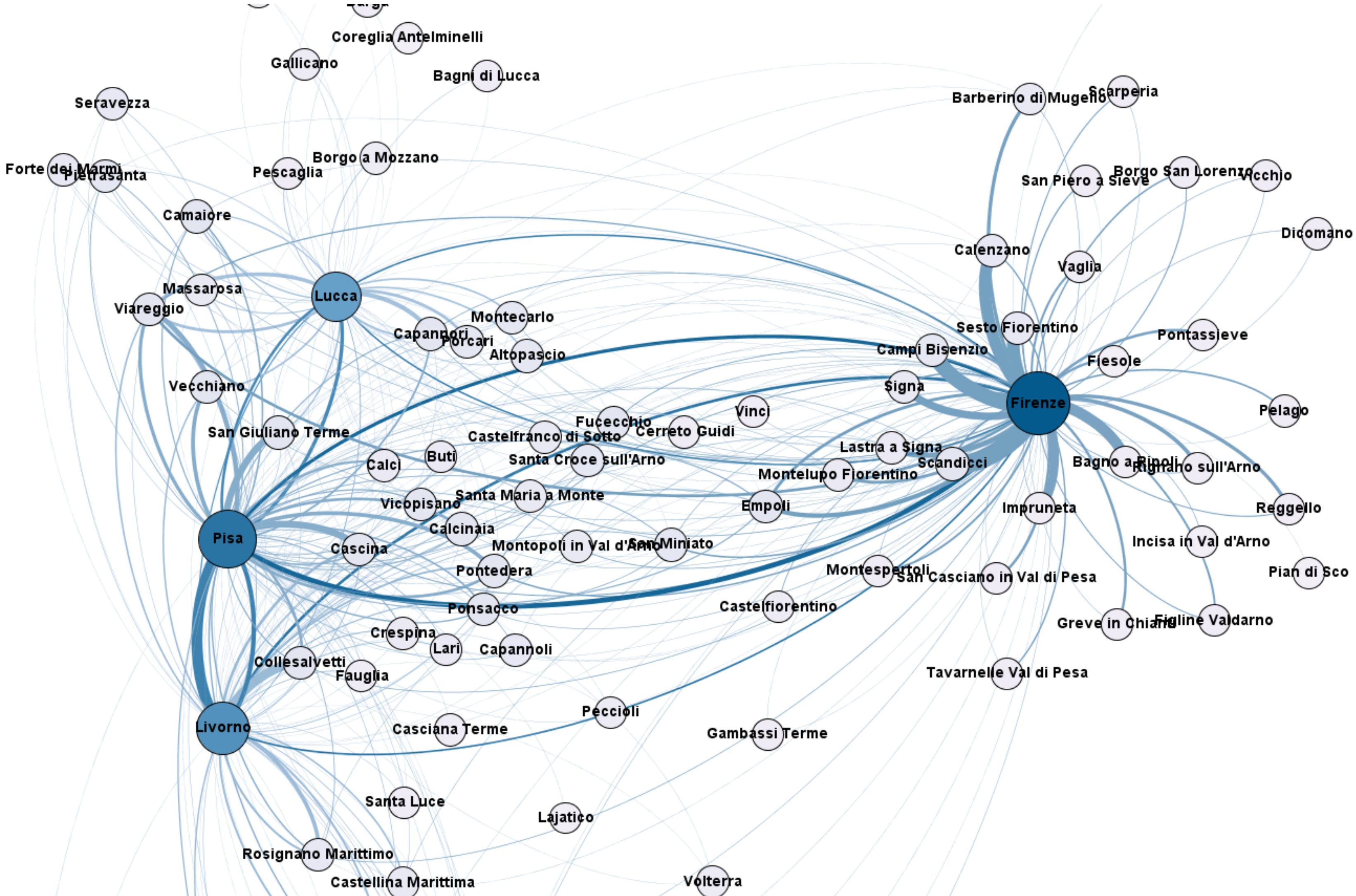
Sample results / 1

Home-Work



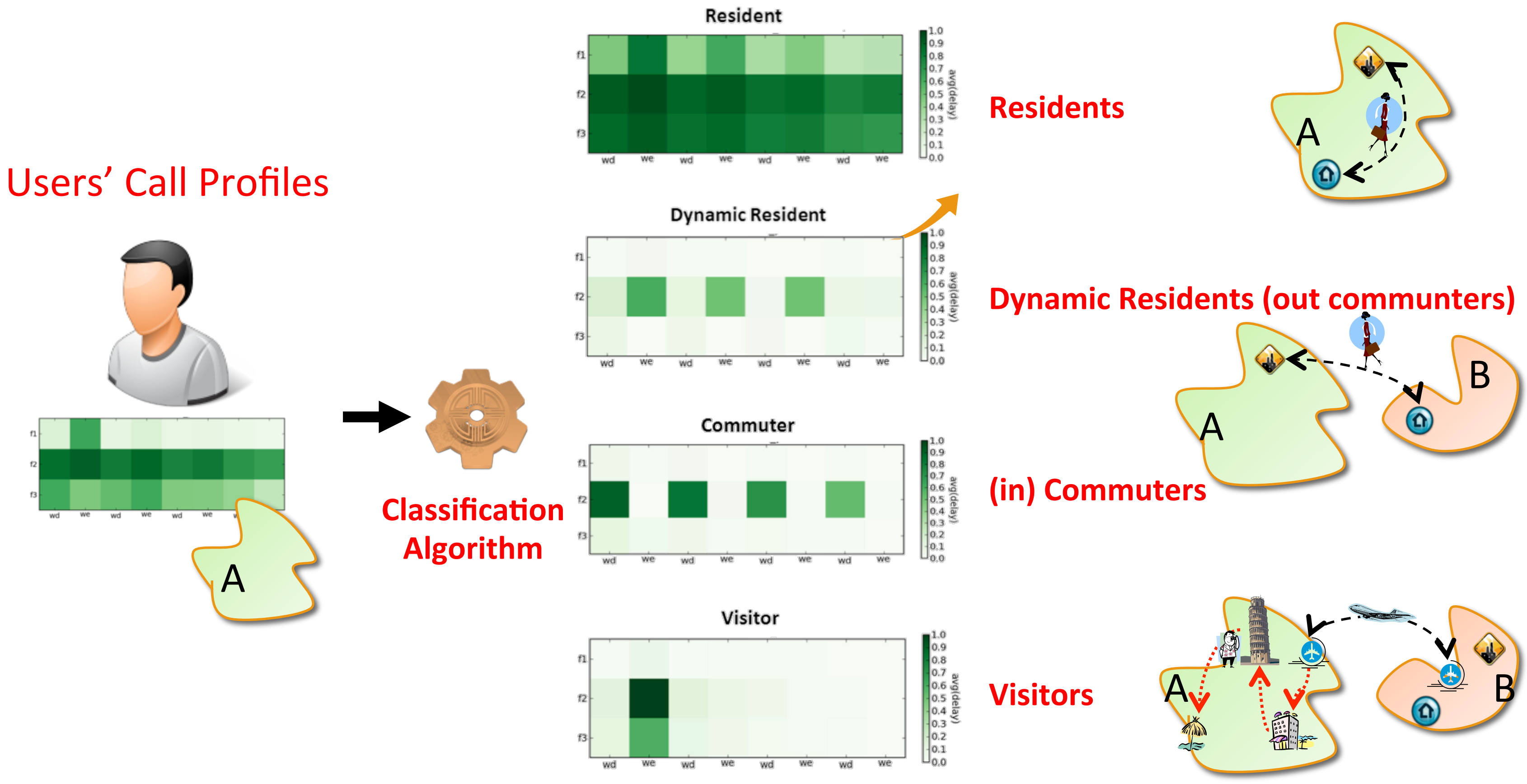
Sample results / 2

Home-Visits

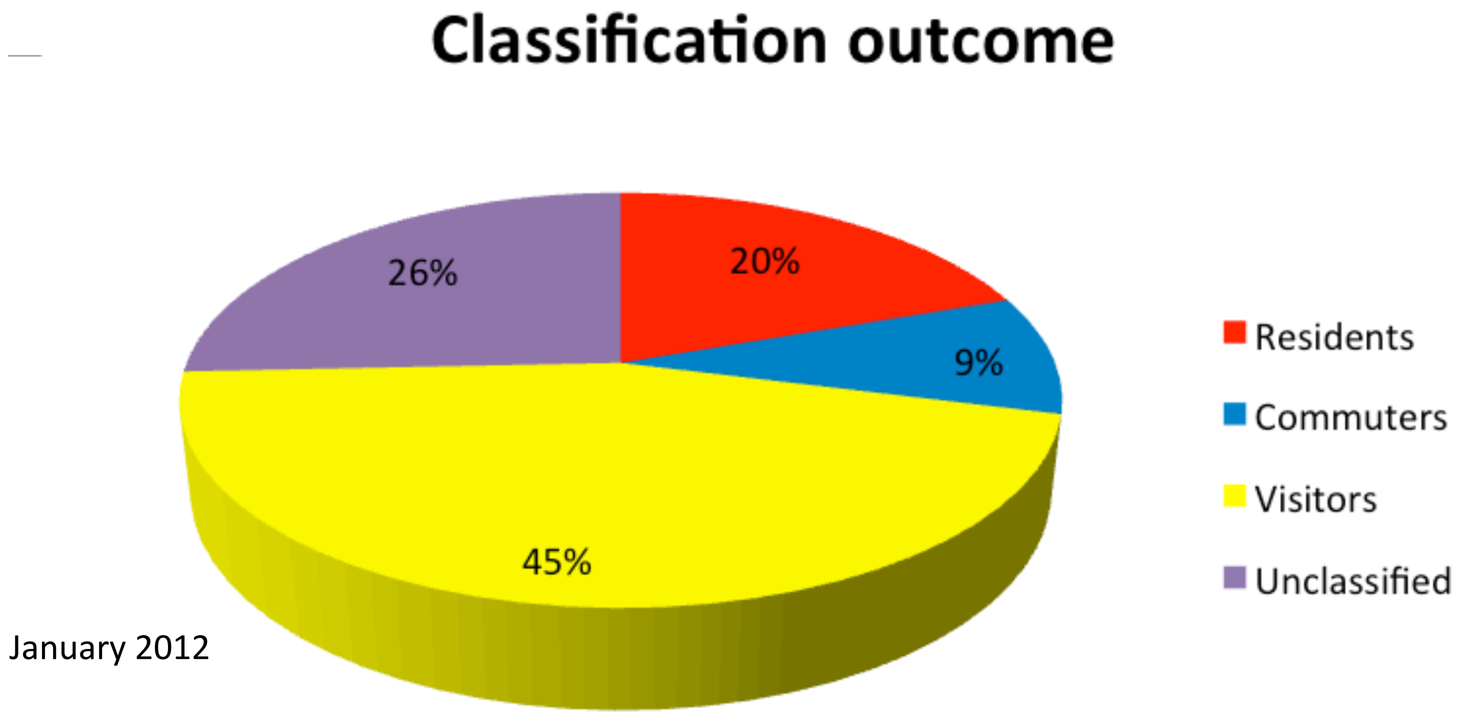


REAL TIME DEMOGRAPHY

Sociometer with Mobile Phone Data.



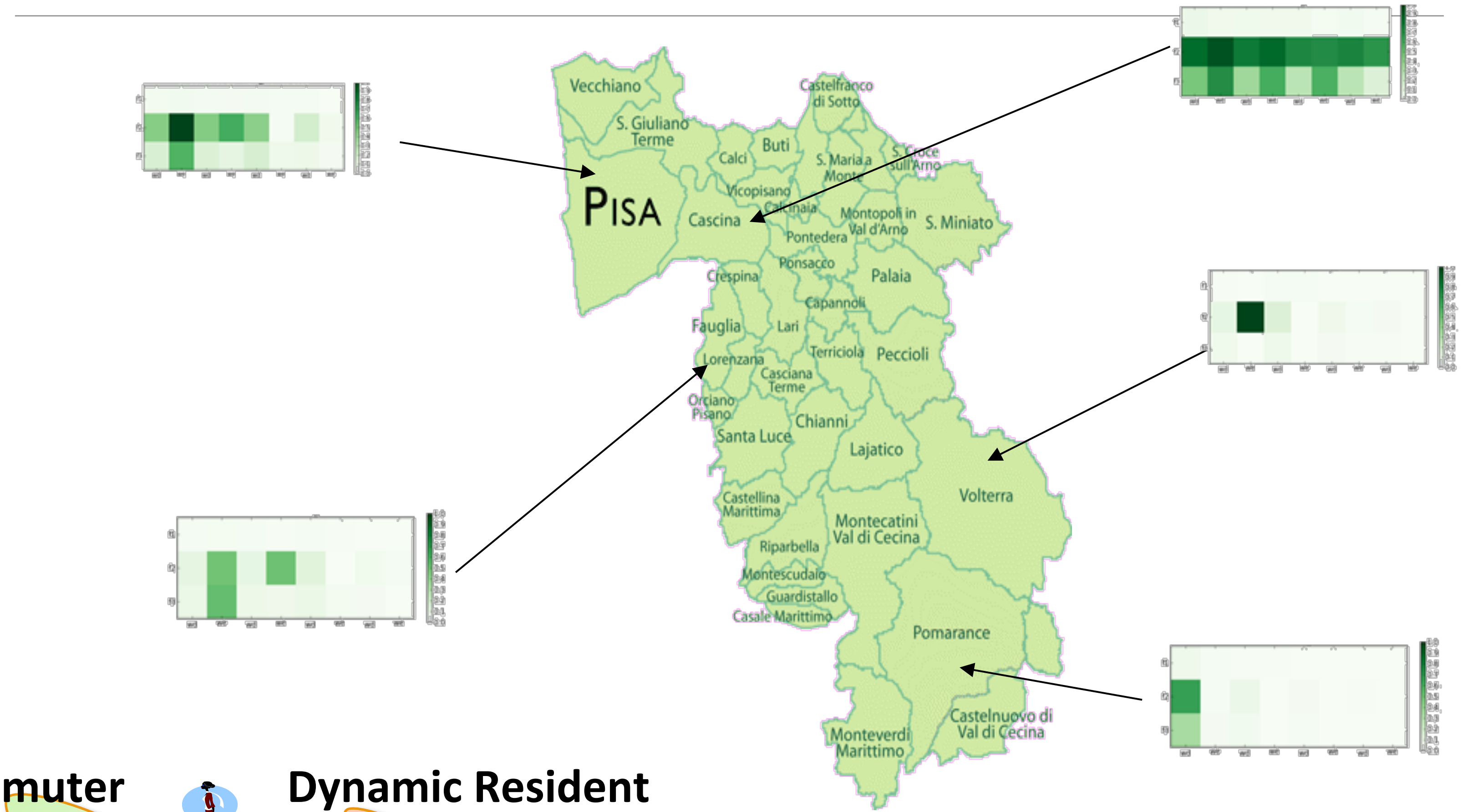
Sociometer: the city user meter



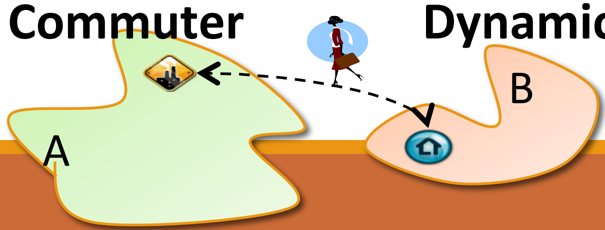
Pisa, January 2012

Analysis of GSM calls data for understanding user mobility behavior
B Furletti, L Gabrielli, C Renso, S Rinzivillo Big Data, 2013 IEEE International Conference on, 550-555

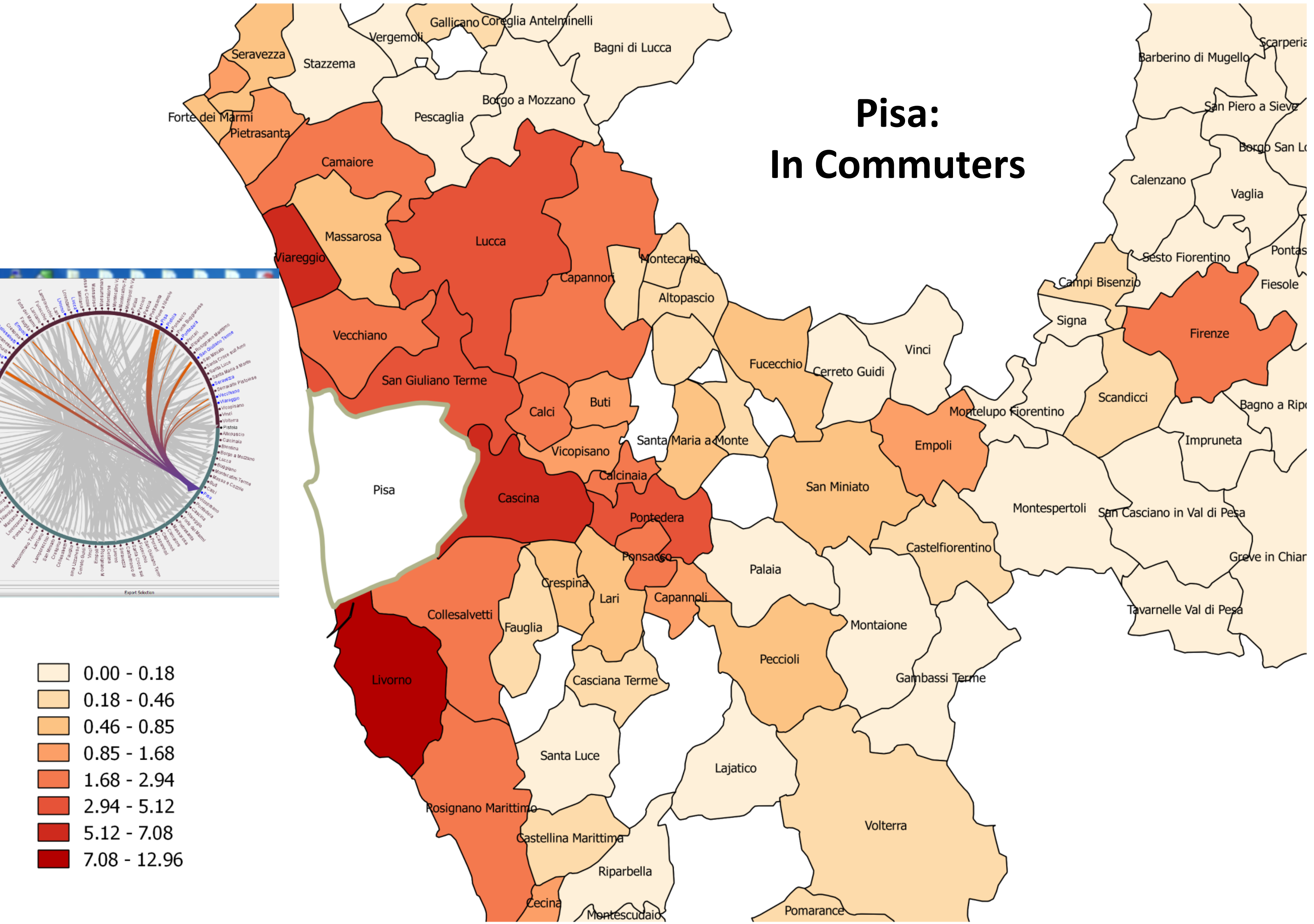
The many profiles of an individual



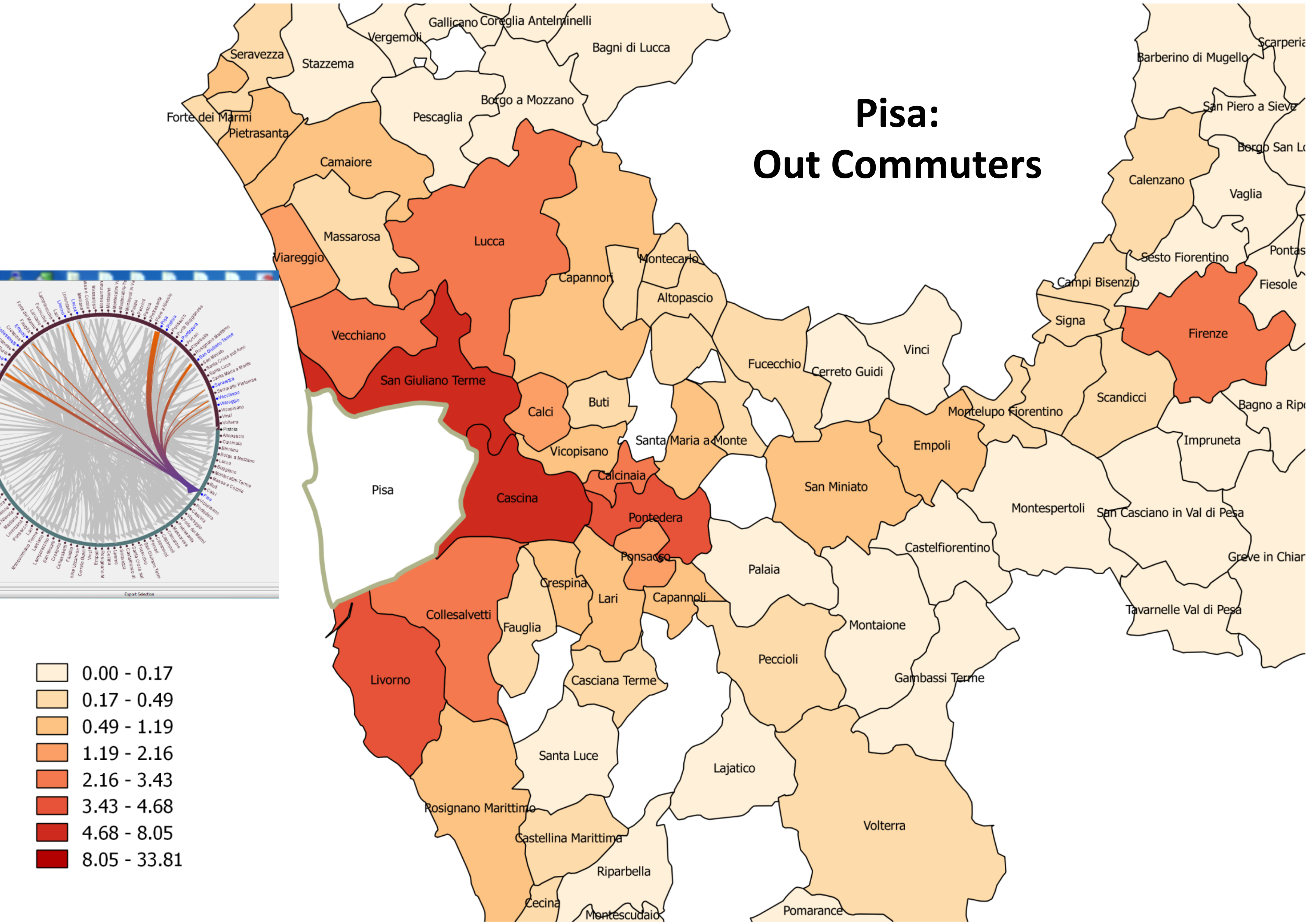
Commuter **Dynamic Resident**



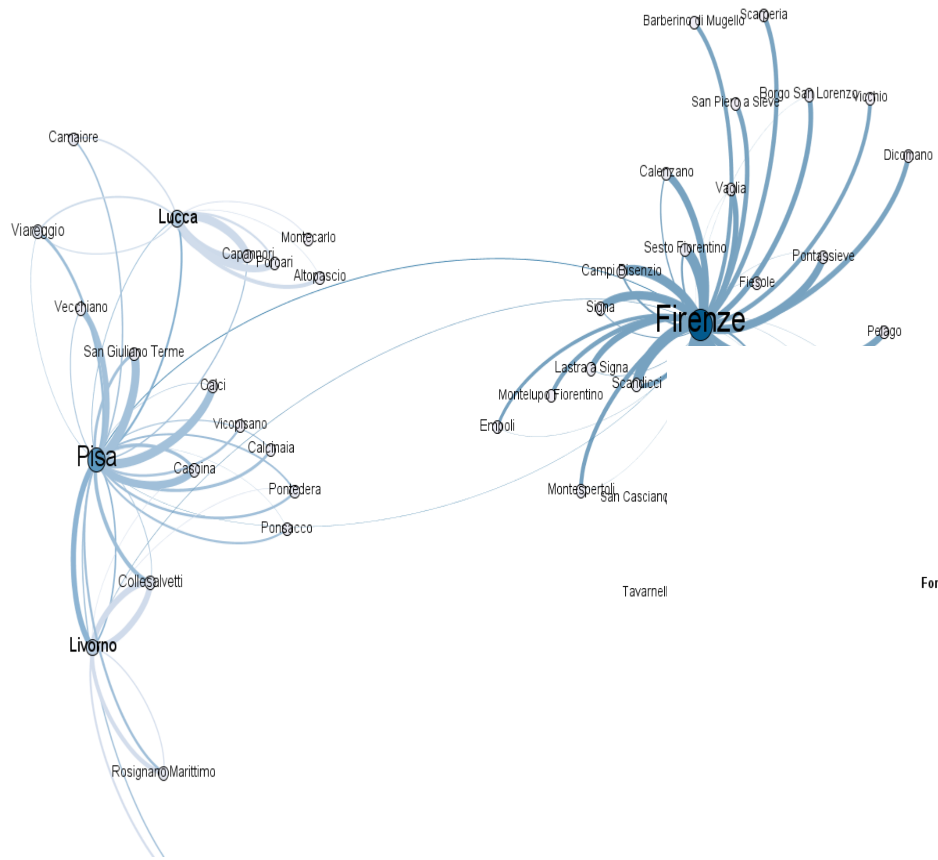
Pisa: In Commuters



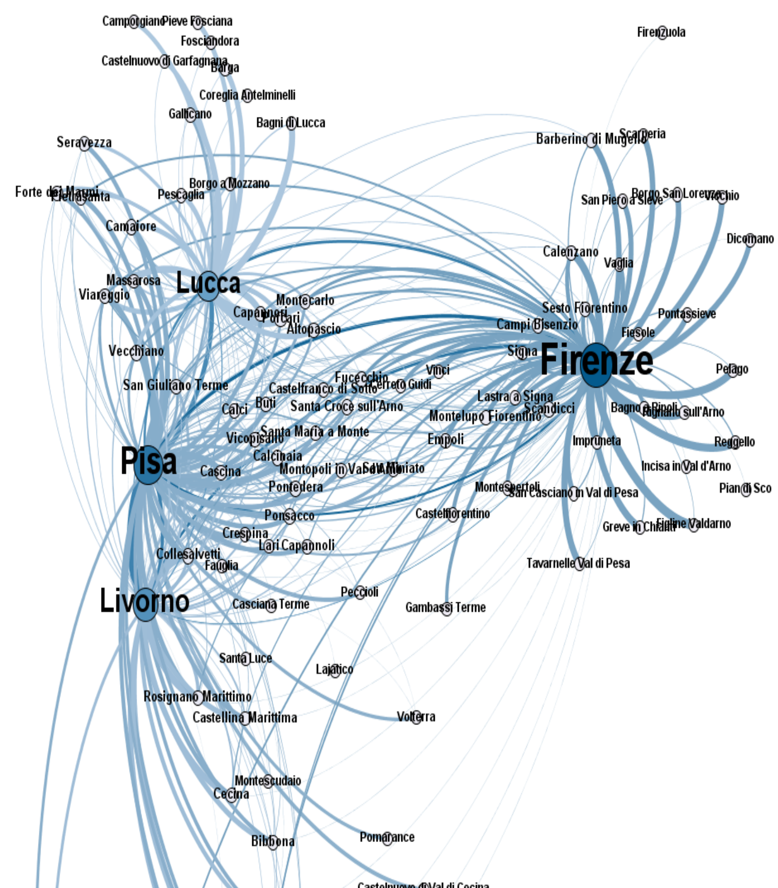
Pisa: Out Commuters



INTER-CITY FLOWS, WITH SEMANTICS

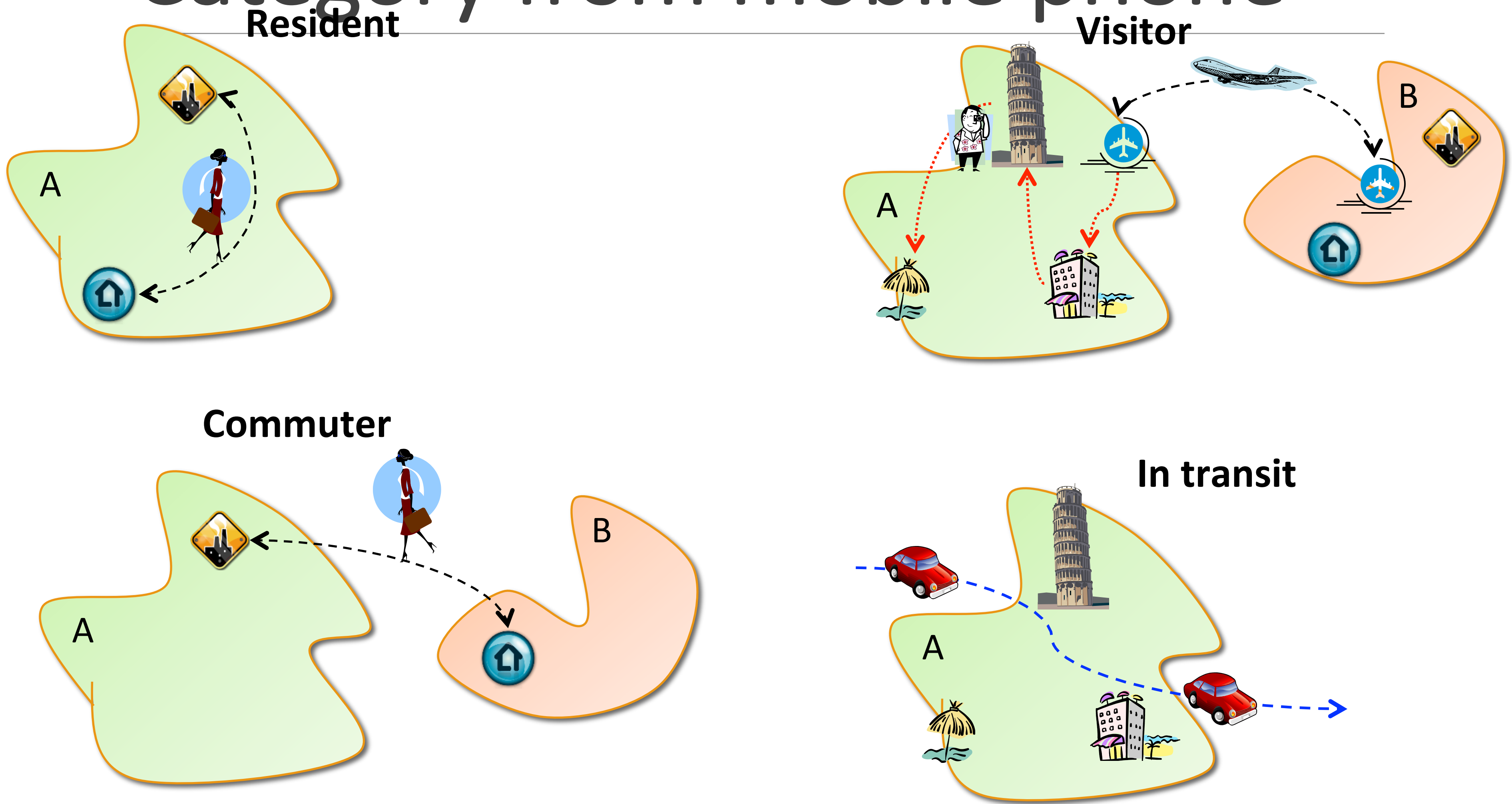


COMMUTER NETWORK

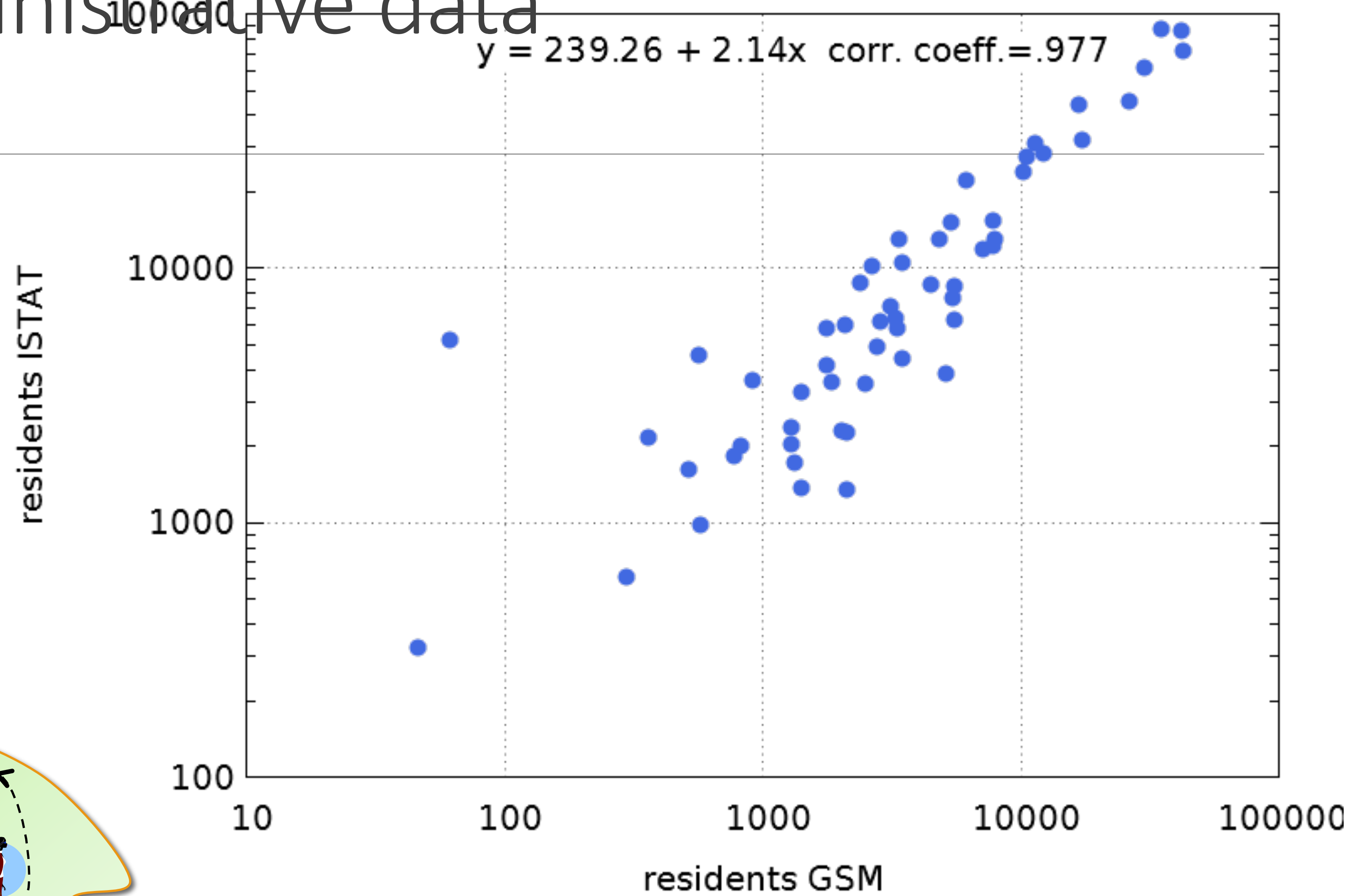


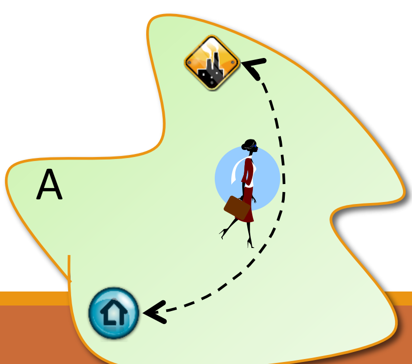
VISITOR NETWORK

Sociometer: Estimating User Category from mobile phone



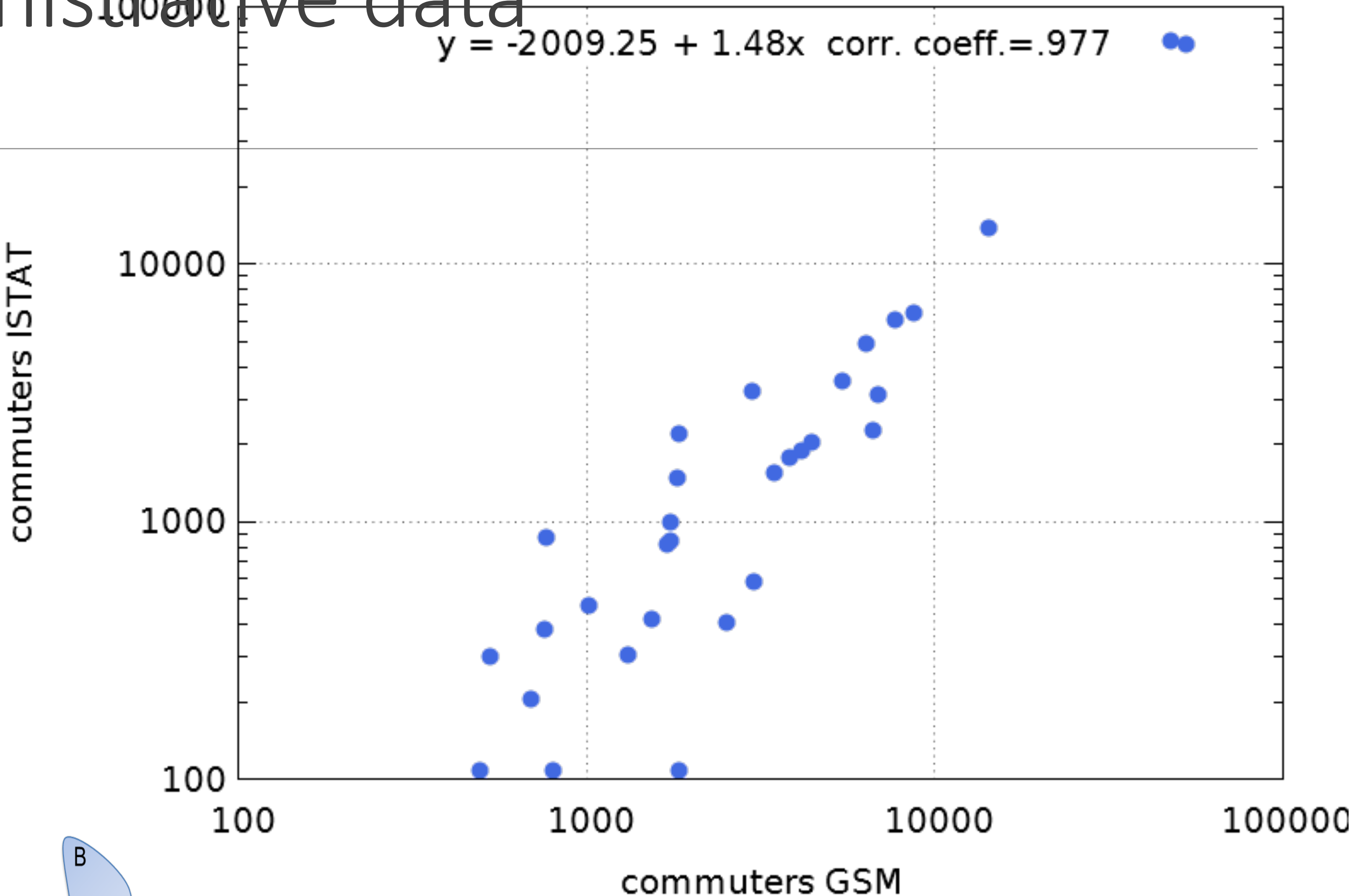
Residents – validation with administrative data

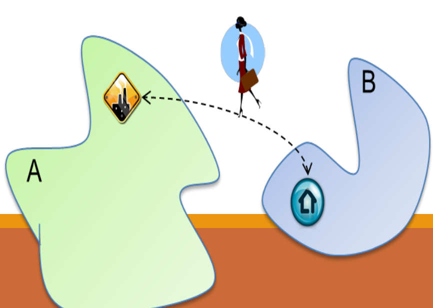




Join work with ISTAT: Barbara Furletti, Lorenzo Gabrielli, Giuseppe Garofalo, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio. Use of mobile phone data to estimate mobility flows. Measuring urban population and intercity mobility using big data in an integrated approach. Italian Symposium on Statistics (2014).

Commuters - Validation with administrative data

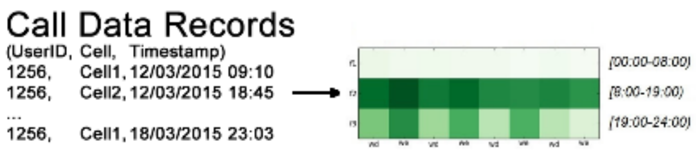




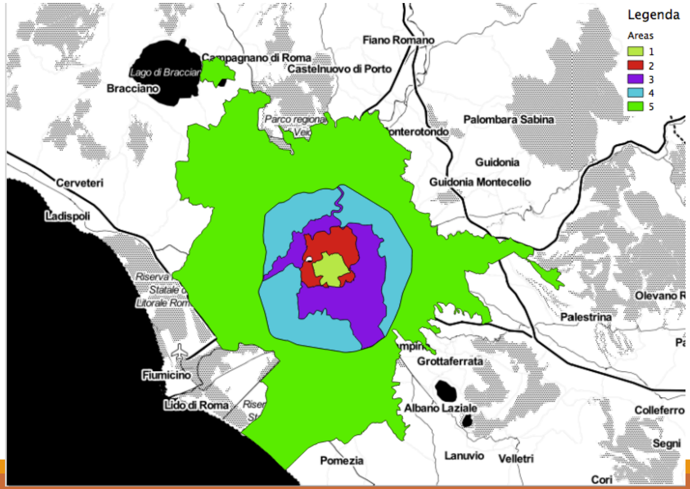
Join work with ISTAT: Barbara Furletti, Lorenzo Gabrielli, Giuseppe Garofalo, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio. Use of mobile phone data to estimate mobility flows. Measuring urban population and intercity mobility using big data in an integrated approach. Italian Symposium on Statistics (2014).

Measuring exceptional events

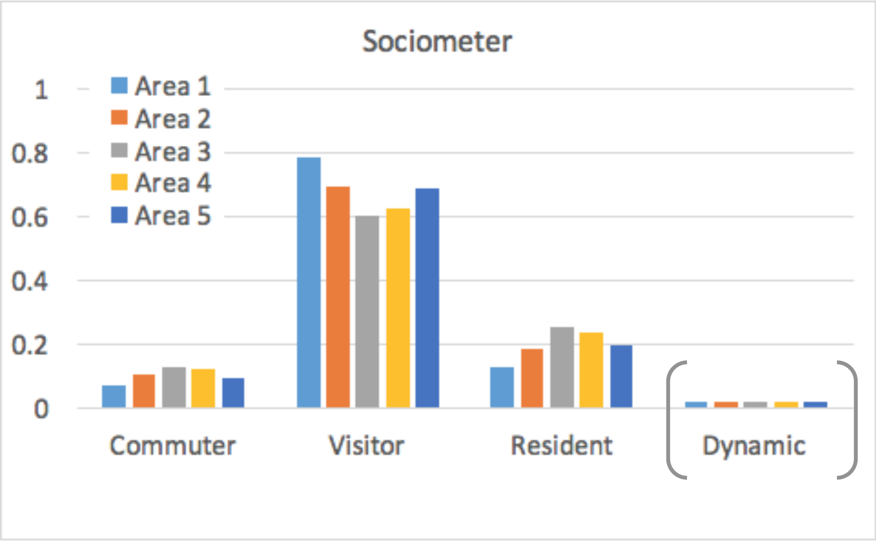
- Presences during Jubilee in Rome (December 2015)
- Continuous monitoring



+



=



San Pietro Square