GSM Data Mining

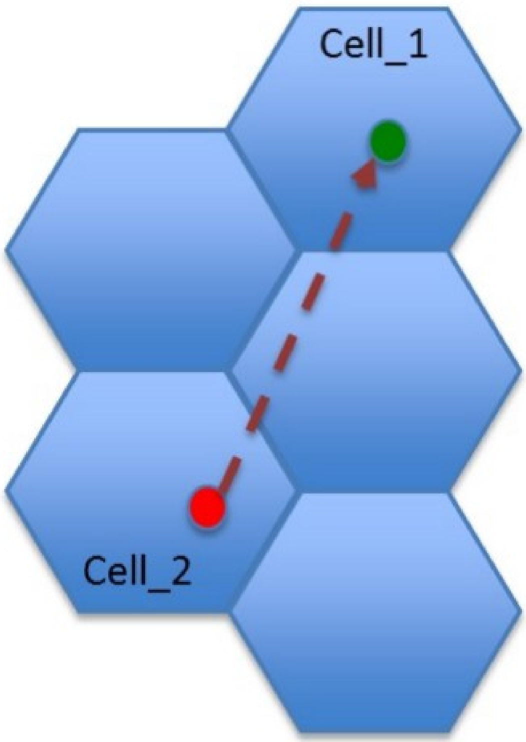
Knowledge Discovery and Data Mining Laboratory

Istituto di Scienza e Tecnologie dell'Informazione, CNR
Dipartimento di Informatica, Università di Pisa

Main data sources

CDR

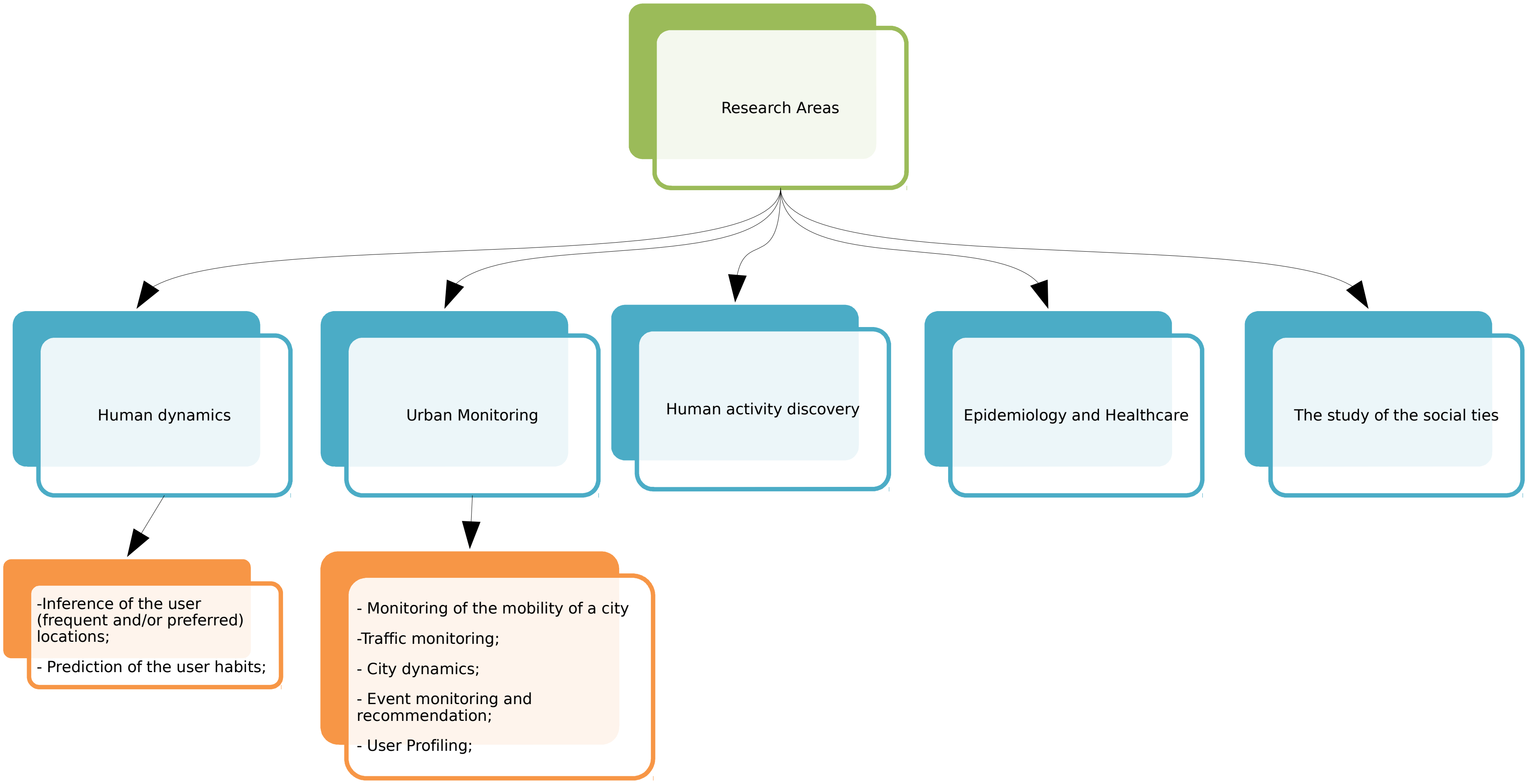
Who calls, **where** and **when**



Call Graph

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

Wide range of applications



Applications

- Tourism
  - Sociometro (Pisa and Cosenza)
  - Visits to attractions (Paris)
- Mobility
  - General laws for human mobility
  - D4D (Ivory Coast)
  - Persons & Places / ISTAT (Pisa)
- People and the territory
  - Presence of people & special events
  - Correlation patterns (Paris)
- Economic dimension
  - Mobility vs. Social vs. Economic status (Paris)
- Social ties
  - Link prediction and mobility

Tourism

City users Sociometer

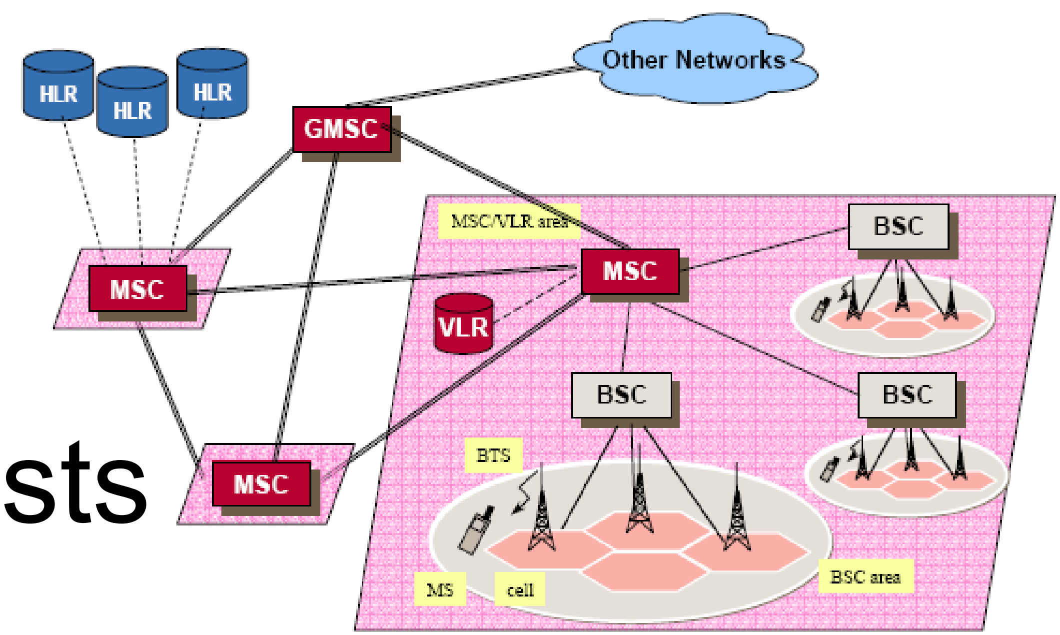
Mobile phone socio-meters

Analyze individual call habits to recognize profiles

–Resident

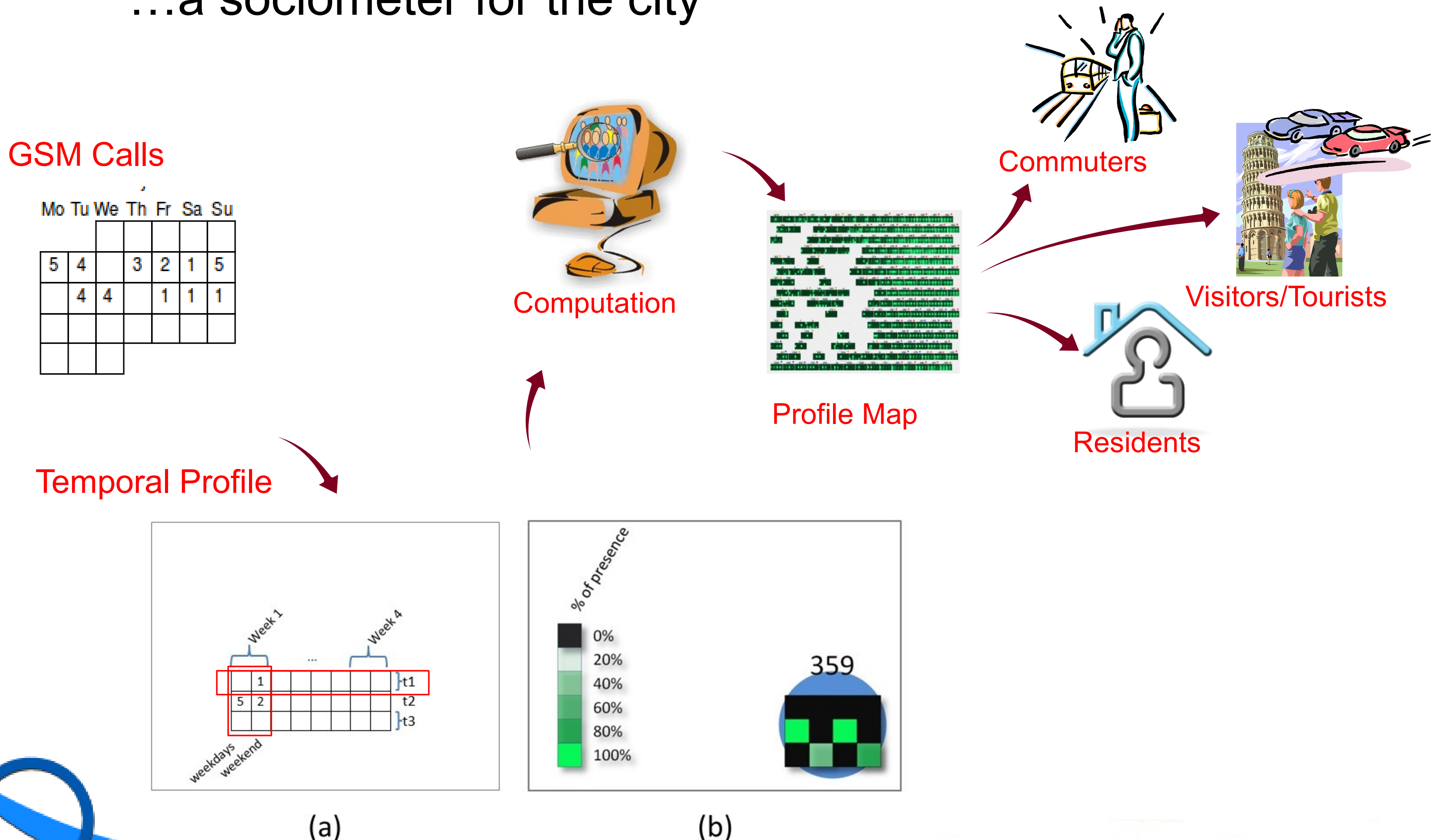
–Commuters

–Visitors/Tourists



GSM: People Profiling

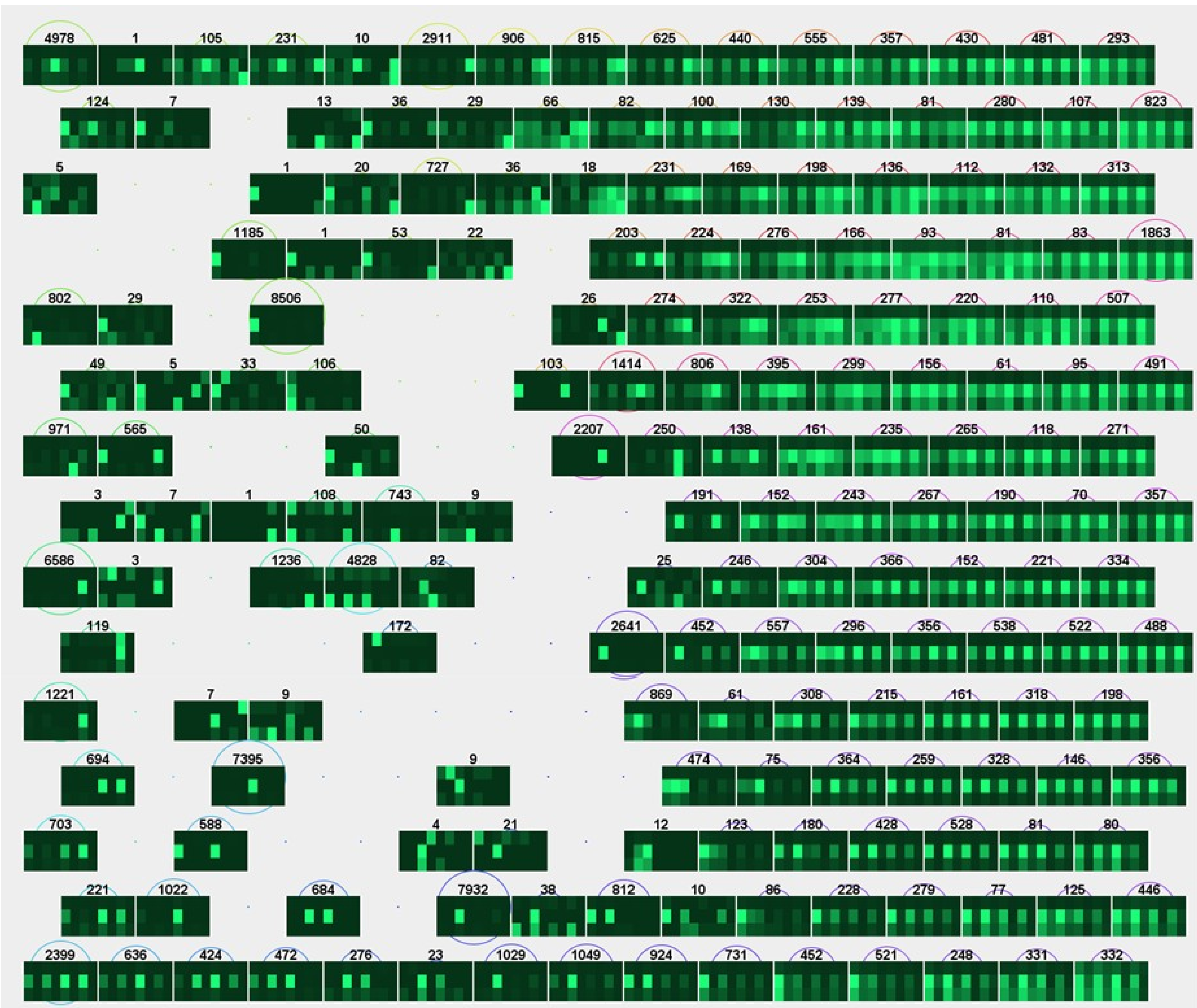
…a sociometer for the city

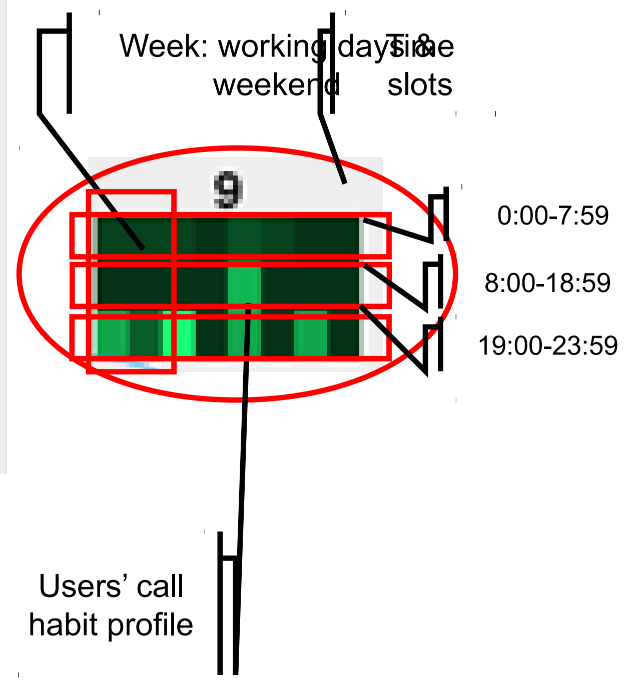


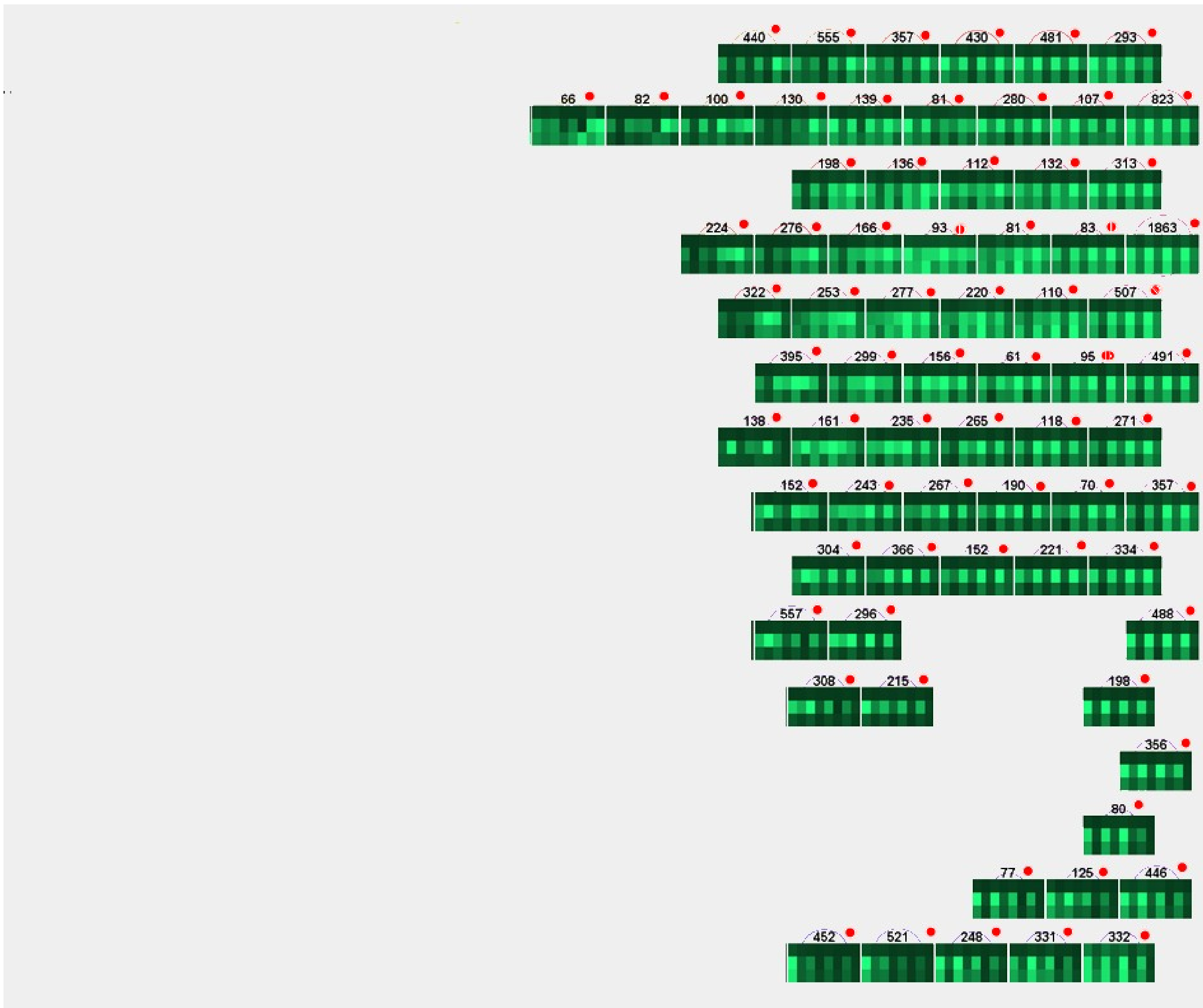
Top-down analysis

- **Resident**
  - C1 - Temporal range: at least 1 call in [19:00 - 6:59] during the weekdays.
  - C2.1 - Daily presence: at least 2 distinct weekdays per week, that satisfy C1.
  - C2.2 - Daily presence: at least 1 day in the weekend without temporal range.
  - C3 - Weekly presence: at least 3 weeks, in which C1, C2.1 and C2.2 are satisfied.
- **Commuter**
  - C1.1 - Temporal range: at least 1 call in [9:00 - 18:59] during the weekdays.
  - C1.2 - Temporal range: no calls in [19:00 - 8:59] during the weekdays.
  - C2.1 - Daily presence: at least 2 distinct weekdays per week, that satisfy C1.1 and C1.2.
  - C2.2 - Daily presence: never during the weekends.
  - C3 - Weekly presence: at least 3 weeks, in which C1.1, C1.2, C2.1, C2.2 and C3 are satisfied.
- **People in Transit**
  - C1 - Temporal range: calls during at most 1 hour.
  - C2 - Daily presence: at most 1 day in which C1 is satisfied.
  - C3 - Weekly presence: at most 1 week, in which C1 and C2 are satisfied.

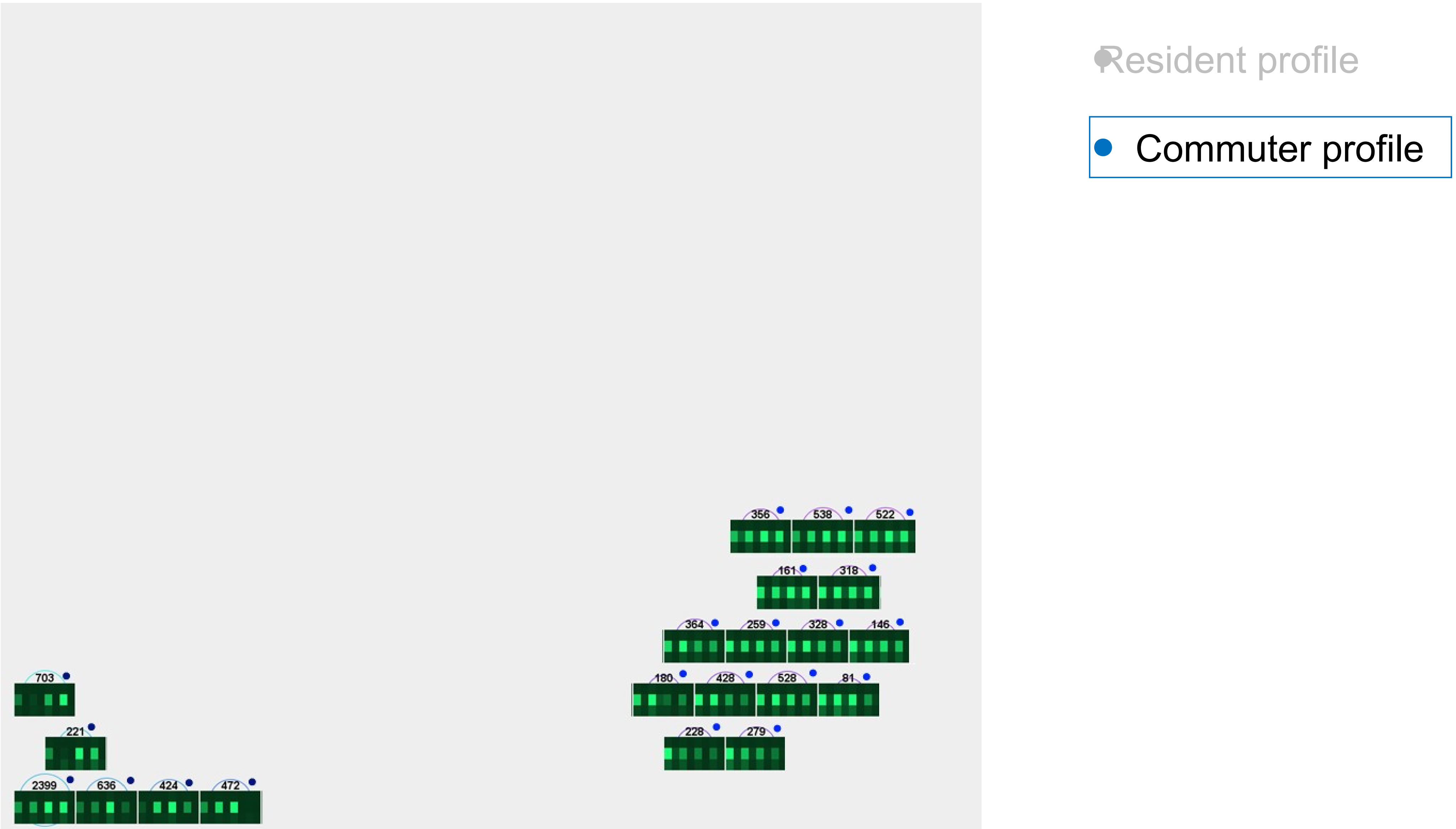
Call Habit Profiles



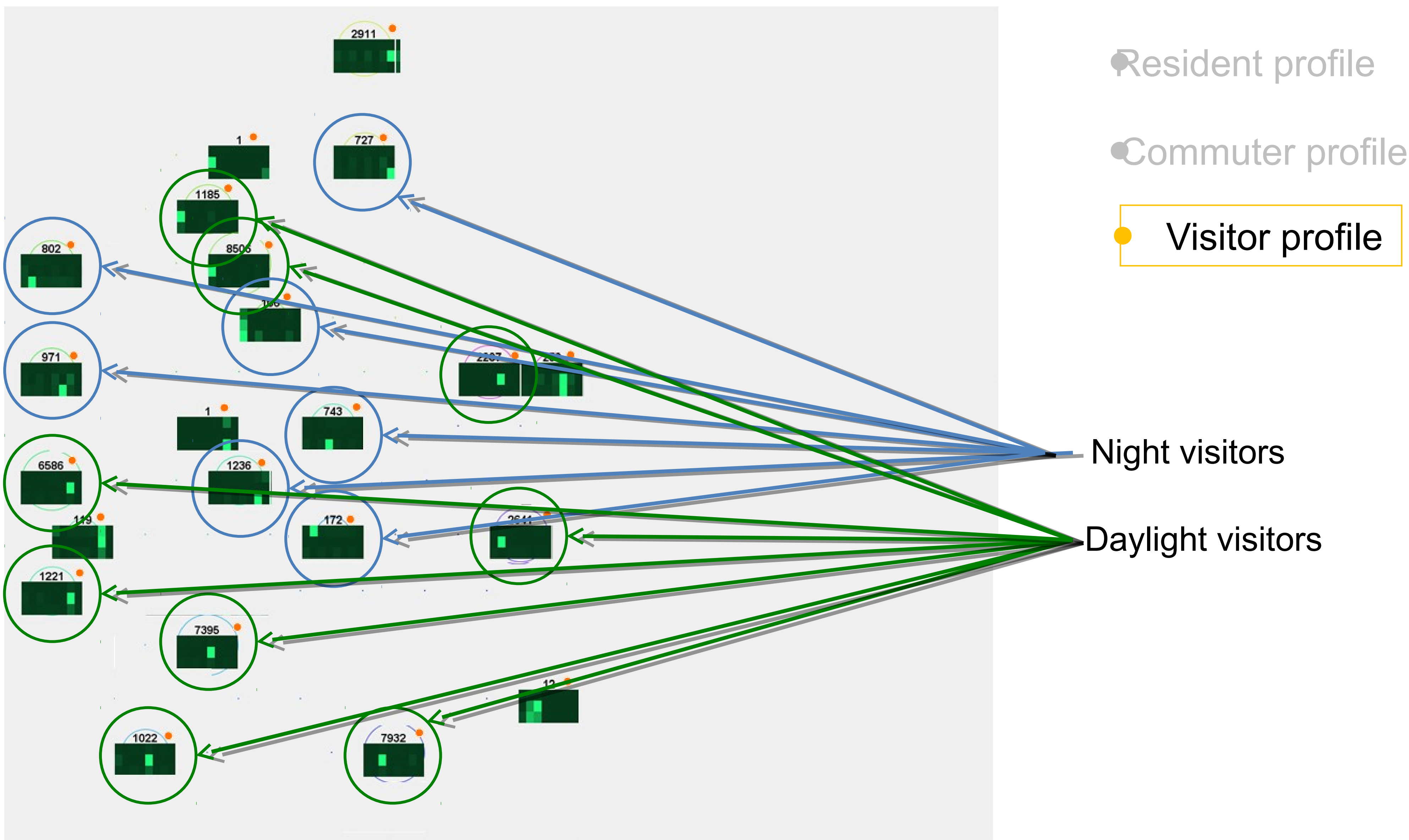




● Resident profile



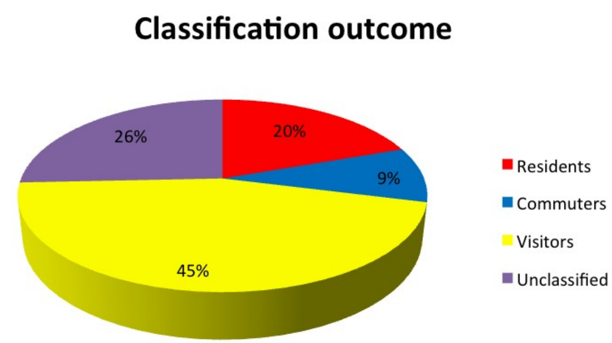
356
538
522
161
318
364
259
328
146
180
428
528
81
228
279
703
221
2399
636
424
472
Resident profile
Commuter profile



Resident profile
Commuter profile
Visitor profile
Night visitors
Daylight visitors

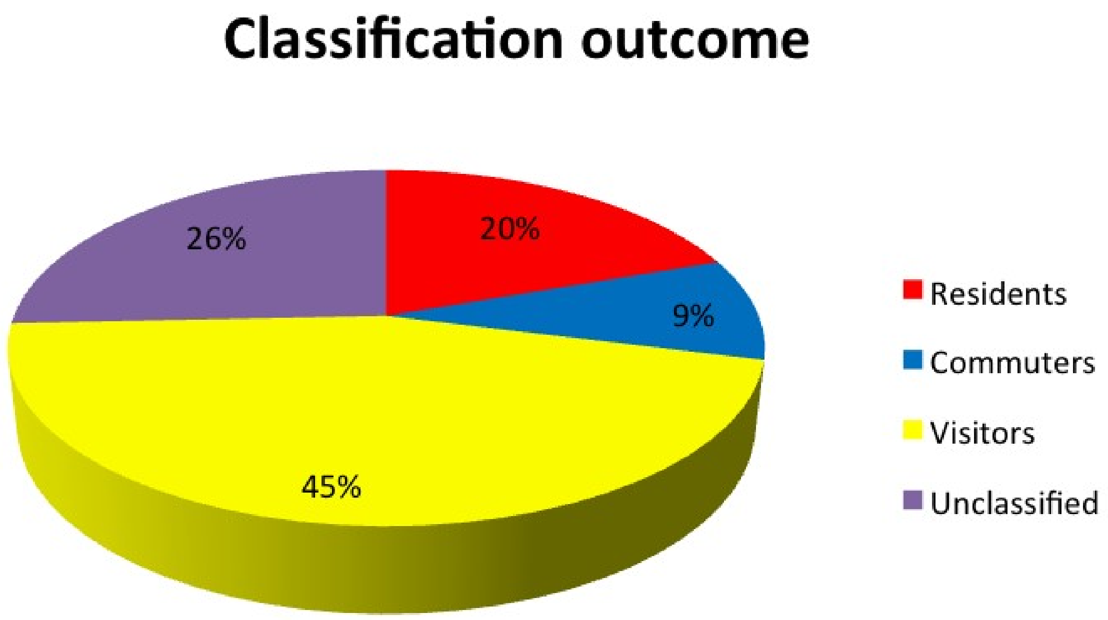
User profile **quantification**

- Resident profile
- Commuter profile
- Visitor profile



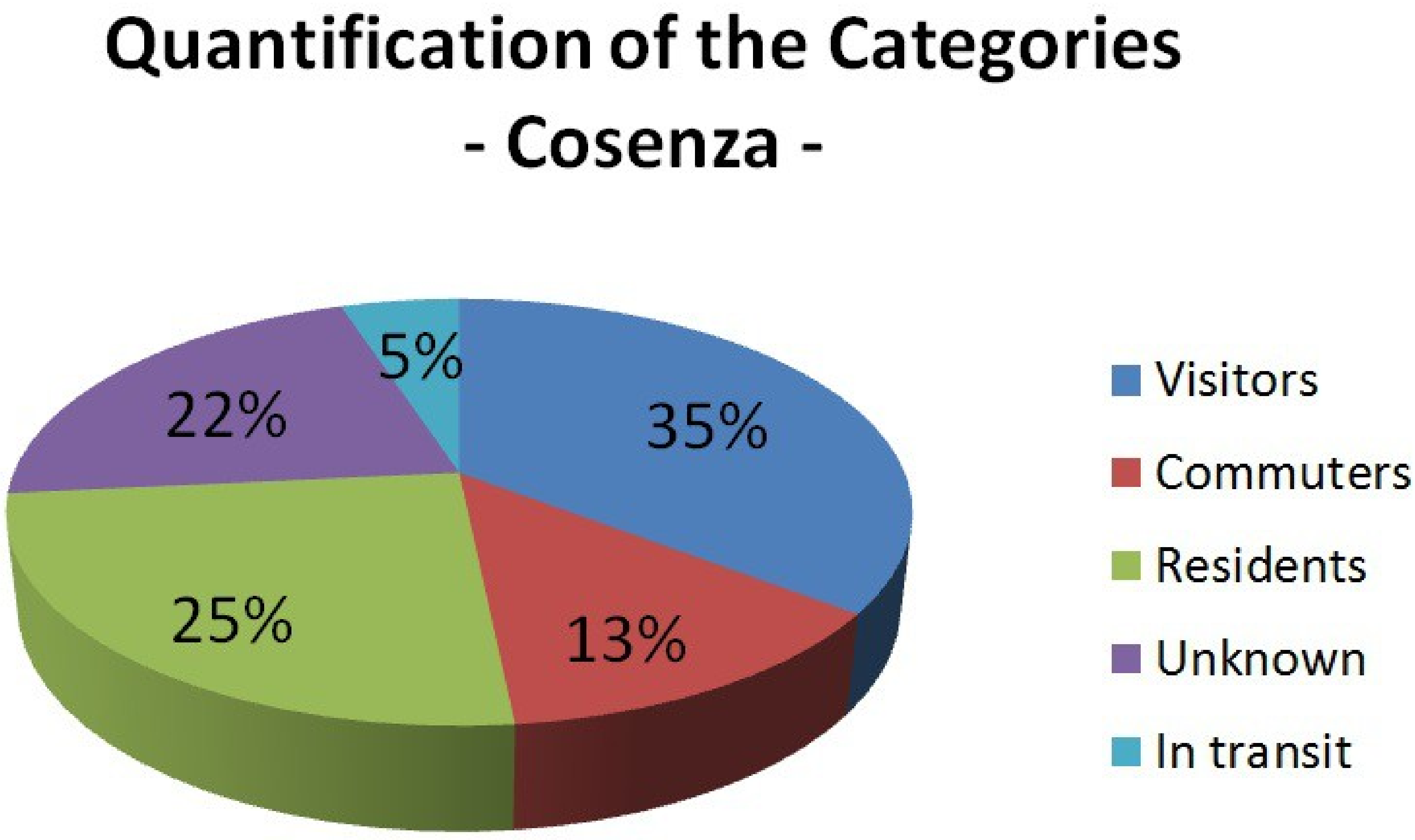
Urban Sociometer indicator: Pisa

Analysing the GSM call habits in Pisa we can find indicators of social profiles

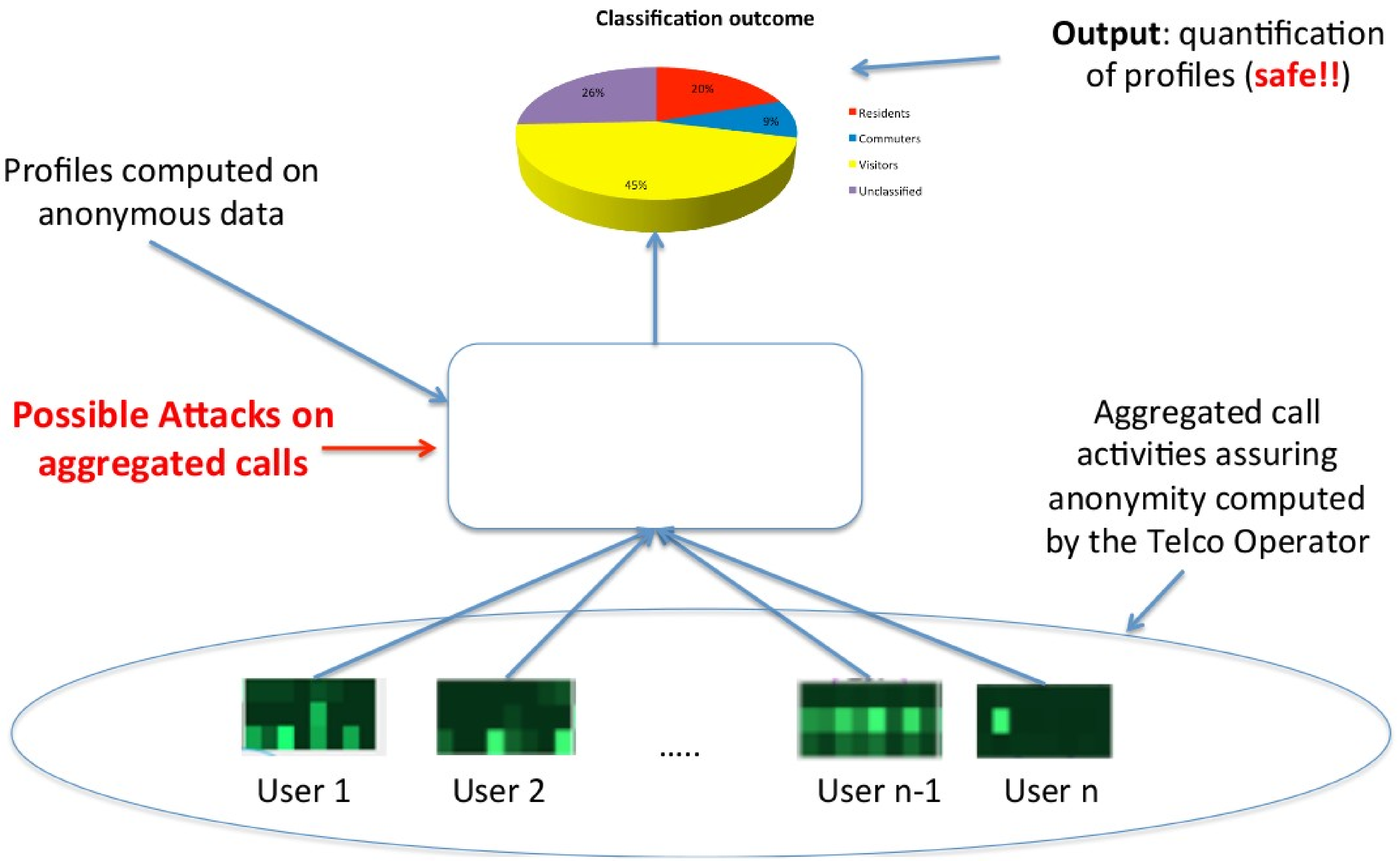


Pisa january 2012

Urban Sociometer indicator: Cosenza – South of Italy



Privacy-Aware socio-meter



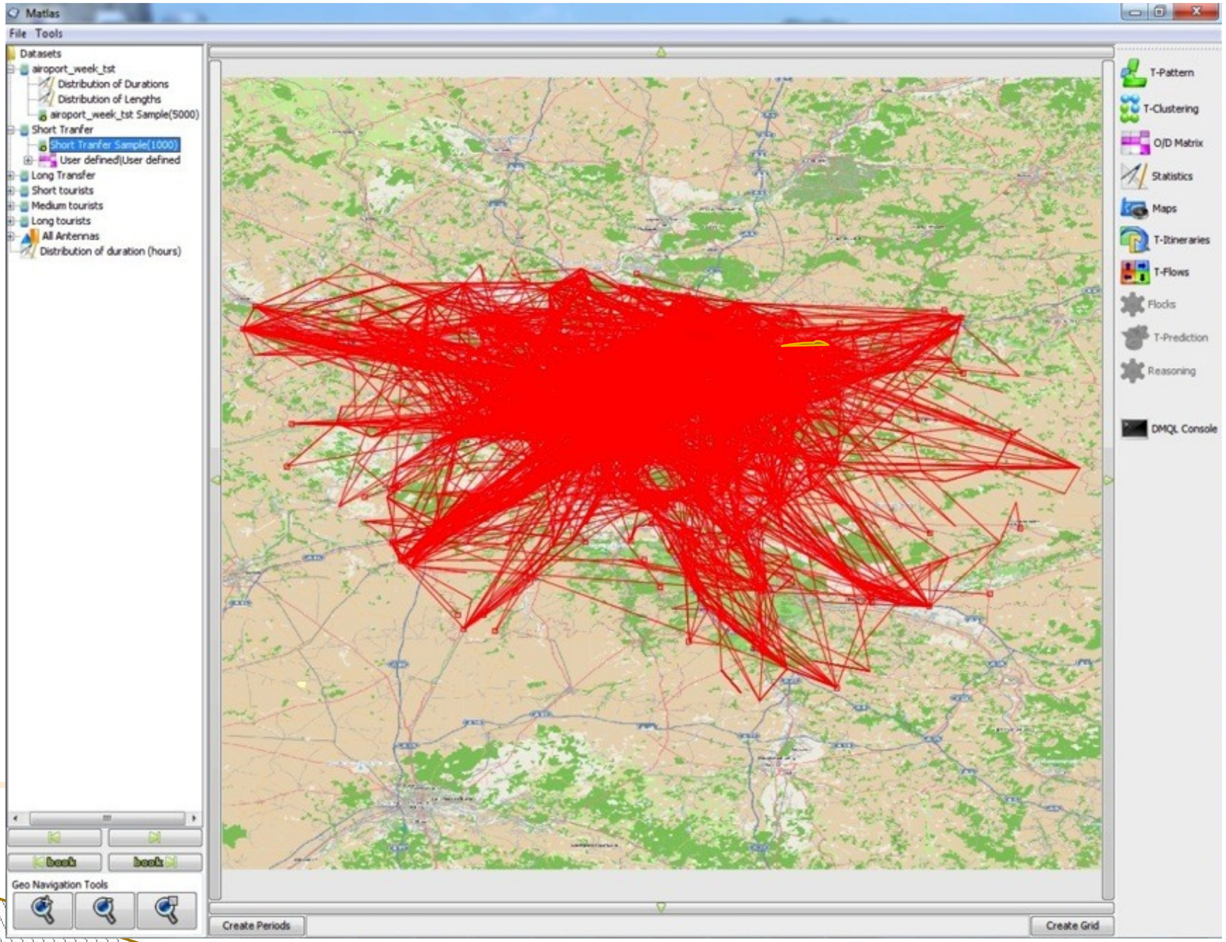
GSM data analysis for tourism application

Ana-Maria Olteanu, Roberto Trasarti, Thomas Couronné, Fosca Giannotti, Zbigniew Smoreda, Mirco Nanni, Cezary Ziemlicki

Analyzing tourist data

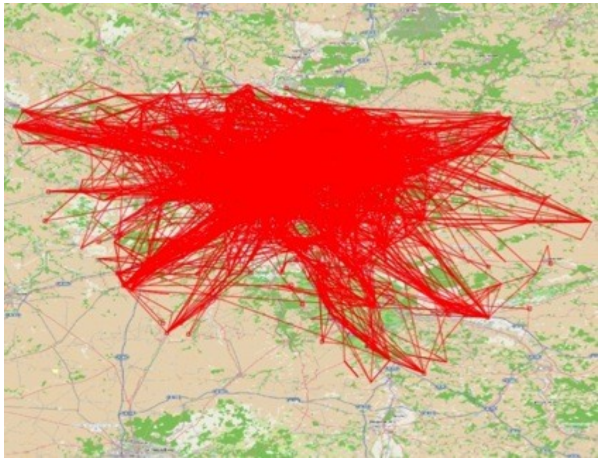
We extracted the foreign (not French) users arriving and leaving at CDG airport in order to classifying them and study their behaviors.

106 000 Users



Distribution of visiting time

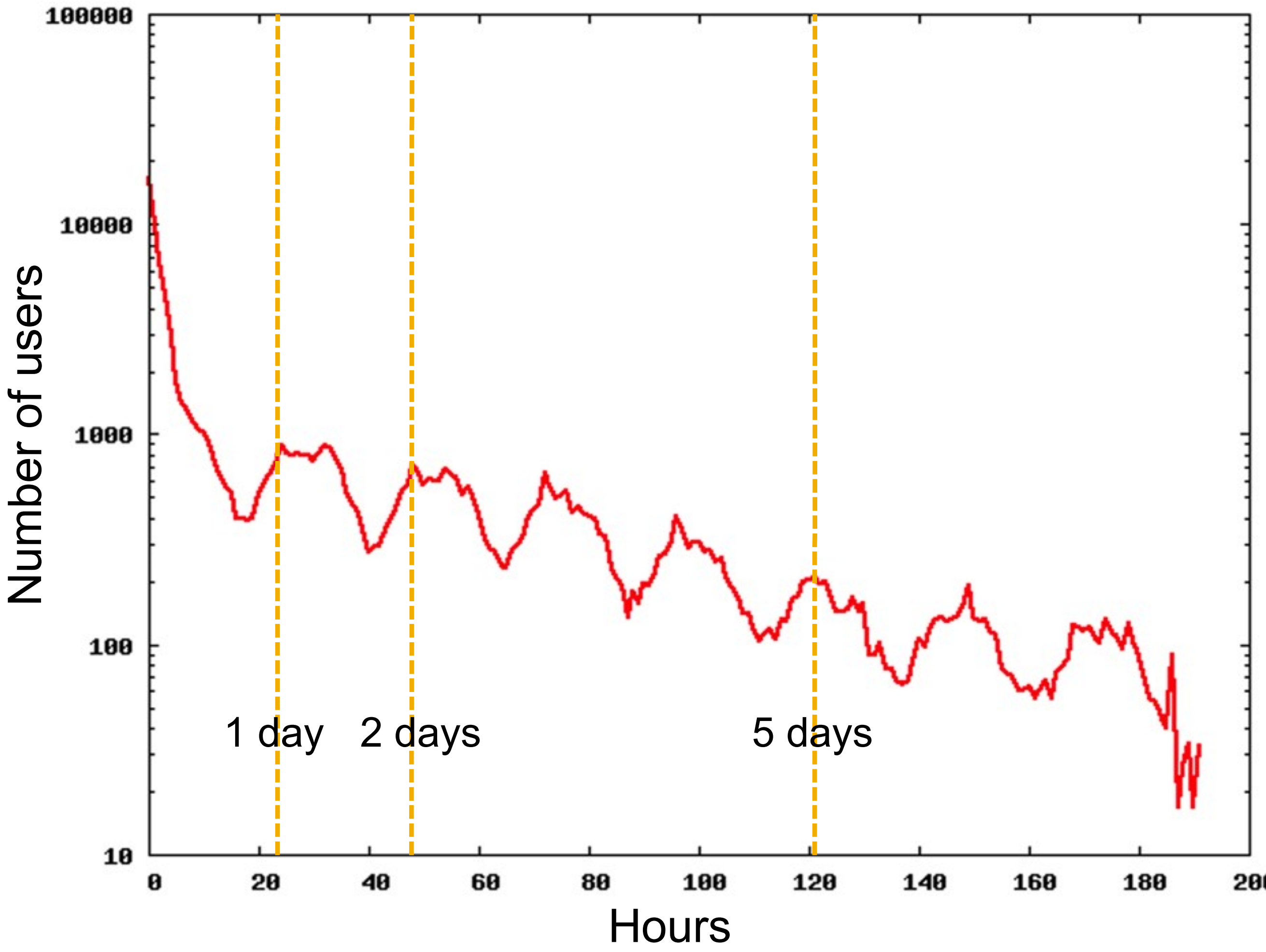
We are interested on the time spent by the tourists in Paris, thanks to the selection of CDG users, we can be sure that the information is complete avoiding disappearances.



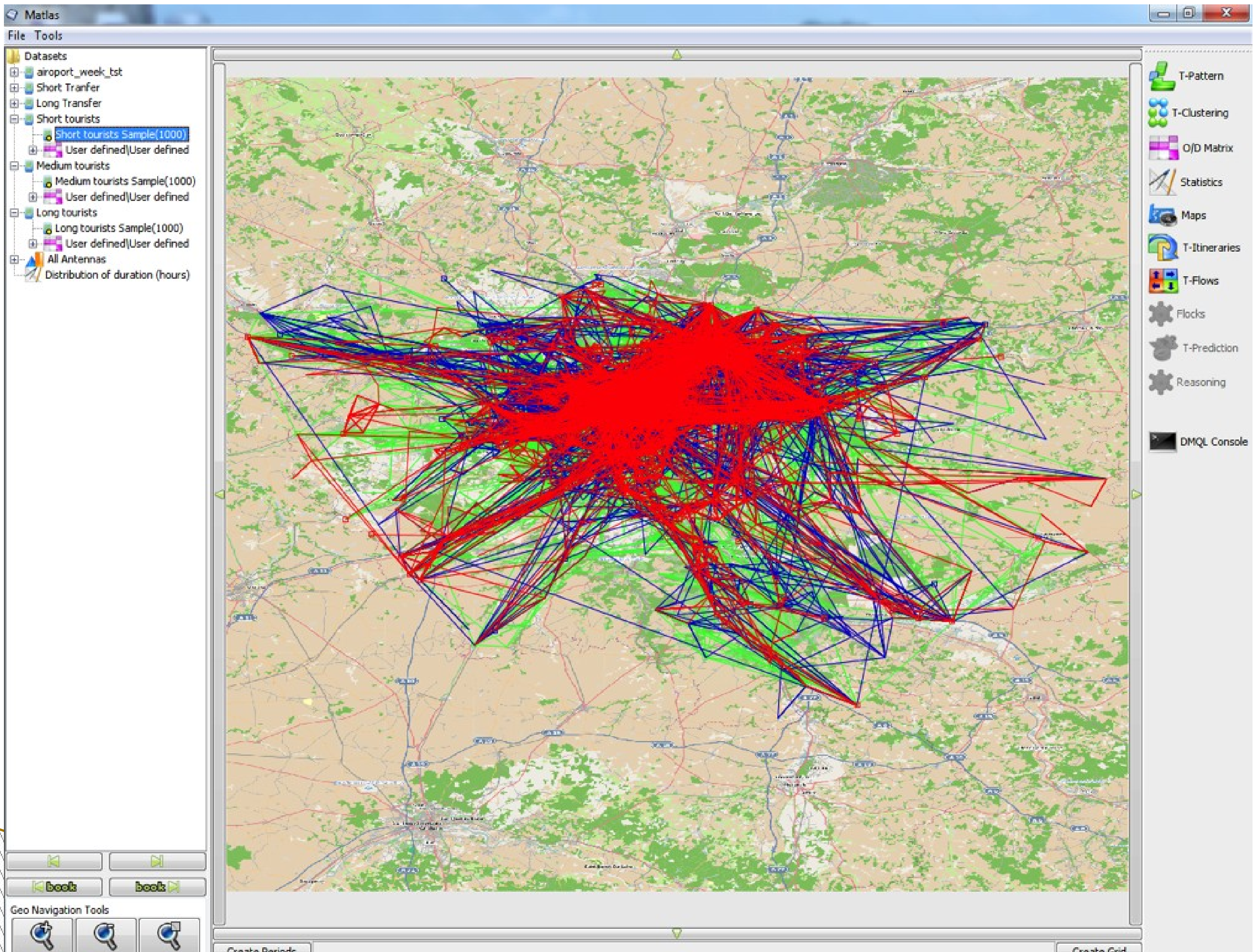
t0

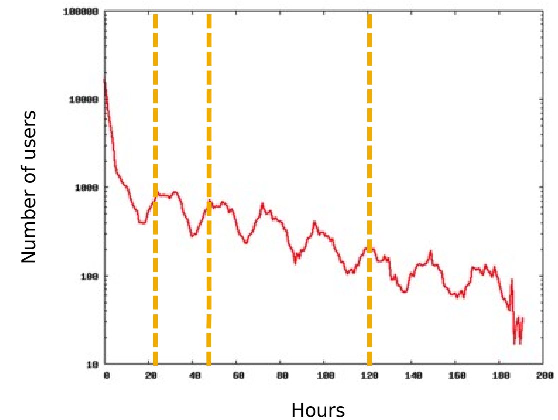
tn

tn-t0 = Visiting time



Categorization of tourists

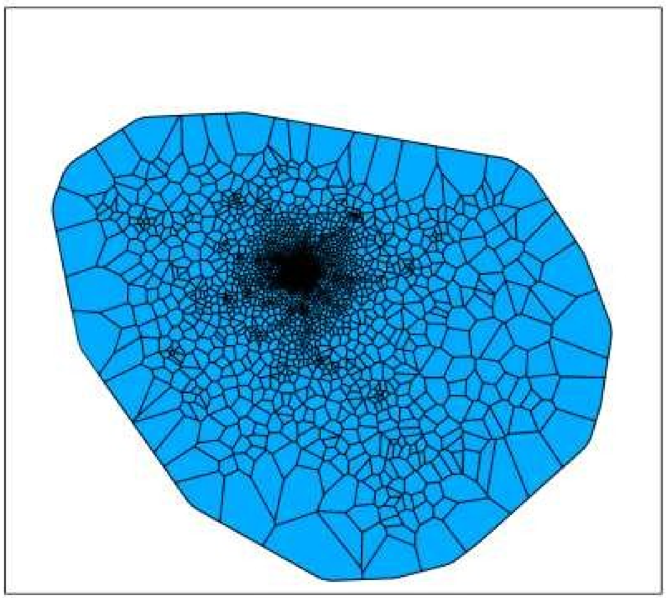


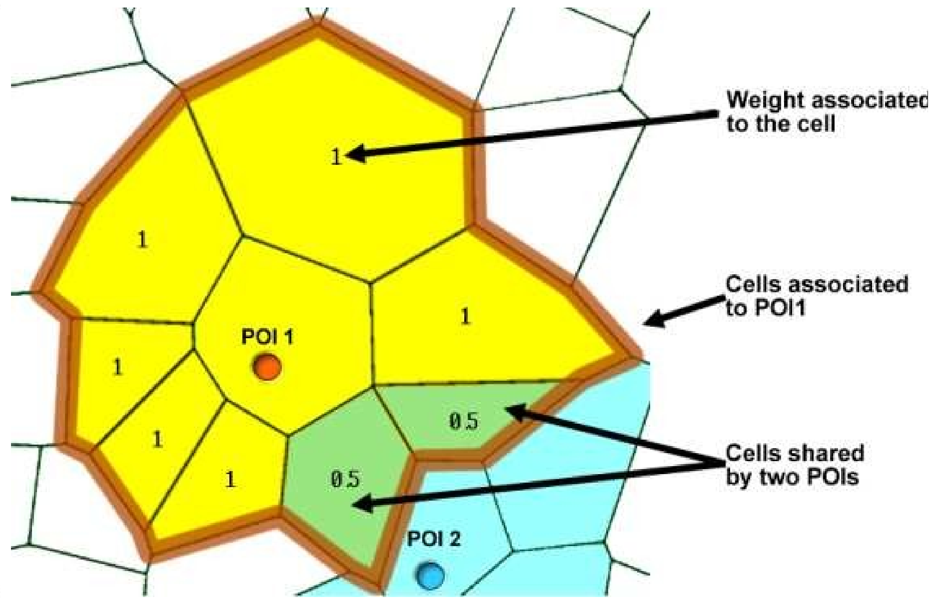


Short period stay Tourist (1 day – 2 days)
Medium period stay Tourist (2 day – 5 days)
Long period stay Tourist (5 day – 7 days)

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:

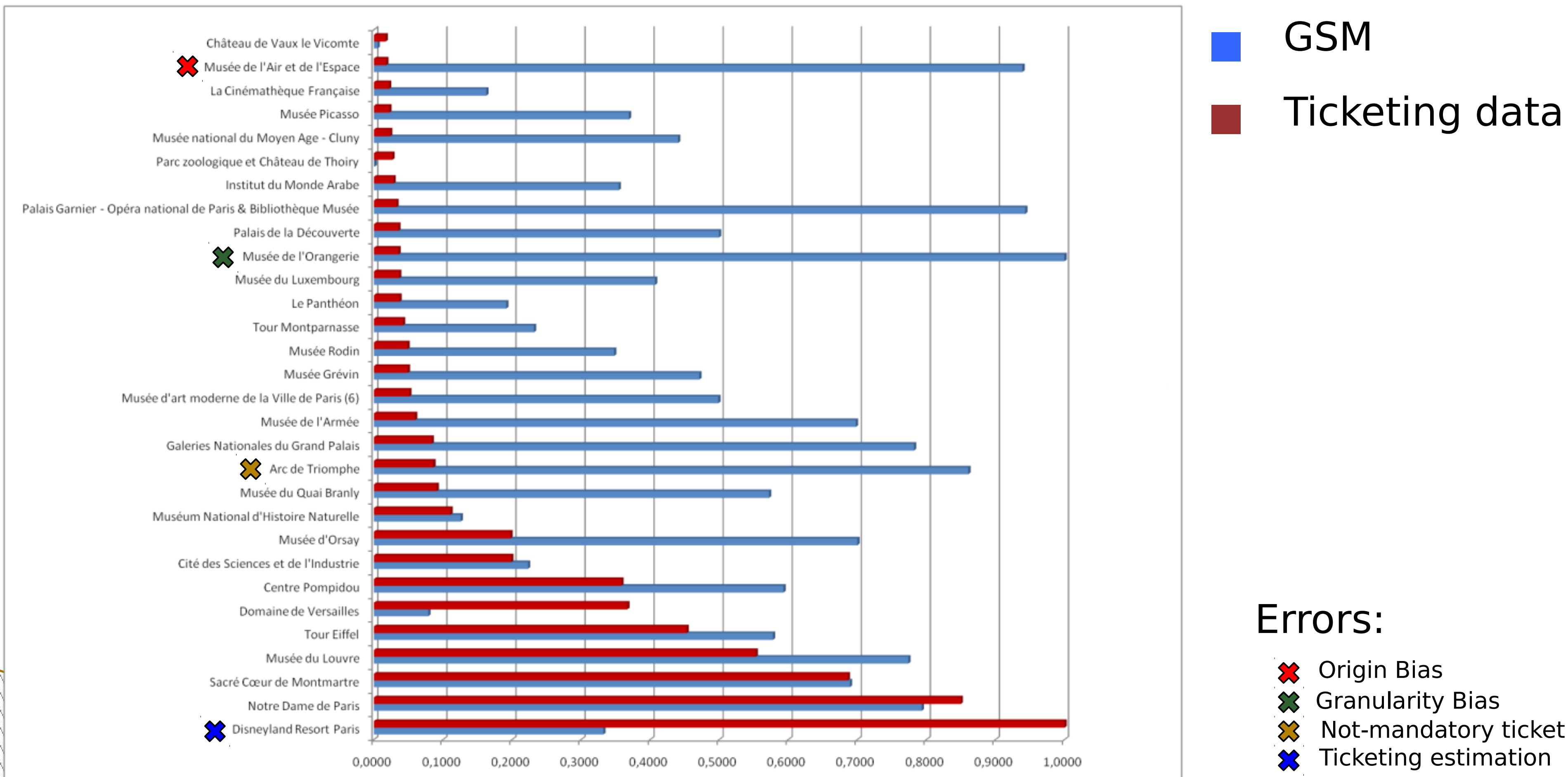




Weight = 1/#neighboring POIs

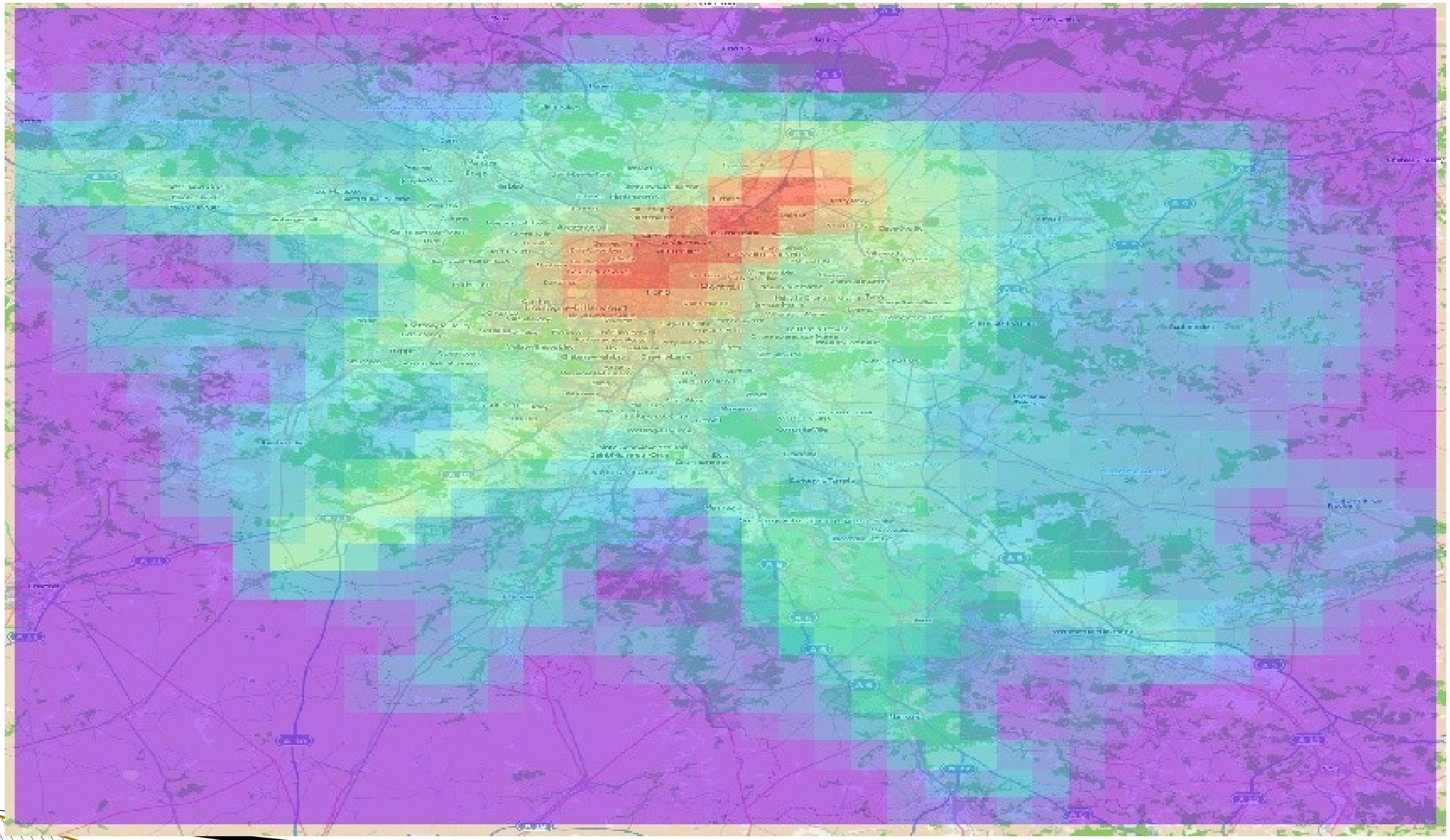
Comparison with Ticketing data

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.



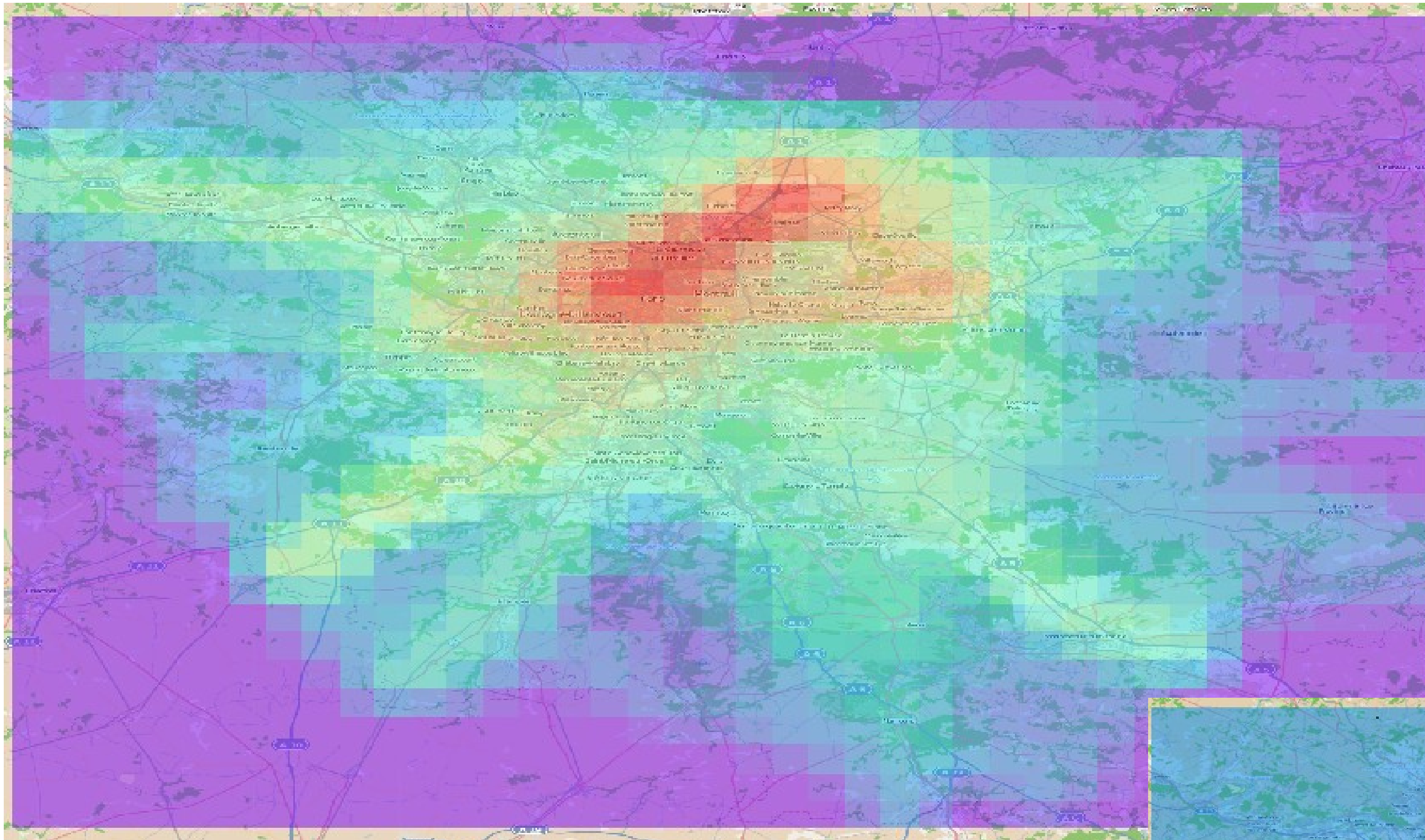
Density map (Short stay)

Having the movements of the users in Paris we can compute a density map of them in space trying to discover they behavior.



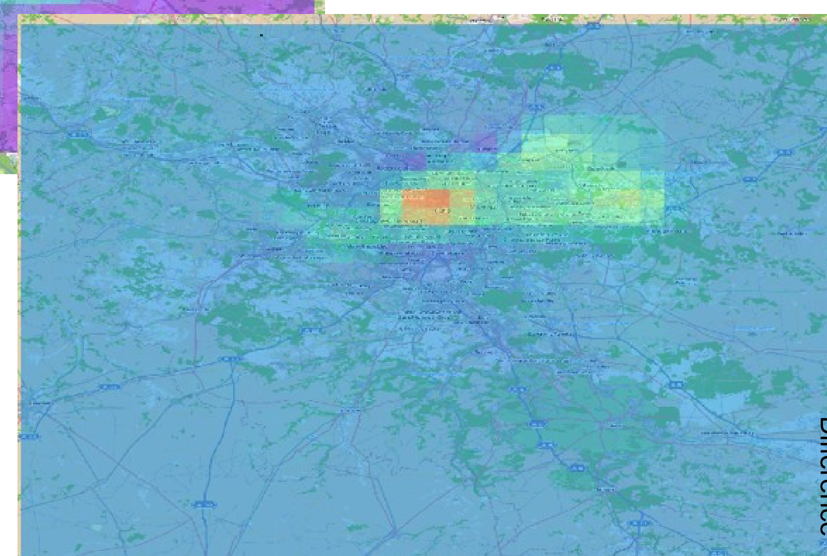
Short stay tourists visit the very center of Paris and go back the airport to leave.

Density map (Medium stay)

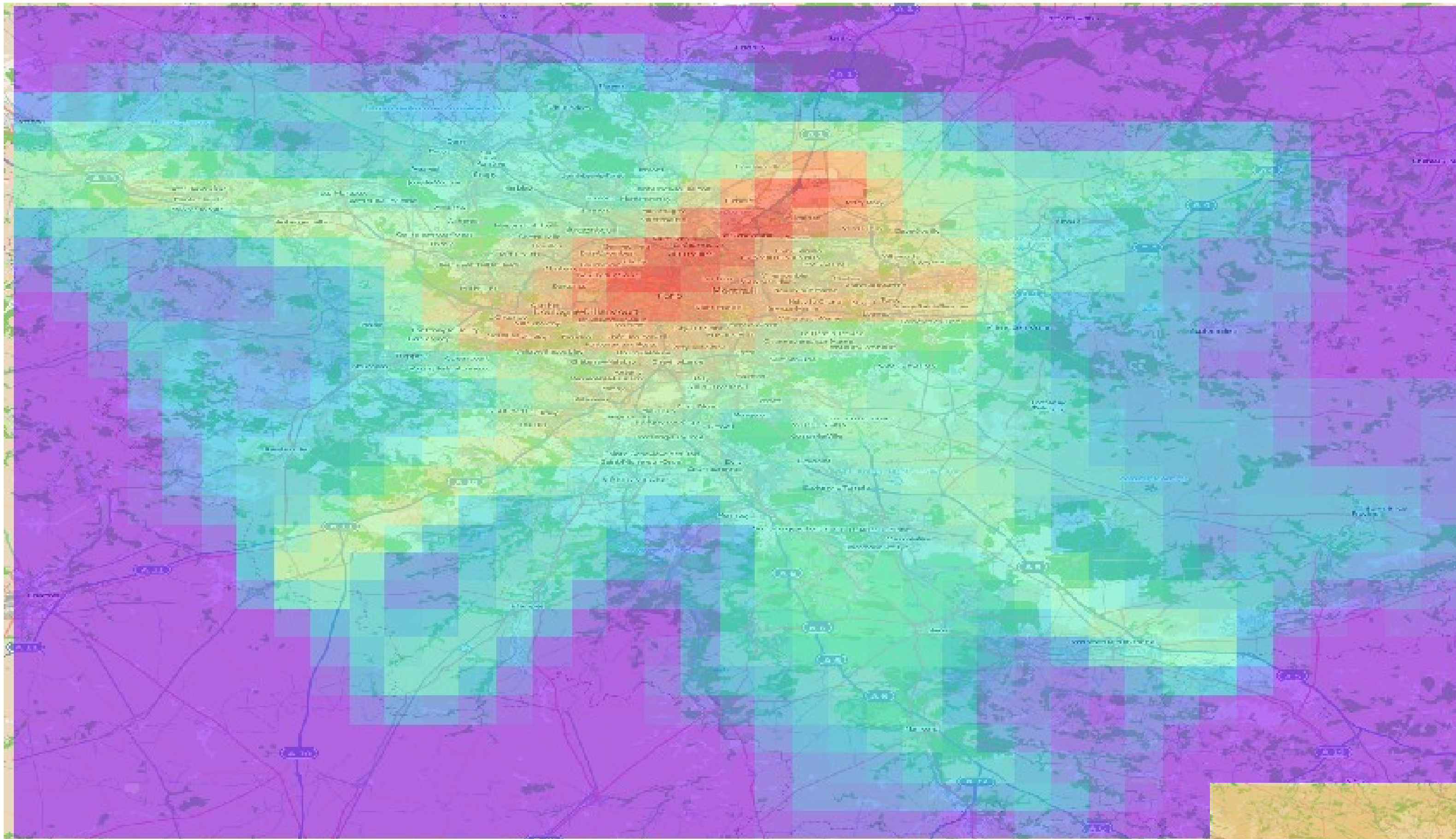


Medium stay tourists visit the center of Paris mostly but Versailles and Disneyland appear as new destinations

Green = Disneyland Paris
Red = Versailles

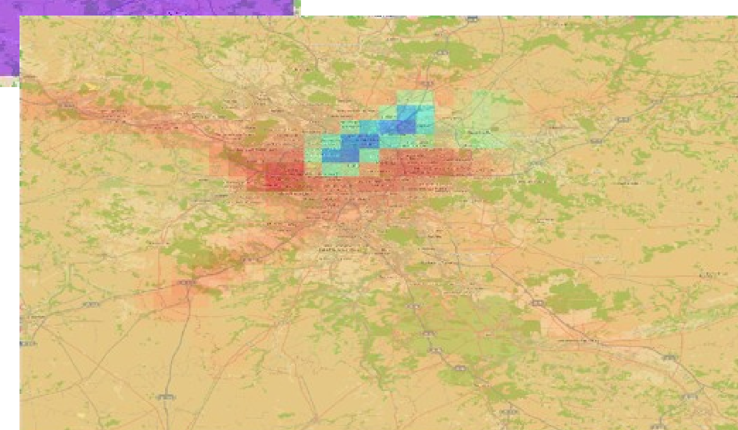


Density map (Long stay)



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

Mobility

D4D Challenge

Data for Development

orange

NetMob 2013

MP4-A Project: Mobility Planning For Africa

A joint work of KDD Lab & mobility consultants Goudappel Coffeng

kdd.isti.cnr.it www.goudappel.nl

KDD Lab: *Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli*

adviseurs mobiliteit Goudappel Coffeng: *Peter Van Der Mede, Joost De Bruijn, Erik De Romph, Gerard Bruil*

The Challenge

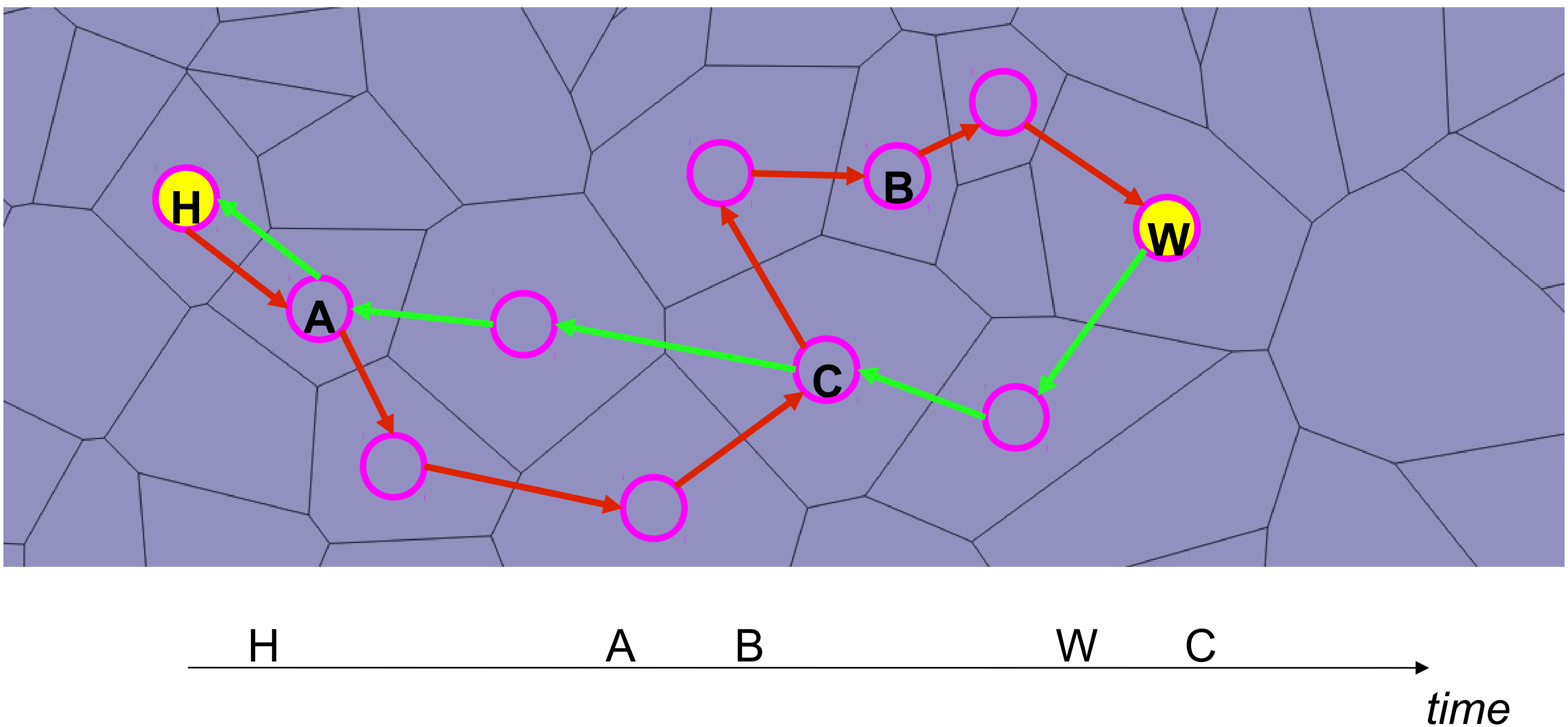
- Incompleteness issue
  - Call Detail Records describe the location of users only during activity (calls, messages)
  - Most individual mobility might be invisible
- Lack of semantics
  - No information about activities and purpose
- Spatial uncertainty issue
  - Location described in terms of cells having dynamic and sometimes large extent

The approach (summary)

- Analyze raw GSM data to
  - infer systematic mobility of individuals
- Build origin-destination matrices
  - Describe (expected) flows between areas
- Build a transportation model
  - Assigns O/D matrix to OSM road network through OmniTRANS system

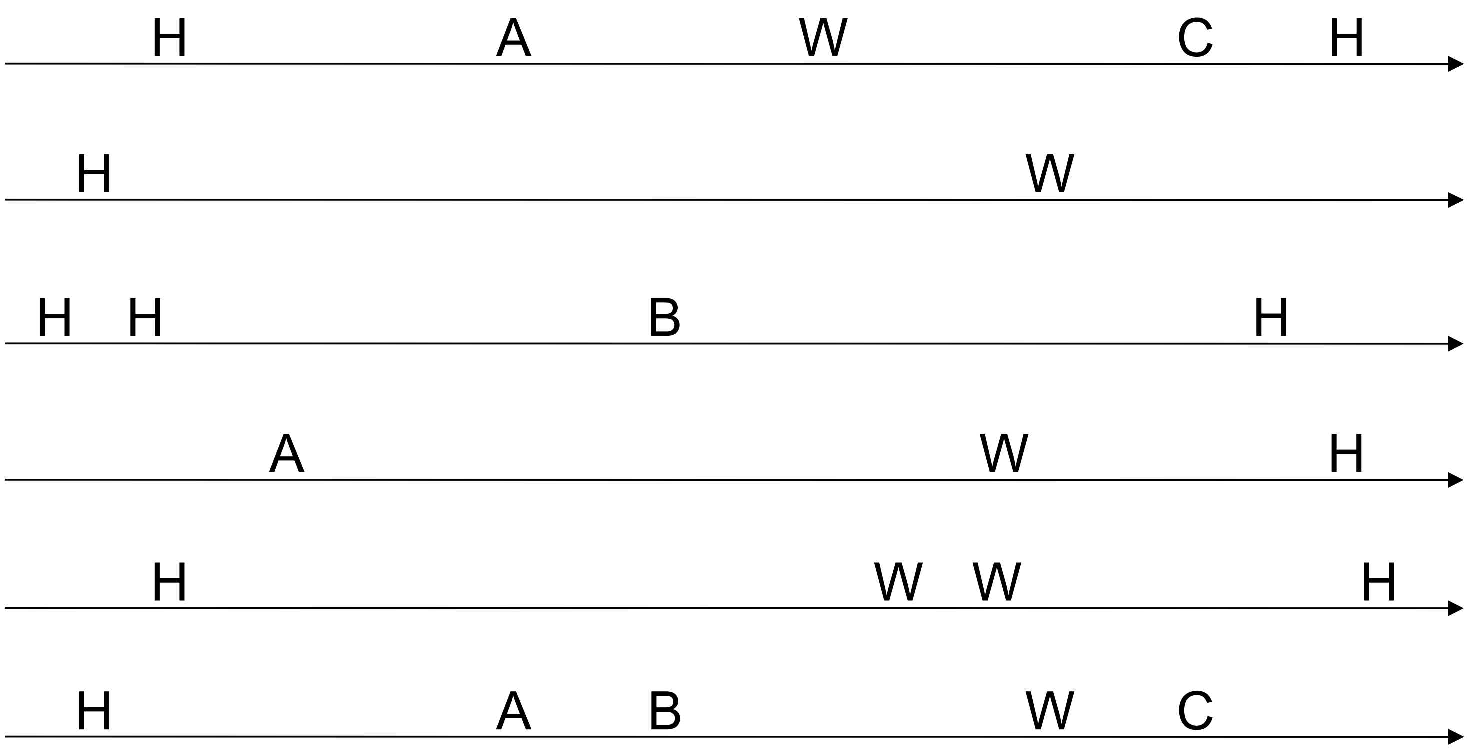
Systematic mobility

- A single trace of an individual can be poorly informative about his/her movements



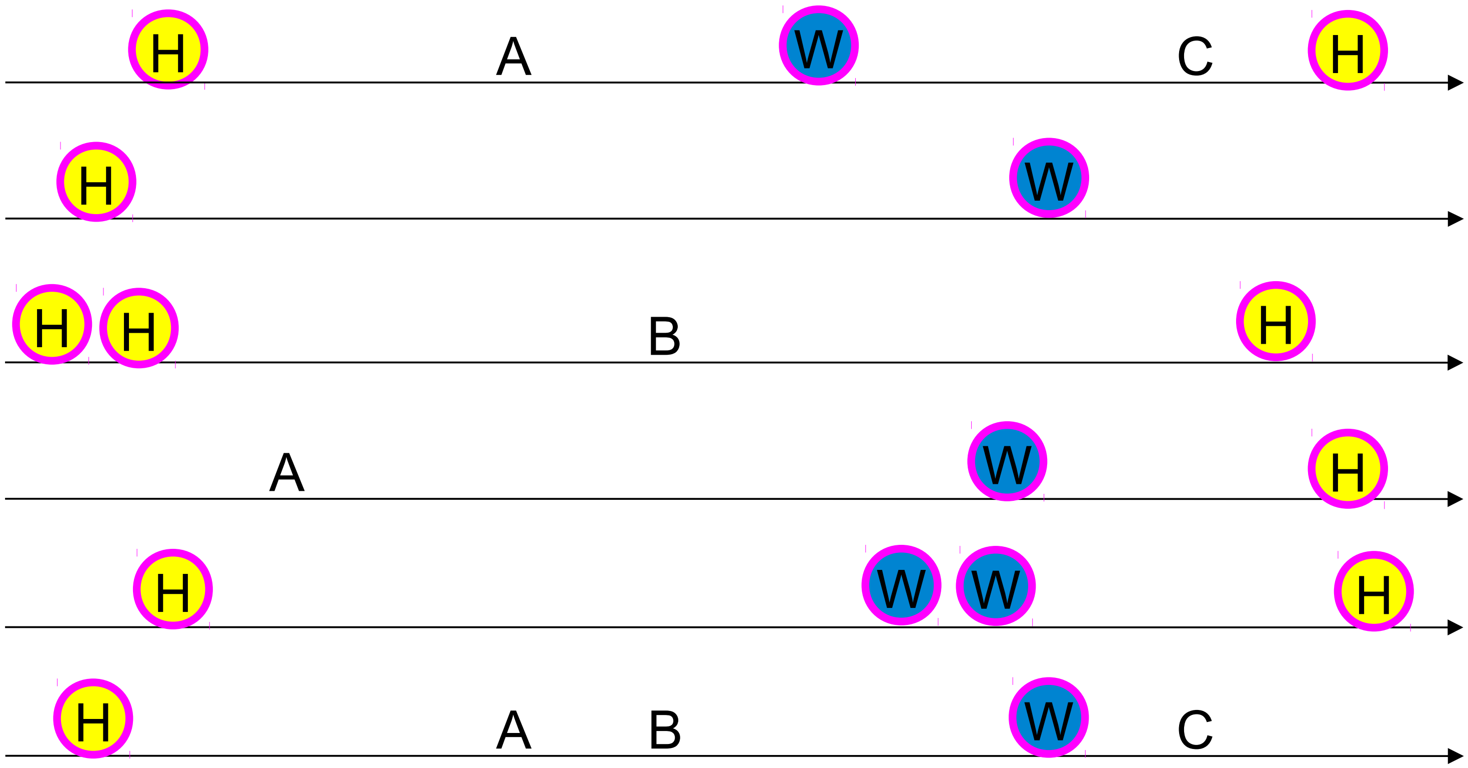
Systematic mobility

- Yet, several daily traces of the same individual might allow to identify regular places



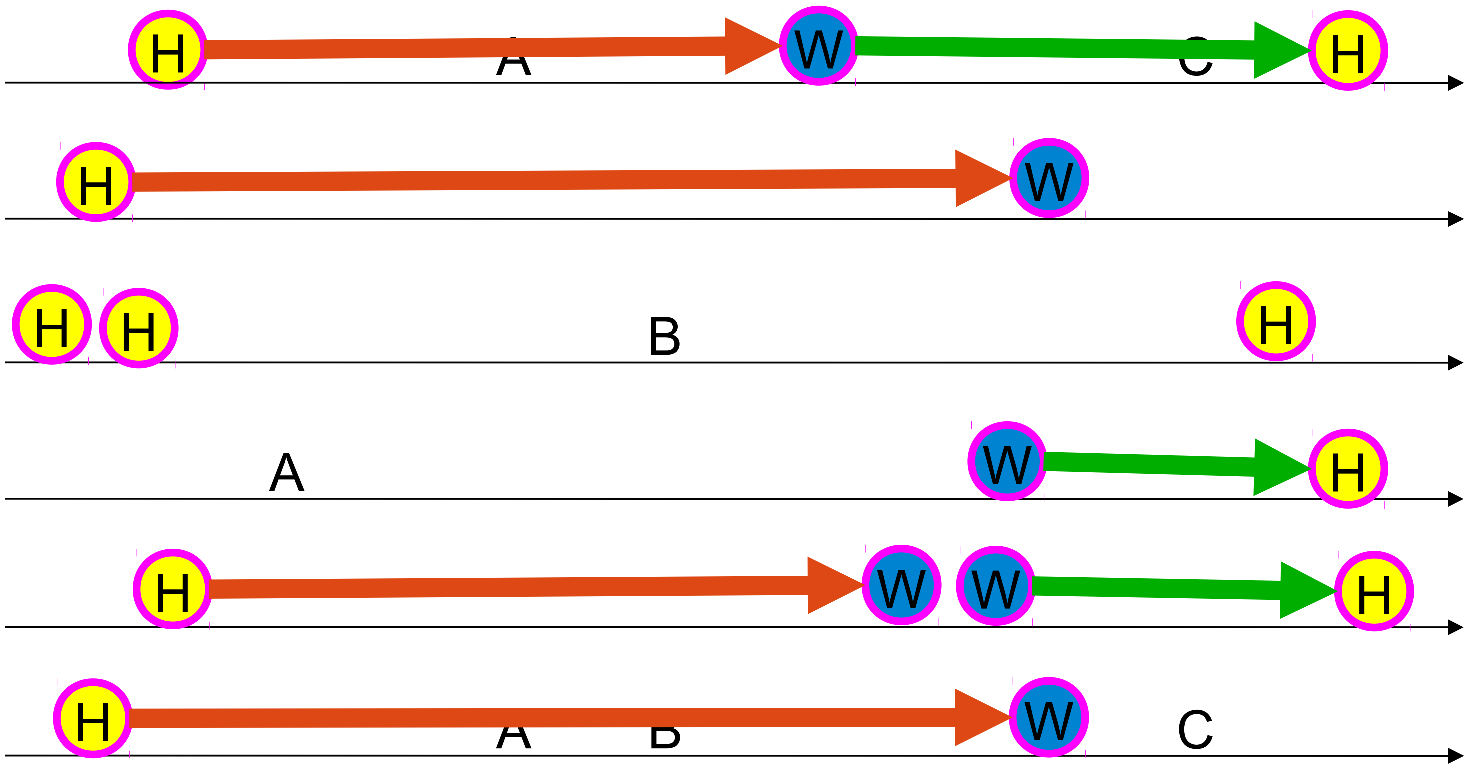
Systematic mobility

- Yet, several daily traces of the same individual might allow to identify regular places



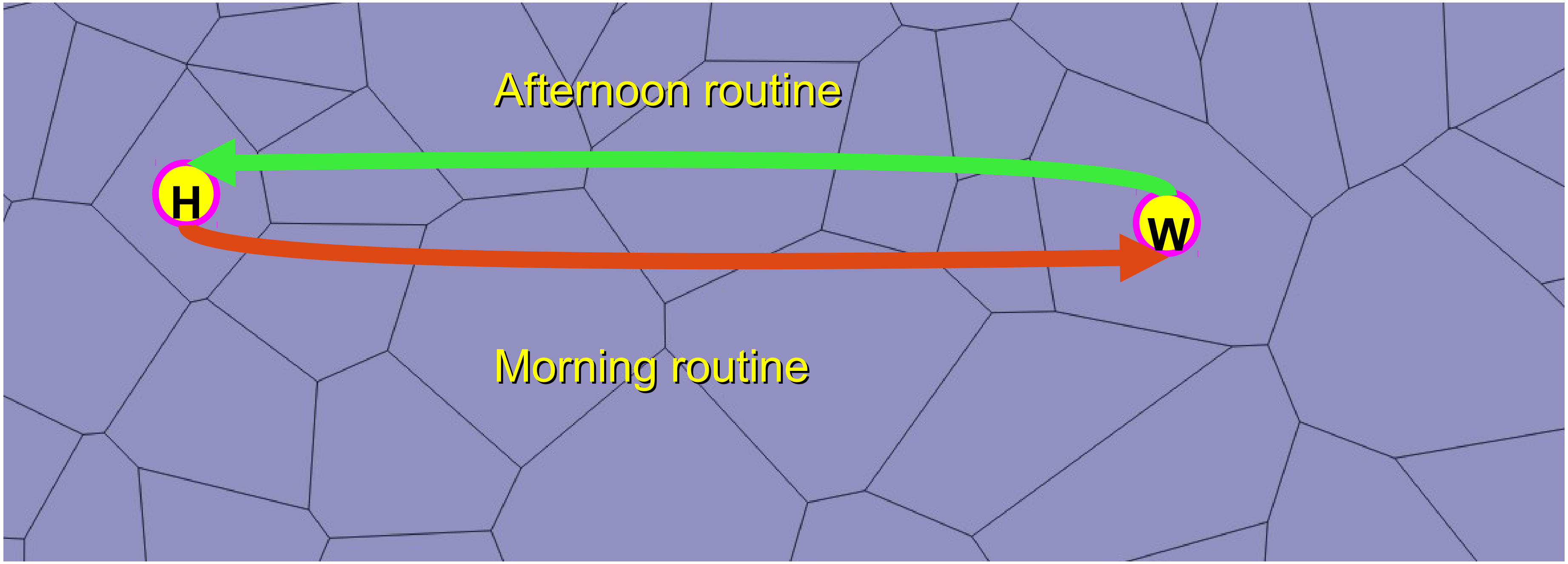
Systematic mobility

- Yet, several daily traces of the same individual might allow to identify regular places and trips



Systematic mobility

- The whole individual mobility is then summarized by its systematic movements

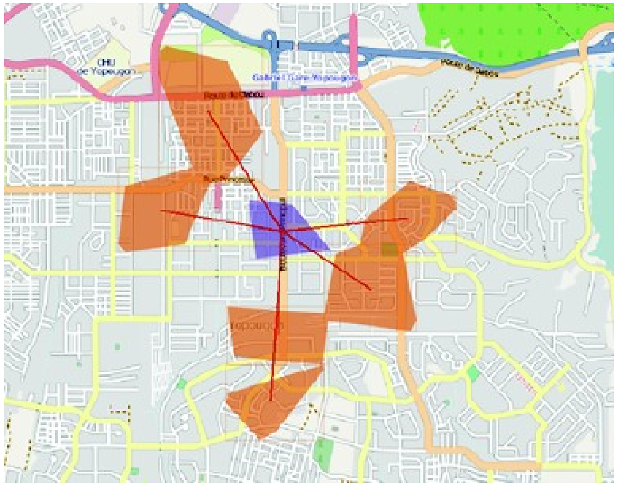


- They will be used as typical daily schedule of the individual

Systematic O/D matrix

- Combine the ten 2-weeks datasets into one
- For each user, extract significant L1 → L2
- Aggregate (individual) systematic movements into (collective) systematic flows
- Examples:

Outgoing traffic



Incoming traffic

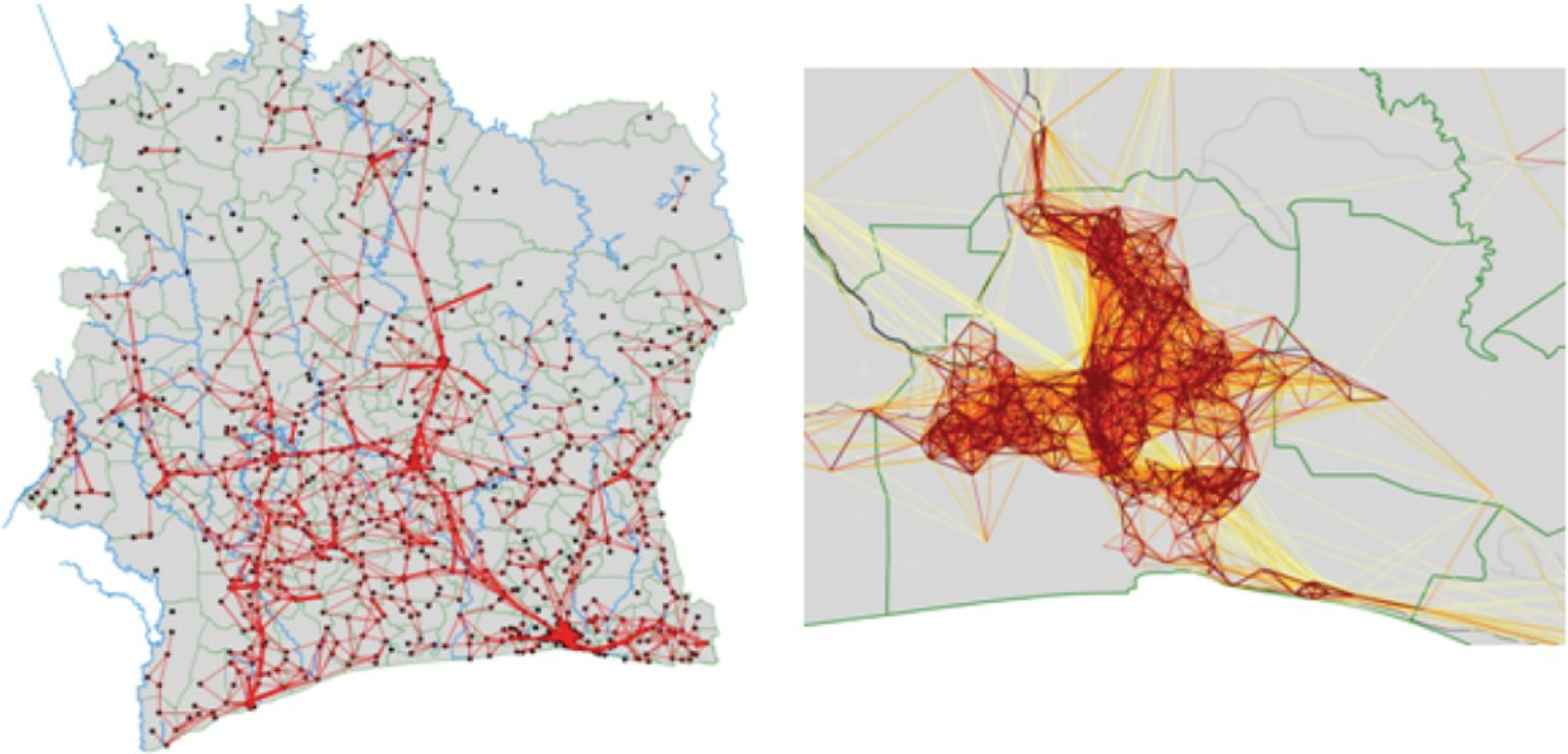
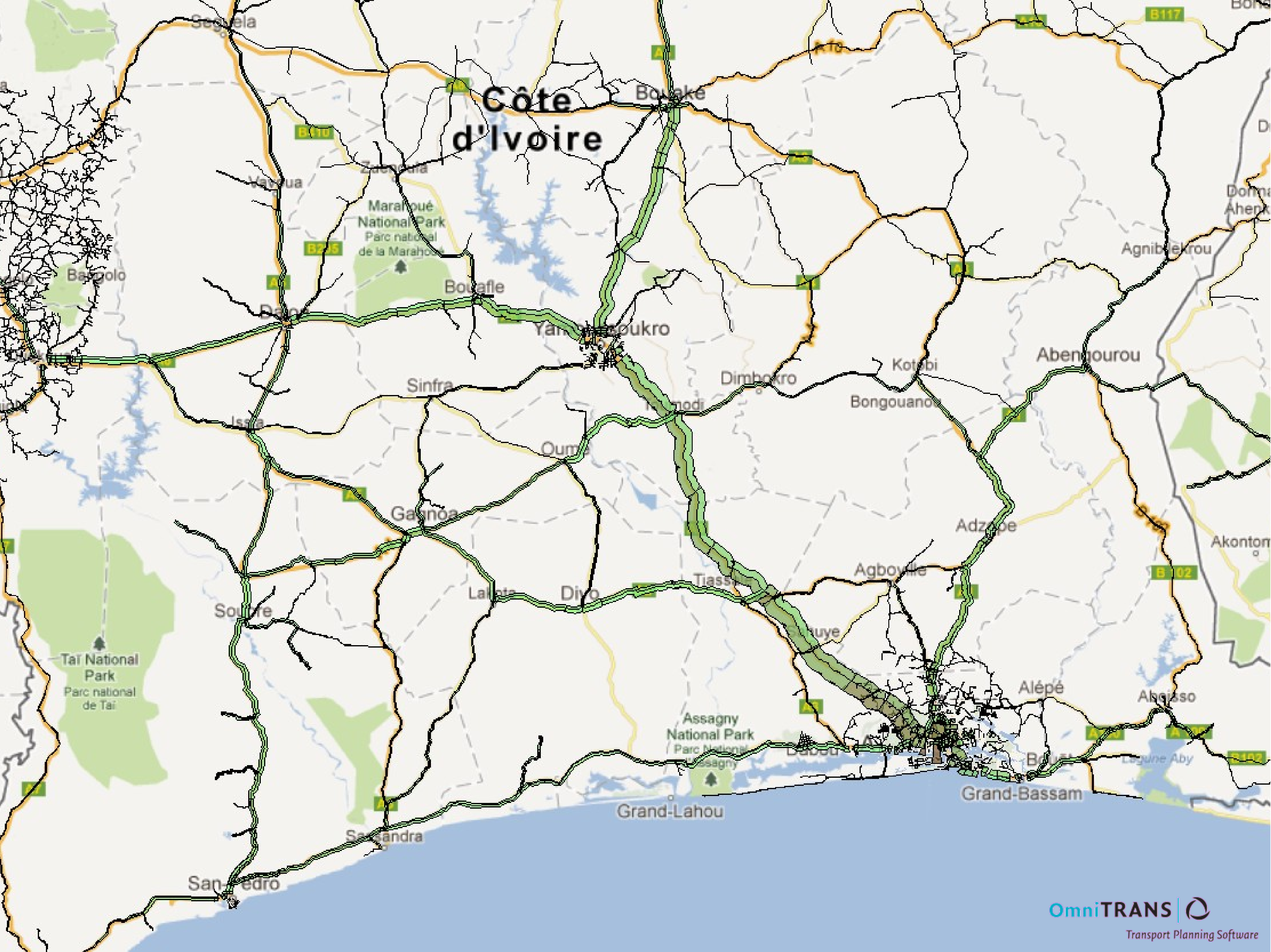


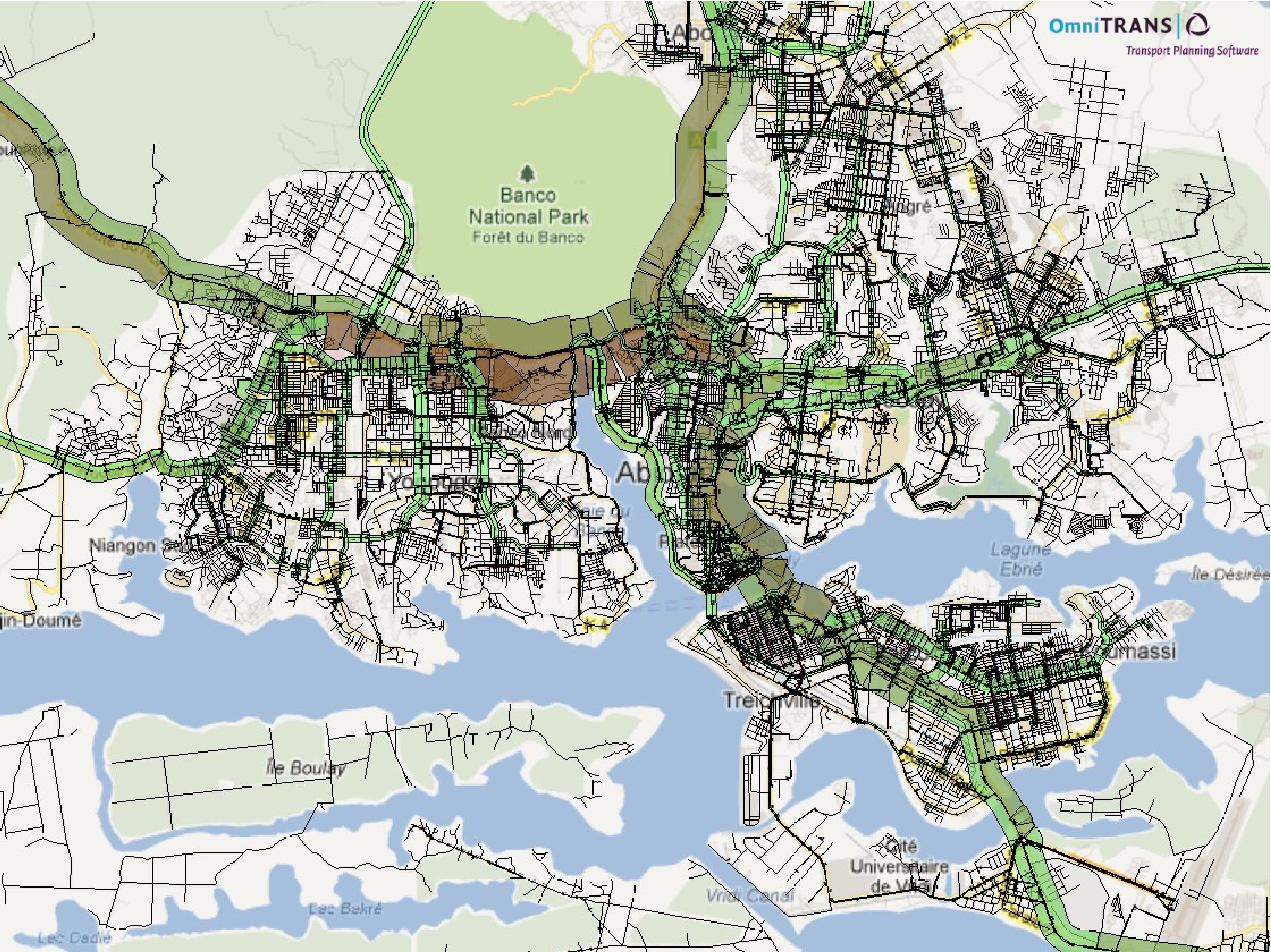
Figure 12: Mobile phone movements in Ivory Coast and Abidjan.

Mirco Nanni, Roberto Trasarti, et al.:
MP4-A Project: Mobility Planning for Africa. “Data for Development” Orange challenge, 2013

Building the transport model

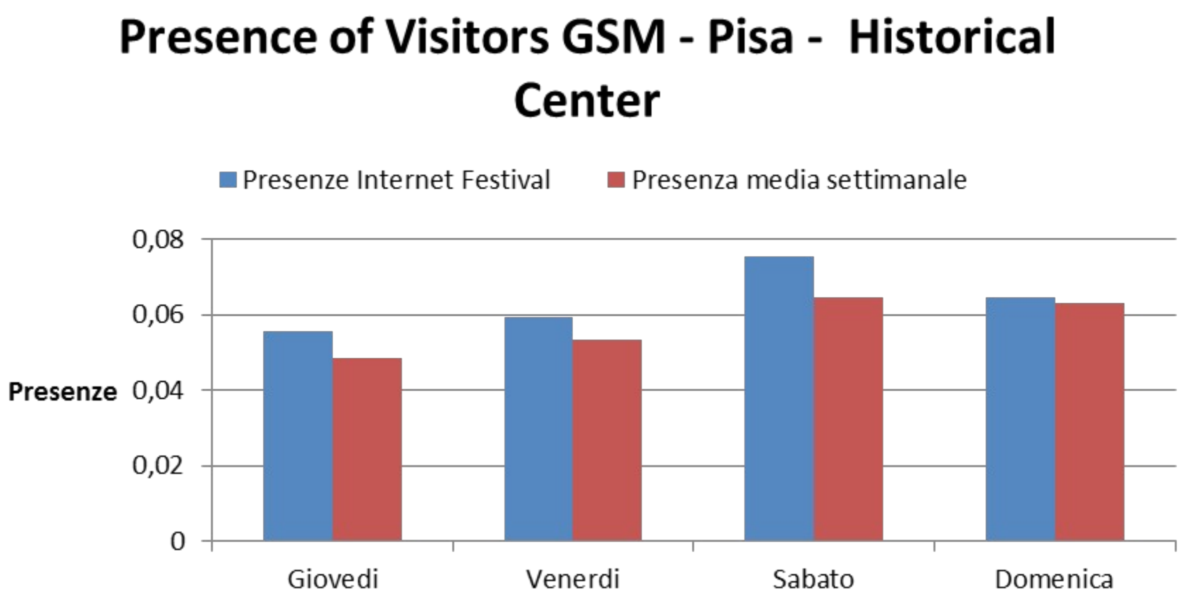
- Traffic assignment
  - Based on OmniTRANS V6 software
- Simulation assumptions
  - Assign each phone tower to the closest road
  - Use OSM information on speed limits
  - Adopt an all-or-nothing assignment

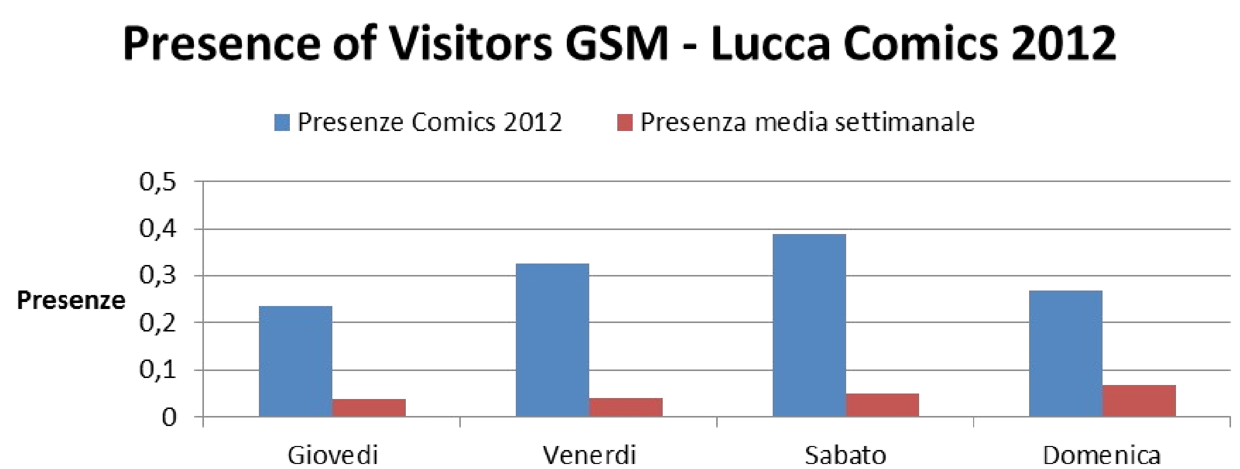

Côte d'Ivoire
Marahoué
National Park
Taï National
Park
Parc national
de Taï
Assagny
National Park
Bouafle
Sinfra
Abengourou
Kotobi
Bongouanou
Grand-Lahou
Grand-Bassam
Alépé
Divo
B117
B102
OmniTRANS
Transport Planning Software


OmniTRANS
Transport Planning Software
Banco
National Park
Forêt du Banco
Lagune
Ebrié
Île Désirée
Île Boulay
Vridi Canal
Lac Bakré
Lac Dadié

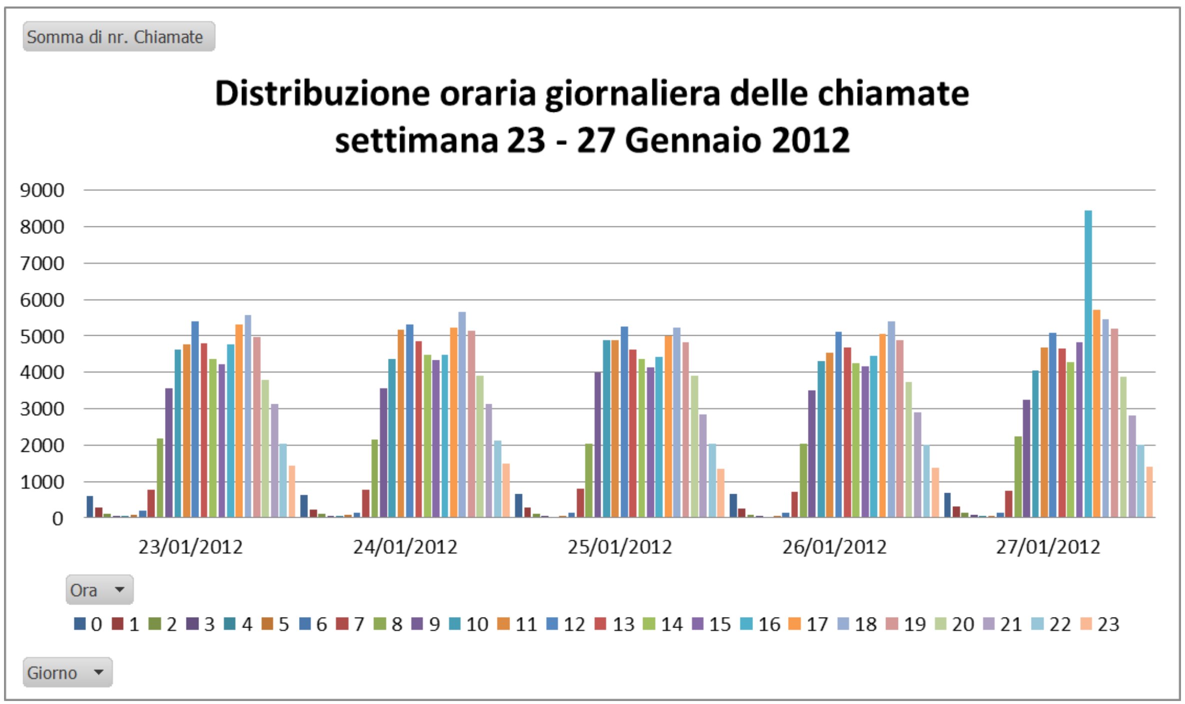
Territory

Measuring exceptional events





Extraordinary events



LA NAZIONE

Firenze / Arezzo / Empoli / Grosseto / La Spezia / Livorno / Lucca / Massa Carrara / Montecatini / Pisa / Pontedera / Pisto

HOME | SPORT | MOTORI | MAGAZINE | LIFESTYLE | SPETTACOLO | TECNOLOGIA | BLOG | MULTIME

Home Toscana | Cinema

HOMEPAGE > Toscana > Terremoto: trema tutta la Toscana Massa, crolla il tetto in chiesa Miracolosamente il in tilt

Terremoto: trema tutta la Toscana Massa, crolla il tetto in chiesa Miracolosamente illesi due fedeli

Evacuato l'Ateneo di Pisa, telefoni in tilt **Terremoto a Massa: buco nel tetto in chiesa** Commenti

L'epicentro del sisma di **magnitudo 5.4** (scala Richter) sull'appennino tosco-emiliano. La scossa è stata avvertita soprattutto in Lunigiana e sulla costa nord della regione

Mi piace 72 | Tweet 1 | +1 1 | Email | Stampa

Terremoto con epicentro nel parmense, scosse avvertite in tutta la Toscana

ARTICOLI CORRELATI

Il terremoto a Empoli

Hai sentito il terremoto? Il questionario INGV

L'allarme dei geologi: "Il 60% delle case è a rischio sismico"

Toscana, 27 gennaio 2012 - Paura in Toscana per la **scossa di terremoto di magnitudo 5.4 della scala Richter**, avvertita in tutto il Nord Italia intorno alle 15 e 53, con epicentro nella zona dell'Appennino tosco-emiliano. Disagi soprattutto

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

orange

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

Riva del Garda, August 30th, 2013

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

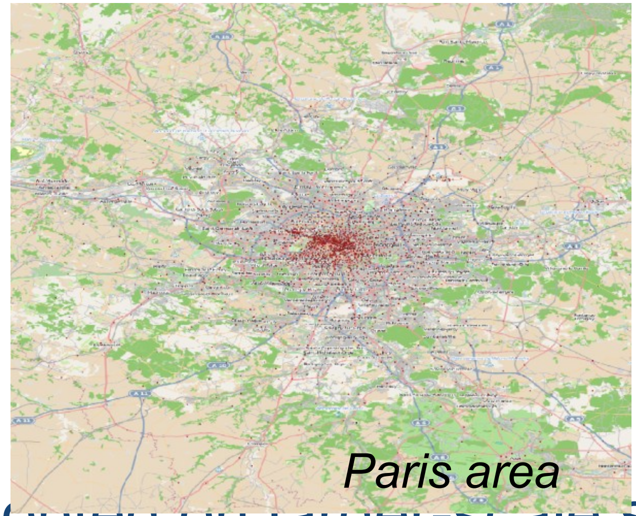
2. Search frequent combinations of **events** across different regions

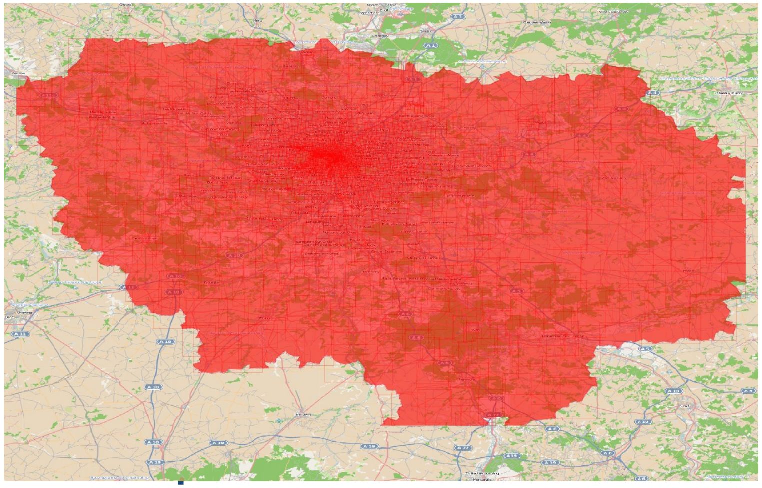
Step 1: estimate density of population

Use Call Detail Records to measure population

- Alternative: heuristics to identify stops

Each GSM tower associated to estimated coverage



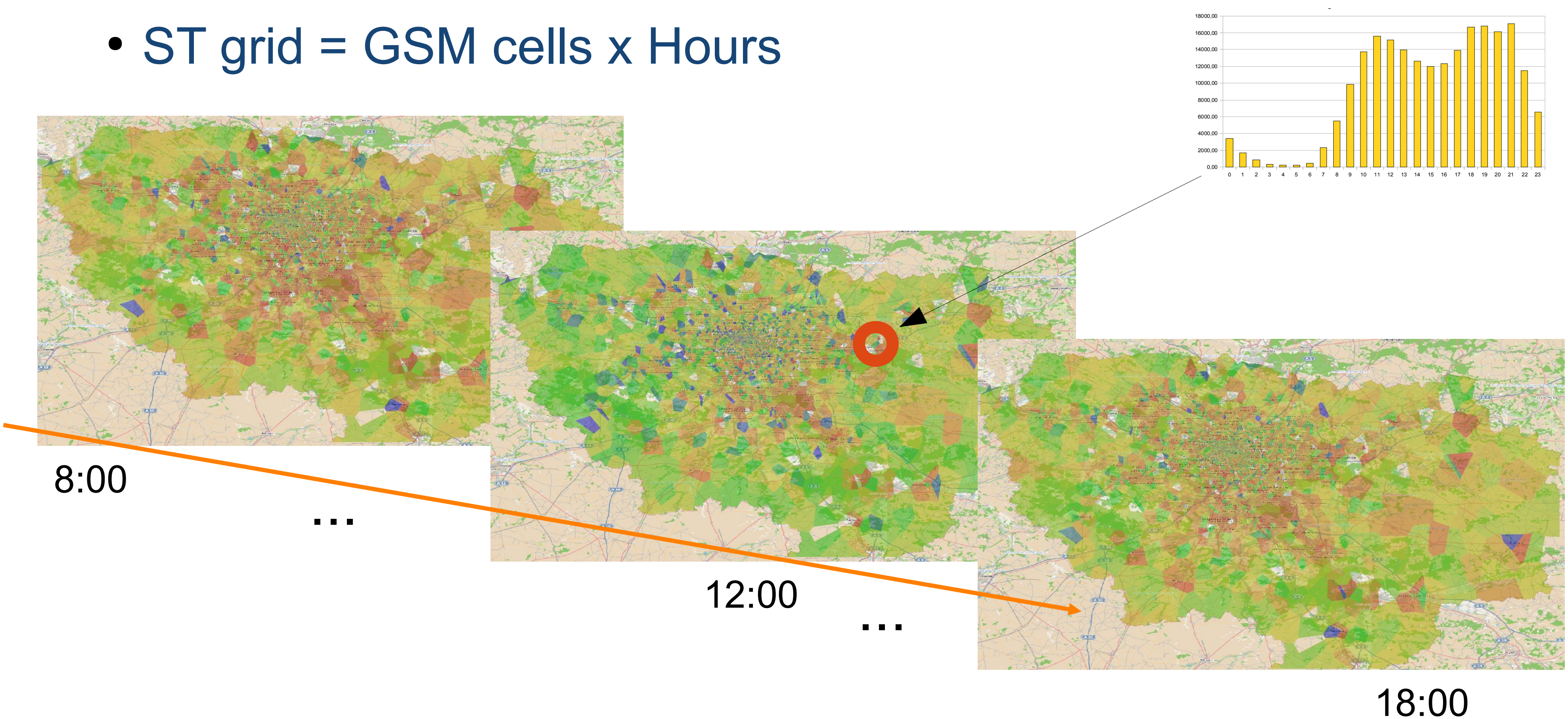


Aggregations adopted on larger-scale scenarios

Step 2: compute density over a space-time grid

Divide the dataset into days, and days into 24h

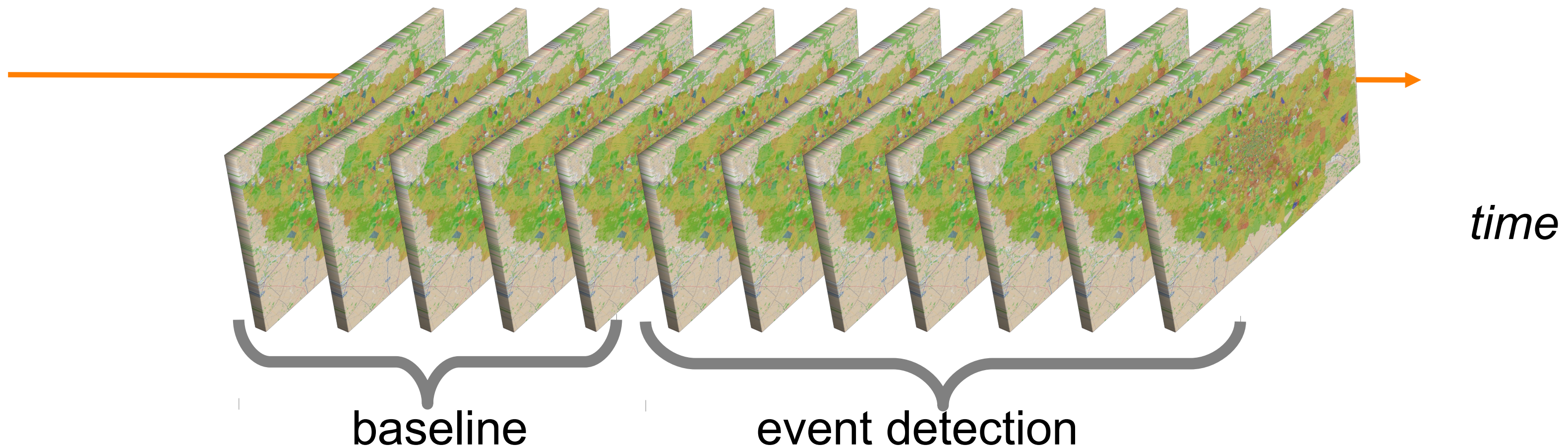
- ST grid = GSM cells x Hours



Step 3: detect events / 1

Split the dataset into temporal segments

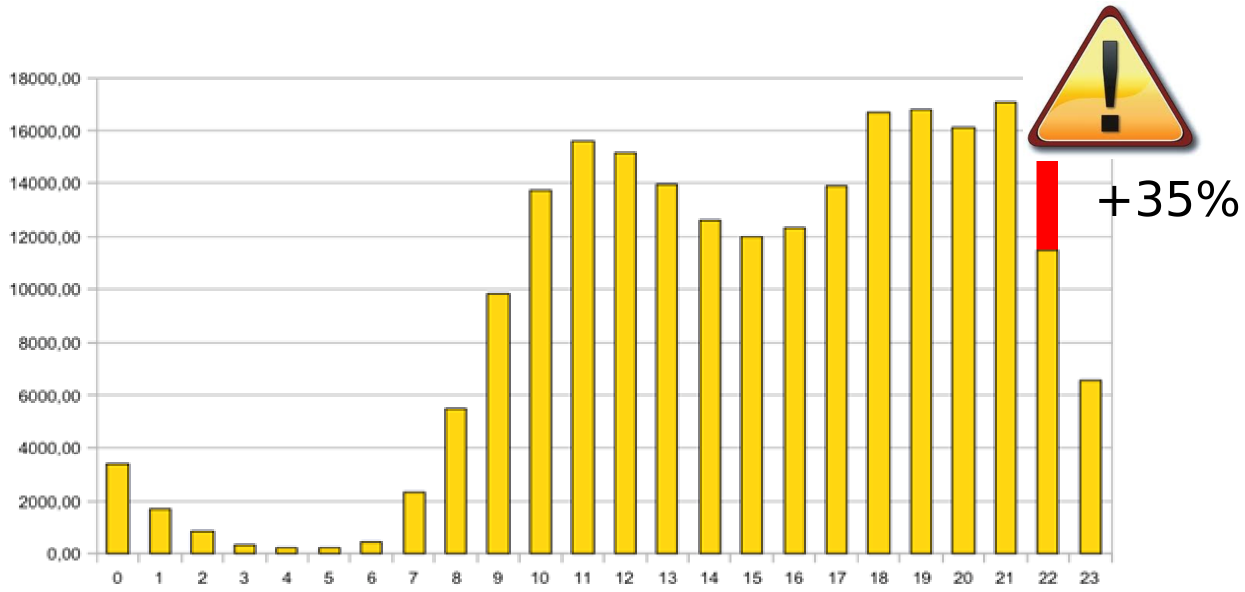
- **Baseline** segment: compute average density values for each hour of each day of the week
- **Event detection** segment: compare values against baseline to detect events



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

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

...

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

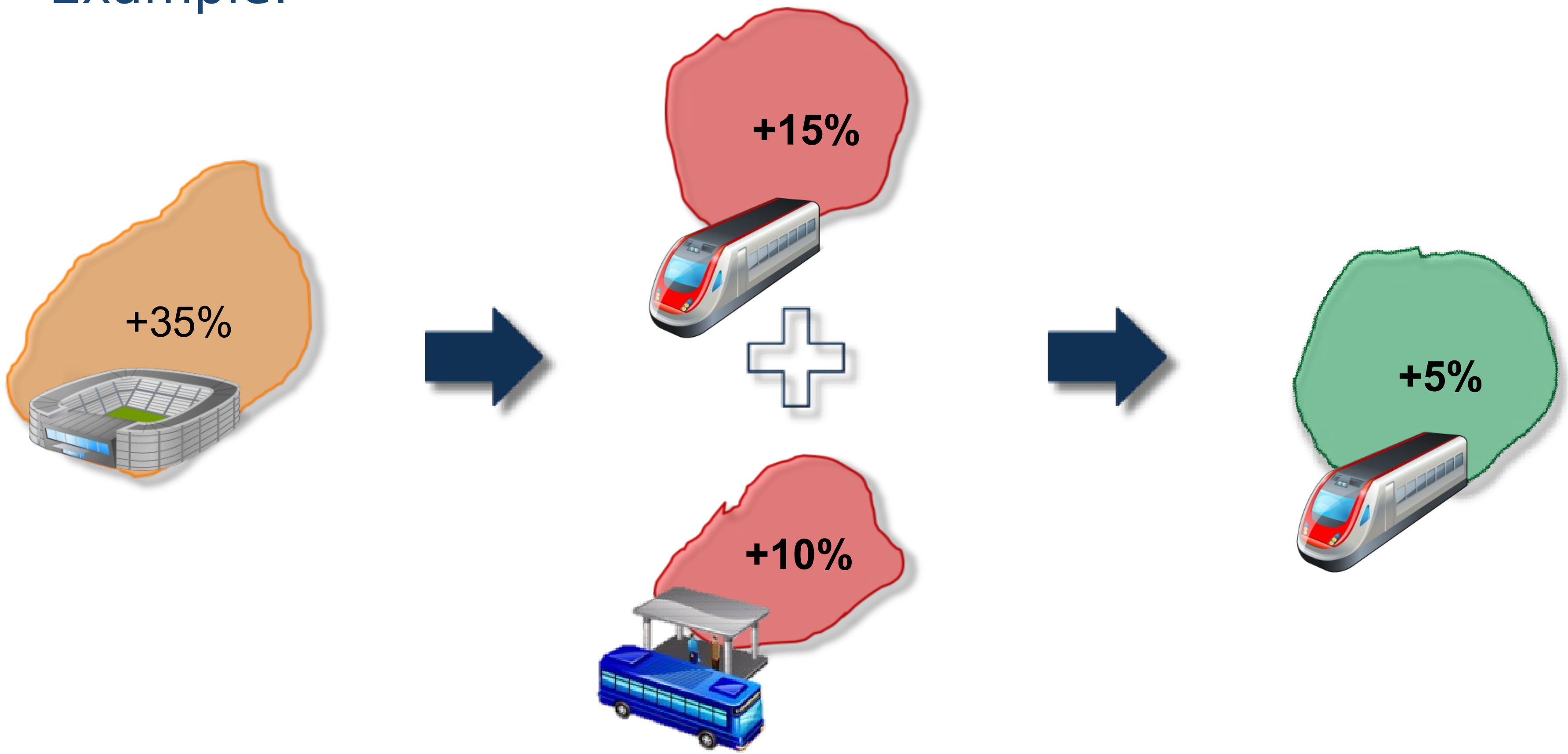
Step 4: correlation patterns/1

- 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

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

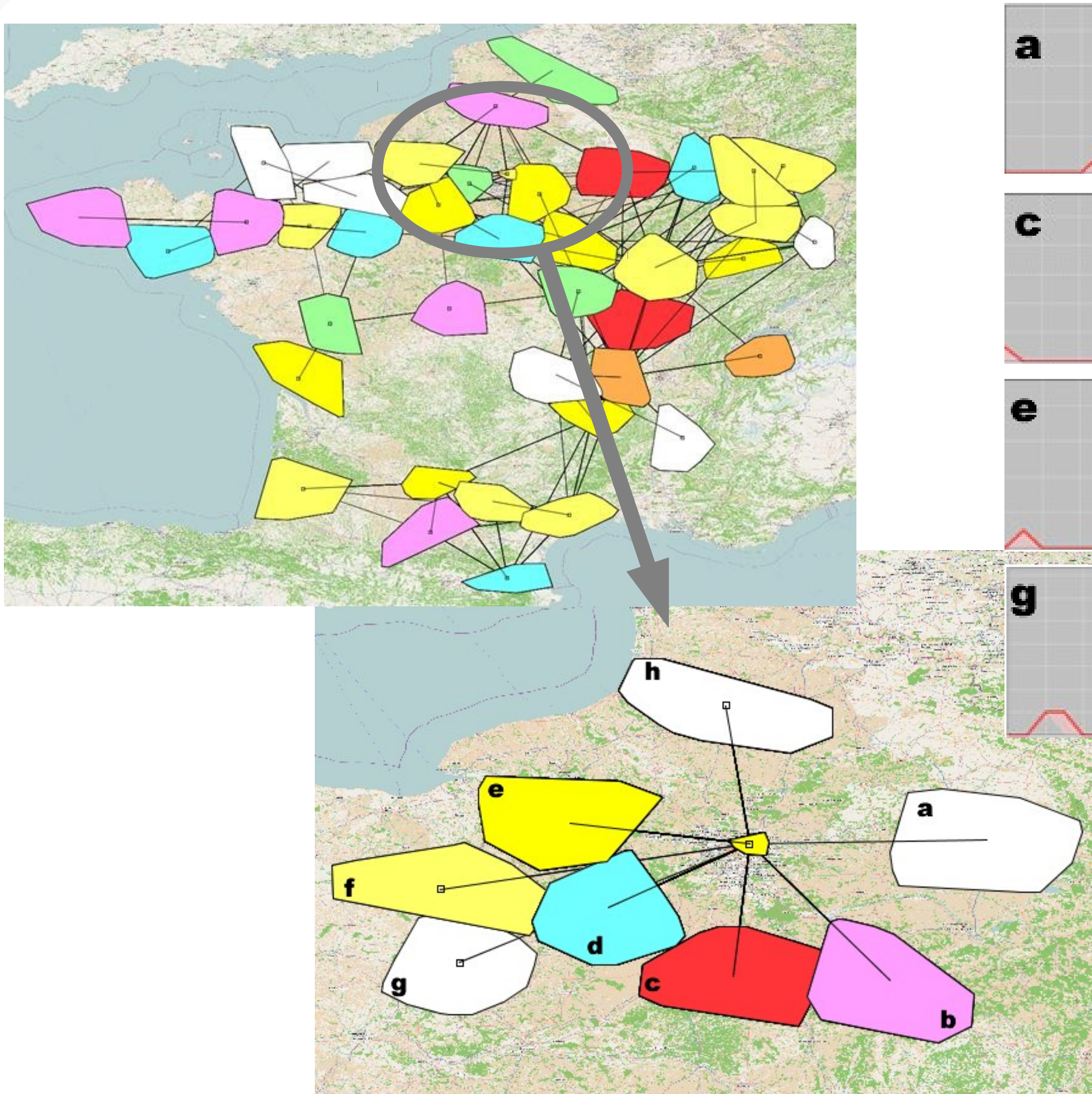
Step 4: correlation patterns/2

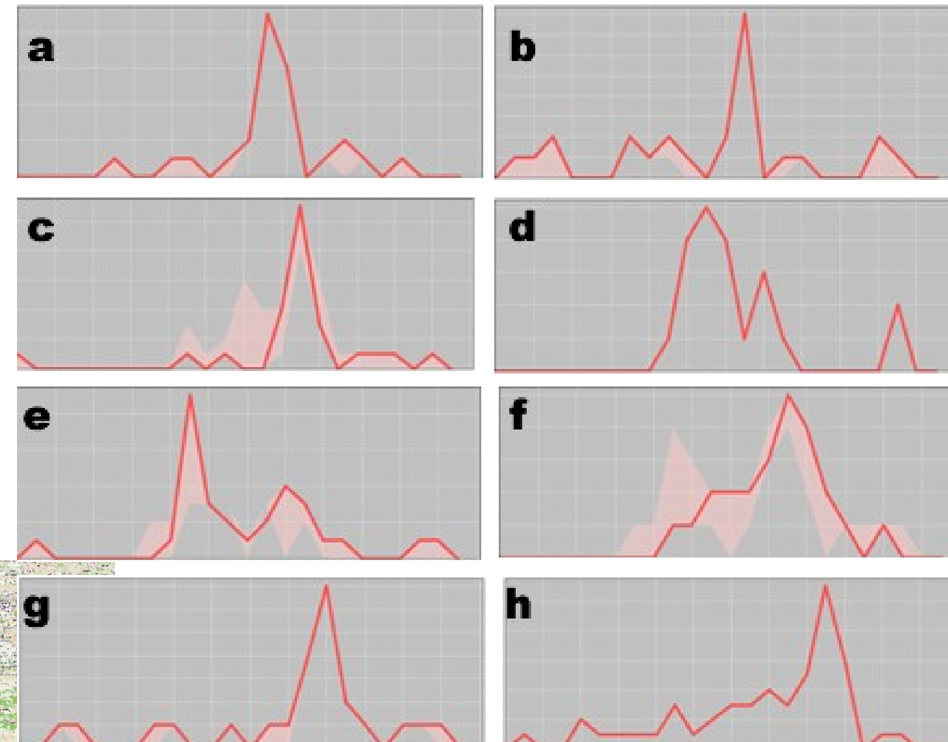
Example:



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

National level example (departments)





Focus on Seine-Saint-Denis