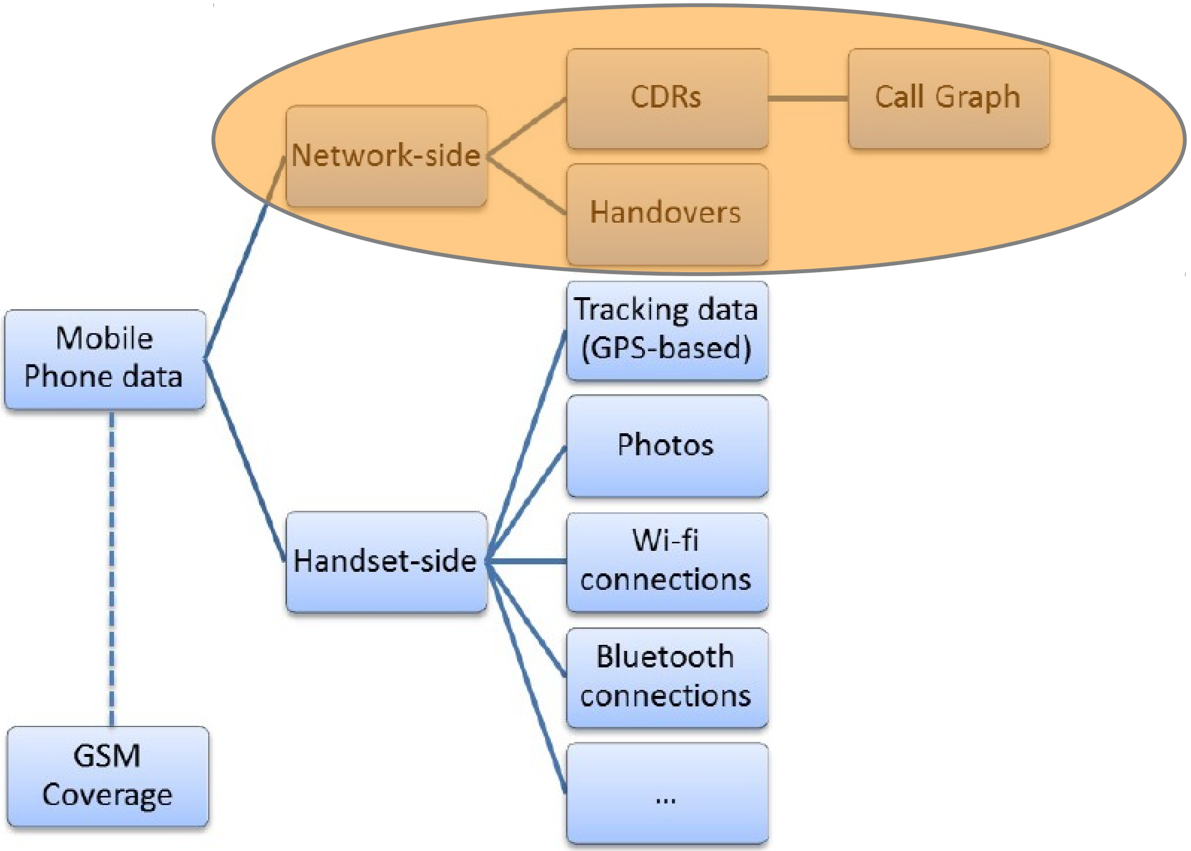
Mobility Data Mining

Mobility Analytics on Mobile Phone data

What are GSM data

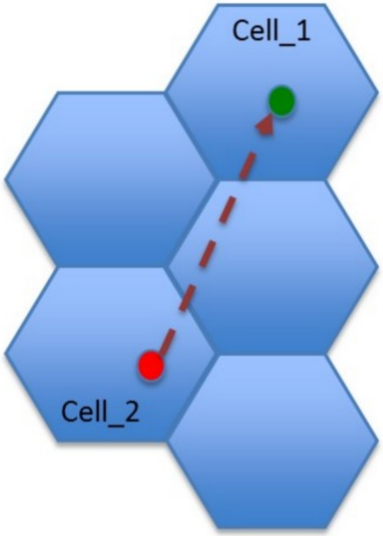
- Most popular resource for mobile phone data
- In principle, several kinds of data



GSM data types

CDR

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

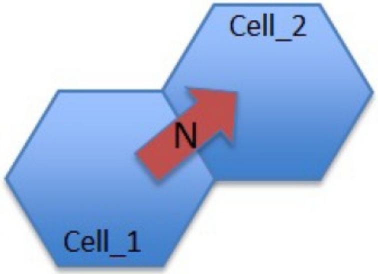


Call Graph

Who calls **whom** and **when**

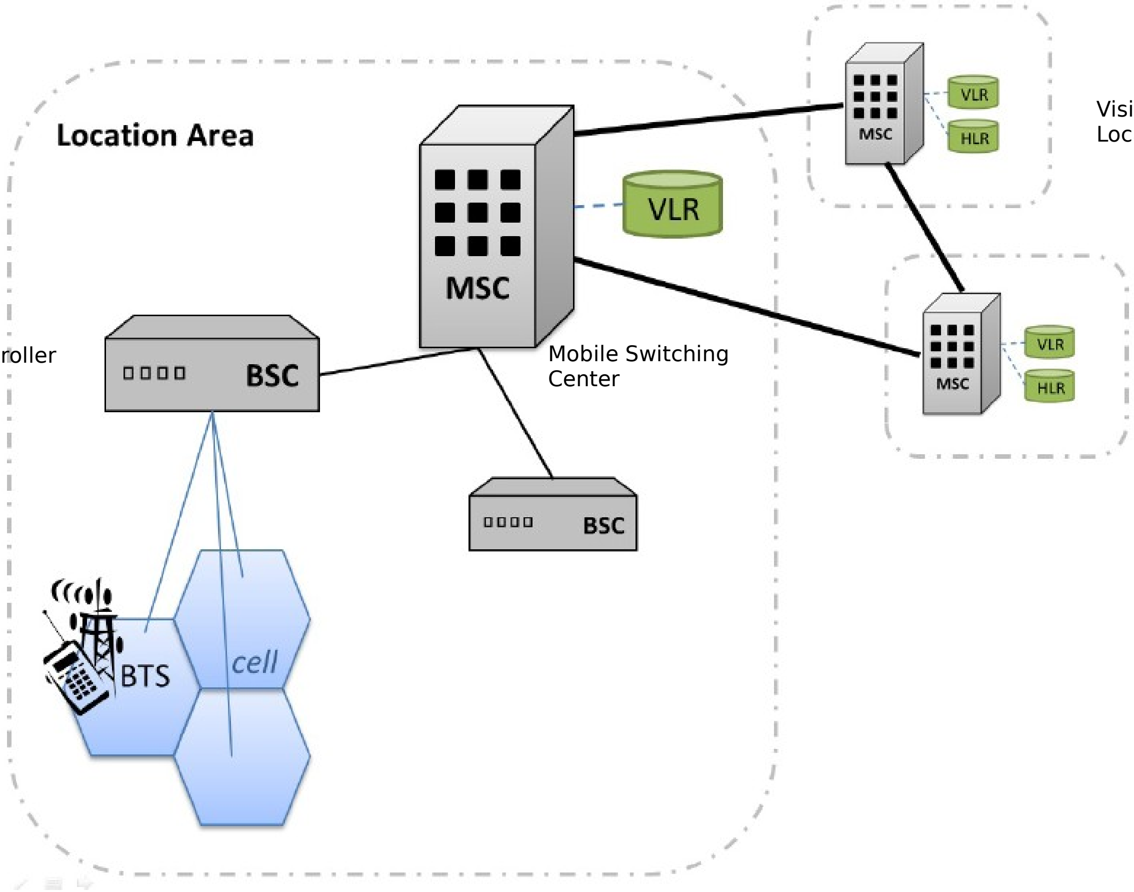
Hand over

Inter-cell flow **counts**



GSM infrastructure

- Aimed at providing voice/data telecom.



Visitor/Home Location Register

Base Station Controller

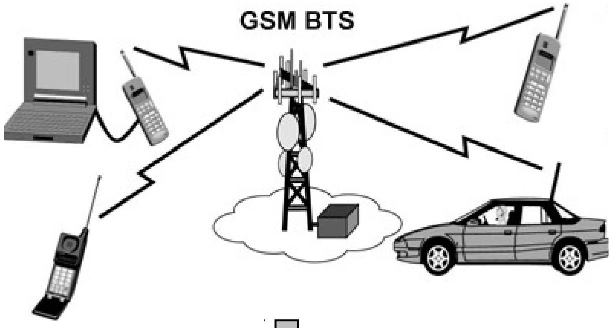
Mobile Switching Center

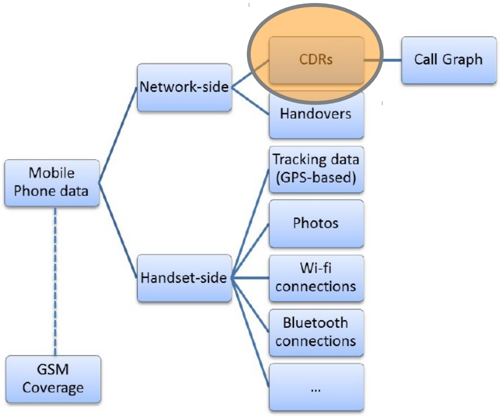
Base Transceiver Station

GSM data - Description

Call Data Record (CDR)

Data gathered from mobile phone operator for billing purpose

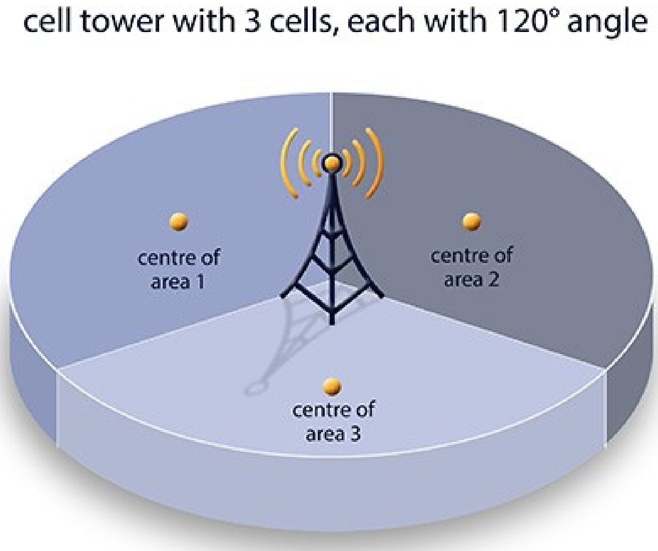




User id	Time start	Cell start	Cell end	Duration
10294595	"2014-02-20 14:24:58"	"PI010U2"	"PI010U1"	48
10294595	"2014-02-20 18:50:22"	"PI002G1"	"PI010U2"	78
10294595	"2014-02-21 09:19:51"	"PI080G1"	"PI016G1"	357

GSM data - Description

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



Pros and cons of using GSM data

Pros

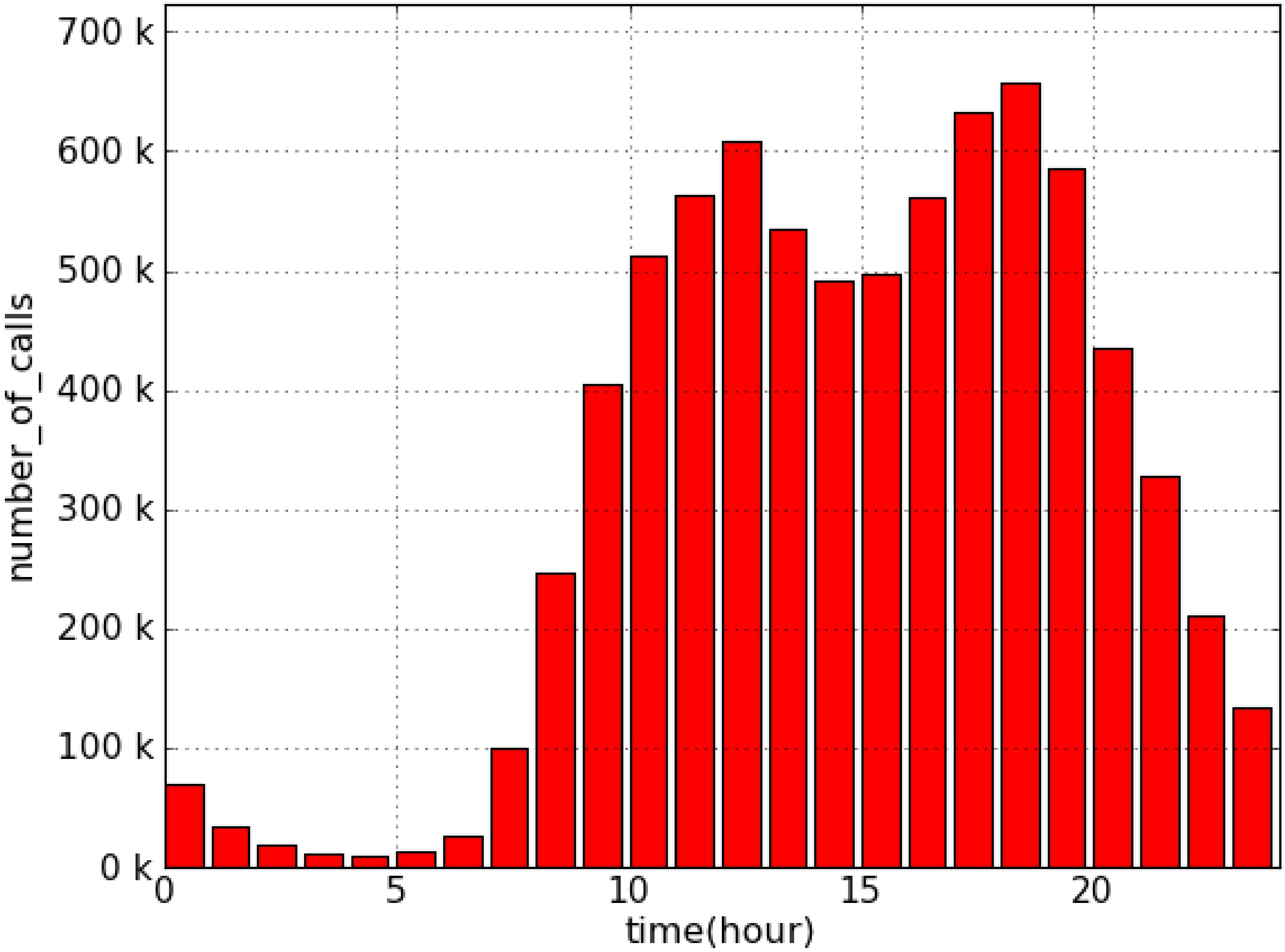
- Passive sensing: does not require an active contribution of the users
- Contains huge amount of information of how, when, with whom we communicate
- Same data format in all the world

Cons

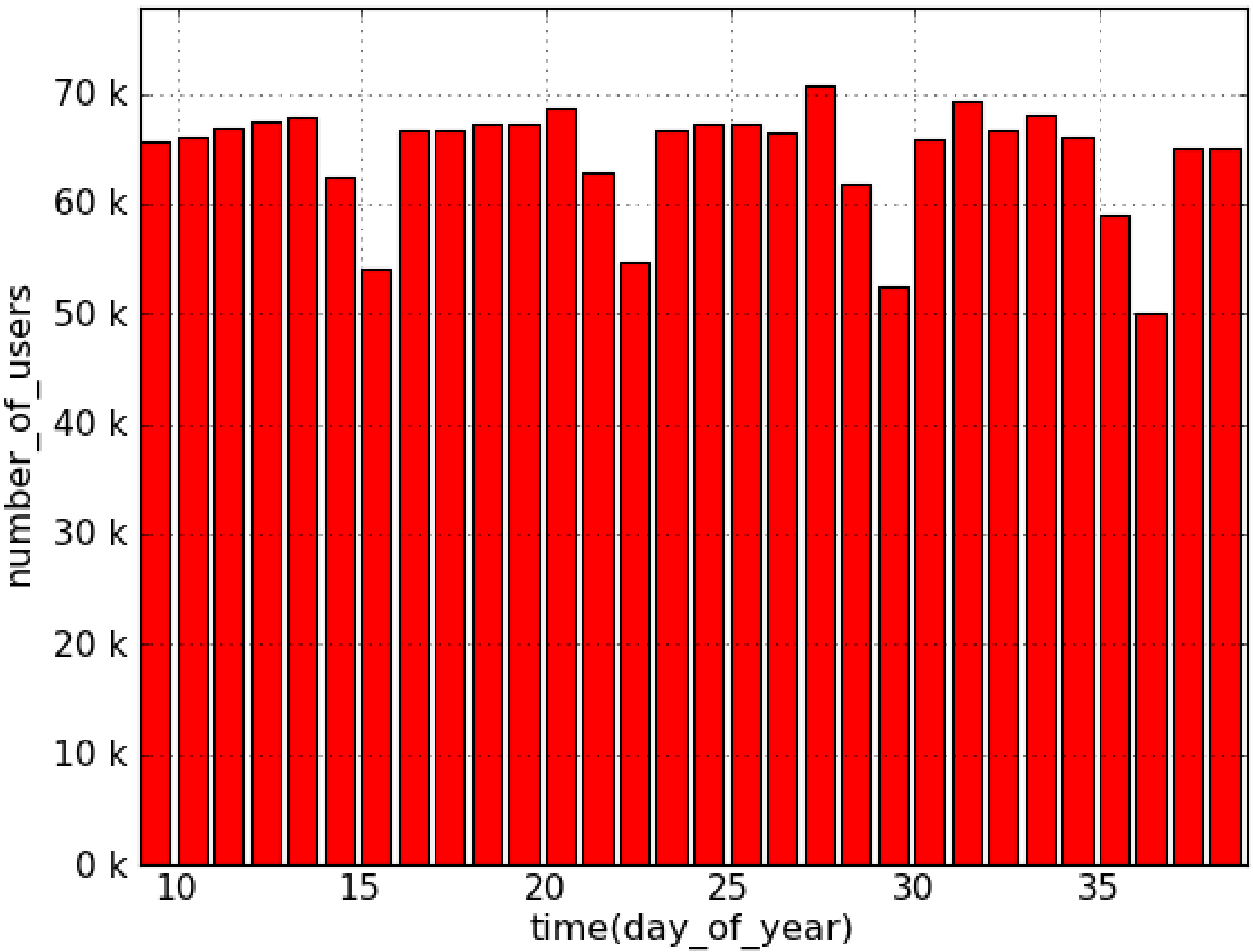
- Poor demographic and economic data
- Privacy concern: different legislations for different countries
- Low sampling: few events of calls for a considerable amount of users

Simple CDR-based statistics

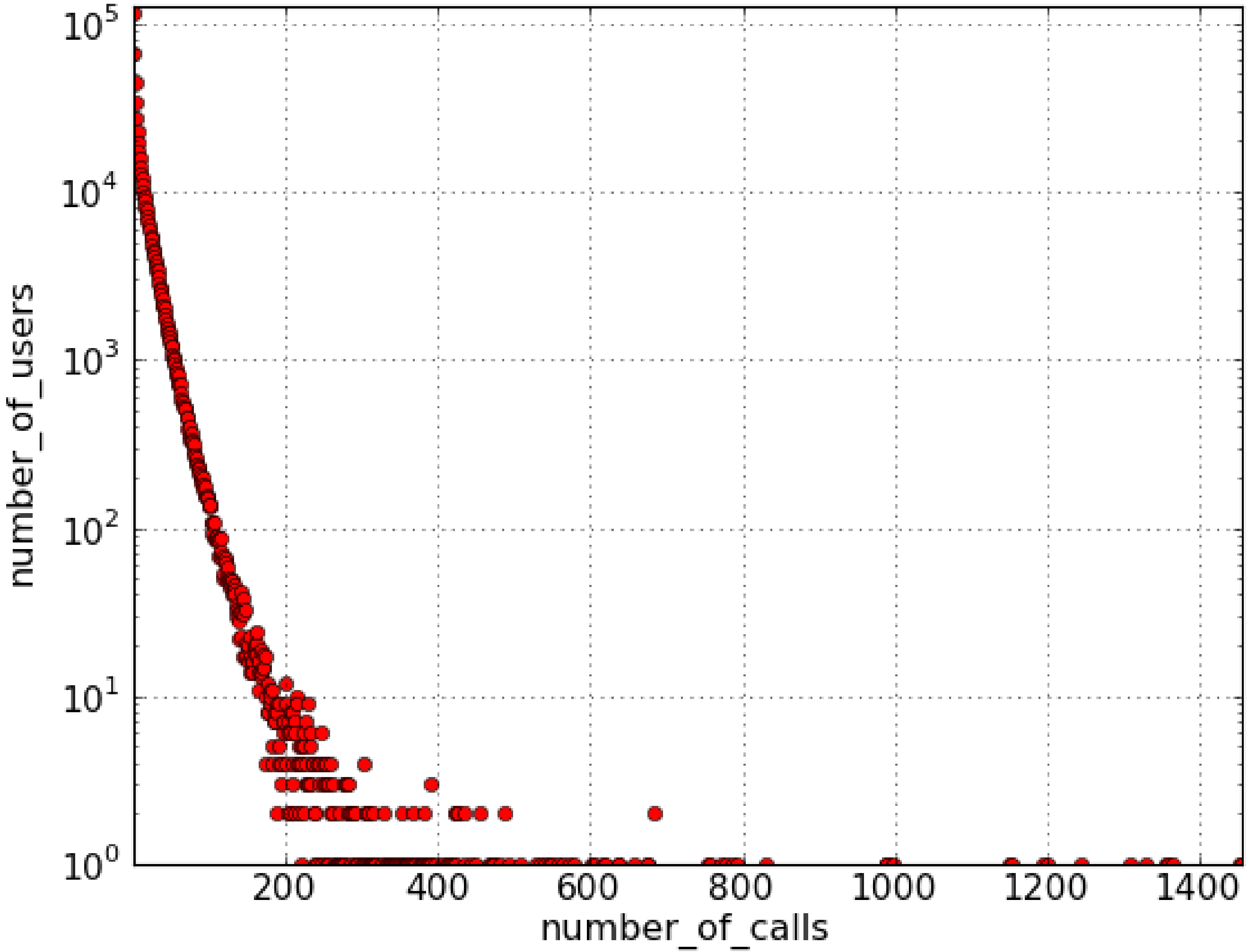
Daily pattern behavior



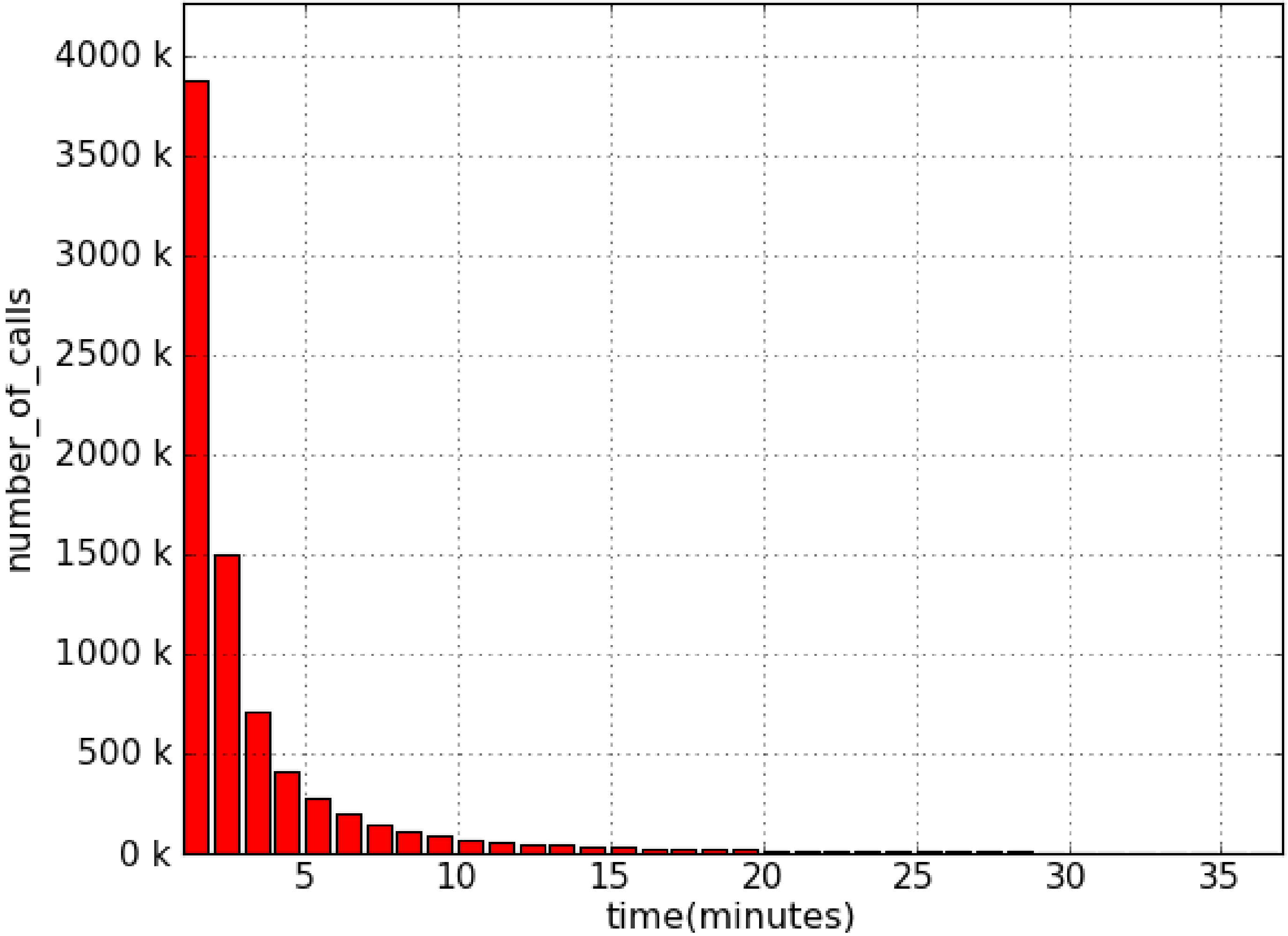
Weekly pattern behavior



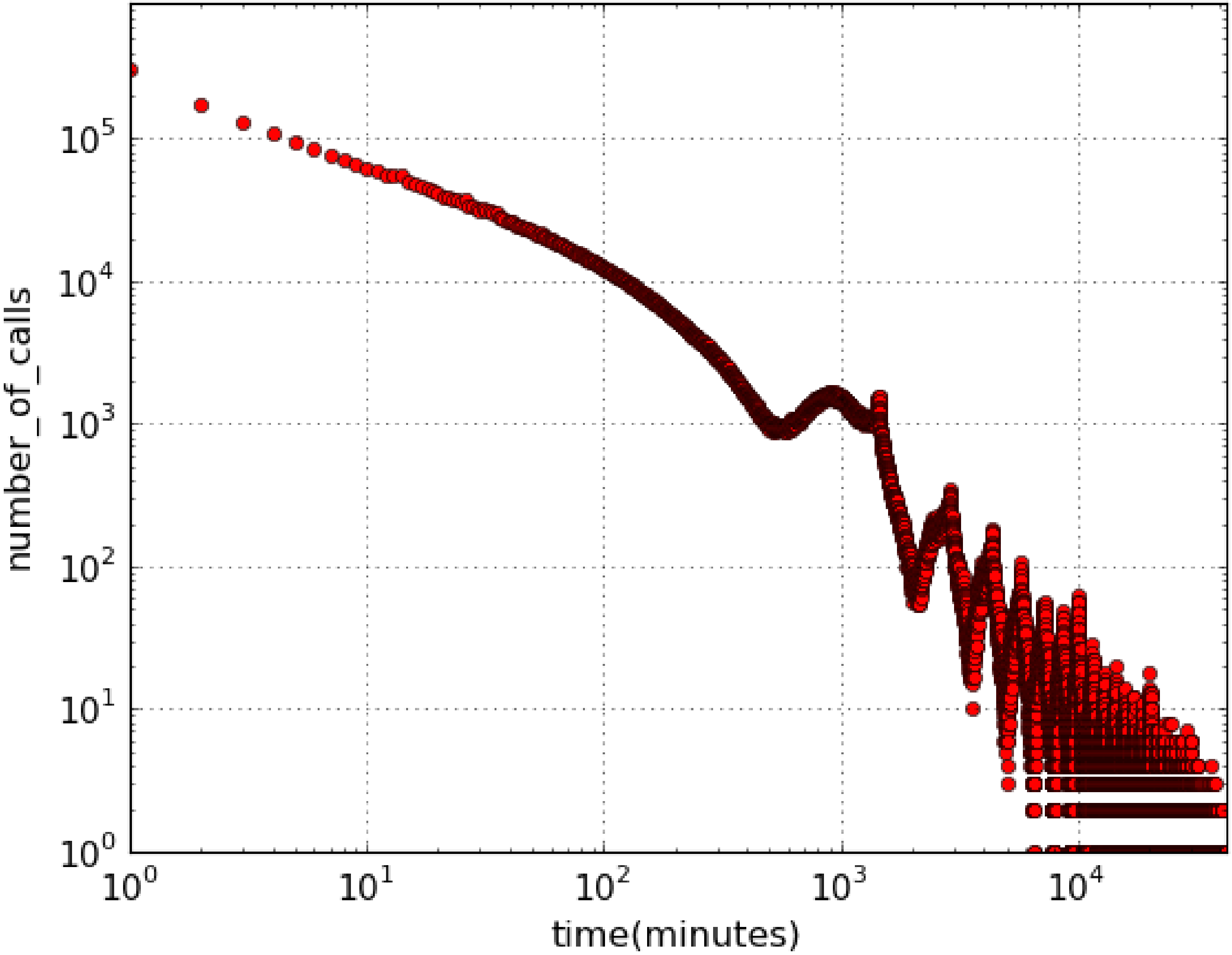
How many times we call?



How long we talk on the phone?



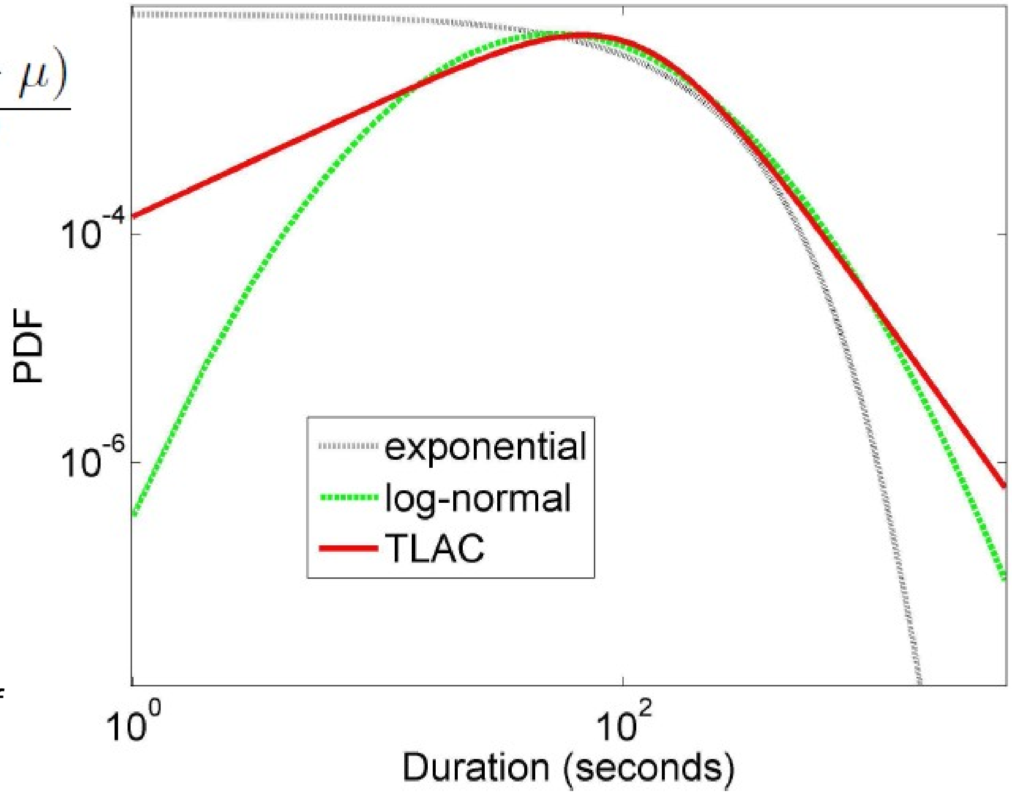
How many minutes goes by a call to the next?



Theoretical model of call durations

- Truncated Lazy Contractor (TLAC)

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

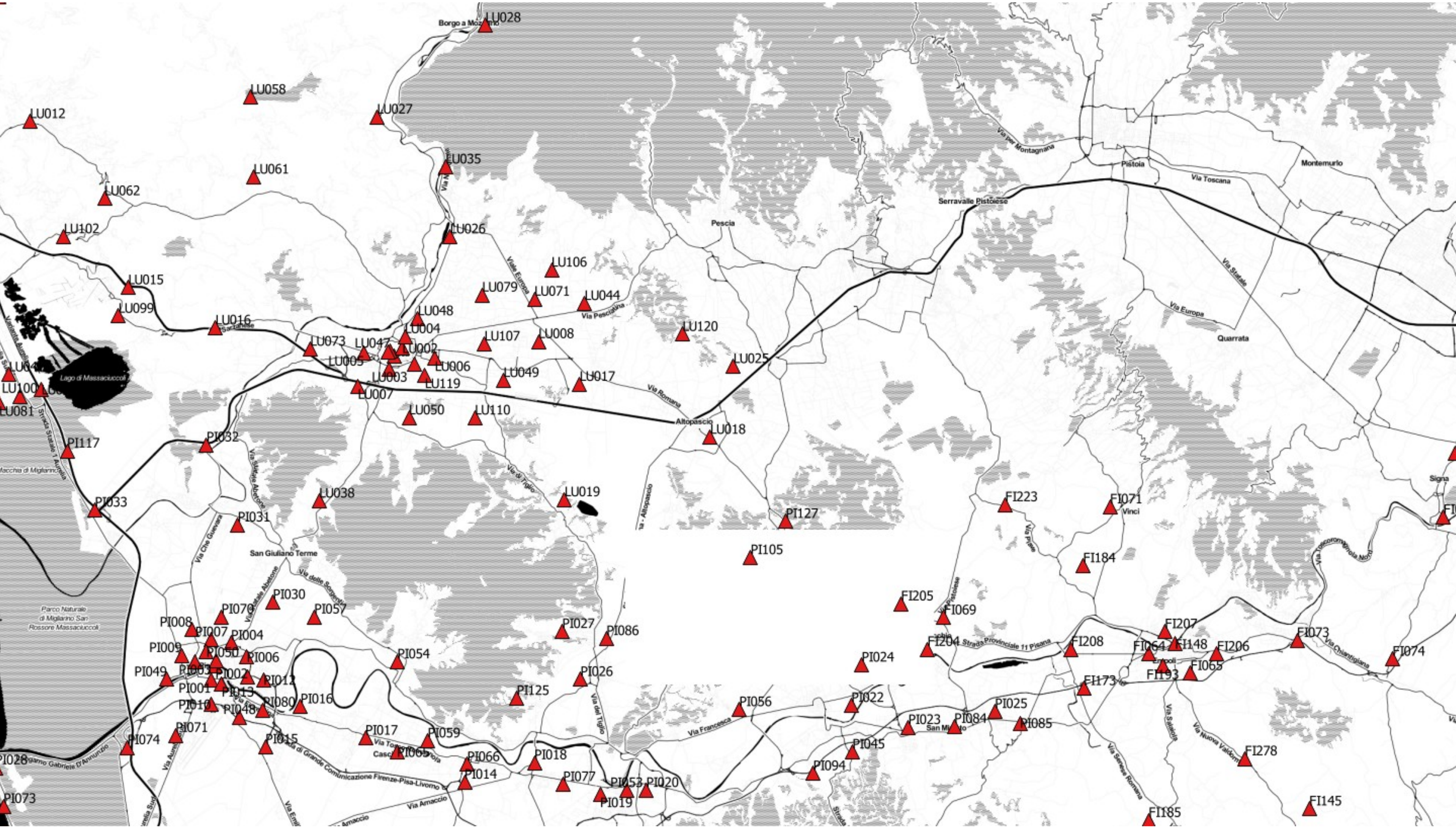


de Melo-Akoglu-Faloutsos-Loureiro.
Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. ECML PKDD 2010.

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

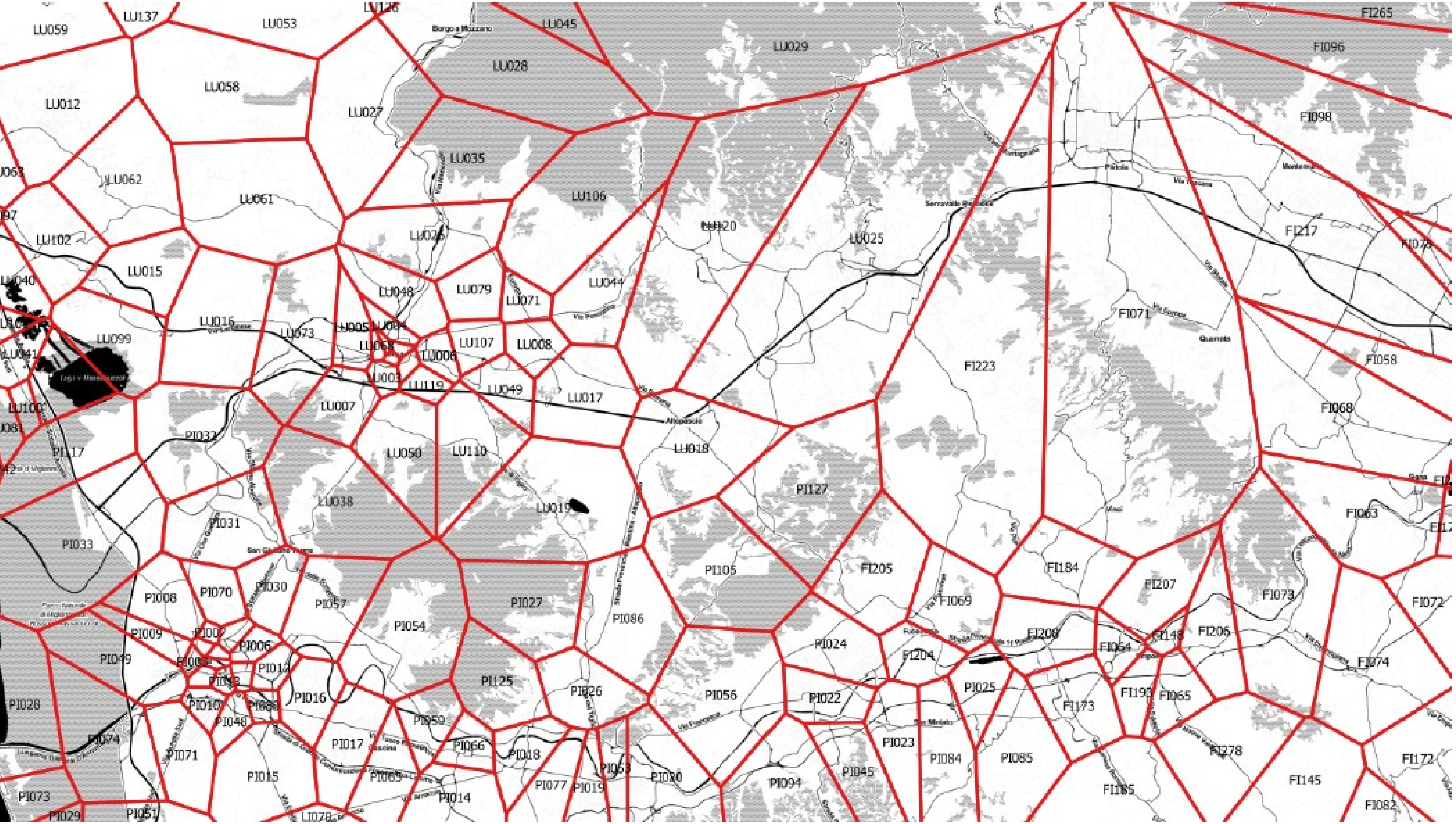
From CDR to Geography:
CDRs describe where the calls started

Antennas

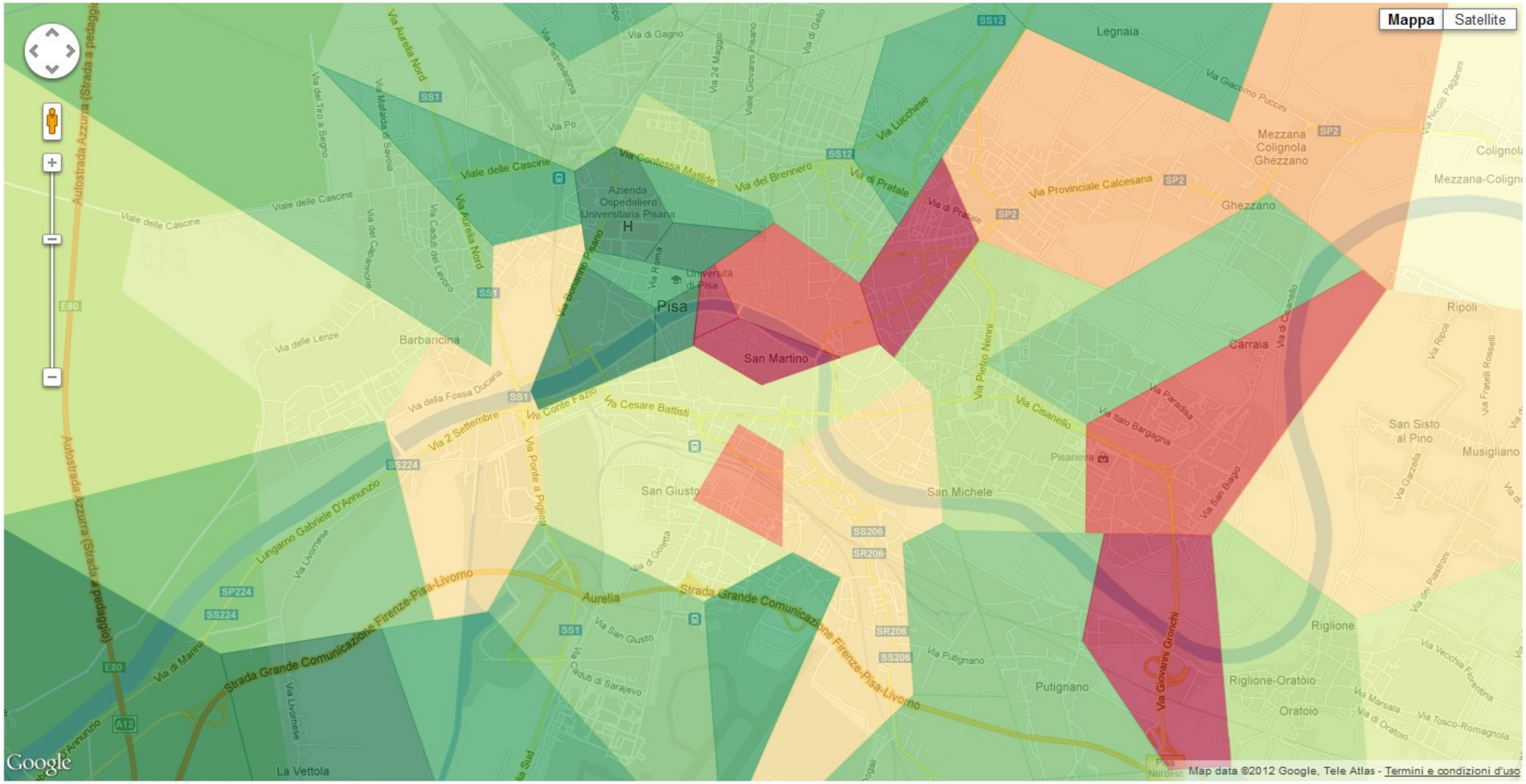


From CDR to Geography:
CDRs describe where the calls started

Voronoi tesselation



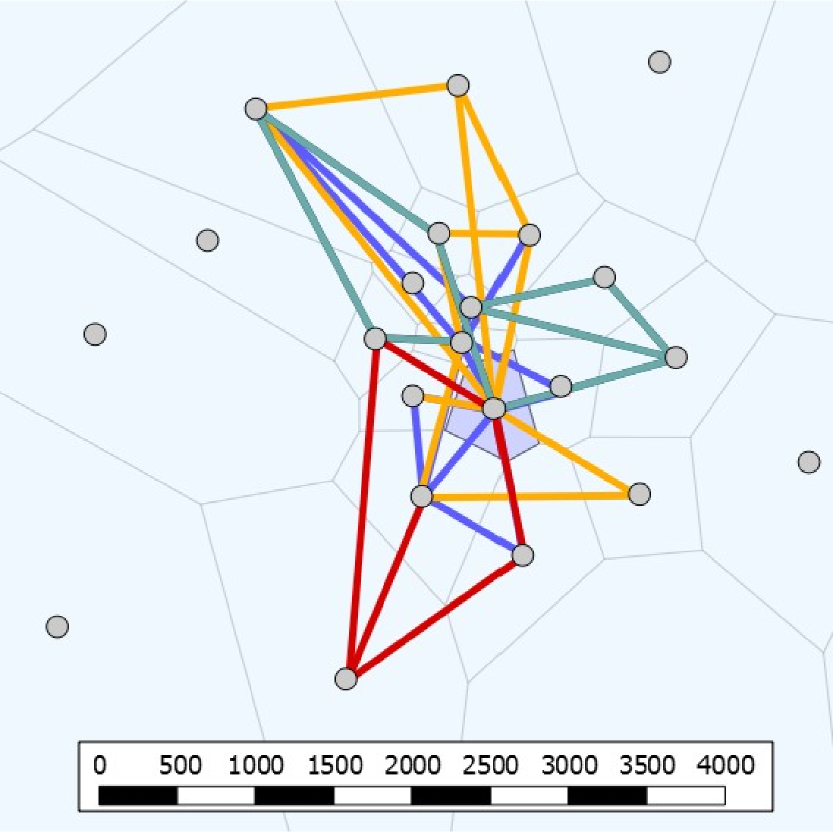
Spatial distribution of calls



High presences of people within the working area of Pisa

Observing the **mobility** of individuals

Mobility Behaviours



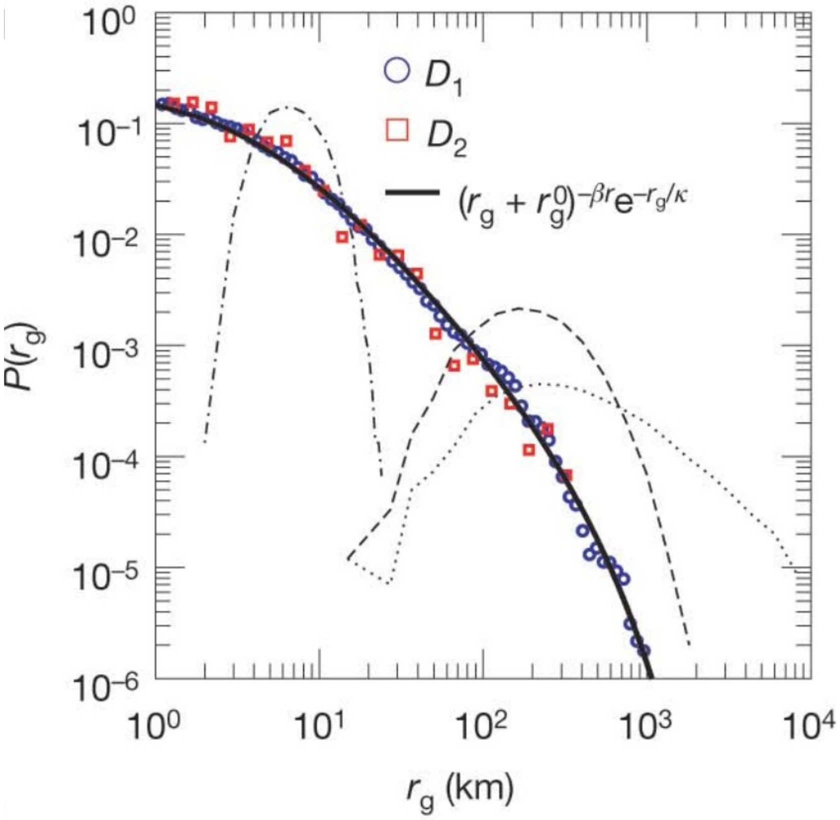
From CDR to how users move within a territory

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

Characteristic distance traveled by an individual

radius of gyration produces heavy tails

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

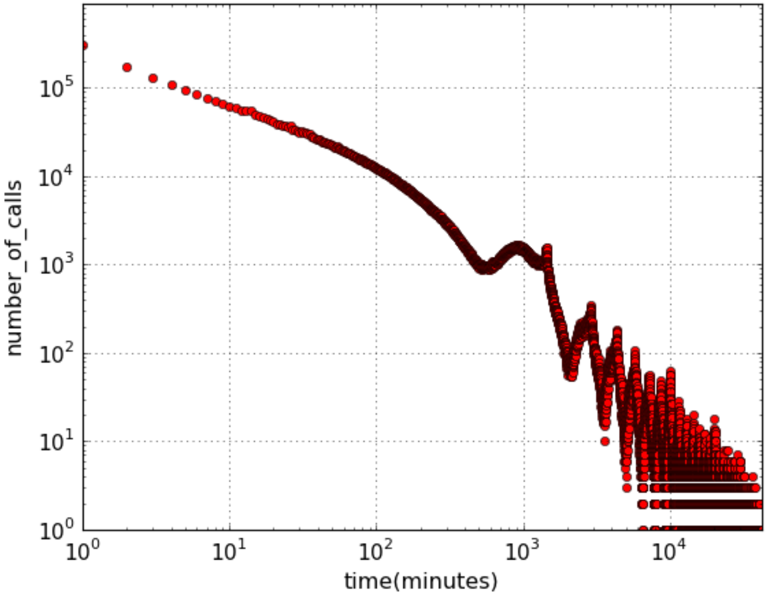


Understanding individual human mobility patterns. Gonzalez, Hidalgo, Barabási. *Nature 453(7196):779--782* (*June 2008*)

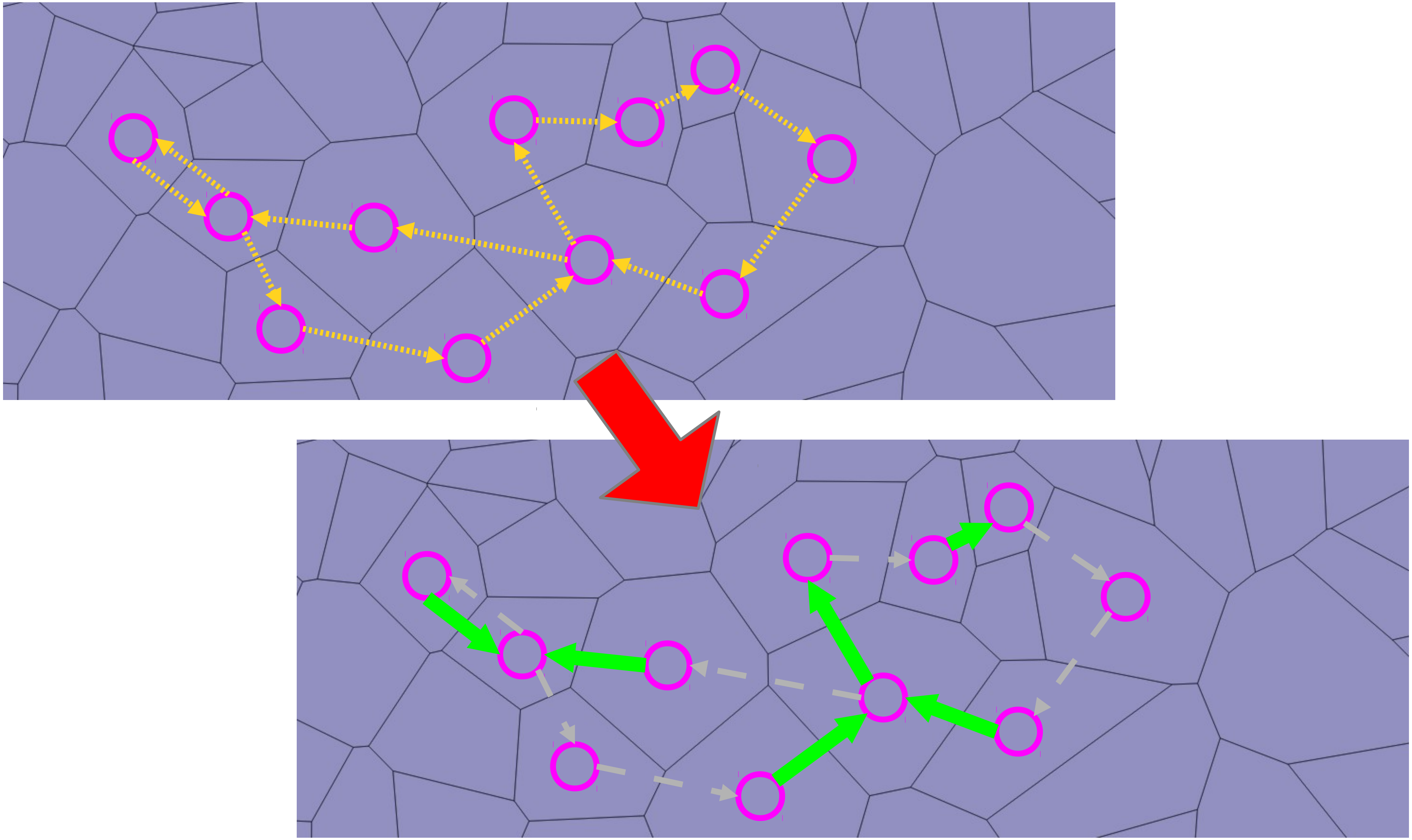
Estimating movements

- 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

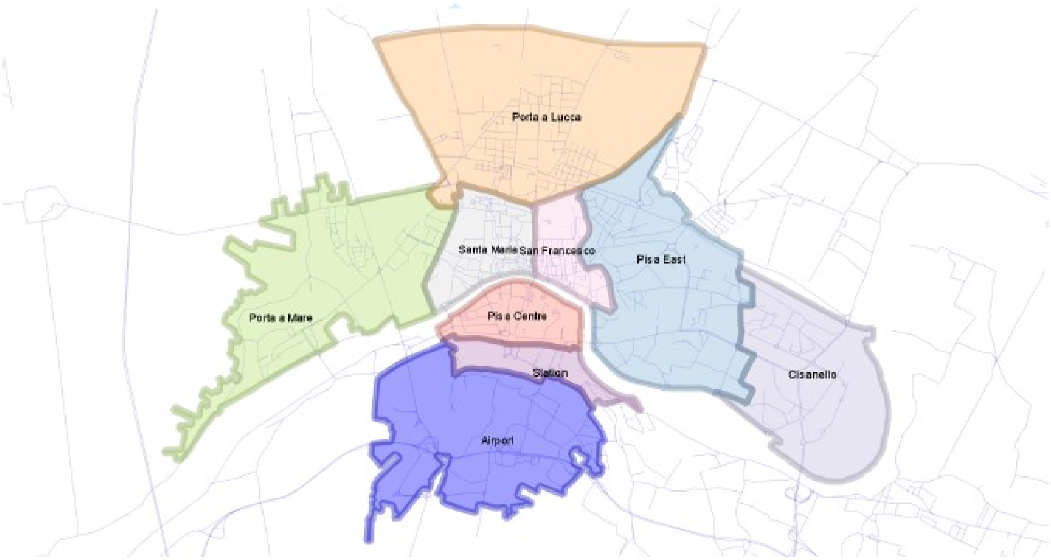


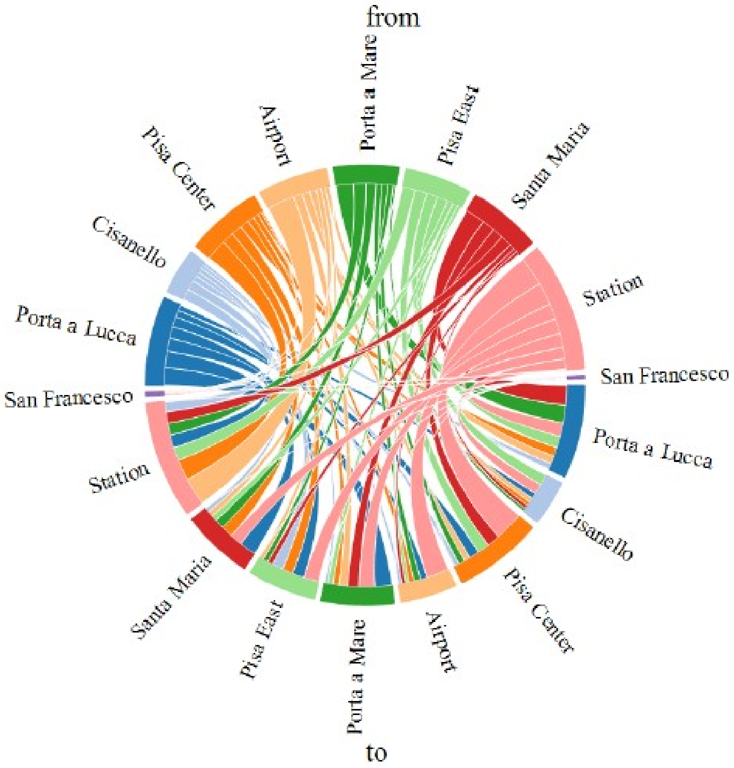
Estimating movements



Estimating movements

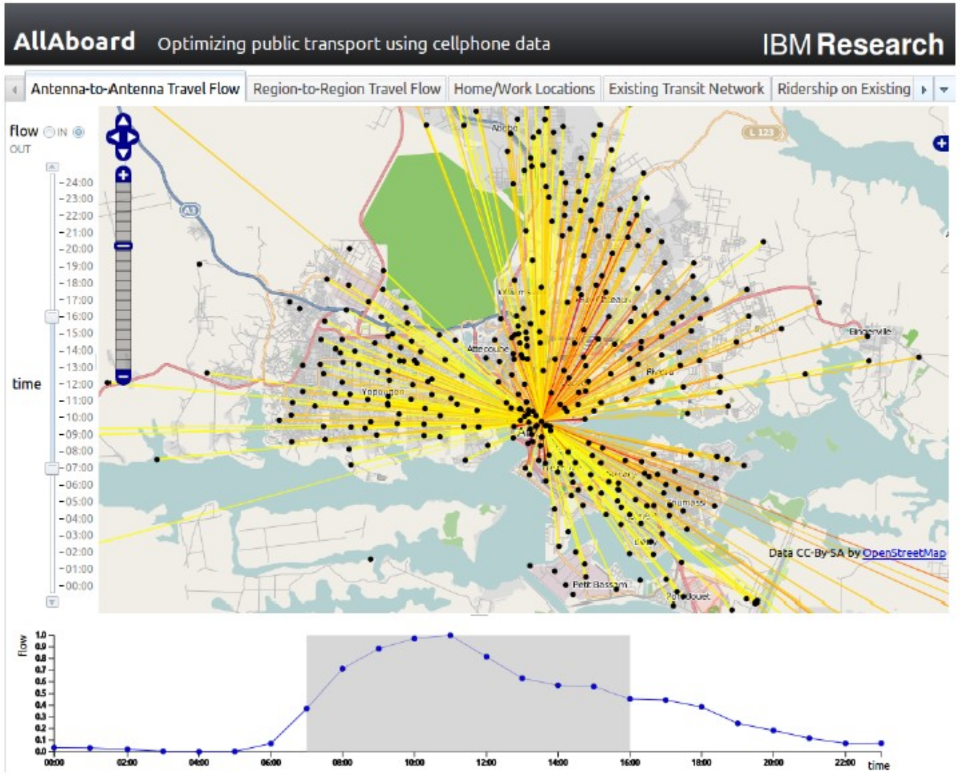
- Example on Pisa city

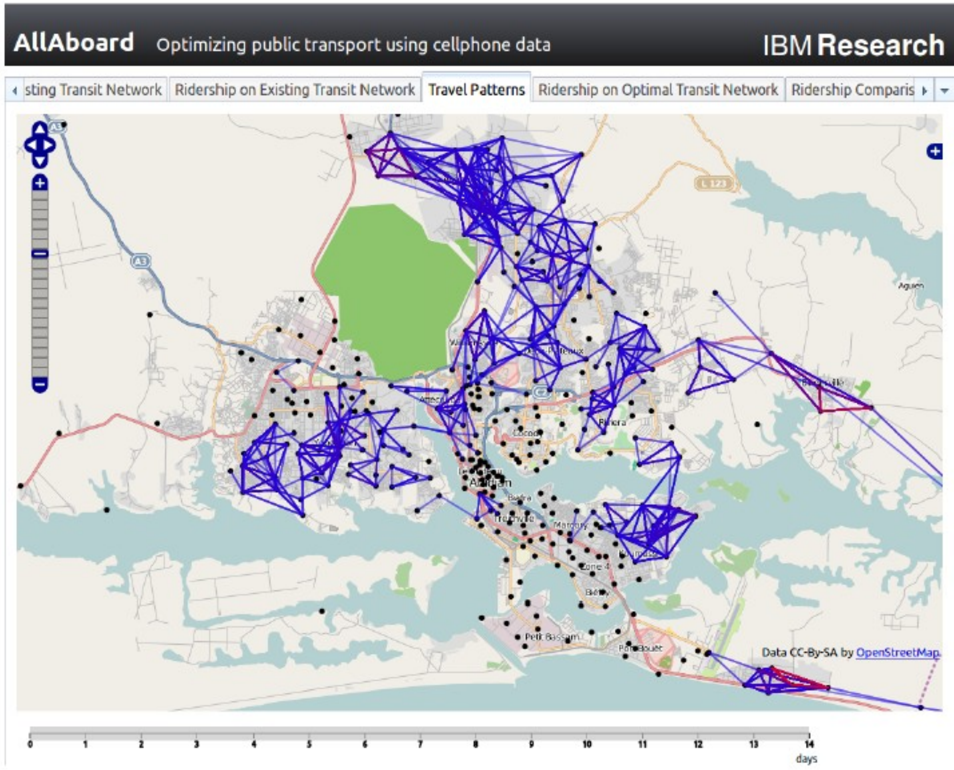




Estimating movements

- Example on Abidjan (Ivory Coast)





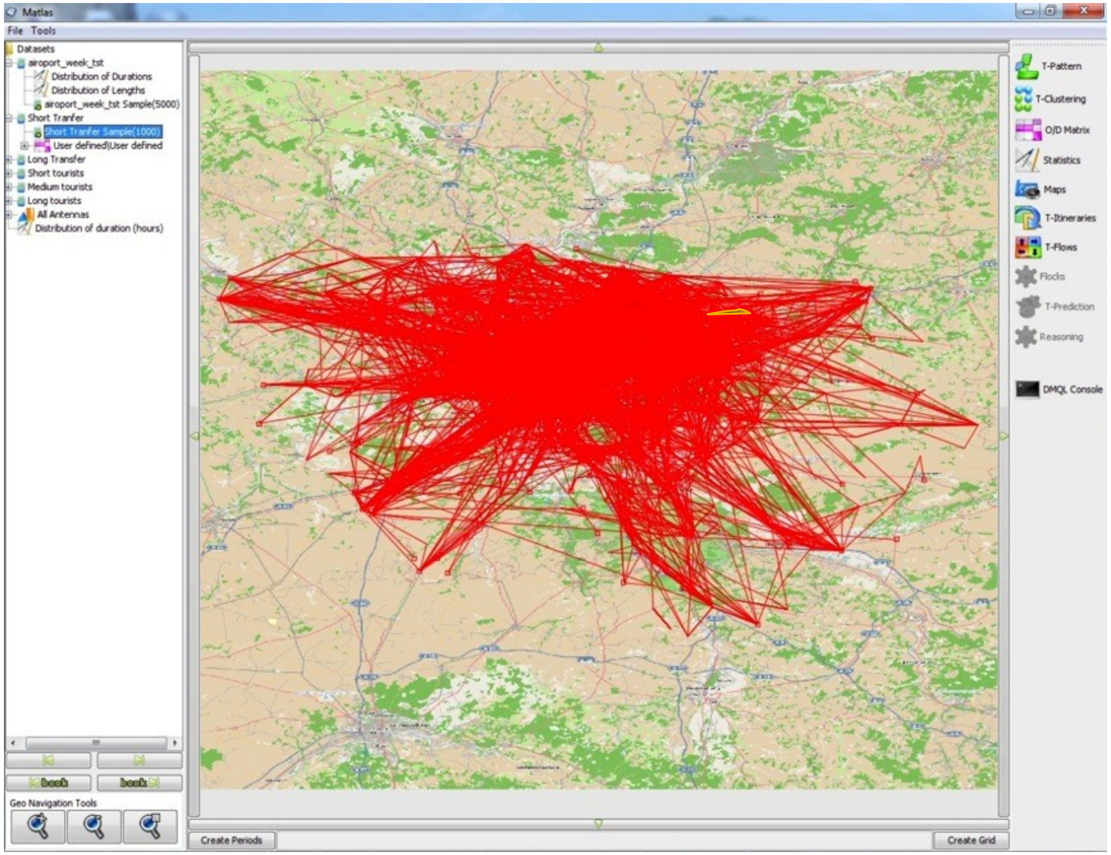
AllAboard
Optimizing public transport using cellphone data
IBM Research
sting Transit Network
Ridership on Existing Transit Network
Travel Patterns
Ridership on Optimal Transit Network
Ridership Comparis
Abidjan
Data CC-By-SA by OpenStreetMap
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
days

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

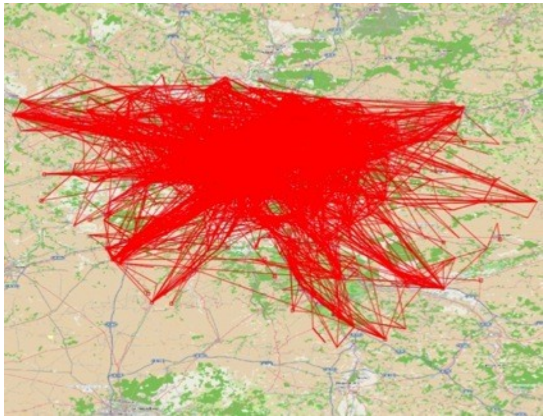
Sample application: Analyzing tourist data

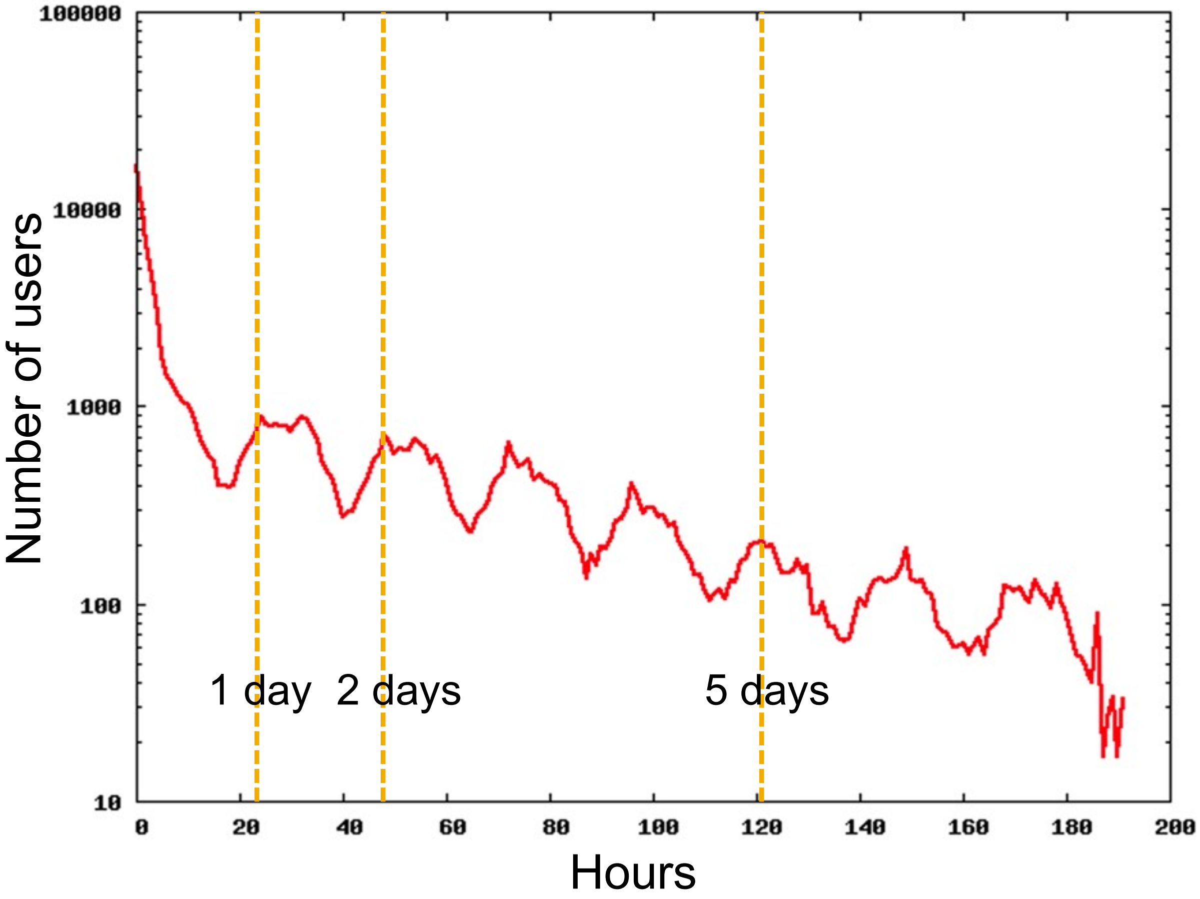
- Case study of foreign (roaming) visitors of Paris area
- Users arriving and leaving at CDG airport

106 000 Users

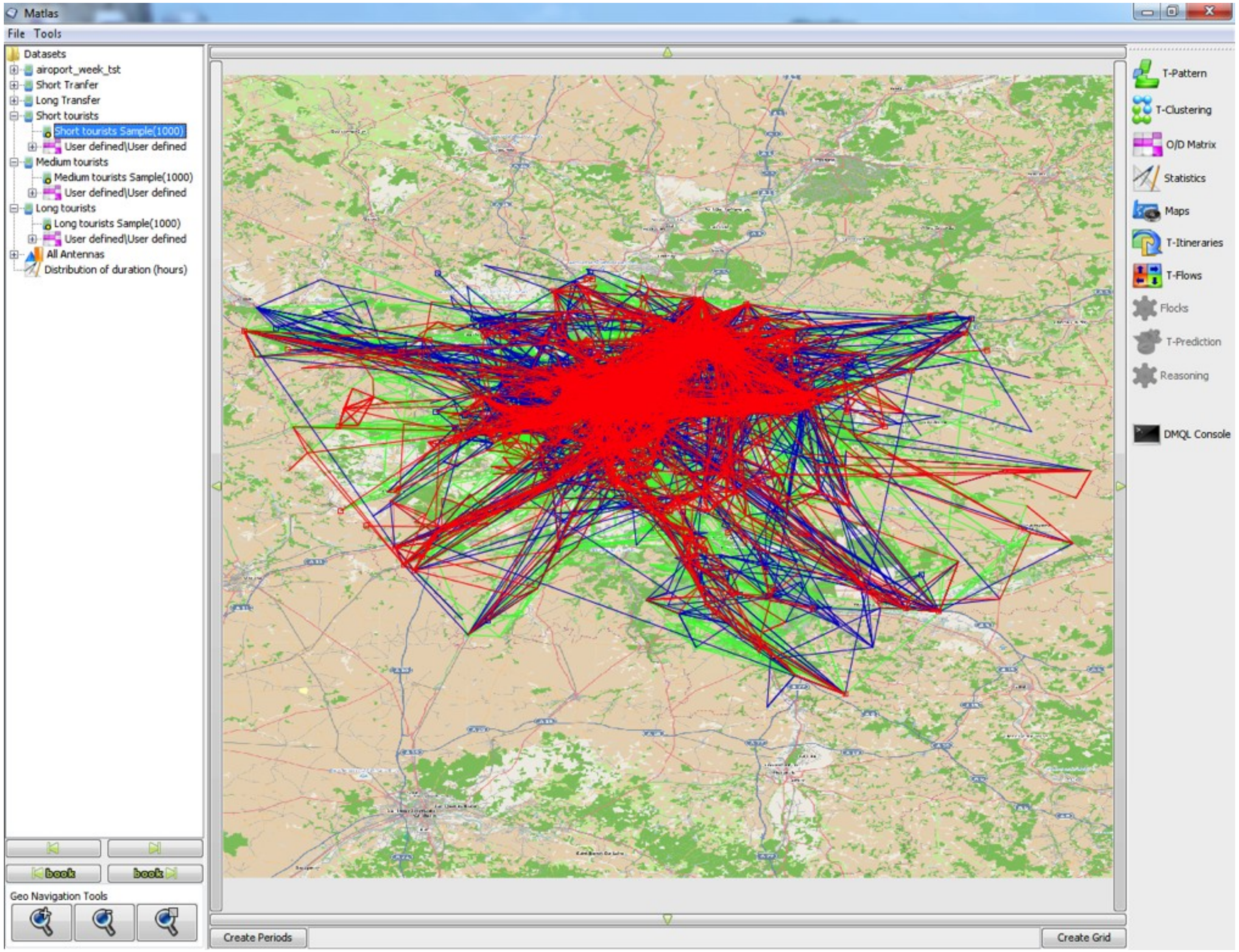


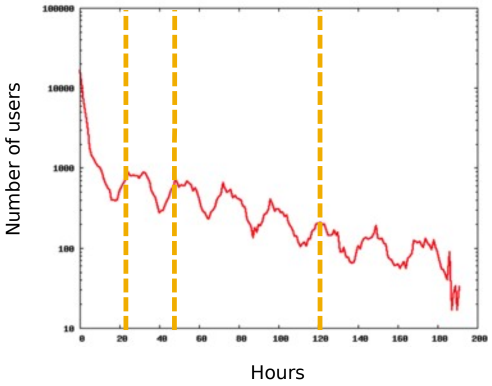
Distribution of visiting time





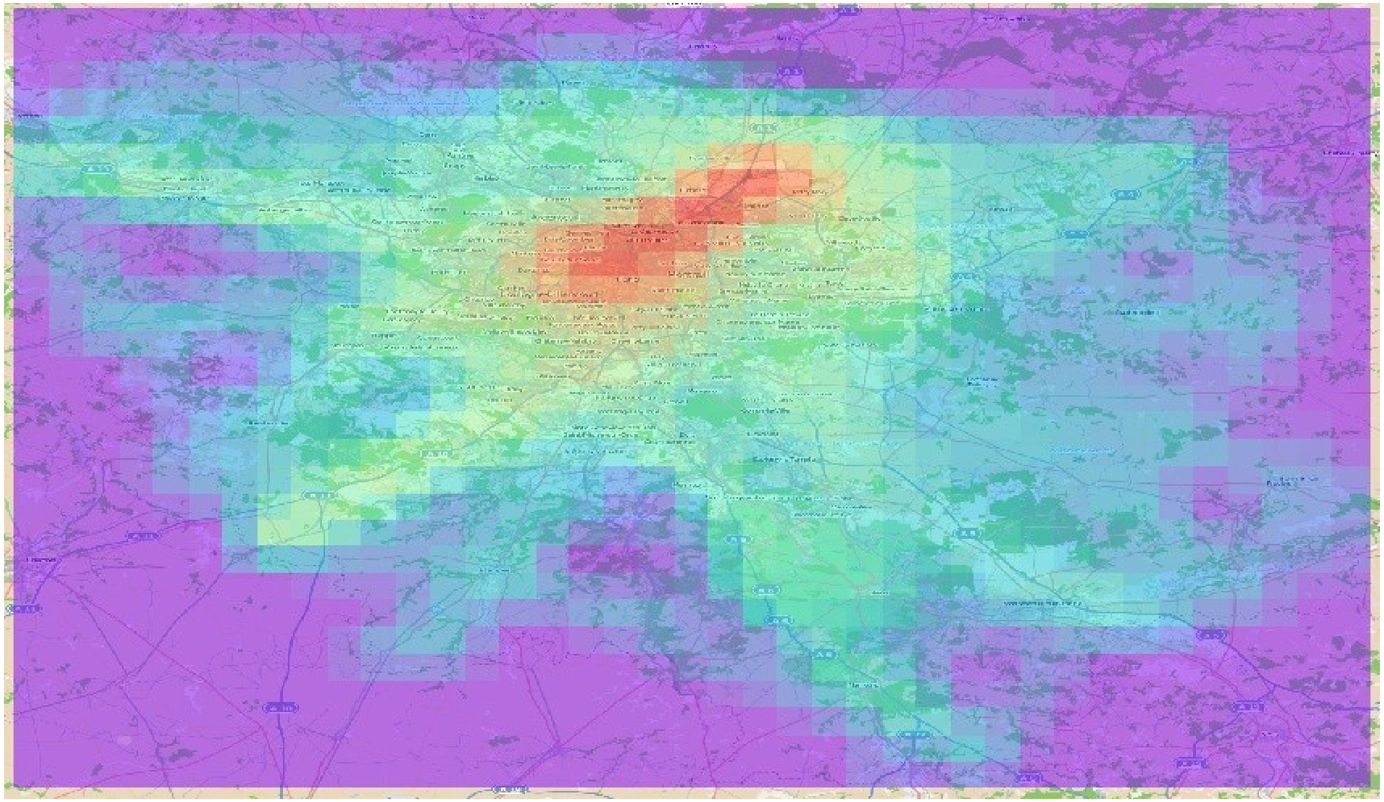
Categorization of tourists





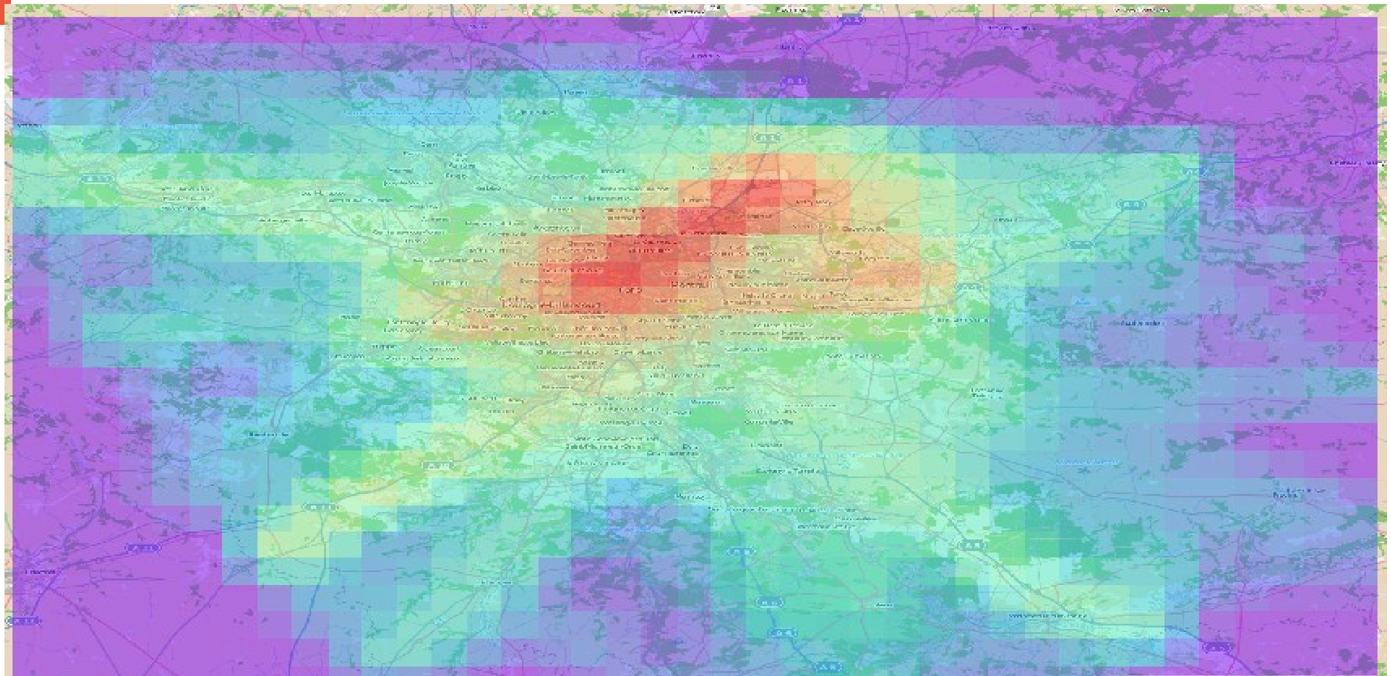
Short period stay Tourist (1 day ▭ 2 days)
Medium period stay Tourist (2 day ▭ 5 days)
Long period stay Tourist (5 day ▭ 7 days)

Density map (Short stay)



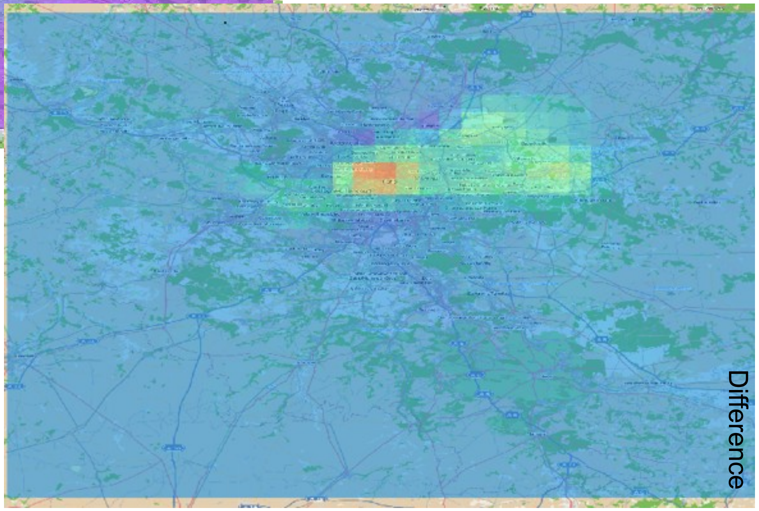
Short stay tourists visit the very center of Paris and go back the airport to leave.

Density map (Medium stay)

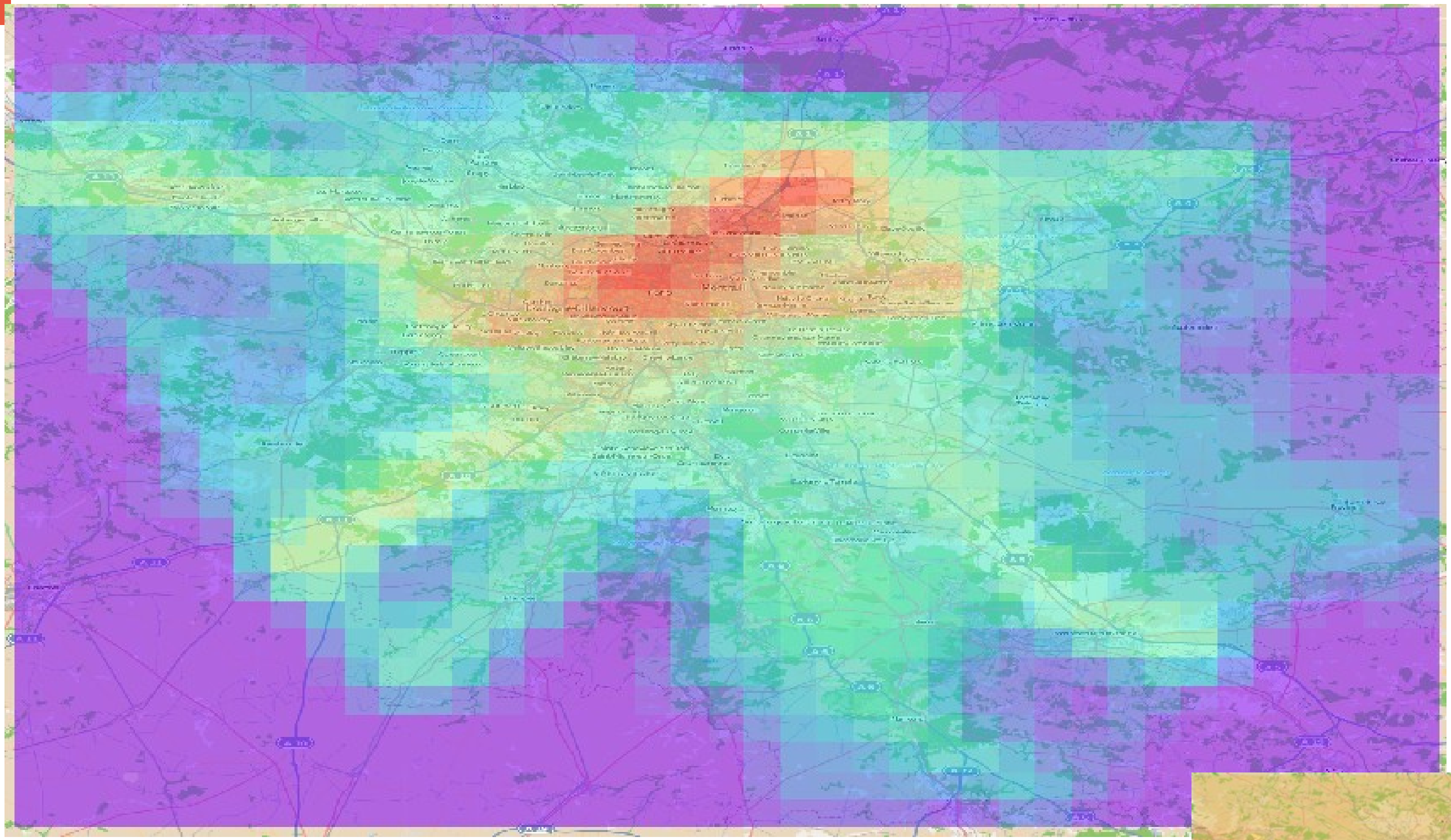


Medium stay tourists visit the center of Paris mostly but Versailles and Disneyland appear as new destinations

Green = Disneyland Paris
Red = Versailles

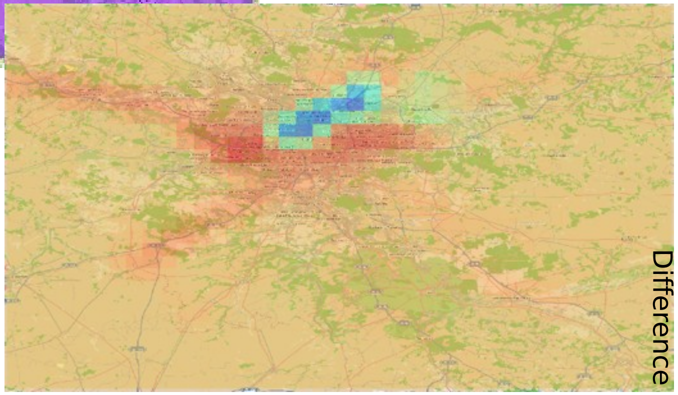


Density map (Long stay)



Long stay tourists visit the center of Paris, Versailles and Disneyland as major destinations, but they also leave Paris toward the surrounding areas.

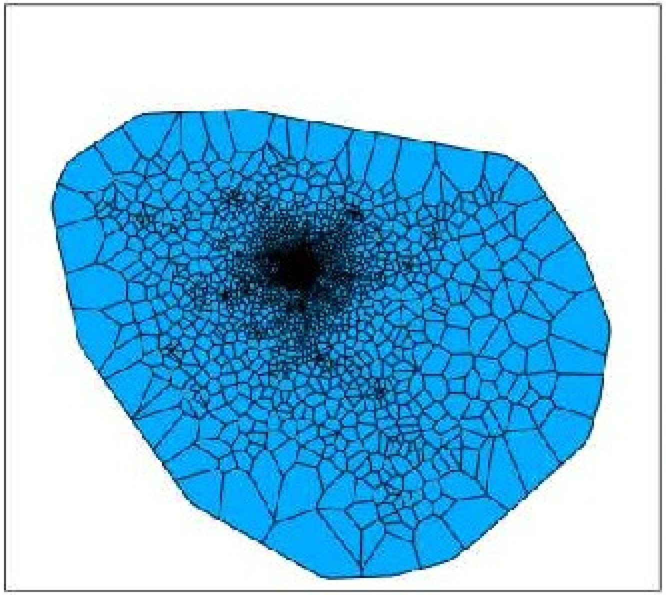
Green = Disneyland Paris
Red = Versailles
Blue = Highway/Train to Mante la jolie
Black = Highway to South-West

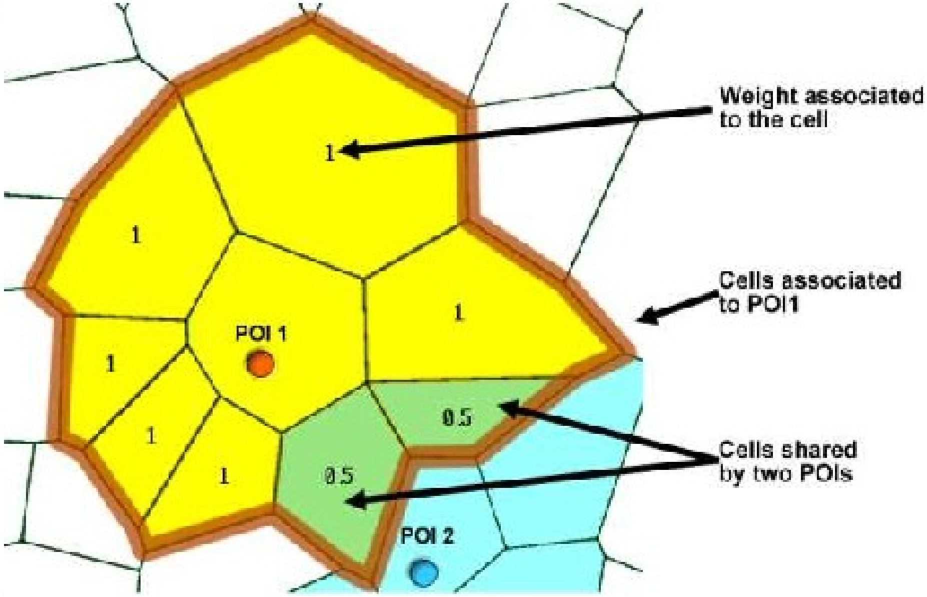


Difference

Point of Interests and Towers

The trajectories jump between towers which do not correspond to the exact position of the POIs. To perform the mapping we defined a mapping between the towers and POIs:

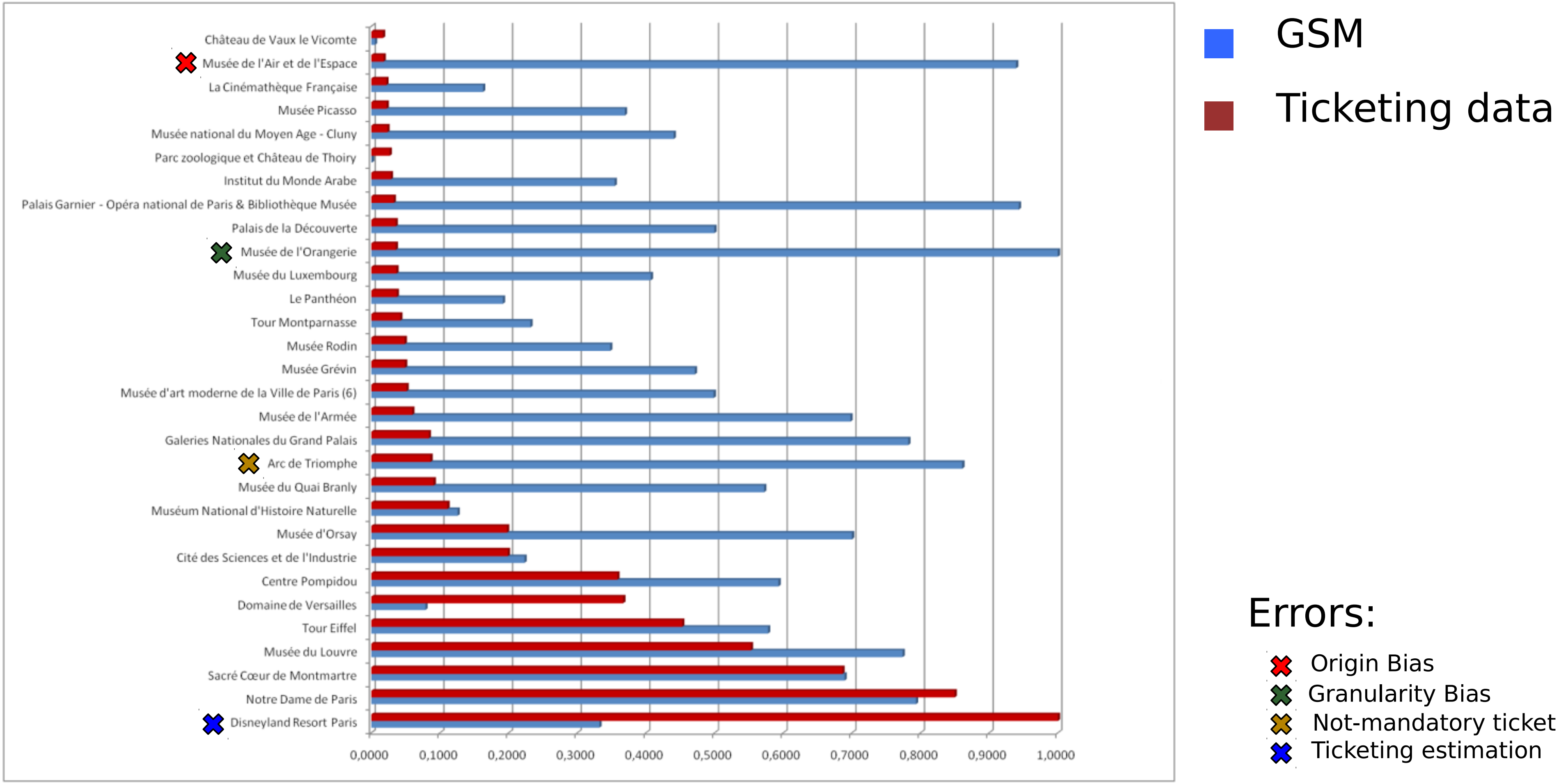




Weight = 1/#neighboring POIs

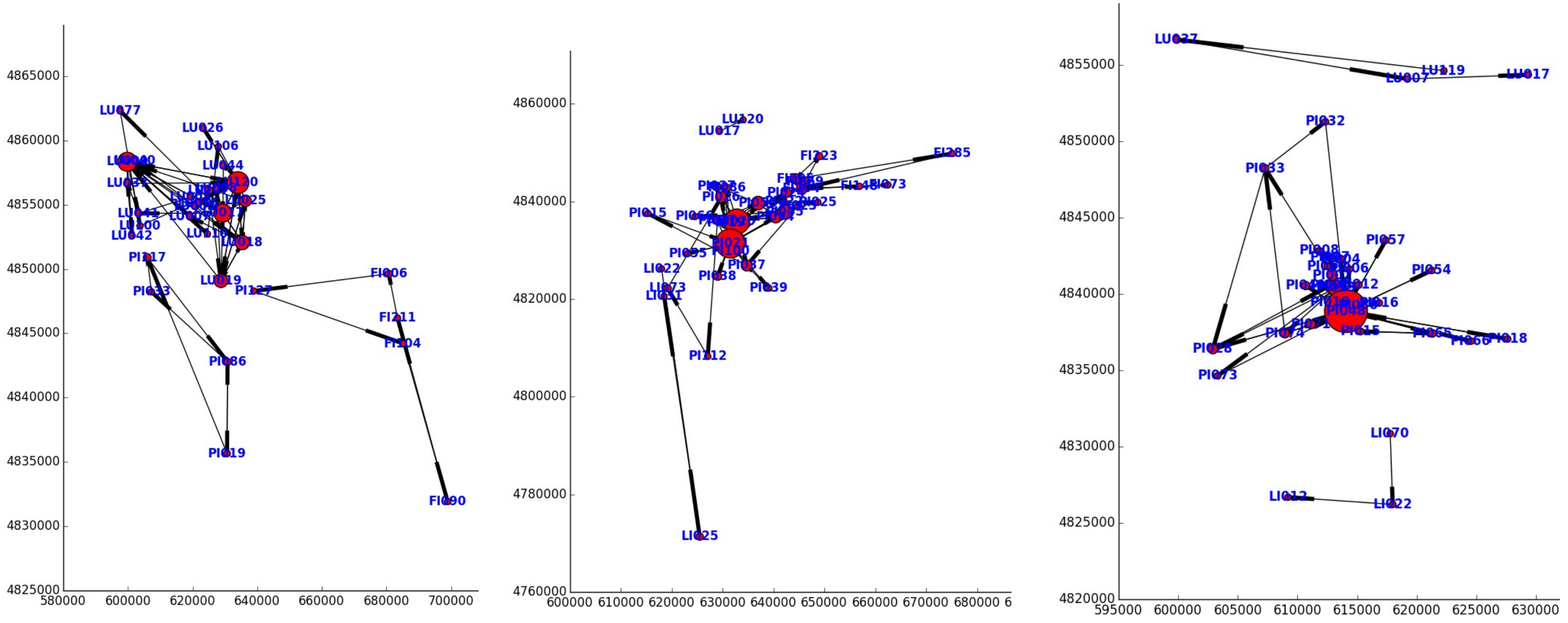
Comparison with Ticketing data

There are differences between the ticketing data and GSM-based density, we discovered that they are comparable only in the places where the ticket is necessary and the data is not estimated.



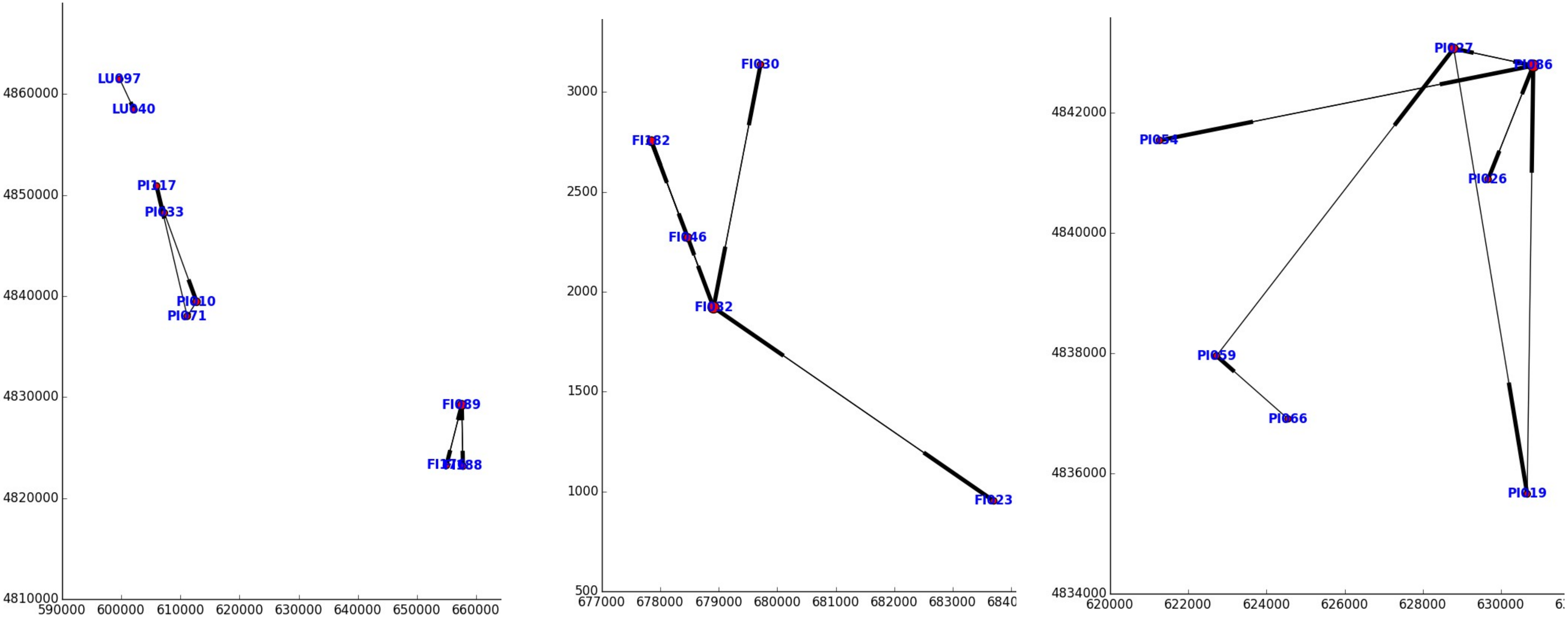
Understanding Individual Mobility

- Difficult task: high variability of behaviours



Understanding Individual Mobility

- Difficult task: several low frequency users

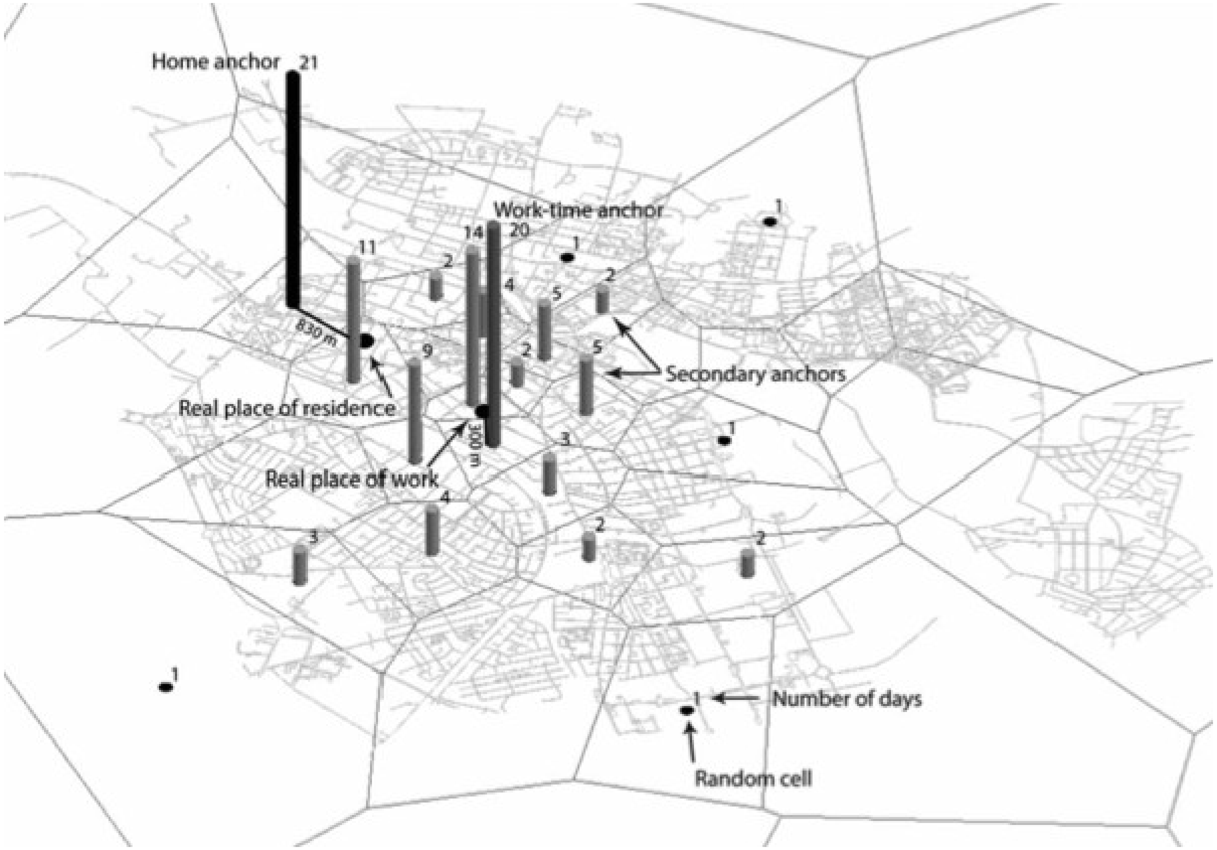


Identifying important locations

- 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.

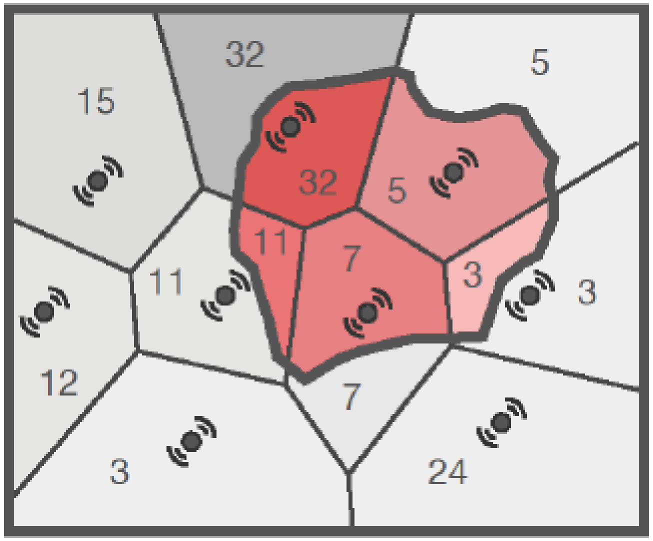
Identifying important locations

- 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

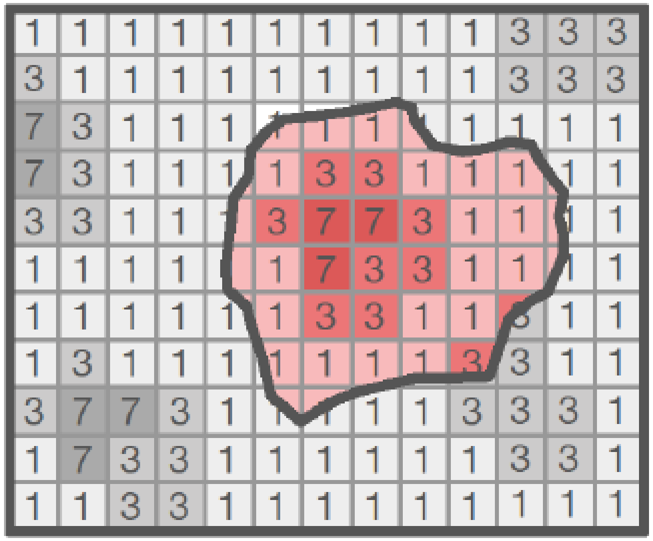


$$\sigma_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} \sigma_{v_j} A_{(c_i \cap v_j)}$$

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

Identifying important locations

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



1	1	1	1	1	1	1	1	1	1	3	3	3
3	1	1	1	1	1	1	1	1	1	3	3	3
7	3	1	1	1	1	1	1	1	1	1	1	1
7	3	1	1	1	1	3	3	1	1	1	1	1
3	3	1	1	1	3	7	7	3	1	1	1	1
1	1	1	1	[illegible]	1	7	3	3	1	1	1	1
1	1	1	1	1	1	3	3	1	1	3	1	1
1	3	1	1	1	1	1	1	1	3	3	1	1
3	7	7	3	1	1	1	1	1	3	3	3	1
1	7	3	3	1	1	1	1	1	1	3	3	1
1	1	3	3	1	1	1	1	1	1	1	1	1

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

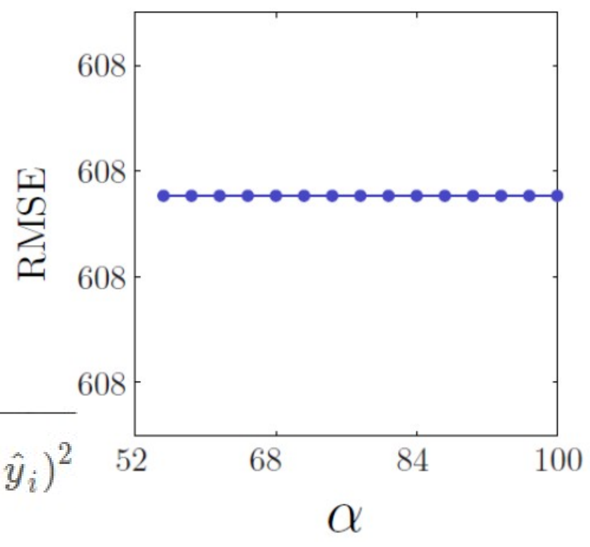
- Approach: integrate external indicators of relative density – e.g. from environment and infrastructures – to distribute cells' contrib.

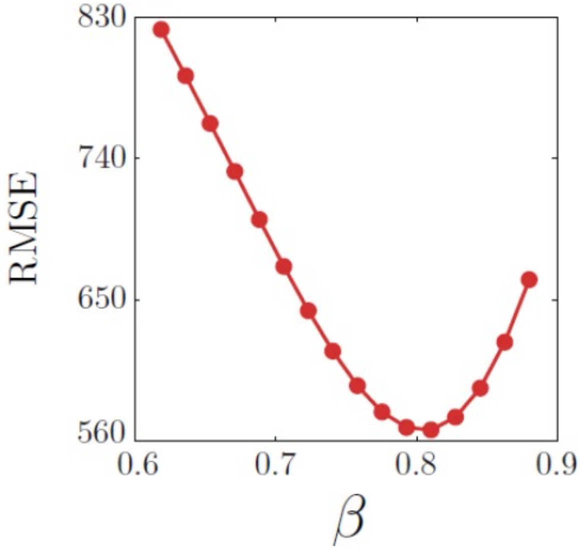
Identifying important locations

- 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)

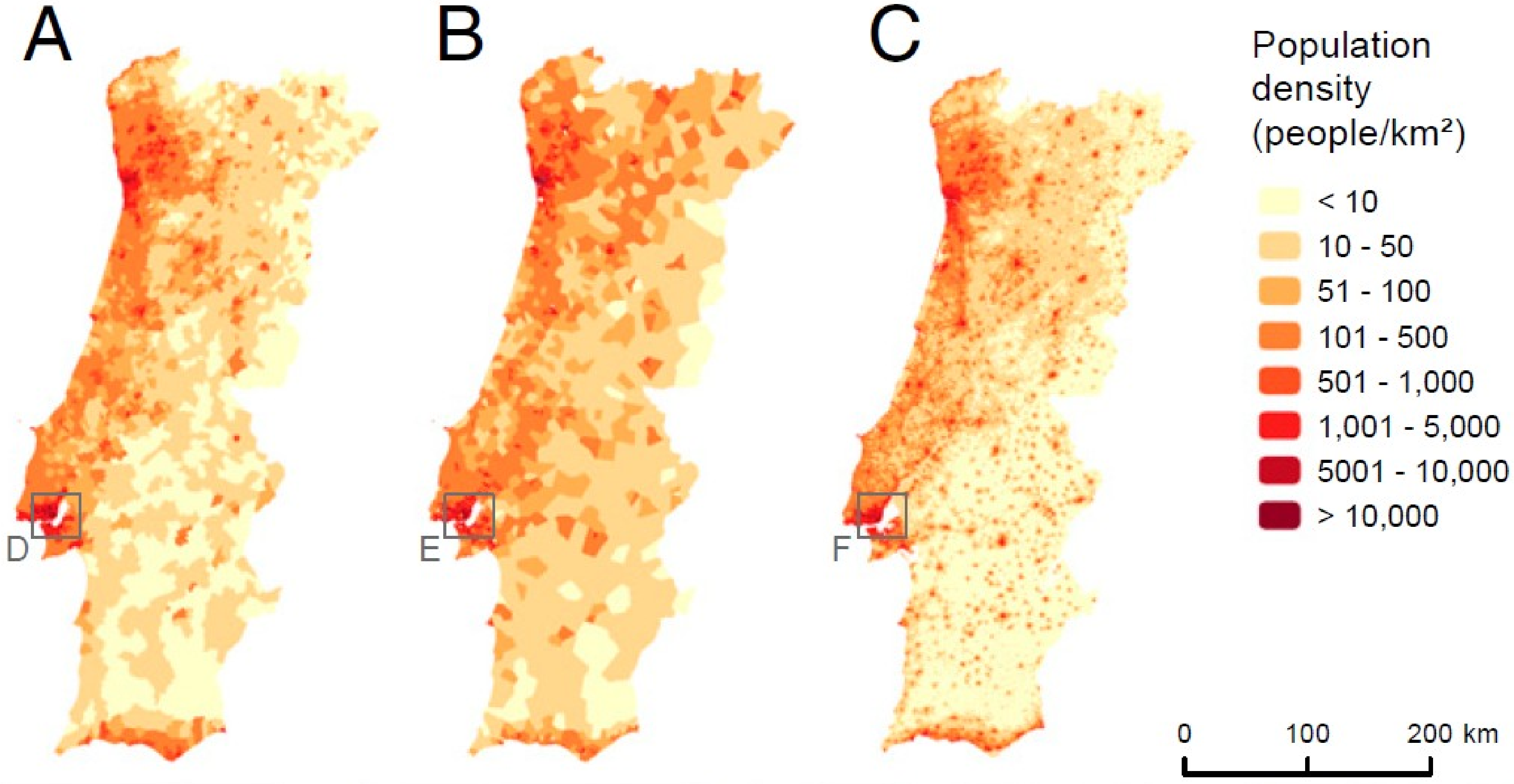




$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Identifying important locations

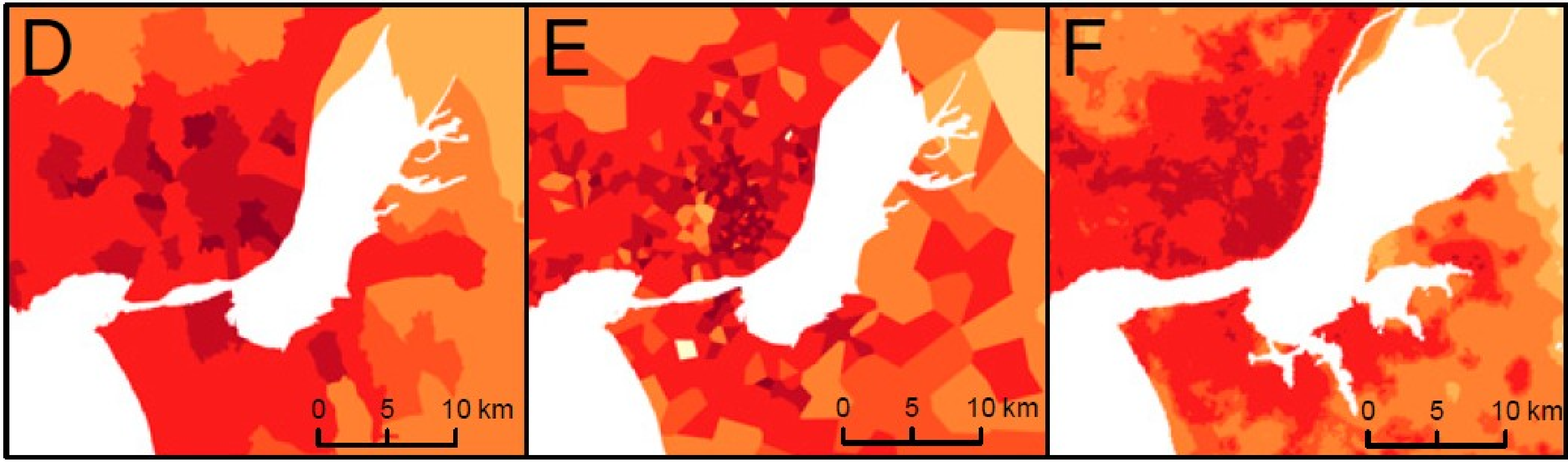
- Sample results on Portugal



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

Identifying important locations

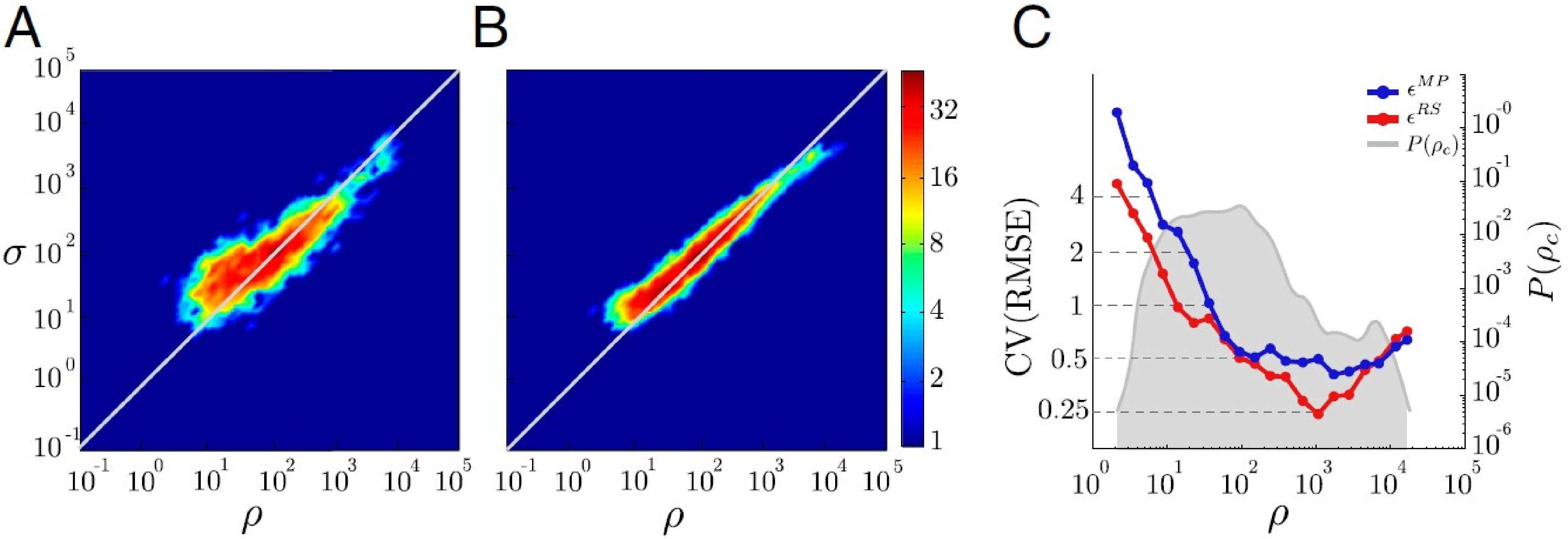
- Sample results on Portugal (close-up)



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

Identifying important locations

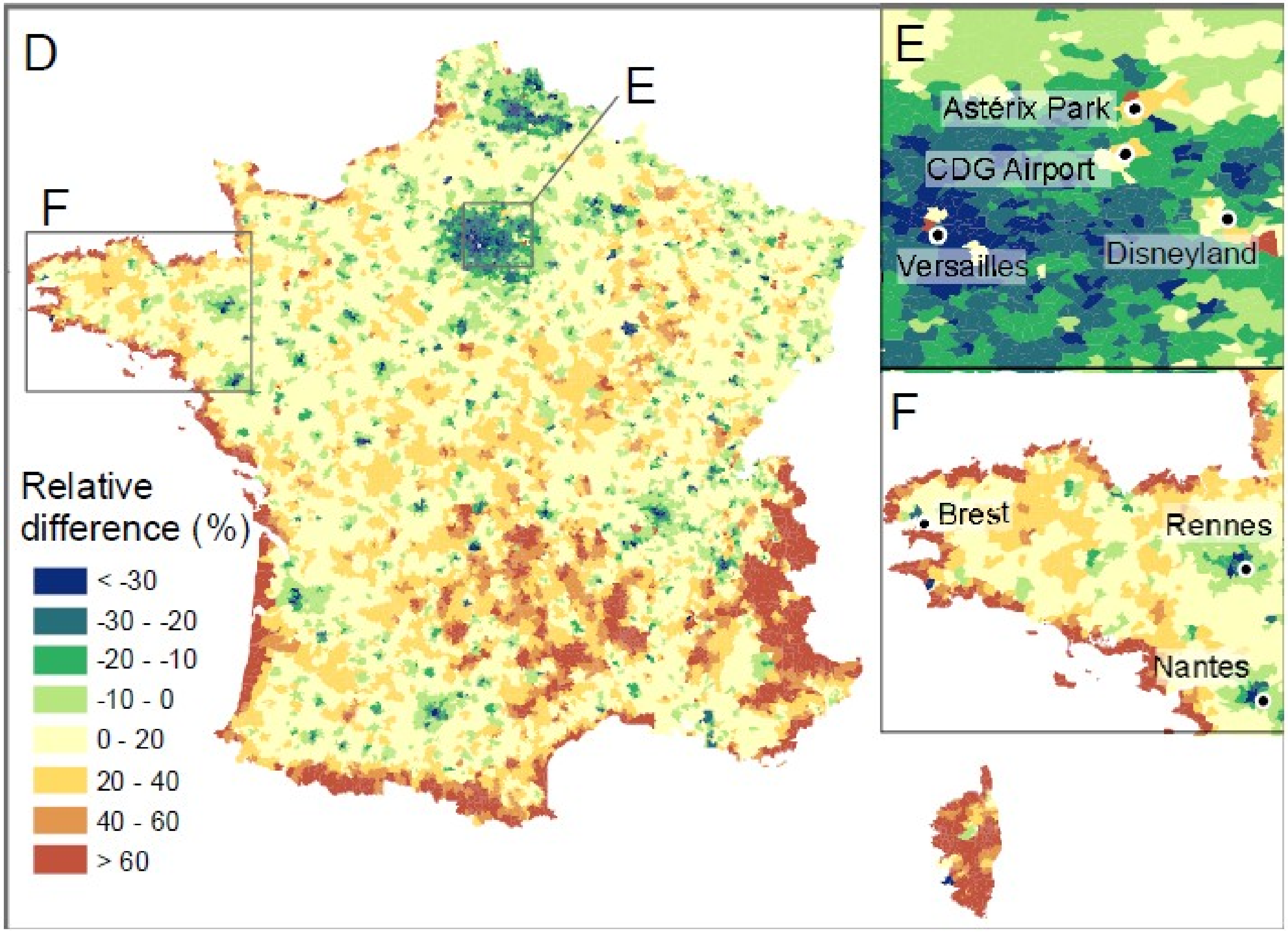
- Sample results



A = GSM data B = Environment/Infrastructures-based

Identifying important locations

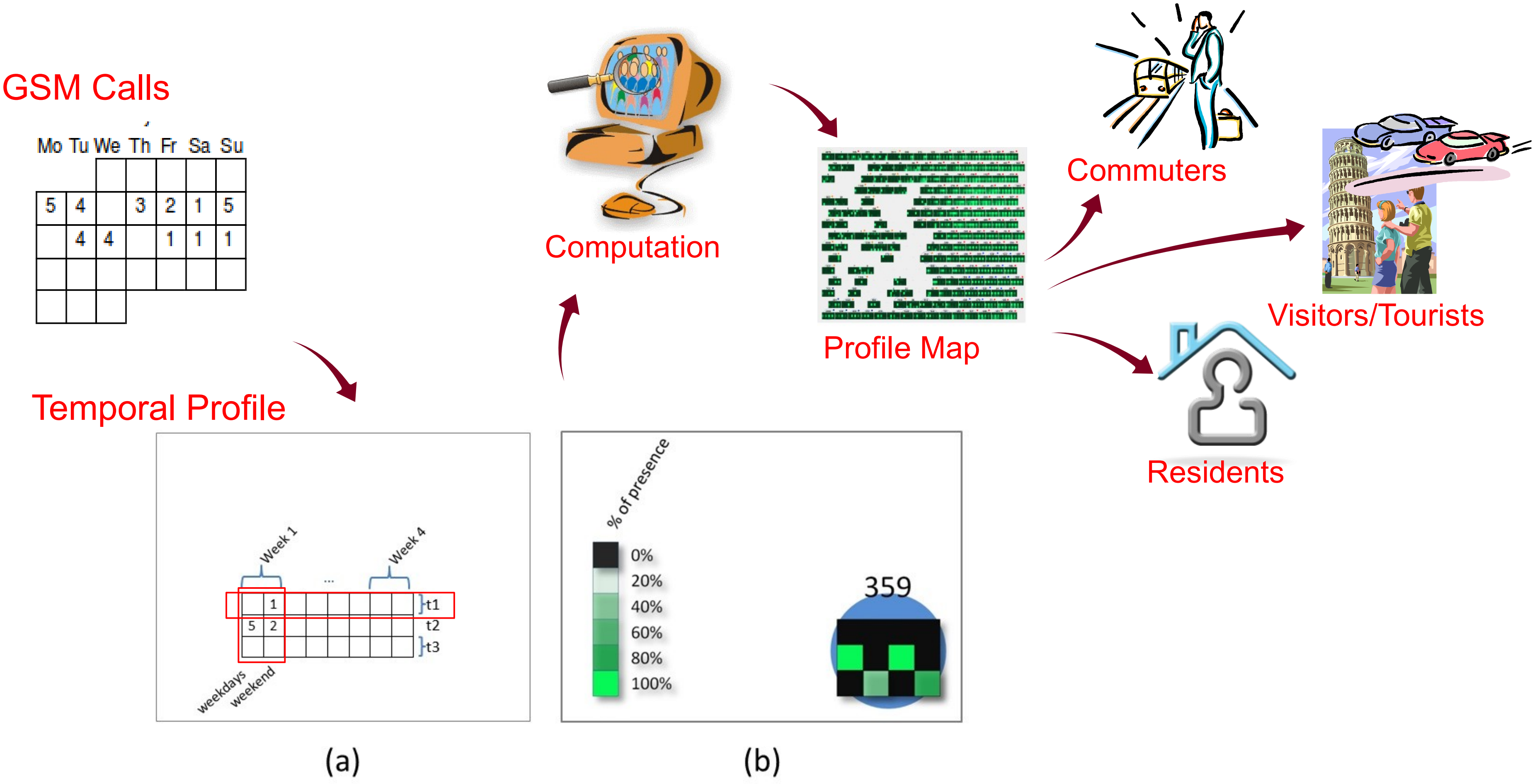
- Sample usage: evaluate seasonal changes
  - Summer variations vs. Winter period



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



GSM Calls

Mo	Tu	We	Th	Fr	Sa	Su
5	4		3	2	1	5
	4	4		1	1	1

Temporal Profile

Computation

Profile Map

Commuters

Visitors/Tourists

Residents

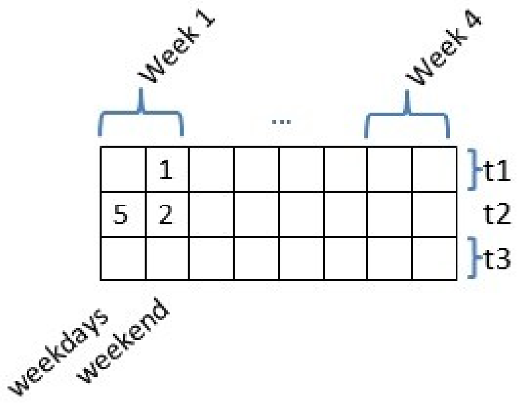
Sociometer

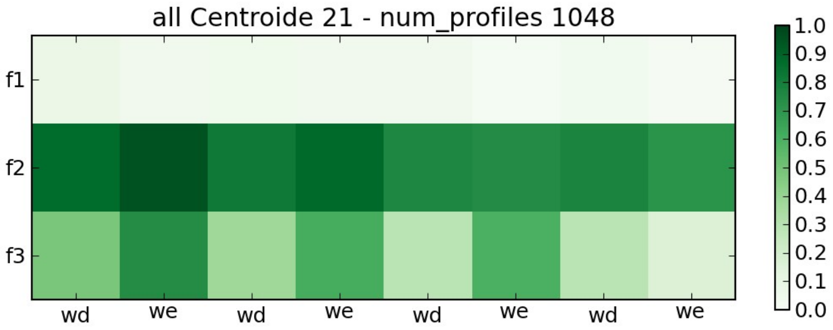
Step 1: build individual profiles

- 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
….
```

t1 = [00:00-08:00)
t2 = [8:00-19:00)
t3 = [19:00-24:00)

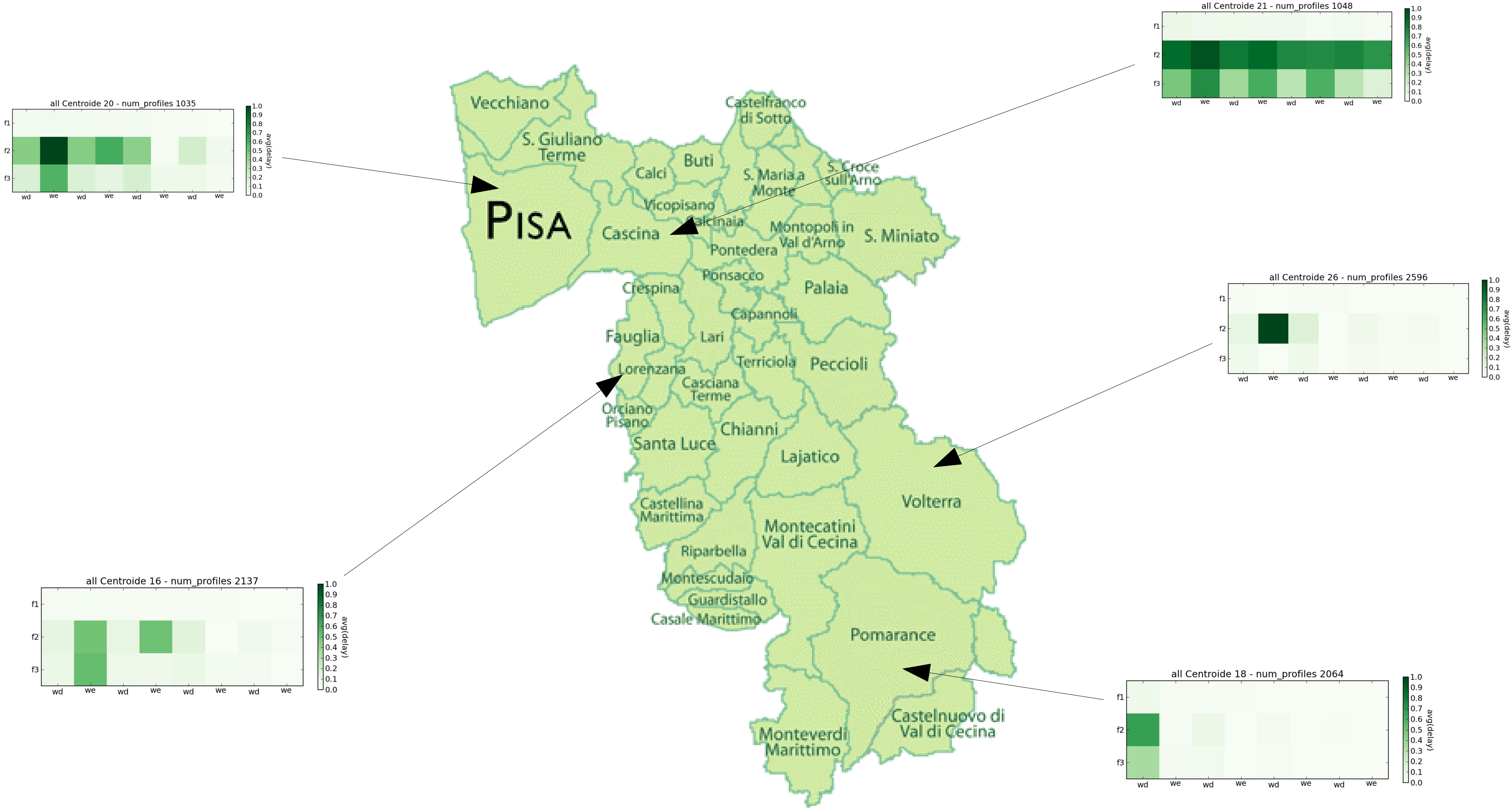




Sociometer 2.0

Step 1: build individual profiles

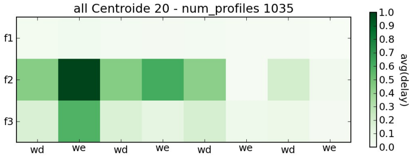
- Result for each user: set of individual profiles:

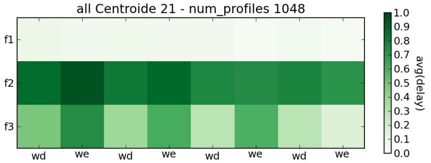


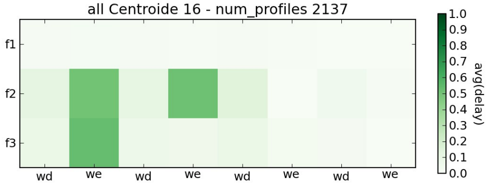
Sociometer 2.0

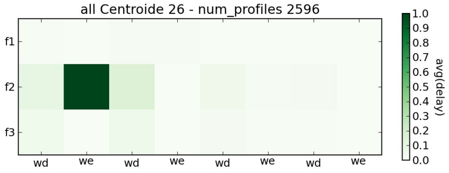
Step 2: find representative profiles across all dataset

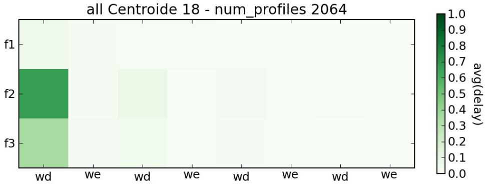
- Based on clustering
  - simple k-means: start with K random representatives, and iteratively refine them
  - in our experiments, k=100
- Output: set of reference (unlabelled) profiles







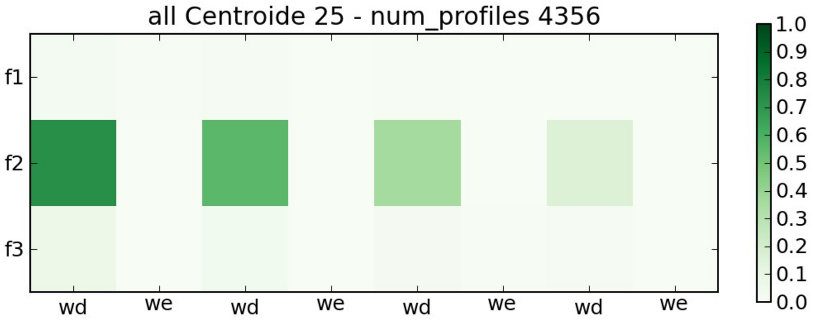




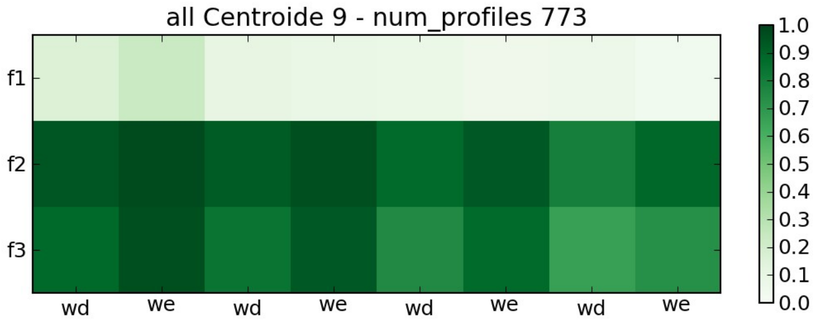
Sociometer 2.0

Step 3: associate representative profiles to categories

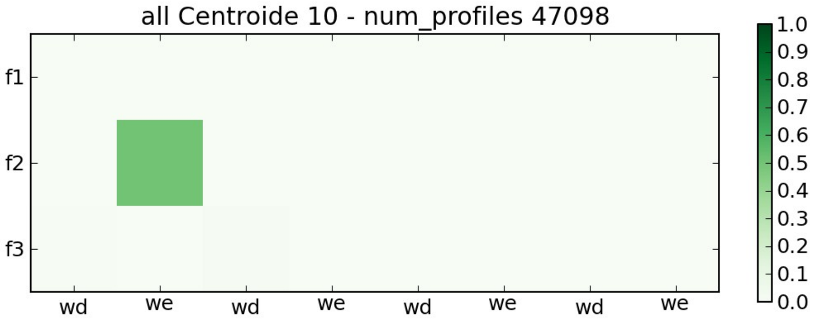
- Manual labelling
  - Use fuzzy rules, difficult to formalize
  - Crisp classification, no weights (reliability of labels)



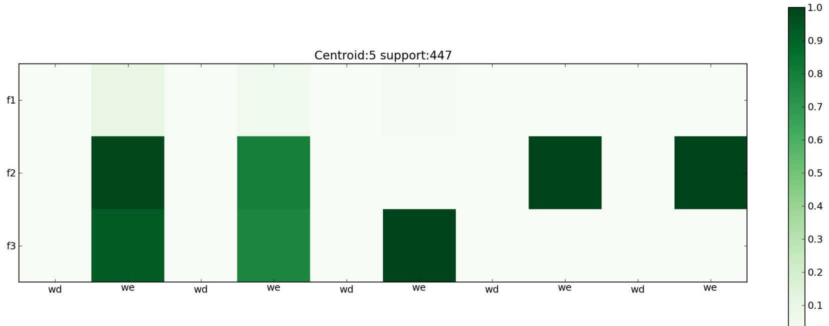
Commuter



"Static" resident



Occasional

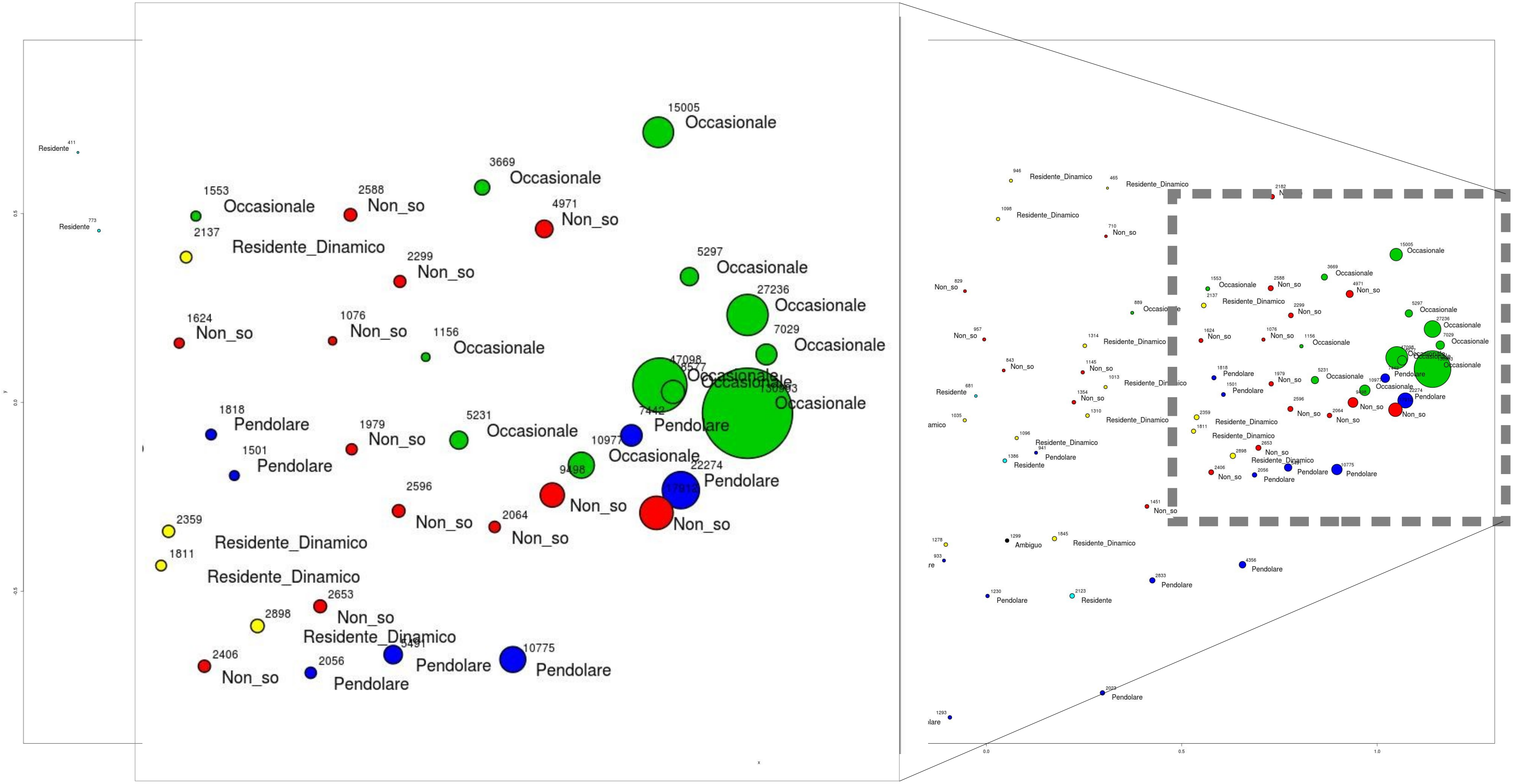


"Dynamic" resident

Sociometer 2.0

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

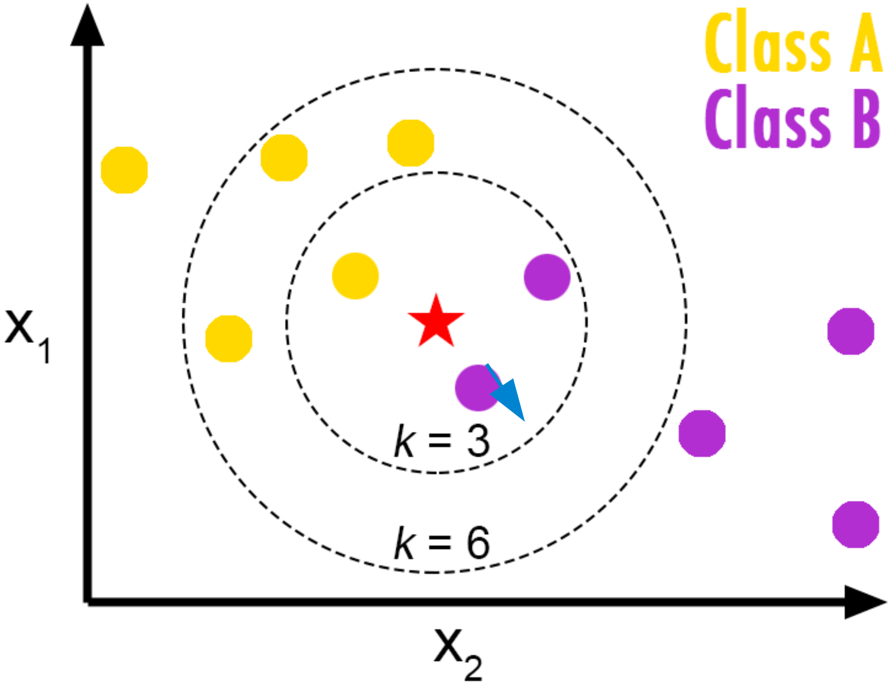
- Profiles (individual and representative) are 24-dimensional
- MDS (24 → 2) to visualize them



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

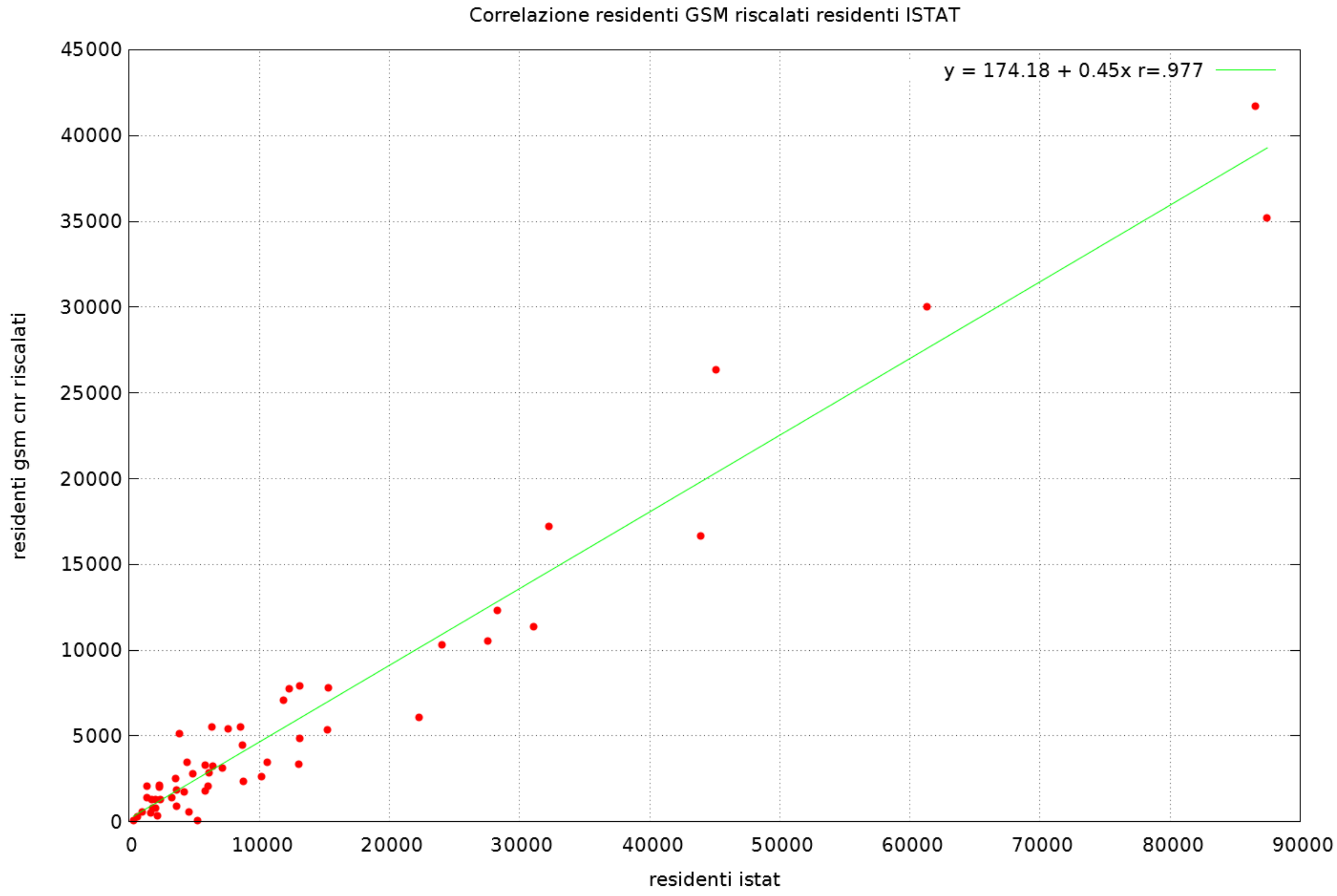
Step 5: aggregate into presence stats and O/D flows

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

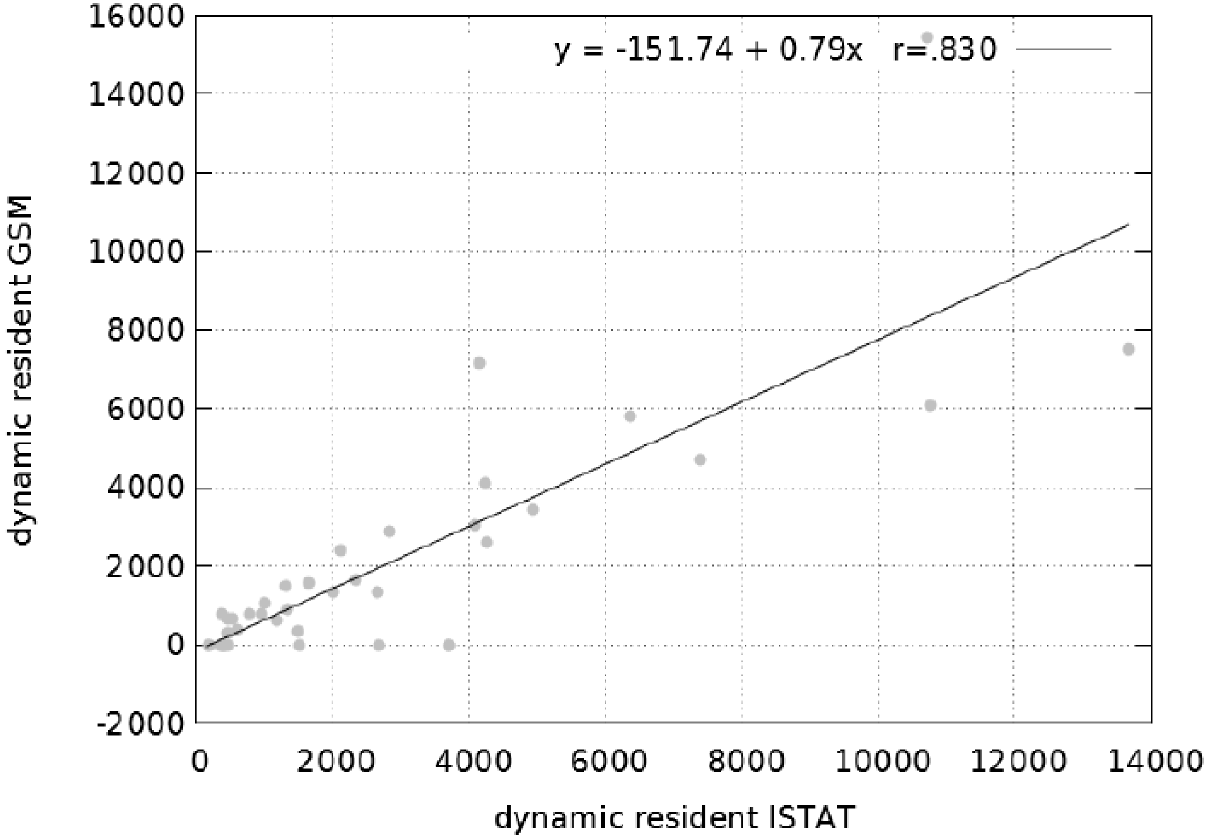
ISTAT Persons & Places project

- 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

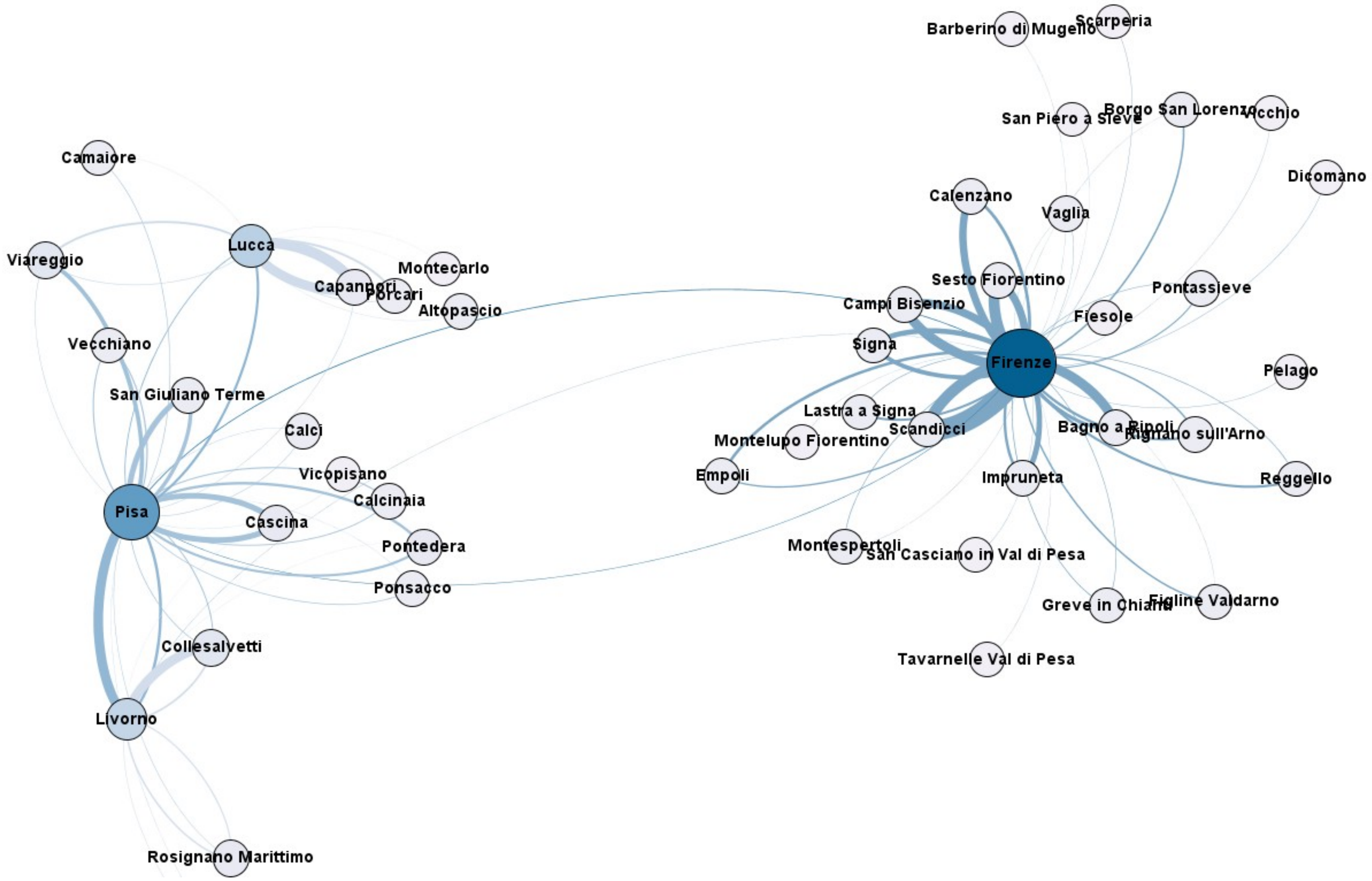


Dynamic residents (outgoing)



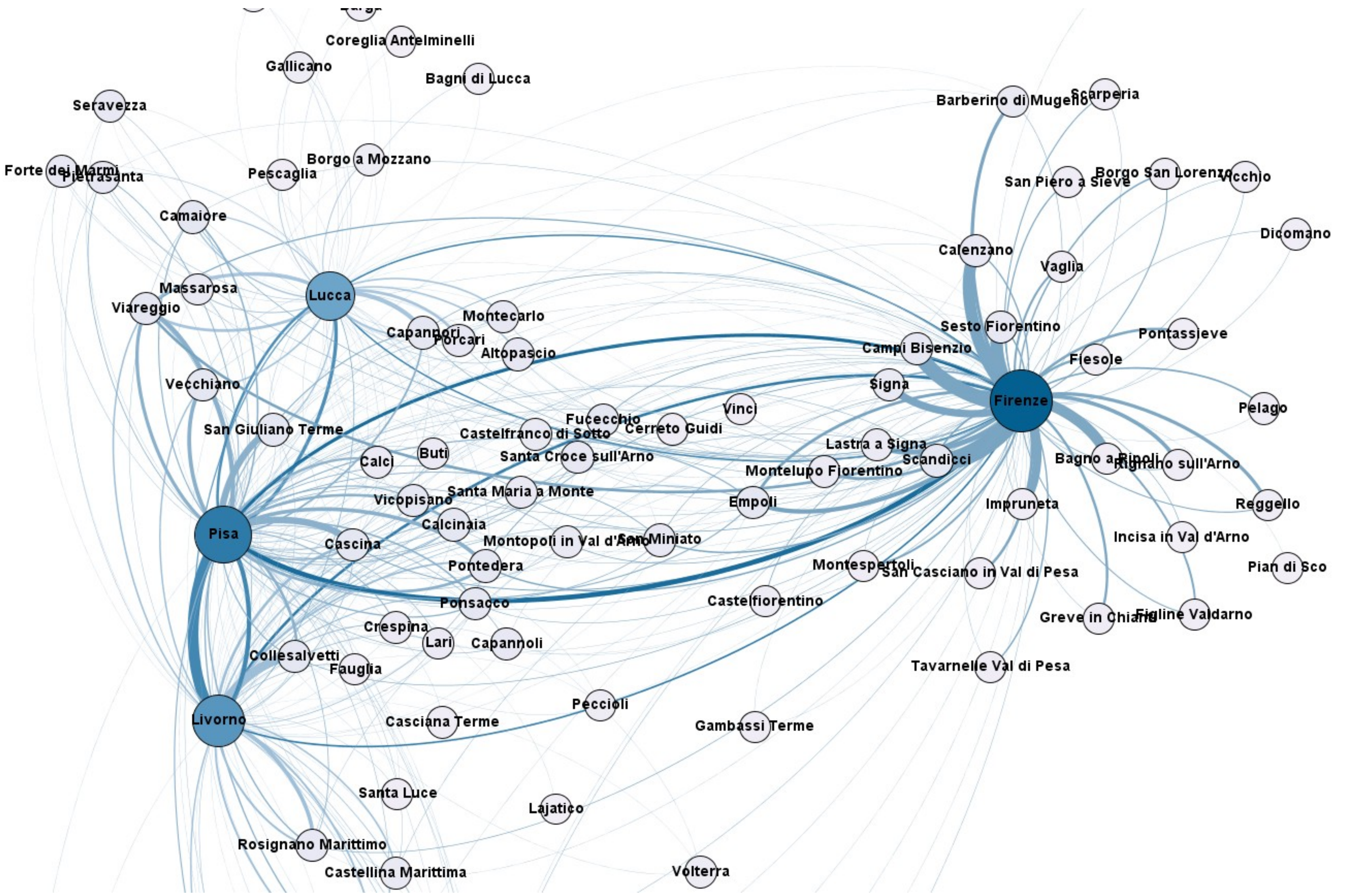
Sample results / 1

Home-Work



Sample results / 2

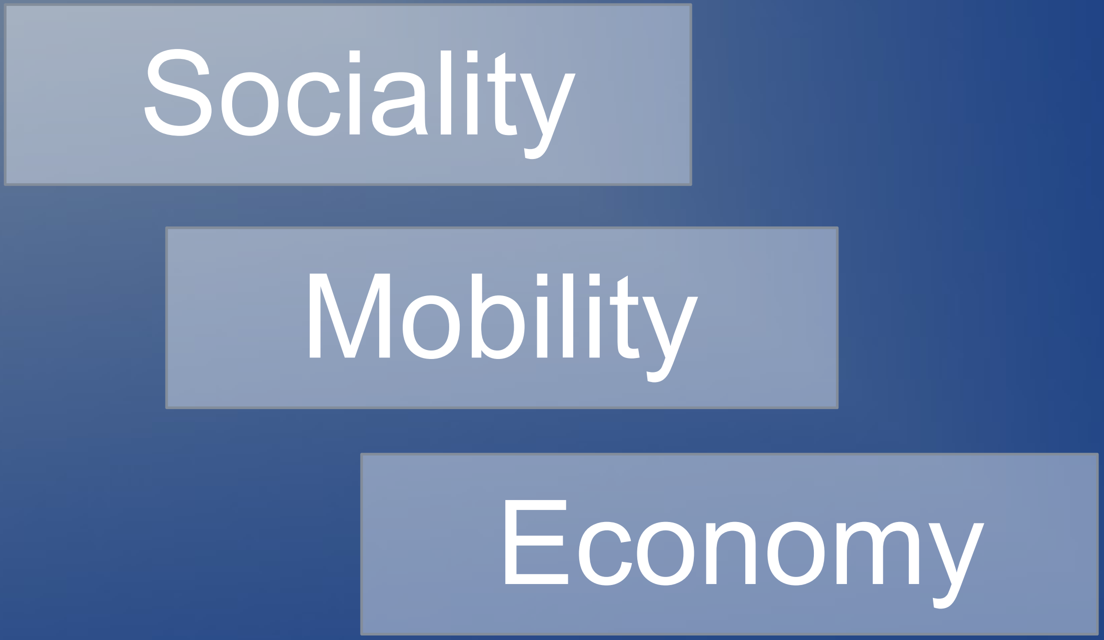
Home-Visits



A multidimensional data driven study of human behavior

Goals

- Understanding the complex relationships between several social aspects:



Sociality

Mobility

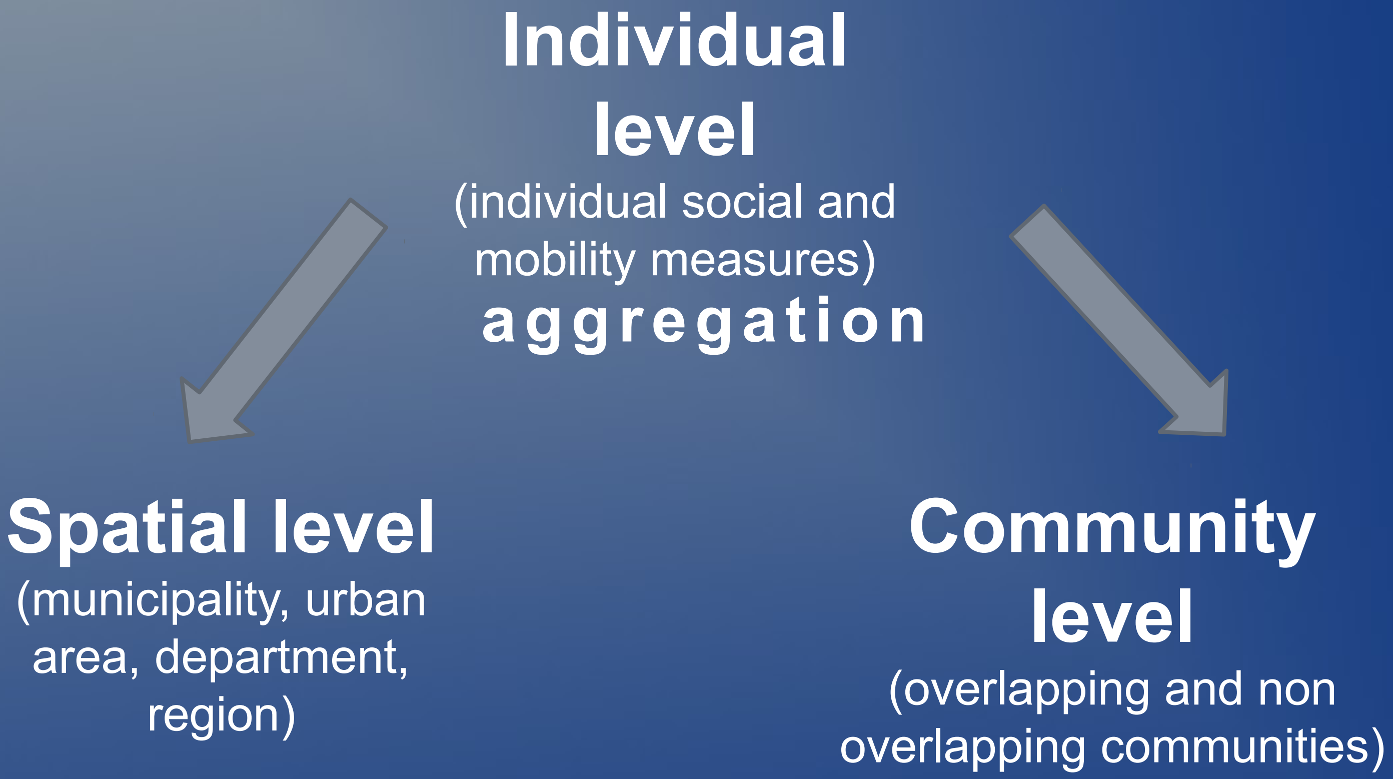
Economy

Goals

- Mobile phone data are used as a proxy for both human mobility and social interactions.

- The economic dimension (at municipality level) is provided by INSEE (French National Institute of Statistics and Economic Studies).

Goals



Individual level

(individual social and mobility measures)

aggregation

Spatial level

(municipality, urban area, department, region)

Community level

(overlapping and non overlapping communities)

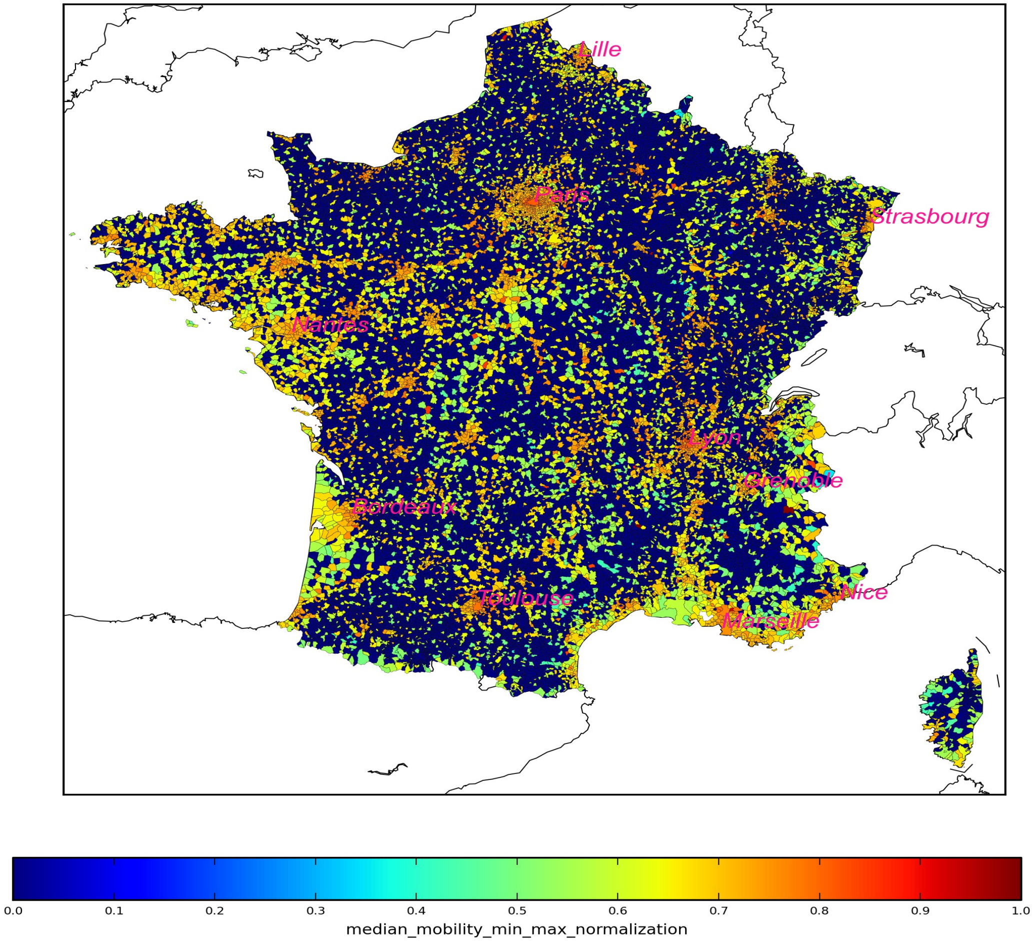
Mobility measures

- The **radius of gyration** of a user is the characteristic traveled distance, a measure of how far she is from her center of mass.

$$\vec{r}_{cm} = \frac{1}{N}\sum_{i \in L} n_i \vec{r}_i$$

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

Mobility entropy



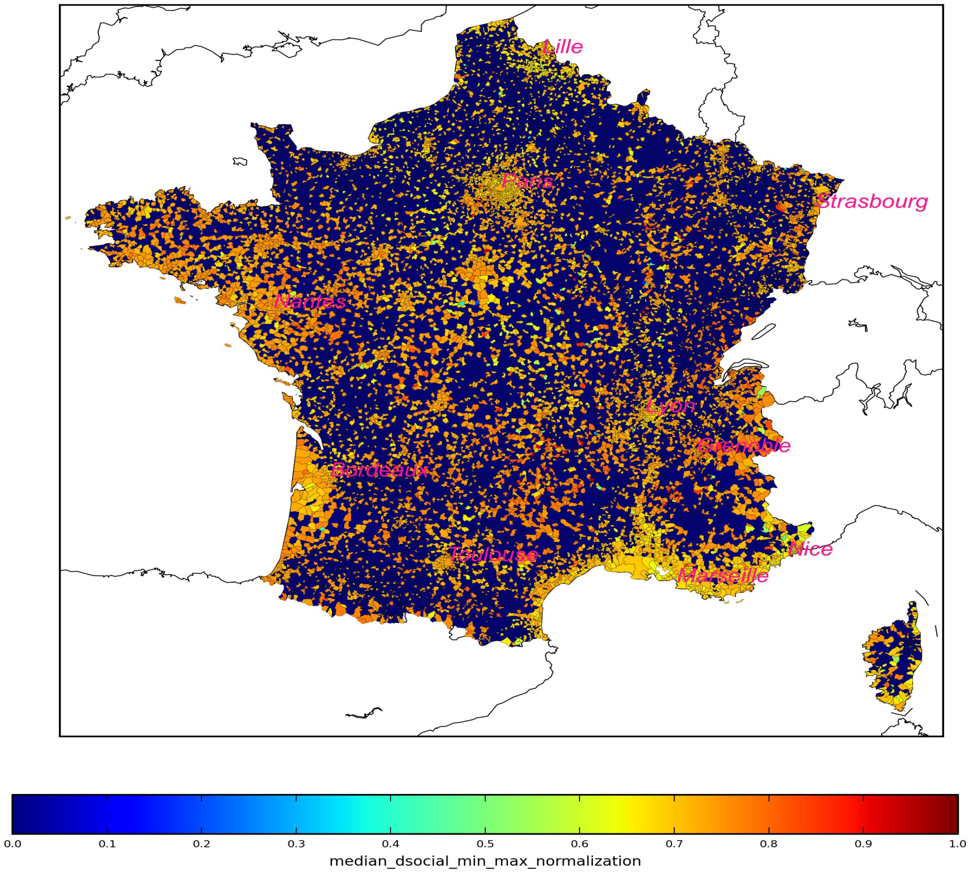
Social measures

- **Social diversity** captures the social diversity of communication ties within an individual’s social network. We quantify topological diversity as a function of the Shannon entropy.

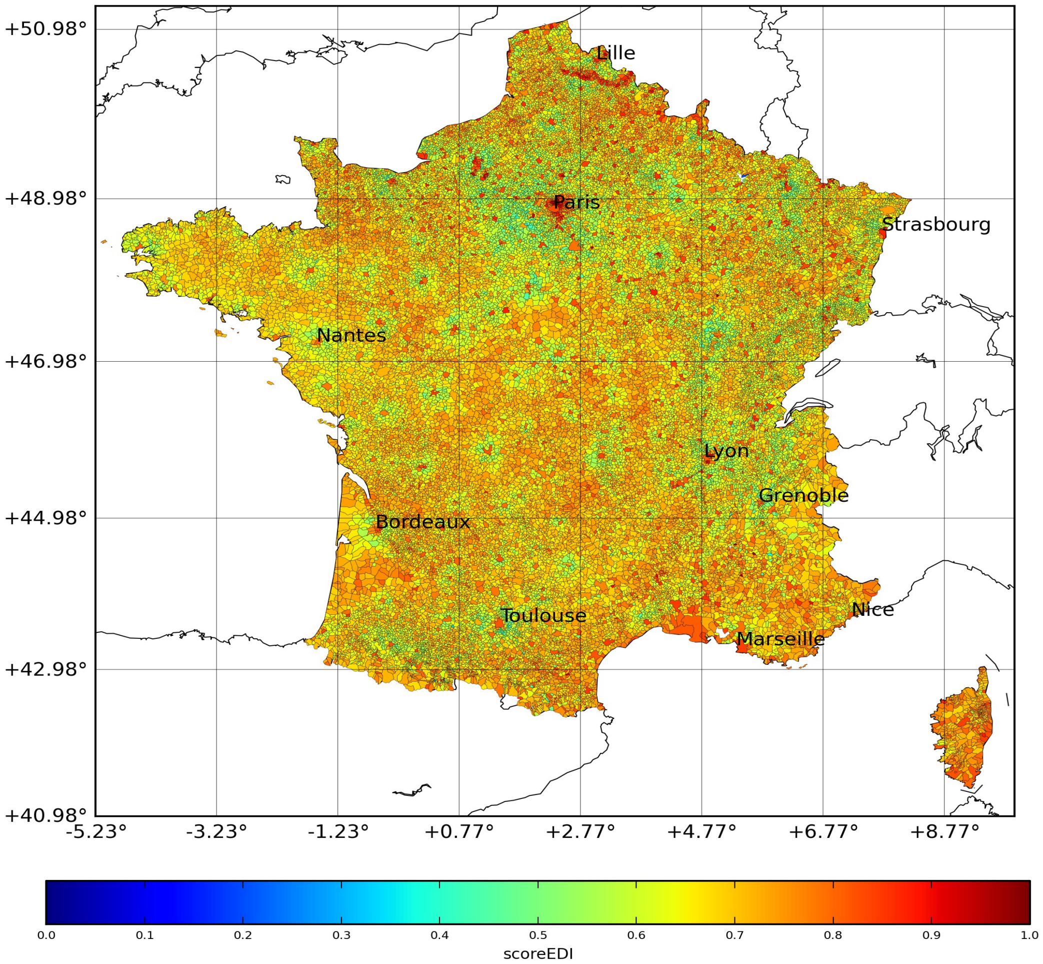
$$D_{social}(i) = \frac{-\sum_{j=1}^{k} p_{ij} \log(p_{ij})}{\log(k)},$$

$$p_{ij} = \frac{V_{ij}}{\sum_{j=1}^{k} V_{ij}},$$

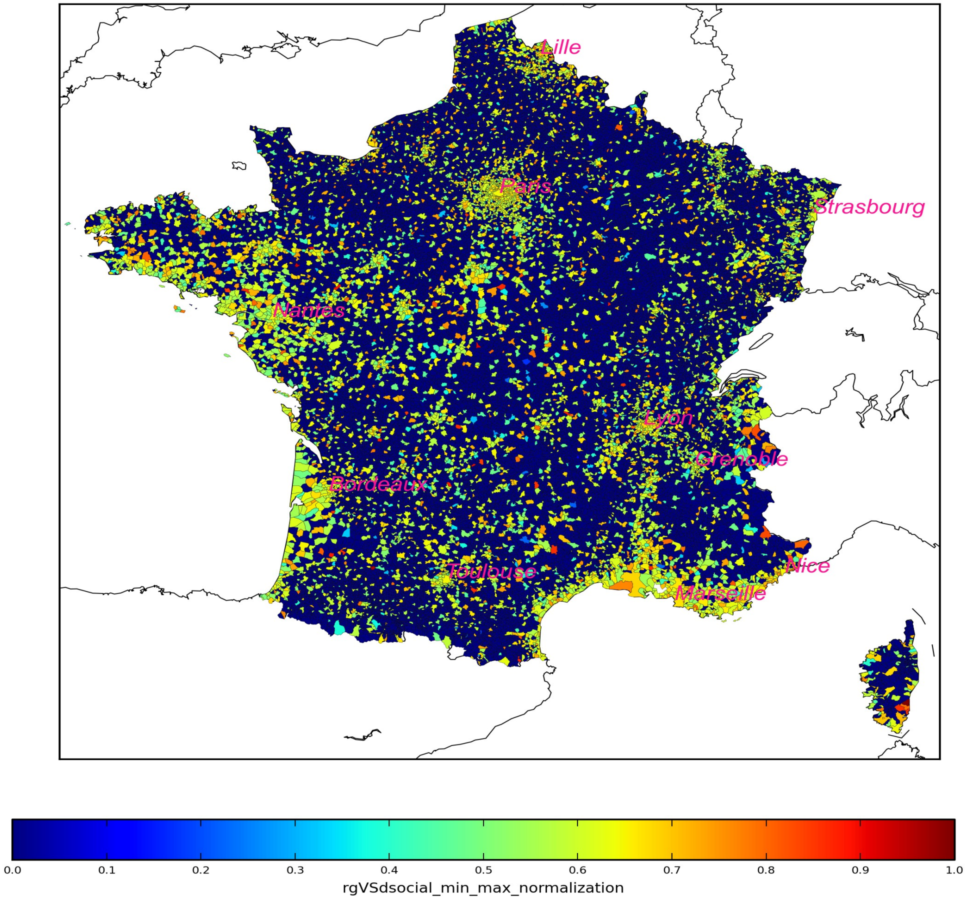
Social Diversity



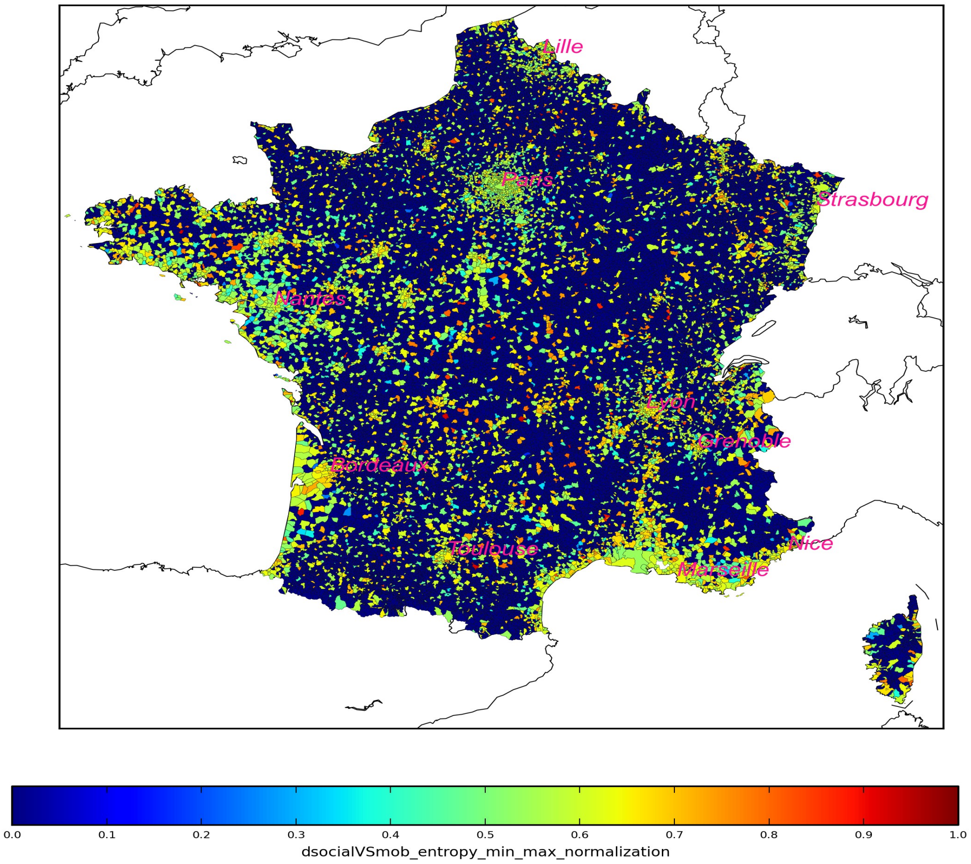
Deprivation Index



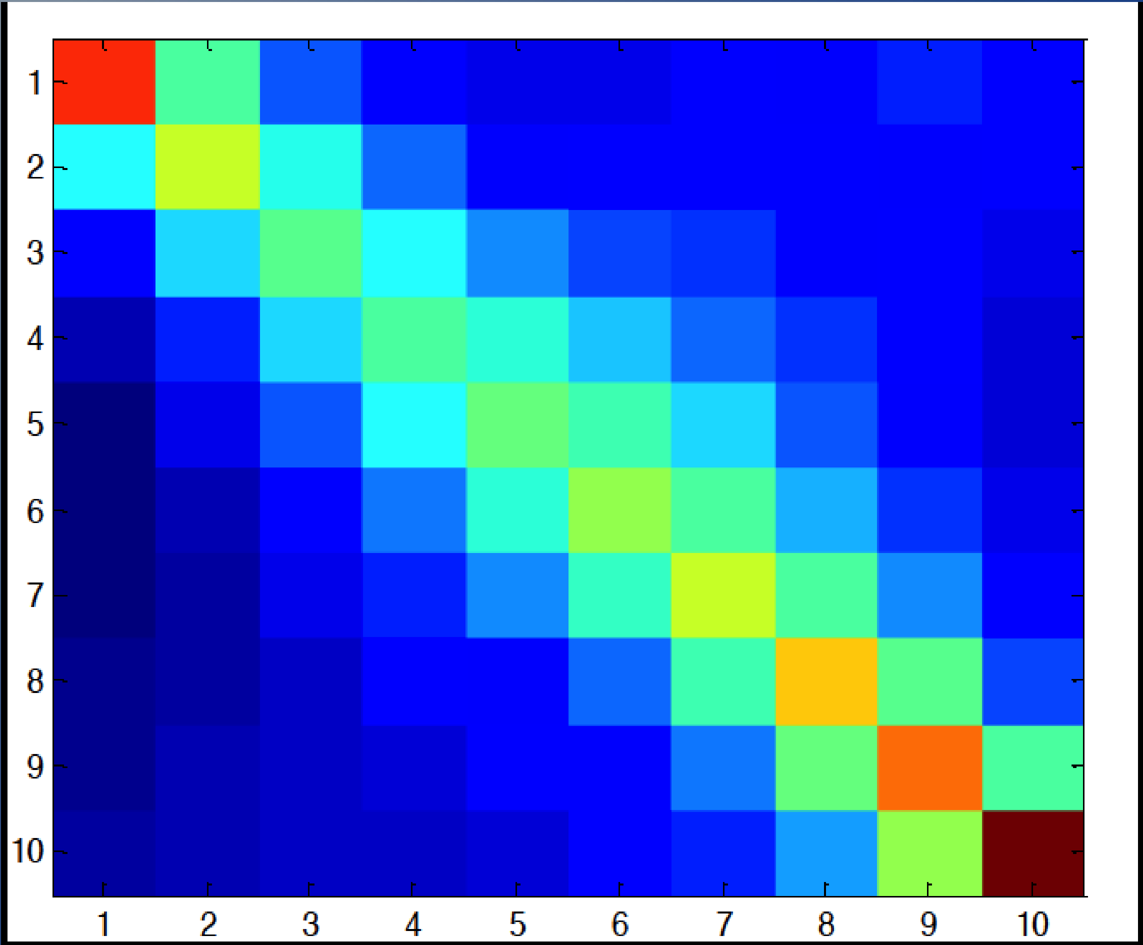
What did we do…
Correlation rg vs dsocial



What did we do…
Correlation dsocial vs mobility



People tend to connect with individuals having similar radius of gyration



Decile radius of gyration / linkage frequency

Correlations/dependencies between areas

Discovering urban and country dynamics from mobile phone data with spatial correlation patterns

kdd.isti.cnr.it

Roberto Trasarti
Mirco Nanni
Barbara Furletti, Fosca Giannotti

Ana-Maria Olteanu-Raimond
Thomas Couronné
Zbigniew Smoreda, Cezary Ziemlicki

General objective

Focus: observe the way the population density behaves in different areas of the city/region

Objective: spot statistically significant, yet potentially hidden, collective regularities

Approach: discover groups of regions that consistently behave in a coordinated way, suggesting the existence of some kind of connection among them

Examples/1

Set of events frequently happening at same time

- Regions that are tightly connected or all react to some (external) factor
- E.g.: people might tend to concentrate in specific areas during leisure time whenever the weather conditions are exceptionally good

Examples/2

- Sequence of events that frequently happen in a specific order
  - Existence of a reaction chain or external factors answered with different reaction times
  - E.g. (a chain of events): a large increase of people at a central train station frequently followed by an increase in an other station within a few hours

Analysis process

1. Extract **events related to population density** from raw data

- Density peaks & valleys might be not meaningful because physiologic to the region
    - E.g., rush hours, crowded stations, etc.
- Focus on **deviations** w.r.t. typical population density levels in each region

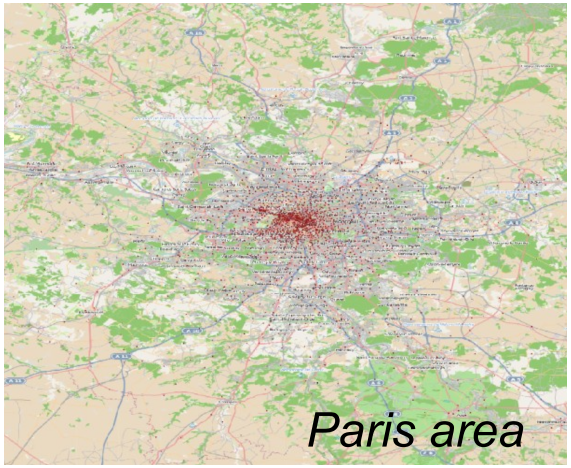
2. Search frequent combinations of **events** across different regions

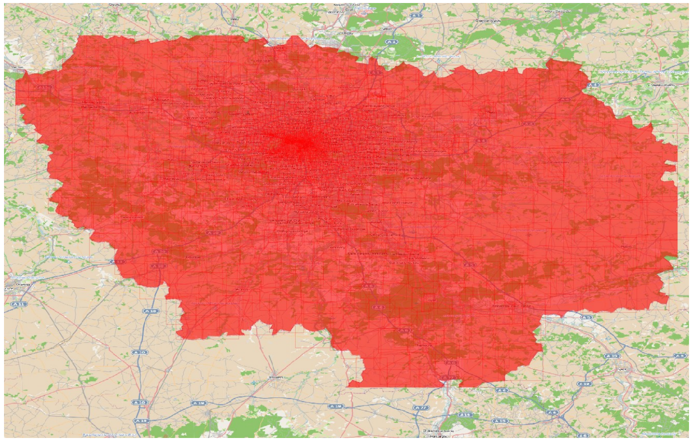
Step 1: estimate density of population

Use Call Detail Records to measure population

- Alternative: heuristics to identify stops

Each GSM tower associated to estimated coverage



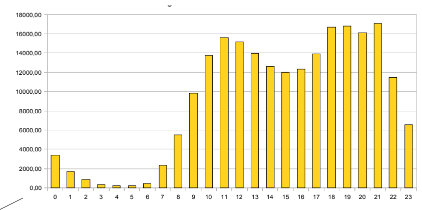


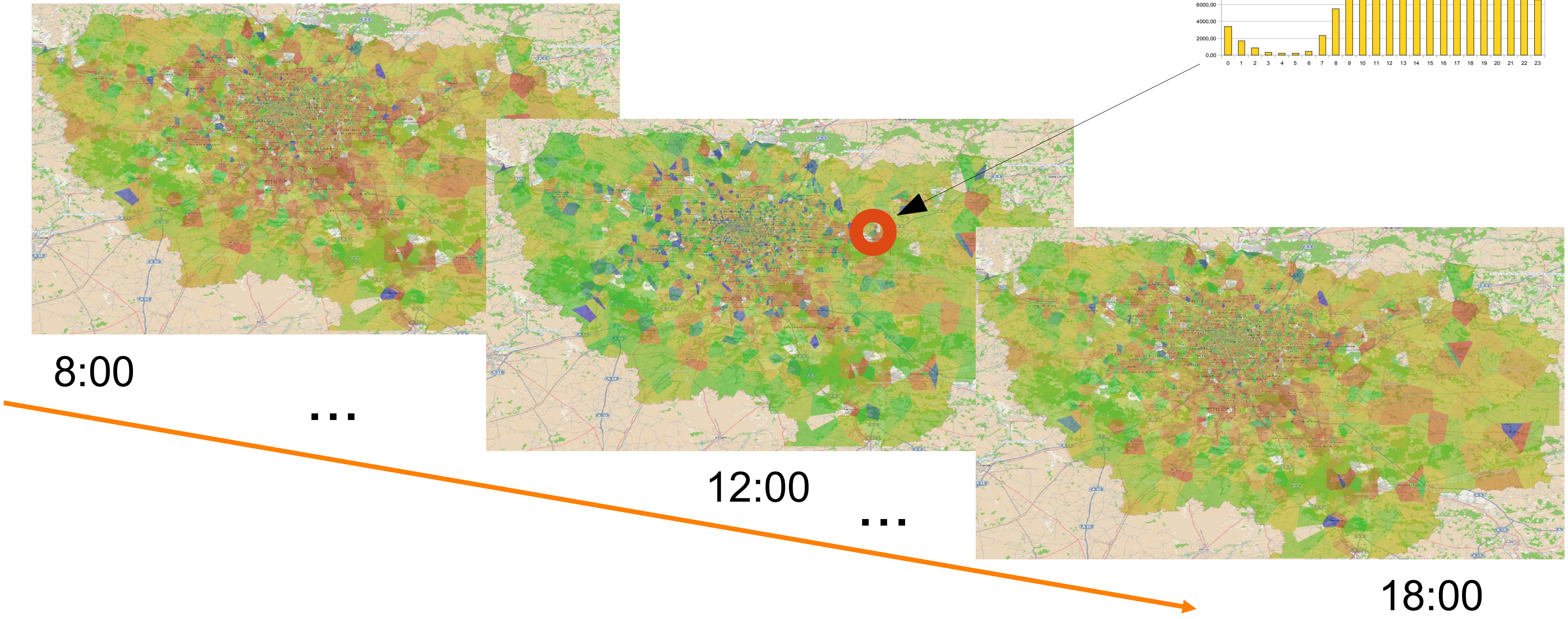
Aggregations adopted on larger-scale scenarios

Step 2: compute density over a space-time grid

Divide the dataset into days, and days into 24h

- ST grid = GSM cells x Hours

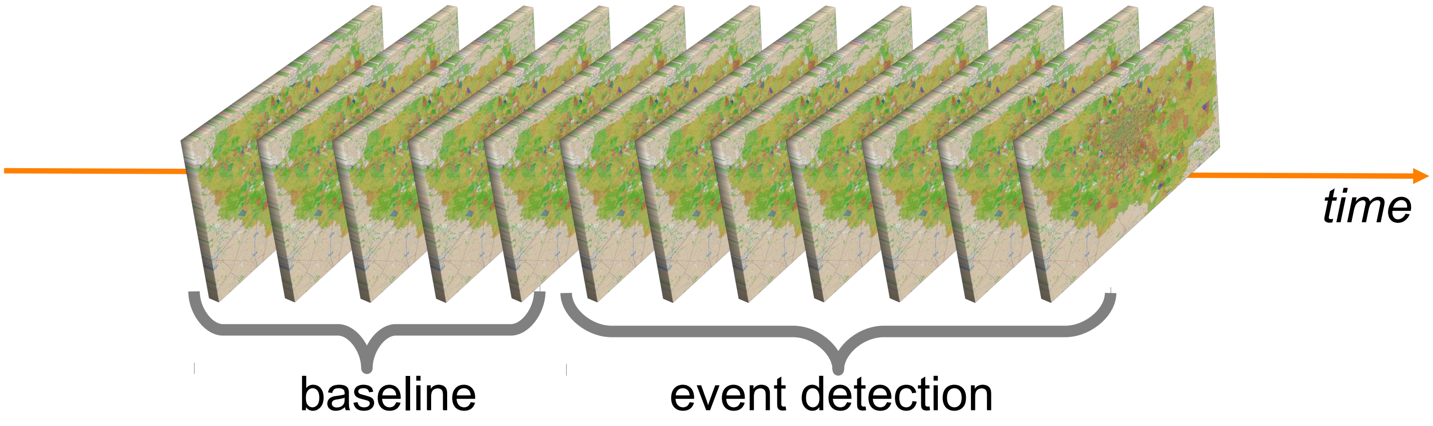




Step 3: detect events / 1

Split the dataset into temporal segments

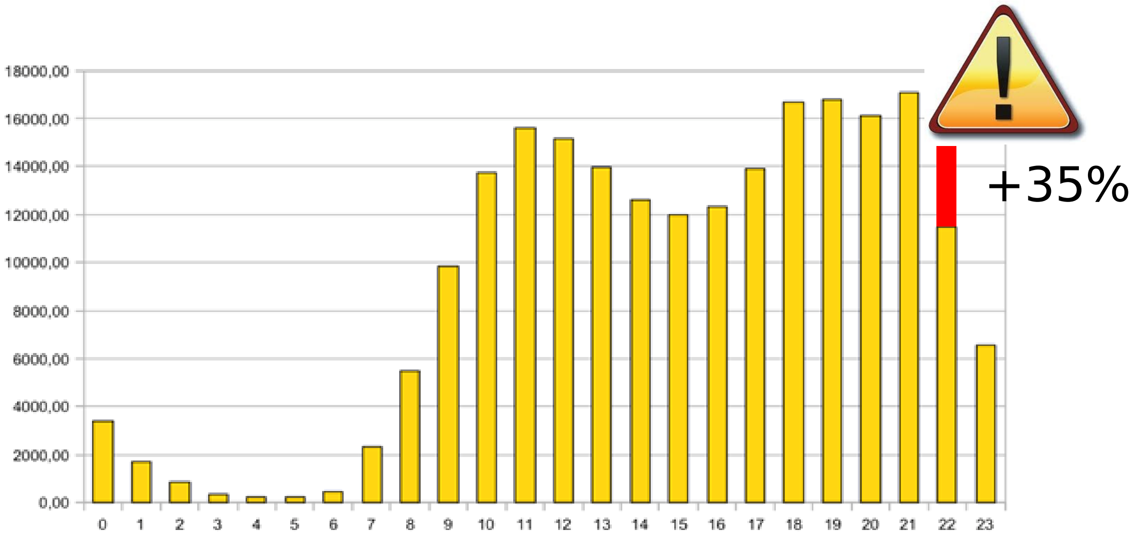
- **Baseline** segment: compute average density values for each hour of each day of the week
- **Event detection** segment: compare values against baseline to detect events



Step 3: detect events / 2

Event = significant deviation from average

- Deviations are discretized into bins (e.g., 5% bins)
- Deviations smaller than a threshold are neglected



Step 3: detect events / 3

Output: dataset of event sequences:

Day 1: {(Cell13,+20%),(Cell5,-15%)}$_{\text{1A.M.}}$ → {(Cell8,-20%)}$_{\text{2A.M.}}$ → ...

Day 2: {(Cell3,-30%)}$_{\text{1A.M.}}$ → {(Cell16,+20%)}$_{\text{5A.M.}}$ → ...

...

Day N: {(Cell270,-10%)}$_{\text{2A.M.}}$ → {(Cell71,+20%),(Cell5,-10%)}$_{\text{4A.M.}}$ → ...

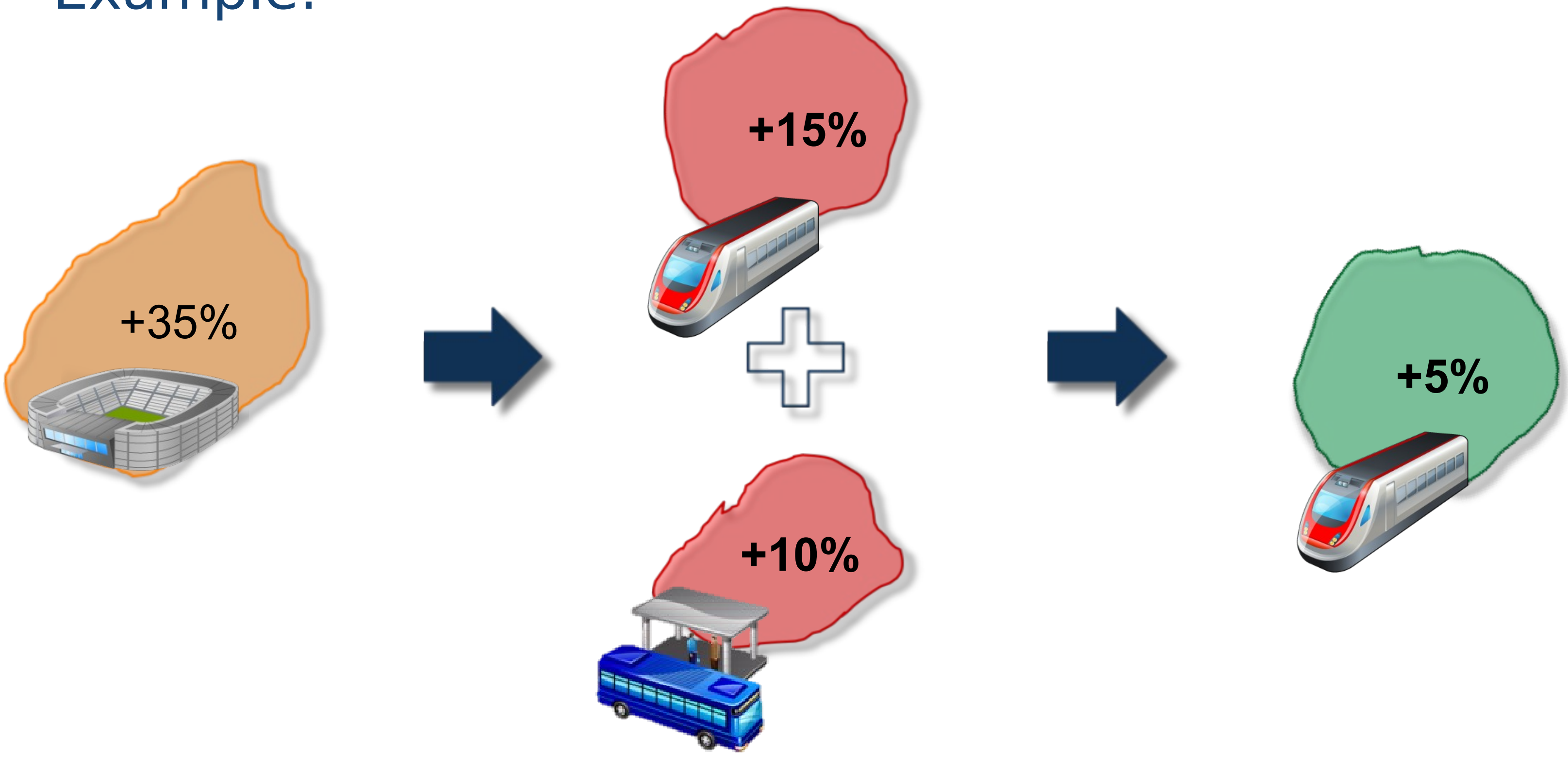
Step 4: correlation patterns/1

- Extract **frequent sequential patterns** of events
  - Frequent itemsets model relations between events that happen at the same time (co-occurrence)
  - Sequential patterns extend that by including ordered sequences of events (chain of events)
- Filter frequent patterns based on a **correlation index**:
  - Comparison against a simplified null model

$$c{-}index(D) = \frac{supp(D)}{\prod_i \prod_{d \in D_i} supp(d)}$$

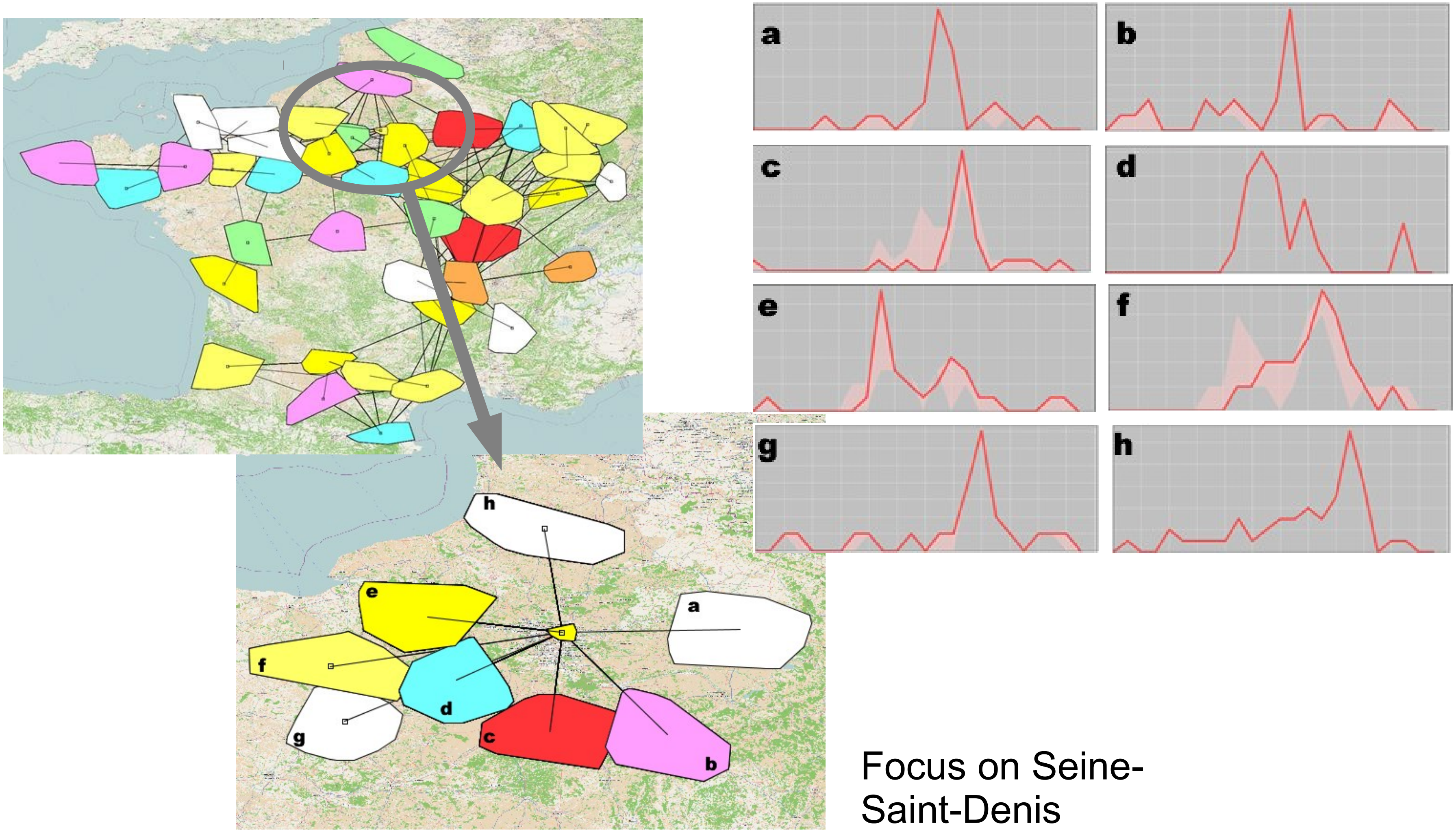
Step 4: correlation patterns/2

Example:



{(Cell27,+35%)} → {(Cell7,+15%),(Cell5,+10%)} → {(Cell13,+5%)}

National level example (departments)



Focus on Seine-Saint-Denis