DM& ML …..for smart cities

Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

Master MAINS 2018

Suggested Bibliography

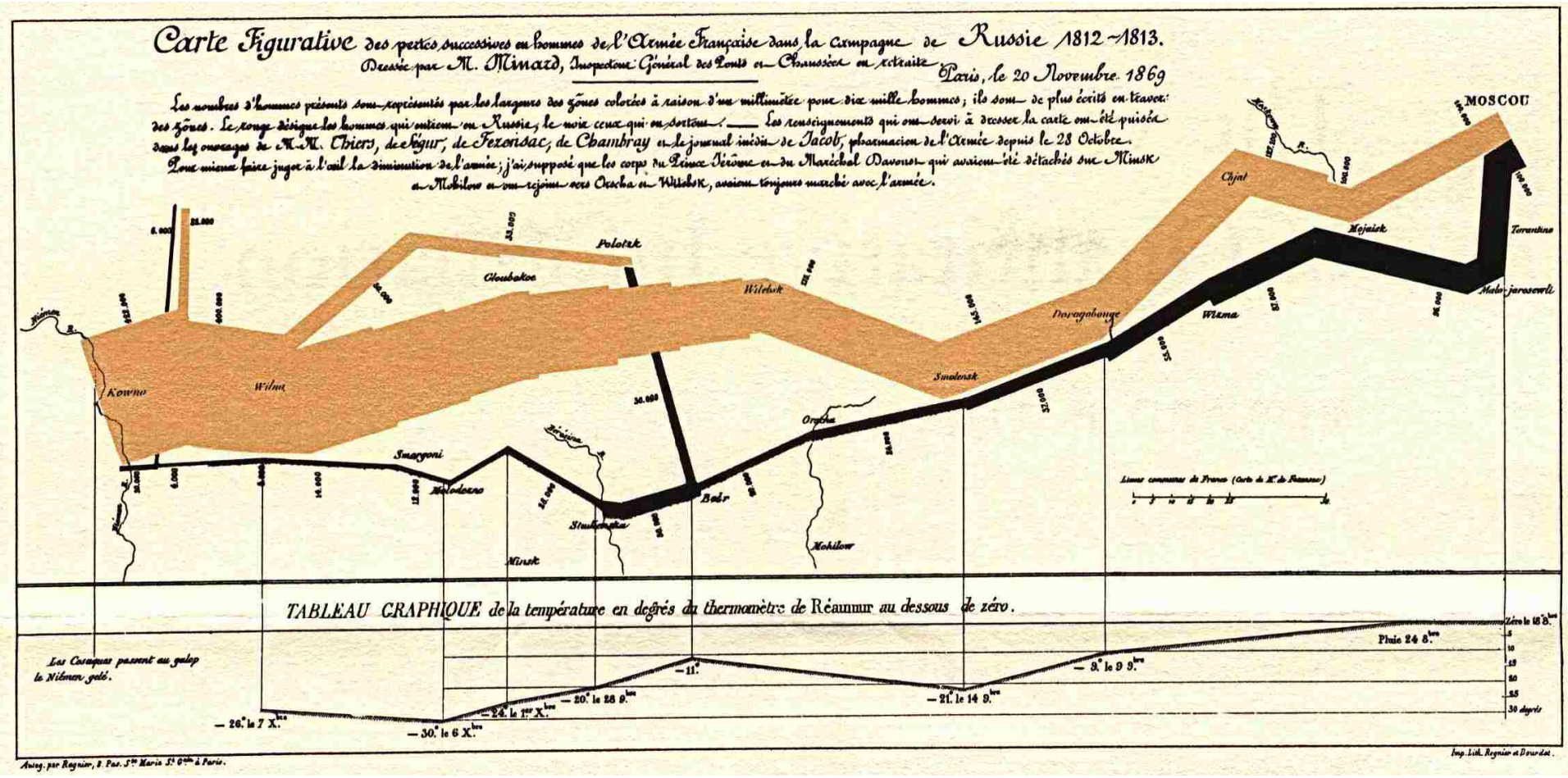
- Mobility Data: Modeling, Management and Understanding. Ed. Edited by Chiara Renso, Stefano Spaccapietra, Esteban Zimanyi. Cambridge Press. chapter 10 **Car Traffic Monitoring. Nanni, Rinzivillo**
- M Batty, KW Axhausen, F Giannotti, A Pozdnoukhov, A Bazzani, M Wachowicz. **Smart cities of the future**. The European Physical Journal Special Topics 214 (1), 481-518, 2012
- F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. **Unveiling the complexity of human mobility by querying and mining massive trajectory data**. VLDB J. , 2011
- Miao Lin, Wen-Jing Hsu, **Mining GPS data for Mobility Pattern: A Survey**. Pervasive Computing 2013

Summary

- Big mobility data & Methods in a nutshell
- Mobility data mining methods
  - Trajectory Clustering
  - Trajectory pattern mining
  - Trajectory classification / prediction
- Understanding human mobility with GPS
  - Trajectory reconstruction
  - Sensing the movemnt: exploring OD – Matrix
  - Discover collective and individual patterns (Sistematic vs Non-sistematic behavior)
  - Building territory Indicators (Urban Mobility Atlas)
  - Activity Recognition
  - Building services towards corporate(Geomarketing, Driving profile)
  - Building services towards citizens (adaptive car pooling)
  - Data validation
- Understanding city (territory) dynamics with GSM
  - Exploring the possible dimensions:
    - call behavior, presence in space, mobility
  - Classifying city users from call behavior
  - Measuring Economic Development
- Models of Human Mobility
    - From Levy Flight to preferential return
    - Explorers & returners
- Sport Analytics

BIG DATA AS A PROXY OF HUMAN MOBILITY

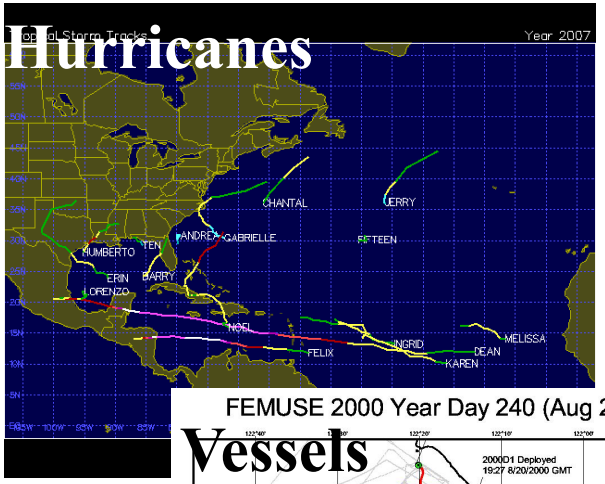
Understanding Human Mobility: a long path

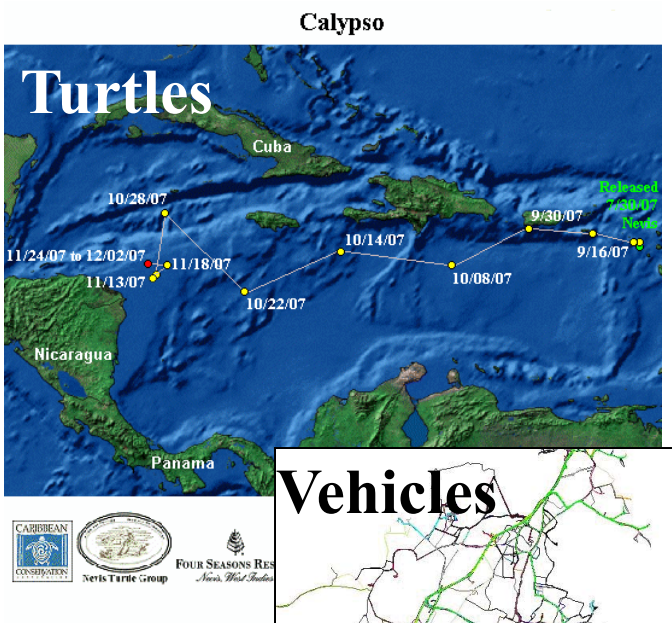


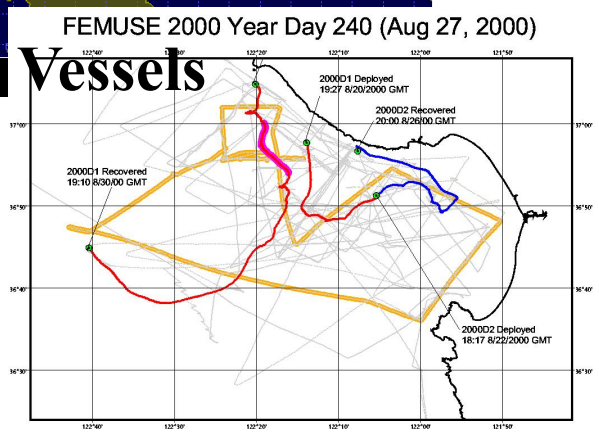
Charles Minard. "Carte figurative des pertes successives en hommes de l'Armée Francaise dans la campagne de Russie 1812-1813". 1869.

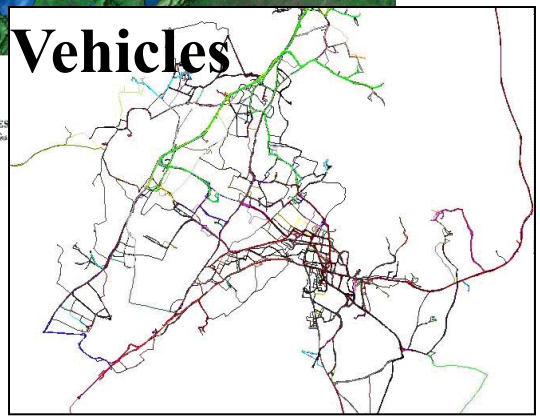
Moving Object Data

- Several domains:

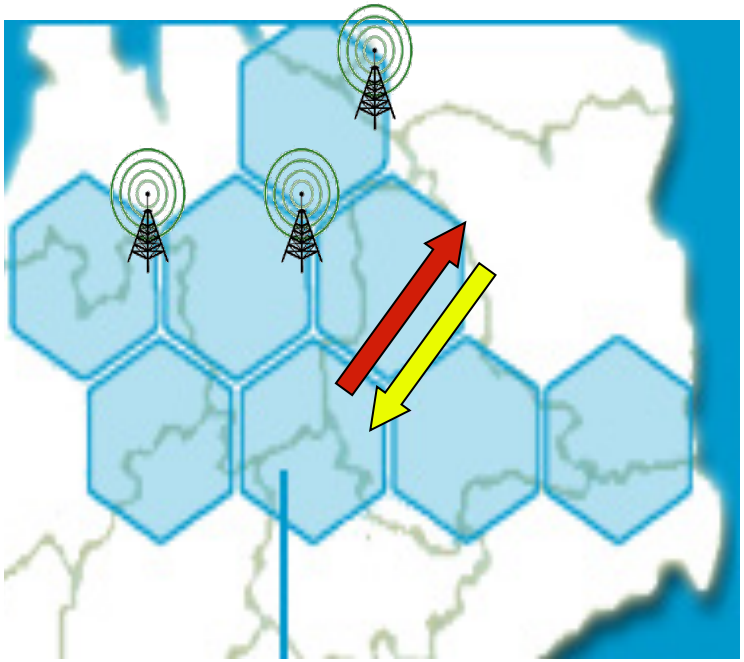


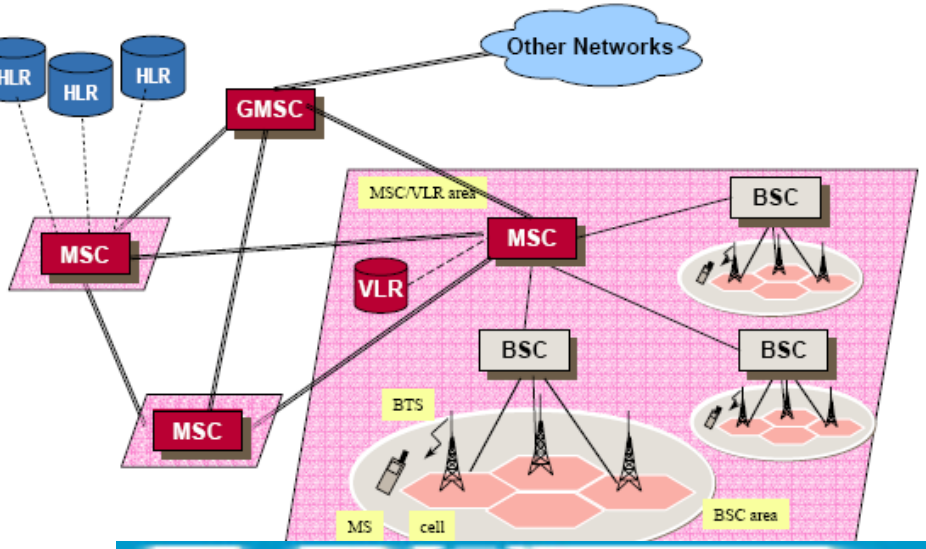


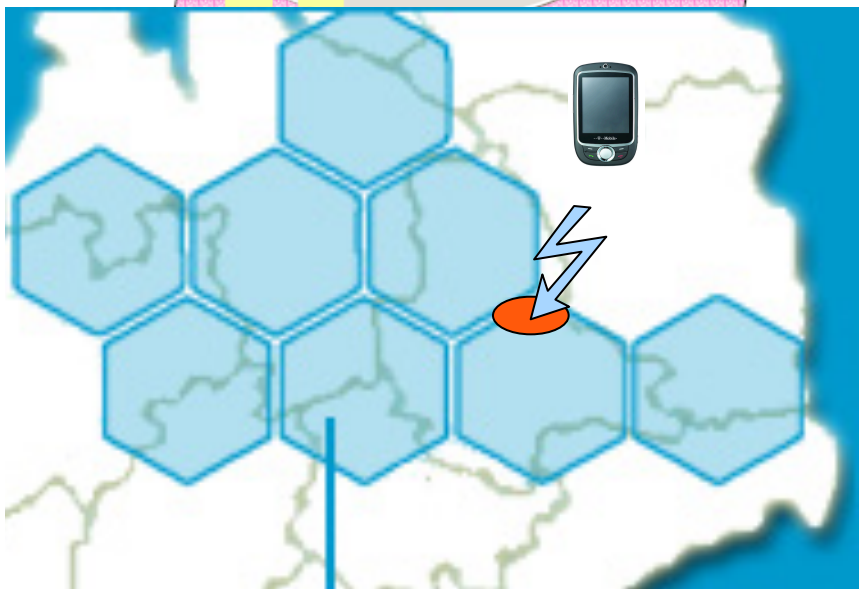




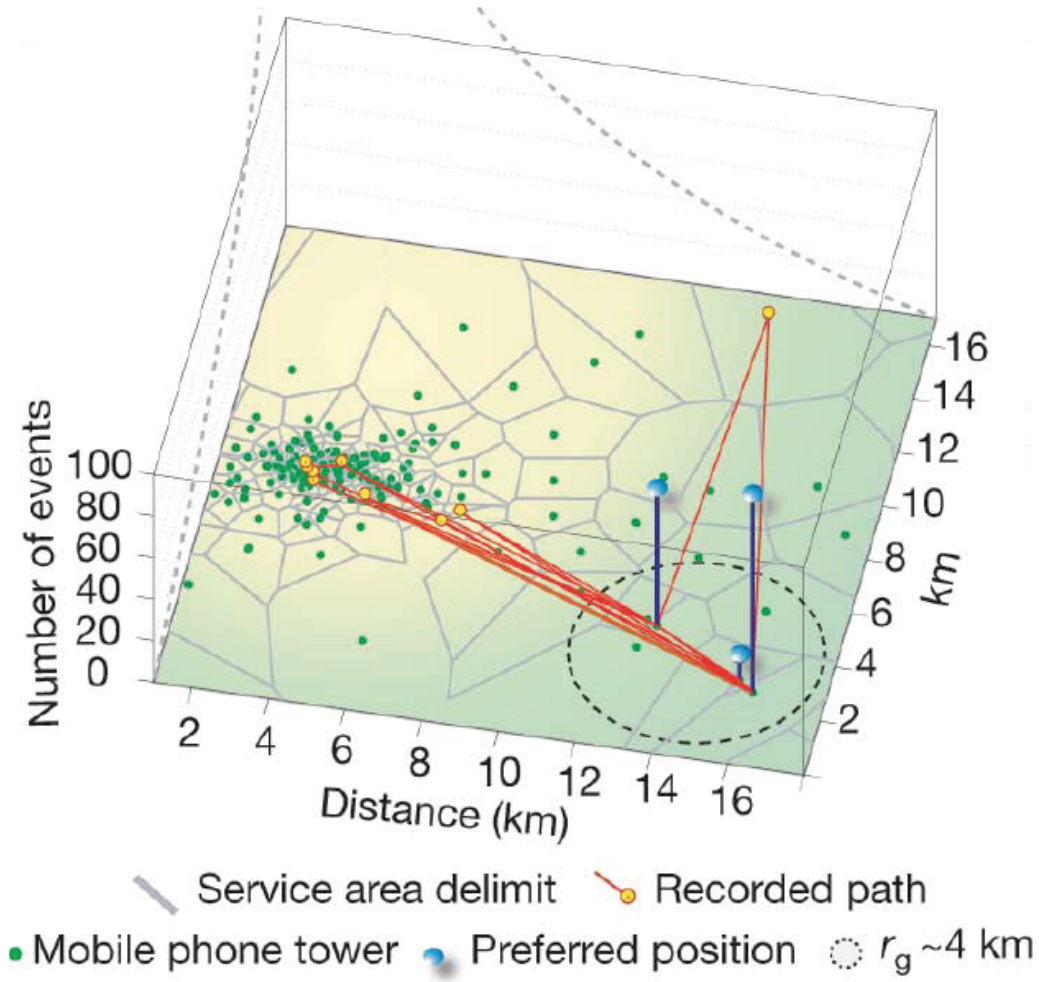
GSM roaming CDR data –







Country-wide mobile phone data



when you call

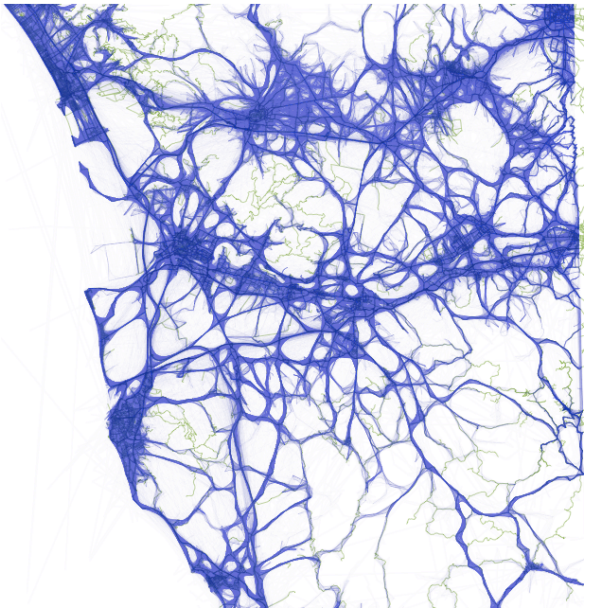
where you call

who you call

GPS tracks

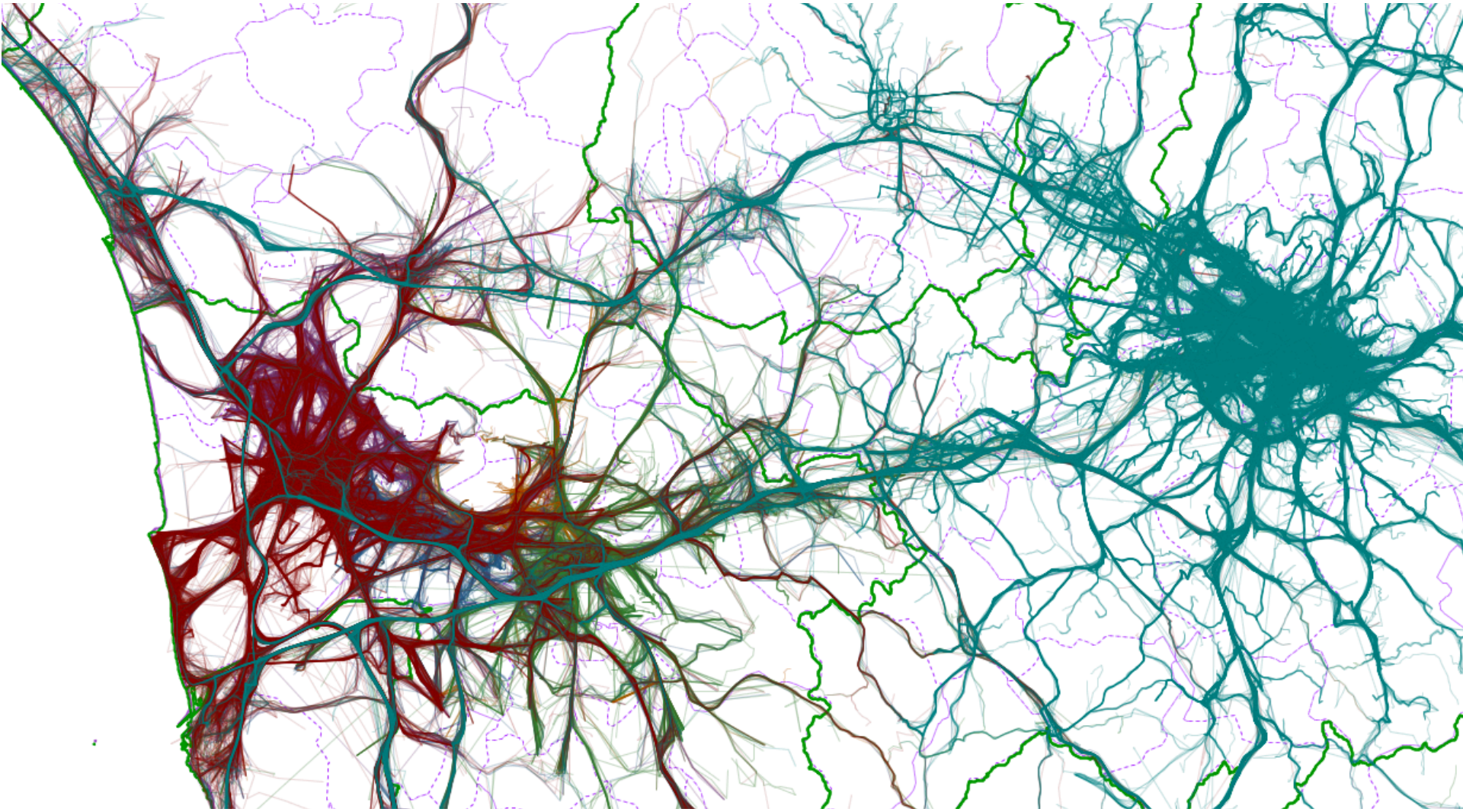
Onboard navigation devices send GPS tracks to central servers

```
Ide;Time;Lat;Lon;Height;Course;Speed;PDOP;State;NSat
…
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
8;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
8;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
8;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
8;22/03/07 08:52:06;50.777185;7.206202; 68.1;124.1;30.058;3.8;1808;4
8;22/03/07 08:52:09;50.777057;7.206522; 67.9;117.7;34.003;3.8;1808;4
8;22/03/07 08:52:12;50.776925;7.206858; 66.9;117.5;37.151;3.8;1808;4
8;22/03/07 08:52:15;50.776813;7.207263; 67.0;99.2;39.188;3.8;1808;4
8;22/03/07 08:52:18;50.776780;7.207745; 68.8;90.6;41.170;3.8;1808;4
8;22/03/07 08:52:21;50.776803;7.208262; 71.1;82.0;35.058;3.8;1808;4
8;22/03/07 08:52:24;50.776832;7.208682; 68.6;117.1;11.371;3.8;1808;4
…
```

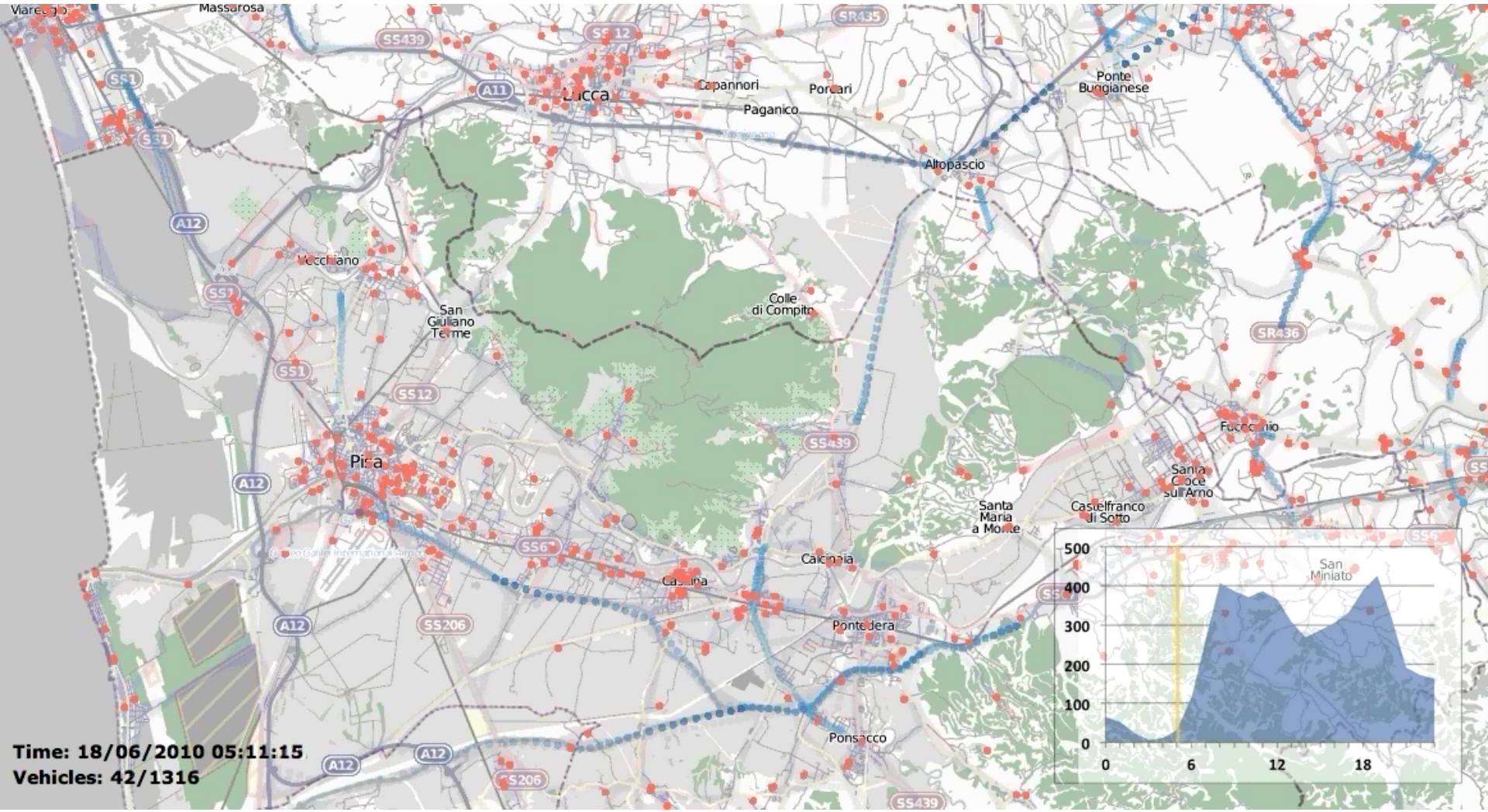


Sampling rate ~3 secs

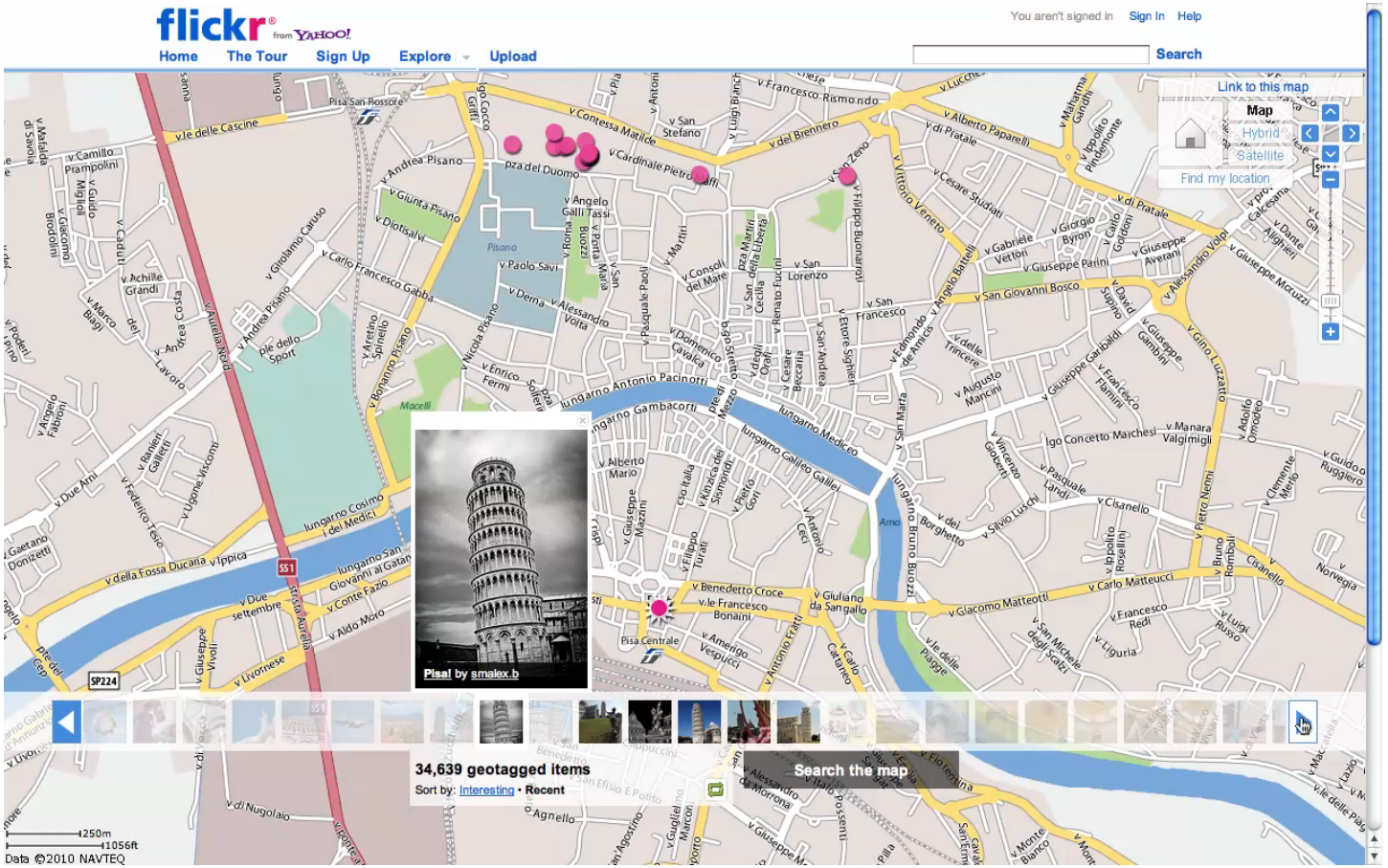
Spatial precision ~ 10 m



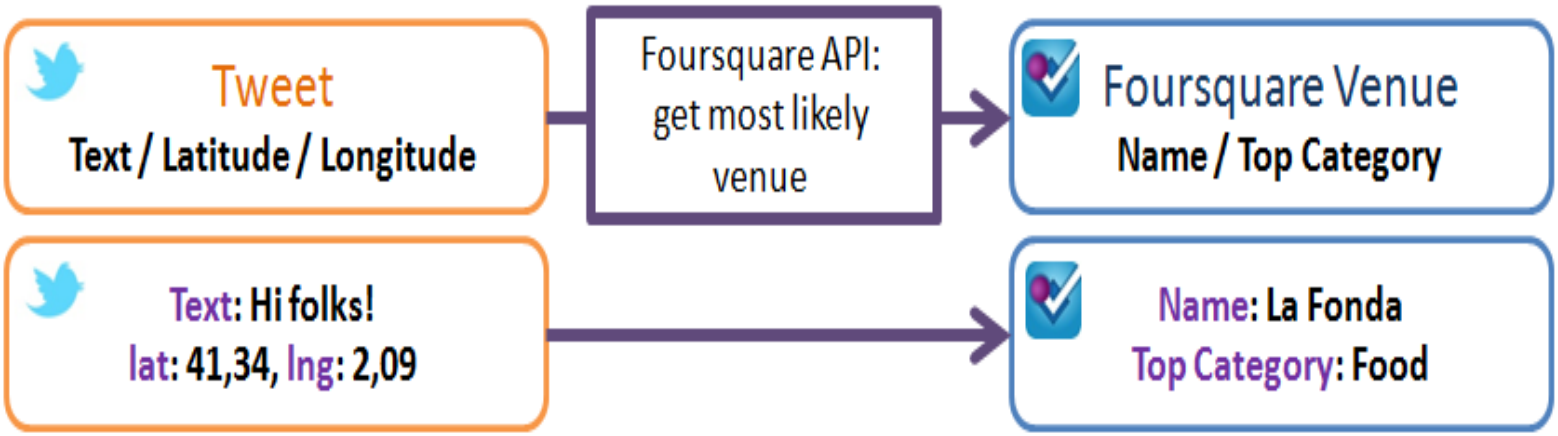
GPS: detailed movements within an



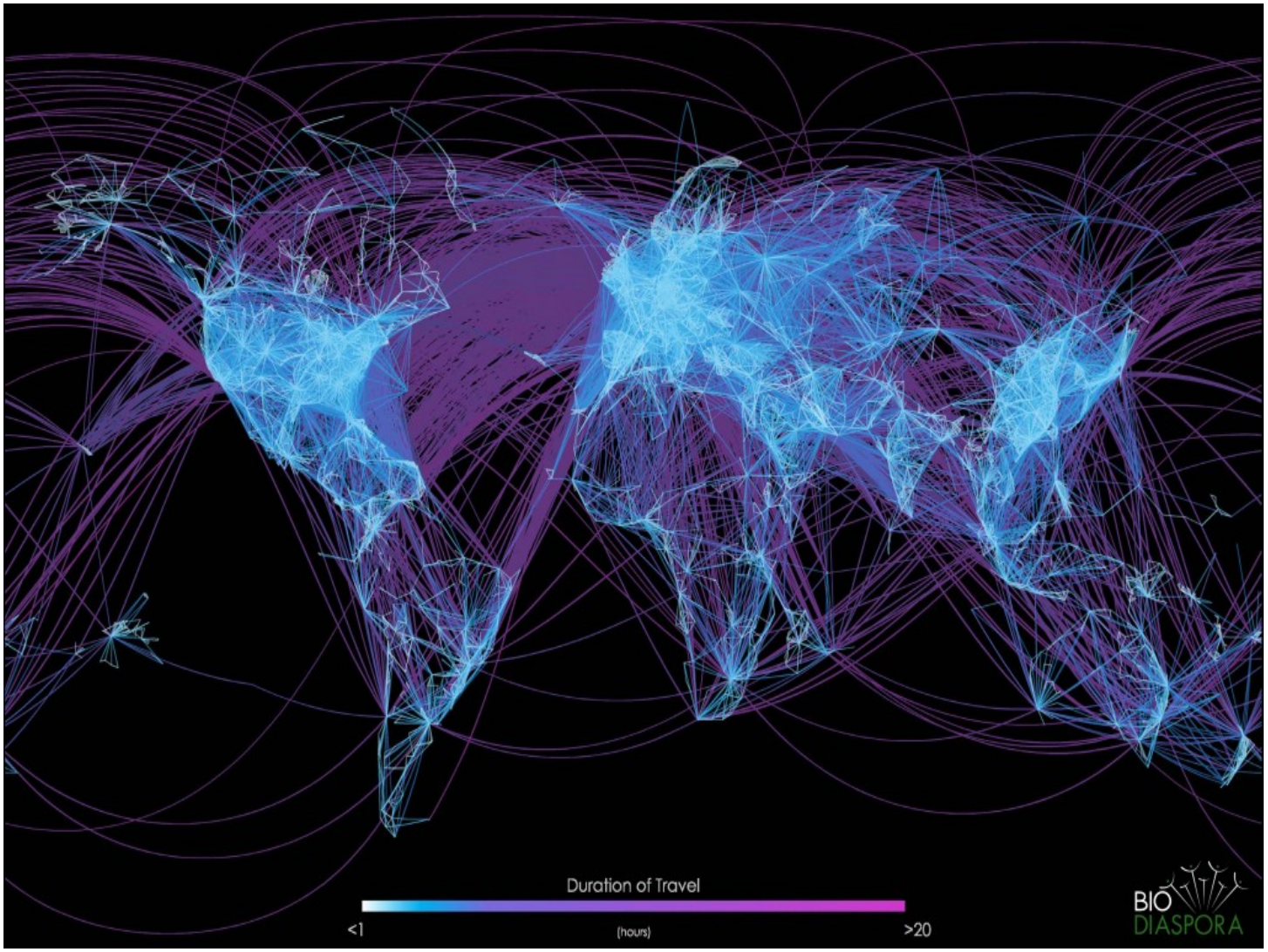
Social Networks: goal of the movement



Twitter


Tweet
Text / Latitude / Longitude
Foursquare API: get most likely venue
Foursquare Venue
Name / Top Category
Text: Hi folks!
lat: 41,34, lng: 2,09
Name: La Fonda
Top Category: Food

Airlines Flights



Quality of Mobility Data

	data type		
	mobile phone data	wireless network traces	GPS data
resolution	km	m	m
scale	metropolitan area	campus/work place	global
velocity	no	no	yes
pause time	approximate	approximate	exact
ict	exact	labels needed	labels needed
path info	rough	rough	exact

Mining GPS Data for Mobility Patterns: A Survey

Miao Lin and Wen-Jing Hsu
Nanyang Technological University

Complexity of Mobility data

- Uncertainty
  - Sampling rate could be inconstant: From every few seconds transmitting a signal to every few days transmitting one
  - Data can be sparse: A recorded location every 3 days
- Noise
  - Erroneous points (e.g., a point in the ocean)
- Background
  - Cars follow underlying road network
  - Animals movements relate to mountains, lakes, ...
- Movement interactions: Affected by nearby moving objects

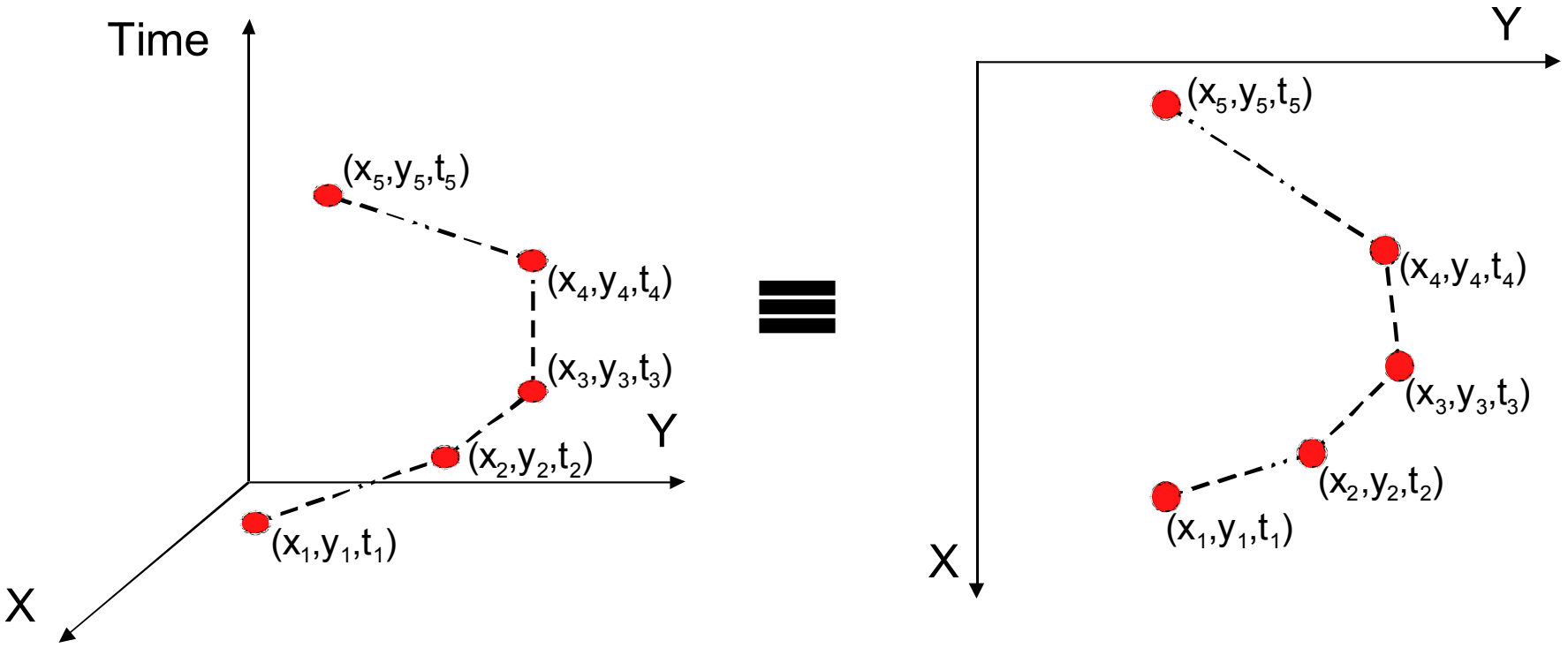
Application Domains

- Moving object and trajectory data mining has many important, real-world applications driven by the real need
  - Ecological analysis (e.g., animal scientists)
  - Weather forecast
  - Traffic control
  - Location-based services
  - Homeland security (*e.g.*, border monitoring)
  - Law enforcement (*e.g.*, video surveillance)
  - …

MOBILITY DATA MINING METHODS IN SHORT

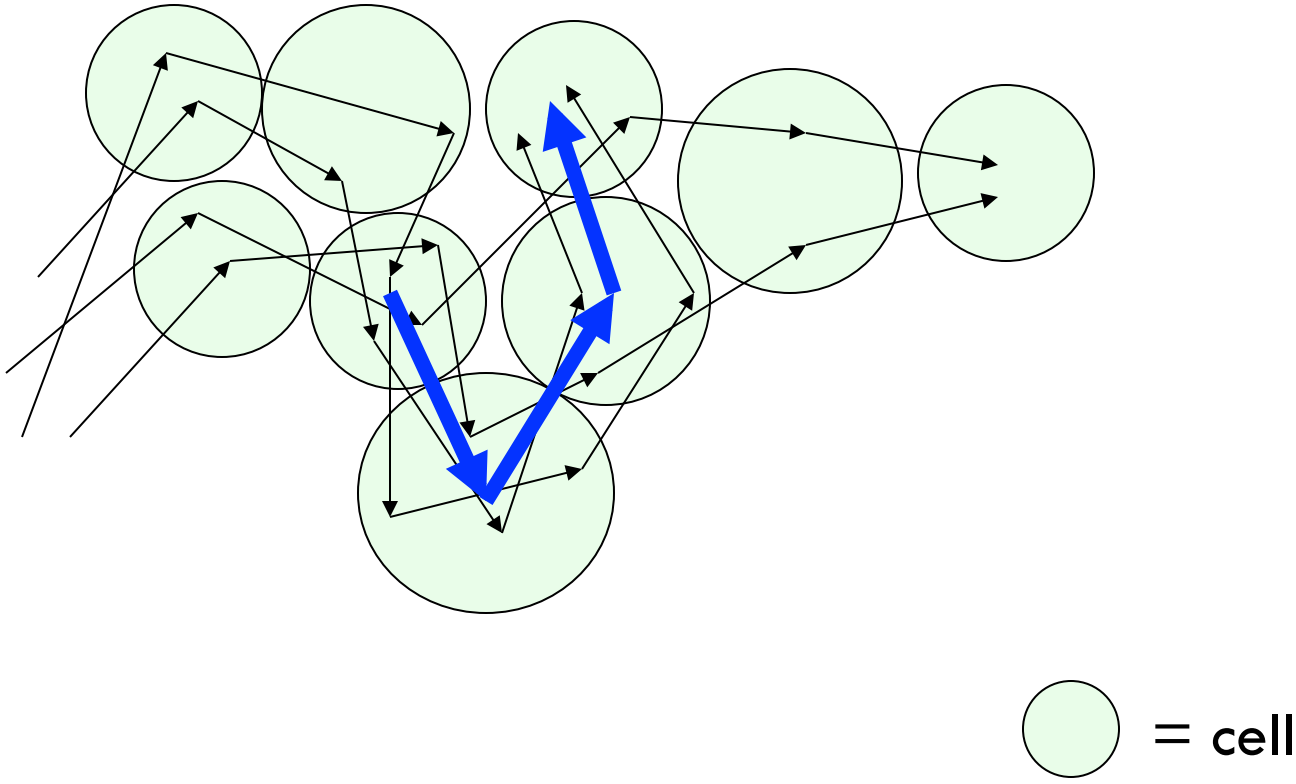
Trajectory data

- Mobility of an object is described by a set of trips
- Each trip is a trajectory, i.e. a sequence of time-stamped locations



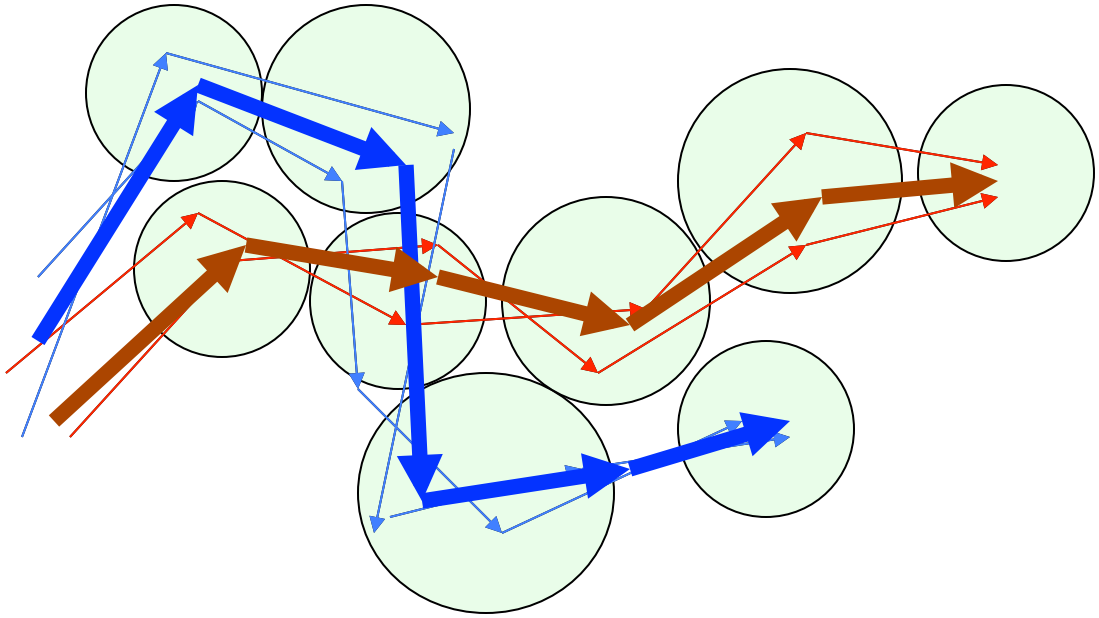
Trajectory patterns

- Discover frequently followed itineraries



Trajectory Clustering

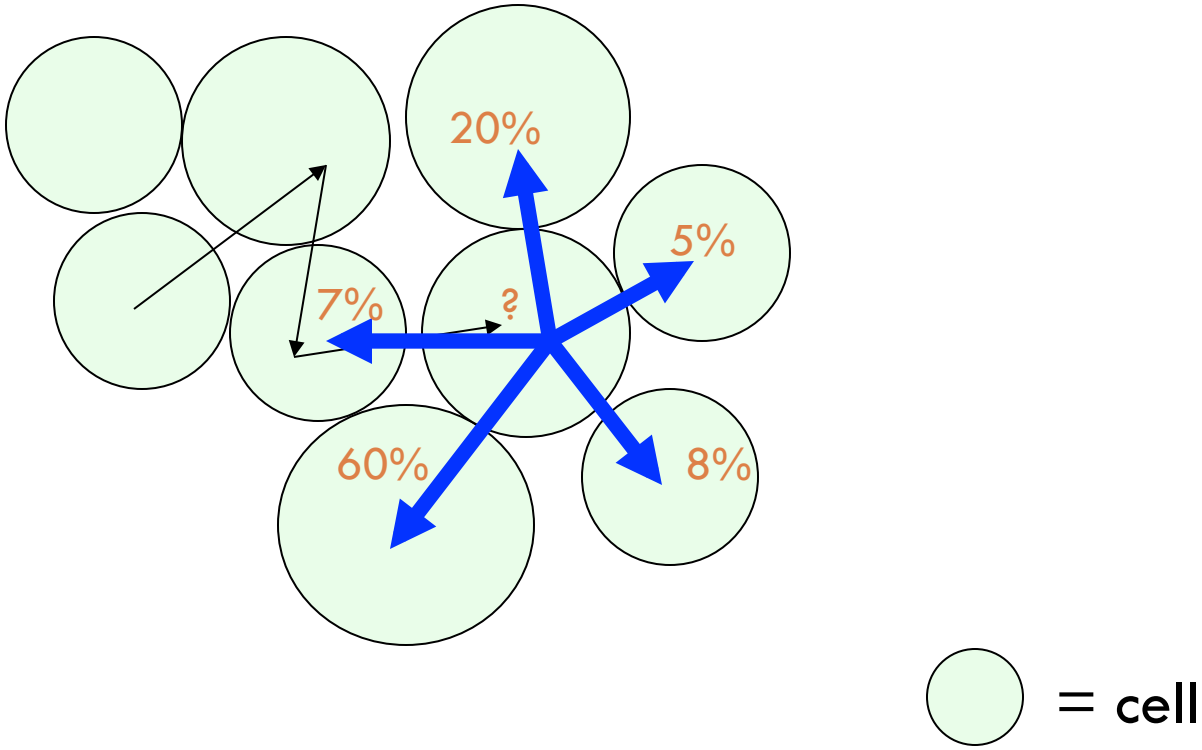
- Group together similar trajectories
- For each group produce a summary



= cell

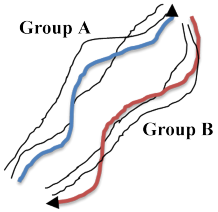
Trajectory classification and prediction

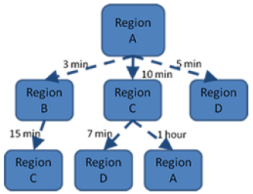
- Extract behaviour rules from history
- Use rules to predict behaviour of future users



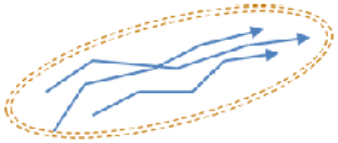
4-stage mobility data mining

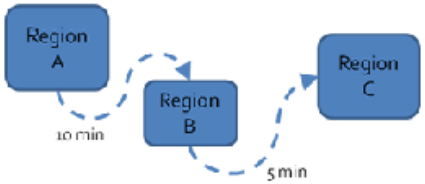
semantics

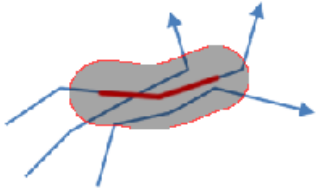




derived models



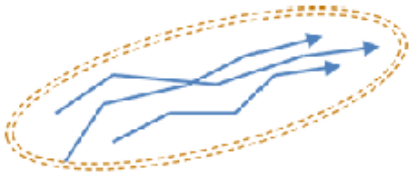




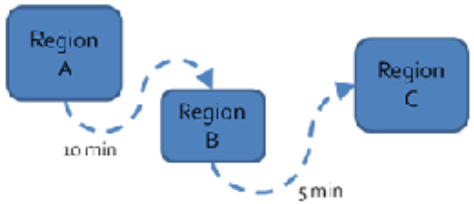
basic trajectory patterns and models

raw trajectory data

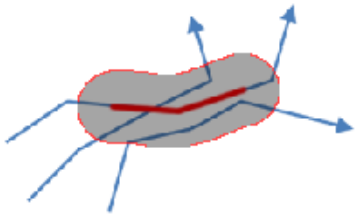
Basic mobility patterns and models



- T-Cluster: represents a group of similar trajectories

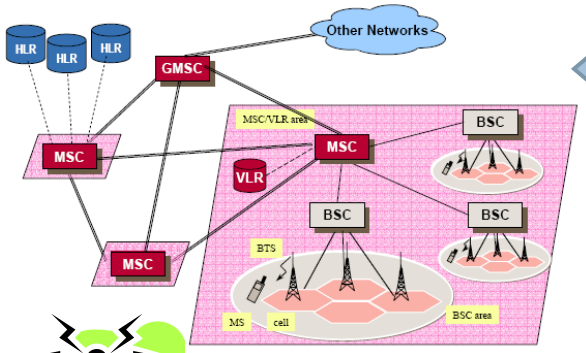


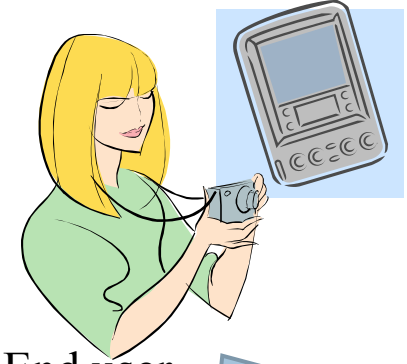
- T-Pattern: represents trajectory segments that visit a sequence of regions with similar transition times



- T-Flock: represents trajectory segments that move together for a time interval

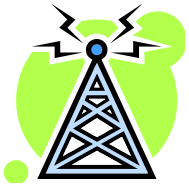
Mobile phone data, GPS tracks

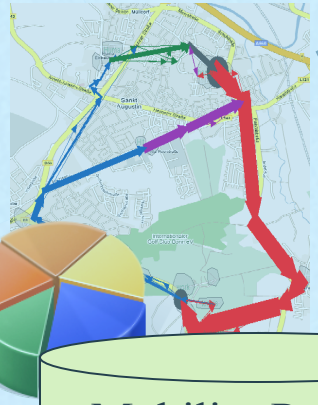




End user

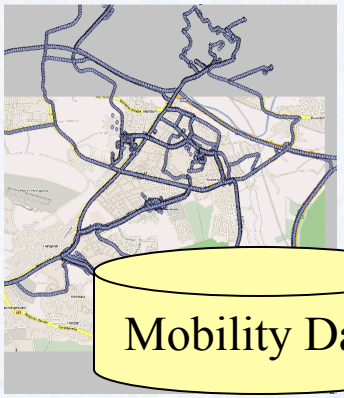
Mobility manager





Mobility Patterns

Mobility Data Mining



Mobility Data

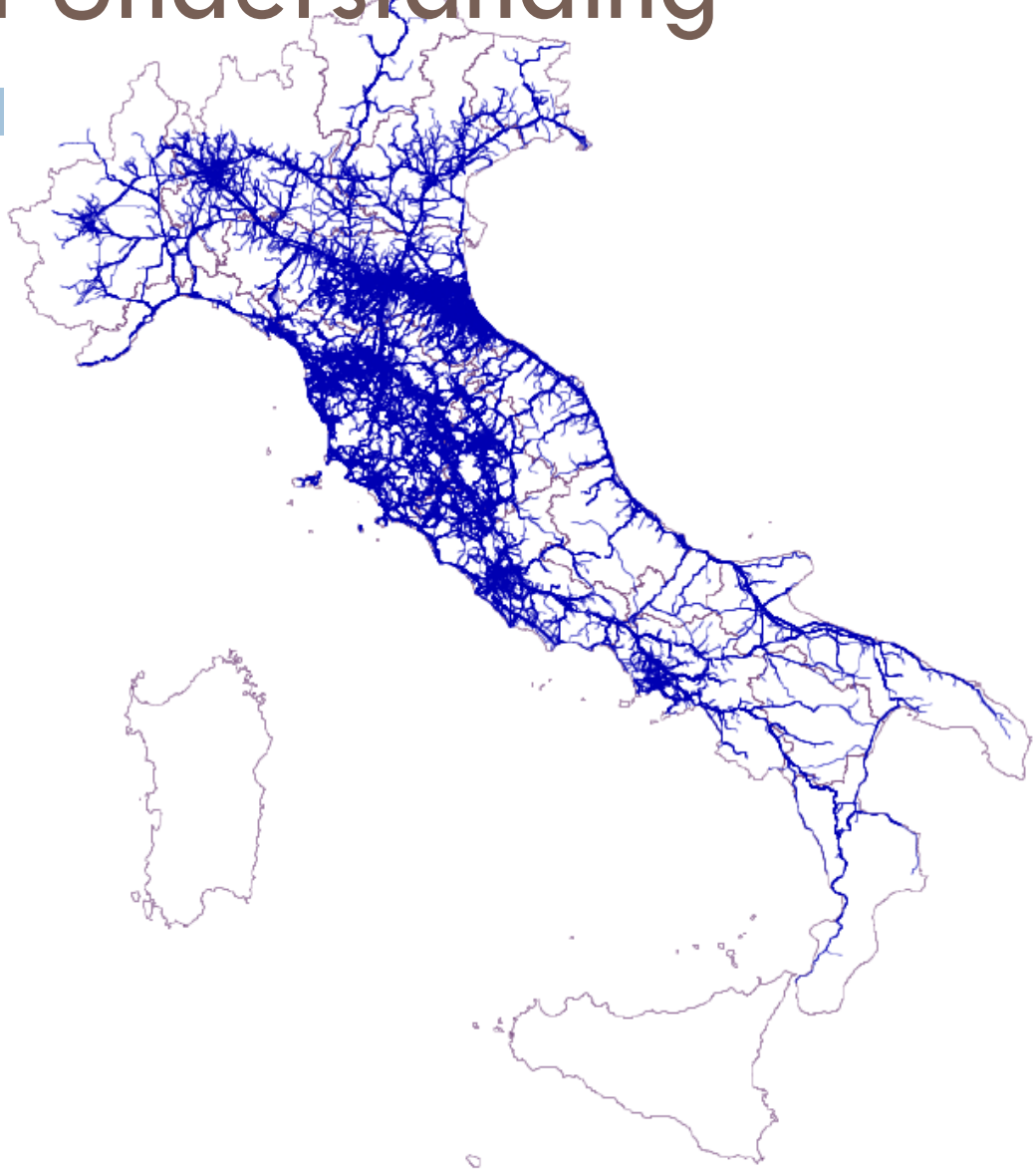
```
name|date|y|x
Prinzessin|08.20.1998|52.118|12.087
Prinzessin|08.23.1998|51.019|15.309
Prinzessin|08.26.1998|47.723|22.786
Prinzessin|08.29.1998|43.040|27.119
Prinzessin|08.31.1998|38.715|32.165
Prinzessin|09.01.1998|37.195|35.255
Prinzessin|09.03.1998|37.070|36.021
Prinzessin|12.08.1998|12.562|34.777
Prinzessin|12.10.1998|9.124|35.644
...
```

Raw data

Privacy and anonymity protection

UNDERSTANDING HUMAN MOBILITY WITH GPS

Data Understanding



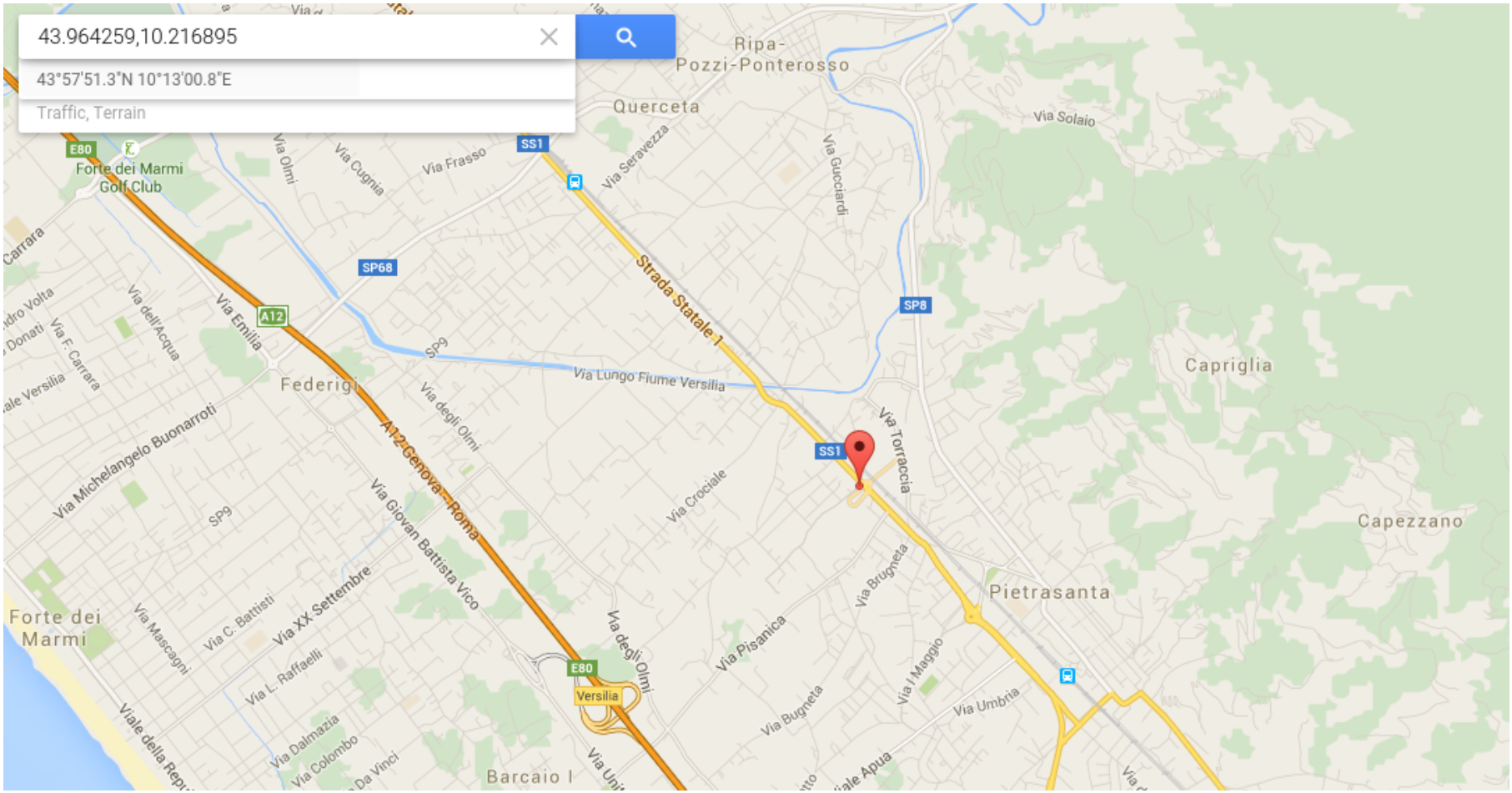
Raw GPS Data

ID Timestamp Latitude Longitude Others (optional)

- 946826,14/06/10 14:08:54,43964259,10216895,0,0,1,0,0
- 457380,13/06/10 22:05:27,43682201,10408320,0,0,3,0,0
- 457380,13/06/10 22:06:00,43682688,10408501,10,10,3,1,33
- 457380,13/06/10 22:06:34,43683609,10409146,14,24,3,1,115
- 457380,13/06/10 22:07:09,43685653,10410117,52,18,3,1,241
- 457380,13/06/10 22:07:43,43689775,10412032,50,18,3,1,484
- 457380,13/06/10 22:08:19,43692906,10413910,32,356,3,1,401
- 457380,13/06/10 22:08:53,43690801,10415016,60,126,3,1,279
- ...

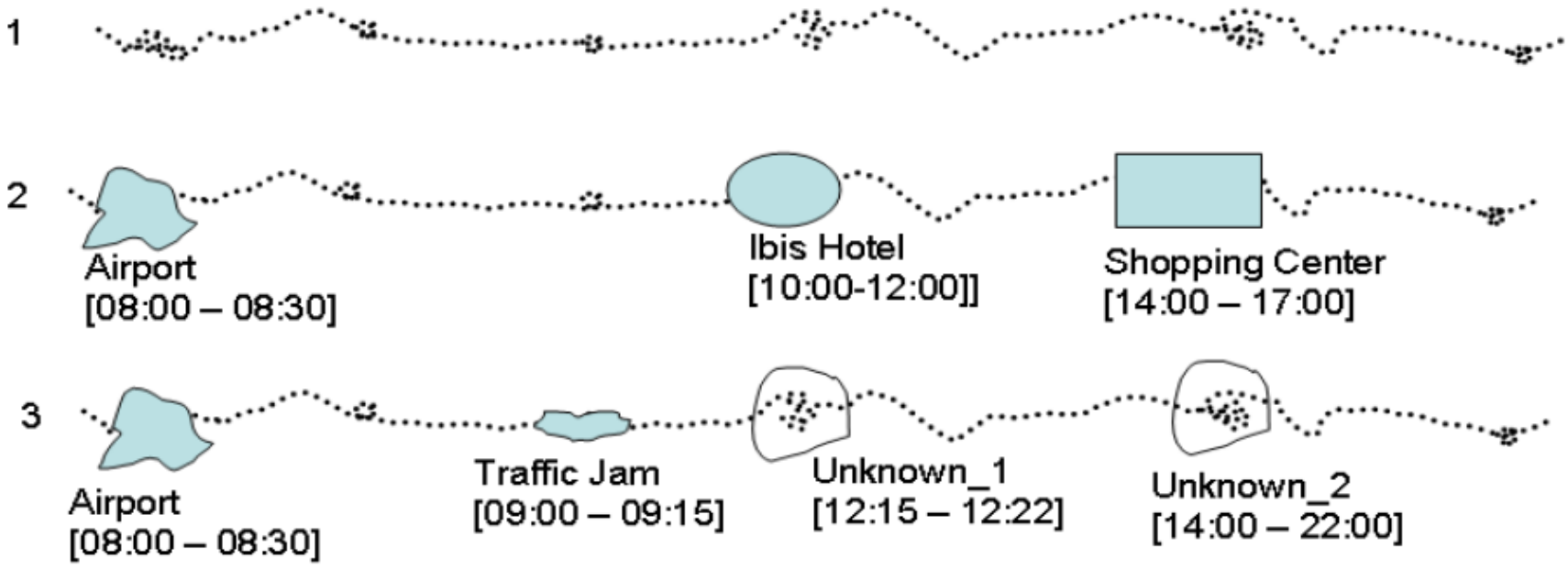
Sample point on the map

- 946826,14/06/10 14:08:54,43964259,10216895,0,0,1,0,0



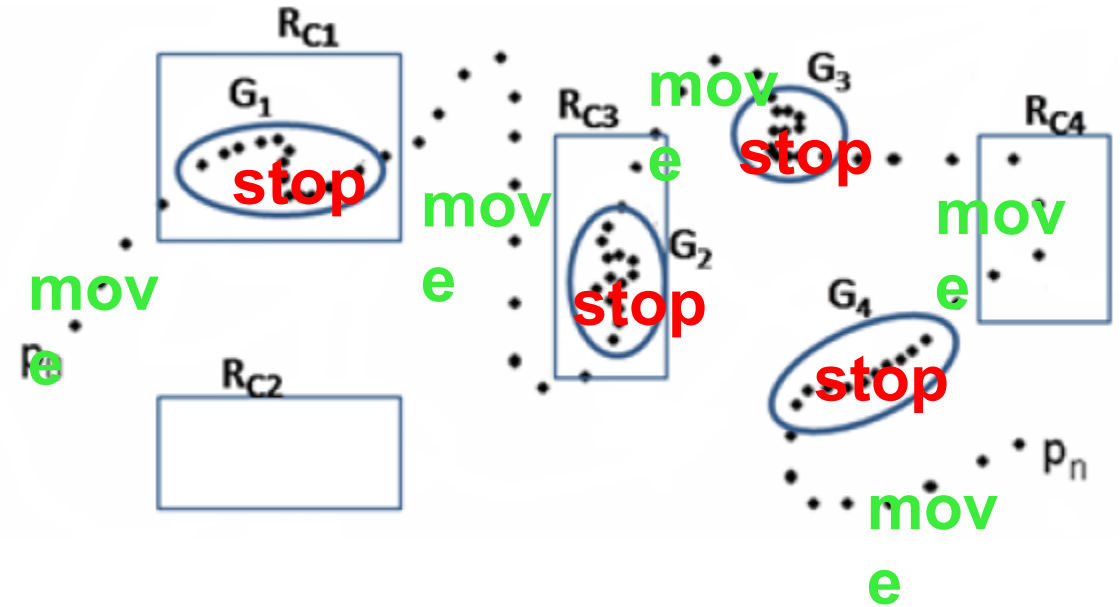
Trajectory reconstruction

- Raw data forms a continuous stream of points
- How to cut it into stops and trips?
  - Example on smart phone traces :



Trajectory reconstruction

- General criteria based on speed
  - If it moves very little (threshold Th_S) over a significant time interval (threshold Th_T) then it is practically a stop
  - Trajectory (trip) = contiguous sequence of points between two stops



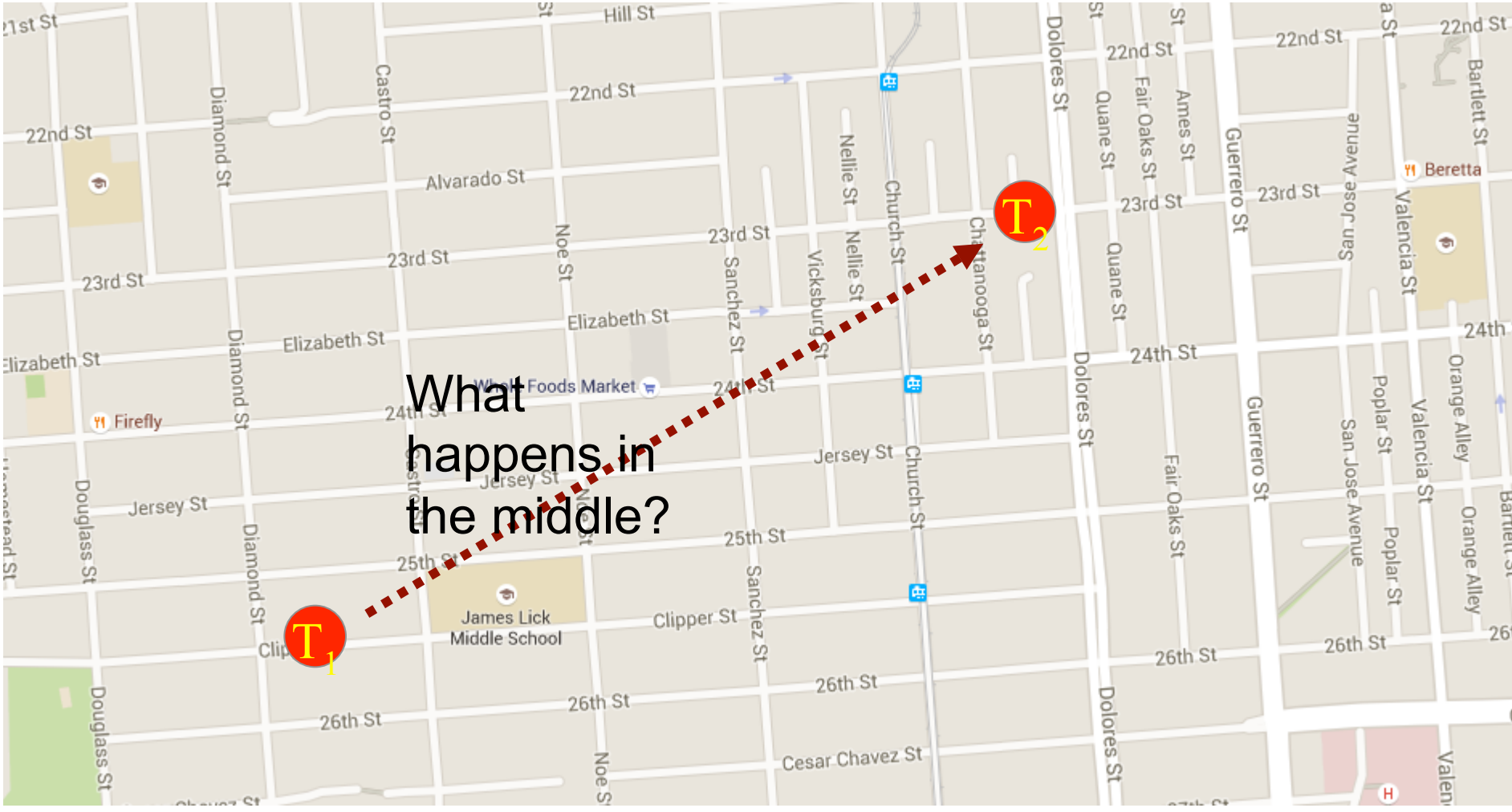
Trajectory reconstruction

- Special cases, easier to treat
  - Stop explicitly in the data: e.g. engine status on/off
    - Simply "cut" trajectories on status transitions

  - Device is off during stops:
    - Typical of cars data
    - A stop results in a time gap in the data
    - Exceptions: short stops might remain undetected

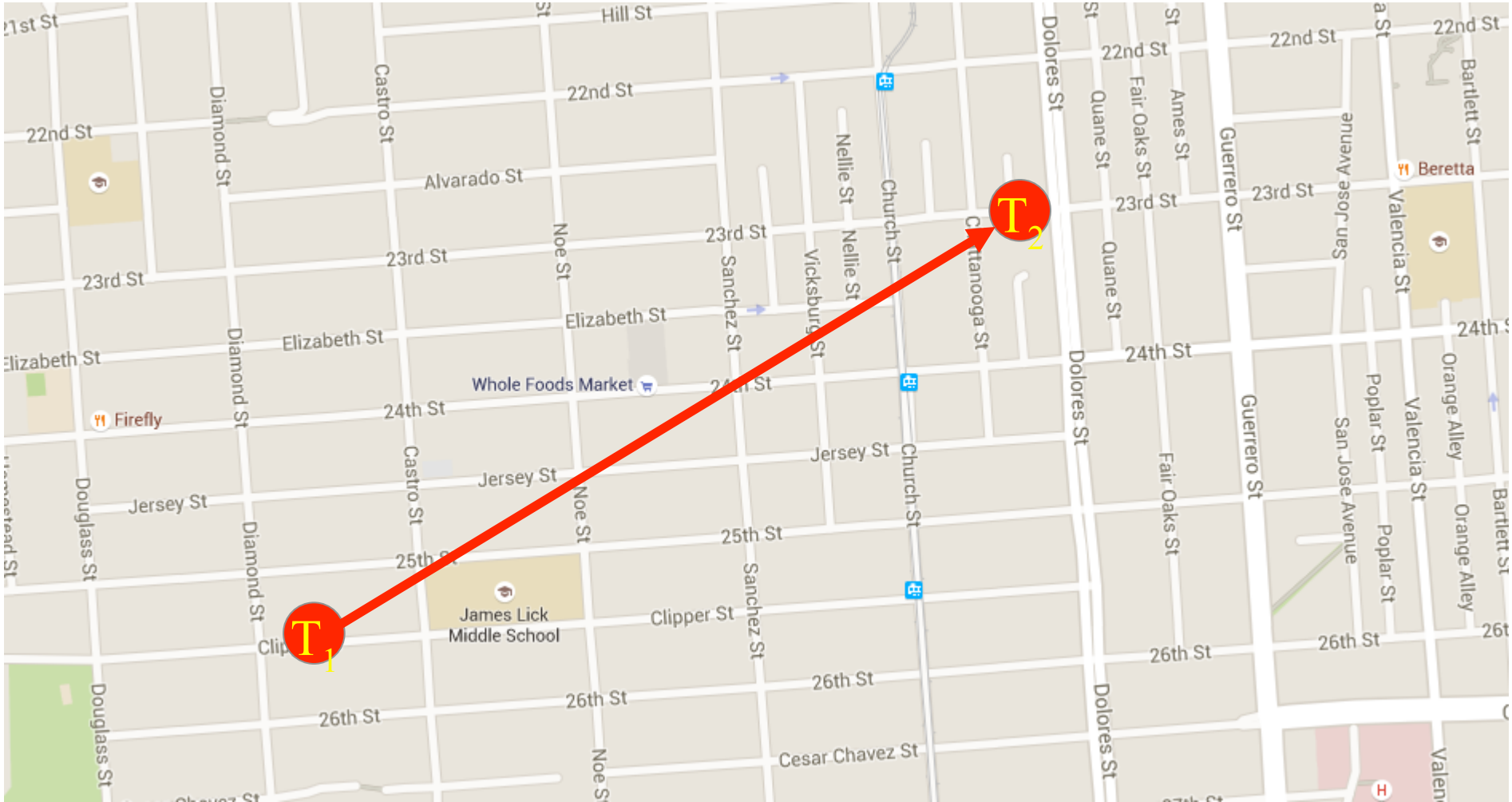
Gaps

- Sometimes the space/time gap between consecutive points is significant



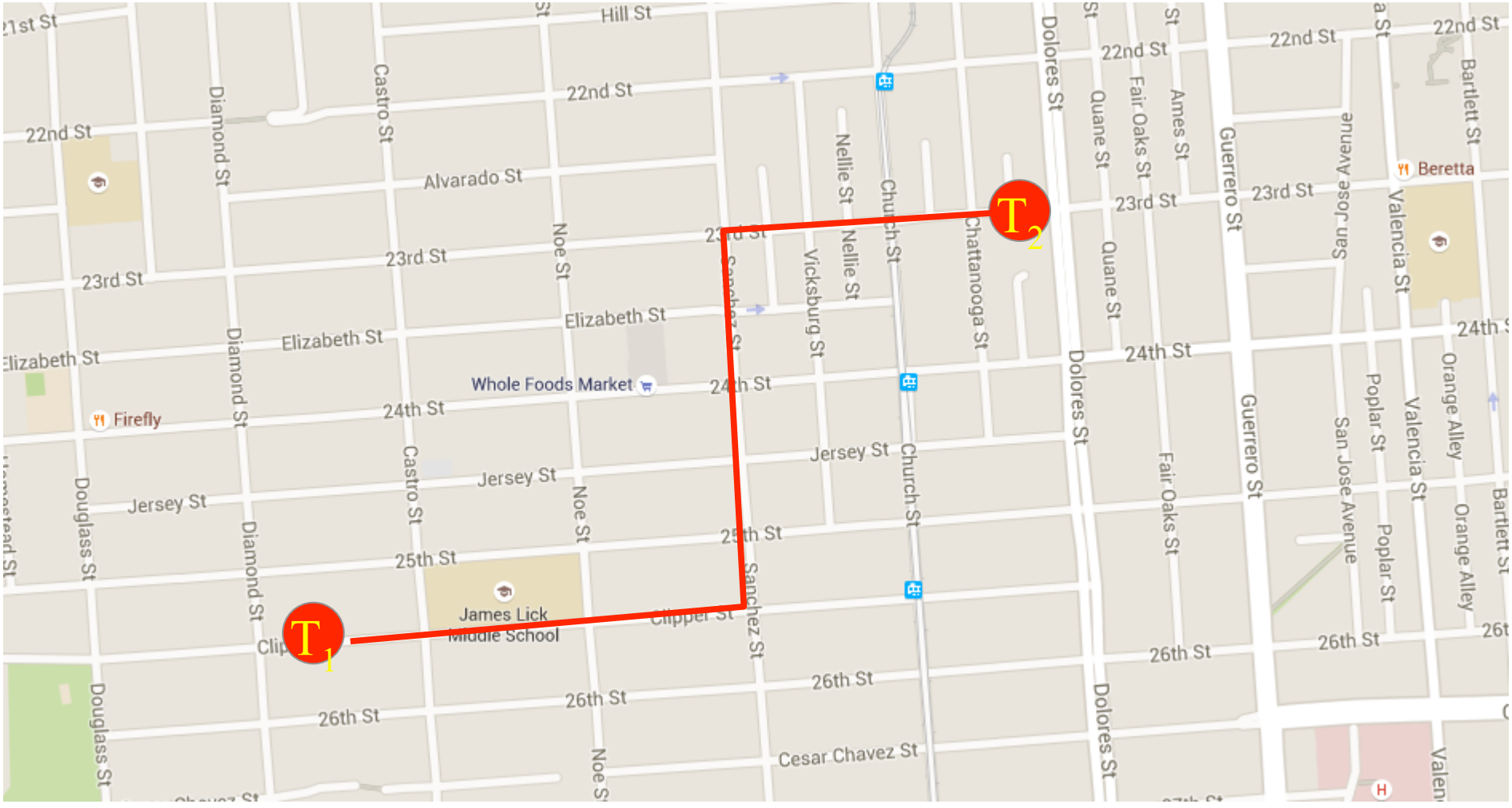
Free vs. constrained movement

- Typical solutions:
  - Free movement => straight line, uniform speed



Free vs. constrained movement

- Typical solutions:
  - Constrained movement => shortest path

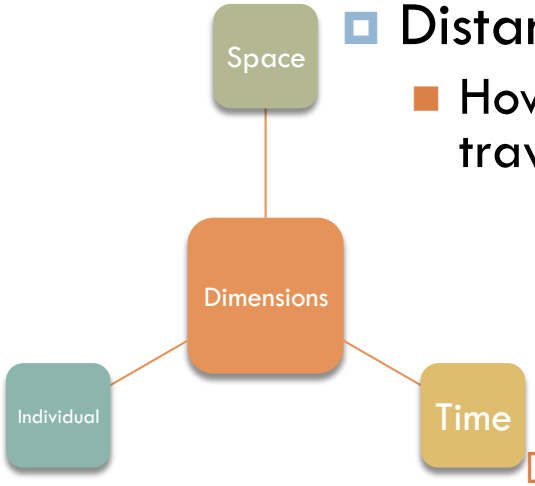


Understanding collective and individual mobility

F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. , 2011.

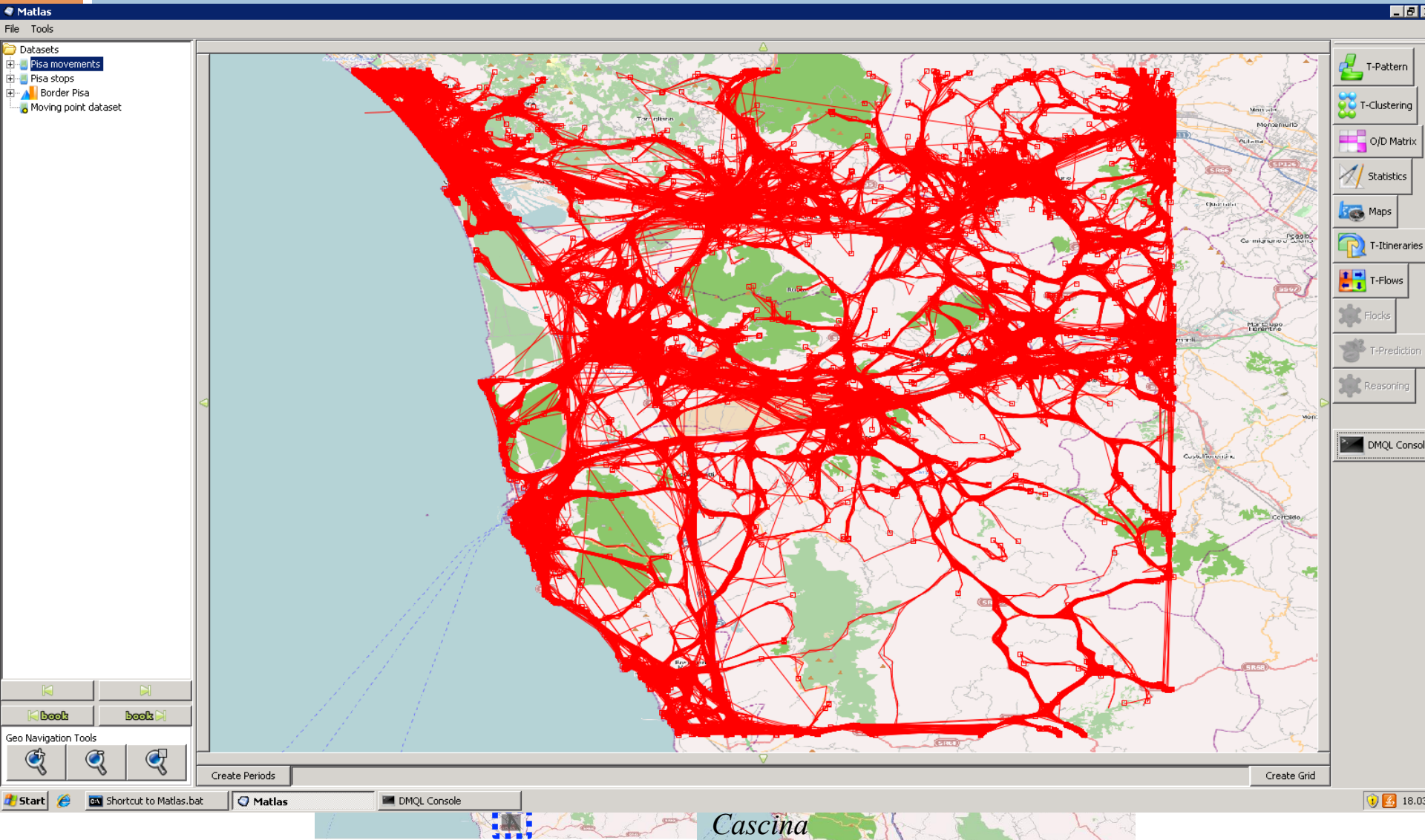
Dimensions to explore

- Space
  - Administrative borders
    - E.g.: city
  - Distance travelled
    - How much a person is travelling

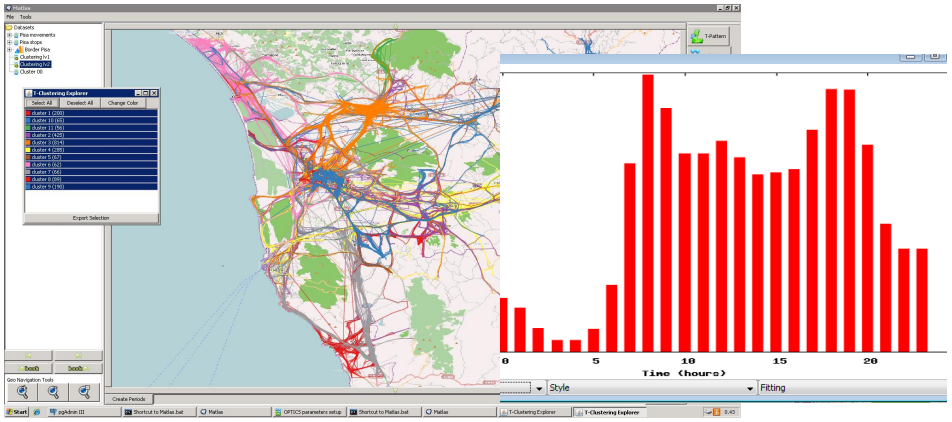


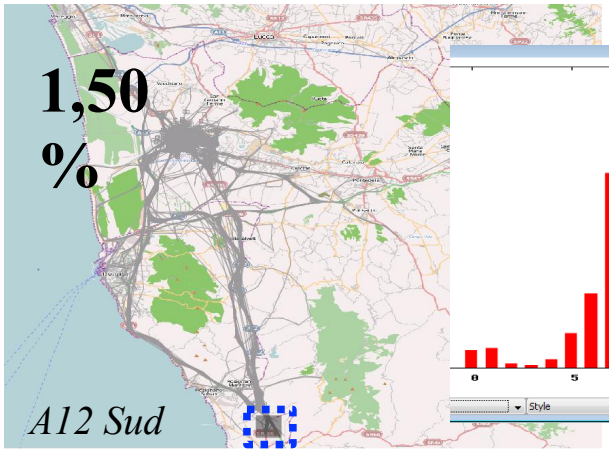
- Individual
  - Preferred locations
  - Activities
- Time
  - Hour of day
  - Day of week
  - Weekdays/weekends

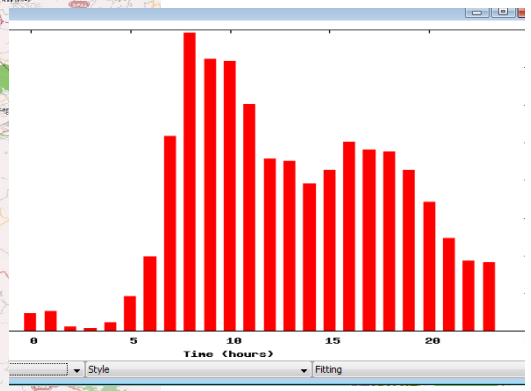
Access patterns using T-clustering

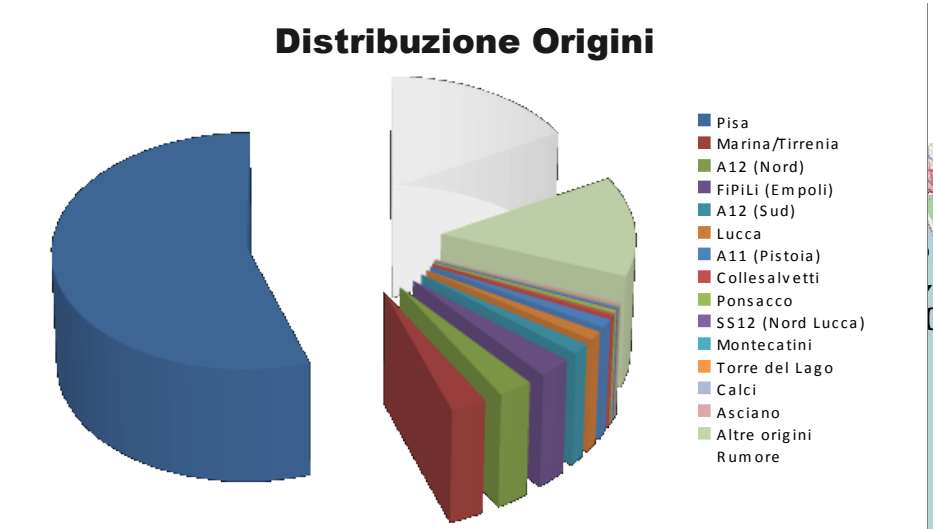


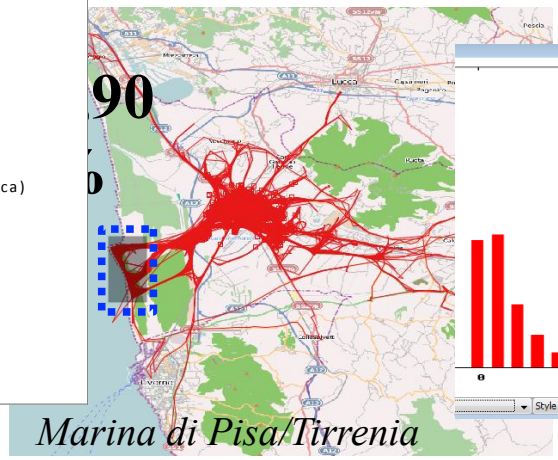
Characterizing the access patterns: origin & time

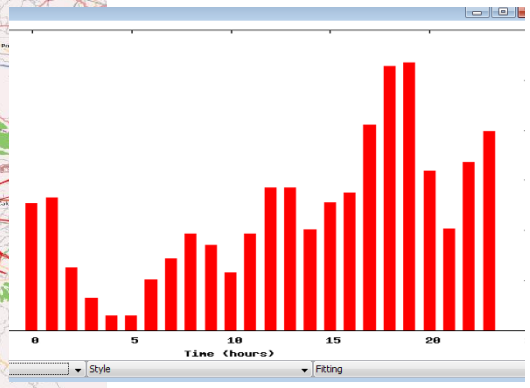




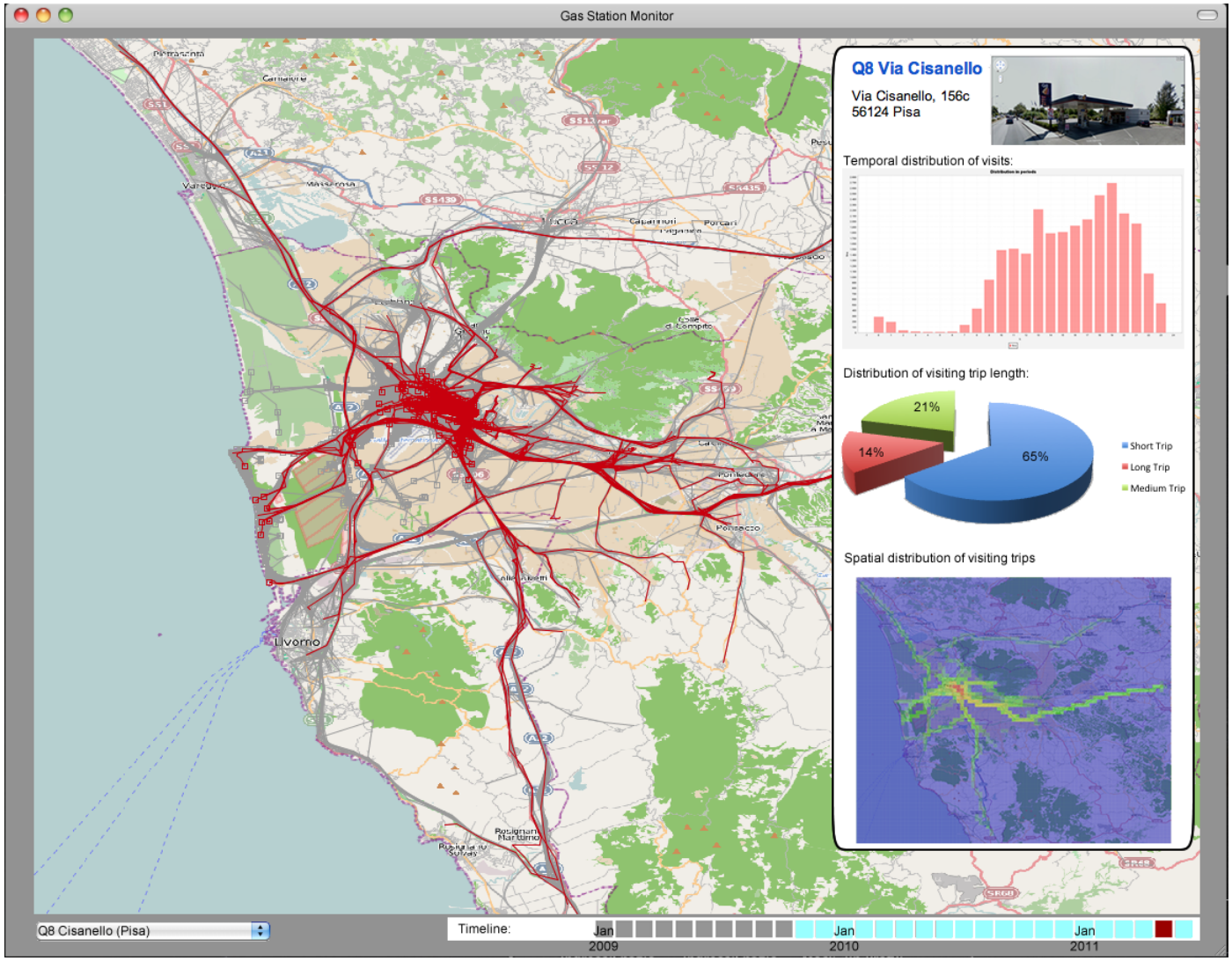








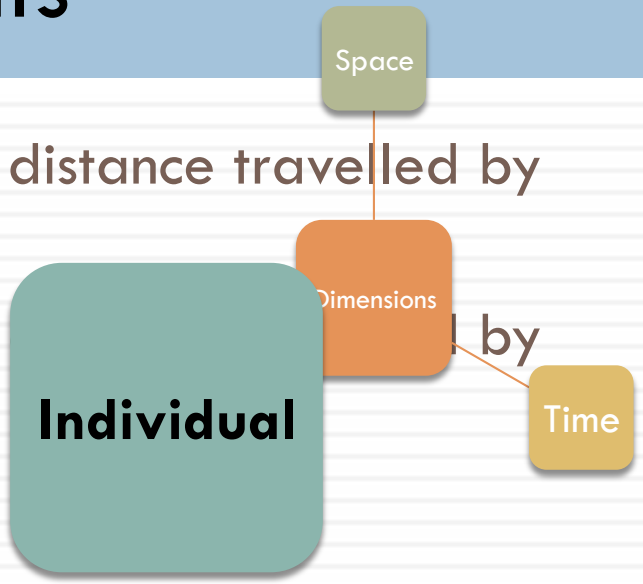
Studying the attractiveness/efficiency of a service with GPS tracks



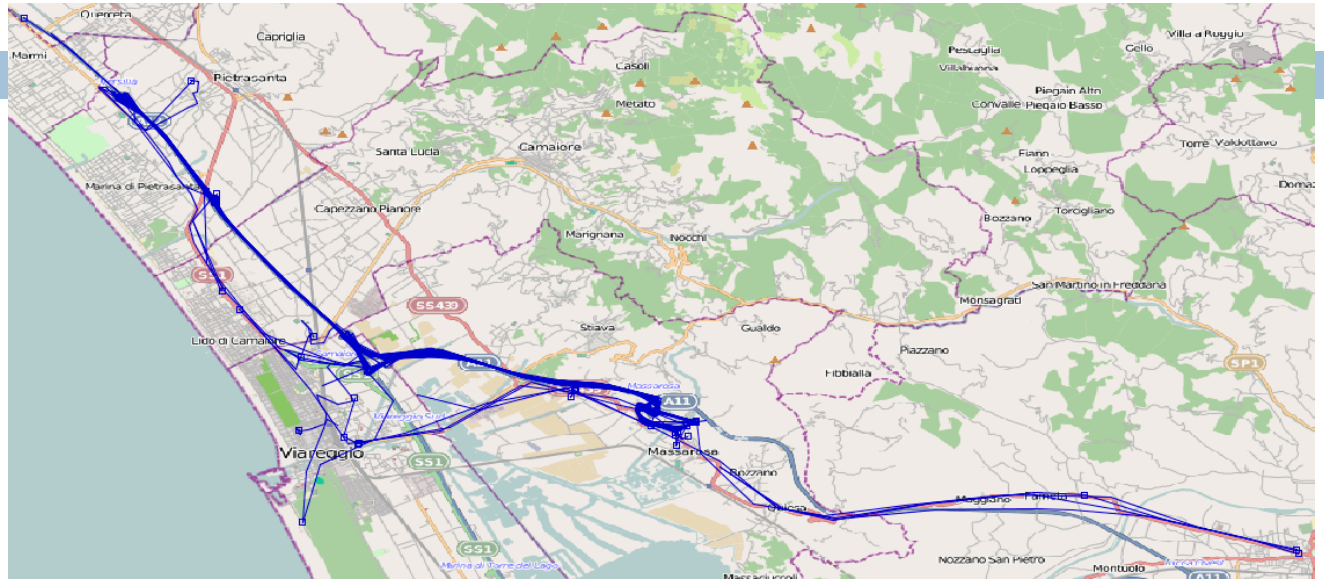
Systematic movements

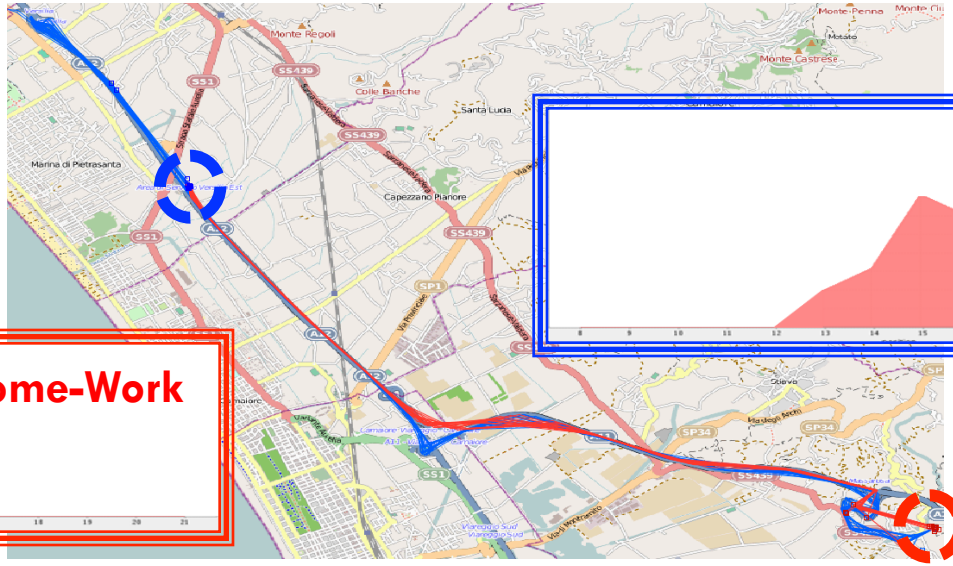
How to measure the characteristic distance travelled by an individual?

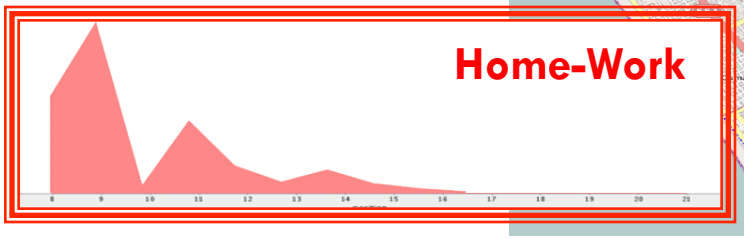
How to measure the characteristic [illegible] by the people of a territory



Discovering individual systematic movements

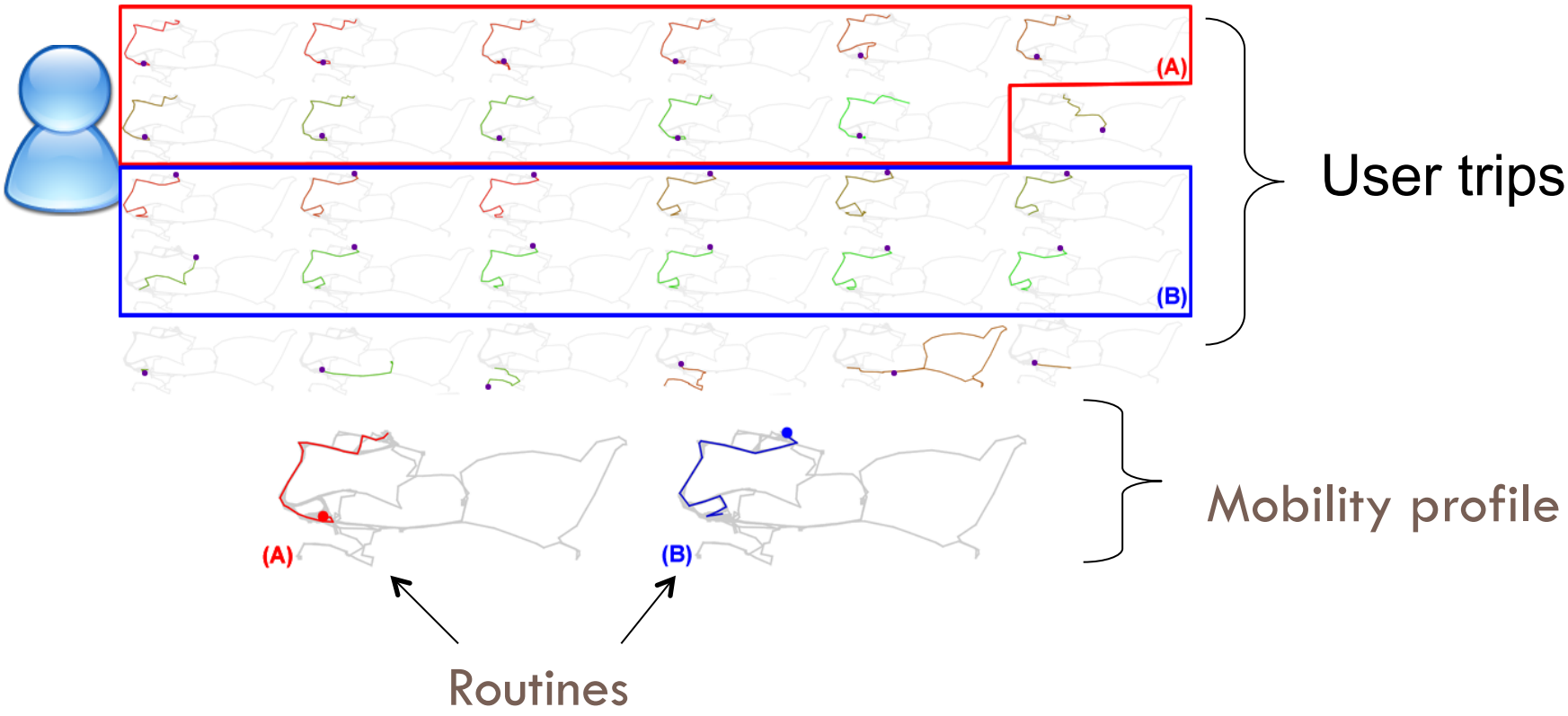




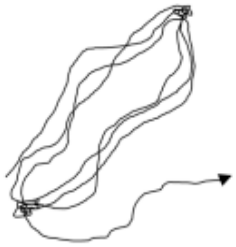


Extracting travellers profiles

- Analysis focused on the single individual
- Find his/her systematic mobility

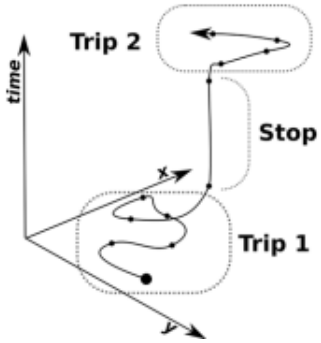


Mobility profiles



User history

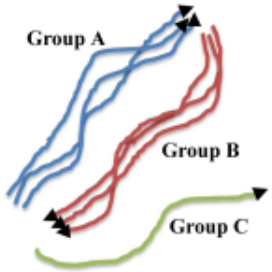
An ordered sequence of spatio-temporal points.



Trips construction

Cutting the user history when a **stop** is detected

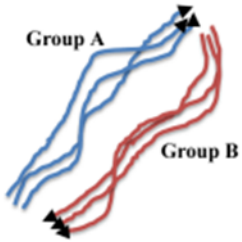
Stops Spatial Threshold
Stops Temporal Threshold



Grouping

Performing a density based **clustering** equipped with a **spatio temporal** distance function

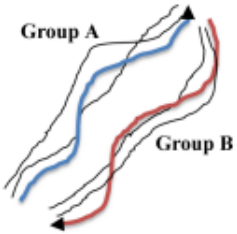
Spatial Tollerance
Temporal Tollerance
Spatio temporal distance



Pruning

Groups with a small Number of trips are Pruned

Support Threshold



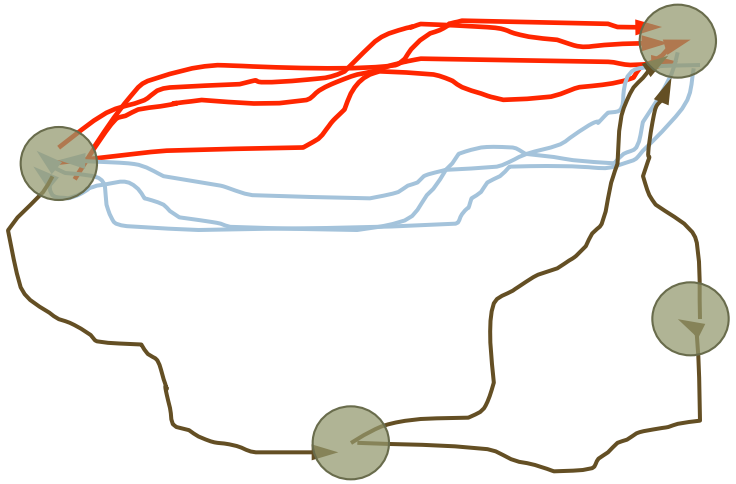
Profile extraction

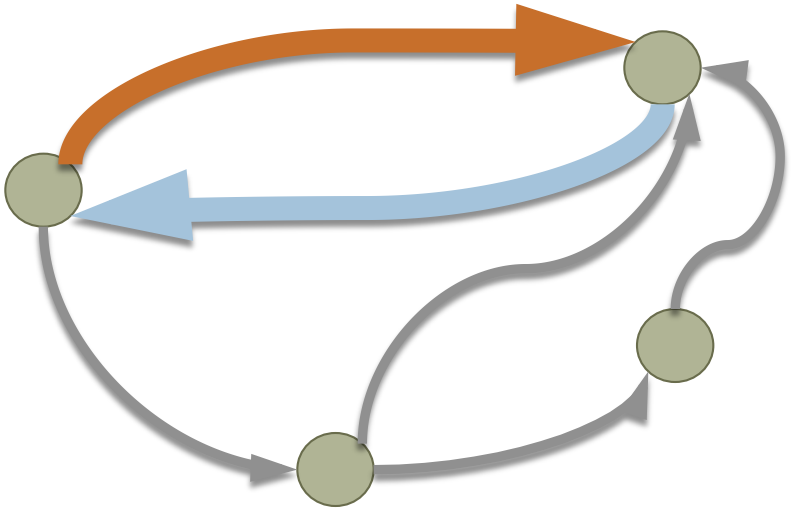
The **medoid** of each group becomes user's **routines** and the all set become the user's **mobility profile**

Trasarti, Pinelli, Nanni, Giannotti.
Mining mobility user profiles for car pooling. ACM SIGKDD 2011

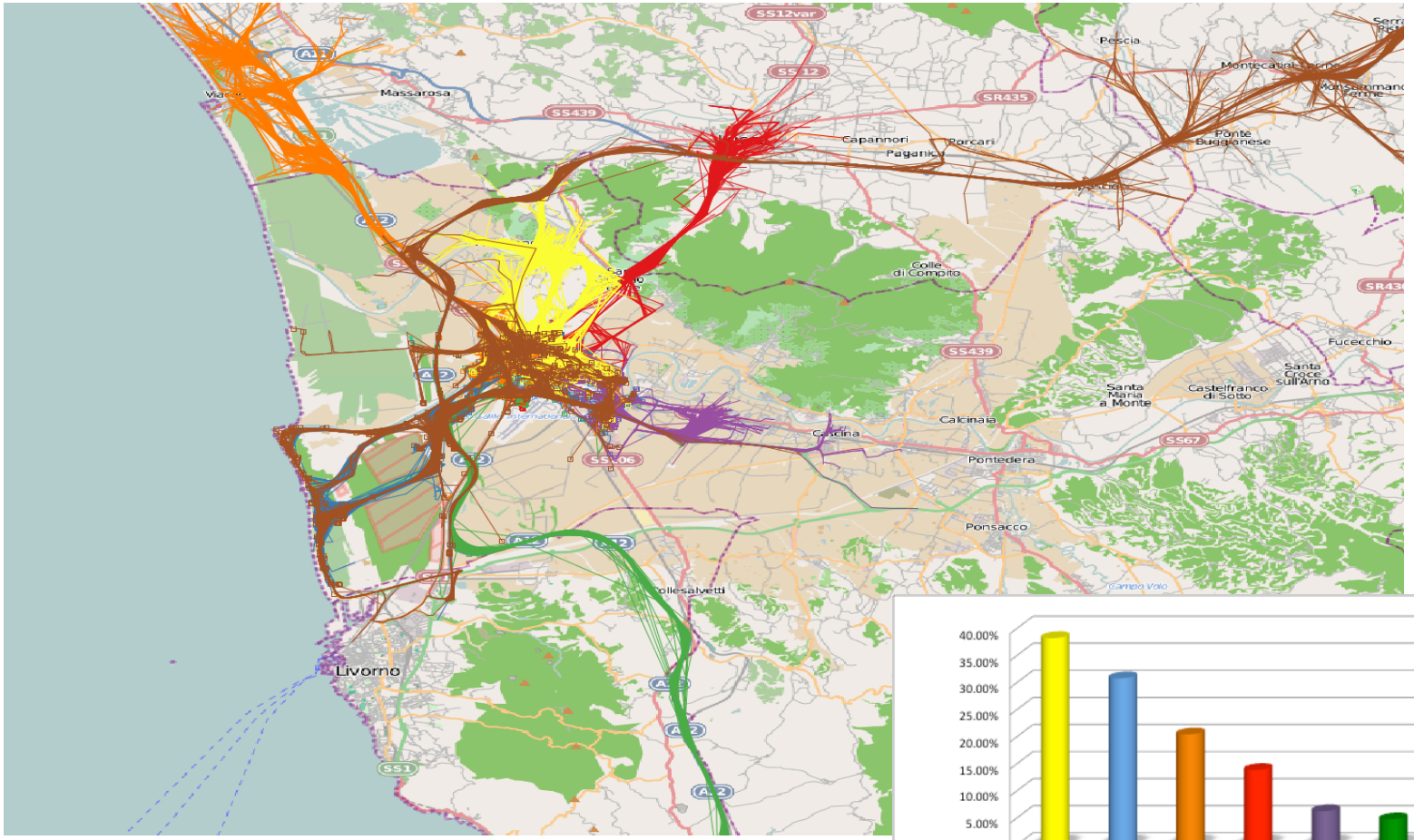
From Profiles to Systematicity Indicator

- Each routine of a profile is associated with a measure of frequency
- Routines are sorted according to their frequency: rank 1, rank 2, rank 3, …
- A minimum frequency threshold allow to distinguish a **systematic** trip from an **occasional** one

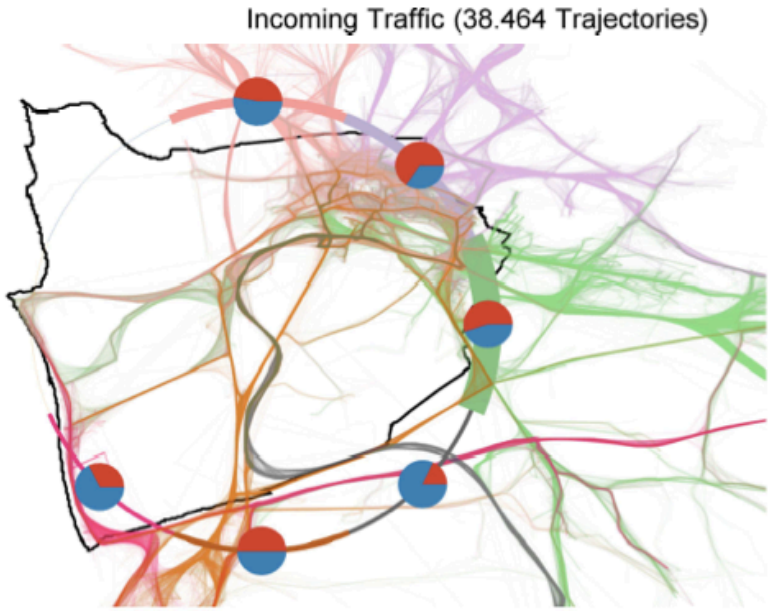




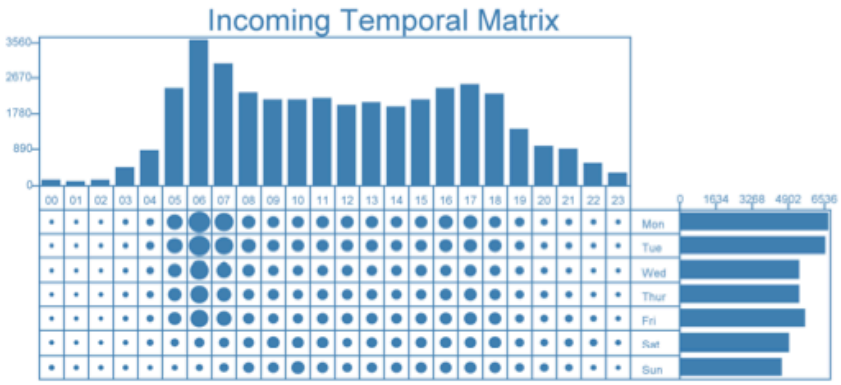
City access paths vs systematic movements

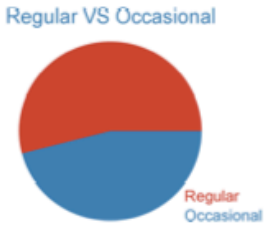


Pisa – Inbound traffic



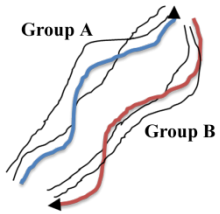
	City	Traj	Perc
NORD 32%	San Giuliano T..	4.816	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	95%
	Calci	795	79%
	Calcinaia	693	92%

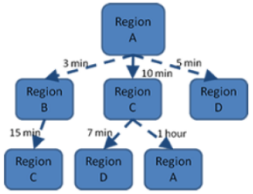




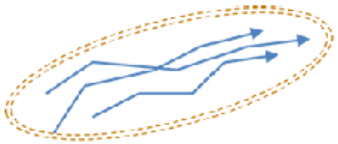
4-stage mobility data mining

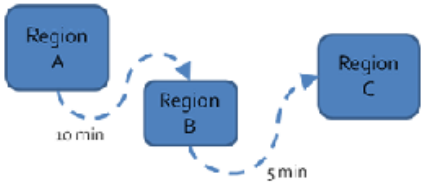
semantics

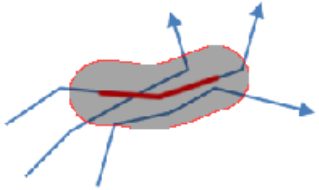




derived models



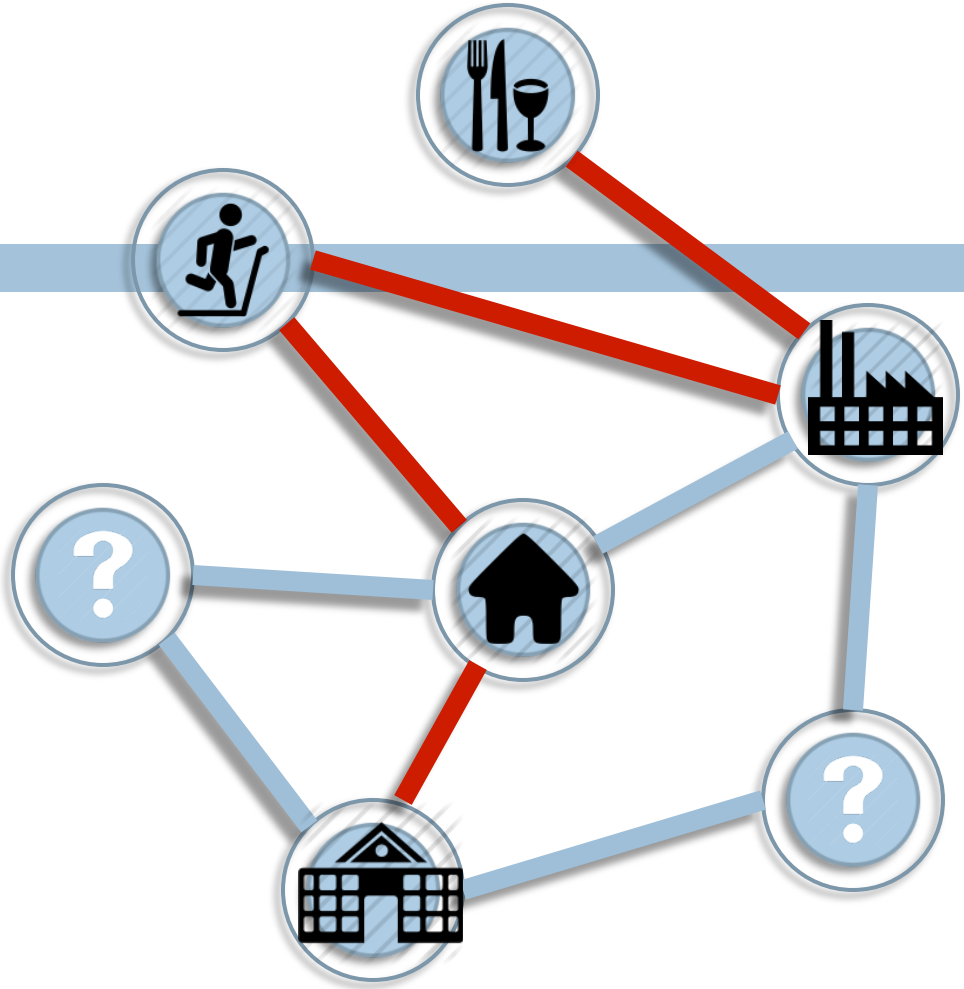




basic trajectory patterns and models

raw trajectory data

Activity Recognition –



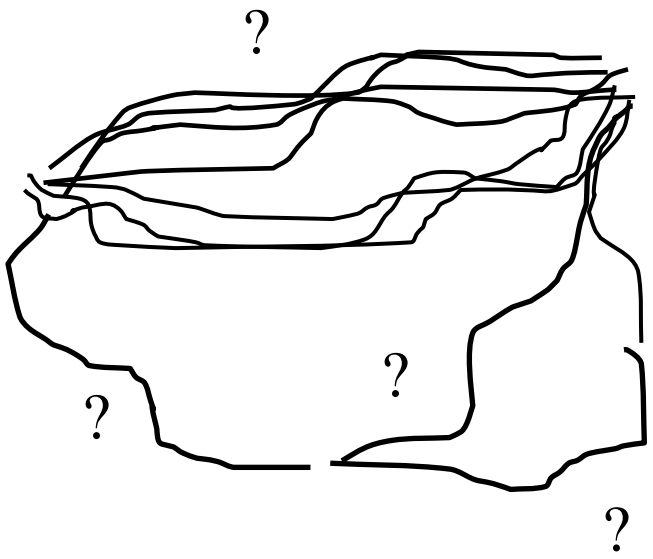
DSAA2014

The Purpose of Motion

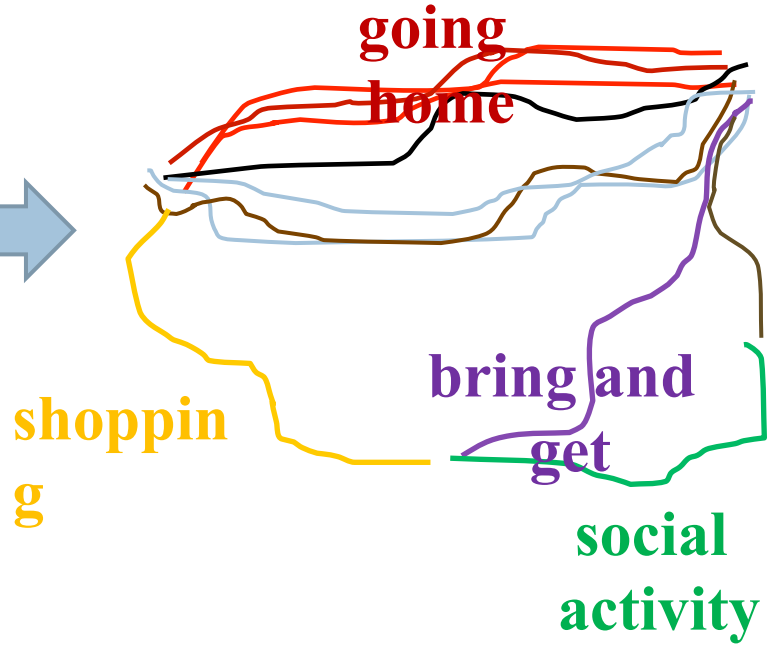
Learning Activities from Individual Mobility Networks

From raw trajectories to activity diaries

Mobility trajectories



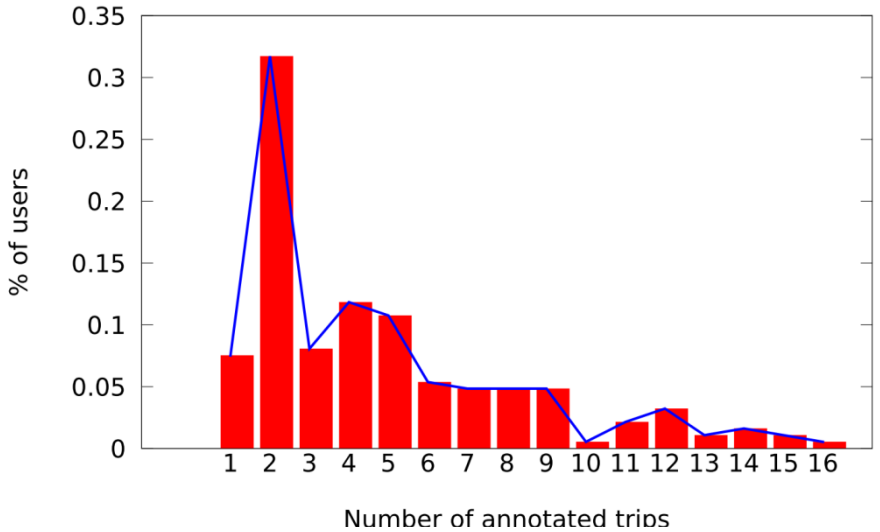
Semantic annotation

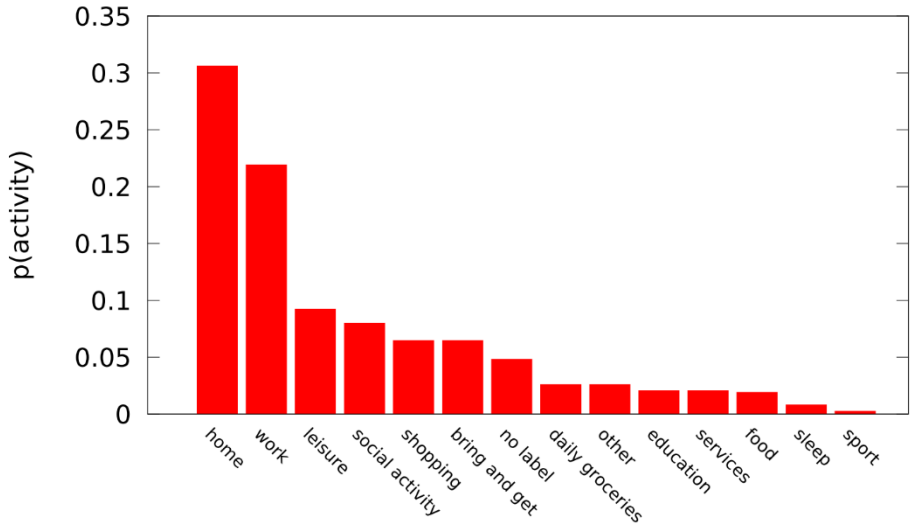


Learn from survey data

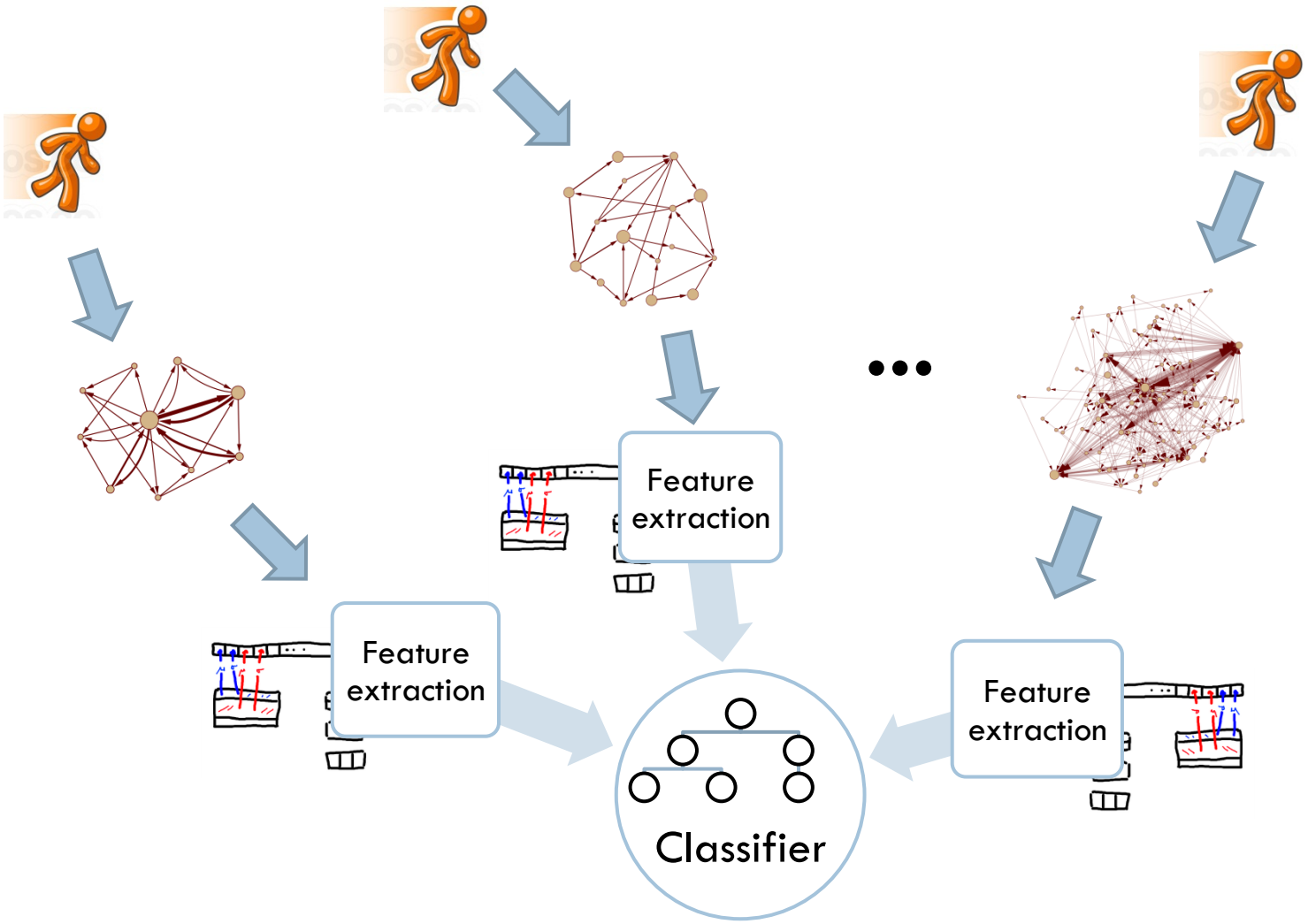
activities

Home	Work	Social
Leisure	Shopping	Daily shopping
Eating	Touring	Sleeping
Education	Services	Bring and get





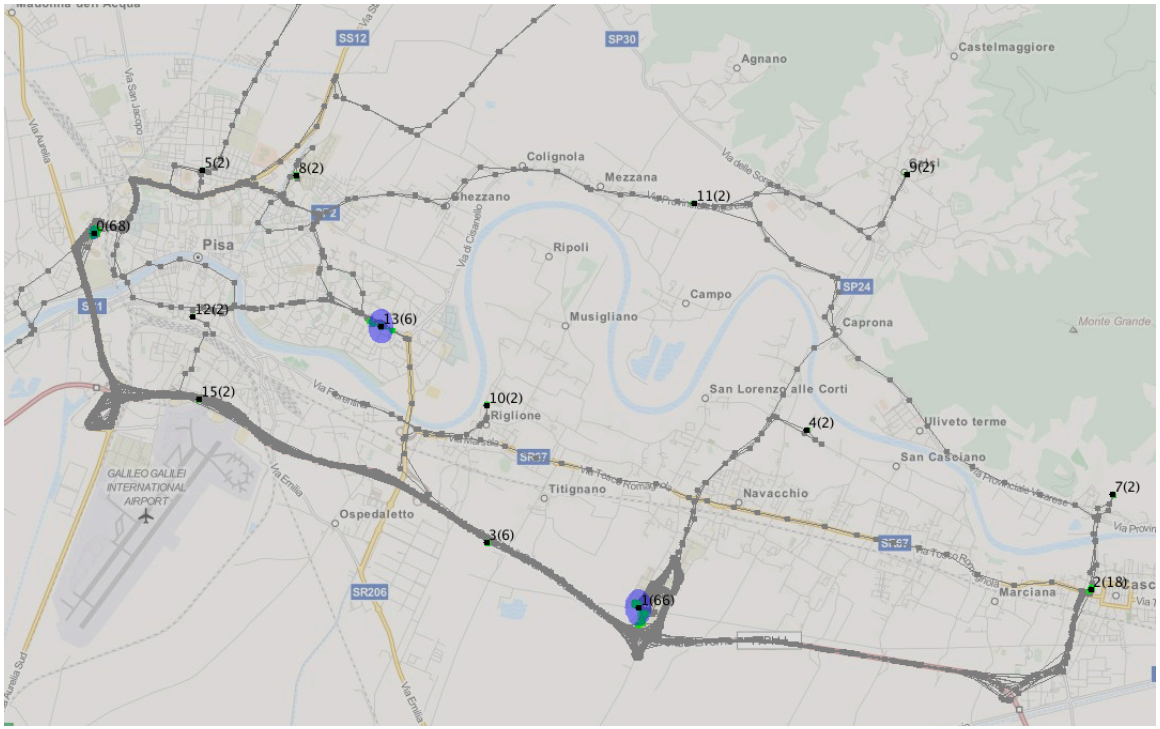
Semantic diary classifier



The process

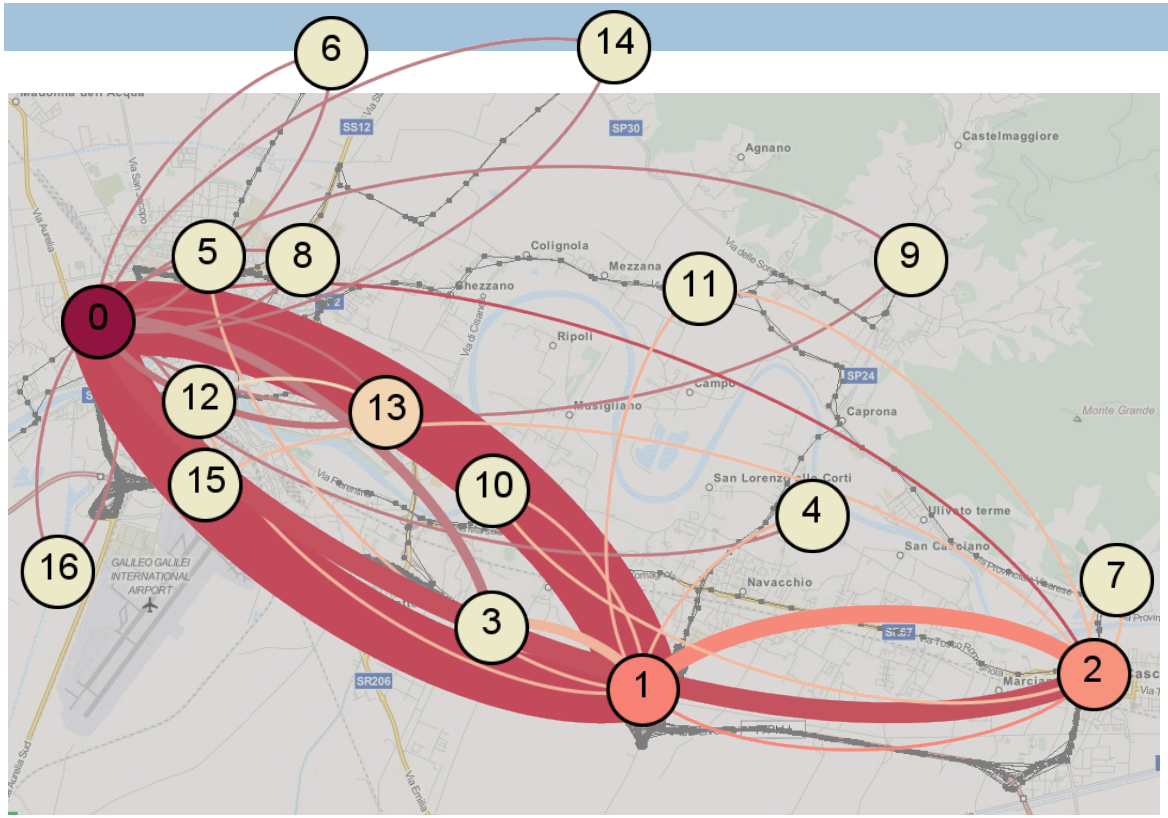
- Build from data an Individual Mobility Network (IMN)
- Extract structural features from the Individual Mobility Network (INM)
- Use survey data for annotate some INM
- Learn a model using **cascading classification with label propagation (ABC classifier)**
- Use the model

How to synthesize Individual Mobility?



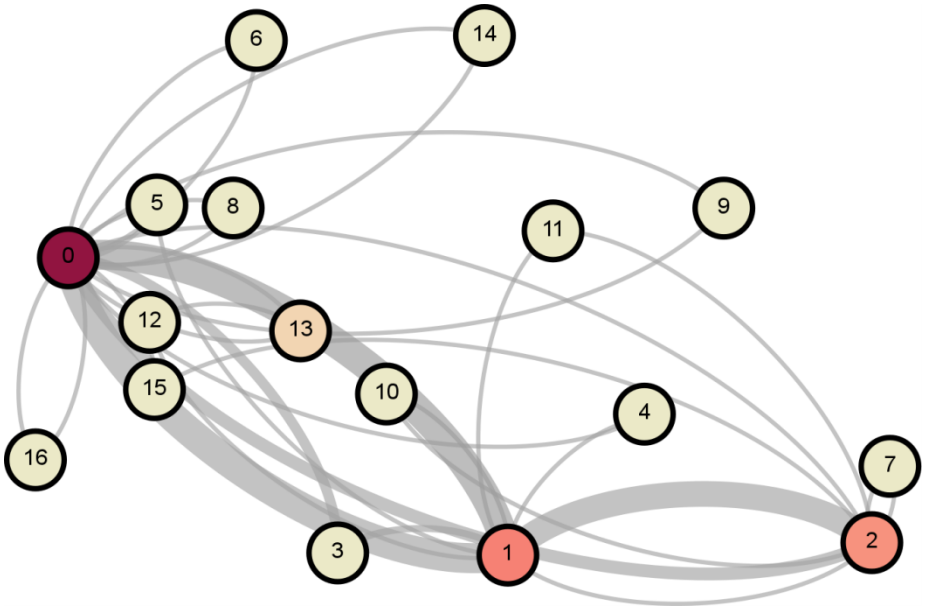
Mobility Data Mining methods automatically extract relevant episodes: locations and movements.

How to synthesize Individual Mobility?



Graph abstraction based on locations (nodes) and movements (edges)

Extracting the IMN

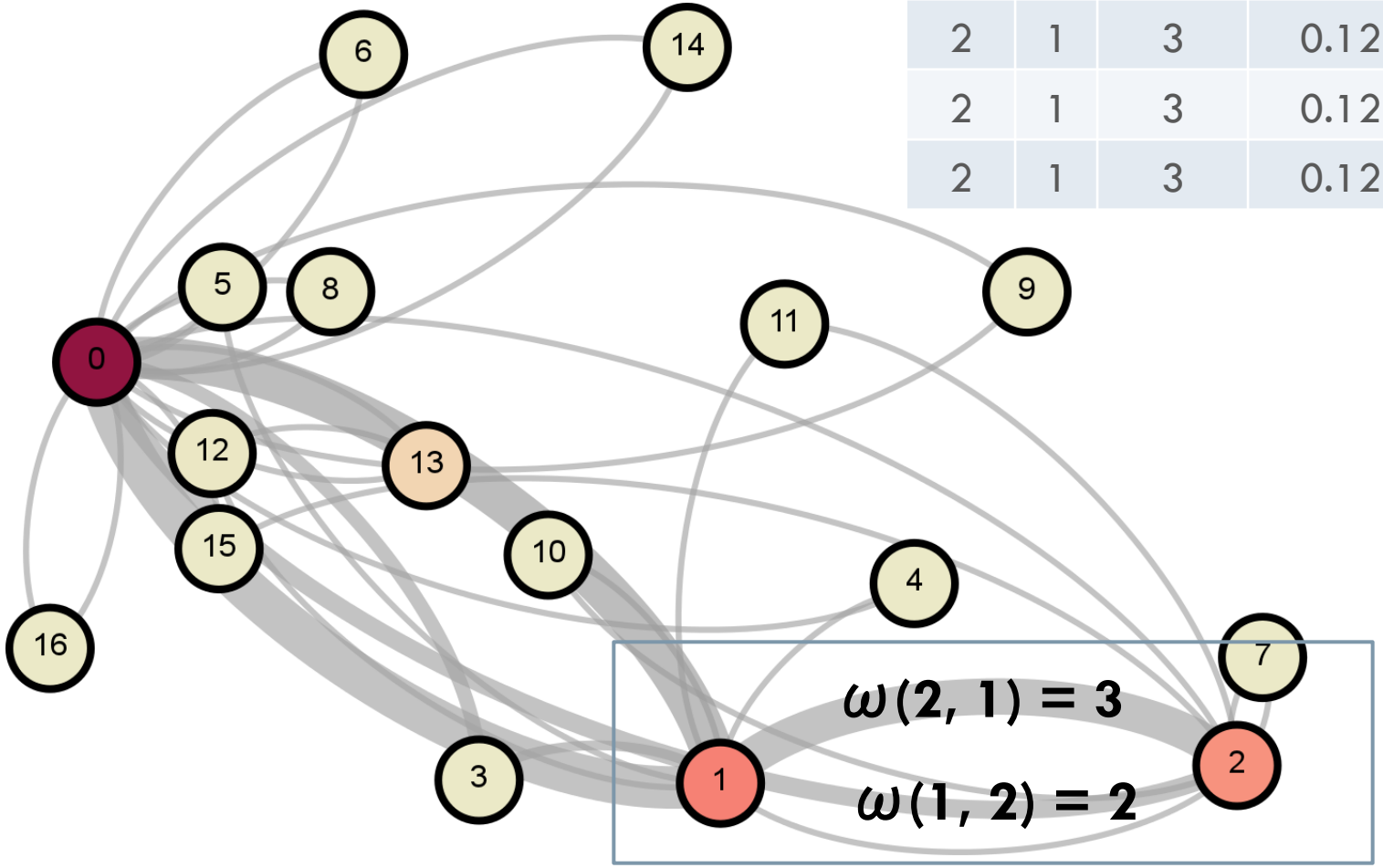


Trip Features
Length
Duration
Time Interval
Average Speed

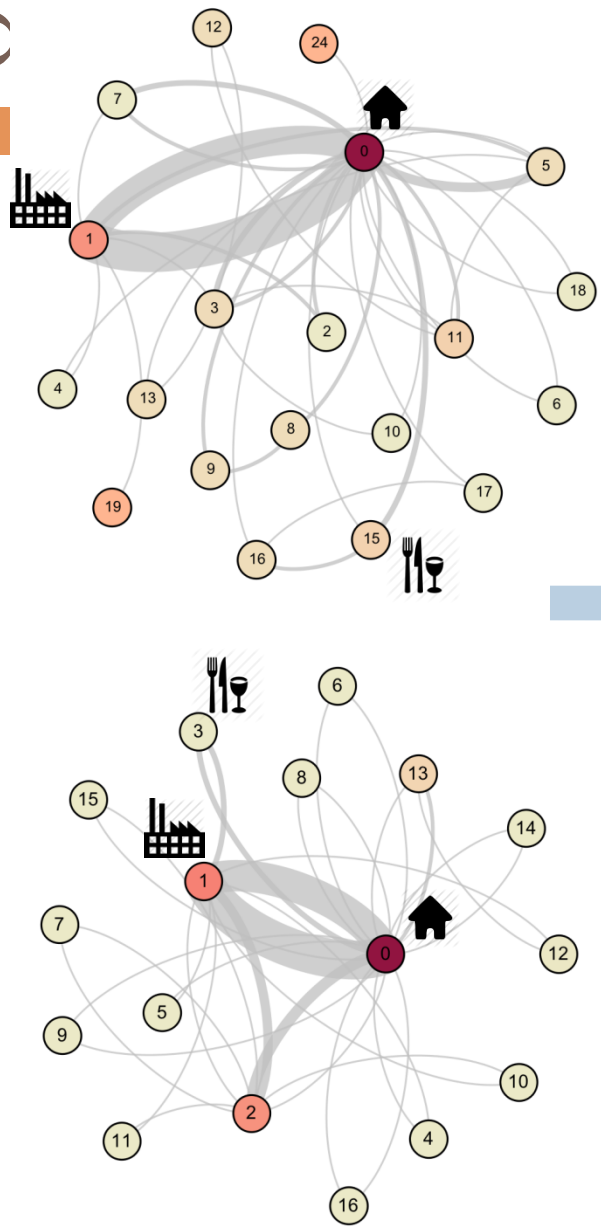
Network Features	
centrality	clustering coefficient average path length
predictability	entropy
hubbiness	degree betweenness
volume	edge weight flow per location

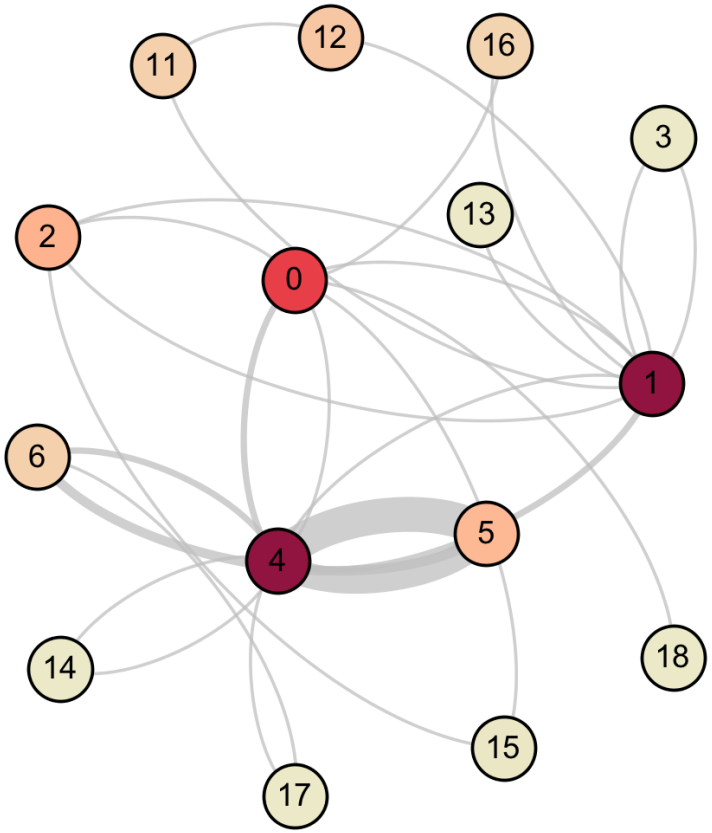
Extracting the IMN

from	to	weight	ccFrom	ccTo	duration
1	2	2	0.22	0.12	10 min
1	2	2	0.22	0.12	5 min
2	1	3	0.12	0.22	4 min
2	1	3	0.12	0.22	6 min
2	1	3	0.12	0.22	4 min

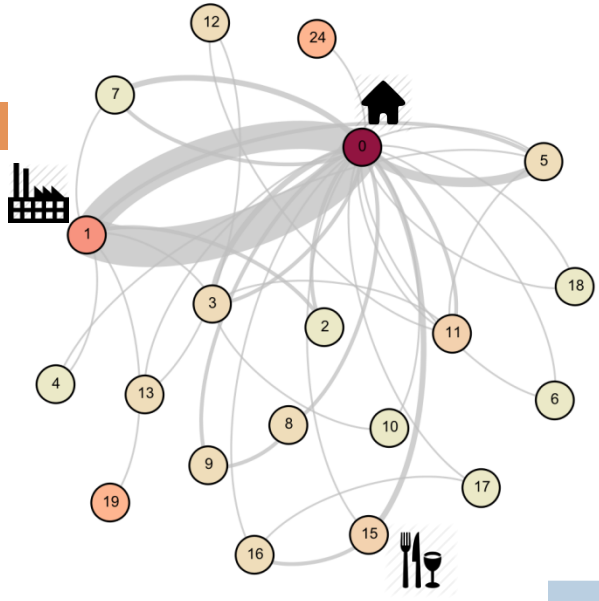


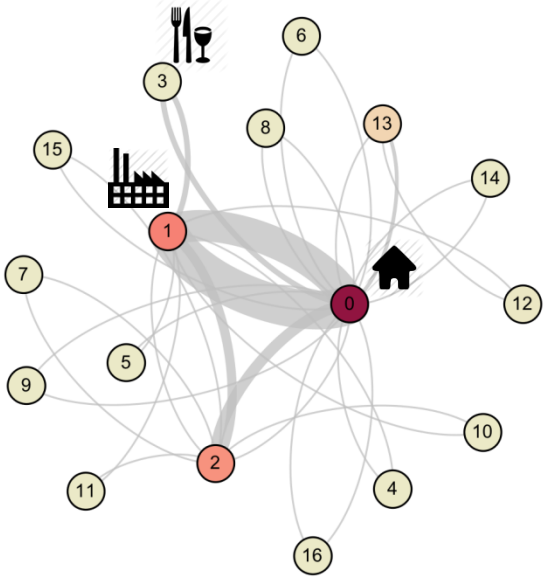
From many annotated IMN learn c

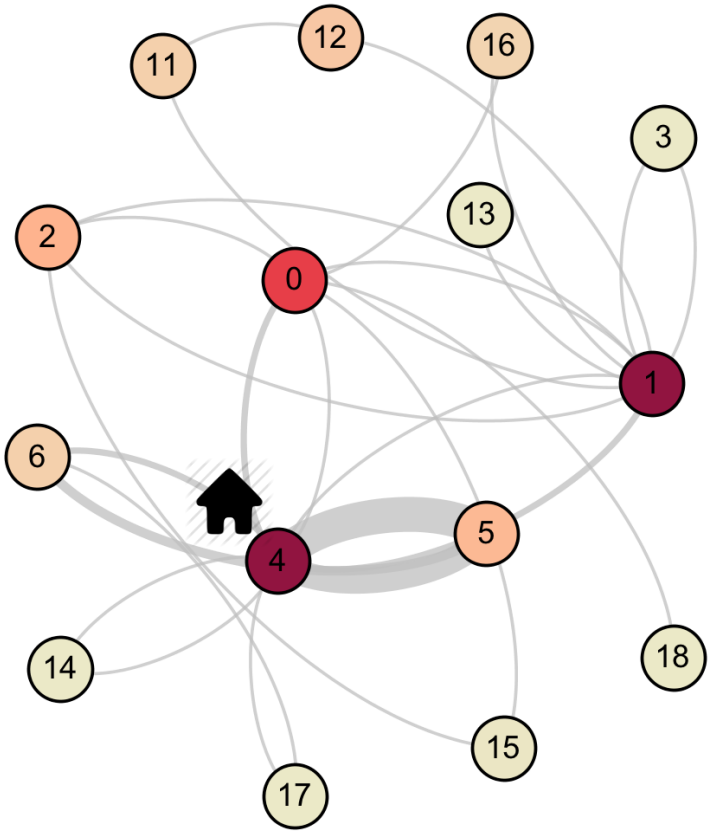




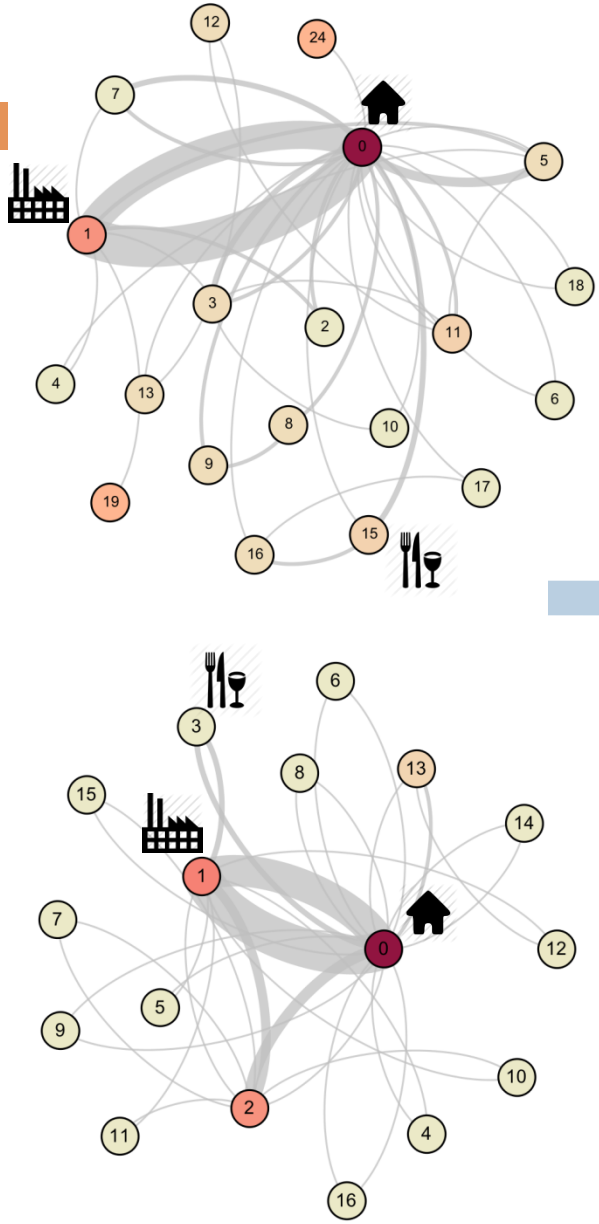
The ABC classifier

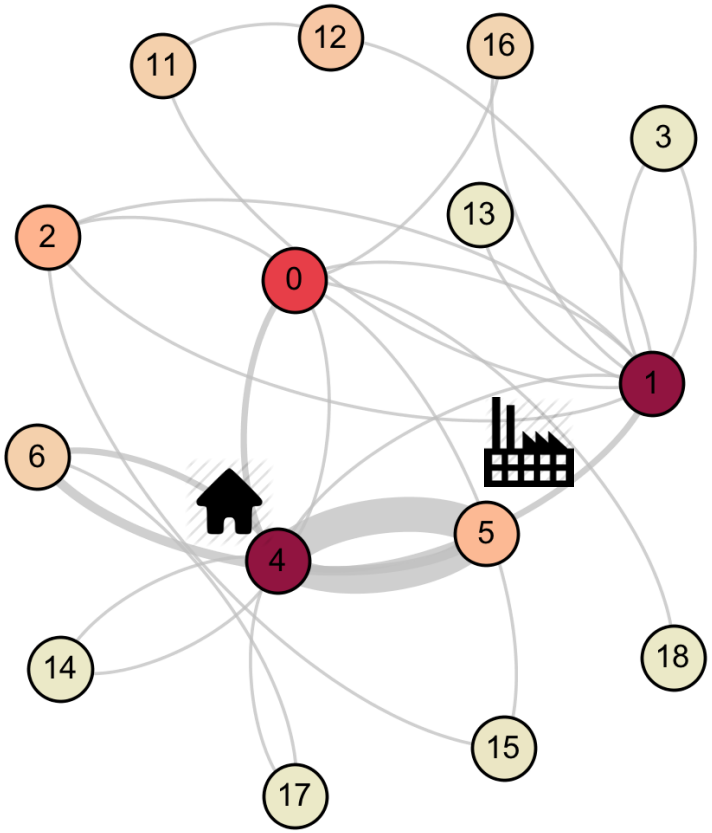




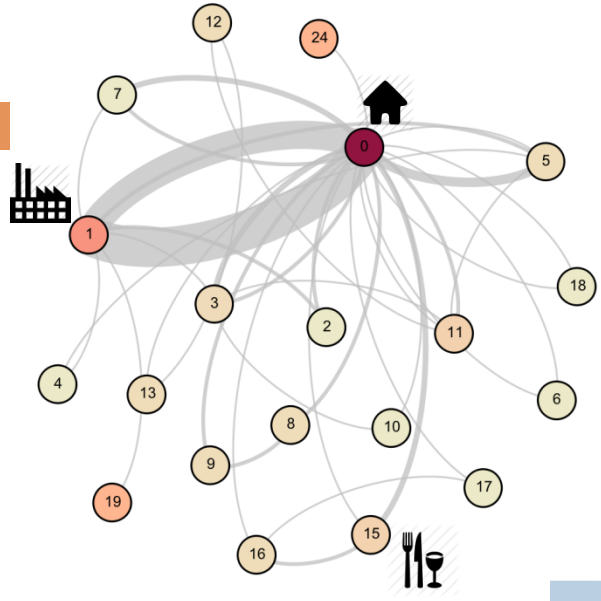


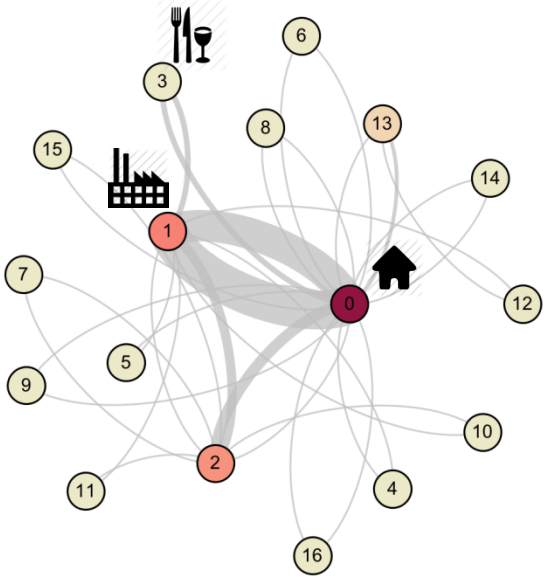
The ABC classifier

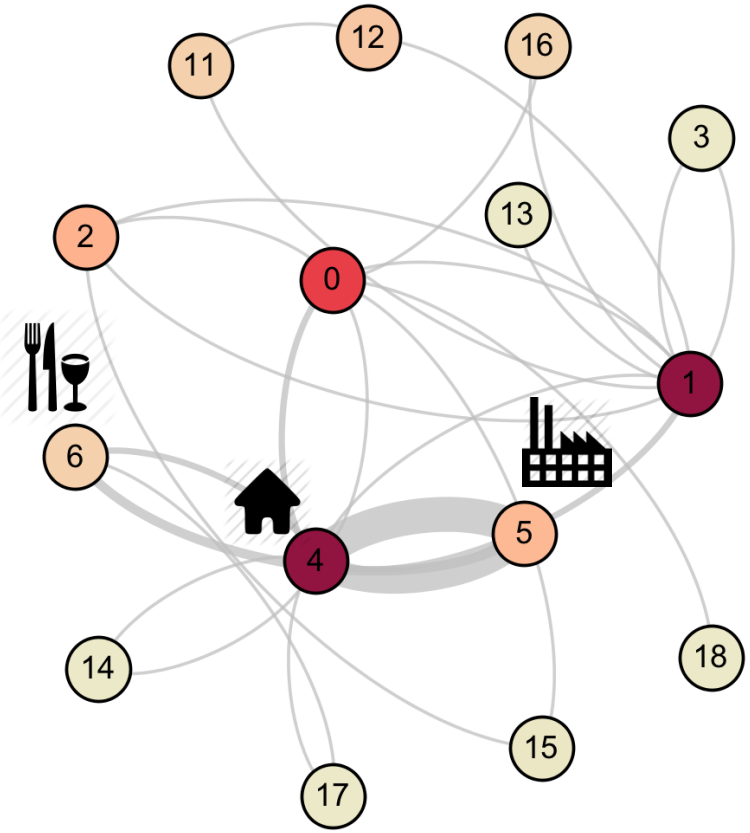




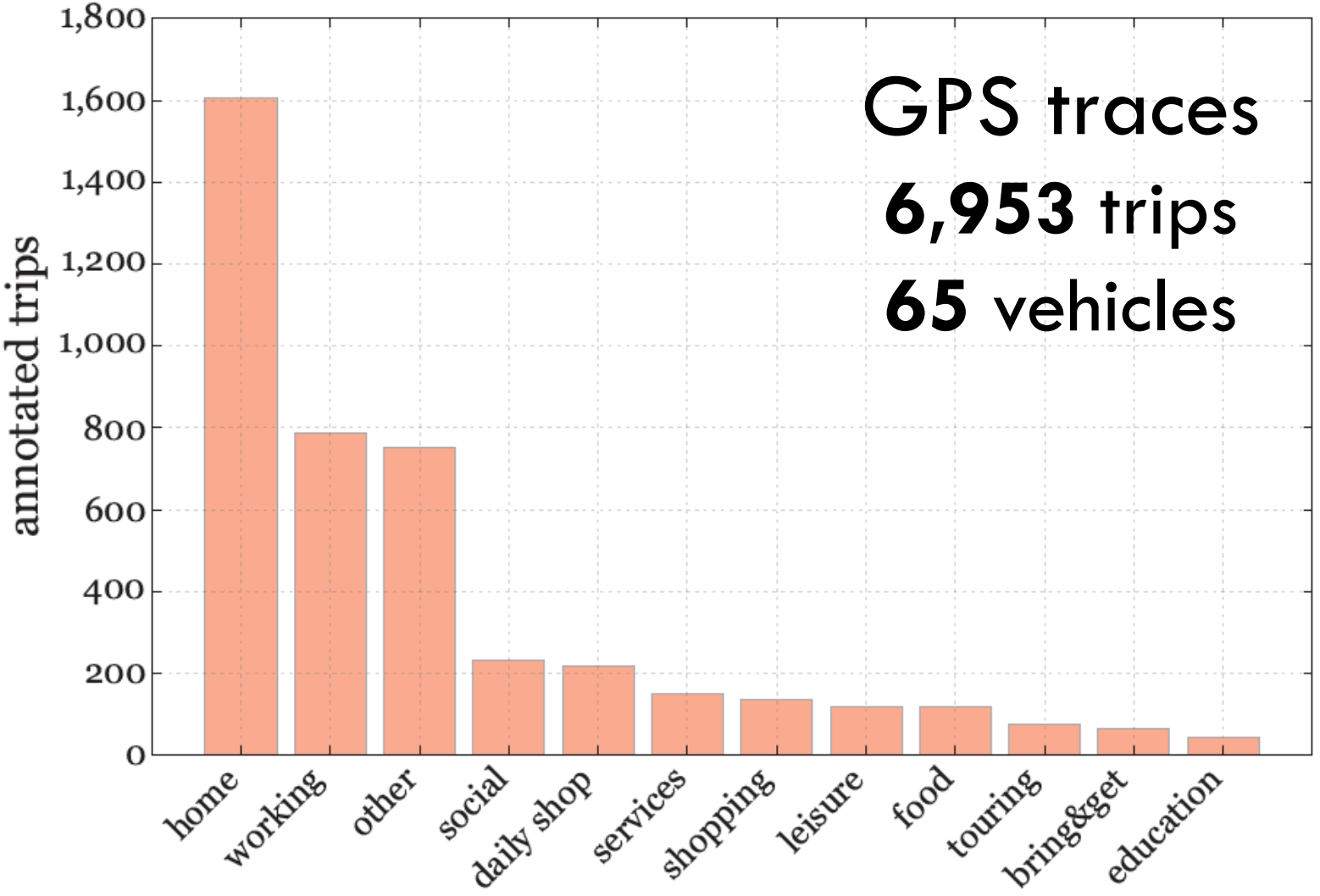
The ABC classifier





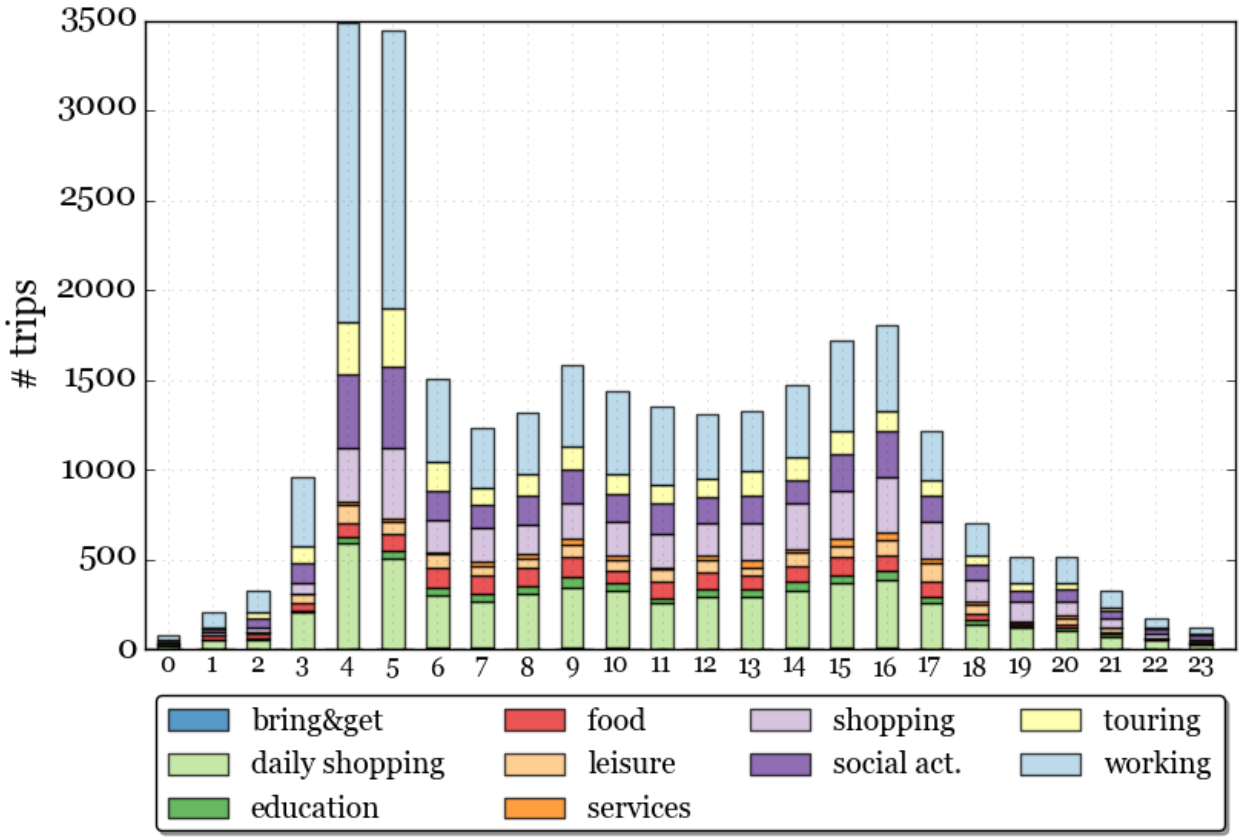


Experiments



Semantic Mobility Analytics

Temporal Analysis



CASE STUDIES

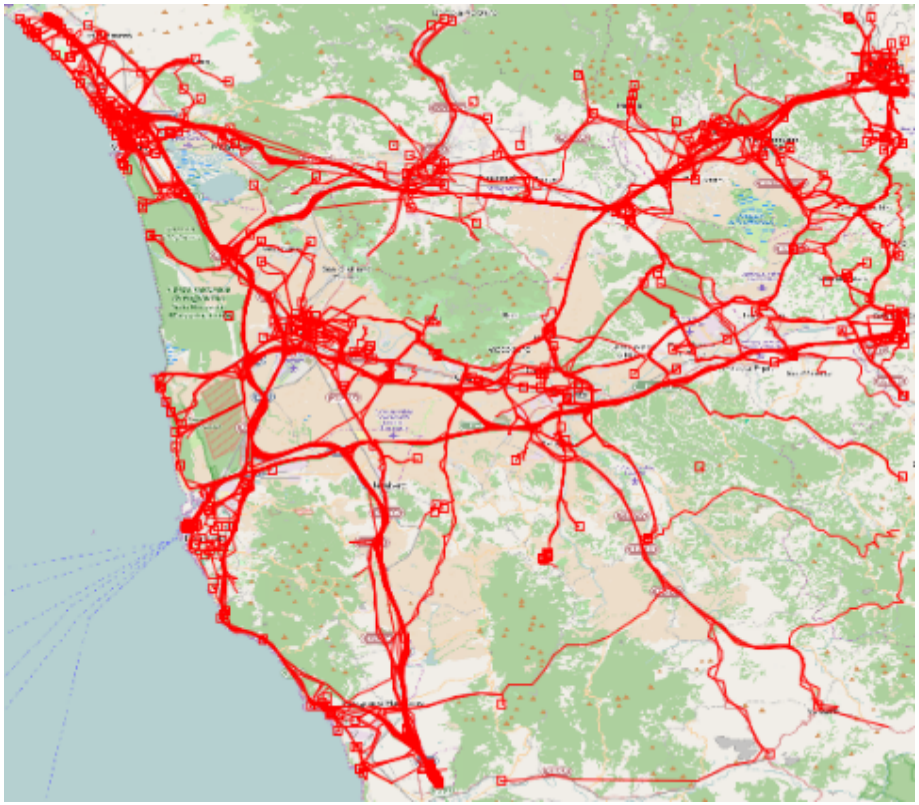
- TOWARDS CORPORATE USERS
  - Geomarketing
  - Monitoring Driving-based Segmentation
- TOWARDS INDIVIDUAL USERS
  - Self-awareness
- TOWARDS PUBLIC SECTOR USERS
  - Urban Mobility Atlas – building territory indicators
  - Studying Actractors' Impact

Services Towards Corporate Users

Geomarketing

Problem definition

Based on the trajectories of a sample of population, what is the best place to open a new shop / mall ?



The “best” place

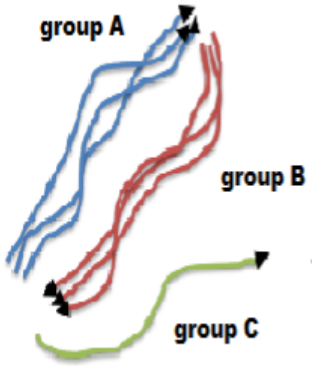
Experts' knowledge: best place to open a mall is where people pass during everyday activities

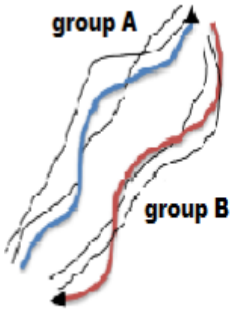
Area crossed by road segments with a high frequency of systematic travels of people

Systematic movements

Step 1: Map-matching

- See users' movements as sequences of road segments.

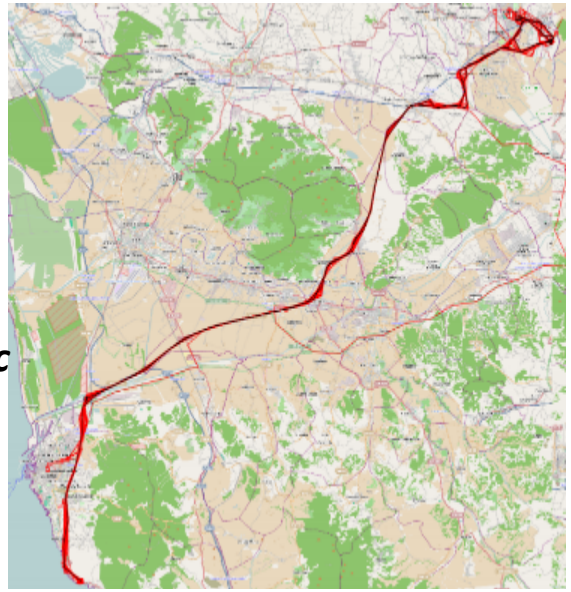




Step 2: Mobility profiles

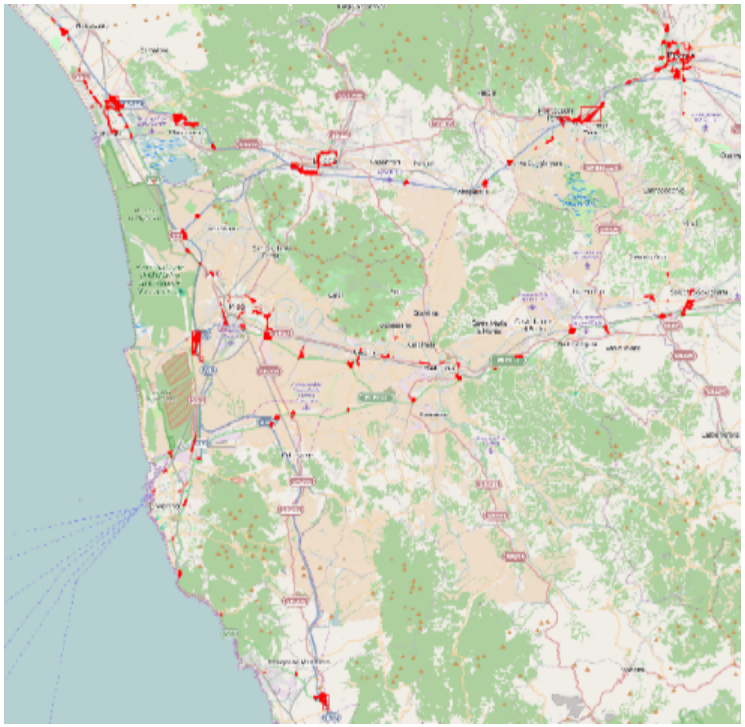
- Select only systematic movements.

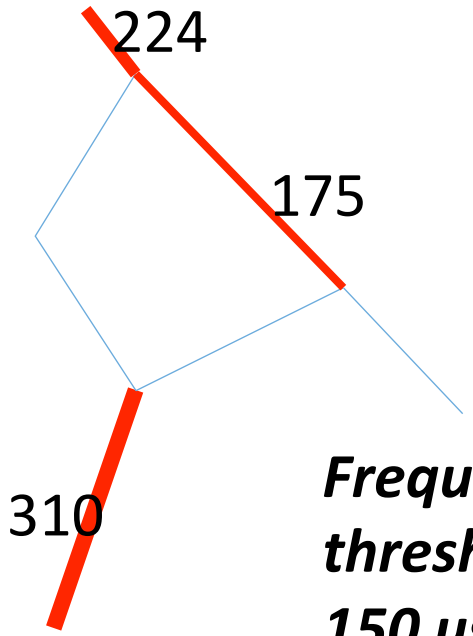
User's systematic movement:
L1 → L2



Frequently visited road segments

- Aggregate systematic movements by road segments
- Set a threshold to select the frequent ones



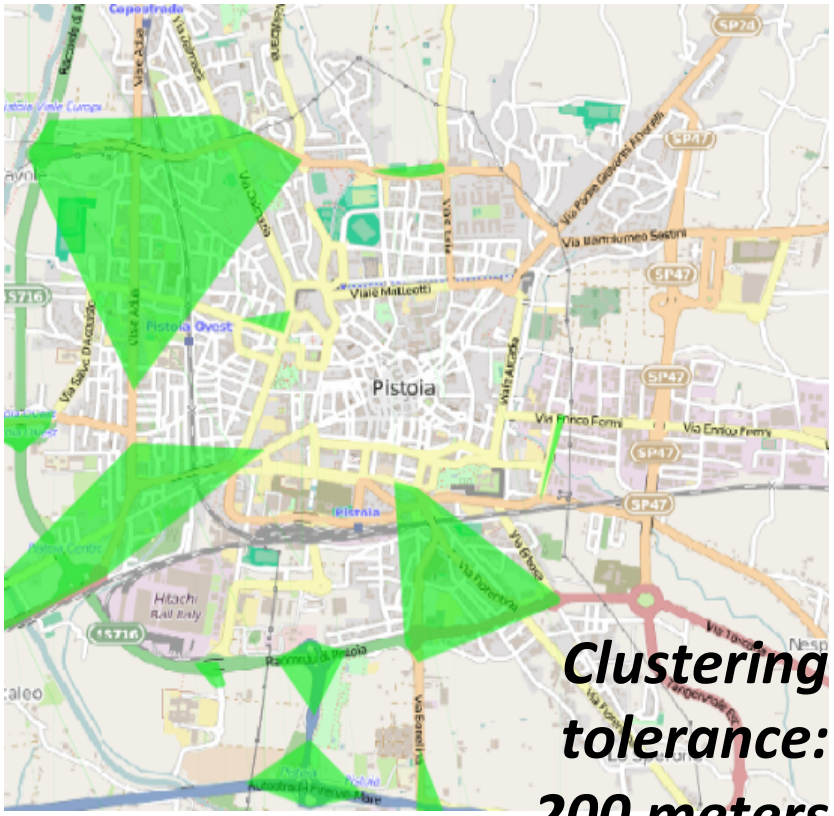


Frequency threshold: 150 users

Candidate areas for a mall

Using a spatial clustering we can extract cluster of frequent road segments which are spatially close each other.

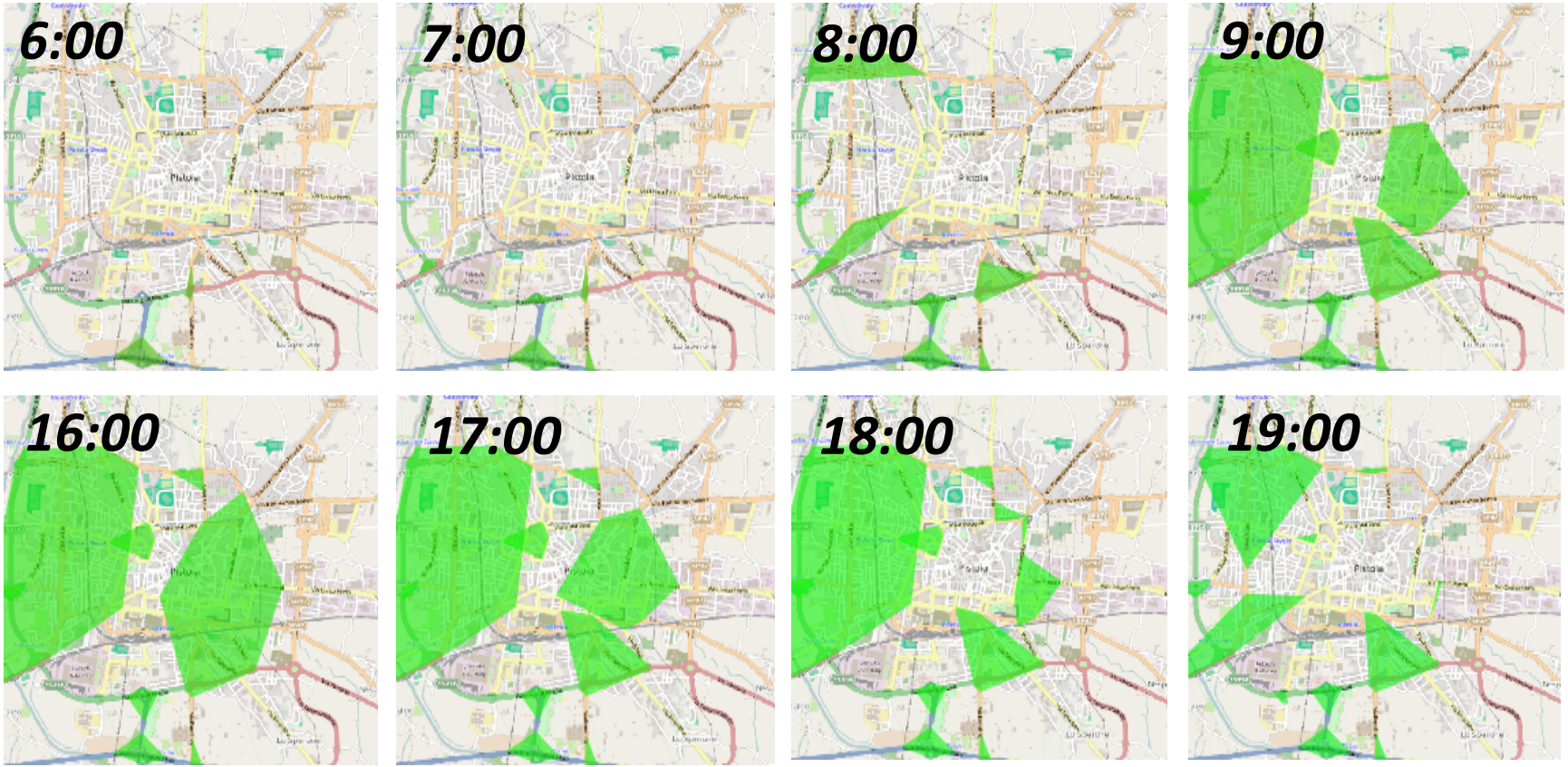
- Distance of 2 segments
  - Compare vertices
- Draw clusters as convex hull



Clustering tolerance: 200 meters

Temporal evolution

Repeat this process for each hour of the day and analyze how they evolve



Monitoring Driving-based Segmentation

Services Towards Corporate Users

Scenario context & motivation

- **Customer segmentation**: a marketing strategy that involves dividing a broad target market into subsets of consumers who have common needs
- **Needs**: car insurance companies would like to define customer segments that capture different driving profiles
  - Each segment could then be offered suitable contract conditions

http://en.wikipedia.org/wiki/Customer_segmentation

- **Opportunities**: the vehicles insured by some companies have on-board GPS devices that can trace their movements
  - They could aggregate such traces into driving habit indicators based on recent history for the driver and transmit them

Scenario description

- Driving indicators
  - **Each vehicle** continuously keeps track of recent movements, compute aggregate indicators and sends them to controller
- Profile extraction
  - **The controller** uses initial indicator values to build clusters of drivers, each corresponding to a "driving profile"
- Profile monitoring
  - **The controller** continuously checks updates to verify that the driving profiles extracted are still good enough

Step 1: Features for individual mobility behaviors

- Indicators for recent mobility behaviors
- Computed over recent history → sliding window

- Include information derivable from standard GPS devices

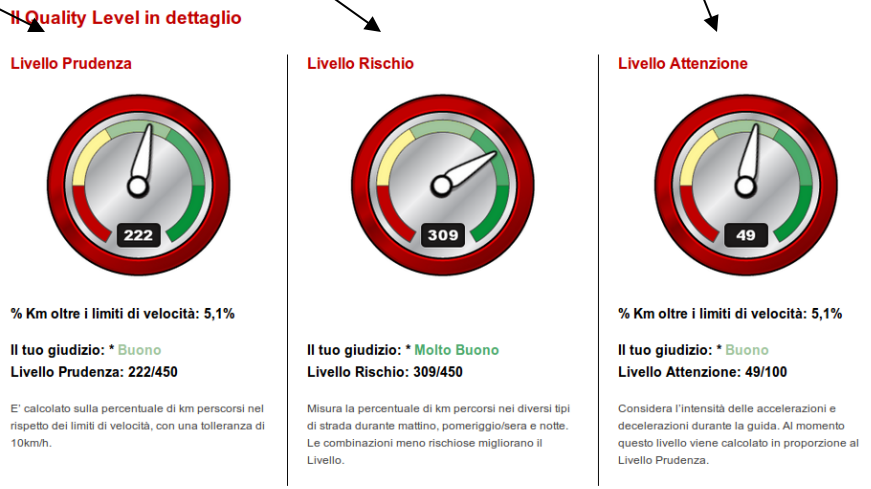
Step 1: Features for individual mobility behaviors

- Which features?
  - Superset of those currently used by insurance companies

How fast I drive w.r.t. speed limits

Where I drive w.r.t. road categories

How dynamic I drive w.r.t. acc-/ decelerations



Features over sliding window

- Length = traveled distance
- Duration = time spent driving
- Count = number of trips
- Phighway = % km on highways
- Pcity = % km inside cities
- Length_arc_crowded = km on 20% most crowded roads
- Pnight = % km in night time
- Pover = % km over speed limit
- Profile = % of km on systematic trips
- Radius_g = radius of gyration
- Radius_g_L1 = radius of gyration w.r.t. L1
- Avg_Dist_L1 = average distance from L1
- TimeL1L2 = % time spent on L1 and L2
- EntropyArc = entropy on road segment frequencies
- EntropyLocation = entropy on location frequencies
- EntropyTime = entropy on hours of the day

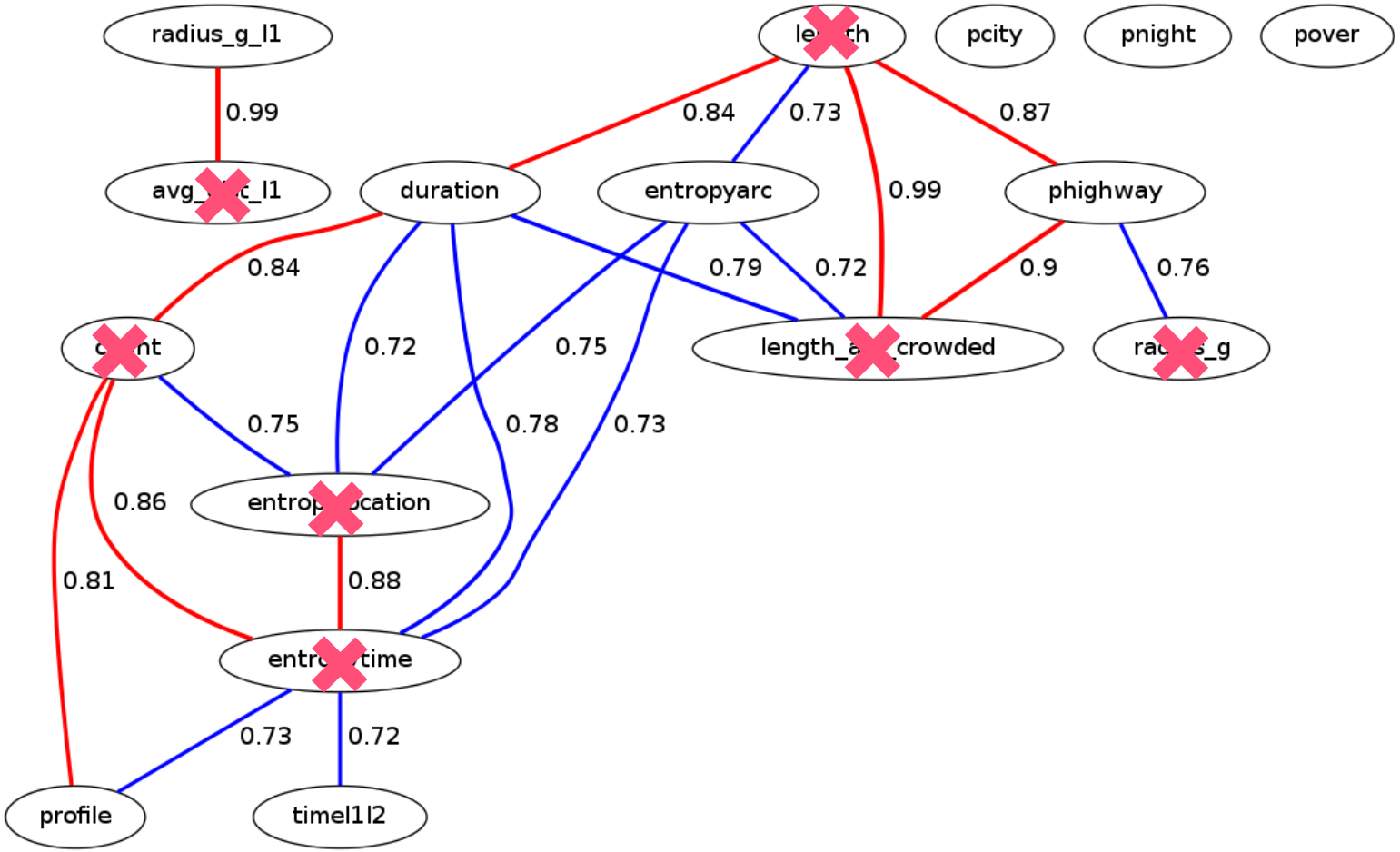
Basic aggregates

Aggregates on spatia
temporal selection

Count of events

Spatial/Temporal distr

Correlation analysis



Features over sliding window

- Length = traveled distance
- Duration = time spent driving
- Count = number of trips
- Phighway = % km on highways
- Pcity = % km inside cities
- Length_arc_crowded = km on 20% most crowded roads
- Pnight = % km in night time
- Pover = % km over speed limit
- Profile = % of km on systematic trips
- Radius_g = radius of gyration
- Radius_g_L1 = radius of gyration w.r.t. L1
- Avg_Dist_L1 = average distance from L1
- TimeL1L2 = % time spent on L1 and L2
- EntropyArc = entropy on road segment frequencies
- EntropyLocation = entropy on location frequencies
- EntropyTime = entropy on hours of the day

Basic aggregates

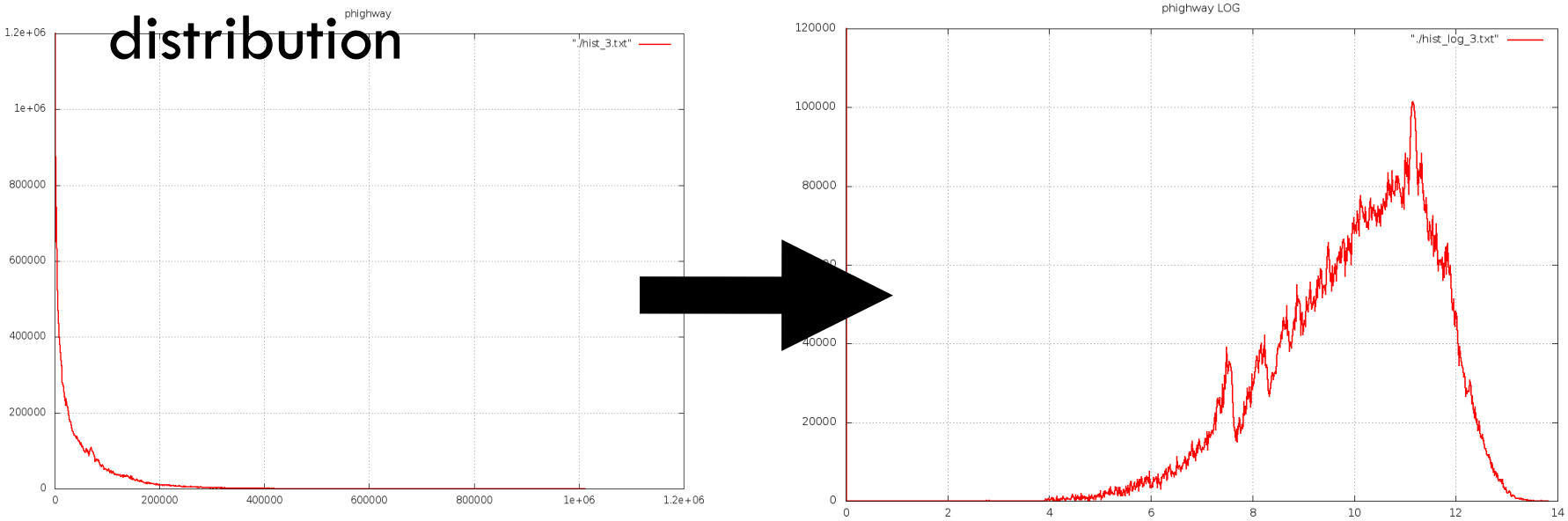
Aggregates on spatia
temporal selection

Count of events

Spatial/Temporal distr

Features normalization

- Log transformation for features with skewed distribution



- Z-score normalization for all features

(2) Compute driving profiles

- Clustering-based definition
  - Profile = representative set of indicators for a large group of drivers with similar behaviors (i.e. similar indicator values)
- Clustering method
  - **K-means** – a partitional, center-based clustering algorithm
  - **Euclidean distance** over driving indicators
  - Refinements: Iterated K-means & select best solution + Noise removal
- Profile = average point of each cluster

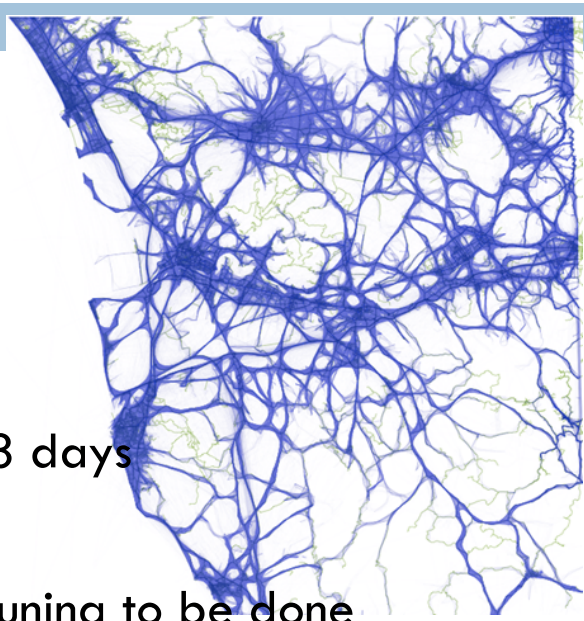
Cluster refinement

- Iterated K-means
  - Run clustering multiple times (→ initial random seeding)
  - Select output with best quality
    - Based on clusters compactness (→ SSE – see definition later)
- Noise removal
  - Performed at postprocessing
  - From each cluster, remove points *p* such that

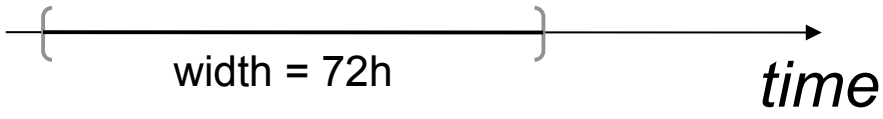
$$d(p,c) > 2 \; median \; \{ d(x,c) \mid x \; in \; cluster \}$$

    where c is the cluster center
  - Alternative solutions are possible
    - e.g.: density-based noise removal

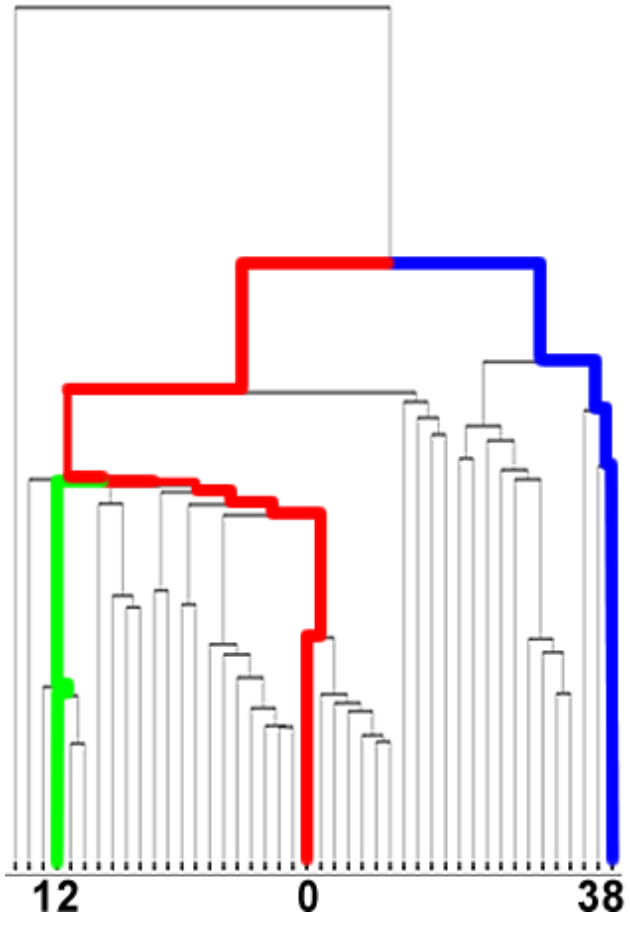
Experimental setting

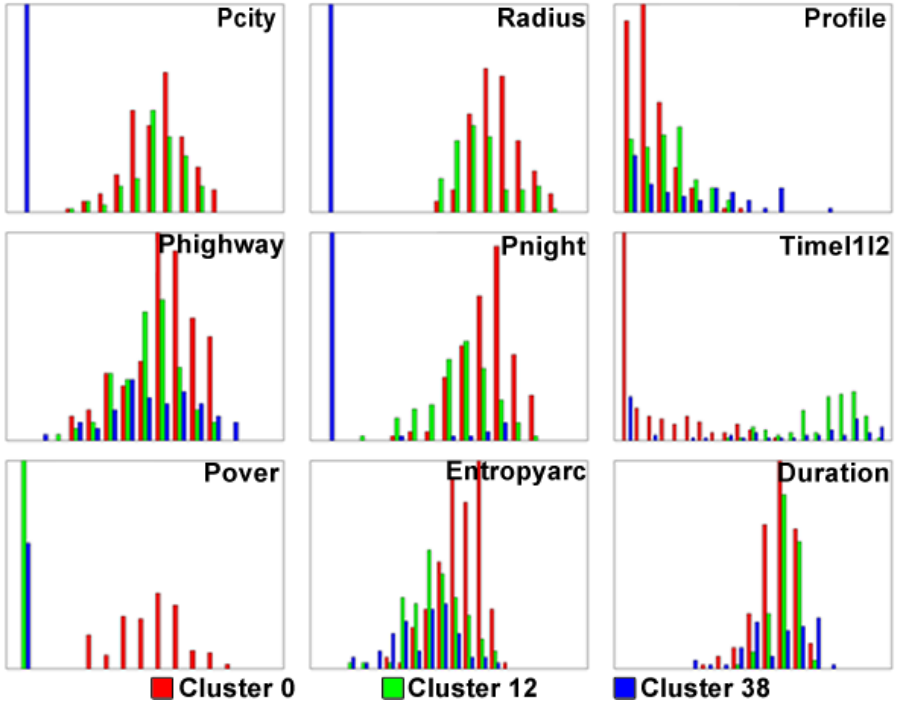


- GSP traces of an insurance company customers
  - 35 days monitoring
- Sample of ~11k vehicles moving in the area
- Short temporal thresholds for testing purposes
  - Compute driving indicators over a sliding window of 3 days
  - Update indicators every 15'
  - Most likely larger in a real application – parameter tuning to be done with domain experts



Experiments: clusters inspection





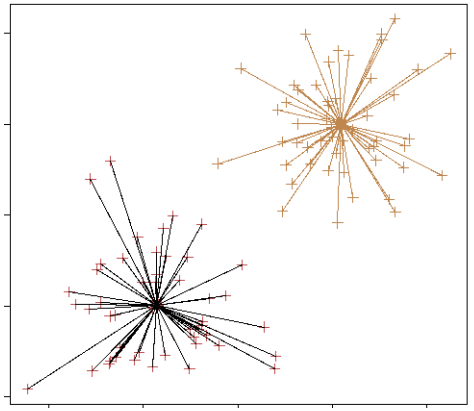
Explorers

Long-range commuters

Sunday drivers

(3) Driving profiles monitoring

- Translated to “cluster quality monitoring”
- Quality measure: SSE = Sum of Squared Errors
  - Given a clustering $C = \{ C_1, \ldots , C_k \}$, and average points m_i for each cluster C_i



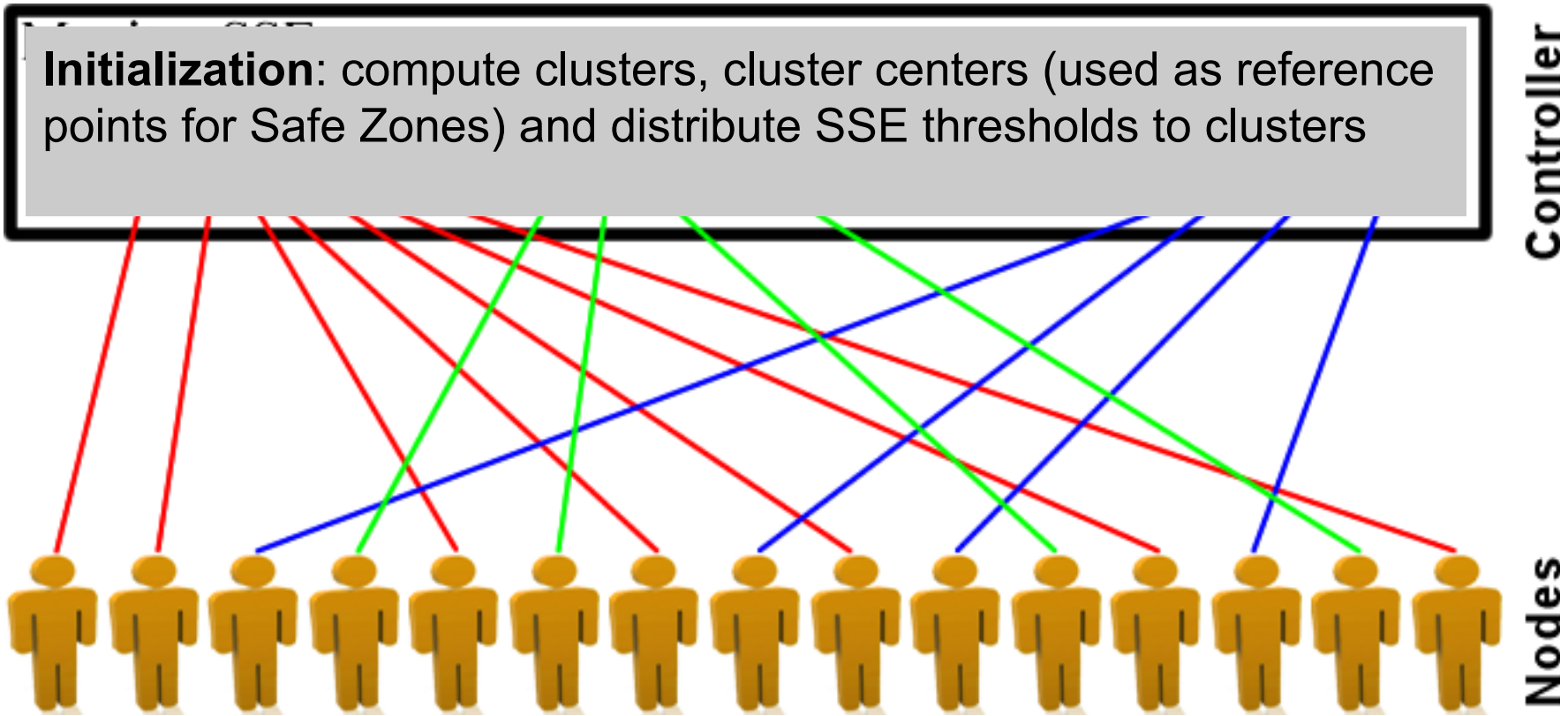
(3) Driving profiles monitoring

DEFINITION 1 (CLUSTER MONITORING PROBLEM). *Given a clustering* $C = \{C_1, \ldots, C_k\}$ *having initial SSE equal to* SSE_0, *and given a tolerance* $\alpha \in \mathcal{R}^+$, *we require to ensure that at each time instant* t *the following holds for the SSE of the (dynamic) dataset* D_t*:*

$$SSE_t \leq (1 + \alpha) SSE_0$$

When that does not happen, a recomputation/update of cluster assignments should be performed.

Monitoring process



Self-awareness

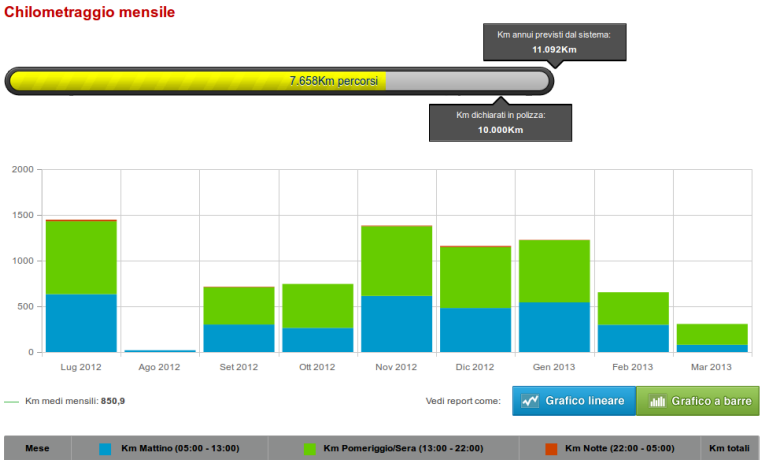
Services Towards Individual Users

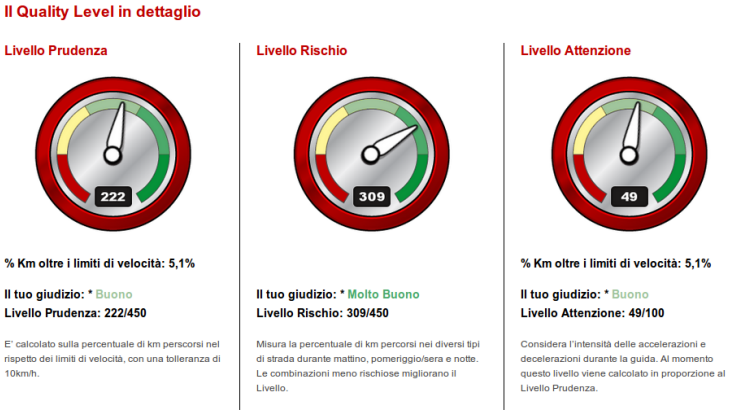
Self-awareness services

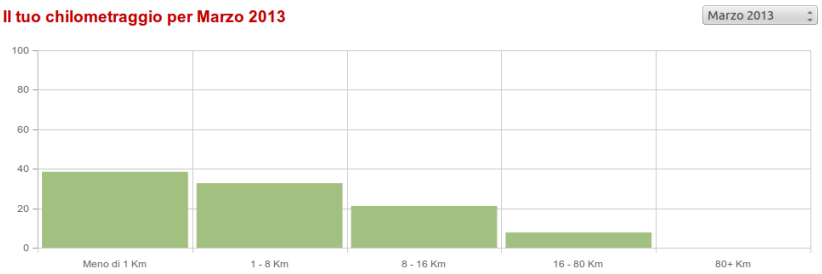
- Mobility-based specialization of self-awareness services for generic users
  - Provide summary of activity of the user
  - Provide comparison against collectivity
- Summaries based on
  - Temporal statistics
  - Spatial statistics / distributions
  - Movement aggregates

User's activity summaries

- An example within Generali

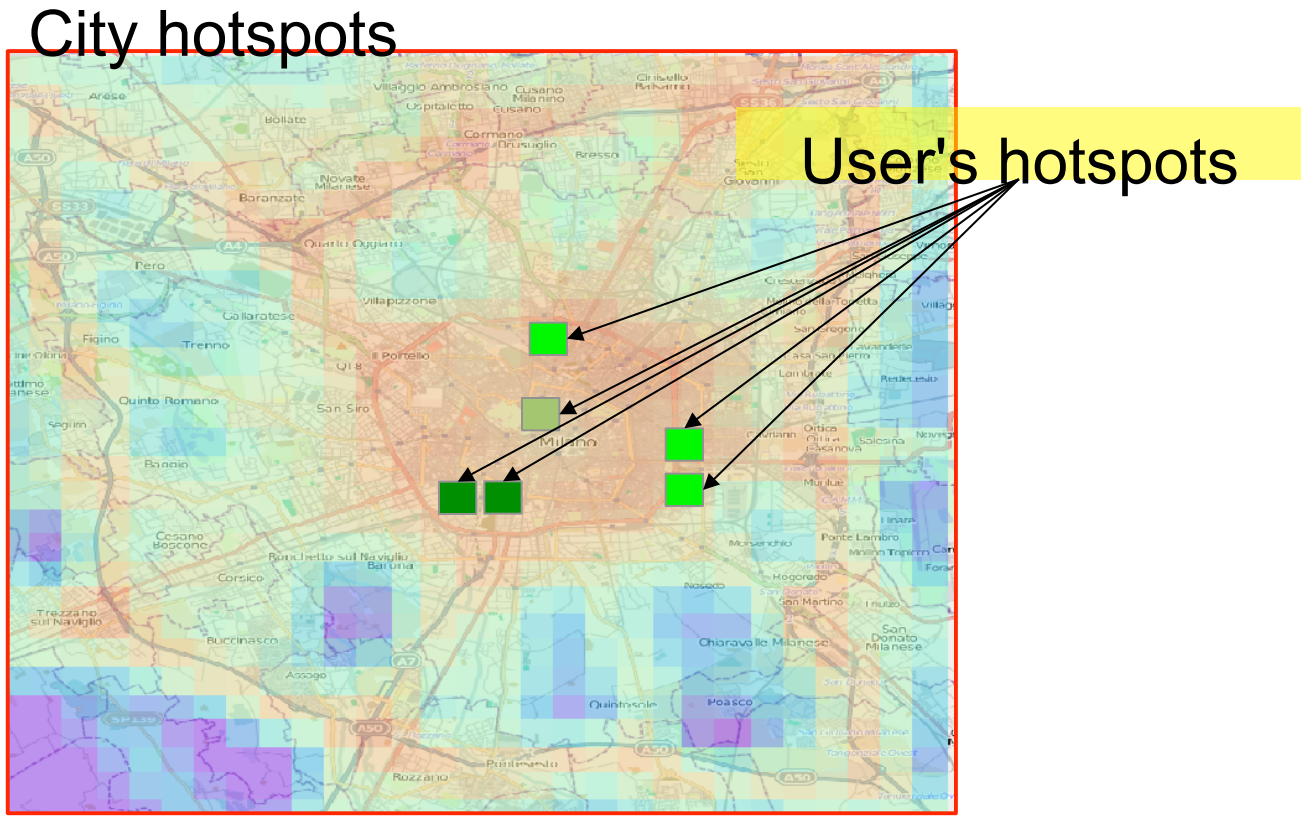






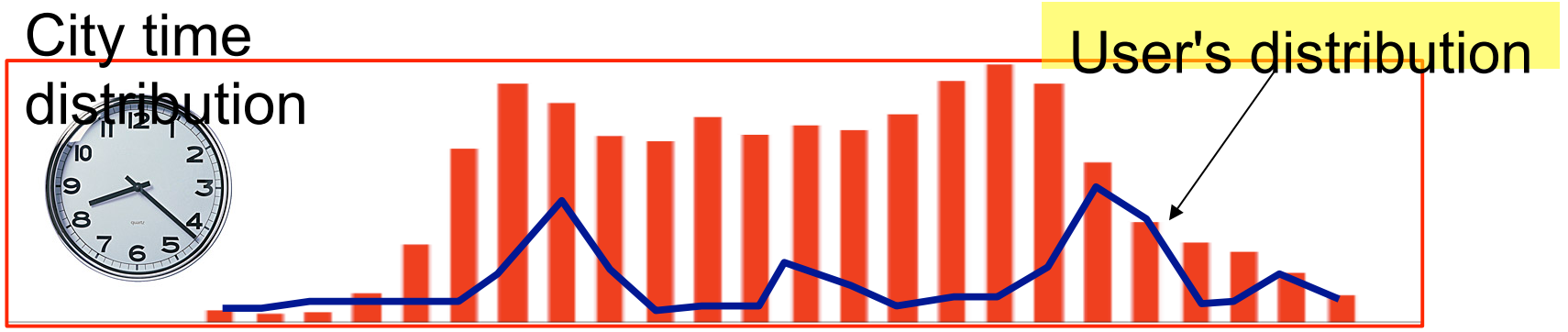
Comparison against collectivity

- In space



Comparison against collectivity

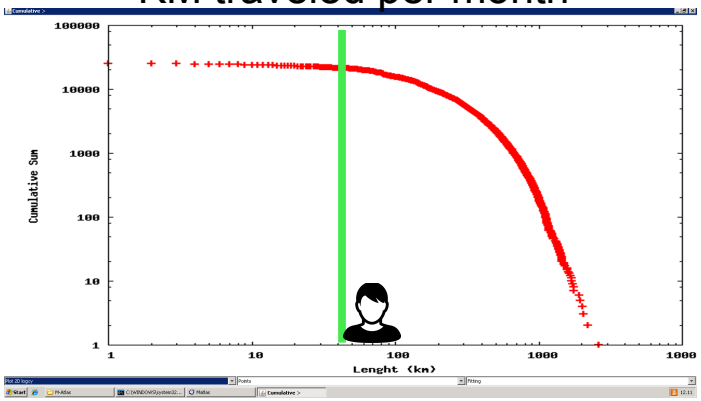
- In time



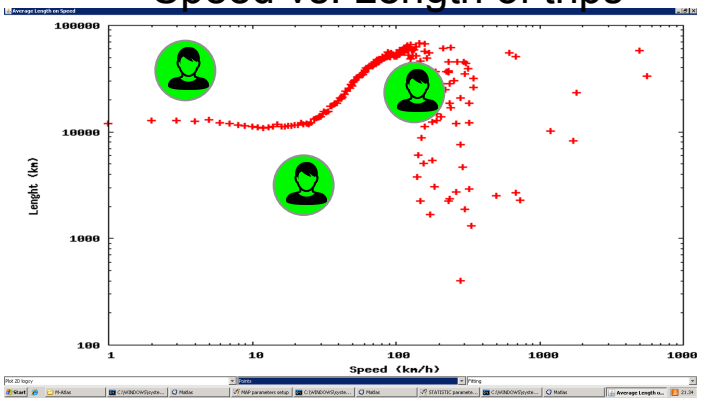
Comparison against collectivity

- On general statistics

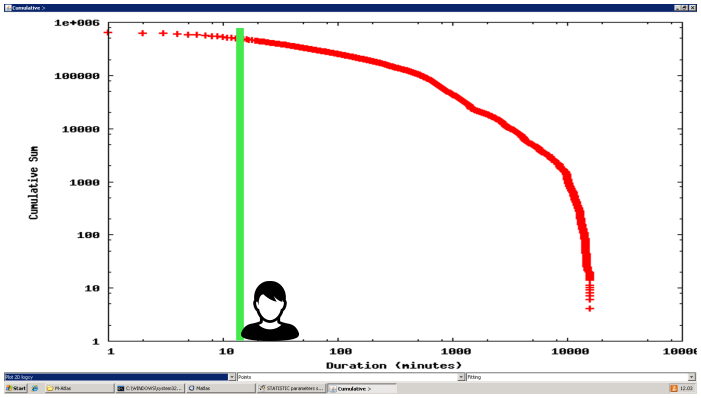
KM traveled per month



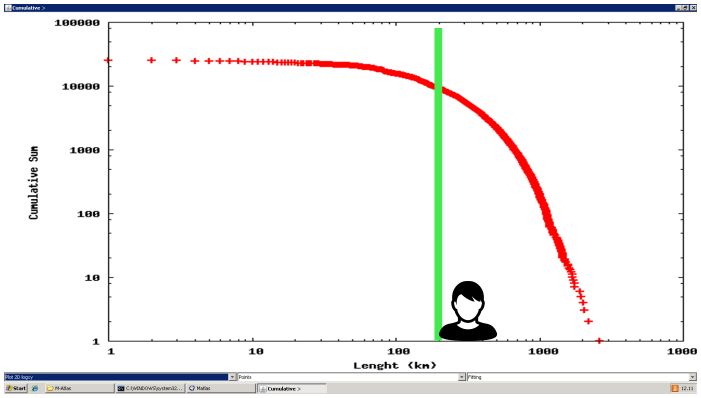
Speed vs. Length of trips



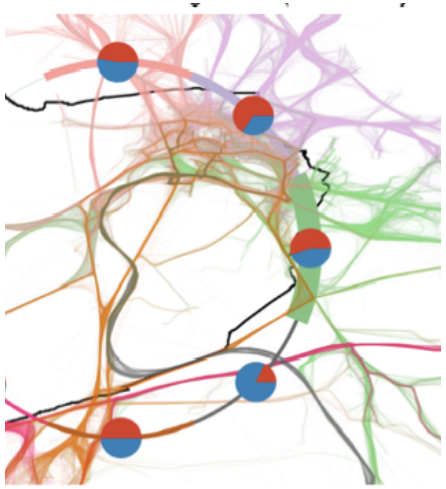
Total duration of travels



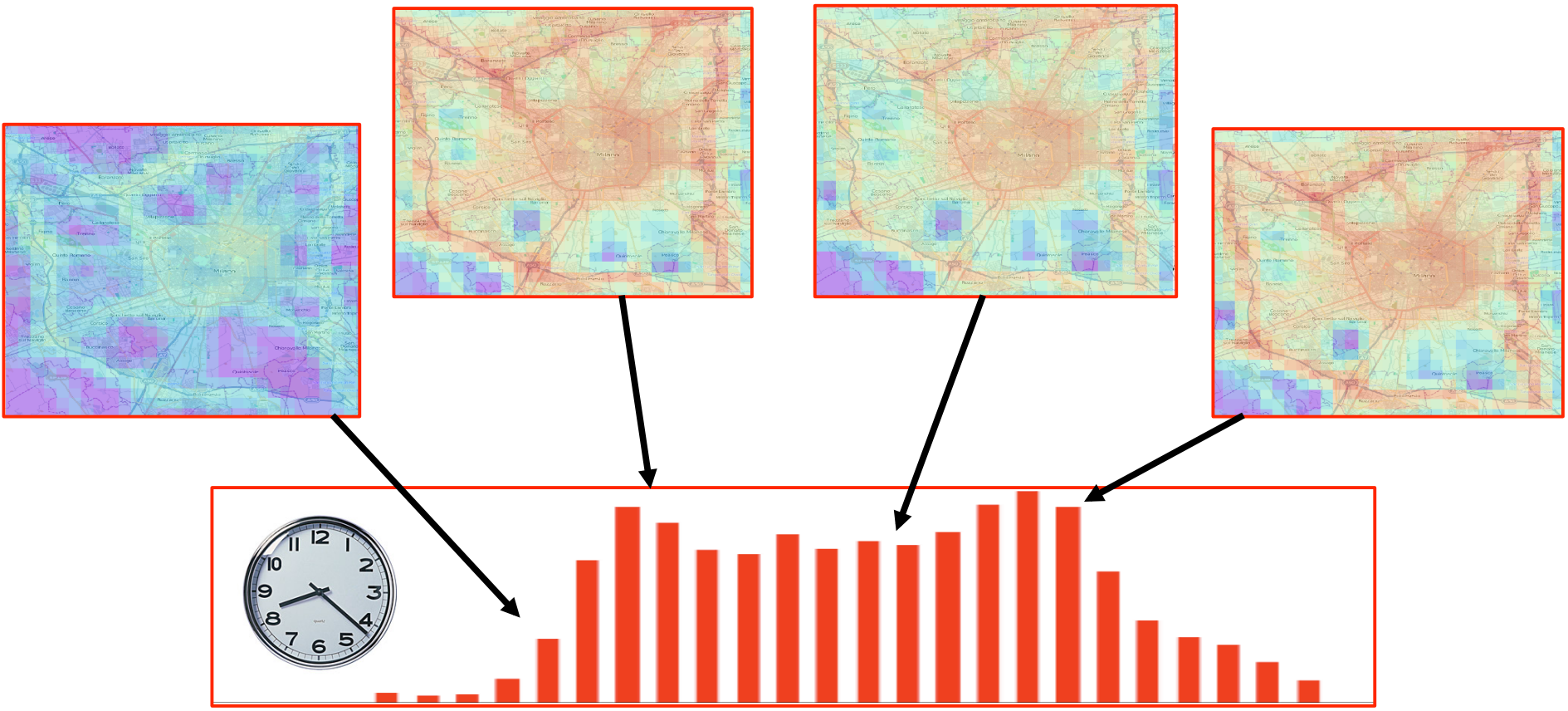
Radius of gyration



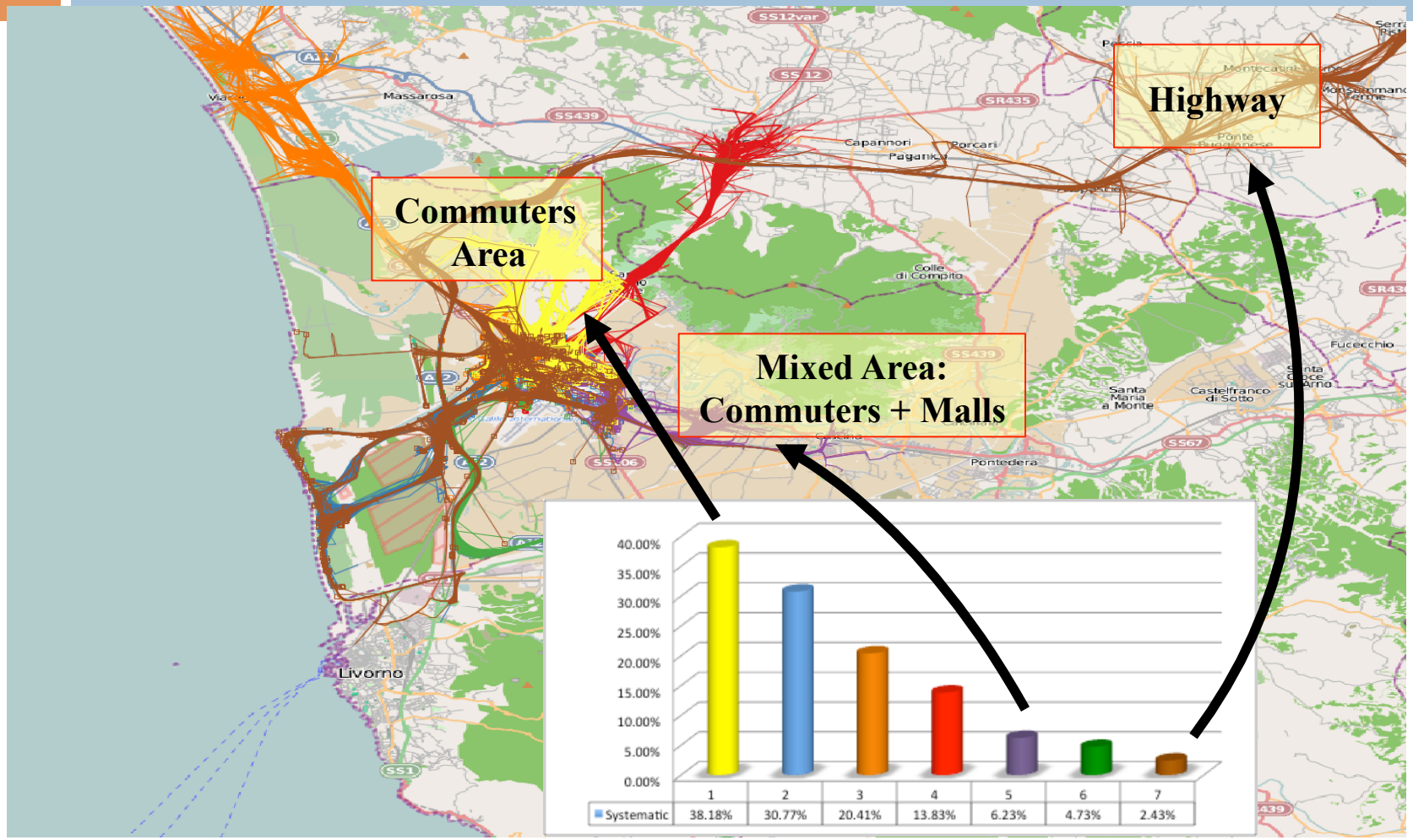
- Services Towards Public Sector
  - *Urban Mobility Atlas*



Dynamics of urban mobility



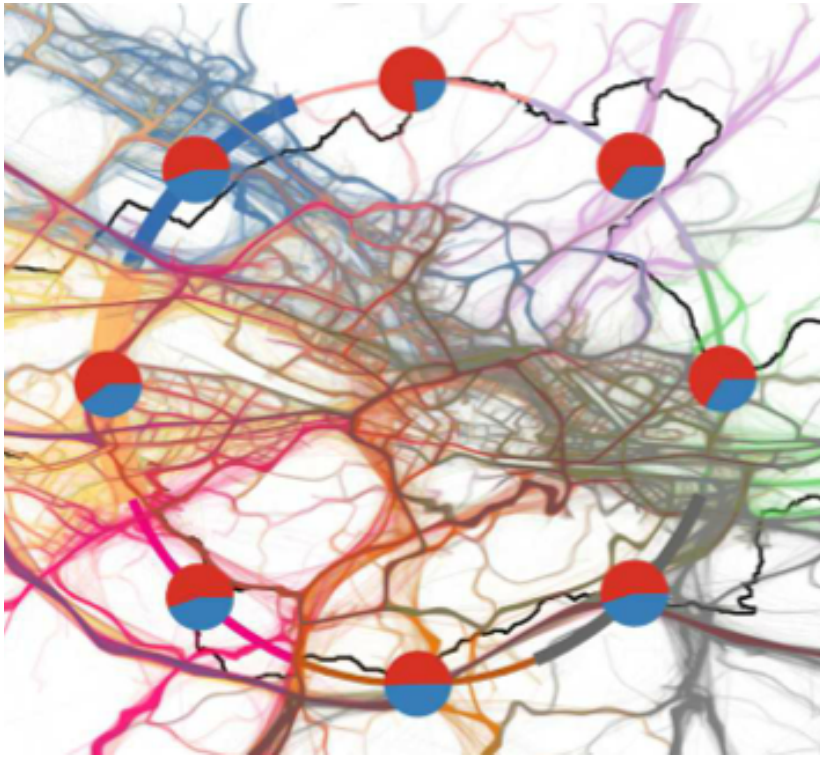
Impact of Systematic Mobility



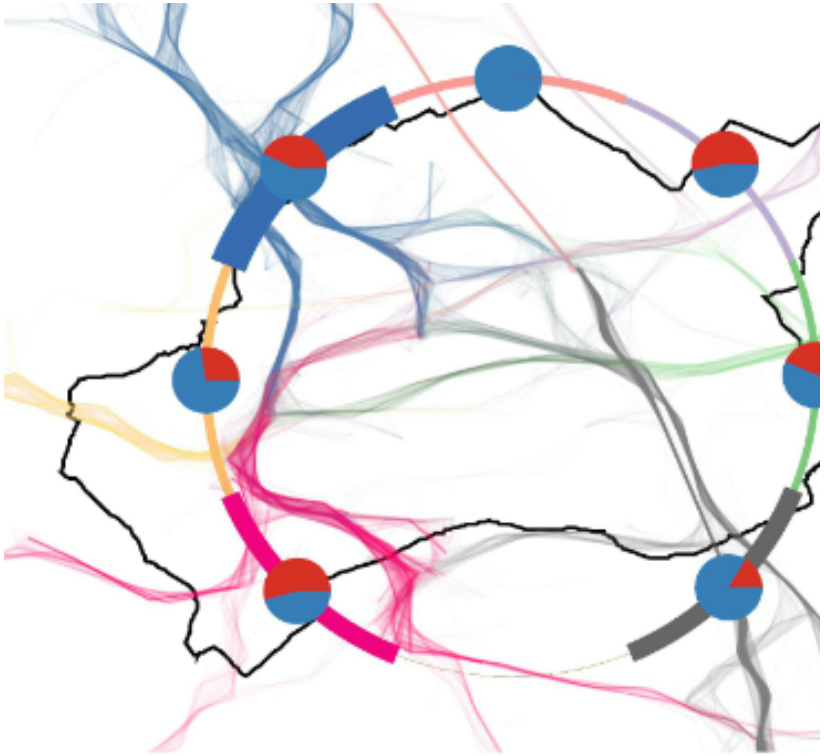
Access Routes
Systematic Mobility (%)

Comparing Cities

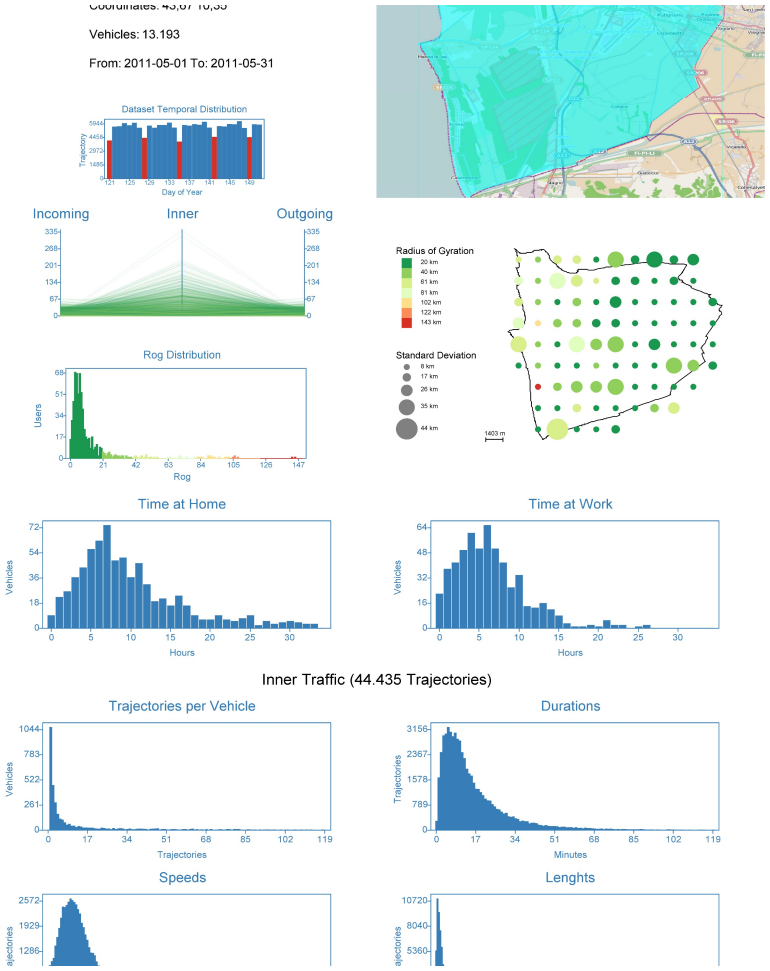
Florence

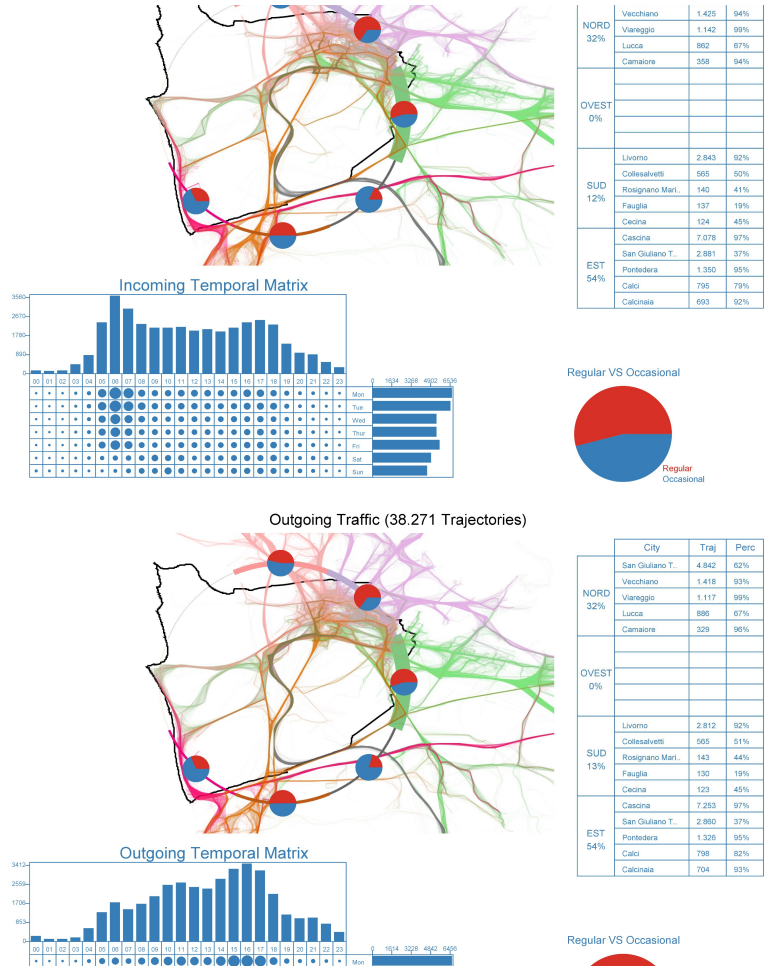


Montepulciano



Atlas of Urban Mobility

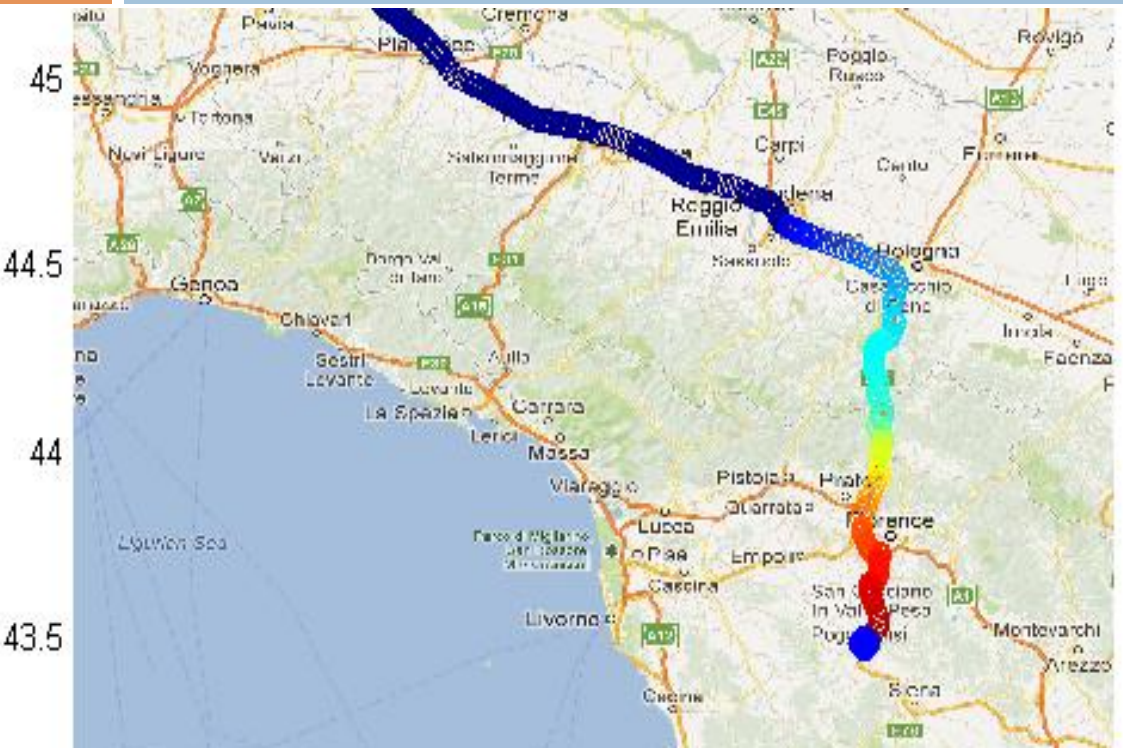




	City	Traj	Perc
NORD 32%	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	95%
	Calci	795	79%
	Calcinaia	693	92%

	City	Traj	Perc
NORD 32%	San Giuliano T..	4.842	62%
	Vecchiano	1.418	93%
	Viareggio	1.117	99%
	Lucca	886	67%
	Camaiore	329	96%
OVEST 0%			
SUD 13%	Livorno	2.812	92%
	Collesalvetti	565	51%
	Rosignano Mari..	143	44%
	Fauglia	130	19%
	Cecina	123	45%
EST 54%	Cascina	7.253	97%
	San Giuliano T..	2.860	37%
	Pontedera	1.326	95%
	Calci	798	82%
	Calcinaia	704	93%

Electrifiability



In Pisa 90.5% of daily trips are electrifiable:
562.061 km electrificable

In Tuscany at least 38% of users have a daily mobility covered at 100% by an electrical vehicle (home to home)

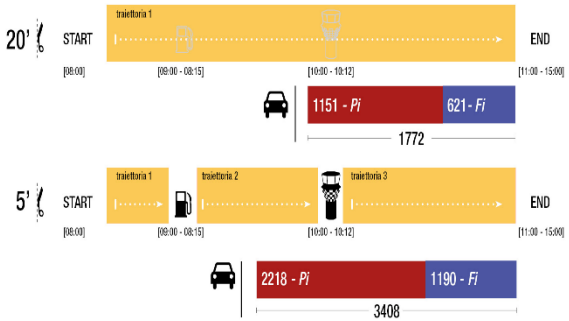
In Florence we notice that in week ends the % of electrifiable trips decreases since people travel further from home

Studying the Airports attraction

Understand how and how much Pisa and Florence airports attract Tuscany residents

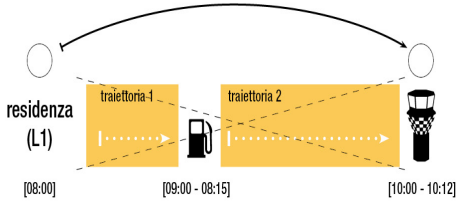
General analysis appraoch

- 1. From raw GPS point to trajectories

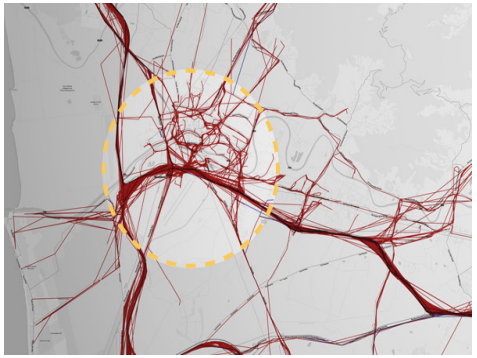


2. Geographic identification of the area of interest and corresponding parking areas

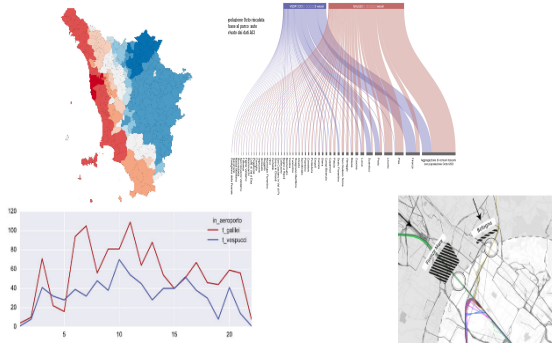
3. Identification of residents: First most frequent location



4. Selection of data



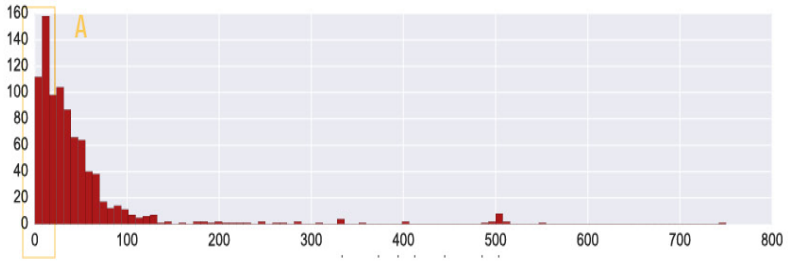
5. Knowledge extraction, Study of the flows and Semantic enrichment

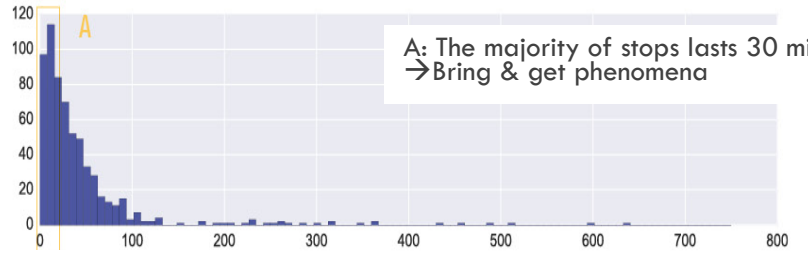


Temporal statistics of the stops at the Airports

Pisa Airport | Florence Airport

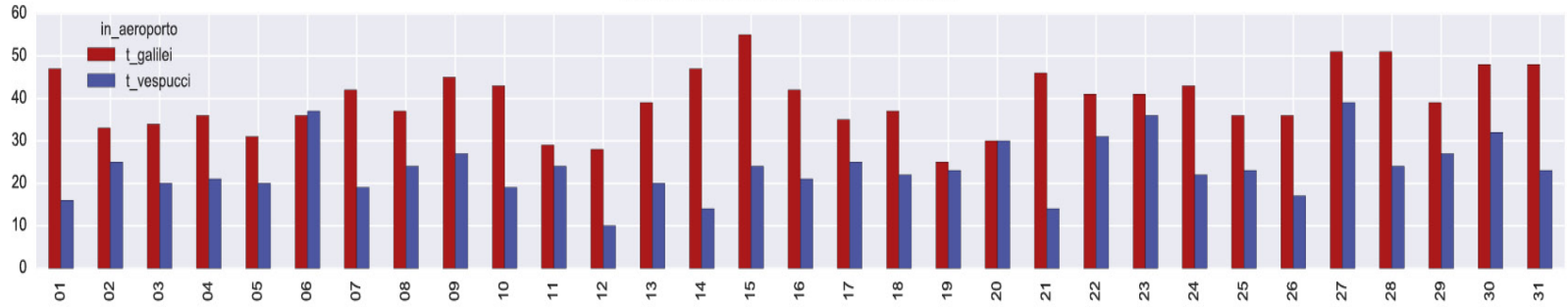
Distribution of the durations of the stops



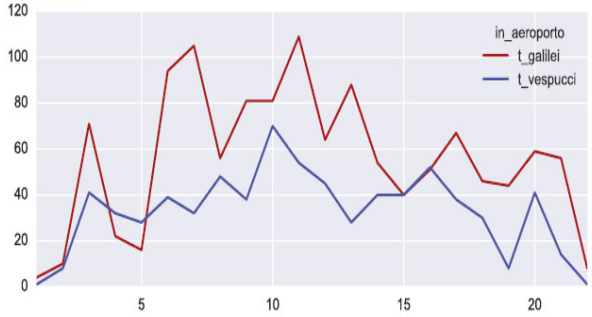


A: The majority of stops lasts 30 minutes
→Bring & get phenomena

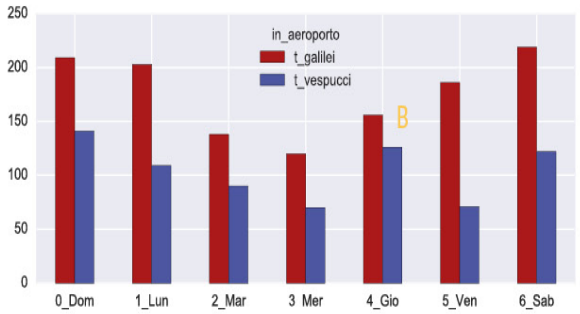
Number of stops per day



Number of stops per hour in a day

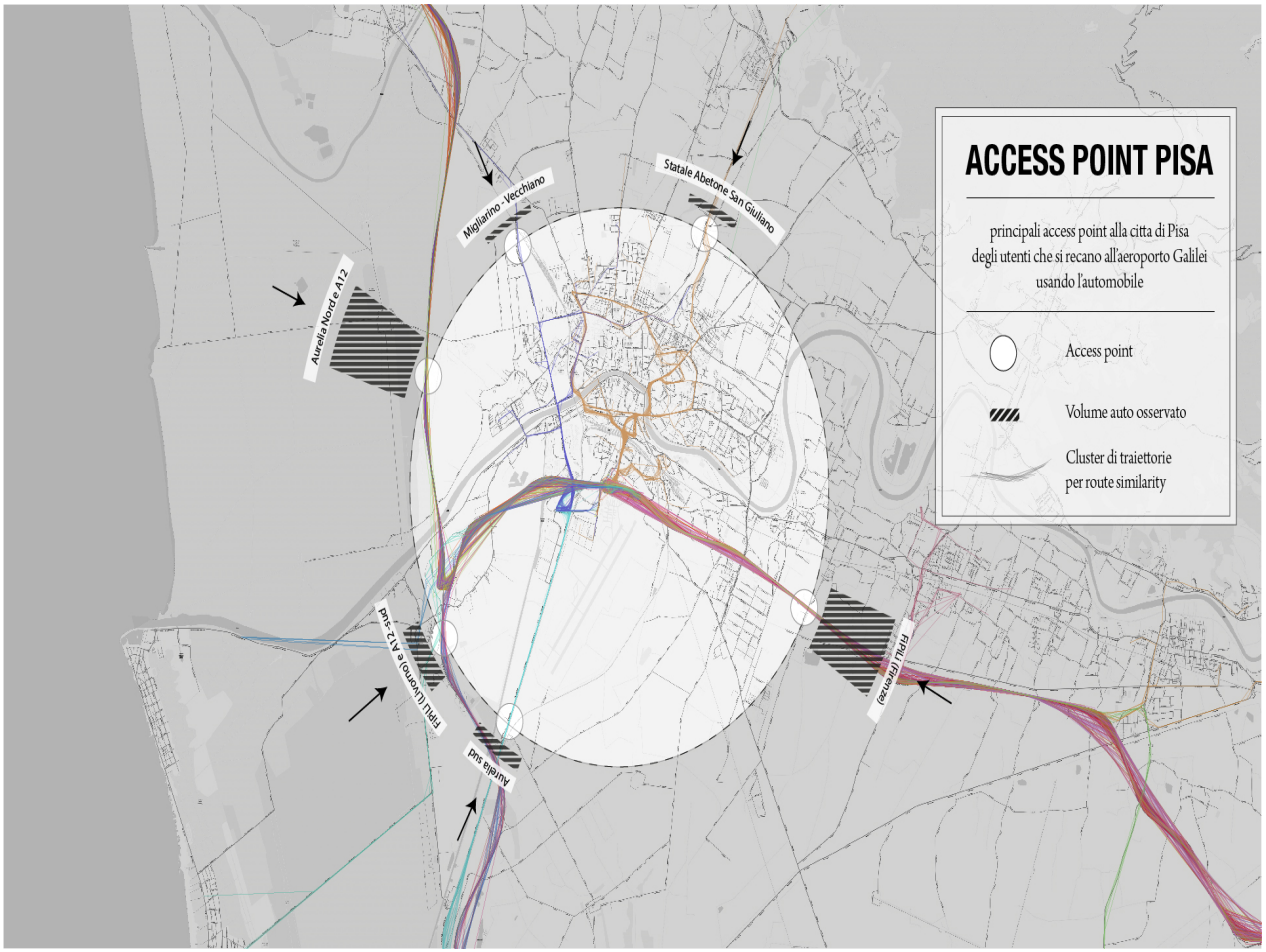


Number of stops per day in the week



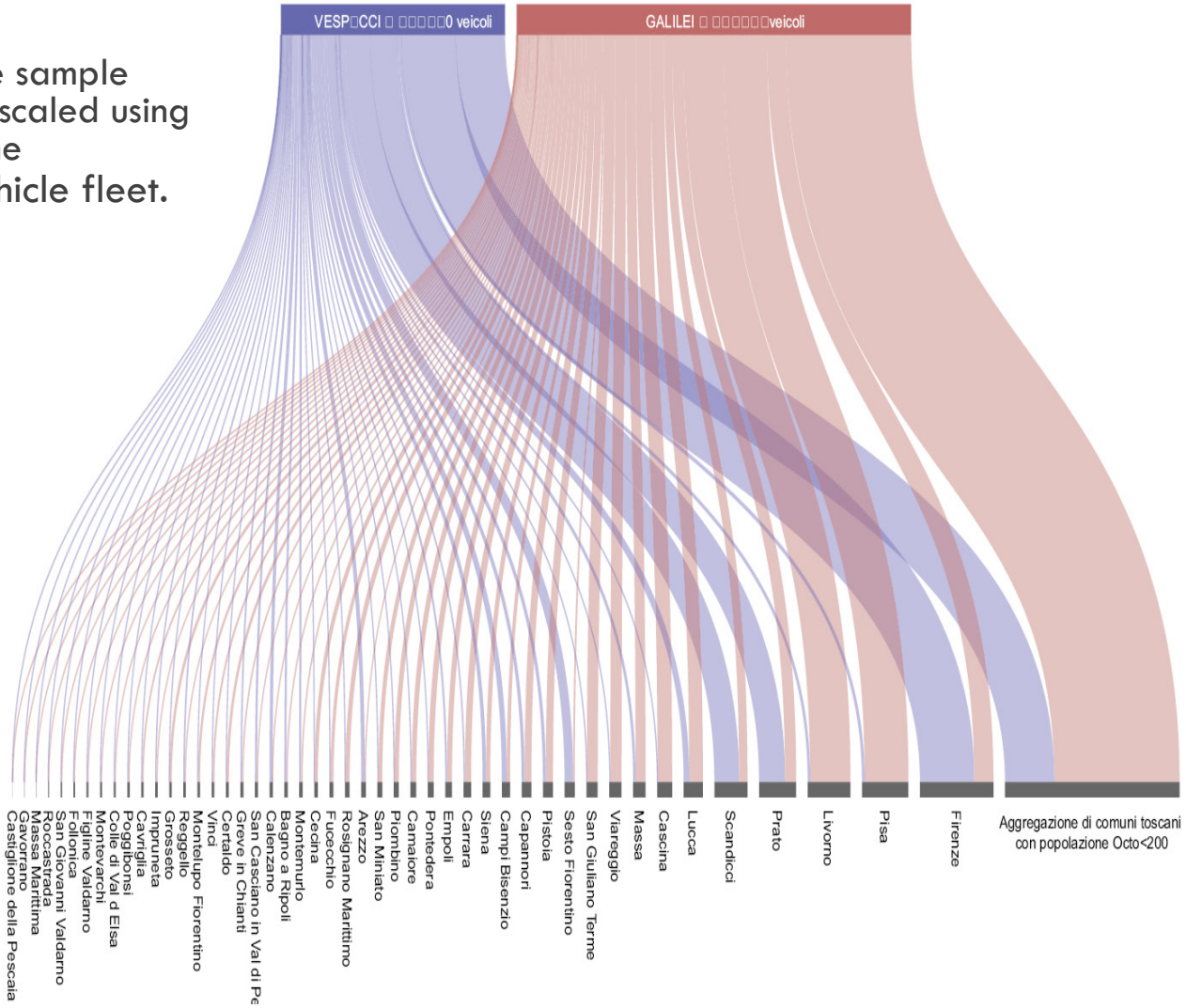
B: The distribution of the stops along the week reflects the flight airport offer

Access pattern to Pisa Airport

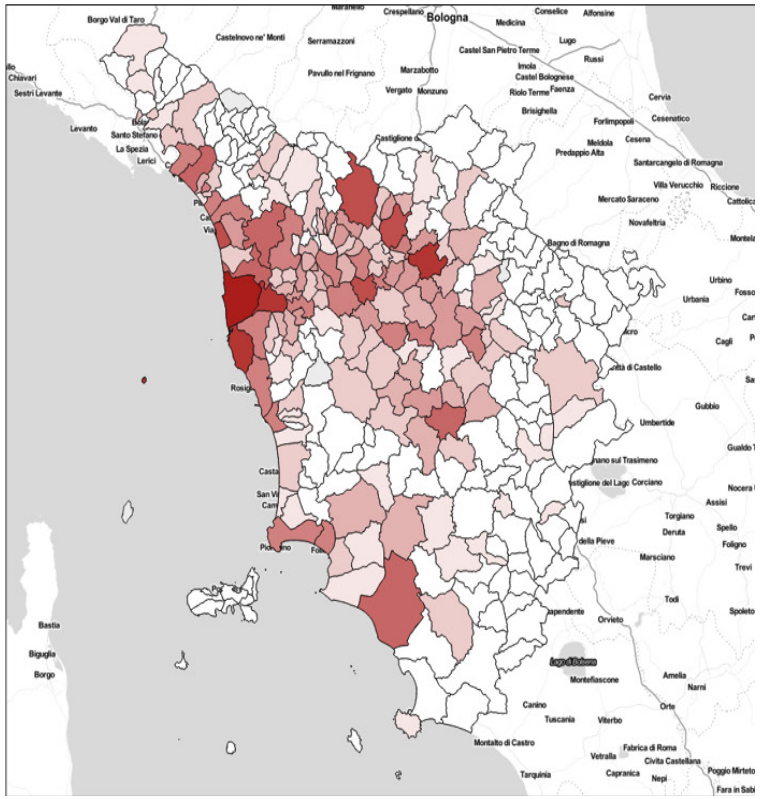


Flows of vehicles to the two airports

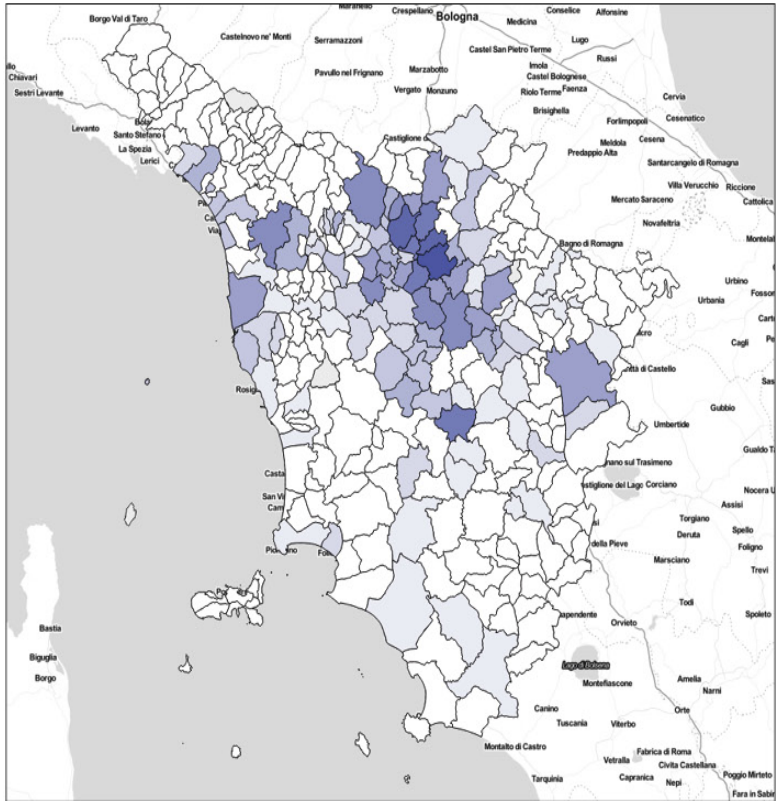
Vehicles of the sample have been re-scaled using ACI data of the circulating vehicle fleet.



Tendency in choosing an Airport: the attractiveness of Galilei vs. Vespucci



Attractiveness of Pisa Airport



Attractiveness of Florence Airport

3

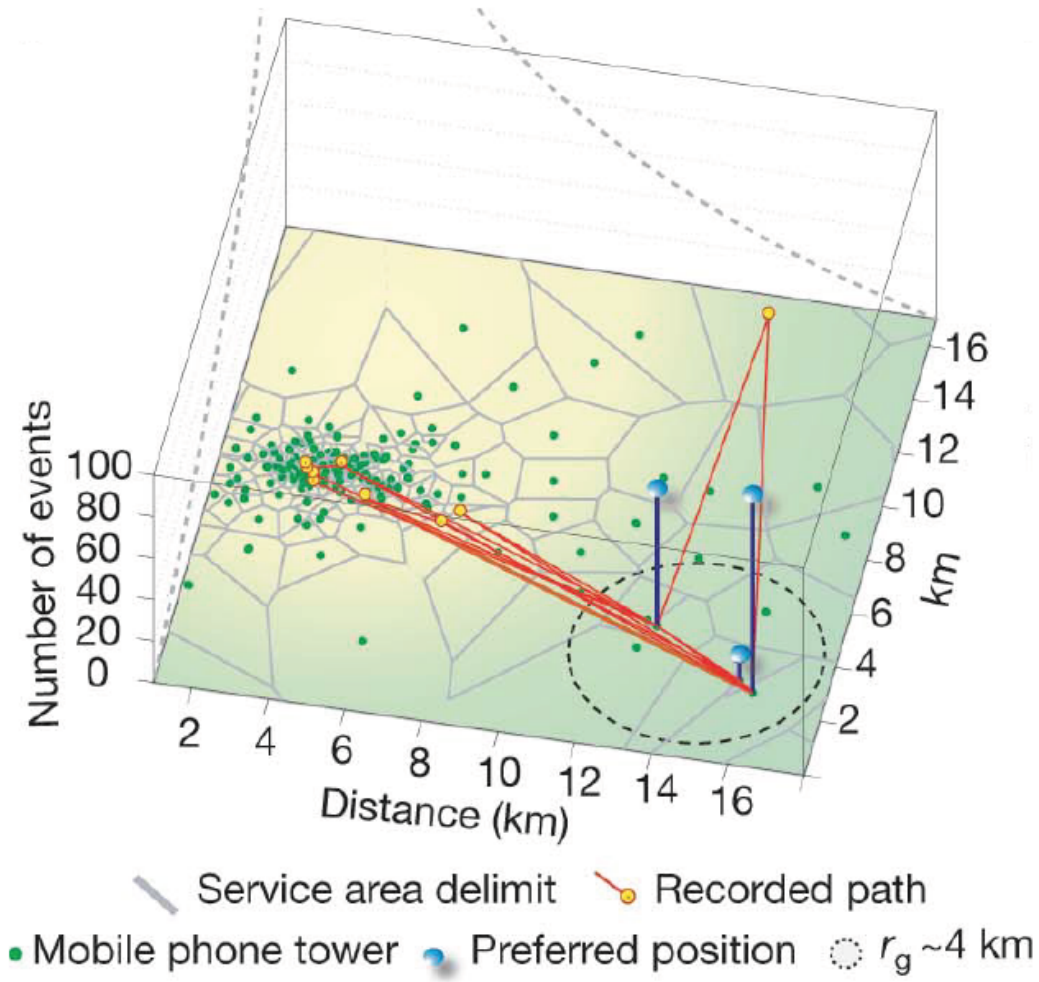
LA MOBILITÀ

GPS data

SUMMARY

- UNDERSTANDING CITY DYNAMICS WITH MOBILE PHONE DATA
  - The data
  - Capturing presence
  - Capturing movement
  - Quantifying city users
  - Building novel demographic and socio-economic indicators
  - Models of human mobility: Explorers&Returners

Focus on country-wide CDR data



when you call

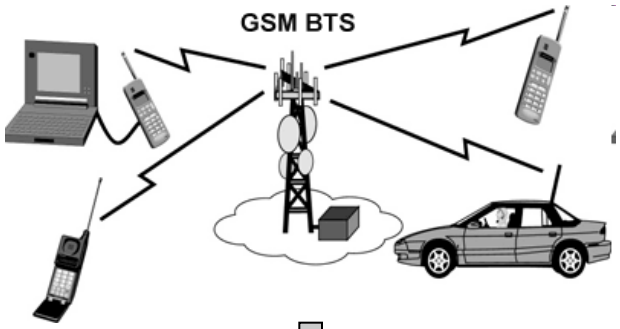
where you call

who you call

GSM data - Description

Call Data Record (CDR)

Data of the users' calls.

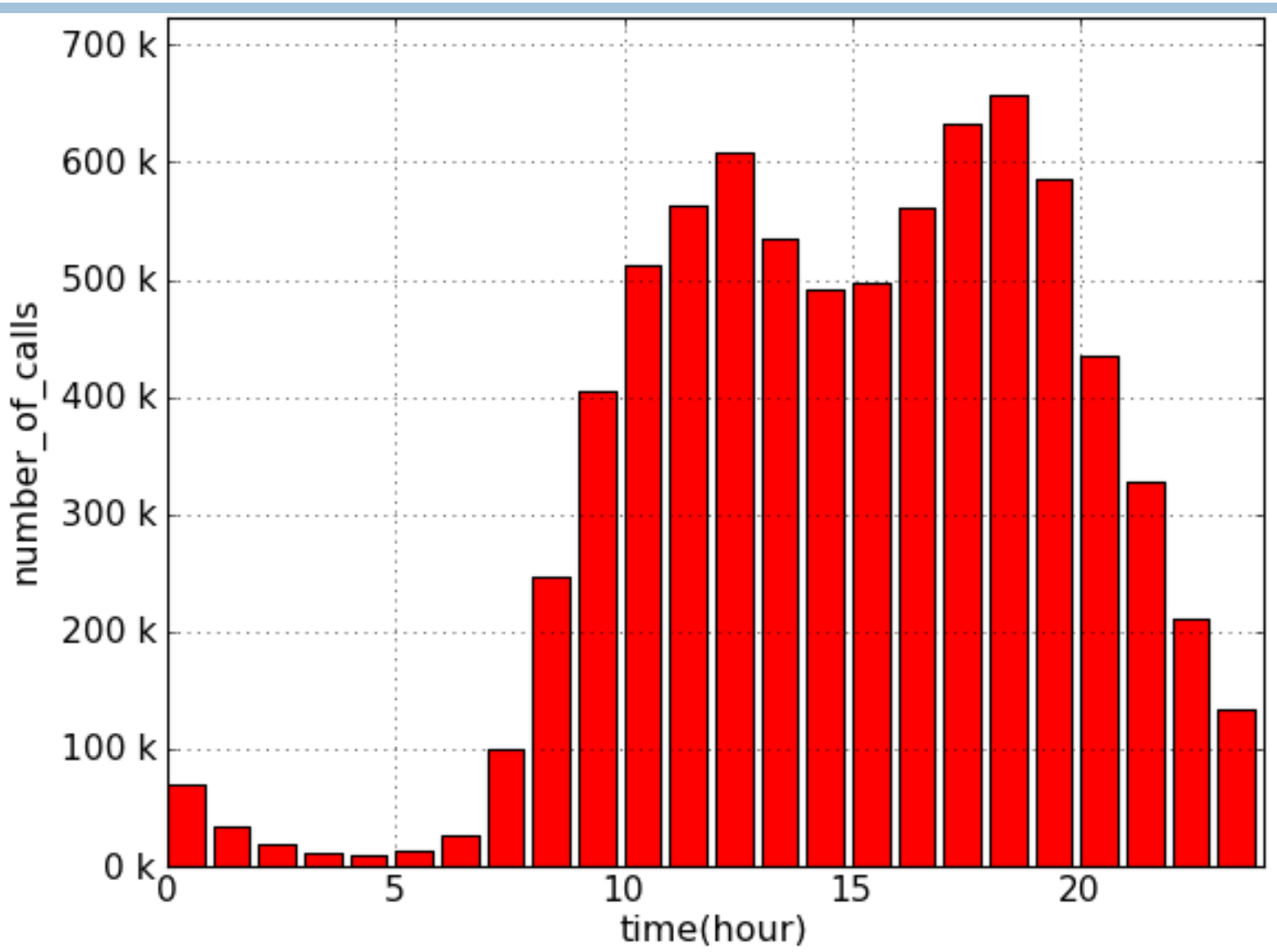


Data gathered from mobile phone operator for billing purpose

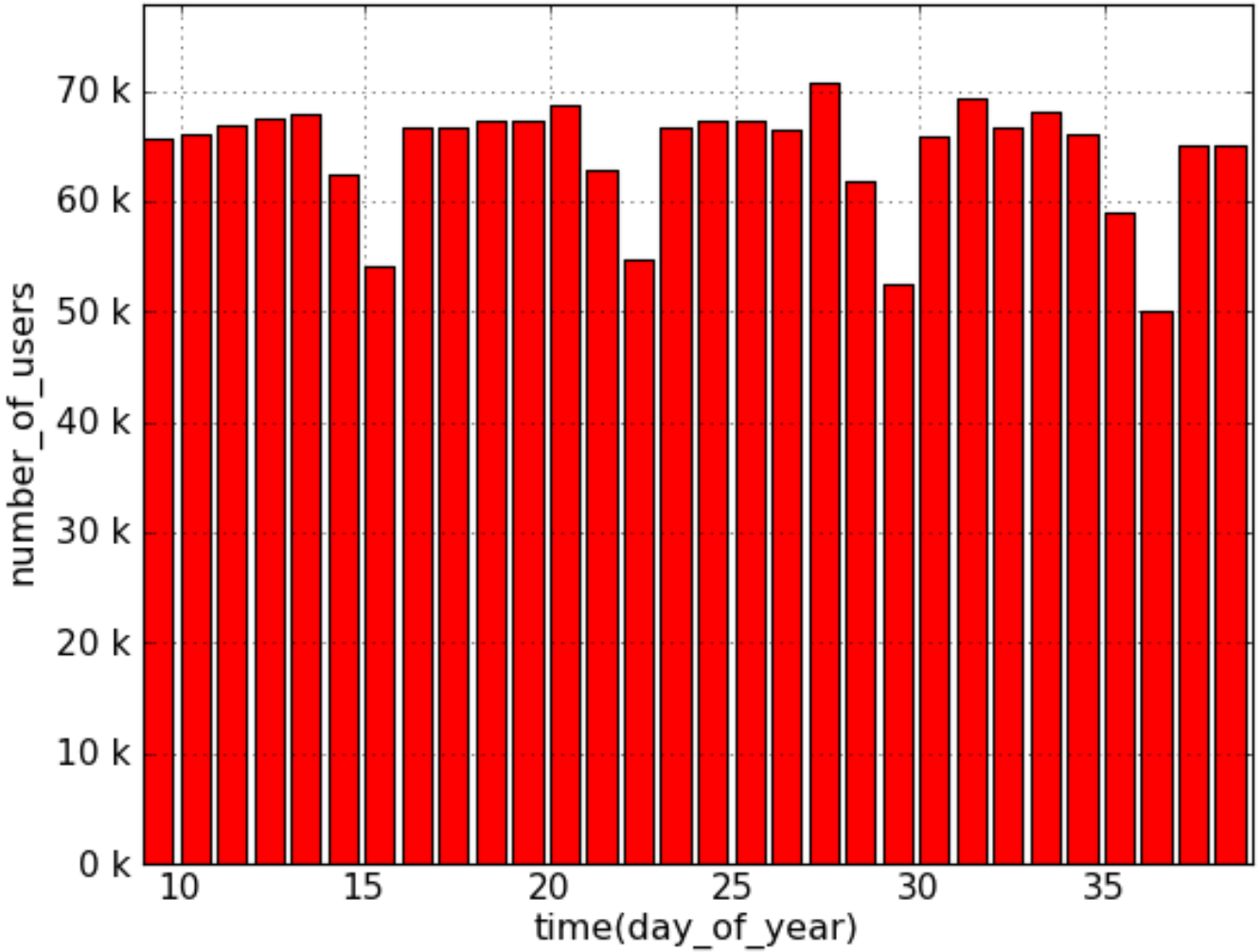
User id	Time start	Cell start	Cell end	Duration
10294595	"2014-02-20 14:24:58"	"PI010U2"	"PI010U1"	48
10294595	"2014-02-20 18:50:22"	"PI002G1"	"PI010U2"	78
10294595	"2014-02-21 09:19:51"	"PI080G1"	"PI016G1"	357

WHICH BASIC INFORMATION CAN BE EXTRACTED FROM CDR?

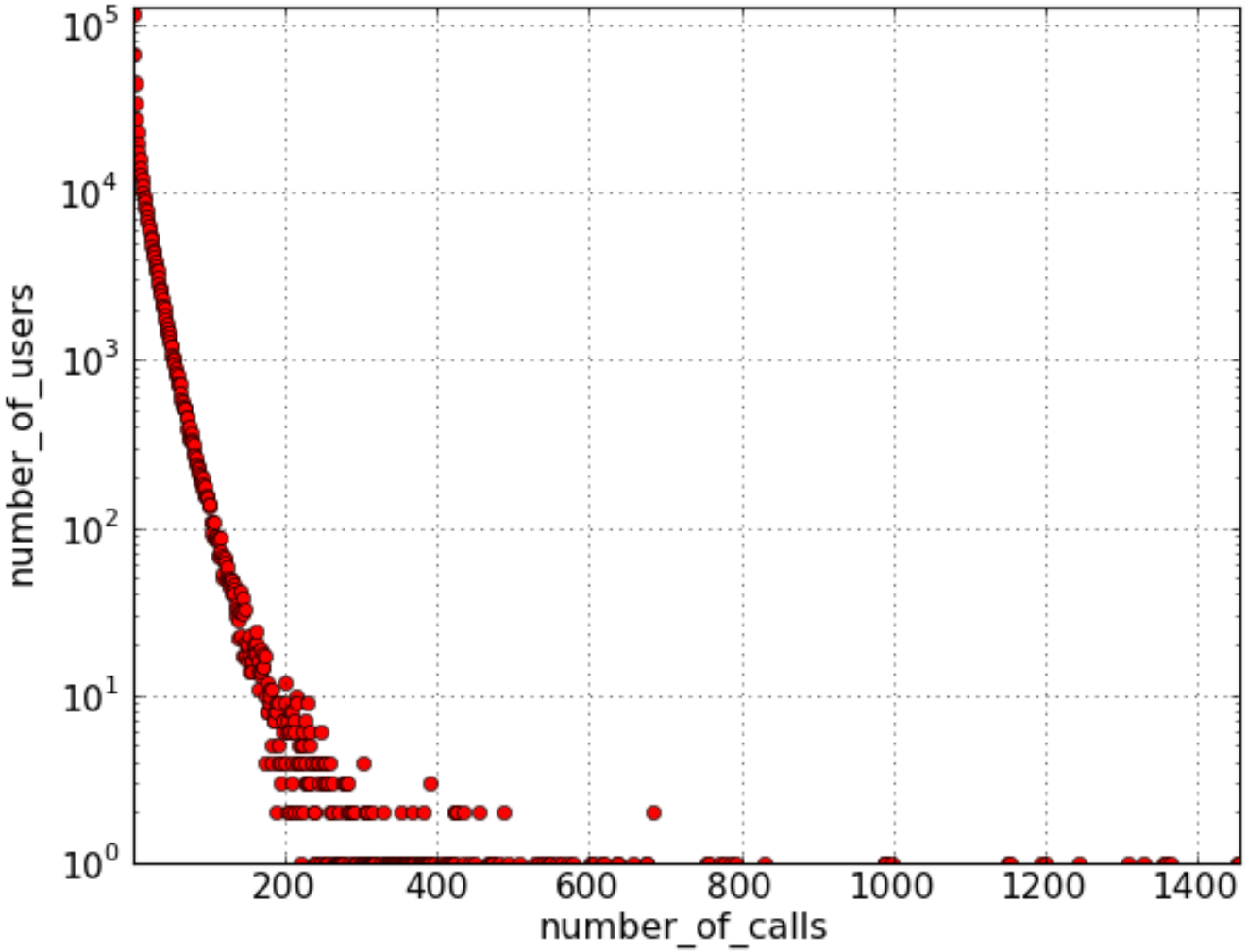
Daily pattern behavior



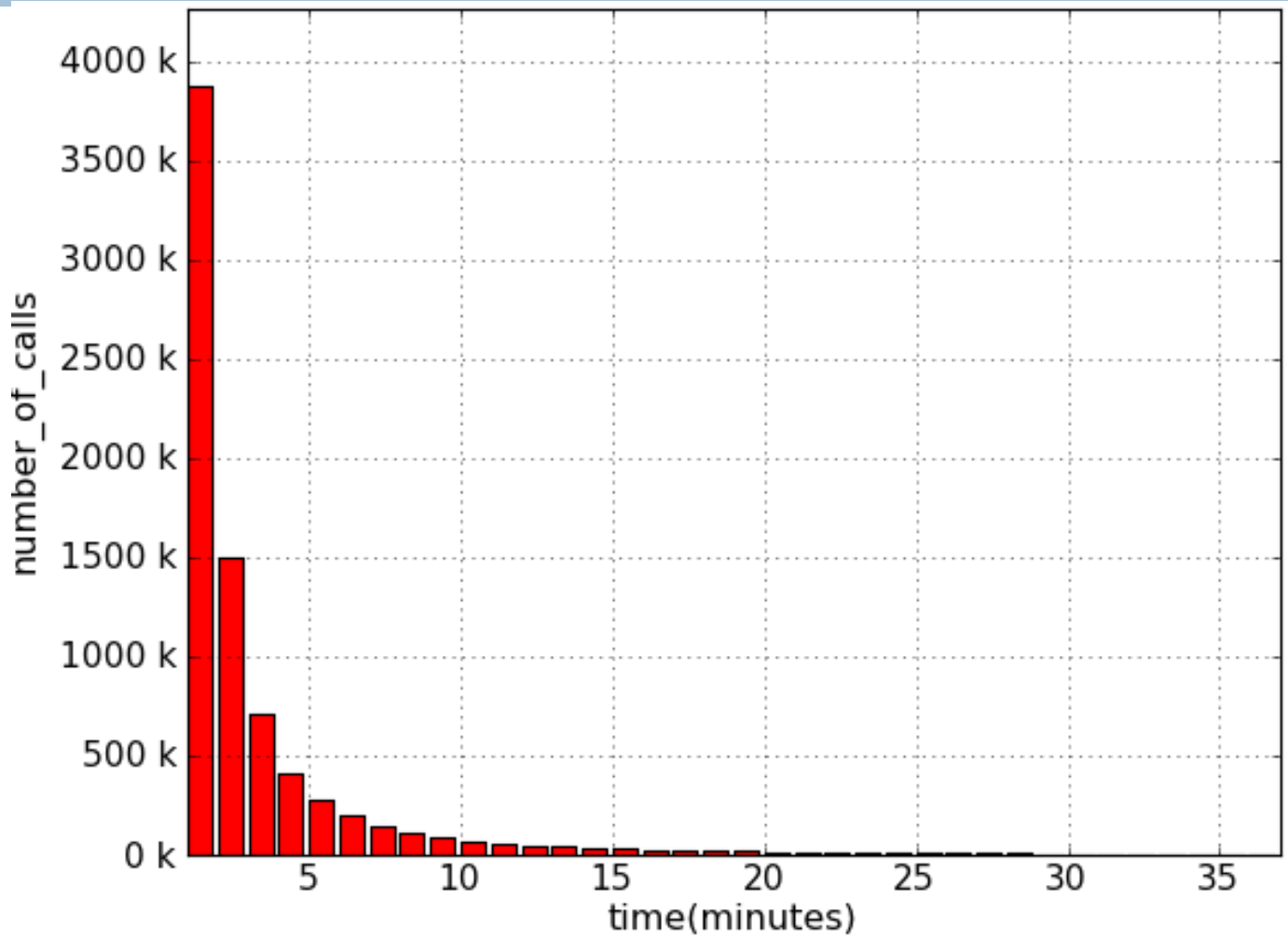
Weekly pattern behavior



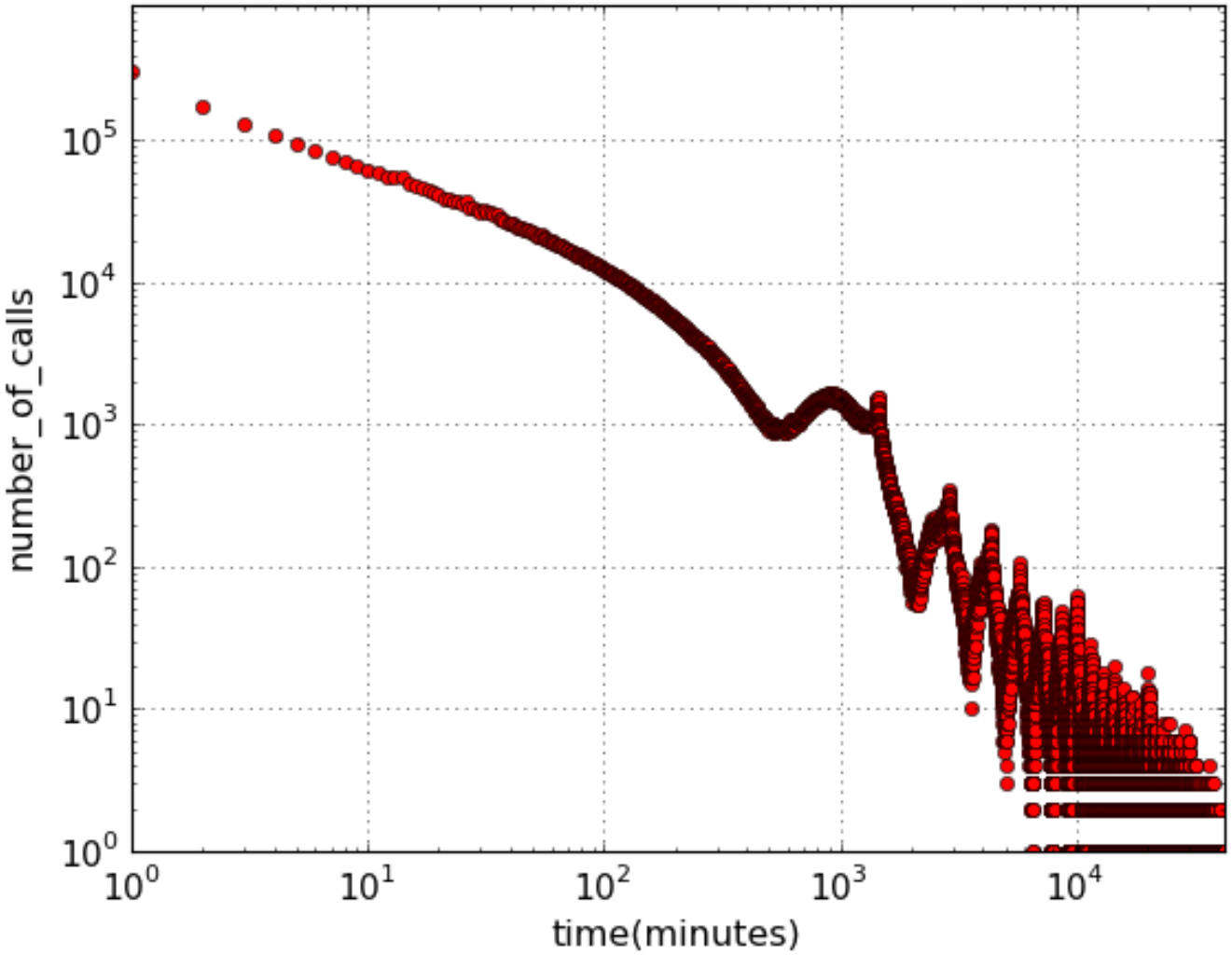
How many times we call?



How long we talk on the phone?



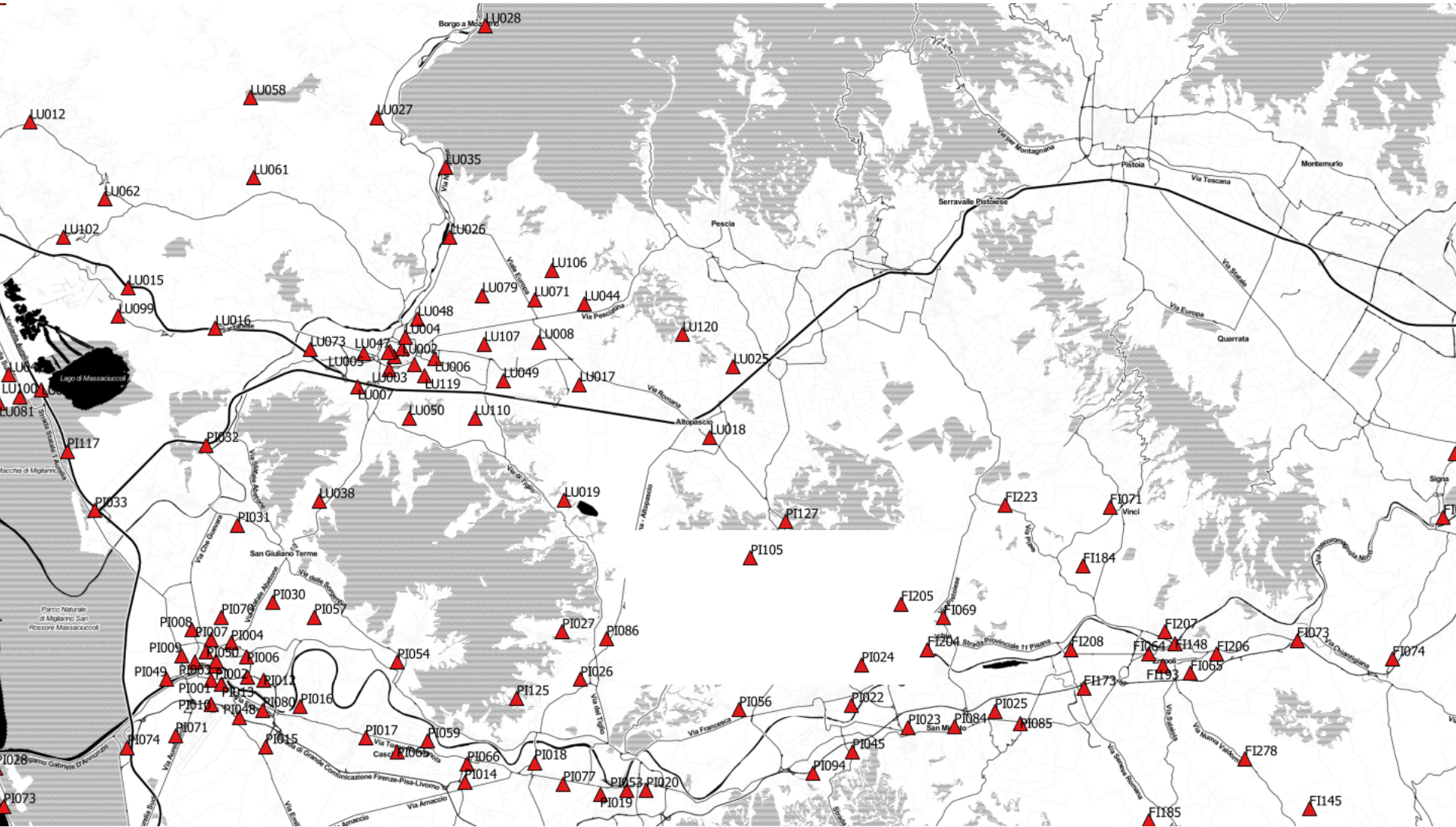
How many minutes goes by a call to the next?



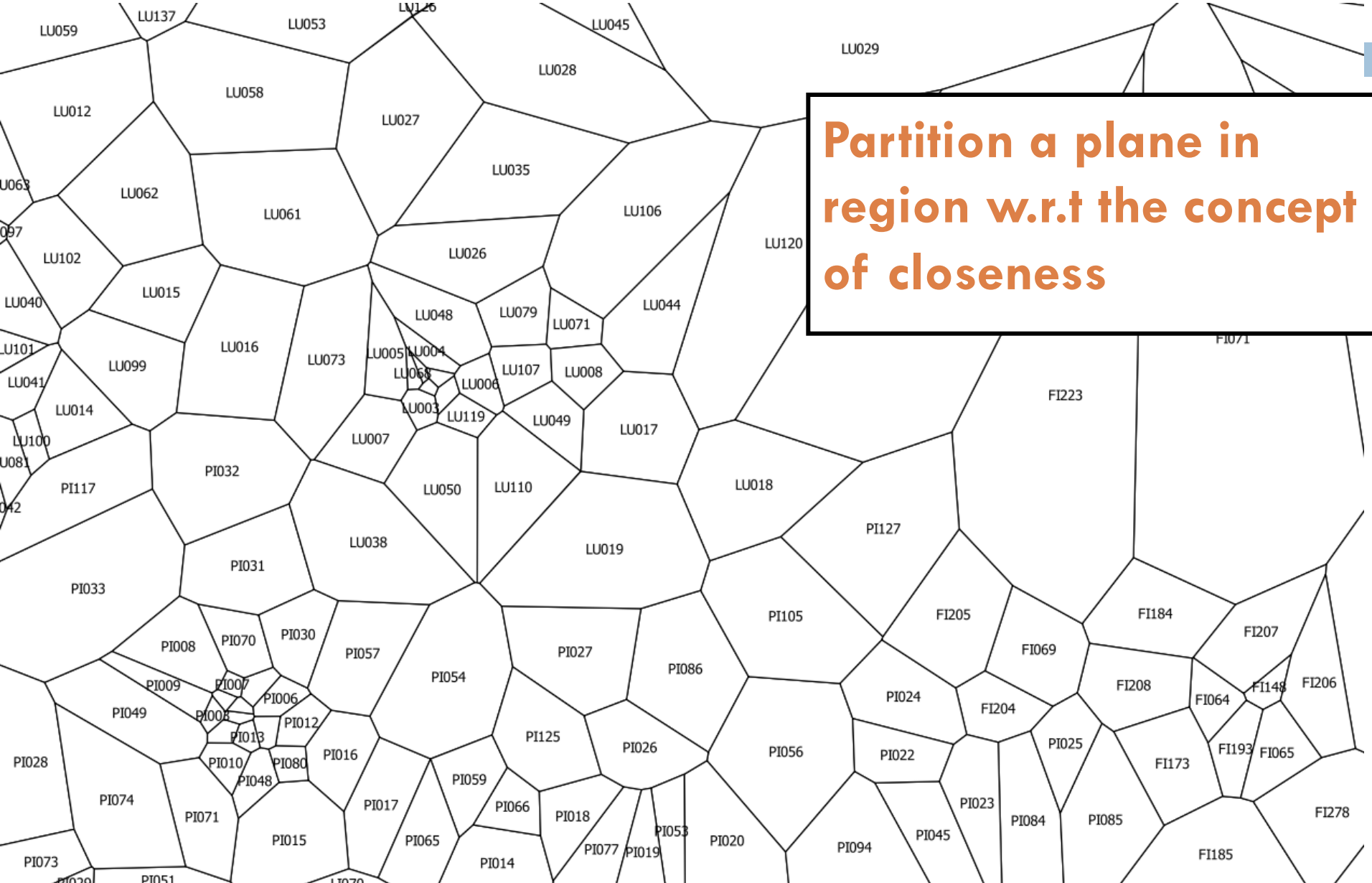
JOIN THE SPATIAL PART OF THE MOBILE PHONE DATA

From CDR to Geography:

CDR contains where the calls started

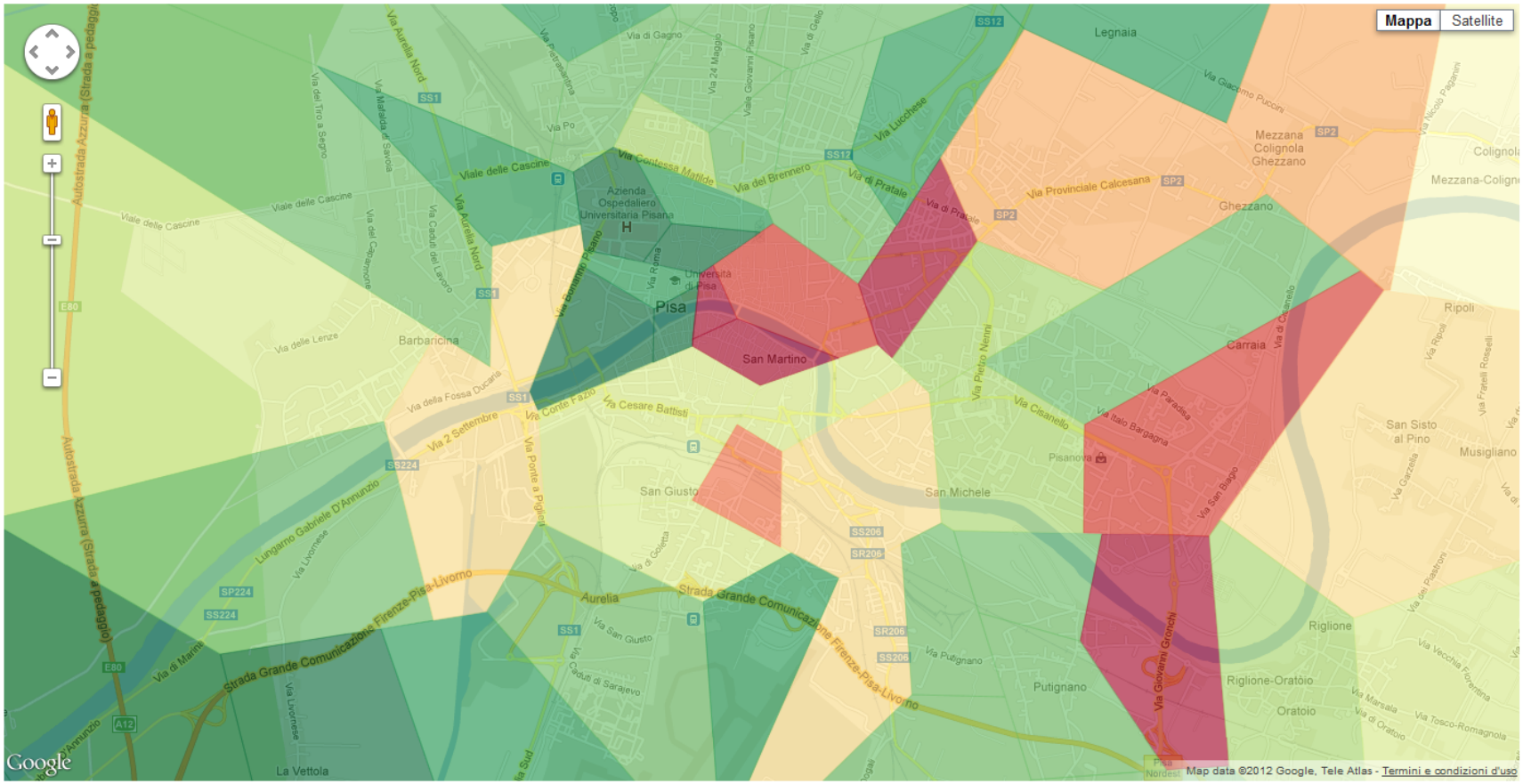


Build the Voronoi tesselation



Partition a plane in region w.r.t the concept of closeness

Spatial distribution of calls

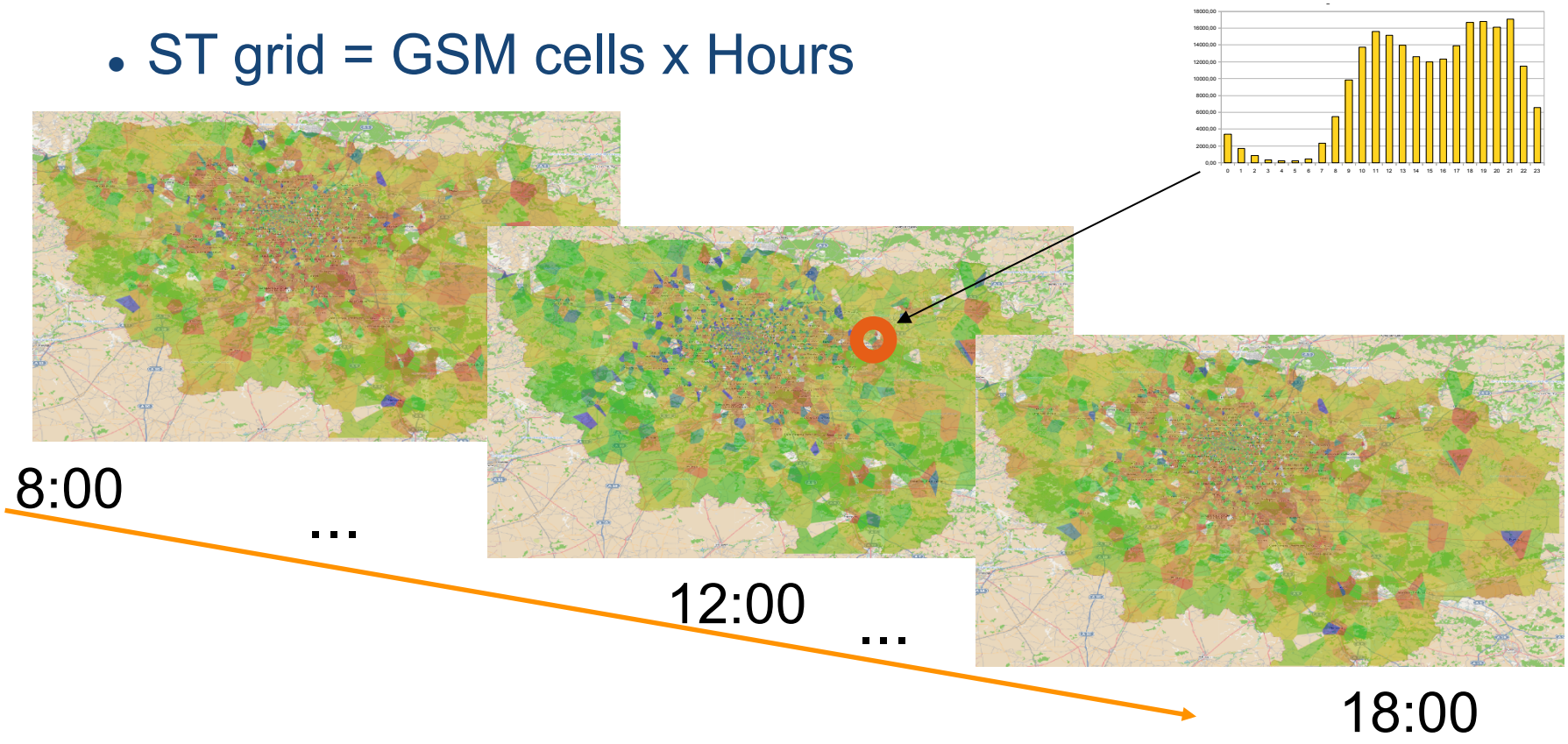


presences of people within the working area of Pisa

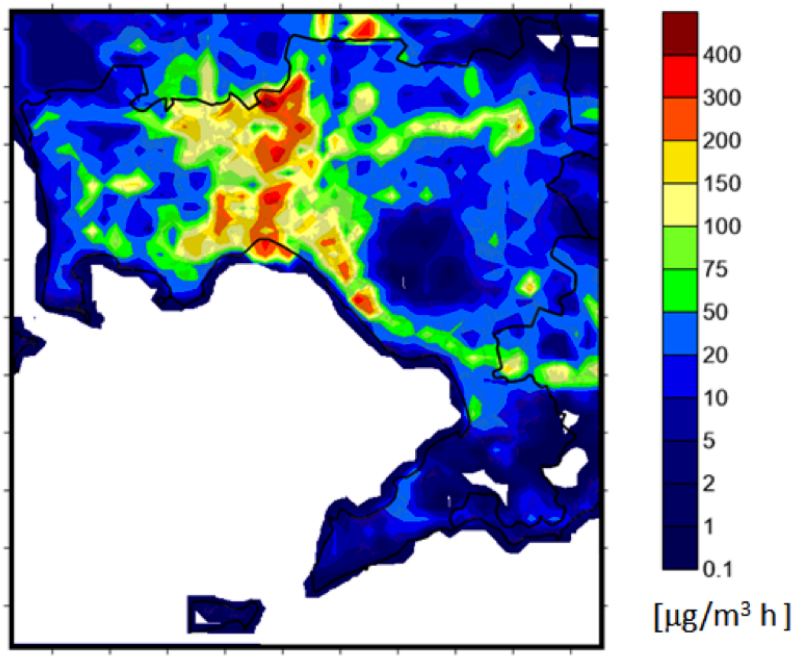
Compute density over a space-time grid

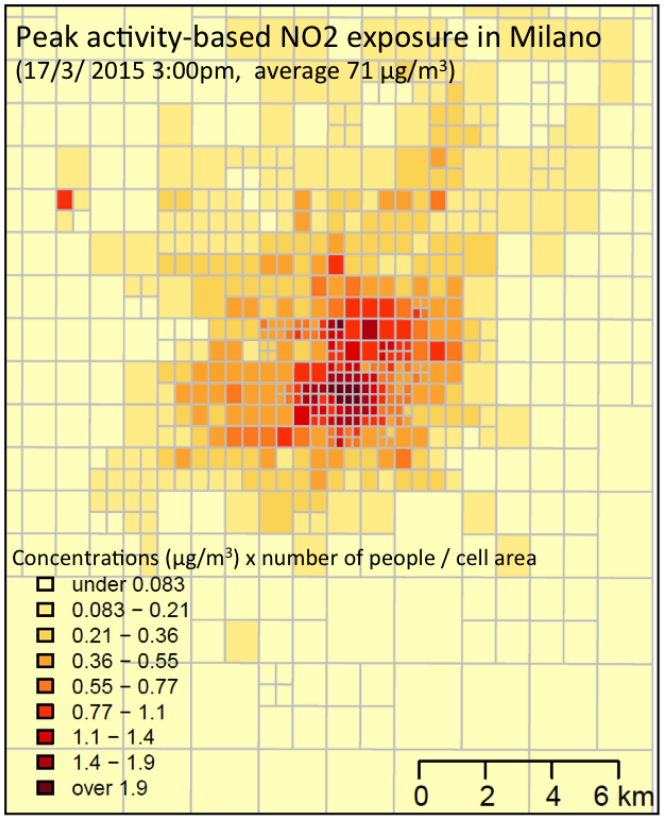
Divide the dataset into days, and days into 24h

- ST grid = GSM cells x Hours



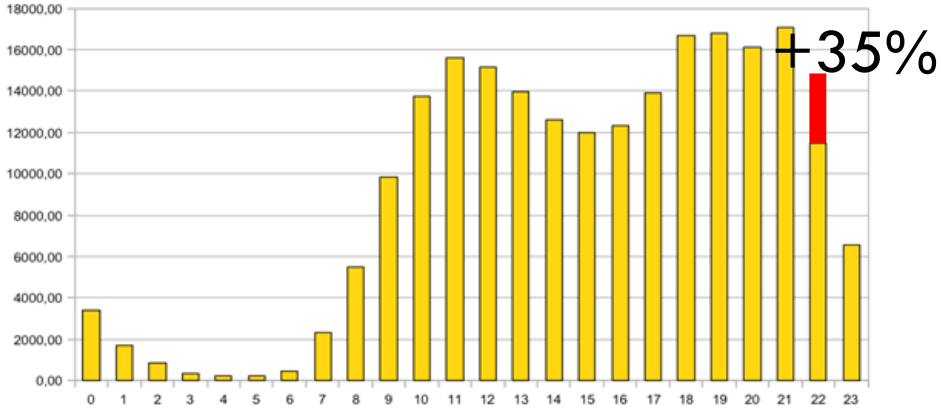
Air pollution/ quality models based on Mobile Phone Activity patterns



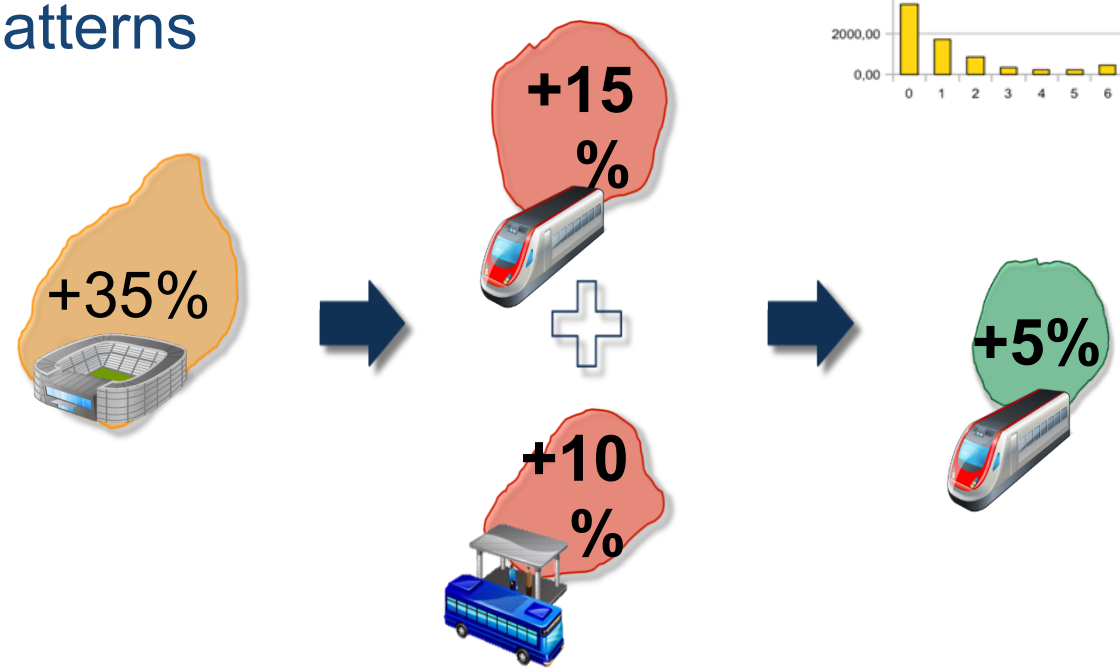


Correletion Patterns: : detect events

1) Detect Event = significant deviation from average



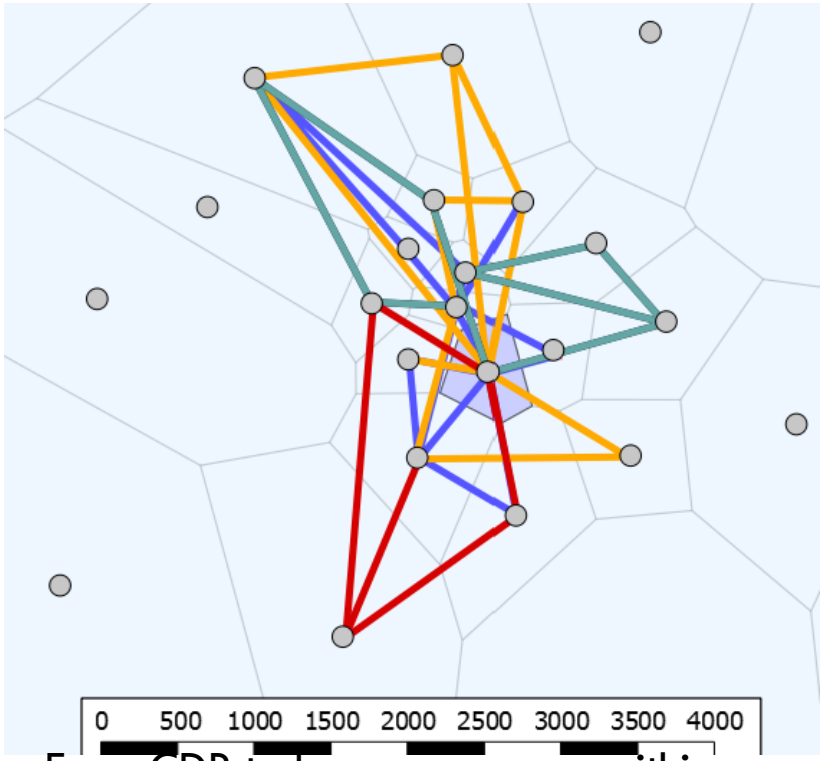
2) Extract correlation patterns



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

A WAY OF 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).

NetMob 2013

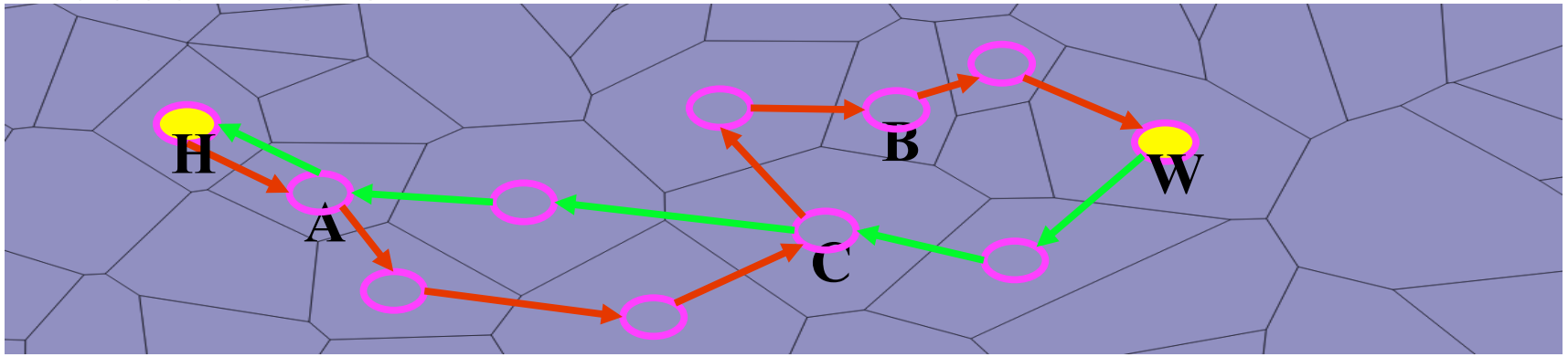
MP4-A Project: Mobility Planning For Africa

Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli

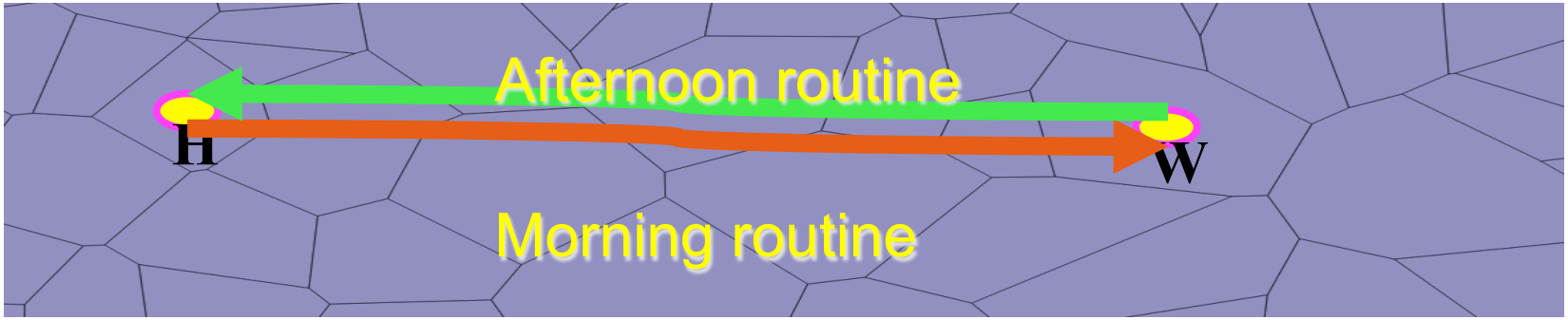
Peter Van Der Mede, Joost De Bruijn, Erik De

Systematic mobility

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



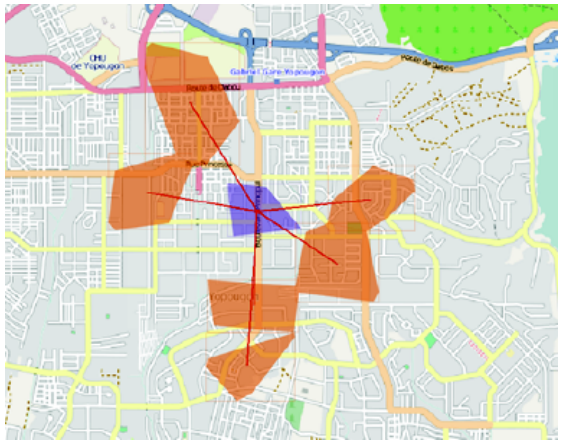
- The whole individual mobility is then summarized by its systematic movements



Systematic O/D matrix

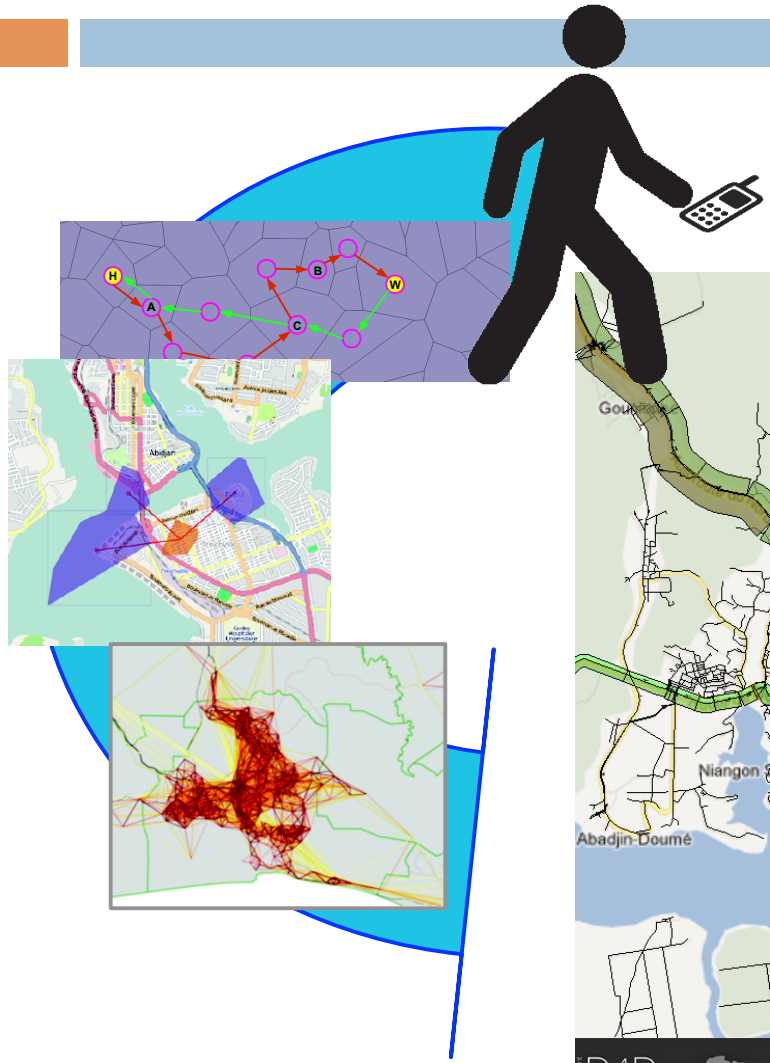
- Combine the ten 2-weeks datasets into one
- For each user, extract significant L1 → L2
- Aggregate (individual) systematic movements into (collective) systematic flows
- Examples:

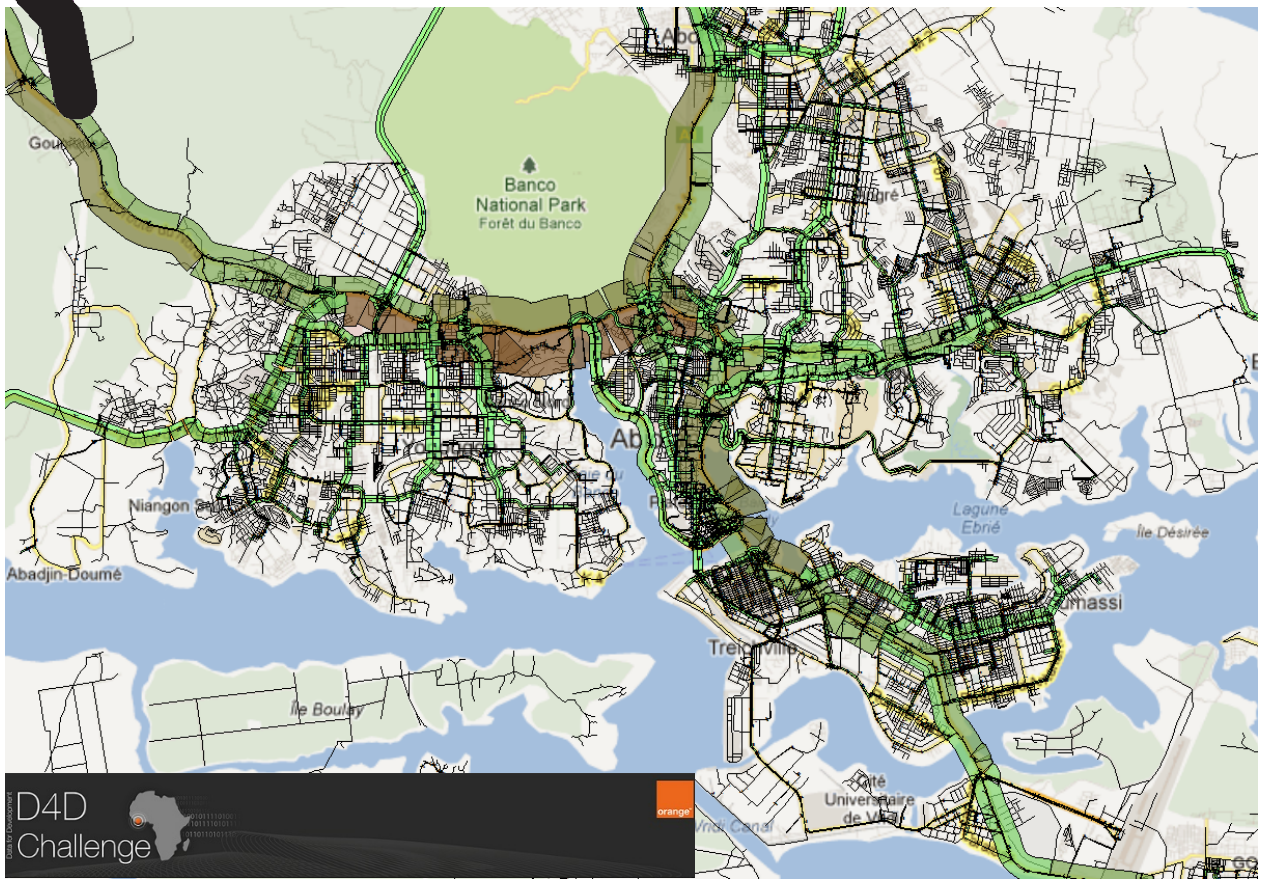
Outgoing traffic

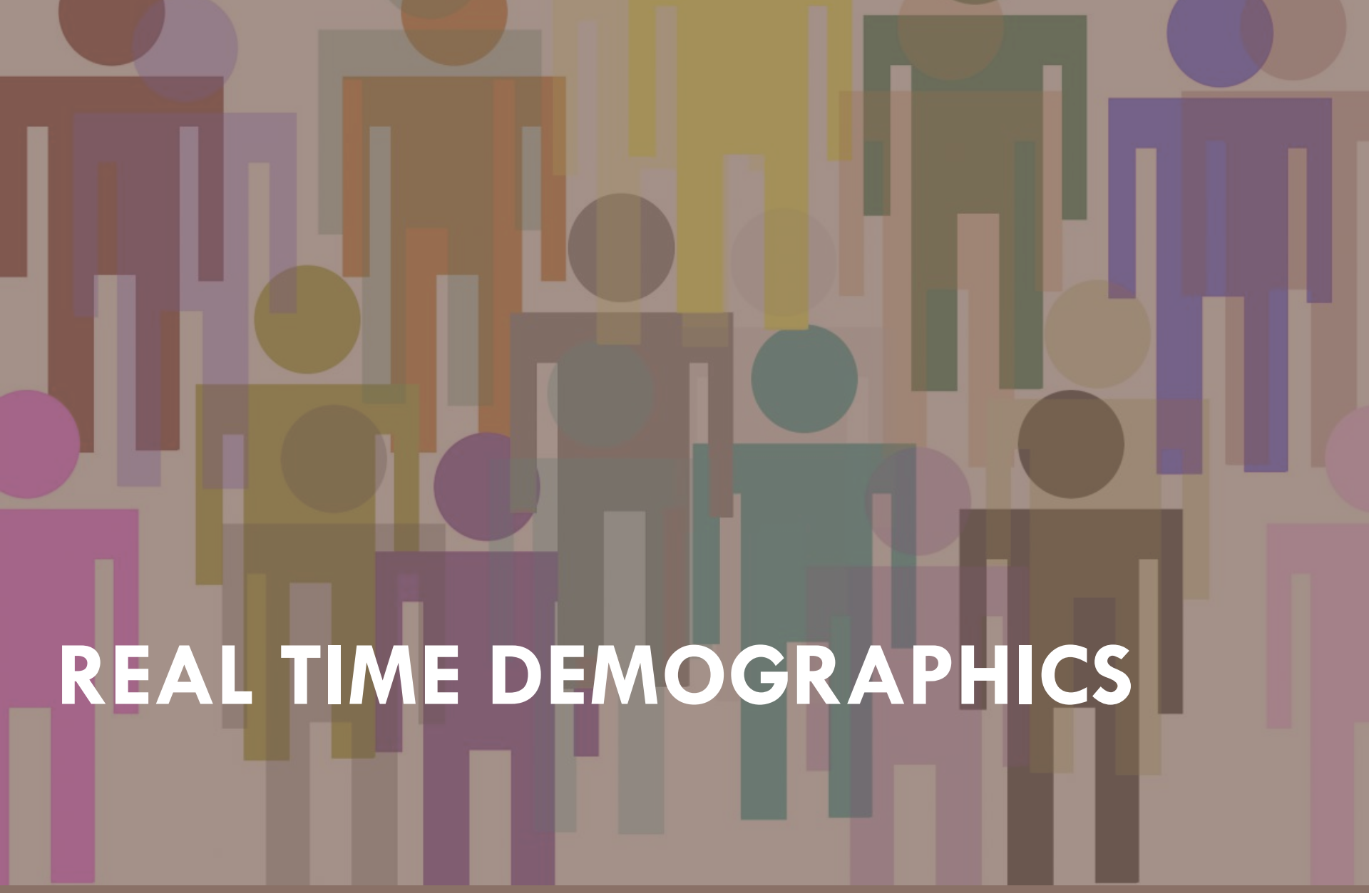


Incoming traffic

Big Data for Developing Countries







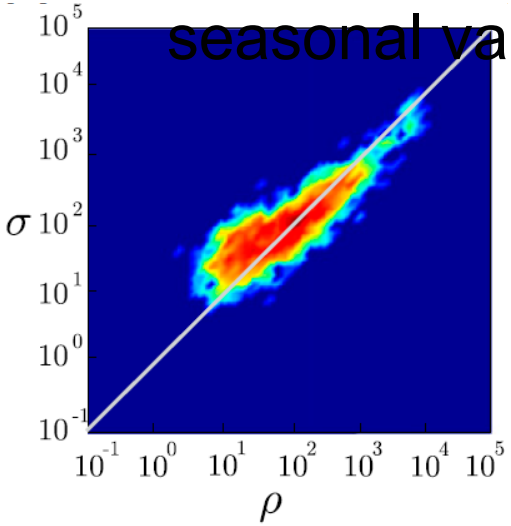
REAL TIME DEMOGRAPHICS

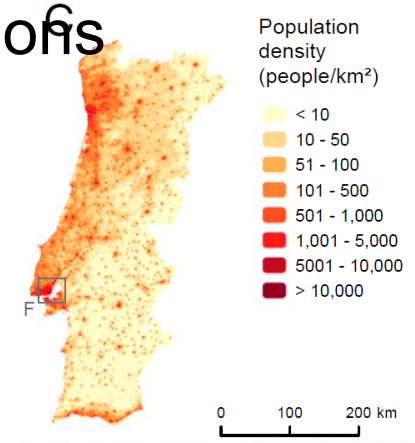
KDD Lab

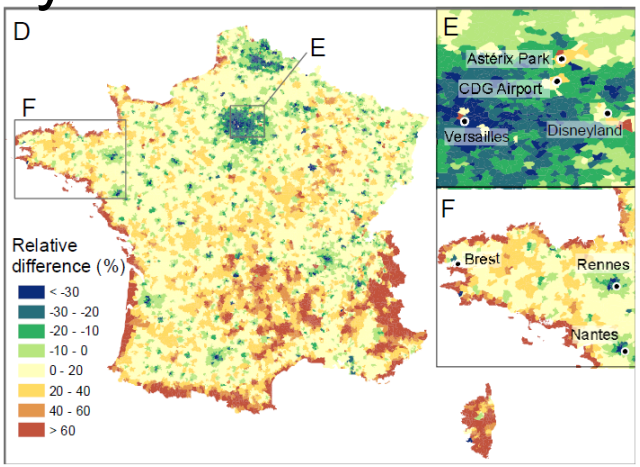
Identifying important locations

- Estimating users' **residence through night activity**
  - Home = region with highest frequency of calls during nighttime

Estimate resident population density … and seasonal variations



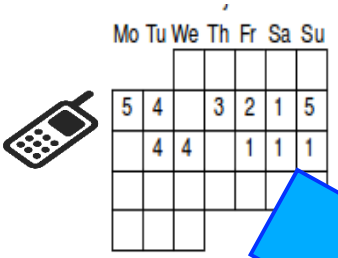




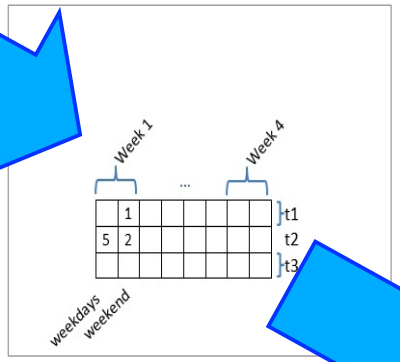
Pierre Deville et al. Dynamic population mapping using mobile phone data. PNAS vol. 111 no. 45, pp. 15888–15893, doi: 10.1073/pnas.1408439111

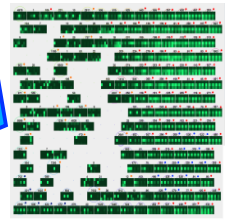
Quantifying city users

GSM Calls

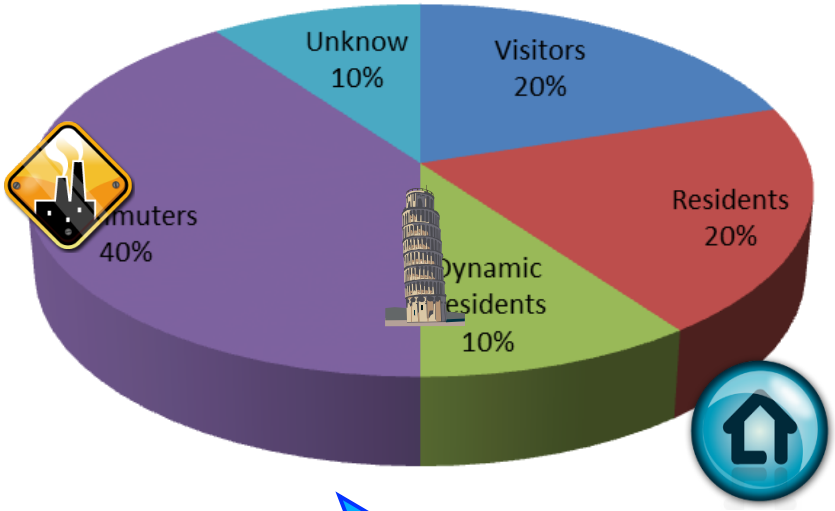


Temporal Profile



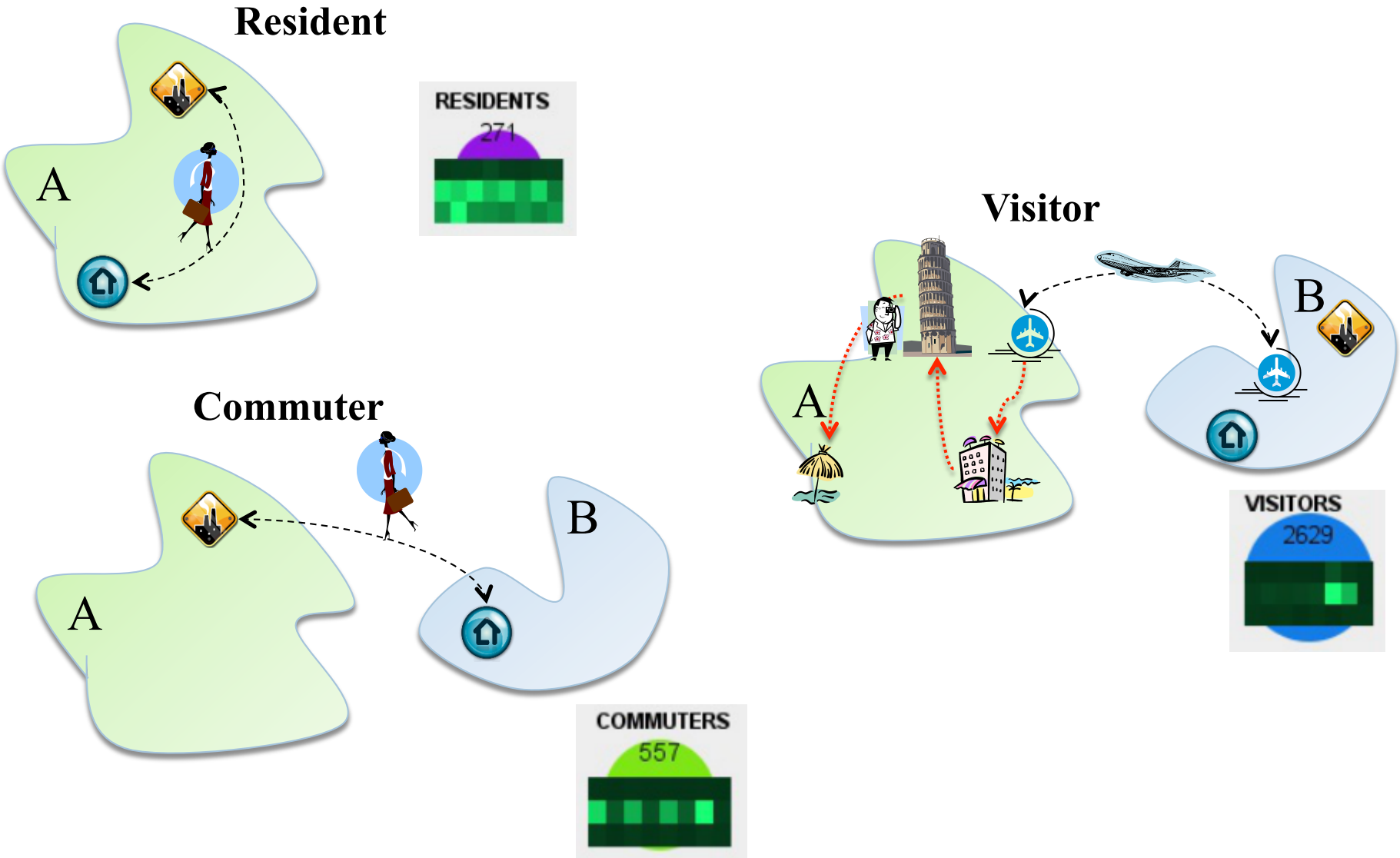


Profile Map



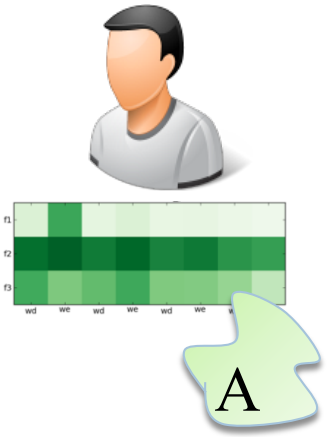
L. Gabrielli, Furletti, B., Trasarti, R., Giannotti, F., and Pedreschi, D., “City users’ classification with mobile phone data”, in IEEE Big Data, Santa Clara (CA) - USA, 2015

Calling profiles

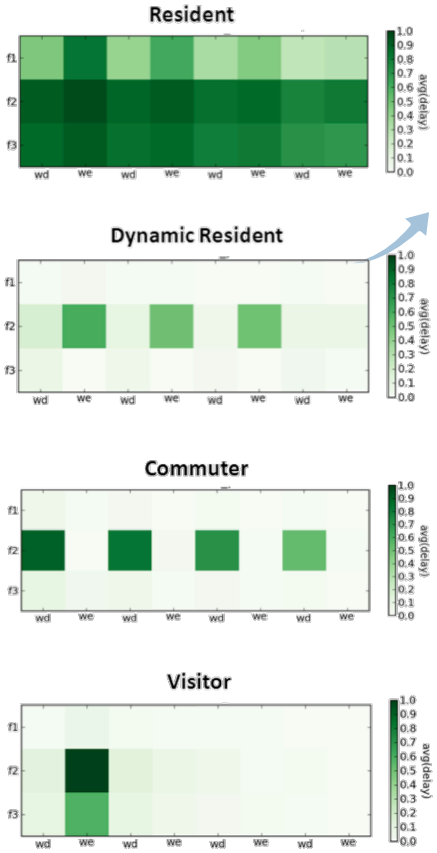


Sociometer with Mobile Phone Data.

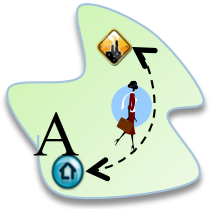
Users' Call Profiles



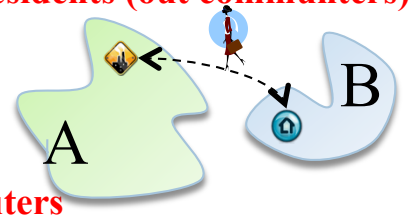
Classification Algorithm



Residents

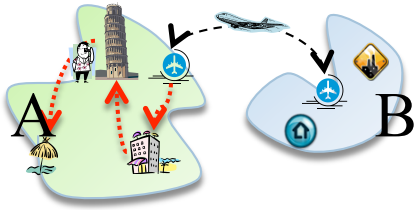


Dynamic Residents (out communters)

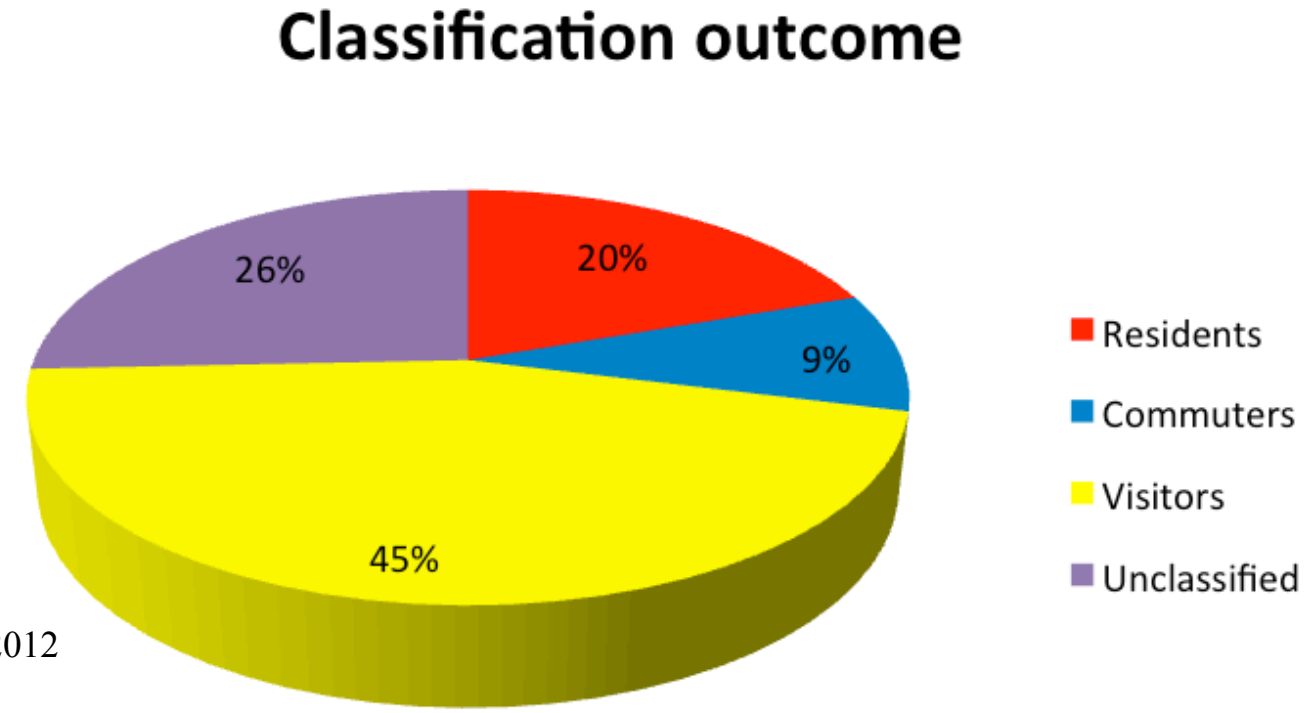


(in) Commuters

Visitors



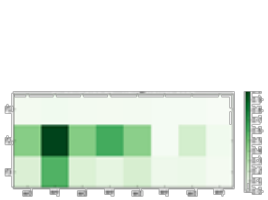
Sociometer: the city user meter

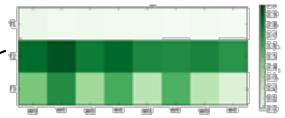


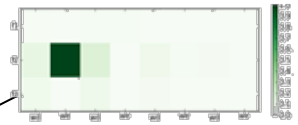
Pisa, January 2012

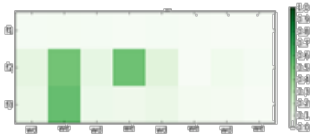
Analysis of GSM calls data for understanding user mobility behavior
B Furletti, L Gabrielli, C Renso, S Rinzivillo Big Data, 2013 IEEE International Conference on, 550-555

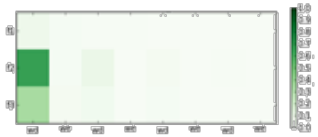
The many profiles of an individual

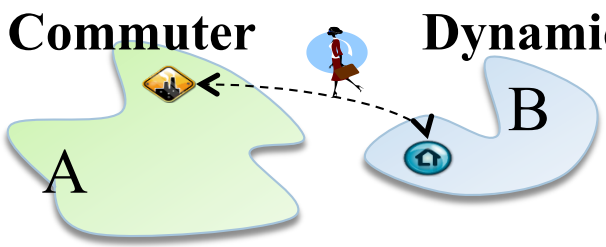




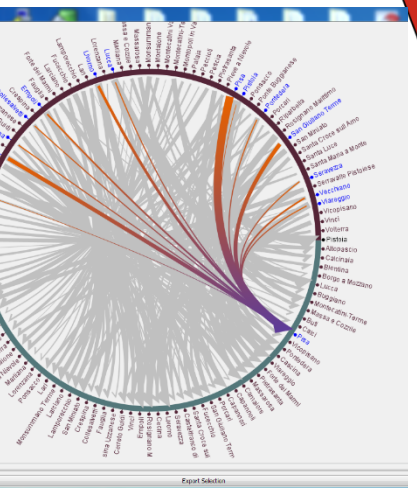








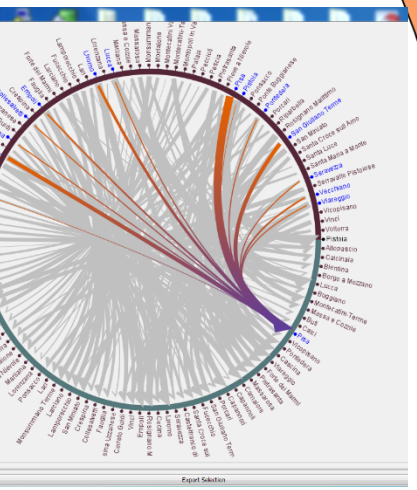
Pisa:
In Commuters



- 0.00 - 0.18
- 0.18 - 0.46
- 0.46 - 0.85
- 0.85 - 1.68
- 1.68 - 2.94
- 2.94 - 5.12
- 5.12 - 7.08
- 7.08 - 12.96

Gallicano
Coreglia Antelminelli
Vergemoli
Bagni di Lucca
Seravezza
Stazzema
Borgo a Mozzano
Pescaglia
Forte dei Marmi
Pietrasanta
Camaiore
Massarosa
Viareggio
Lucca
Capannori
Montecarlo
Altopascio
Vecchiano
San Giuliano Terme
Calci
Buti
Vicopisano
Calcinaia
Pisa
Cascina
Pontedera
Ponsacco
Crespina
Lari
Capannoli
Collesalvetti
Fauglia
Livorno
Casciana Terme
Santa Luce
Rosignano Marittimo
Castellina Marittima
Riparbella
Cecina
Montescudaio
Palaia
Peccioli
Lajatico
Volterra
Pomarance
Montaione
Gambassi Terme
Fucecchio
Cerreto Guidi
Vinci
Santa Maria a Monte
San Miniato
Empoli
Montelupo Fiorentino
Barberino di Mugello
Scarperia
San Piero a Sieve
Borgo San Lo
Calenzano
Vaglia
Sesto Fiorentino
Pontas
Campi Bisenzio
Fiesole
Signa
Firenze
Scandicci
Bagno a Ripo
Impruneta
Montespertoli
San Casciano in Val di Pesa
Castelfiorentino
Greve in Chian
Tavarnelle Val di Pesa

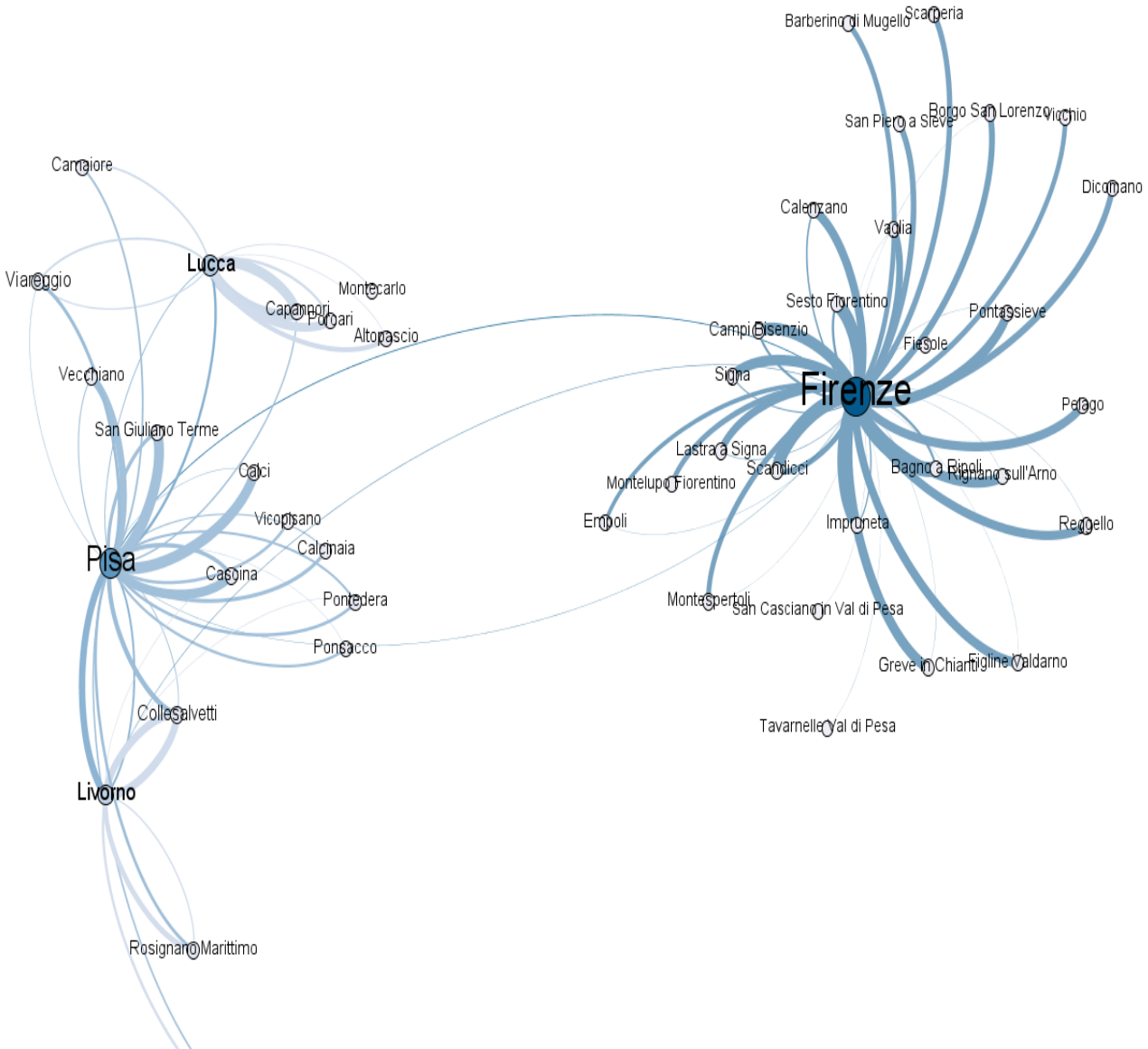
Pisa:
Out Commuters



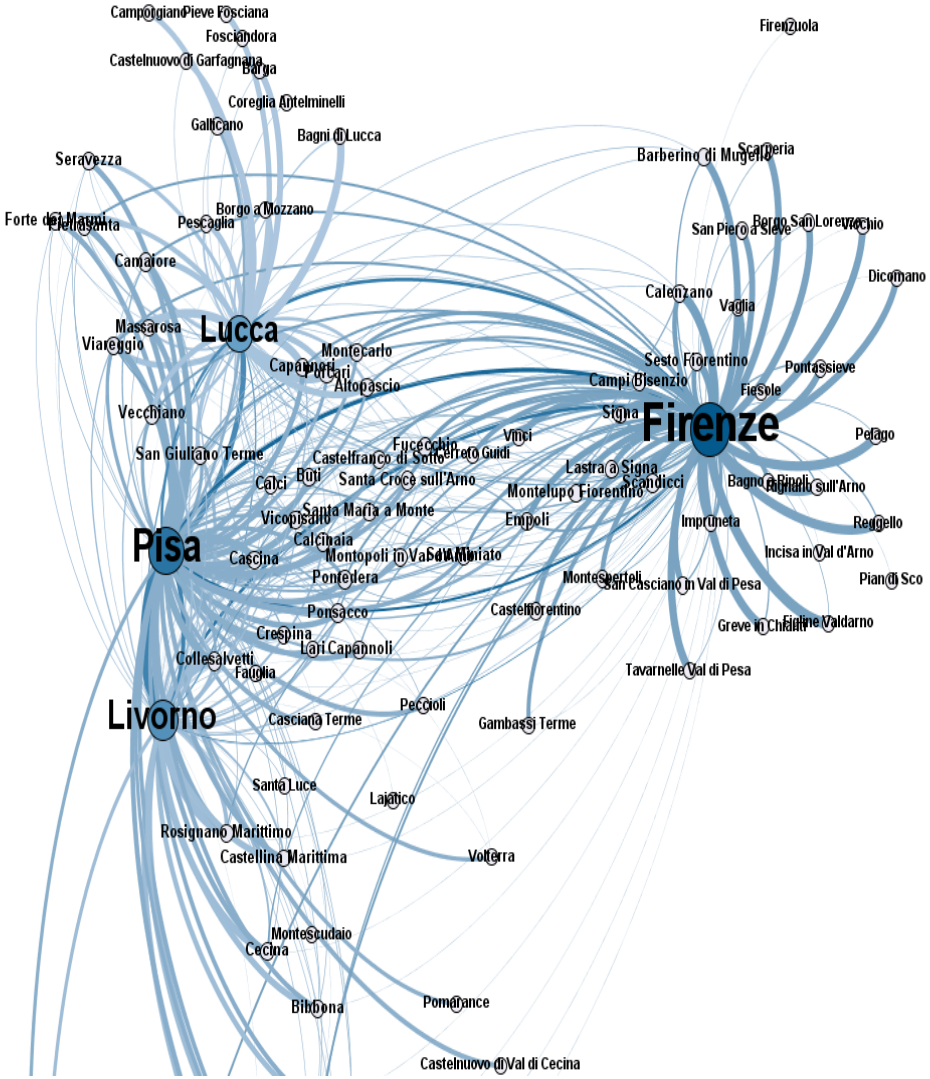
0.00 - 0.17
0.17 - 0.49
0.49 - 1.19
1.19 - 2.16
2.16 - 3.43
3.43 - 4.68
4.68 - 8.05
8.05 - 33.81

Gallicano
Coreglia Antelminelli
Vergemoli
Bagni di Lucca
Seravezza
Stazzema
Forte dei Marmi
Pietrasanta
Borgo a Mozzano
Pescaglia
Camaiore
Massarosa
Viareggio
Lucca
Capannori
Montecarlo
Altopascio
Vecchiano
San Giuliano Terme
Calci
Buti
Vicopisano
Santa Maria a Monte
Pisa
Cascina
Calcinaia
Pontedera
Ponsacco
Palaia
Crespina
Lari
Capannoli
Collesalvetti
Fauglia
Livorno
Casciana Terme
Peccioli
Santa Luce
Lajatico
Posignano Marittimo
Castellina Marittima
Riparbella
Cecina
Montescudaio
Volterra
Pomarance
Fucecchio
Cerreto Guidi
Vinci
Montelupo Fiorentino
Empoli
San Miniato
Montespertoli
Castelfiorentino
Tavarnelle Val di Pesa
Barberino di Mugello
Scarperia
San Piero a Sieve
Borgo San Lo
Calenzano
Vaglia
Sesto Fiorentino
Pontas
Campi Bisenzio
Fiesole
Signa
Firenze
Scandicci
Bagno a Ripo
Impruneta
San Casciano in Val di Pesa
Greve in Chian
Montaione
Gambassi Terme

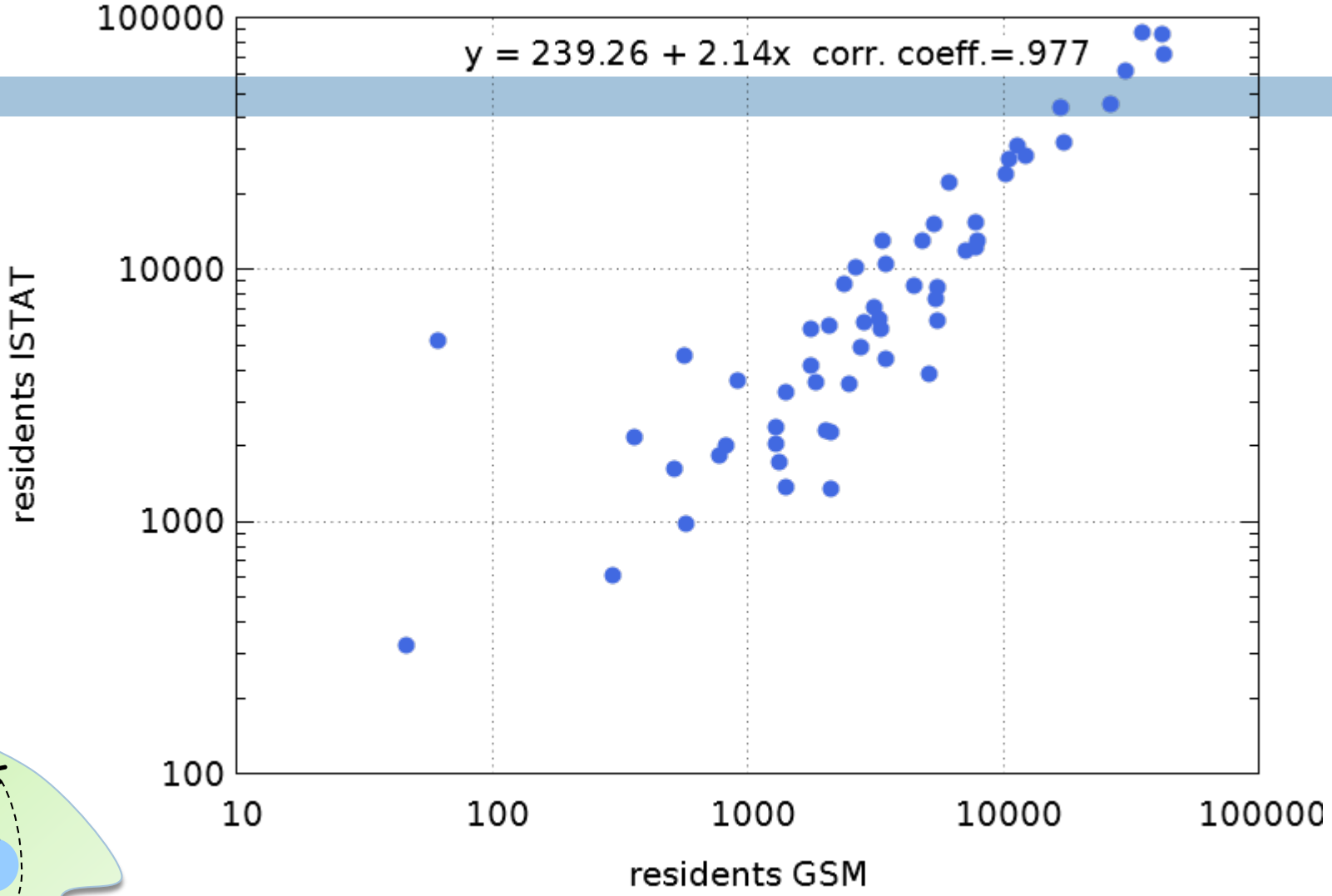
Commuter in-out flows

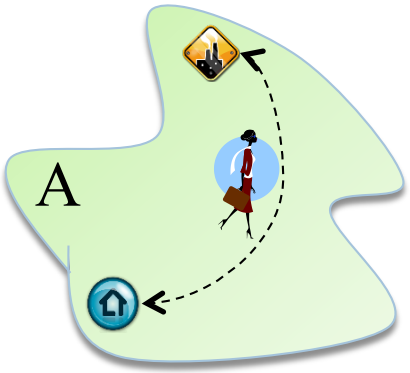


VISITOR in-out flows



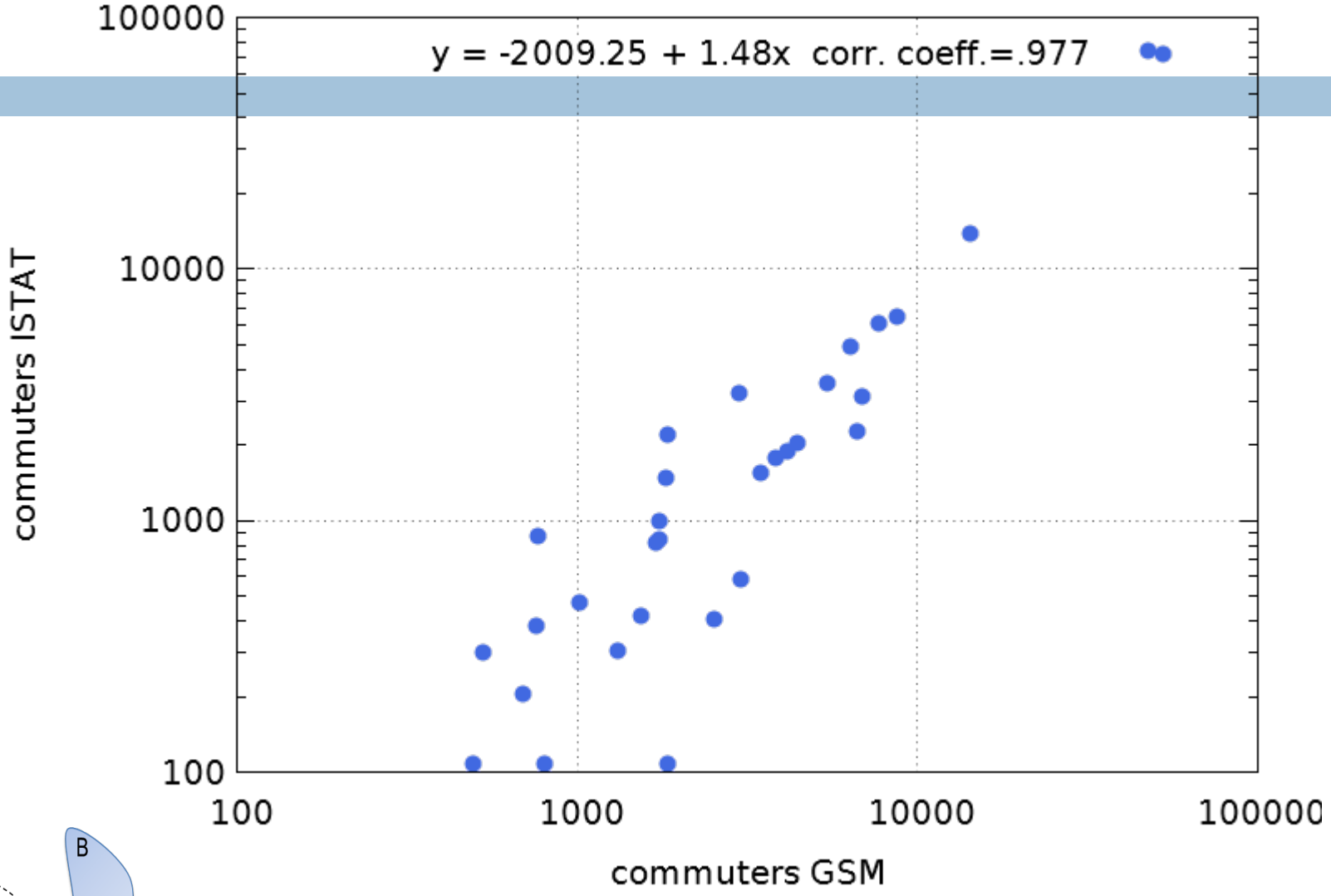
Residents – validation with administrative data

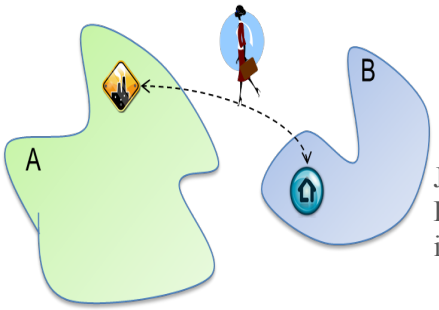




Join work with ISTAT: Barbara Furletti, Lorenzo Gabrielli, Giuseppe Garofalo, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio. Use of mobile phone data to estimate mobility flows. Measuring urban population and intercity mobility using big data in an integrated approach. Italian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

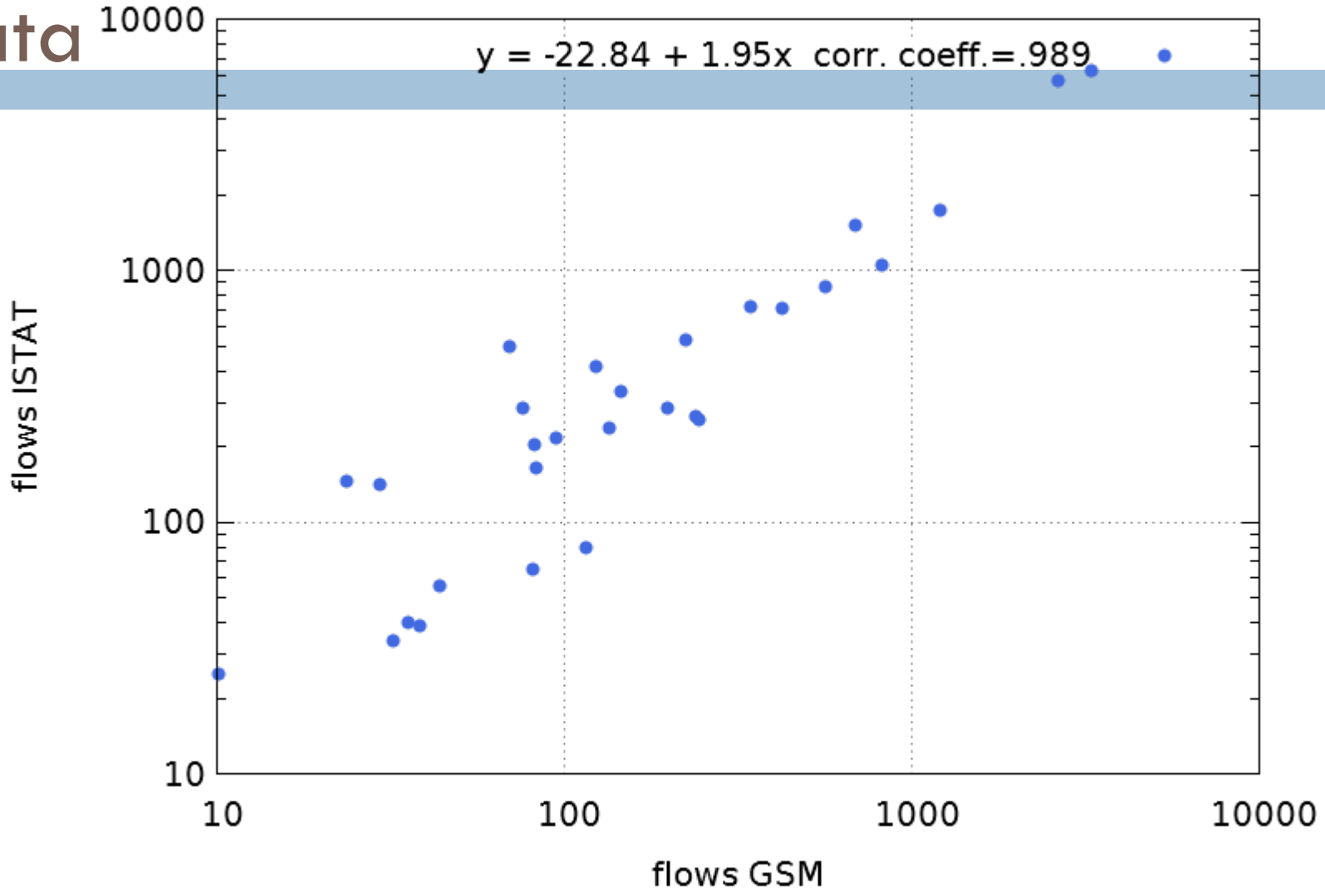
Commuters - Validation with administrative data





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).

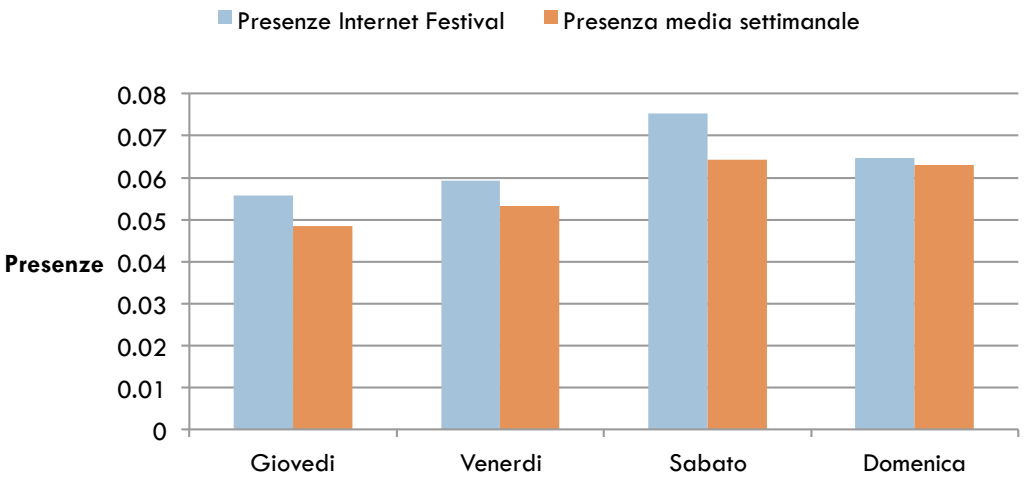
Inter-city flows – validation with ISTAT data



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).

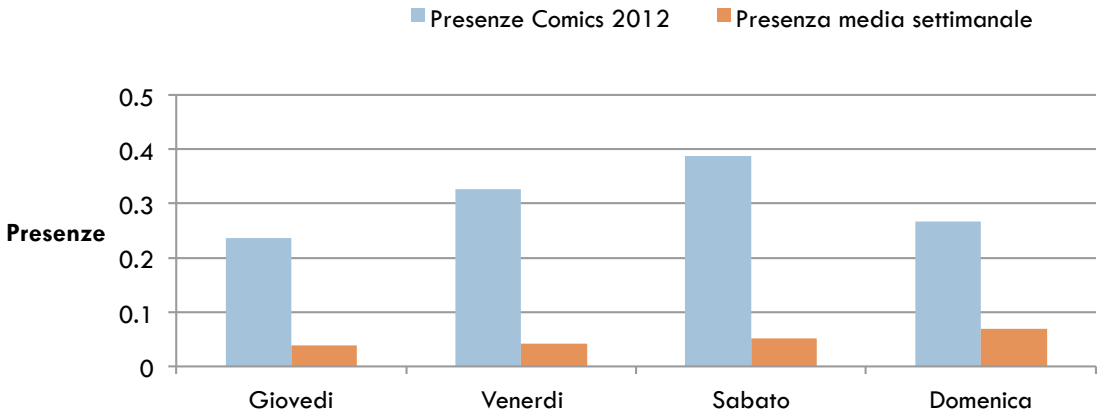
Measuring PRESENCE during events

Presence of Visitors GSM - Pisa - Historical Center



INTERNET FESTIVAL 2013
FORME DI FUTURO

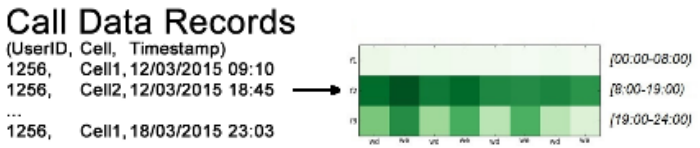
Presence of Visitors GSM - Lucca Comics 2012

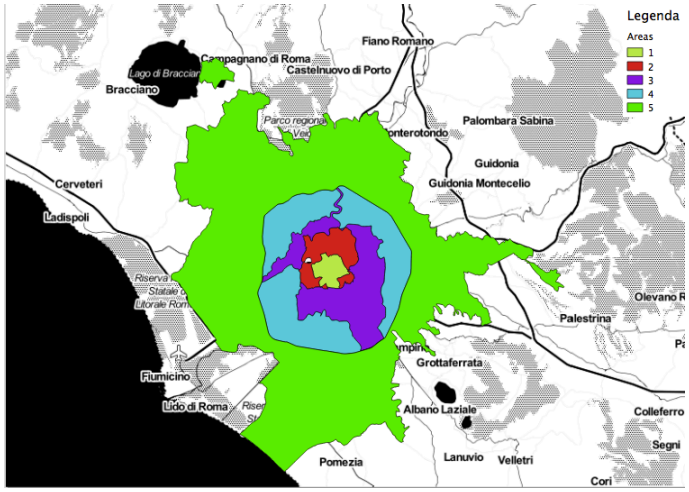


LUCCA COMICS AND GAMES 2012

Measuring exceptional events

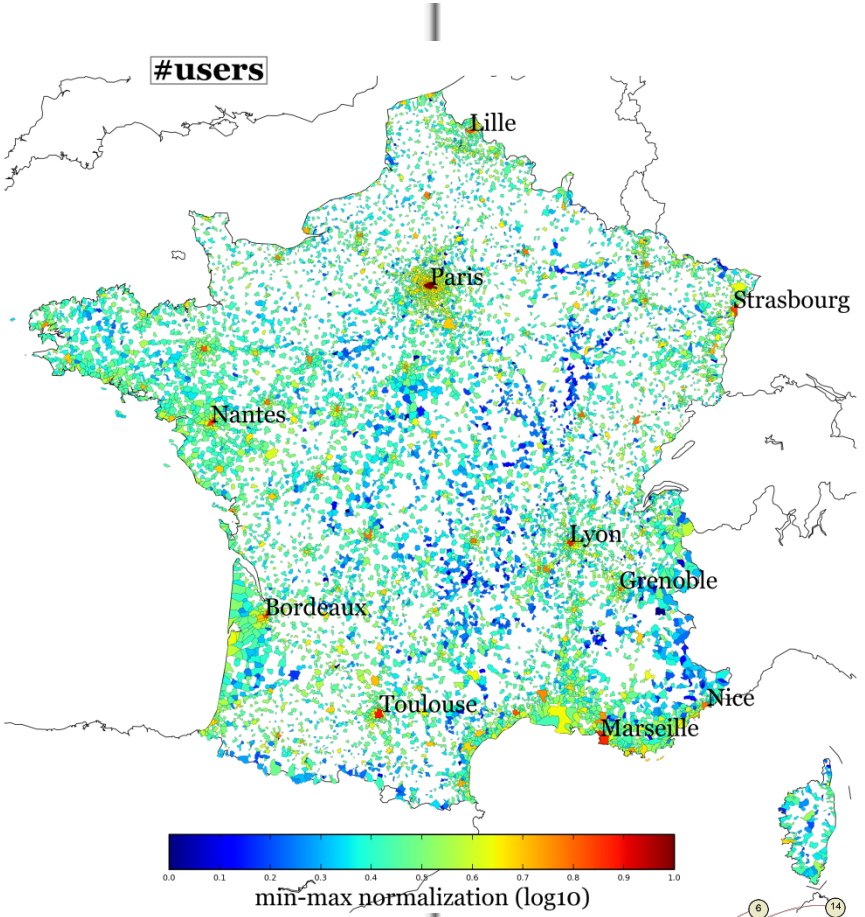
- Presences during Jubilee in Rome (December 2015)
- Continuous monitoring

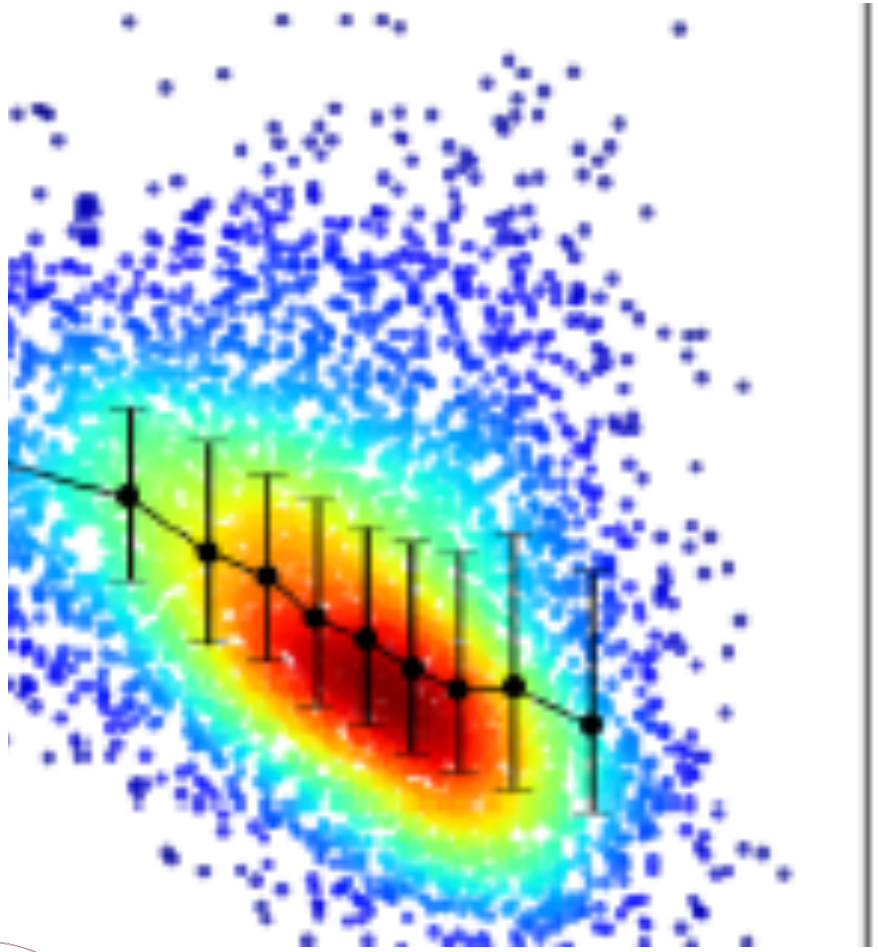


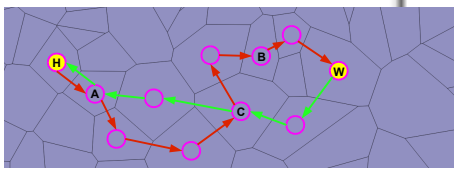


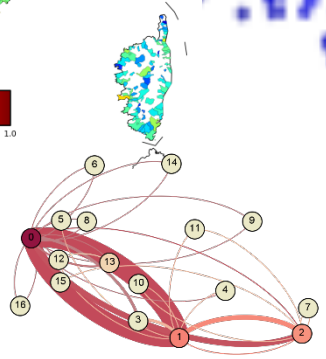
BIG DATA: DIVERSITY AND ECONOMIC DEVELOPMENT

Mobility Diversity and Wellbeing





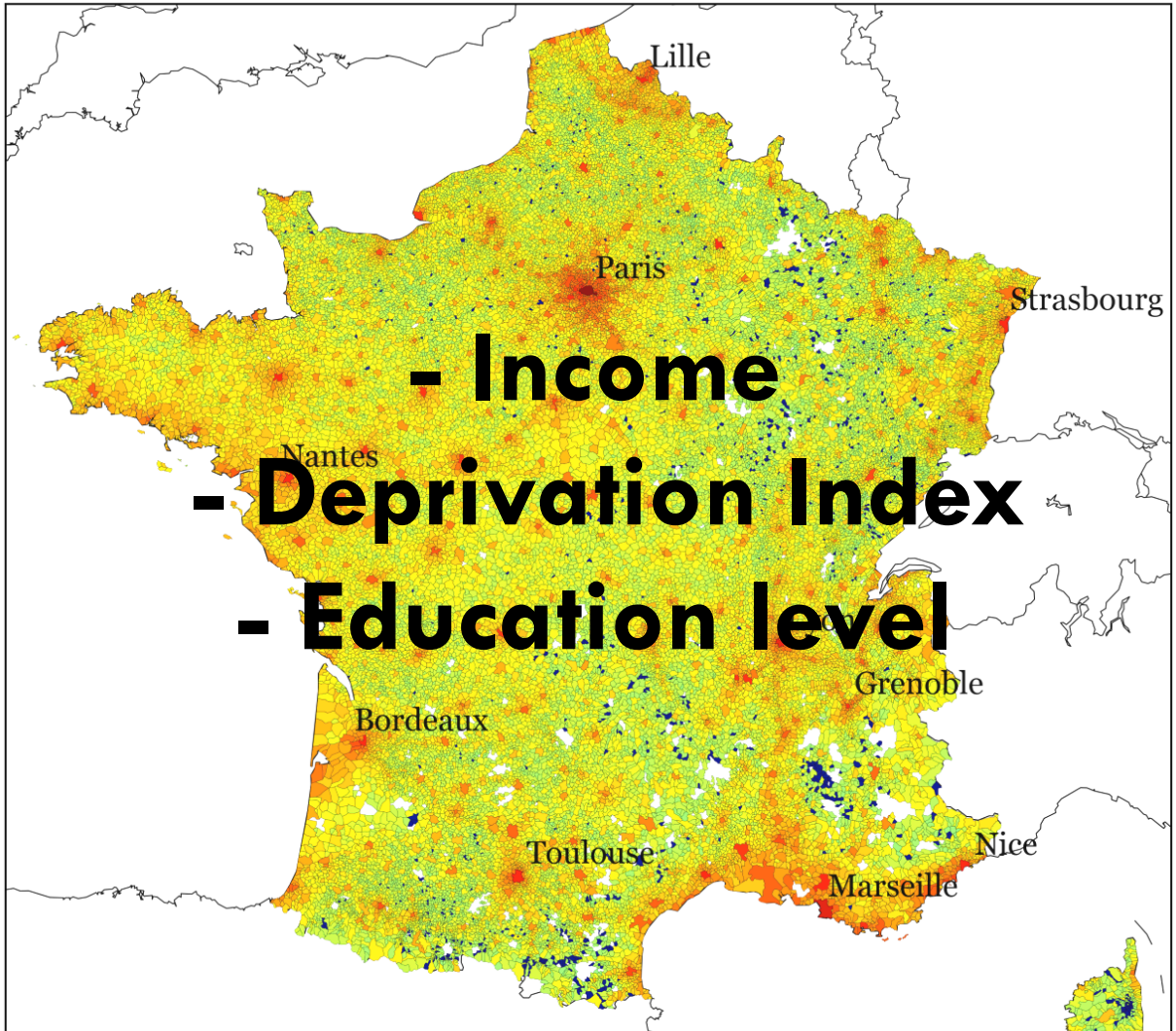




$$-\sum_{T_i' \subset T_i} P(T_i') \log_2 [P(T_i')]$$

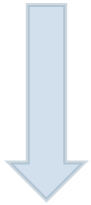
Economic Measures

7,000 French cities



20 million users

200 million calls



user filtering

6 million users

mobility trajs

social network

Four individual measures

- Radius of gyration
- Social degree

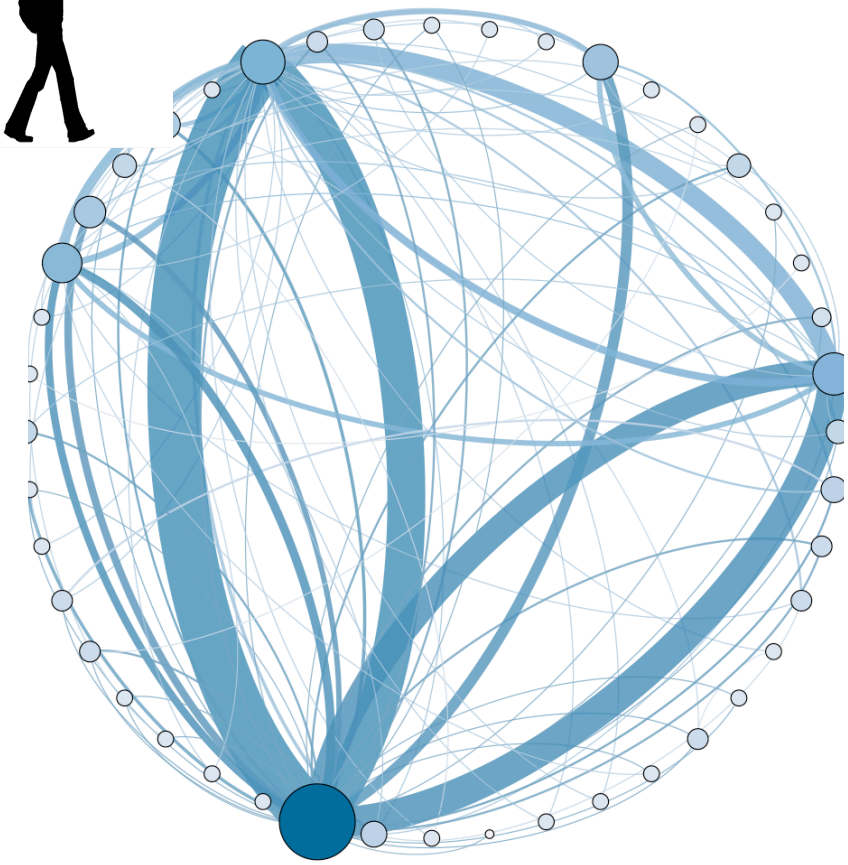
} volume

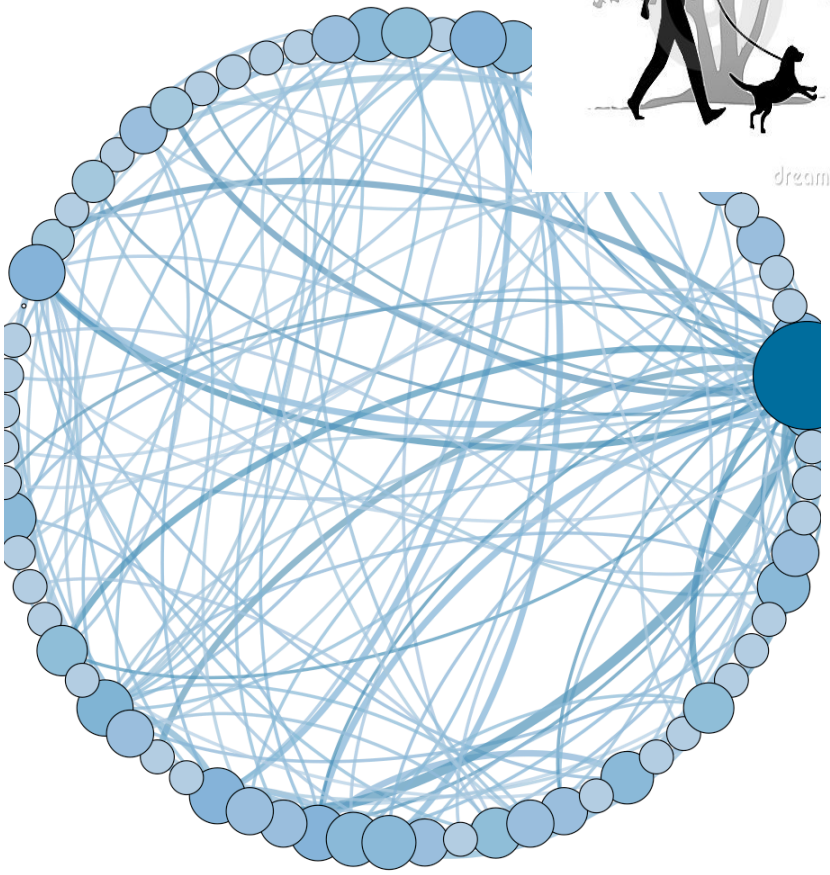
- Mobility entropy
- Social diversity

} diversity

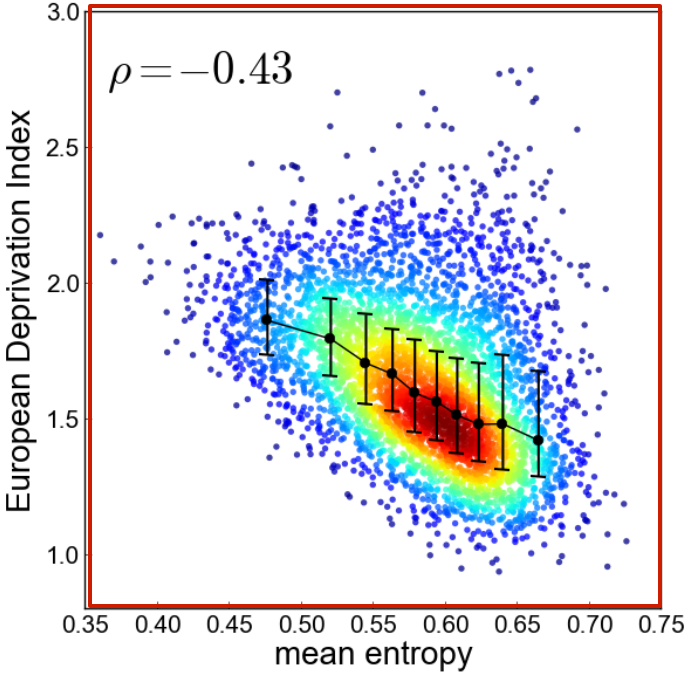
$$-\sum_{T_i' \subset T_i} P(T_i') \log_2 [P(T_i')]$$

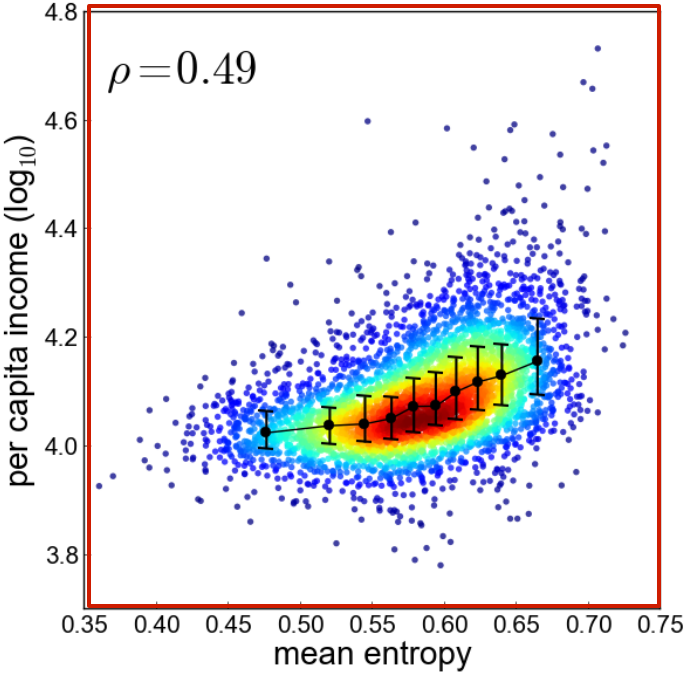
Diversity of individual mobility netw



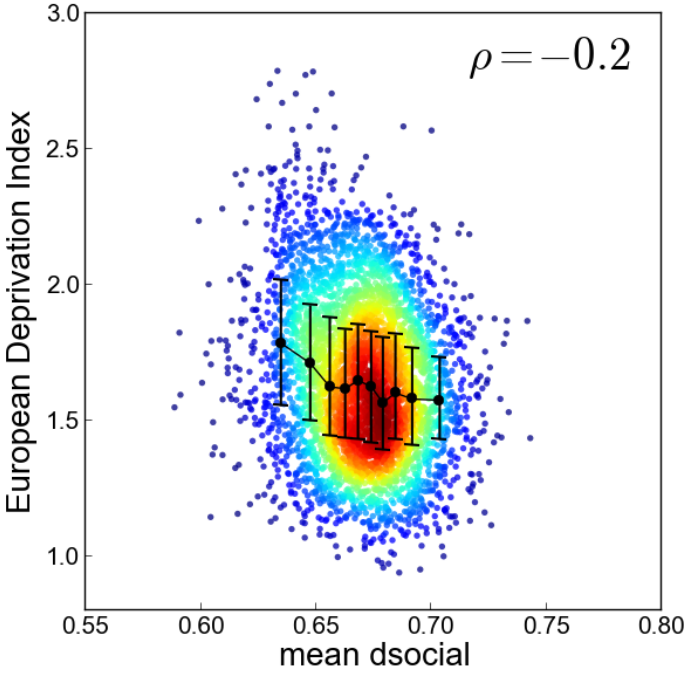


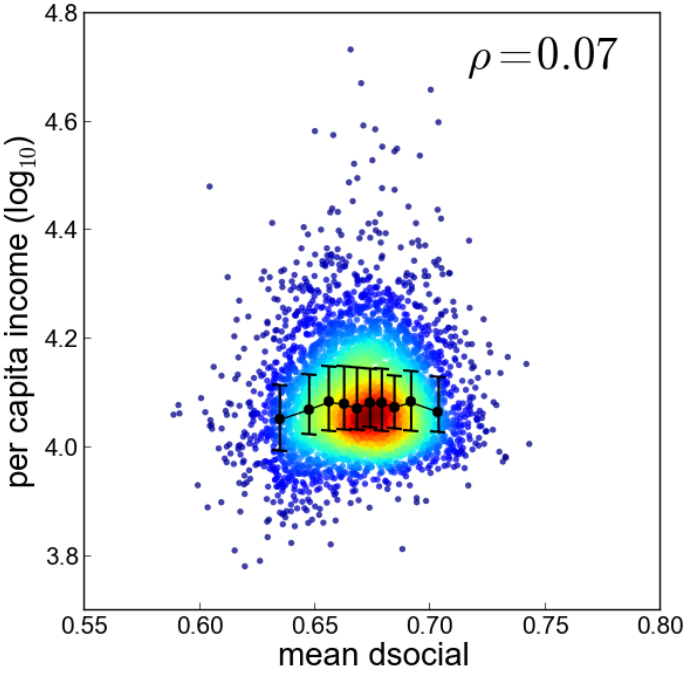
mobility





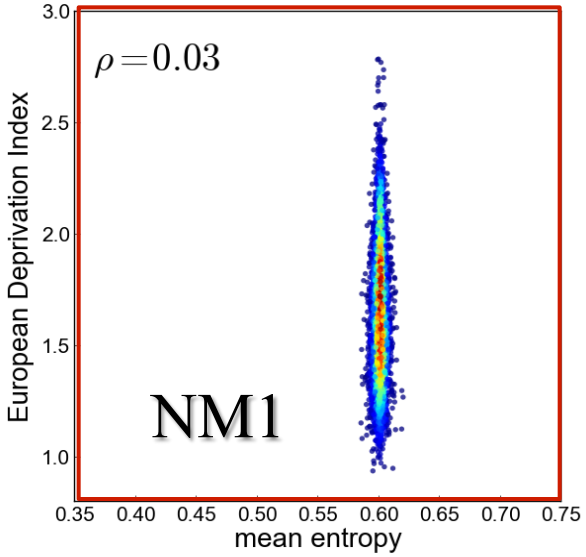
sociality

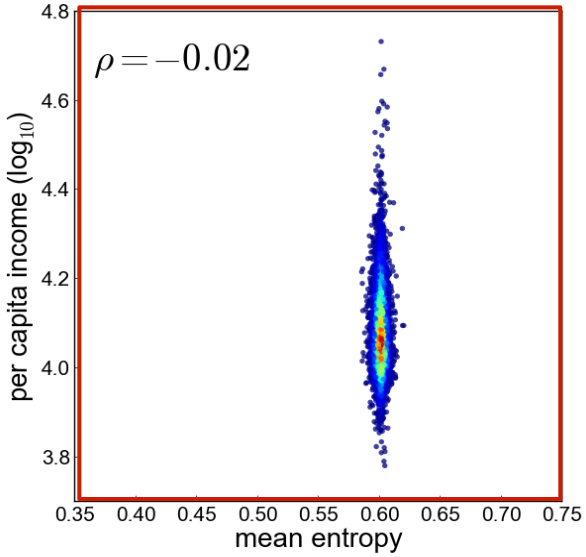




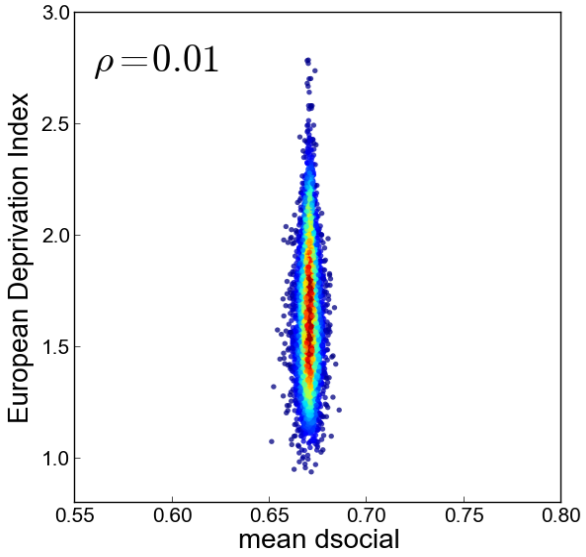
What on a null model: randomizing people

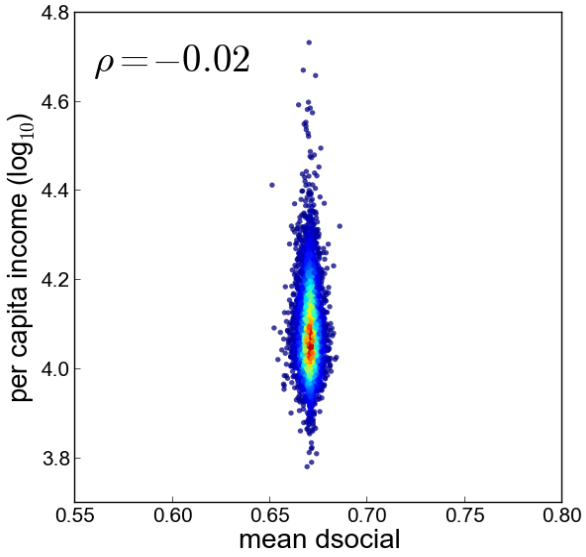
mobility





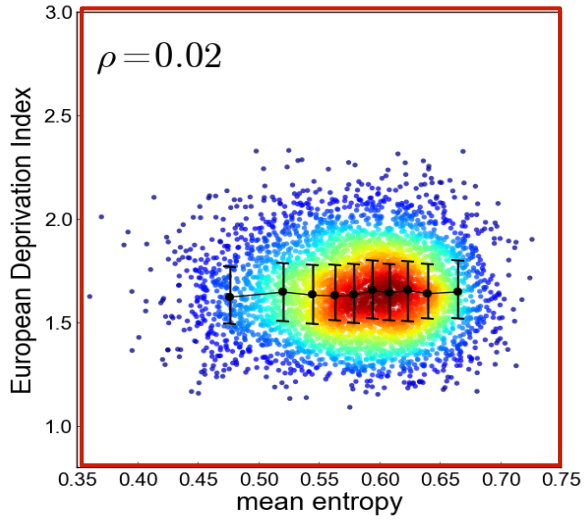
sociality

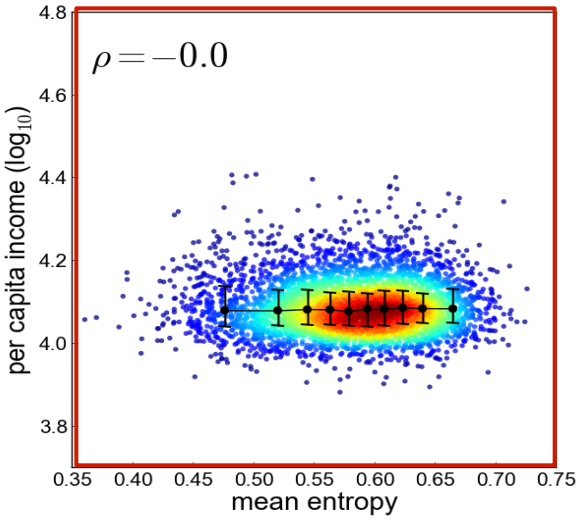




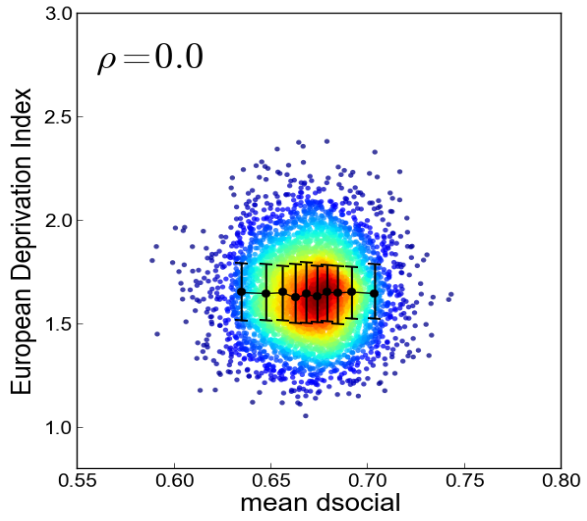
What on a null model: randomizing EDI

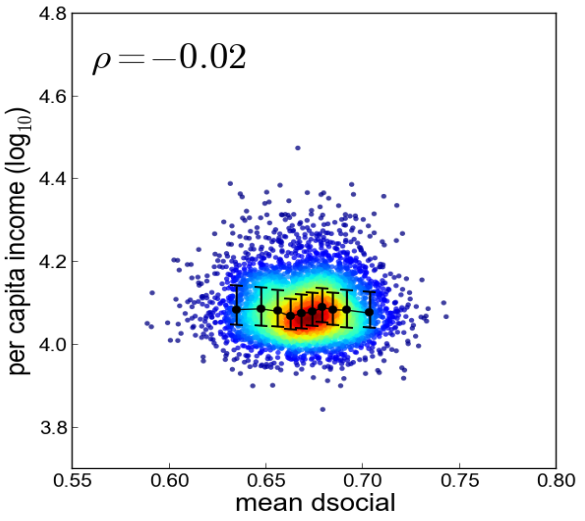
mobility



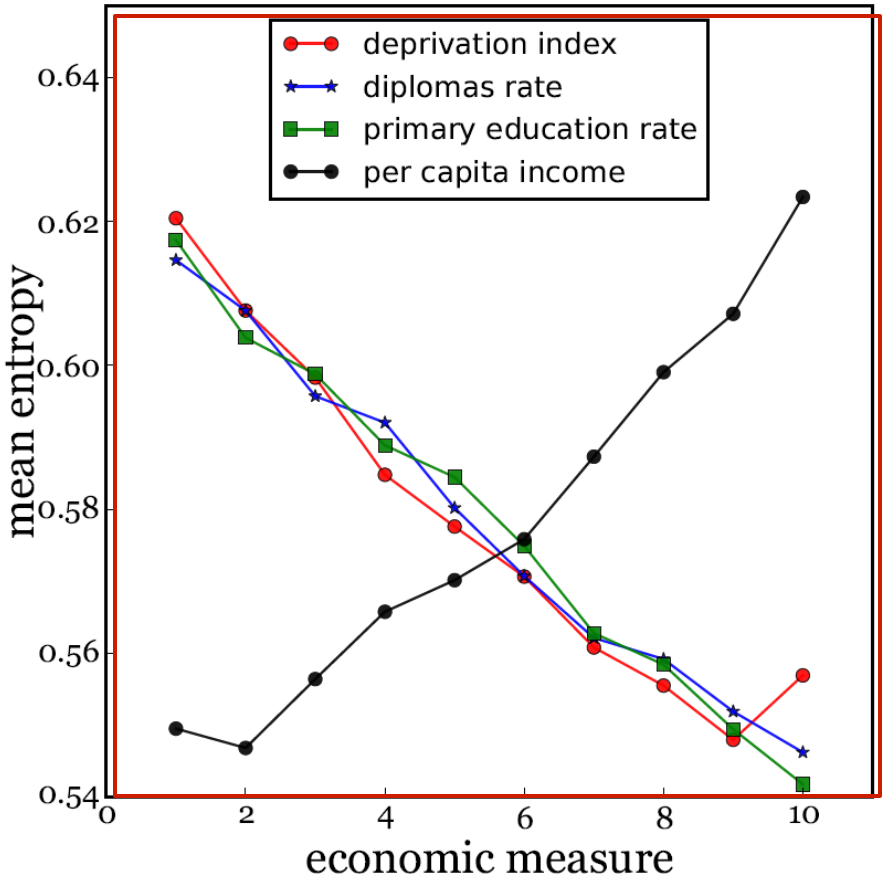


sociality

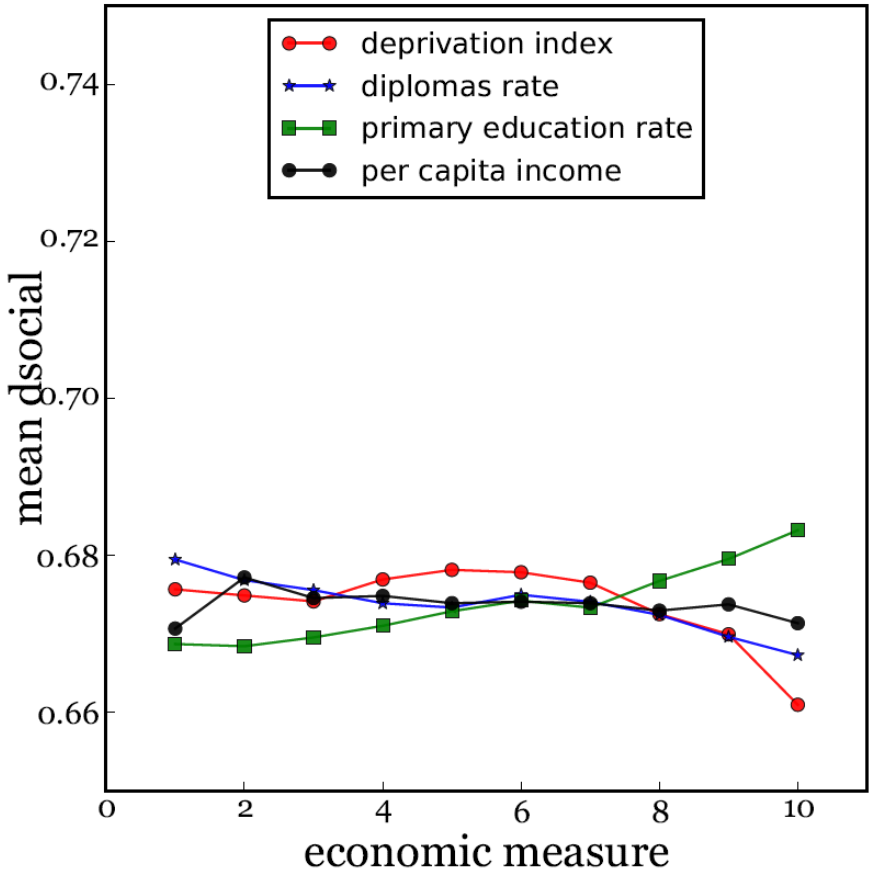




mobility

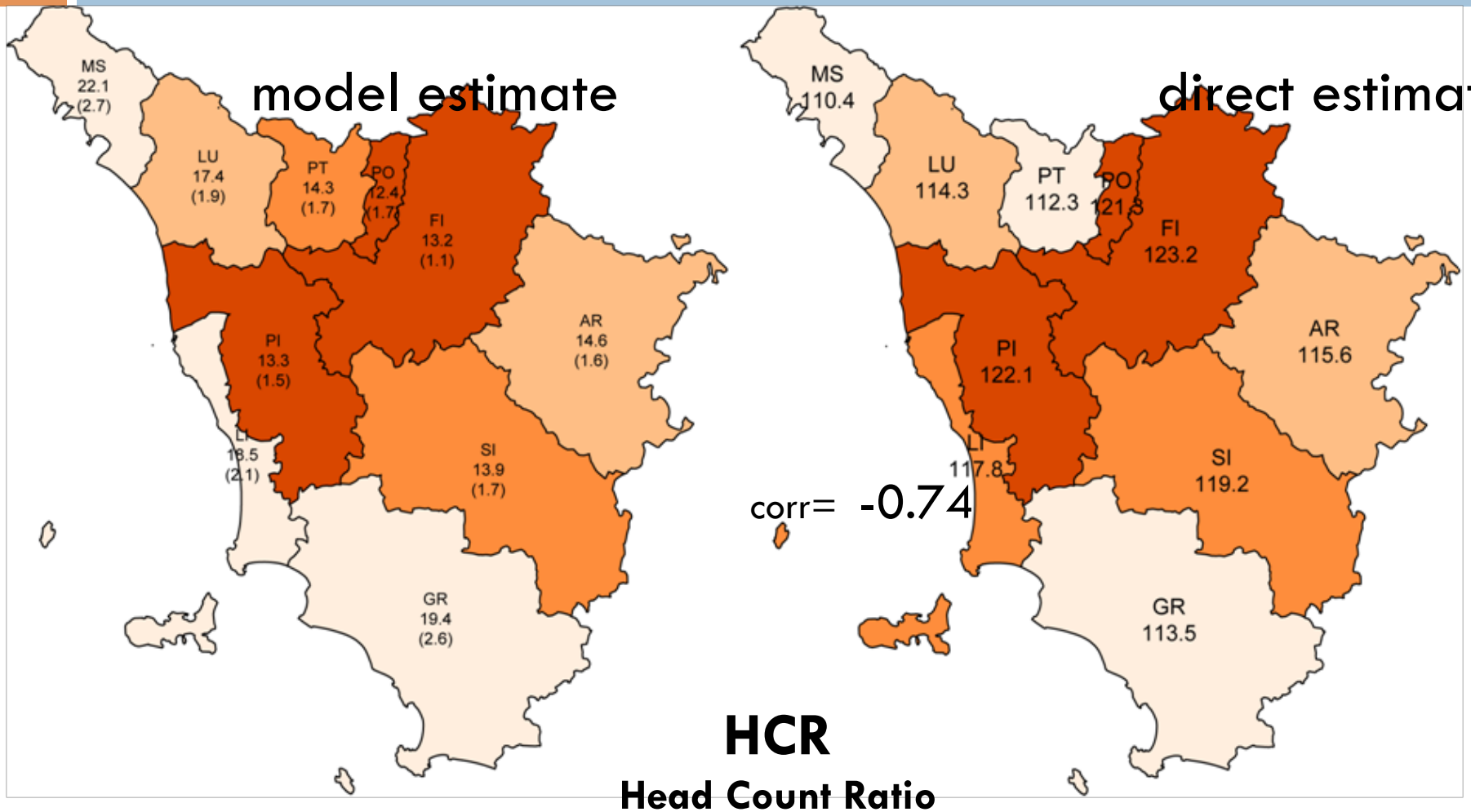


sociality



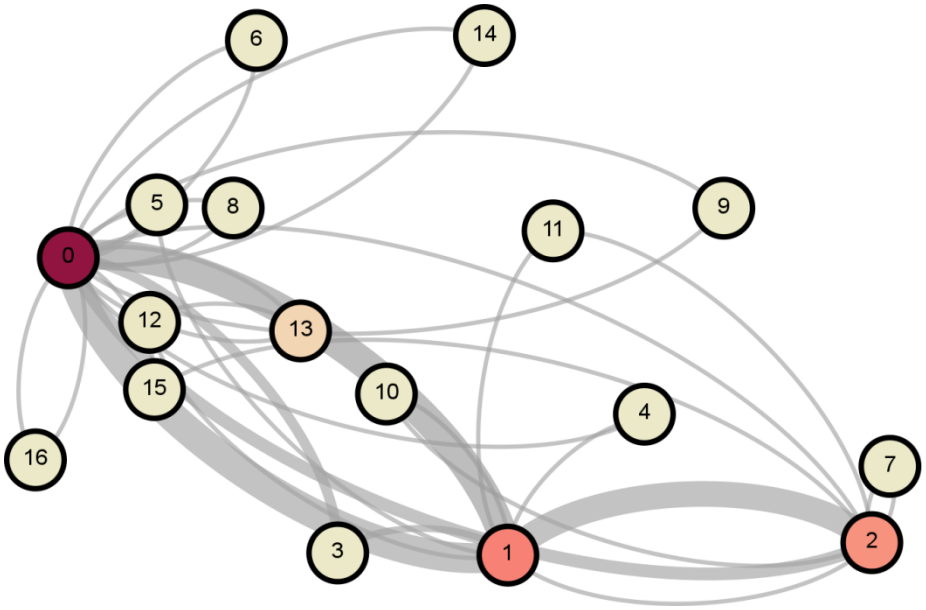
Human Mobility, Social Networks and Economic Development, L. Pappalardo, M. Vanhoof,

…also in Tuscany



Giusti,Marchetti, Pratesi, Salvati, D. Pedreschi, F. Giannotti, Rinzivillo, Pappalardo,Gabrielli. **Small area model based estimators usign Big Data Sources**. Journal of Official Statistics, vol. 31(2) 2015.

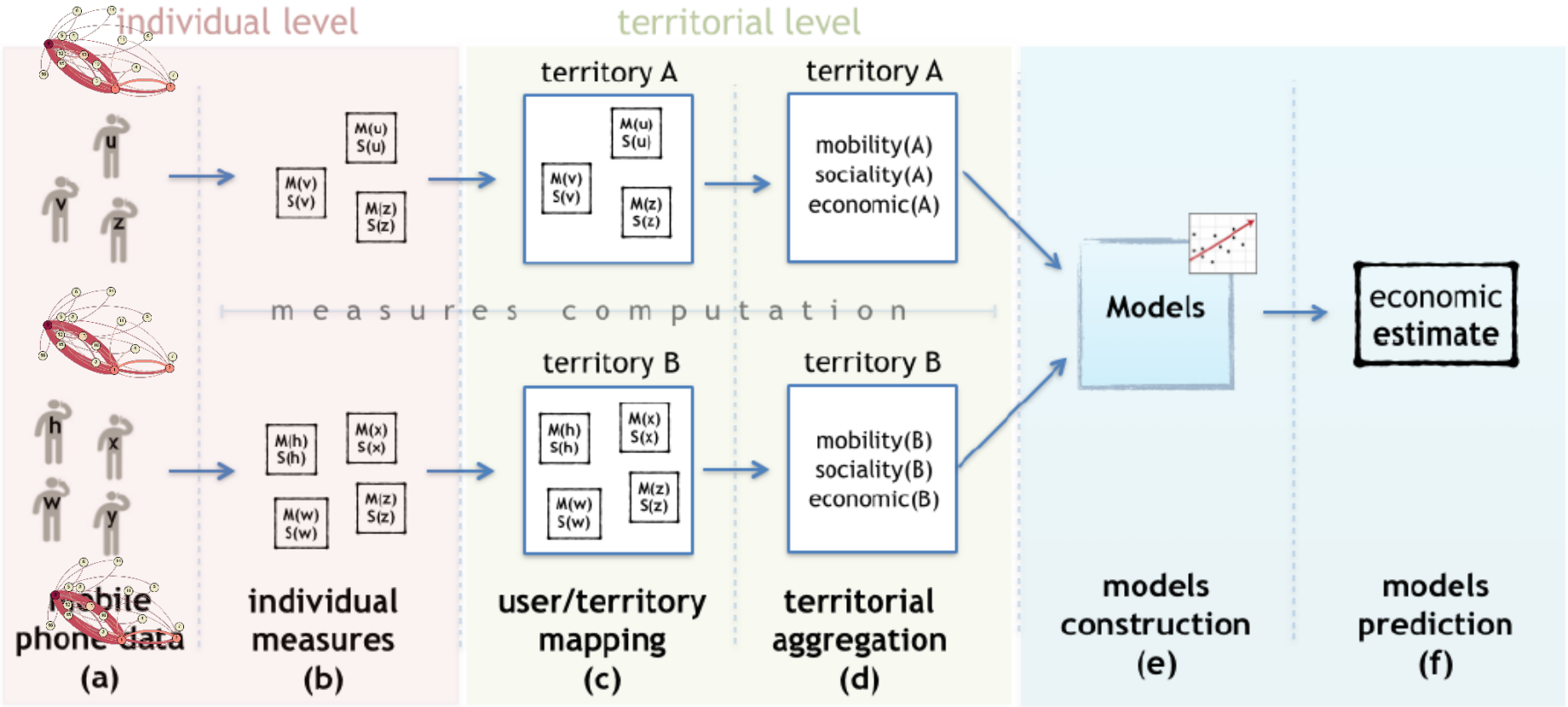
Behind the scene: Individual Mobility Networks



Trip Features
Length
Duration
Time Interval
Average Speed

Network Features	
centrality	clustering coefficient average path length
predictability	entropy
hubbiness	degree betweenness
volume	edge weight flow per location

Behind the scene



L. Pappalardo, Van Hoof, L. Gabrielli, D. Pedreschi, F. Giannotti, Z. Smoreda. **Mobility Diversity & Wellbeing: estimating economic development with mobile phone data**. Submitted

Discussion

1. Mobility diversity is linked to wellbeing
2. Entropy is stable across age/gender but varies with wellbeing
3. Geography matters

4. Big Data as a pillar for official statistics

Possible project

- Suppose to have CDR for an entire nation or all Europe continuosly available or any other source you think is useful
  - Migration fluxes: how to build migration (in-out) indicator and a monitoring system for fluxes among regions at various scale. (are there other resources to be used)
  - Same for turistic fluxes