UNVEILING THE COMPLEXITY OF HUMAN MOBILITY BY MINING & QUERYING MASSIVE TRAJECTORY DATA

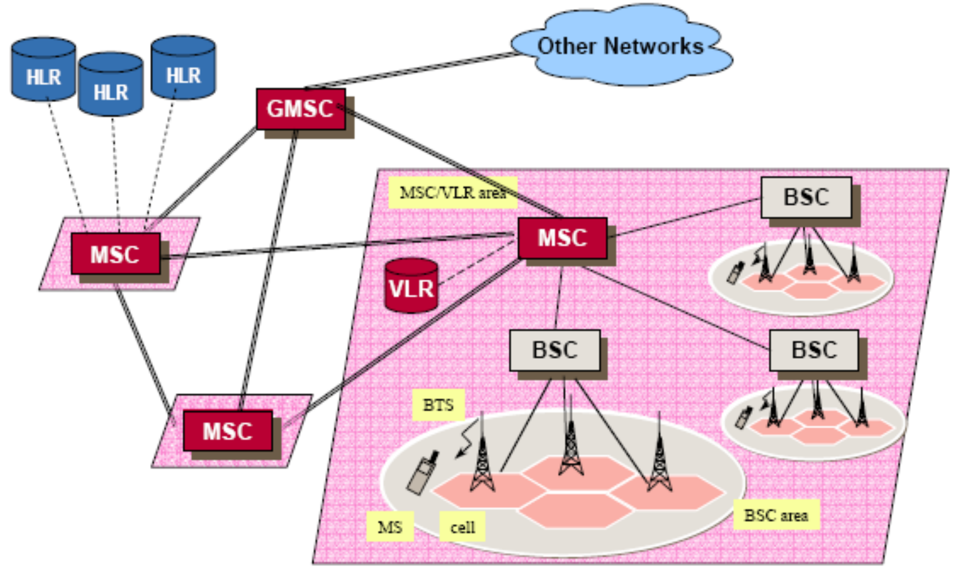
Fosca Giannotti

Knowledge Discovery & Data Mining LAB ISTI-CNR & Università di Pisa

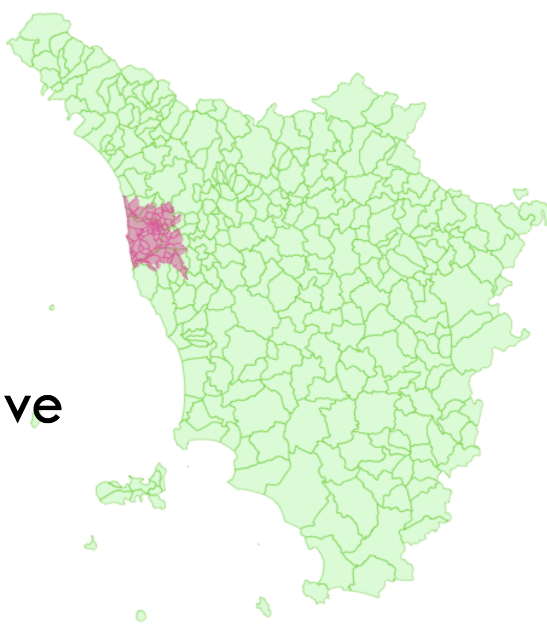
http://kdd.isti.cnr.it

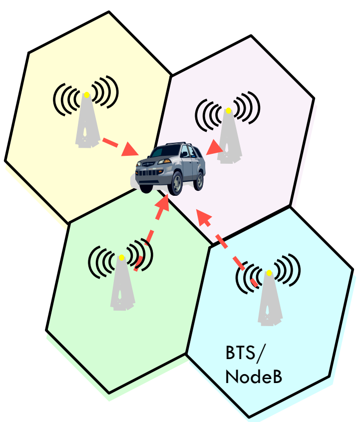
BIG DATA as a proxy of human mobility

GSM data



- Mobile Cellular Networks handle information about the positioning of mobile terminals
  - CDR Call Data Records: call logs (tower position, time, duration,..)
  - Handover data: time of tower transition
- More sophisticated Network Measurement allow tracking of all active (calling) handsets



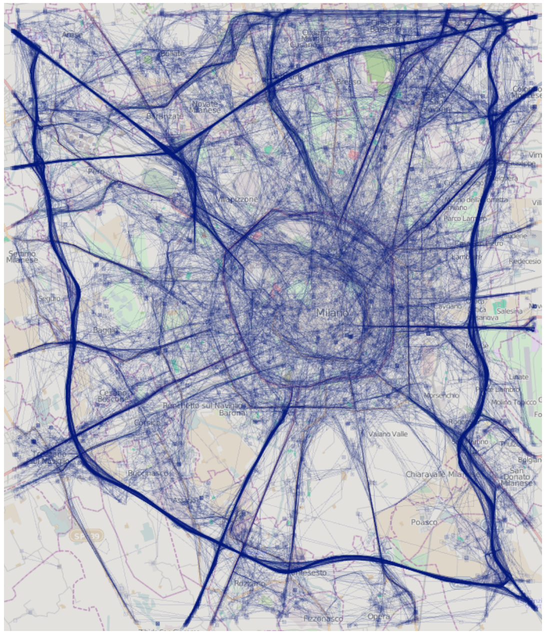


GSM data as a proxy of presence and fluxes

Video: Paris_splines.avi

GPS tracks

- Onboard navigation devices send GPS tracks to central servers
  - Sampling rate ~3 secs
  - Spatial precision ~ 10 m



```
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
…
```

GPS: detailed movements within an area

Video: moves_viz_prov_cut.mov

GPS: movements within the town

Video: moves_viz_city_cut.mov

Social Networks: goal of movements

Video: flickr_cut.mov

Plan of the presentation

- Mastering the overall KDD process
  - M-atlas platform
- Exemplar case studies
  - Advanced OD Matrix browsing
  - Understanding collective patterns
  - Understanding Individual profiles
  - Putting interactions in the game

Mastering the overall KDD process: M-Atlas platform

Fosca Giannotti · Mirco Nanni · Dino Pedreschi · Fabio Pinelli · Chiara Renso · Salvatore Rinzivillo · Roberto Trasarti

Unveiling the complexity of human mobility by querying and mining massive trajectory data

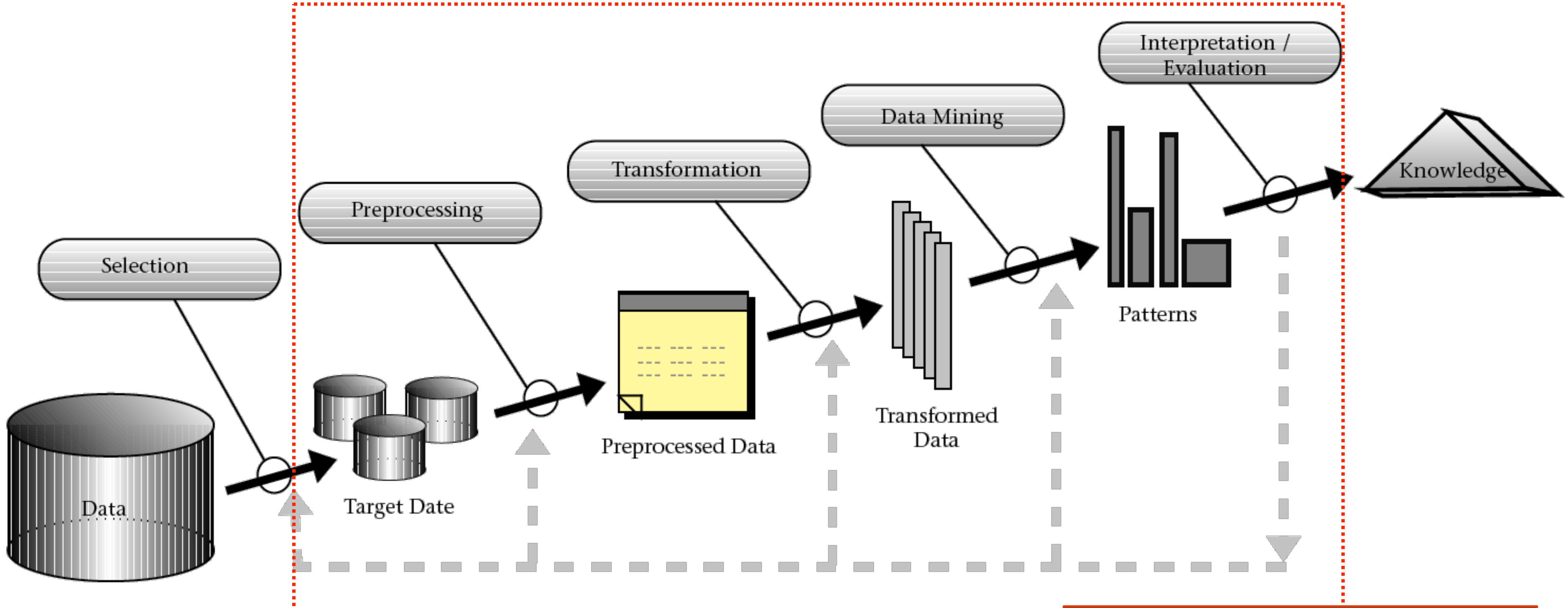
The VLDB Journal, 2011

Roberto Trasarti, Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Chiara Renso.

A Query Language for Mobility Data Mining.

International Journal of Data Warehousing and Mining (IJDWM) 2010

Knowledge Discovery process



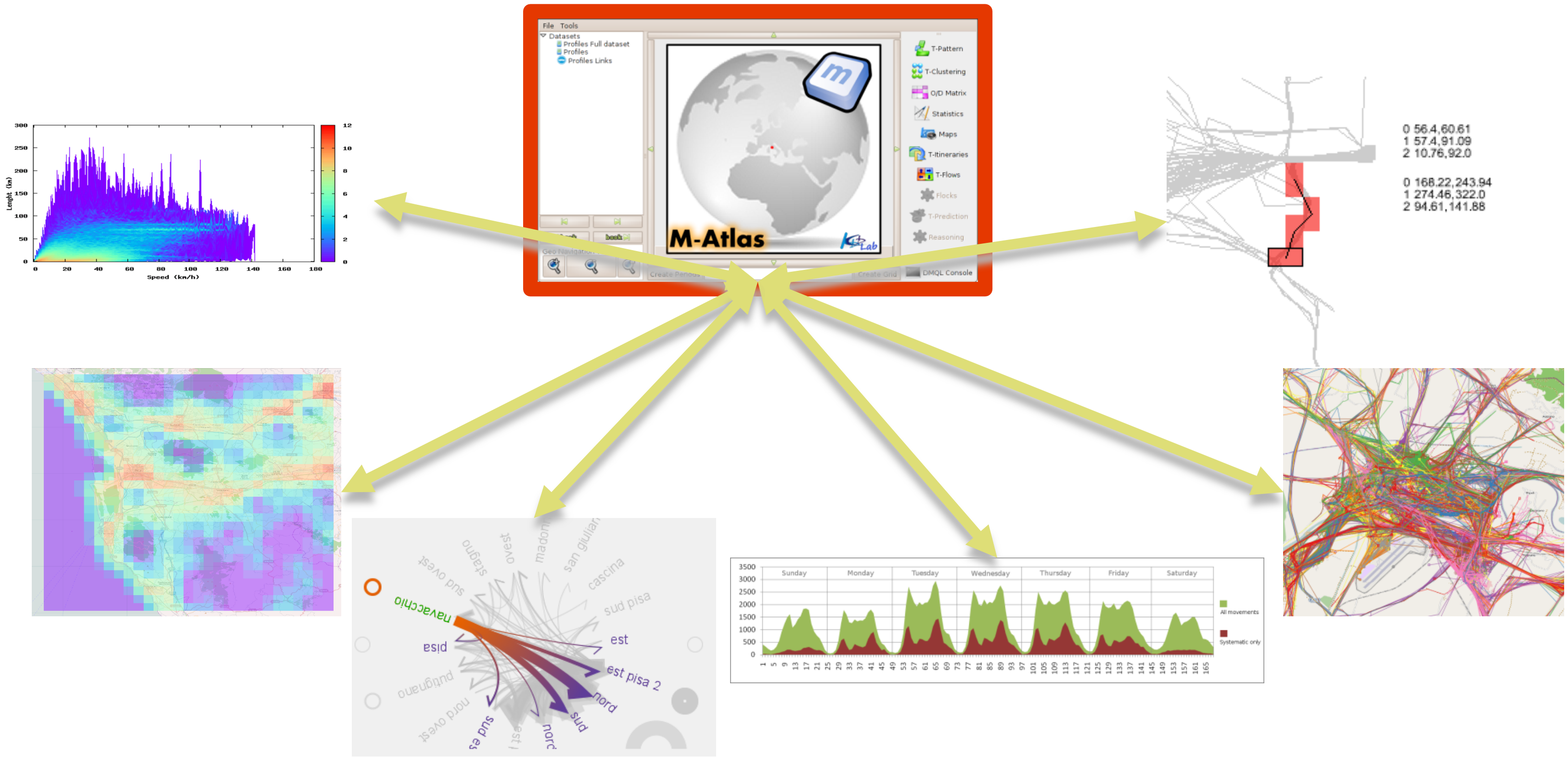
M-Atlas platform

M-Atlas: An analytical system to create and navigate an atlas of urban mobility

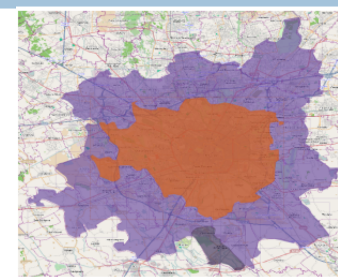
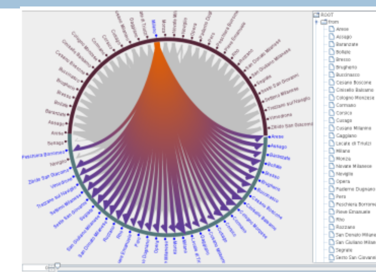
Source data: GPS, GSM, Sensors, Rfid, spatial data

M-Atlas platform

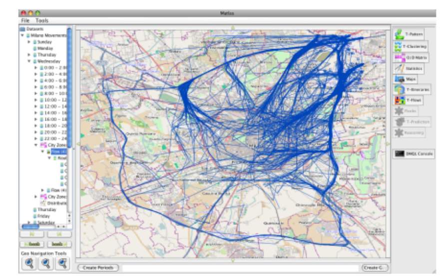
A tool kit to extract, store, combine different kinds of models to build mobility knowledge discovery processes.



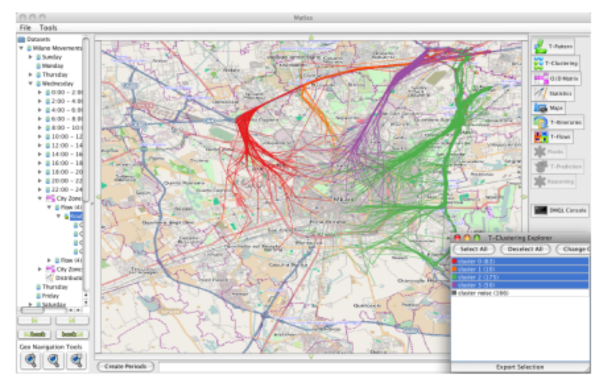
DMQL EXPRESSIVENESS:
How do people leave the city toward suburban areas?

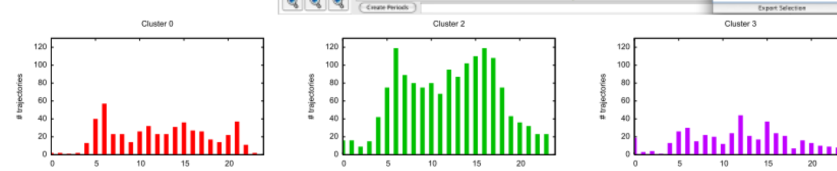
CREATE MODEL MilanODMatrix AS MINE ODMATRIX
FROM (SELECT t.id, t.trajectory FROM TrajectoryTable t),
(SELECT orig.id, orig.area FROM MunicipalityTable orig),
(SELECT dest.id, dest.area FROM MunicipalityTable dest)



CREATE RELATION CenterToNESuburbTrajectories USING ENTAIL
FROM (SELECT t.id, t.trajectory FROM TrajectoryTable t, MilanODMatrix m
WHERE m.origin = Milan AND
m.destination IN (Monza, …, Brugherio))



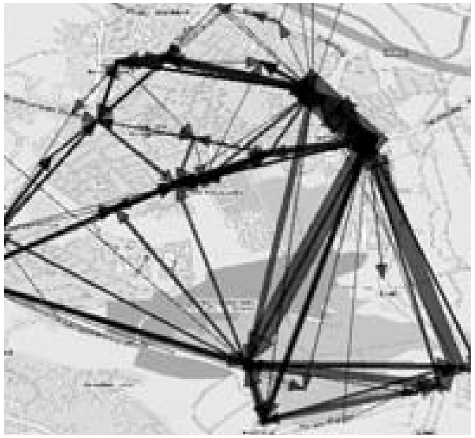
CREATE MODEL ClusteringTable AS MINE T-CLUSTERING
FROM (Select t.id, t.trajectory from CenterToNESuburbTrajectories t)
SET T-CLUSTERING.FUNCTION = ROUTE_SIMILARITY AND
T-CLUSTERING.EPS = 400 AND
T-CLUSTERING.MIN_PTS = 5



A DataWarehouse for OD Matrix

OD Matrix

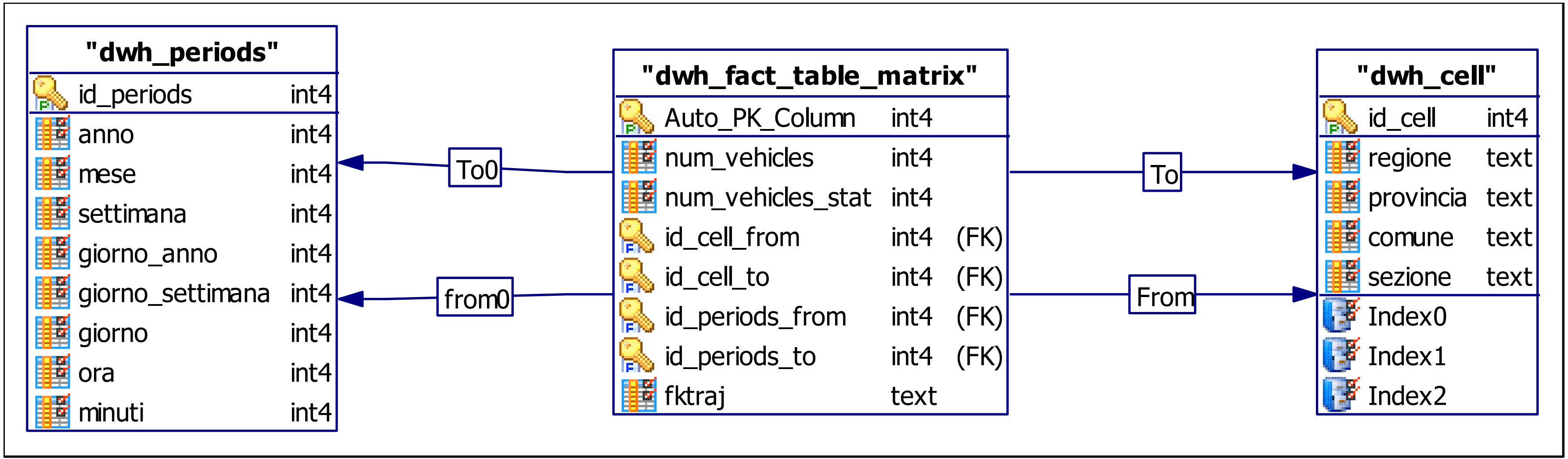
- Model mobility demand by measuring the flows among different areas
- General approach
  - Spatial grid with relevant zones
  - (Estimated) flows of movement from origin to destination



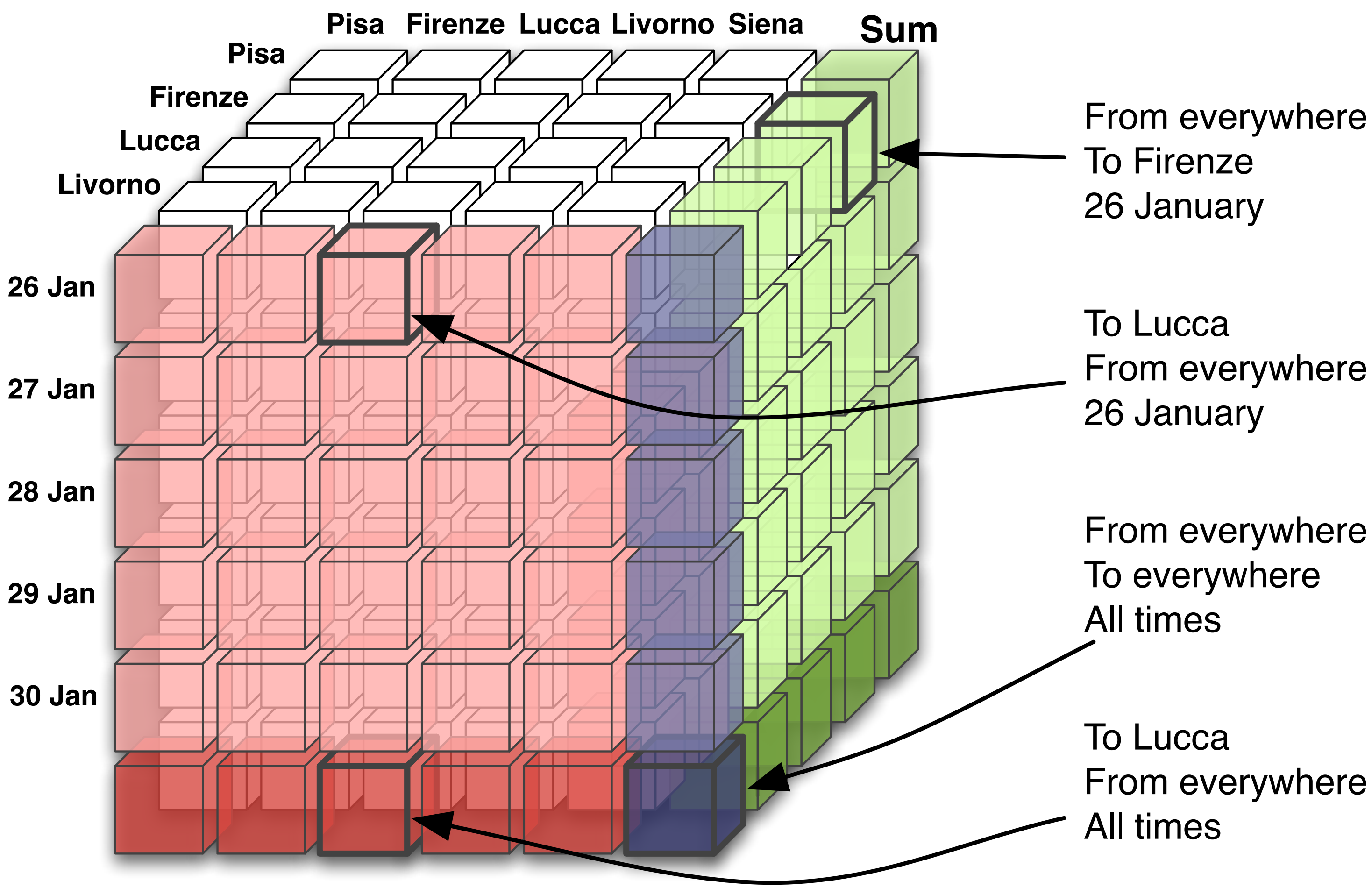
OD Matrix exploration

- OD Matrix should answer the questions
  - From which region?
  - To which region?
  - When?
  - How many?

- DW Concepts
  - Facts: basic observation
    - Aggregated movements from an origin to a destination
  - Dimensions
    - Origins
    - Destinations
    - Time
  - Measures
    - Count
    - Ratio over total



OD Matrix: DW design



The general process

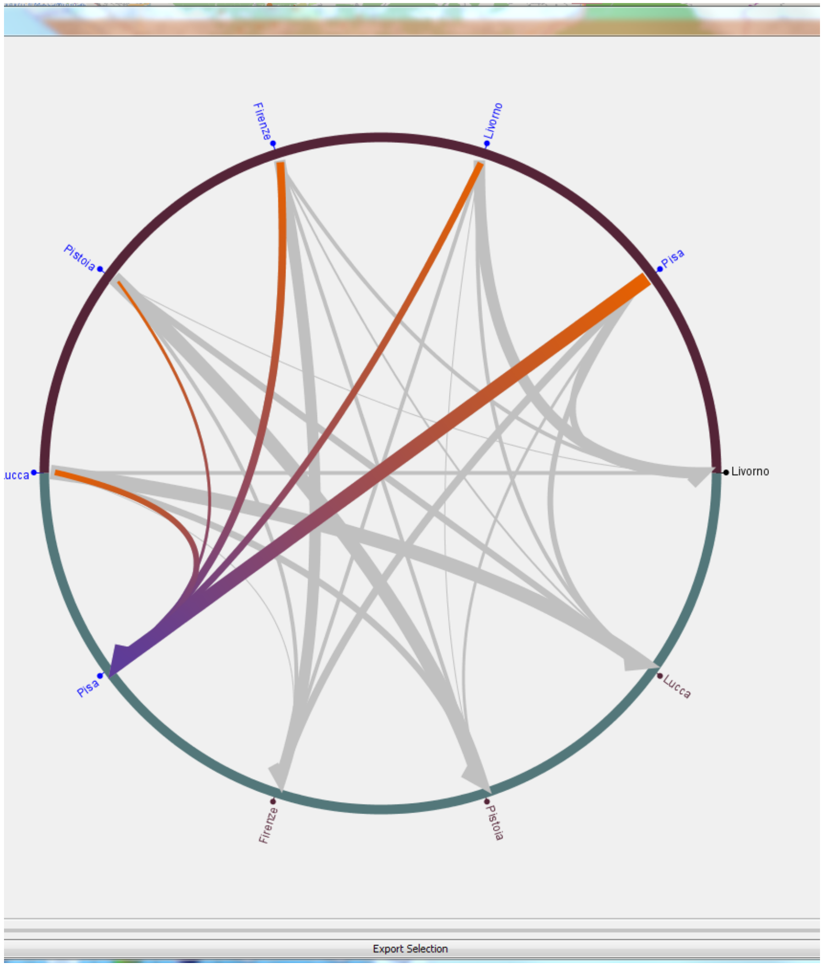
1. Browse DWH at high spatial level (provinces)
2. Identify interesting flows and drill down to specific flow
3. Navigate the cube at finer level (cities)
4. From flow to trajectories: entails original data
5. Do specific analysis on the real trips

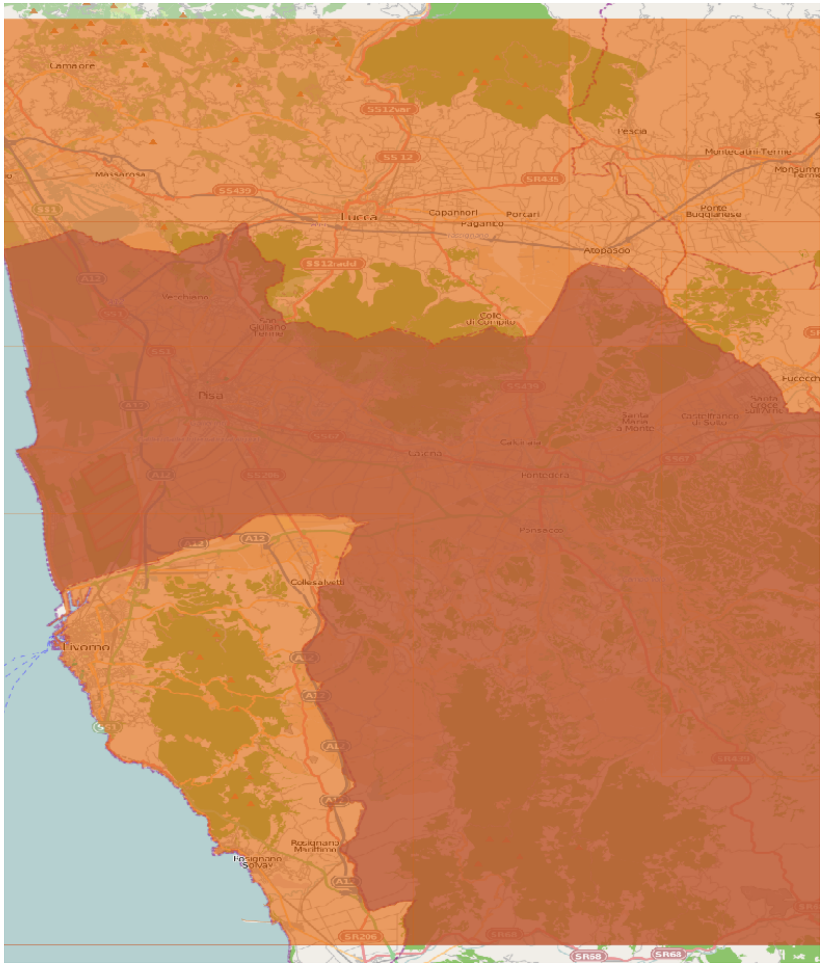
Navigate the cube at higher spatial level (provinces): pivot table

Time to	Cell to		Cell from		Measures	
(All)	Regione	Provincia	Regione	Provincia	Numero Veicoli	Perc
All Time To	Toscana	Pisa	Toscana	Pisa	462.583	86,53%
				Firenze	27.742	13,26%
				Livorno	20.429	10,05%
				Lucca	17.681	04,07%
				Pistoia	5.727	01,30%
		Pistoia	Toscana	Pistoia	405.003	92,05%
				Lucca	19.040	04,38%
				Firenze	7.853	03,75%
				Pisa	5.630	01,05%
				Livorno	2.306	01,13%
		Lucca	Toscana	Lucca	388.854	89,42%
				Pistoia	19.268	04,38%
				Pisa	17.750	03,32%
				Livorno	6.488	03,19%
				Firenze	2.747	01,31%
		Firenze	Toscana	Firenze	163.845	78,34%
				Pisa	27.571	05,16%
				Pistoia	7.769	01,77%
				Livorno	6.617	03,26%
				Lucca	2.650	00,61%
		Livorno	Toscana	Livorno	167.347	82,36%
				Pisa	21.088	03,94%
				Firenze	6.971	03,33%
				Lucca	6.625	01,52%
				Pistoia	2.228	00,51%

- The cube dimensions are flattened by means of a multi-row table
- Example at the province level:
  - How many trips from Lucca province to Pisa province?
  - How many in the other way?

Navigate the cube at higher spatial level (provinces): visual browser





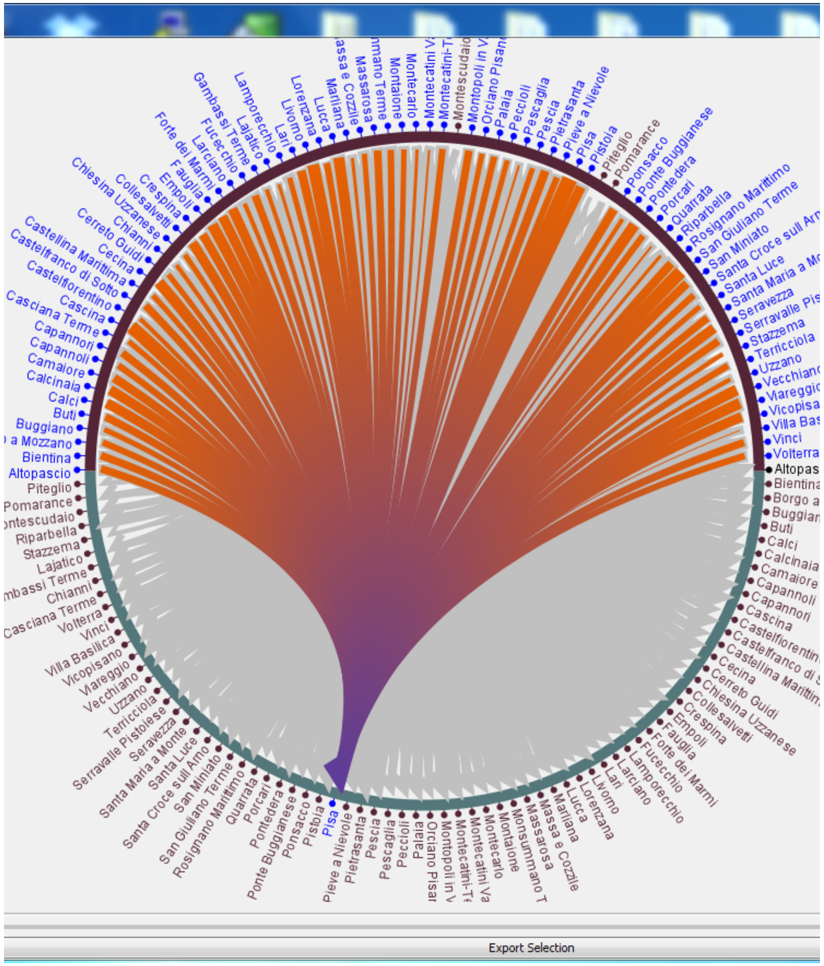
Select origins and destination from the doughnut. The map is linked with the selection. Flow weights are represented by line width

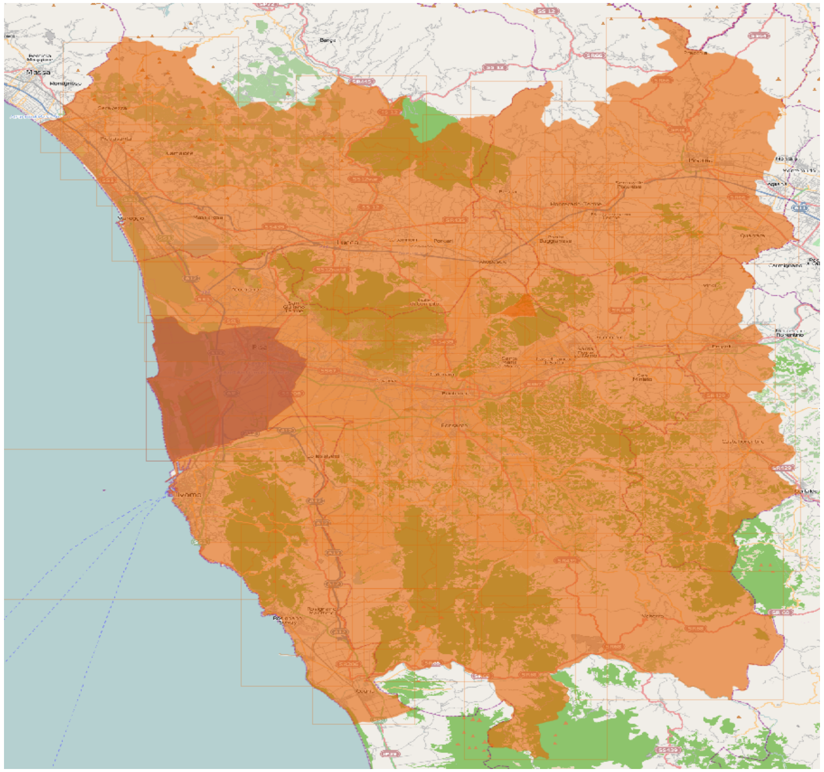
Drill down: from province to single cities

ime to	Cell to			Cell from		Measures	
All)	Regione	Provincia	Comune	Regione	Provincia	Numero Veicoli	Pe
ll Time To	Toscana	Pisa		Toscana	Pisa	462.583	86,53
					Firenze	27.742	13,26
					Livorno	20.429	10,05
					Lucca	17.681	04,07
					Pistoia	5.727	01,30
		Pisa	Pisa	Toscana	Pisa	131.430	28,41
					Livorno	8.066	39,48
					Lucca	7.053	39,89
					Firenze	2.383	08,59
					Pistoia	1.574	27,48
			Cascina	Toscana	Pisa	58.146	12,57
					Livorno	1.777	08,70
					Lucca	1.168	06,61
					Firenze	795	02,87
					Pistoia	305	05,33
			San Miniato	Toscana	Pisa	30.924	06,69
					Firenze	12.018	43,32
					Livorno	459	02,25
					Pistoia	388	06,77
					Lucca	283	01,60
			Pontedera	Toscana	Pisa	37.186	08,04
					Firenze	2.402	08,66
					Livorno	1.180	05,78
					Lucca	611	03,46
					Pistoia	244	04,26
			San Giuliano Terme	Toscana	Pisa	30.331	06,56
					Lucca	1.983	11,22
					Livorno	468	02,29
					Pistoia	345	06,02
					Firenze	124	00,45
			Calcinaia	Toscana	Pisa	18.425	03,98
					Livorno	359	01,76
					Lucca	331	01,87
					Firenze	278	01,00
					Pistoia	194	03,39
			Santa Croce sull Arno	Toscana	Pisa	14.561	03,15
					Firenze	3.893	14,03
					Lucca	347	01,96
					Pistoia	290	05,06
					Livorno	133	00,65
			Vecchiano	Toscana	Pisa	12.931	02,80
					Lucca	2.700	15,27
					Pistoia	1.032	18,02
					Livorno	388	01,90
					Firenze	79	00,28

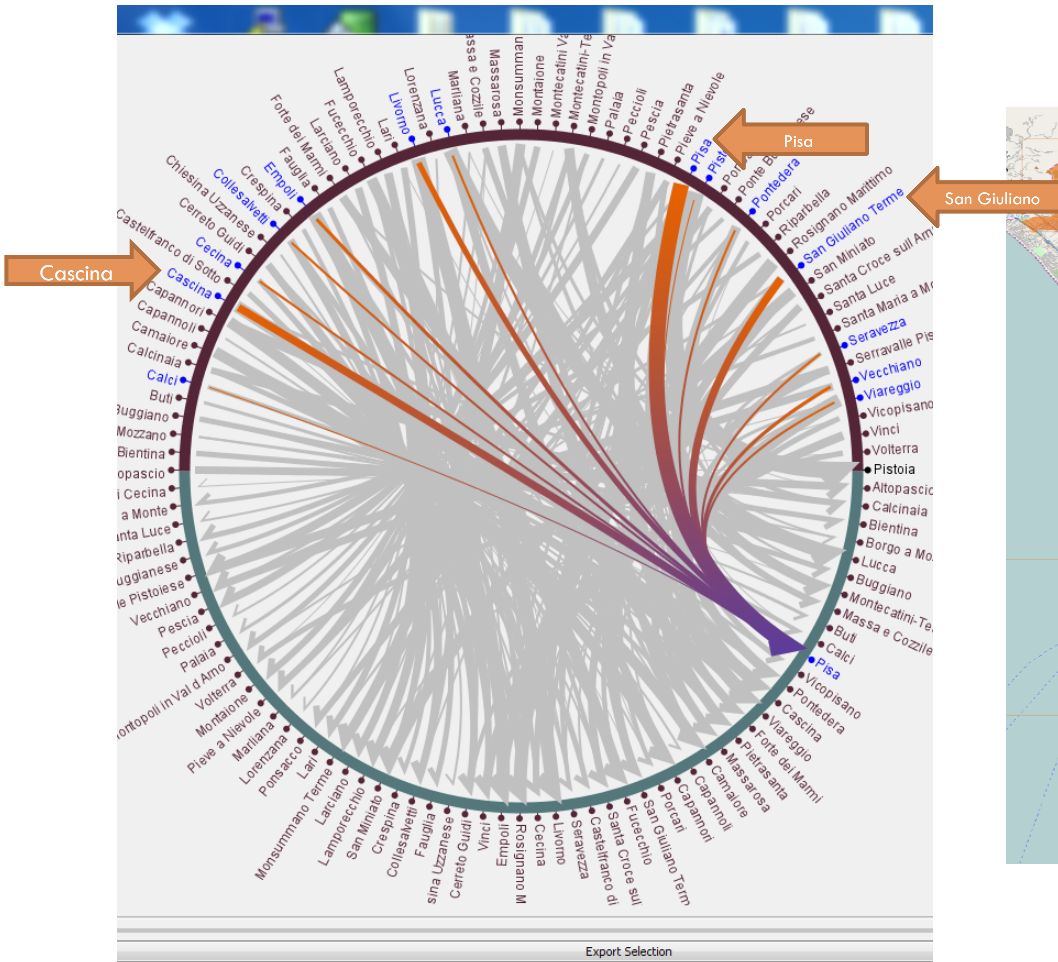
- Explode the destination by specific cities
  - Easy to identify the cities with the higher incomin traffic
  - For each city it is possible to identify the source of traffic

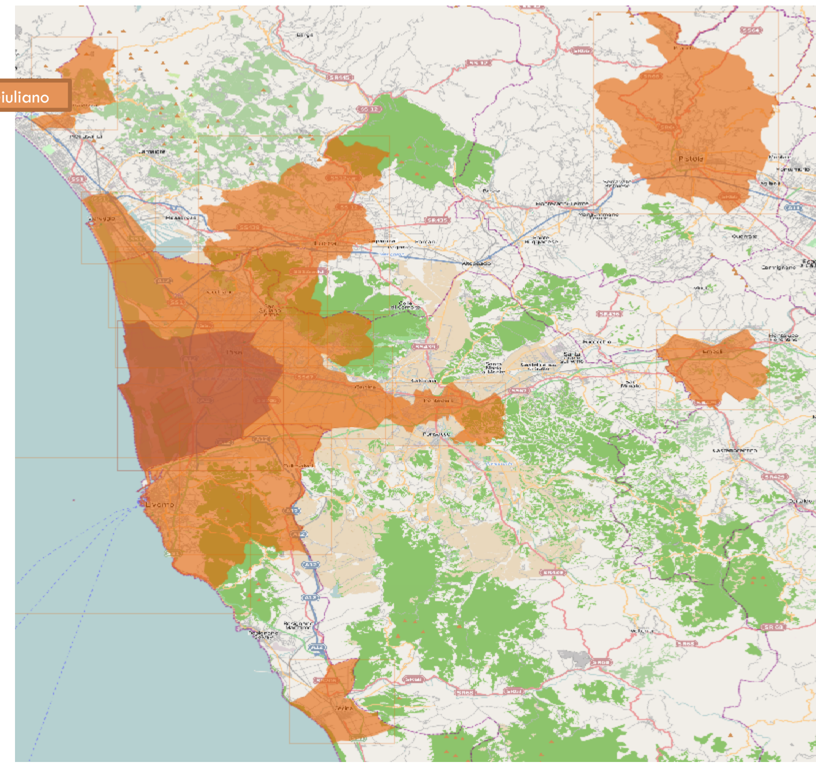
Drill down: from cities to cities





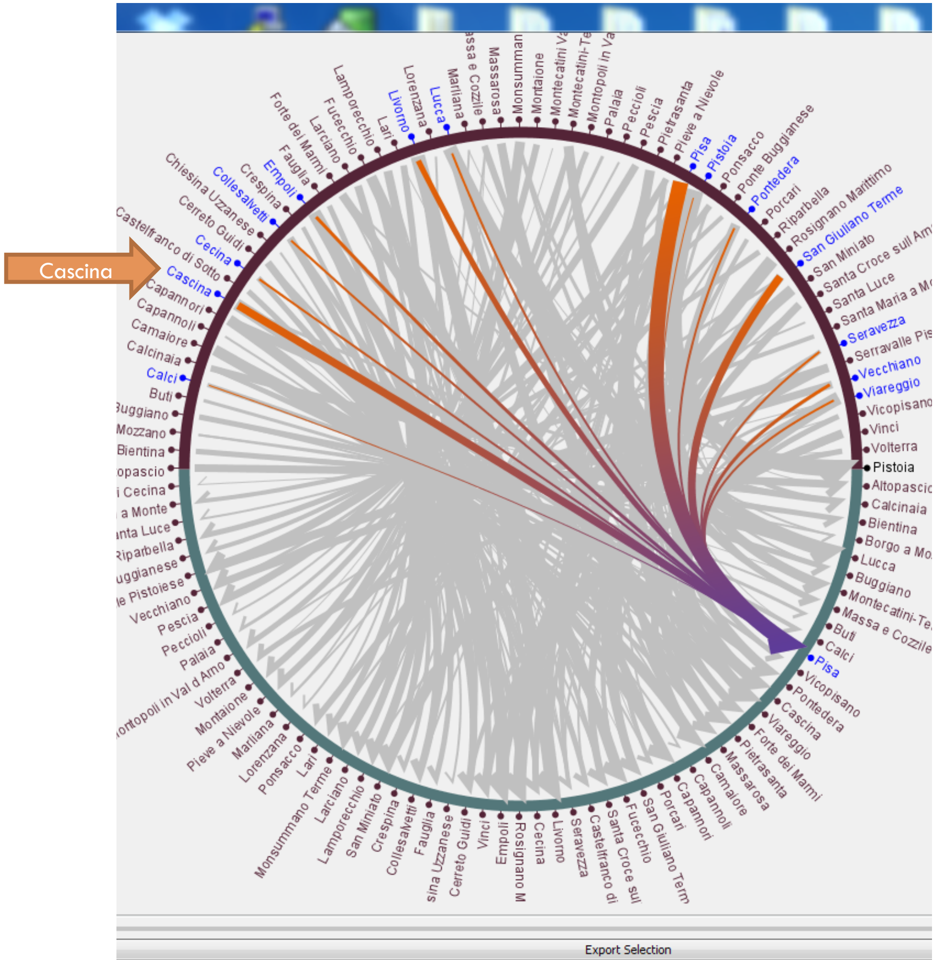
Drill down: from cities to cities (filtered)

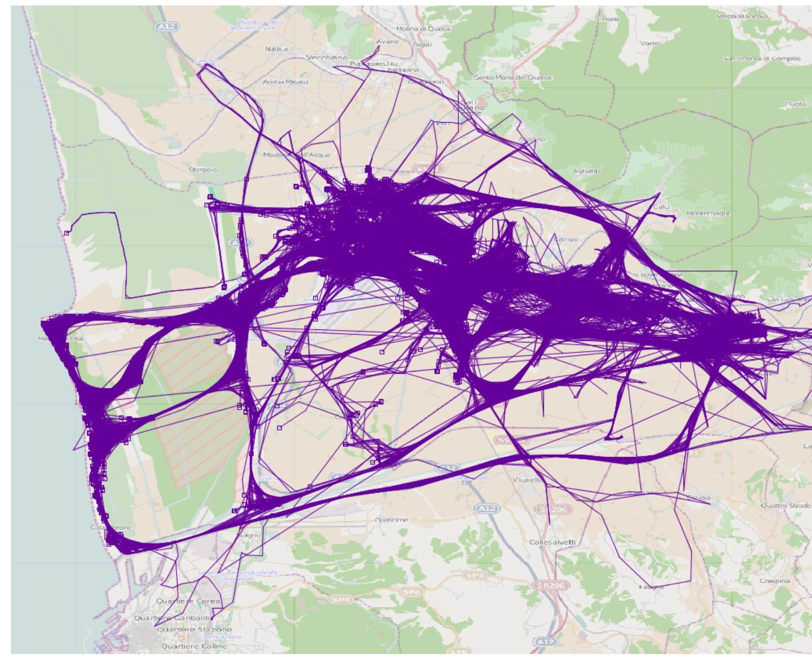




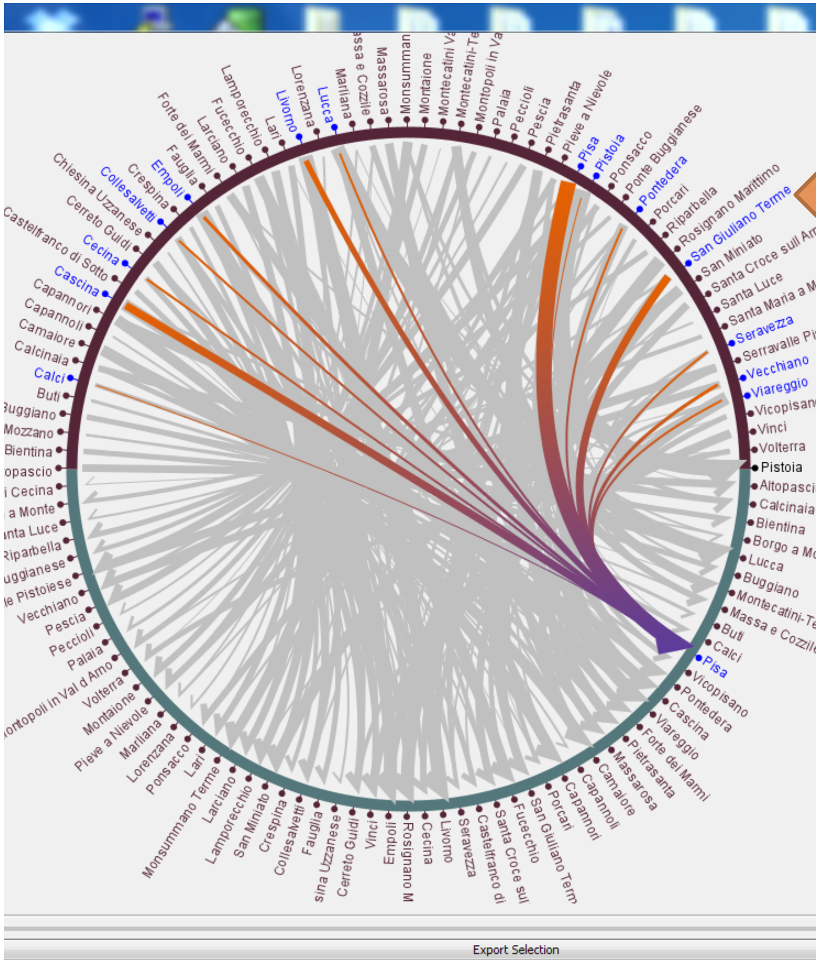
Restrict visualization to flows above a given threshold.
Select specific flows: from Cascina, San Giuliano, and Pisa

Specific flow: from Cascina to Pisa

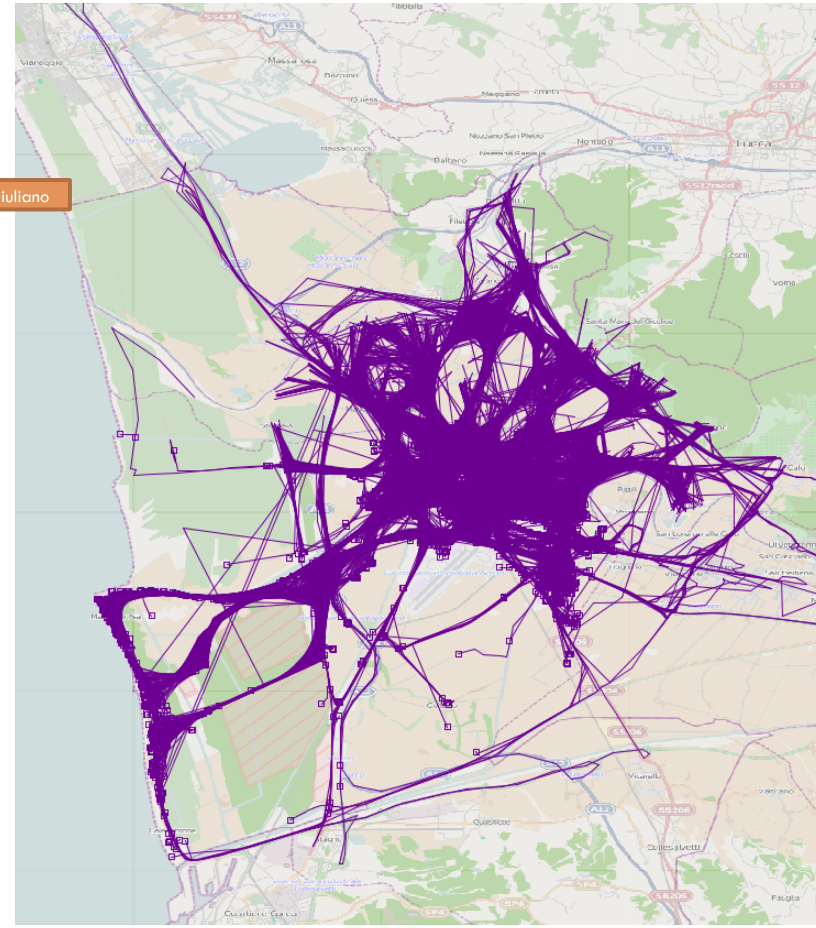




Specific flow: from San Giuliano to Pisa

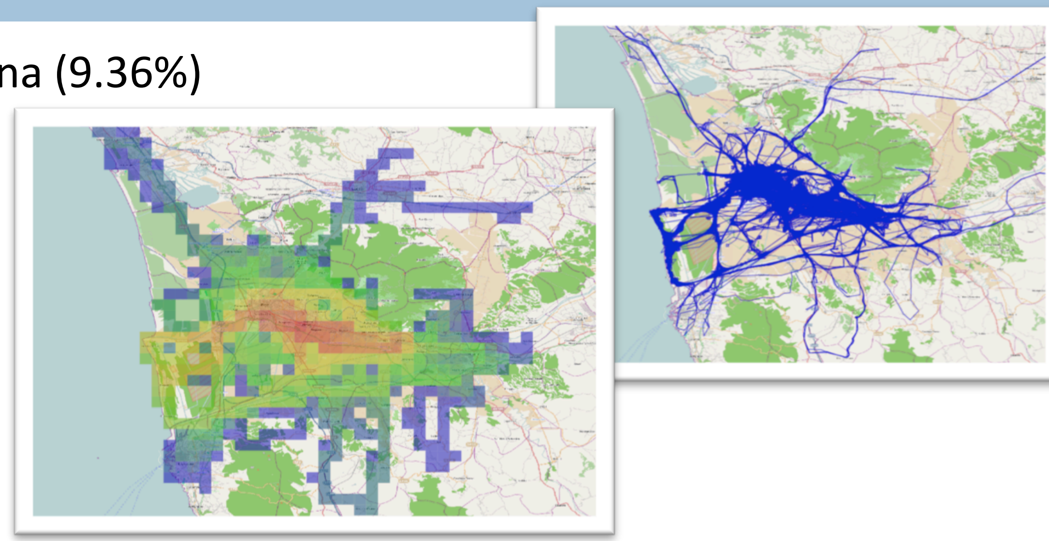


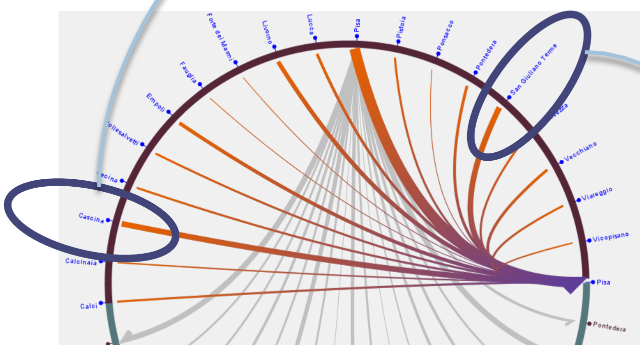
San Giuliano



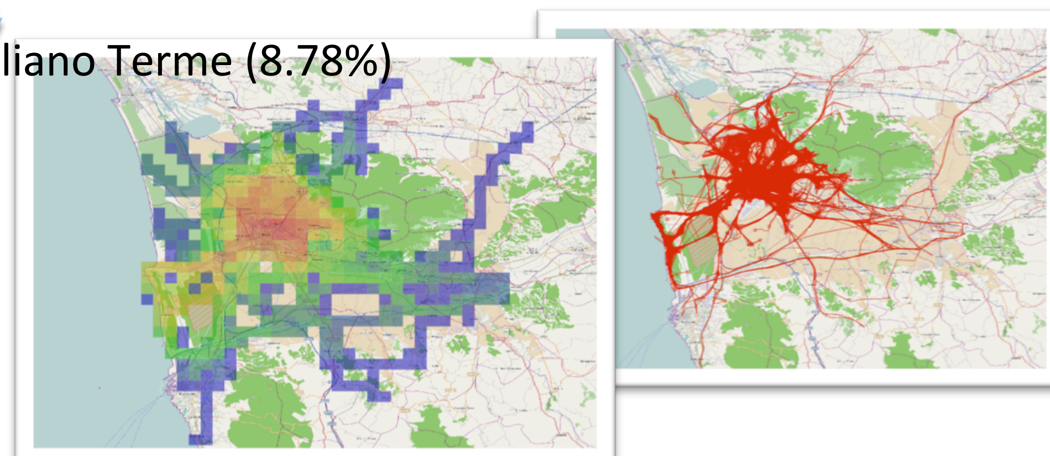
Exploring entailed data

Cascina (9.36%)





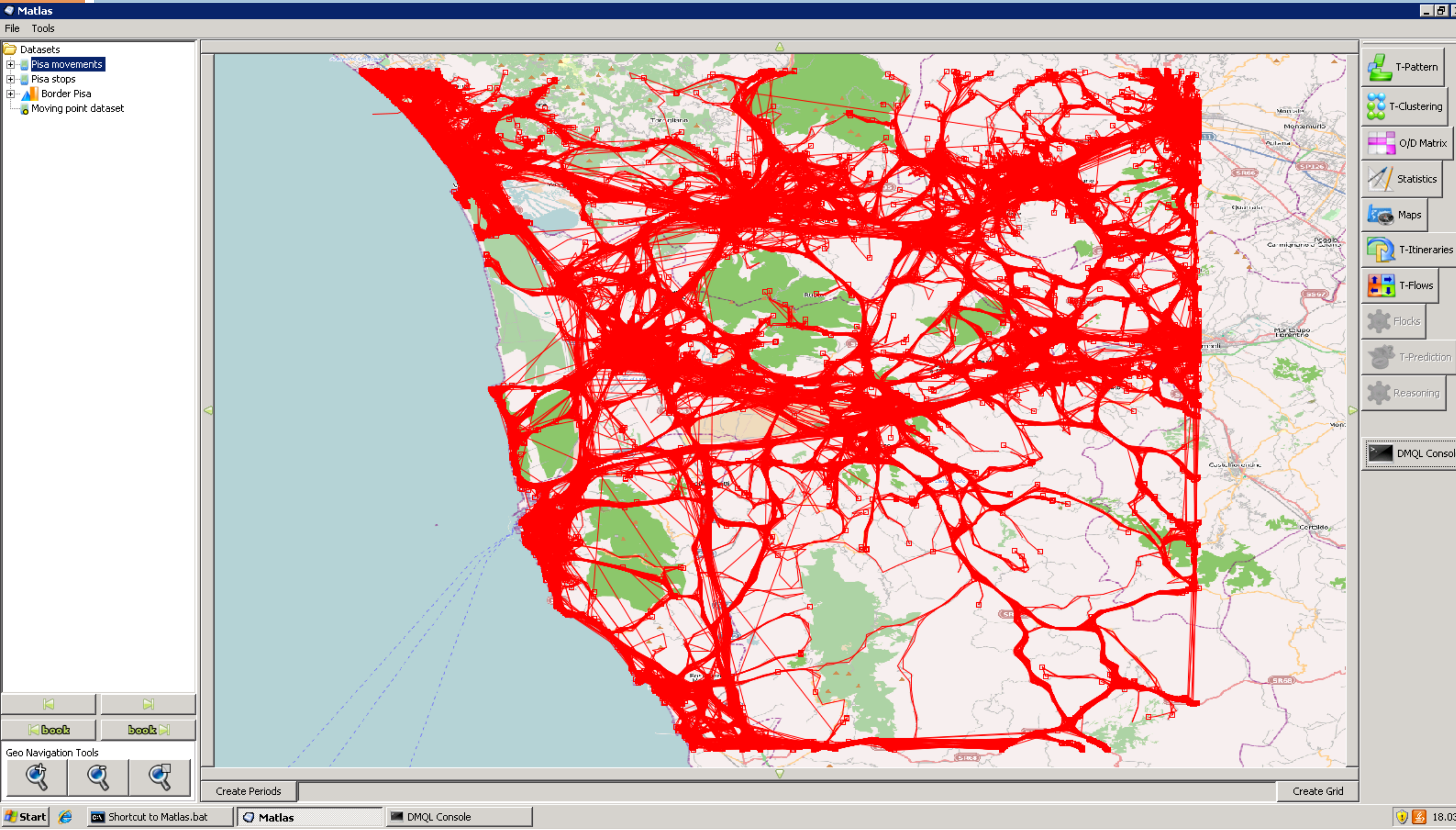
San Giuliano Terme (8.78%)



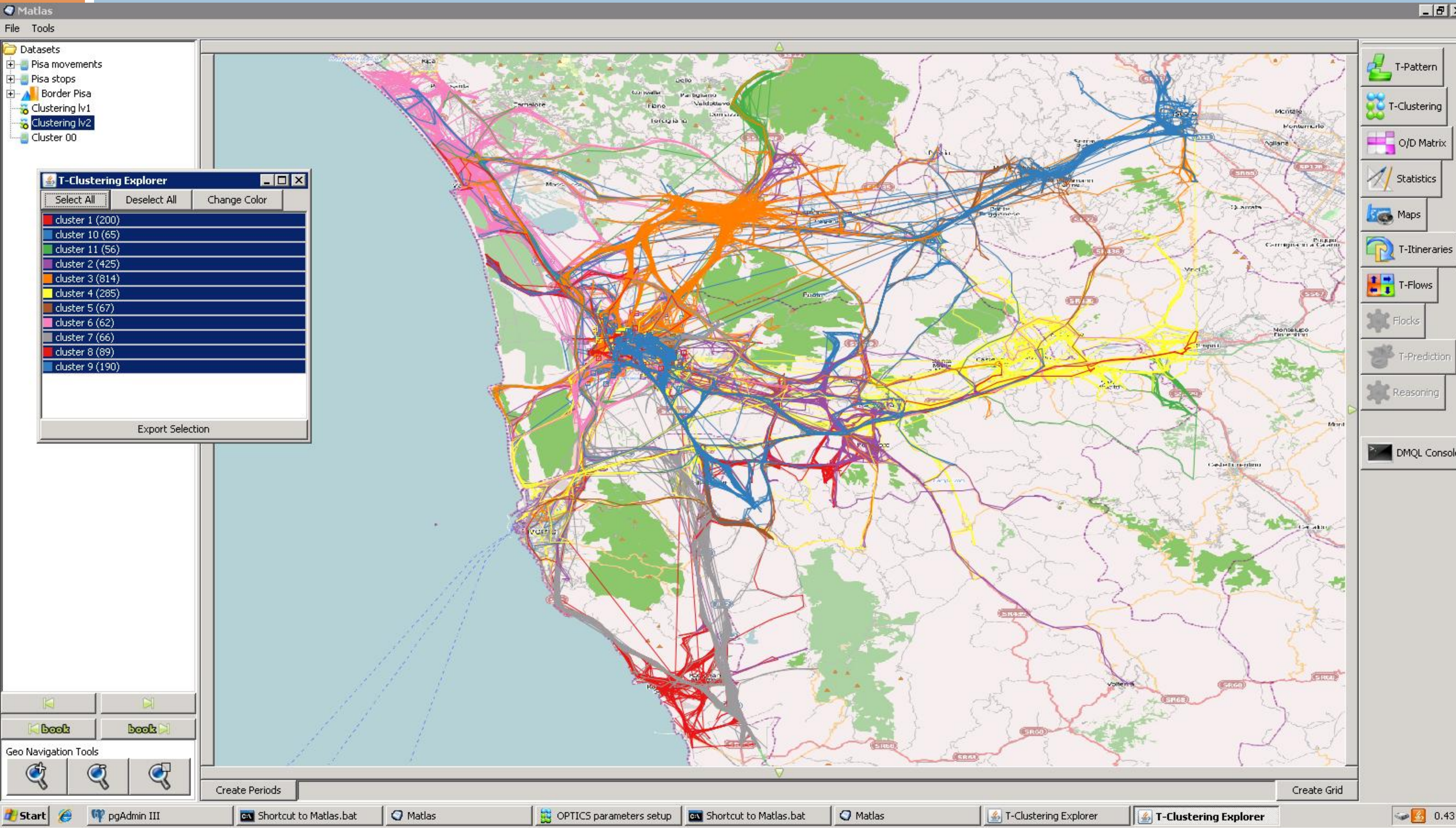
		Measures	
Cell To	Cell From	Num vehicles	%
+Pisa	−Pisa	89.730	84,24%
	+Pisa	63.331	70,58%
	+Cascina	8.402	09,36%
	+San Giuliano Terme	7.877	08,78%
	+Vecchiano	1.869	02,08%
	+Pontedera	1.408	01,57%
	+Calci	1.220	01,36%

Discovering **access patterns** to Pisa with GPS track data

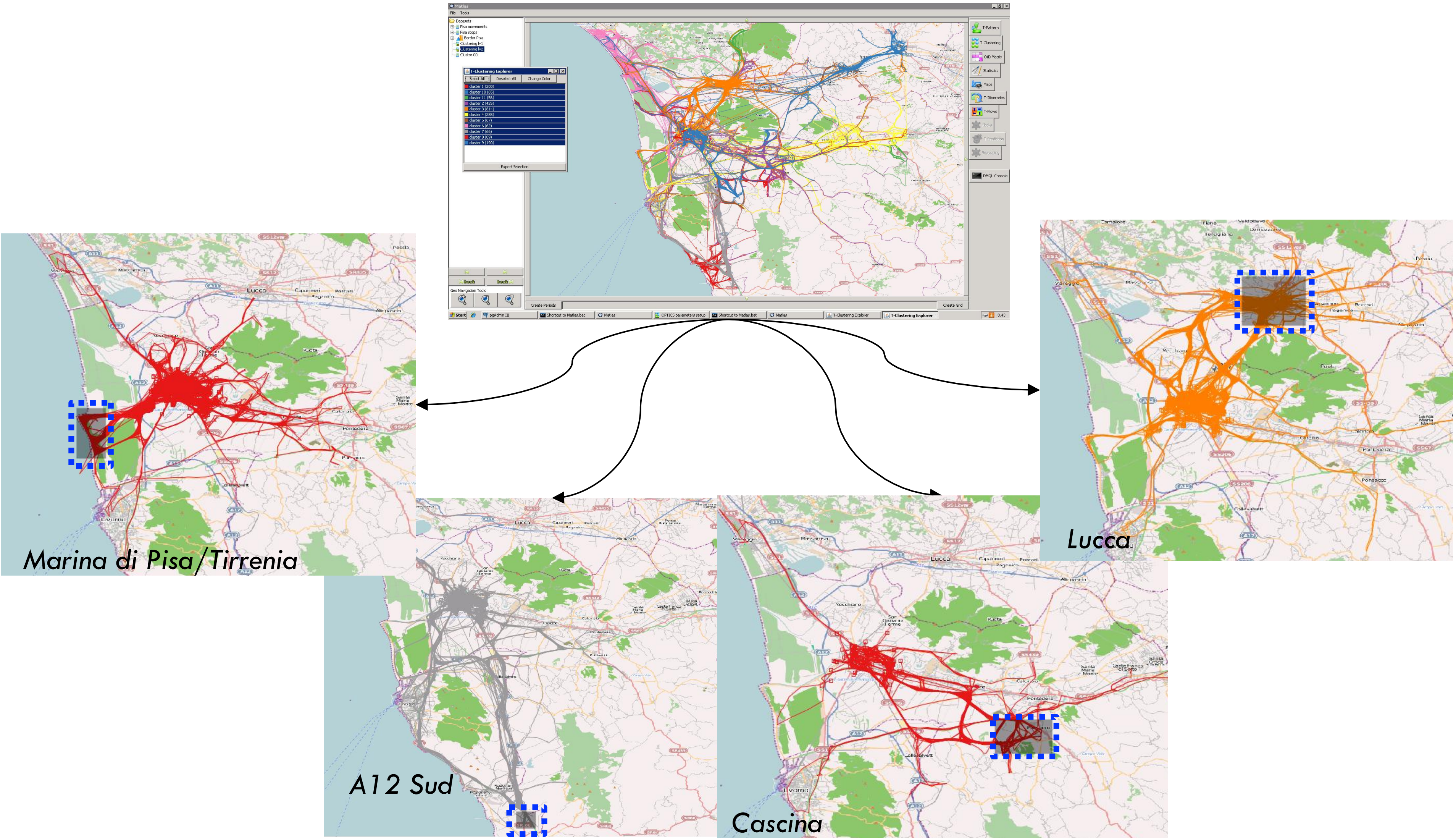
Access patterns using T-clustering



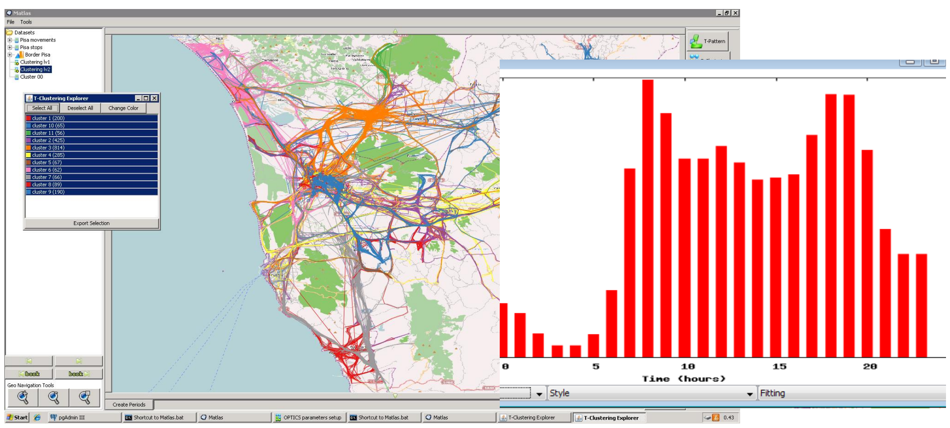
Access patterns using T-clustering

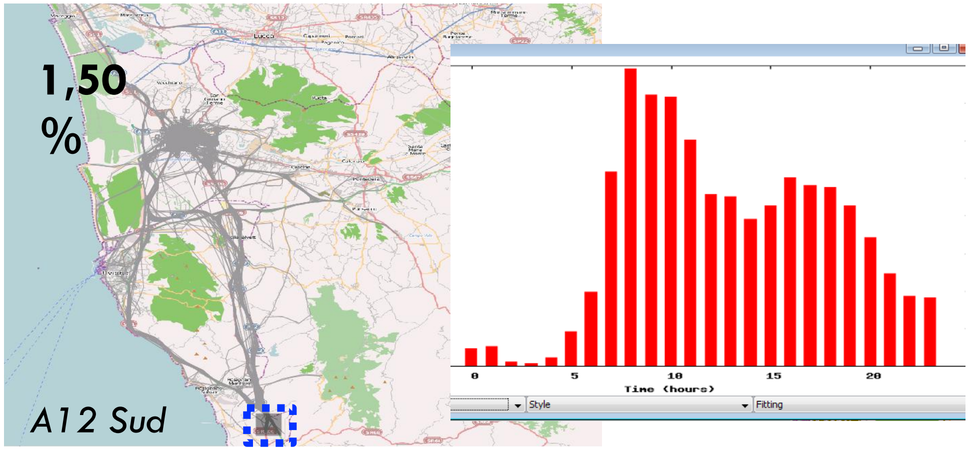


Access patterns using T-clustering

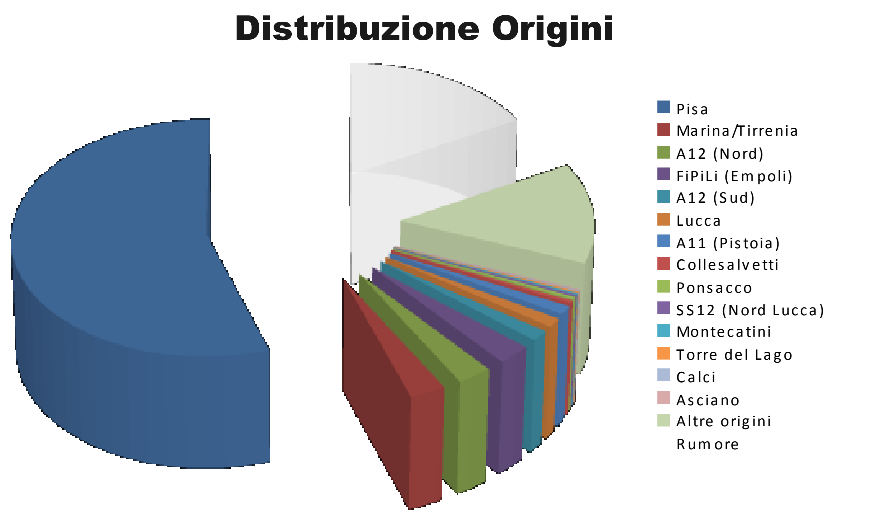


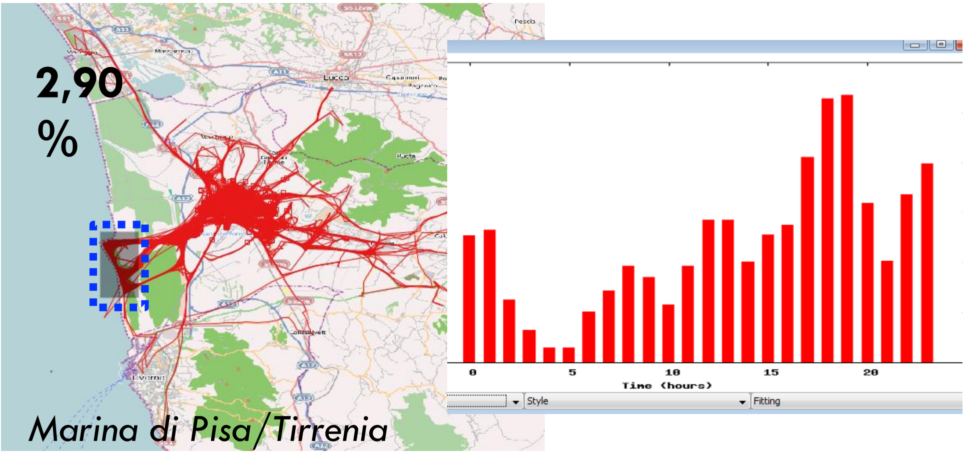
Characterizing the **access patterns**: origin & time



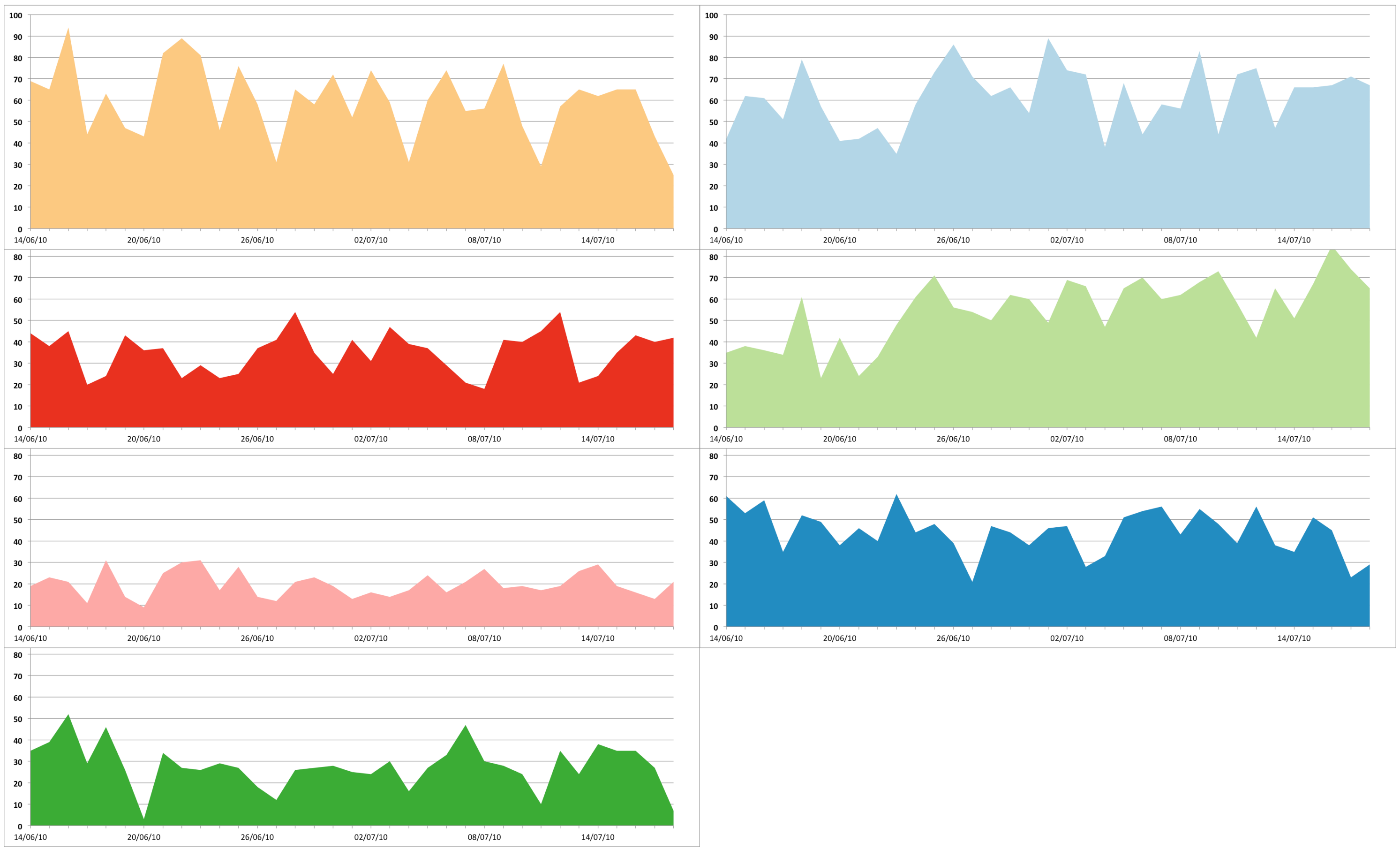


Origin distribution

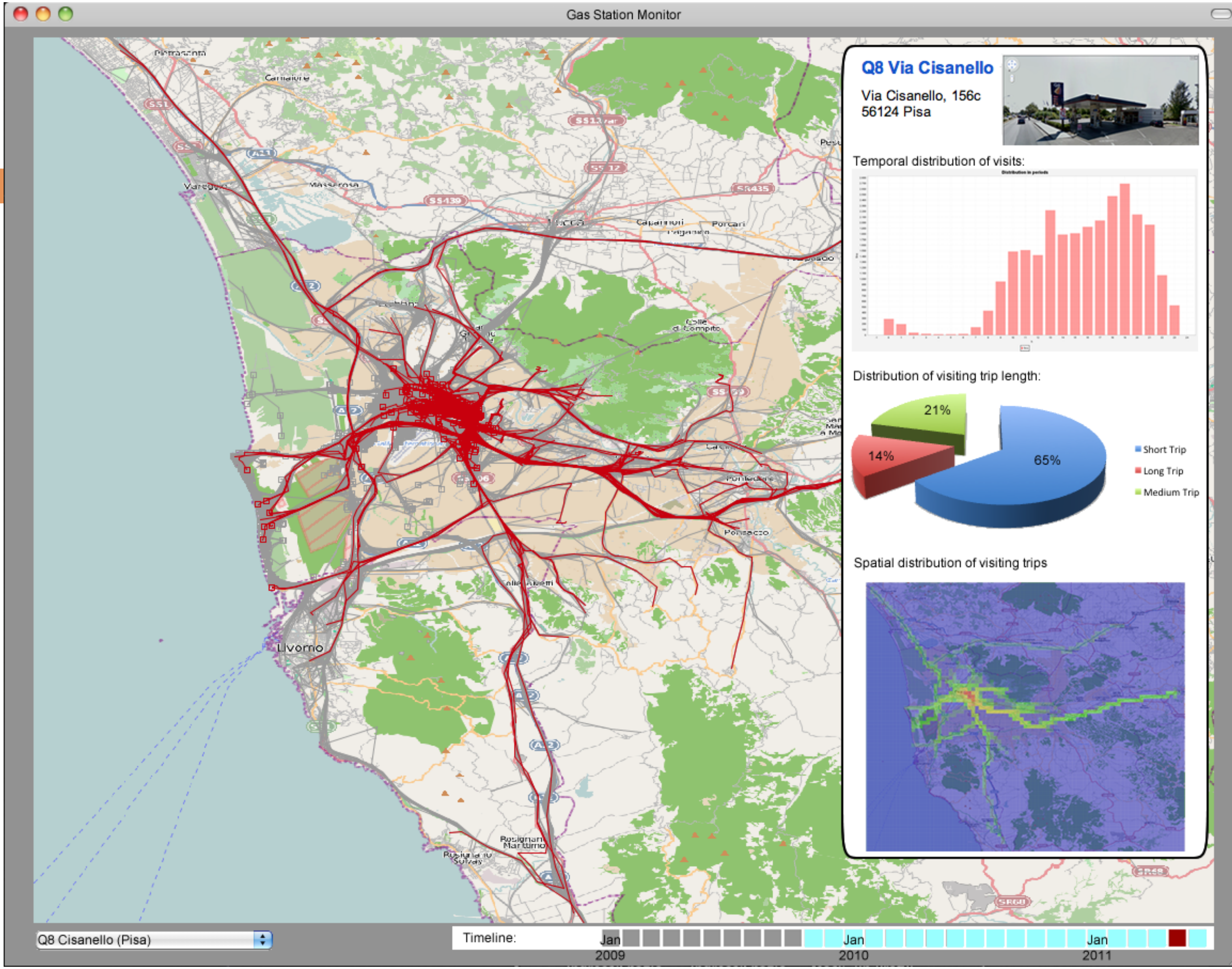




Persistency of access patterns



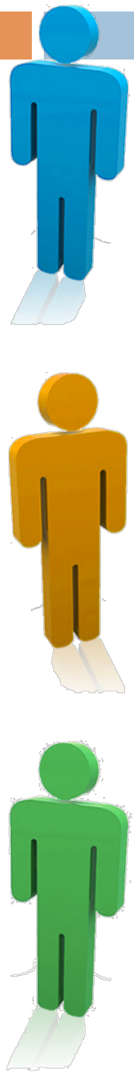
Studying the attractiveness/efficiency of a service with GPS tracks

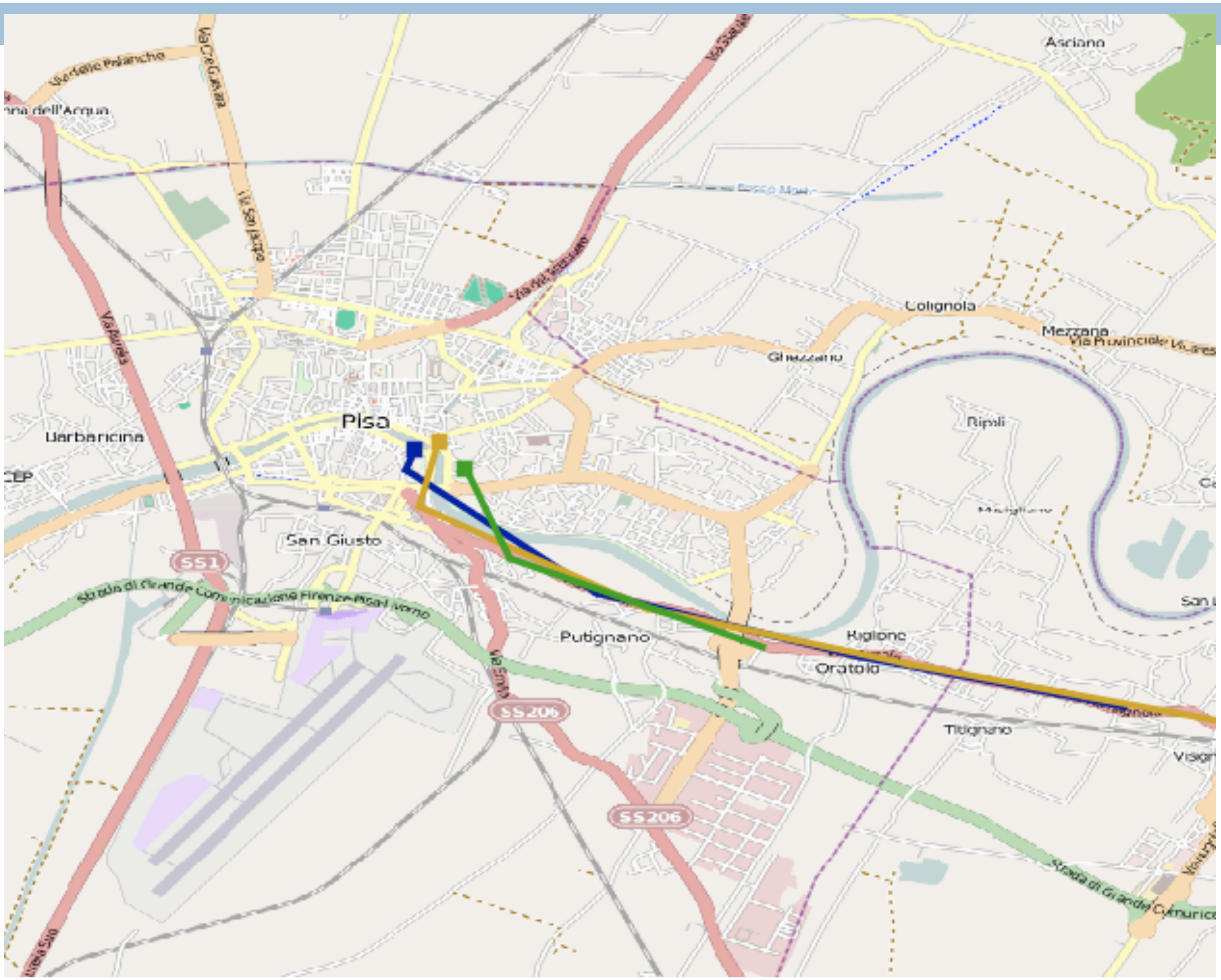


Gas Station Monitor
Q8 Via Cisanello
Via Cisanello, 156c
56124 Pisa
Temporal distribution of visits:
Distribution of visiting trip length:
21%
14%
65%
Short Trip
Long Trip
Medium Trip
Spatial distribution of visiting trips
Livorno
Q8 Cisanello (Pisa)
Timeline:
Jan
2009
Jan
2010
Jan
2011

Discovering mobility profiles with GPS tracks data

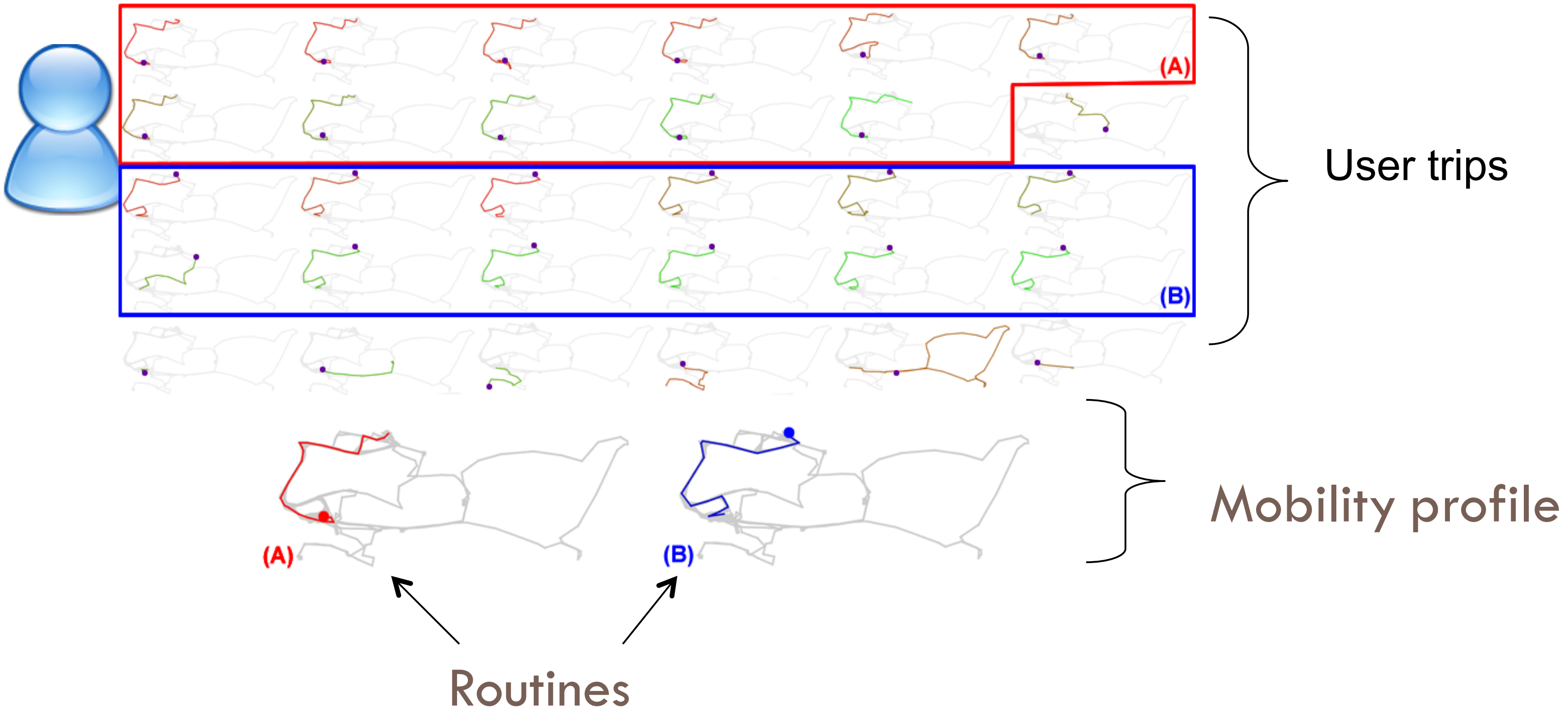
Extract travellers profiles





Extracting travellers profiles

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

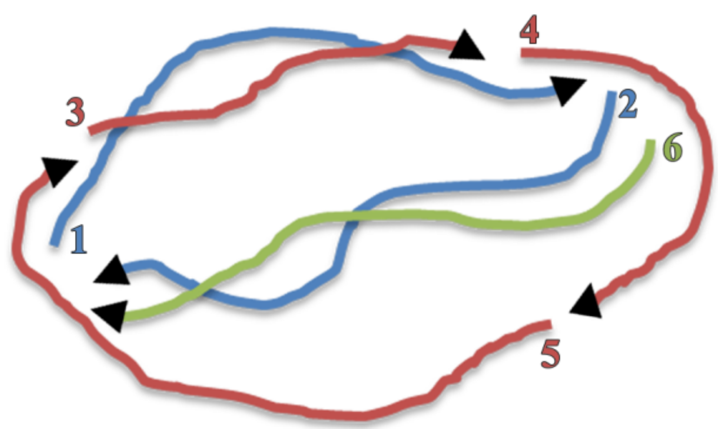


Application: Car pooling

Pro-active suggestions of sharing rides opportunities without the need for the user to explicitly specify the trips of interest.

Matching two routines:

$$\begin{aligned} contained(T_1, T_2, th^{walking}_{distance}, th^{wasting}_{time}) \;\equiv\; & \exists i,j \in \mathcal{N} \mid \\ & 0 < i \leq j \leq m \;\wedge \\ & Dist(p^1_1, p^2_i) + Dist(p^1_n, p^2_j) \leq th^{walking}_{distance} \;\wedge \\ & Dur(p^1_1, p^2_i) + Dur(p^1_n, p^2_j) \leq th^{wasting}_{time} \end{aligned}$$



	1	2	3	4	5	6
1	-	-	F	F	F	F
2	-	-	F	F	F	**T**
3	**T**	F	-	-	-	F
4	F	F	-	-	-	F
5	F	F	-	-	-	F
6	F	**T**	F	F	F	-

	-	0	1/2
	1/3	-	0
	1	0	-

Mobility profile share-ability:

$$mobility\ profiles\ \tilde{T}_1\ and\ \tilde{T}_2$$

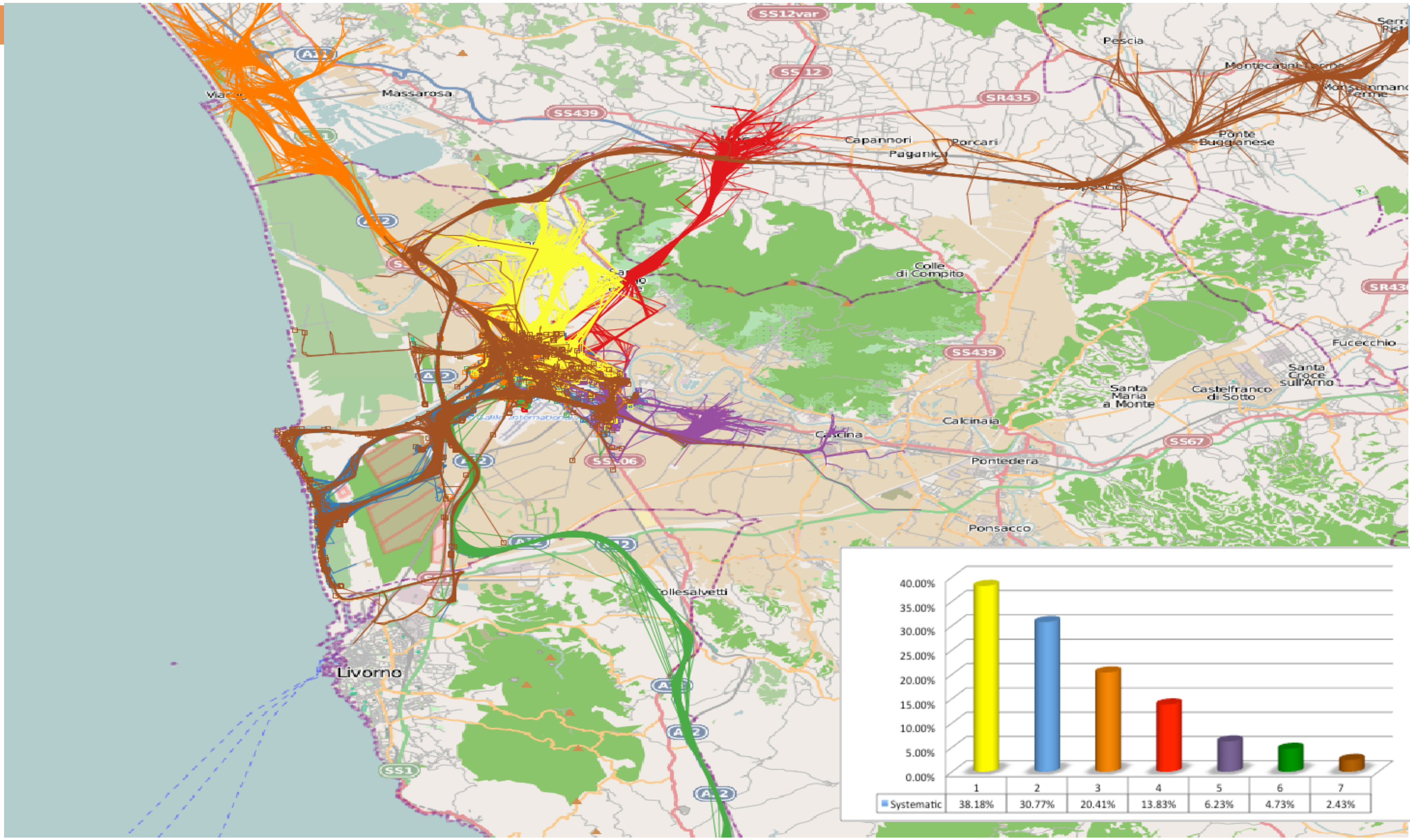
$$profileShare(\tilde{T}_1, \tilde{T}_2, th^{walking}_{distance}, th^{wasting}_{time}) = \frac{\left| \left\{ p \in \tilde{T}_1 \mid \exists q \in \tilde{T}_2 . Share(p, q, th^{walking}_{distance}, th^{wasting}_{time}) \right\} \right|}{\mid \tilde{T}_1 \mid}$$

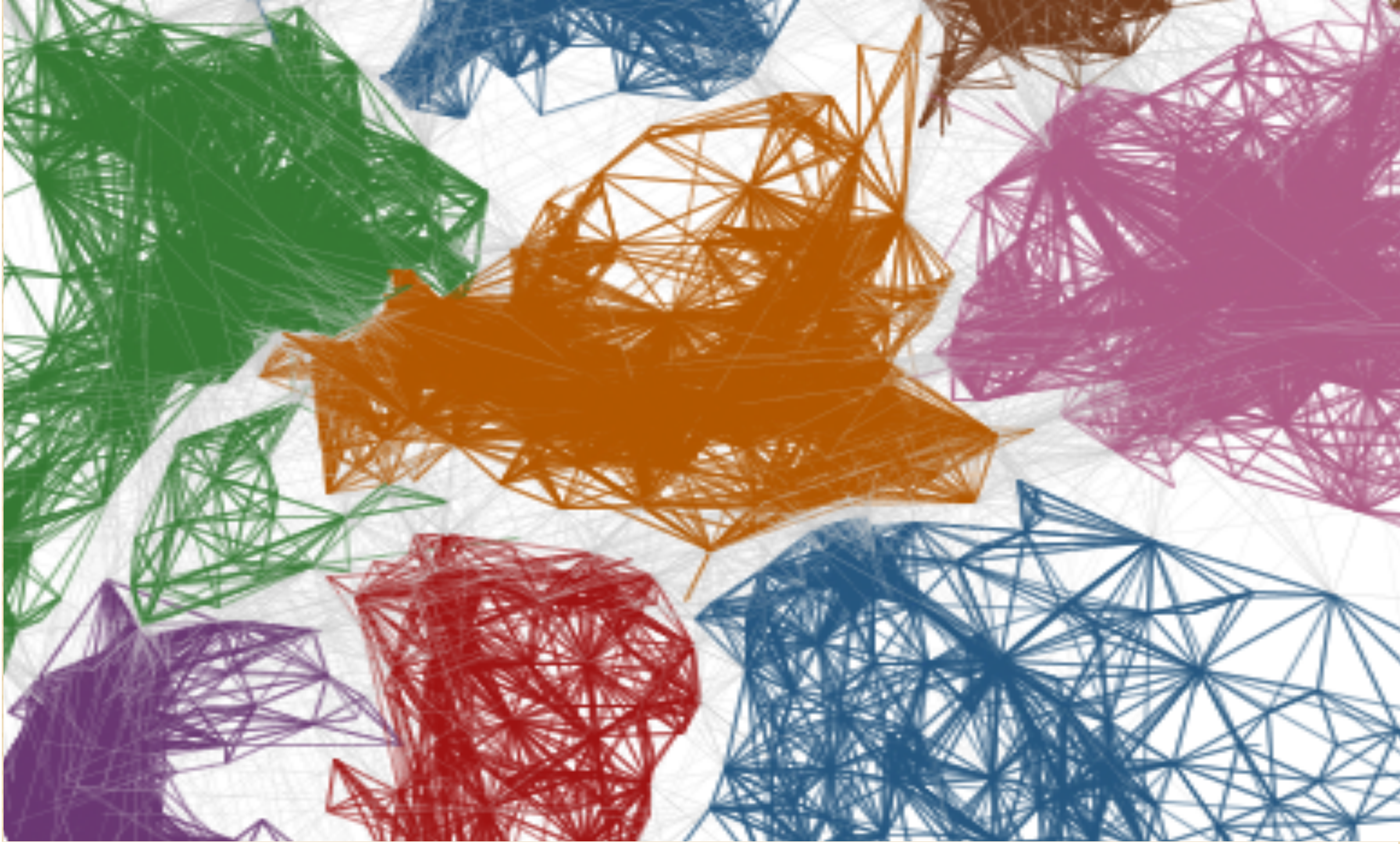
Car pooling potential

67.2% routines match with a routine of other users

32.5% users share one or more routines with other users

Impact of systematic mobility on access patterns



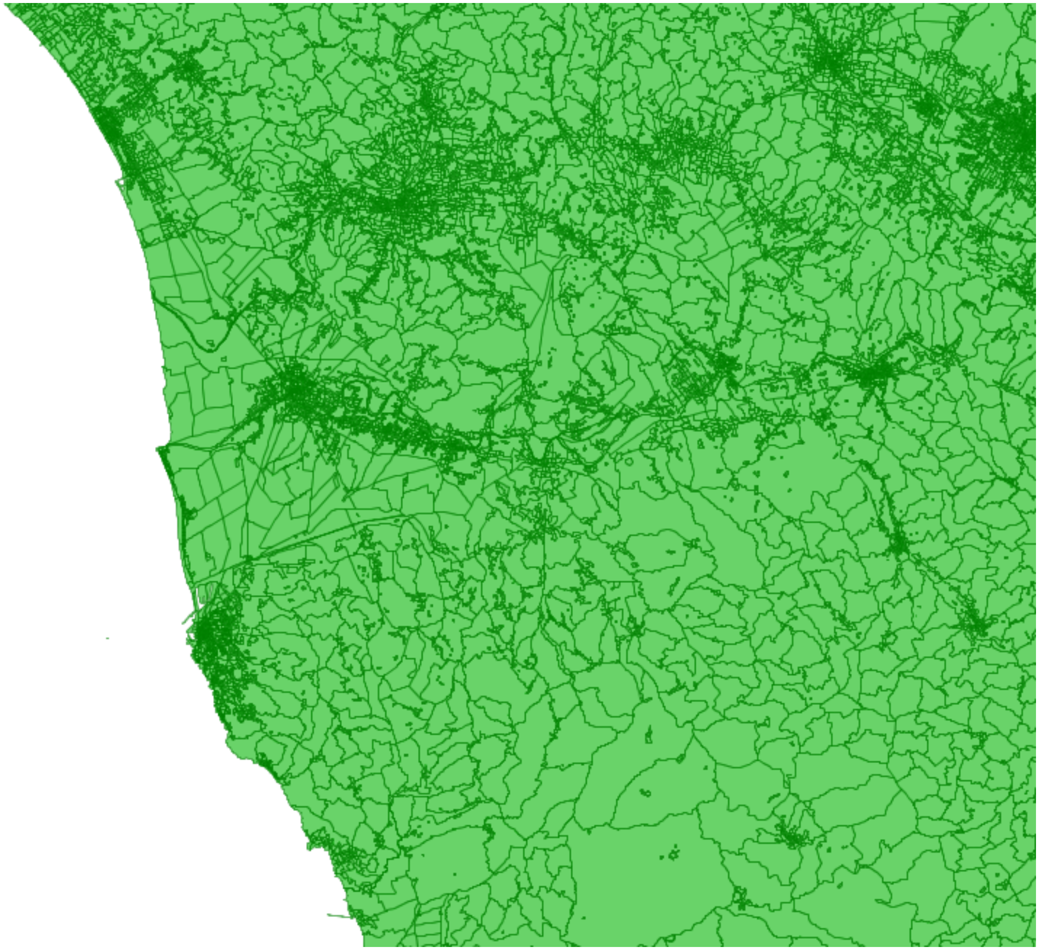


Find border of human mobility

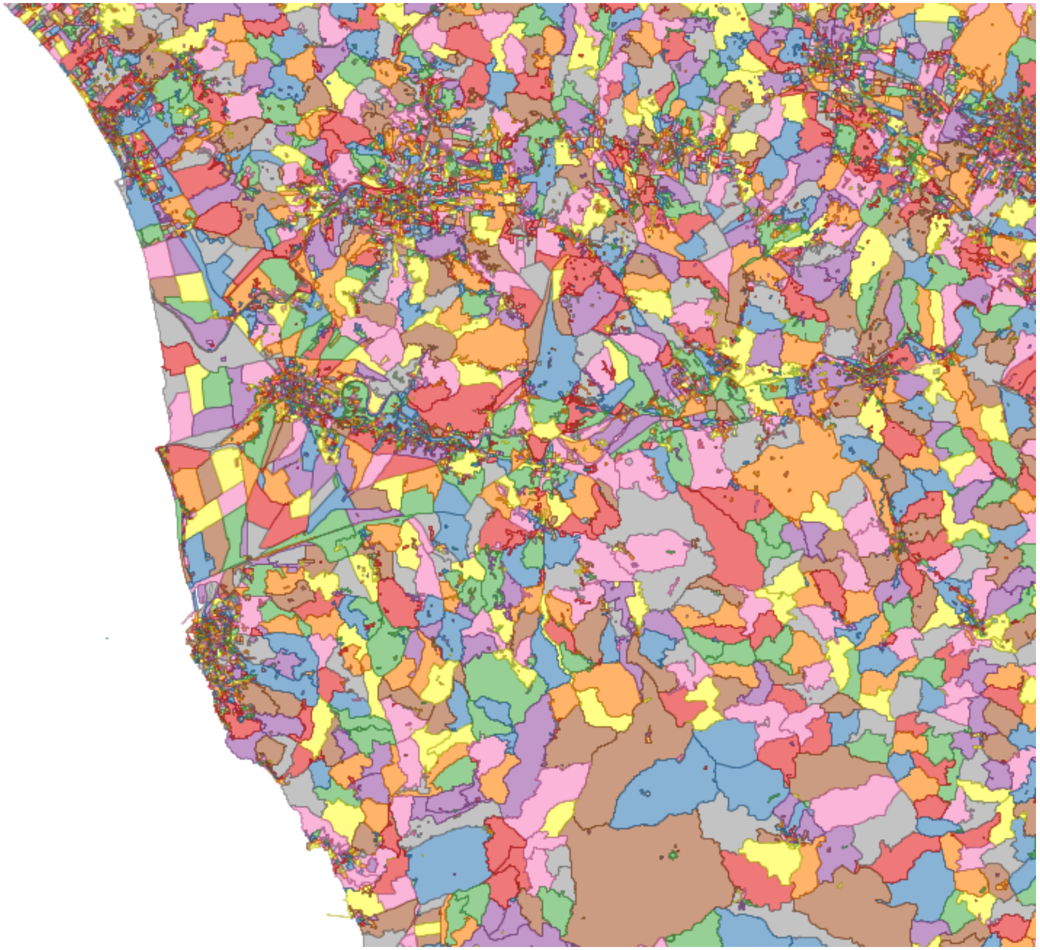
Motivations

- Mobility management offices needs accurate information to handle mobility issues
  - Monitoring: how to predict/manage emergences of special events?
  - Planning: public transportation desisgn, incentives for multi-modal movement, etc.
- Planning involse several entities
  - The city level is not sufficient: the neighbor cities are necessarily influenced
  - The regional level is too general: lost focus for specific/local requirements
  - Does provinces provide the necessary level of details?

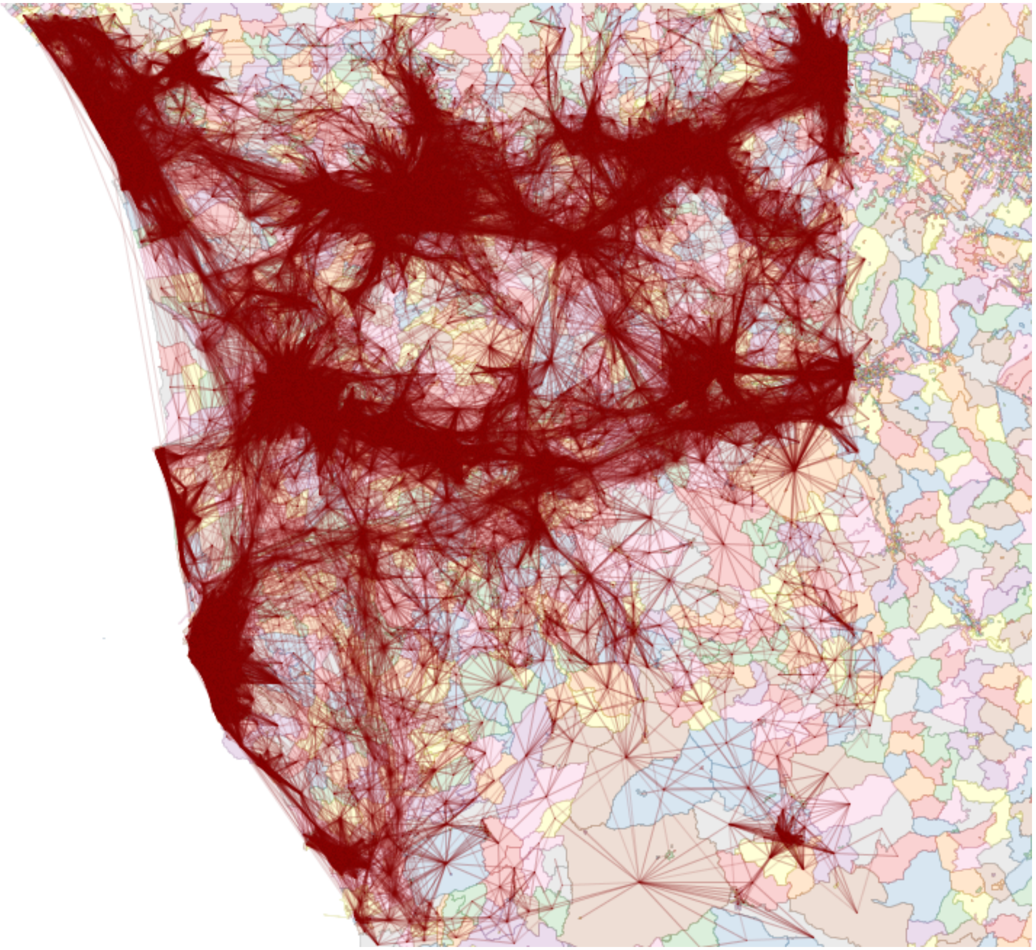
Step 1: spatial regions



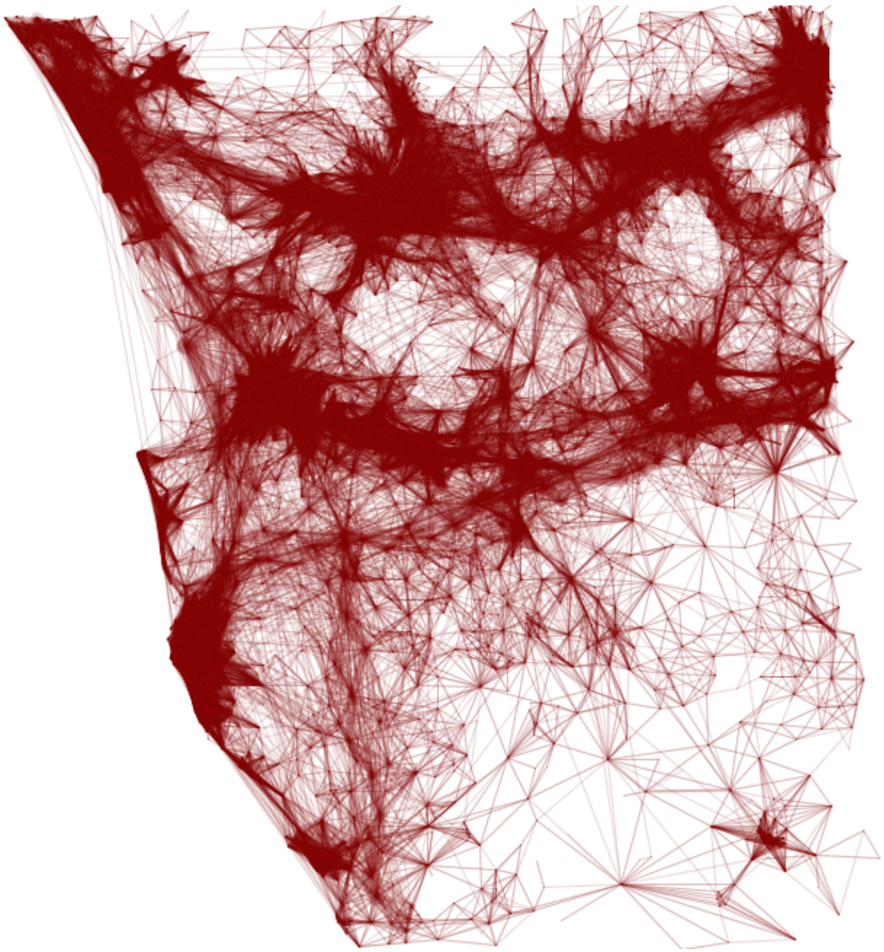
Start from random labeling for region



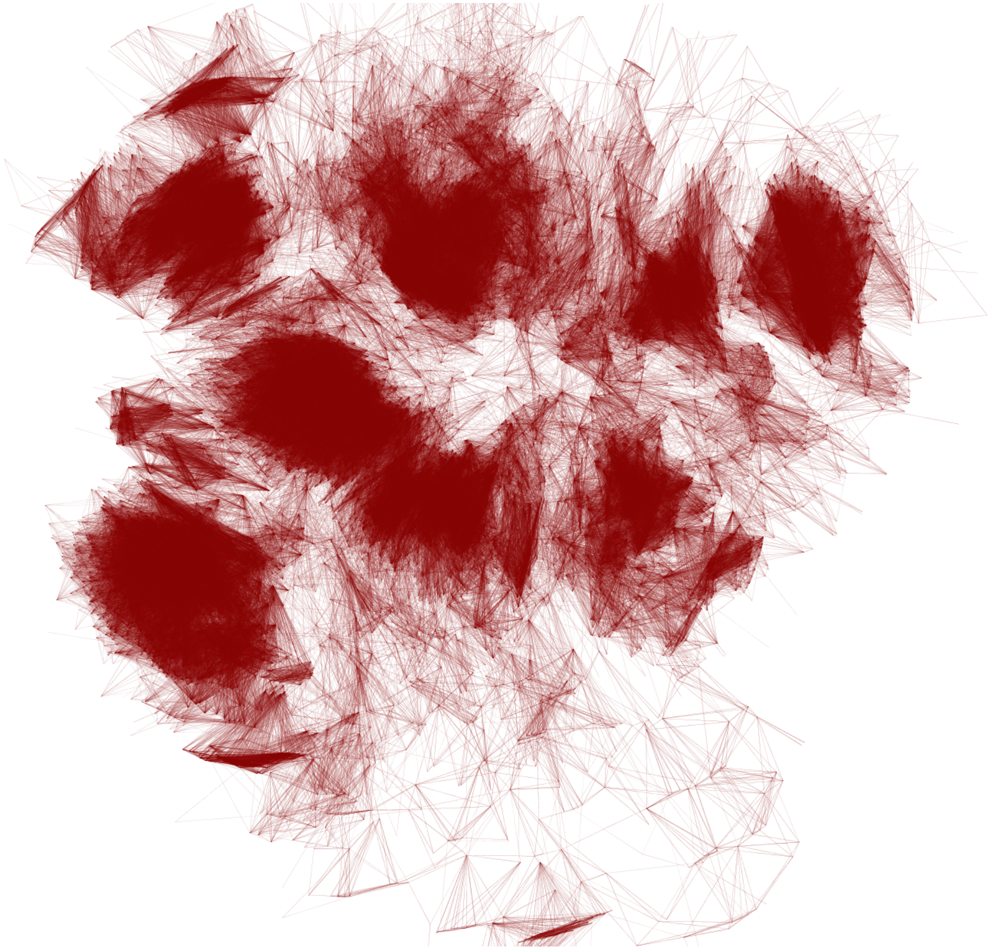
Step 2: evaluate flows among regions



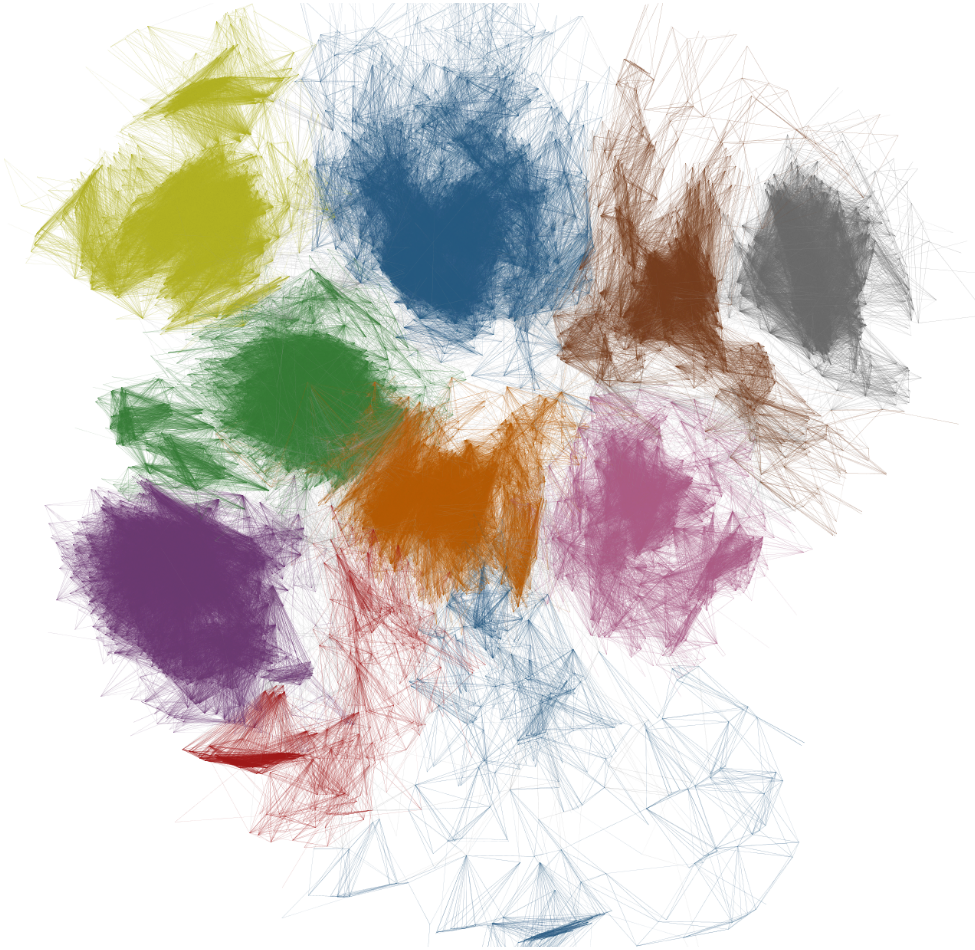
Step 3: consider only the network



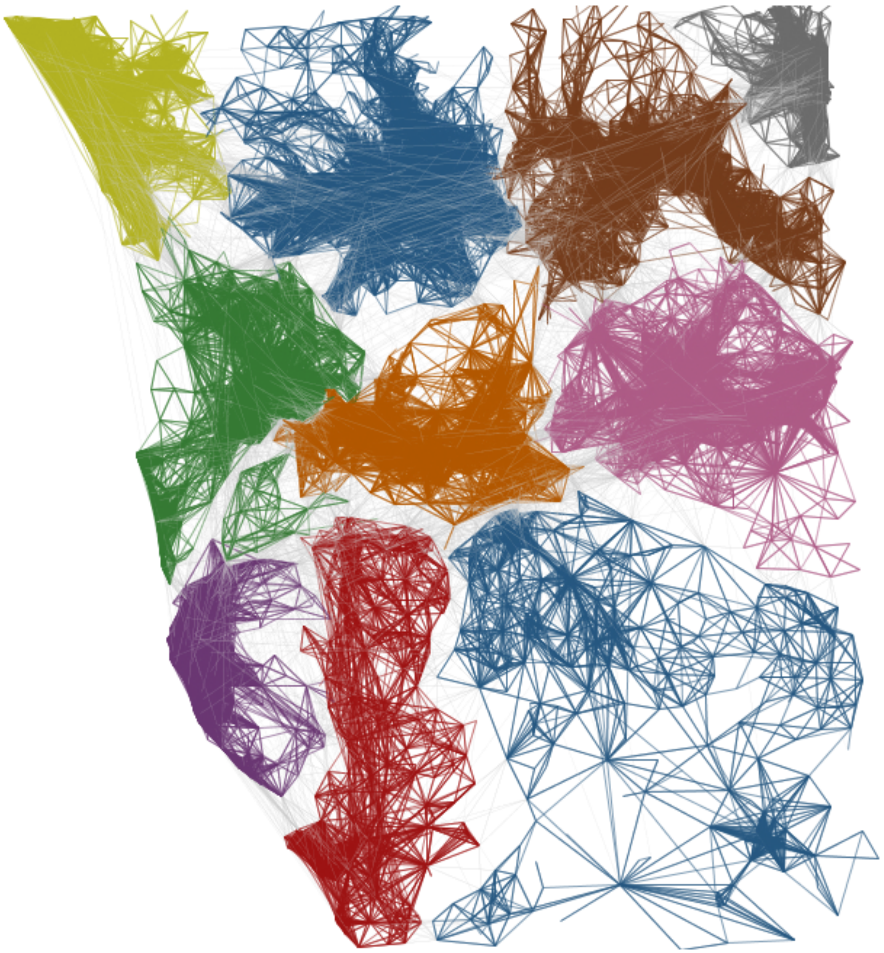
Step 4: perform clustering



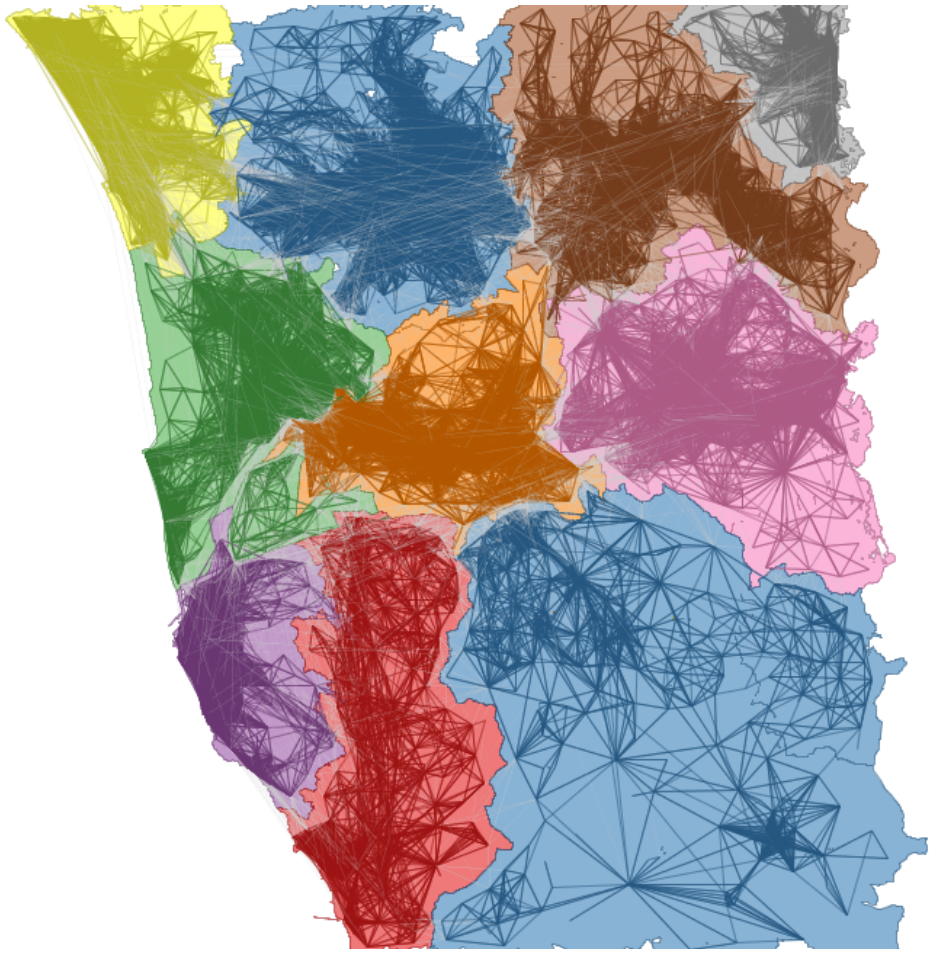
Step 4: perform clustering



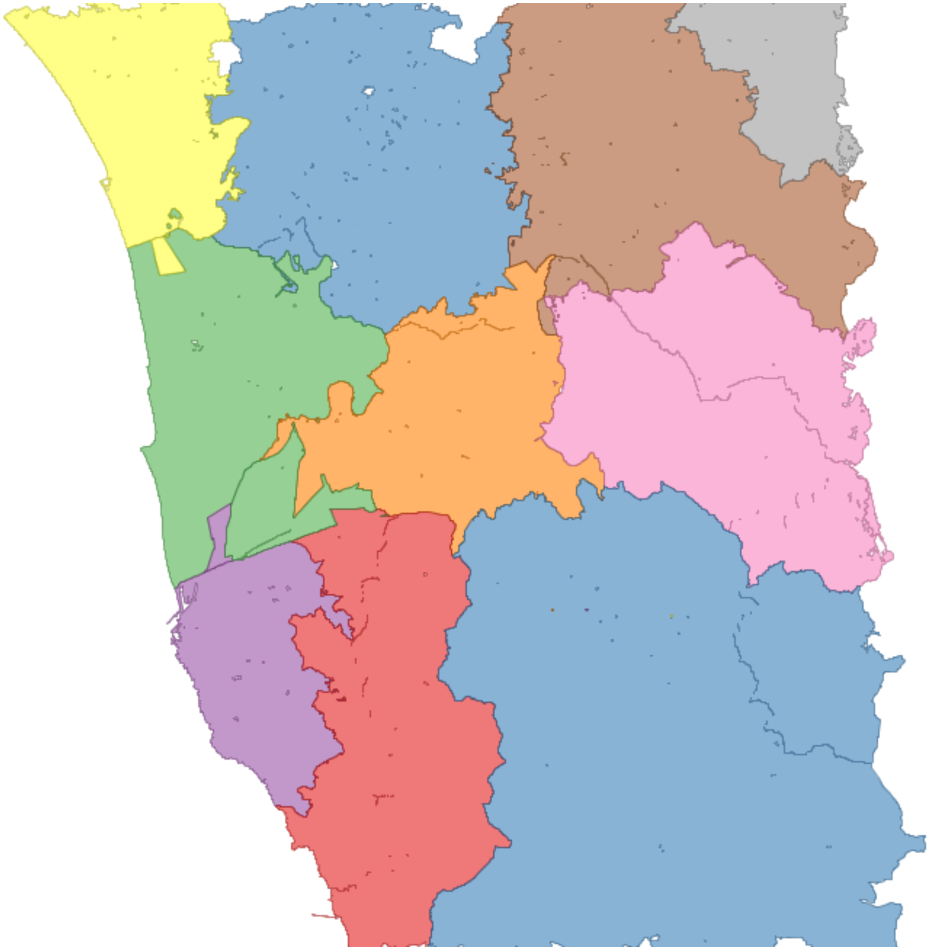
Step 5: map nodes back to geography



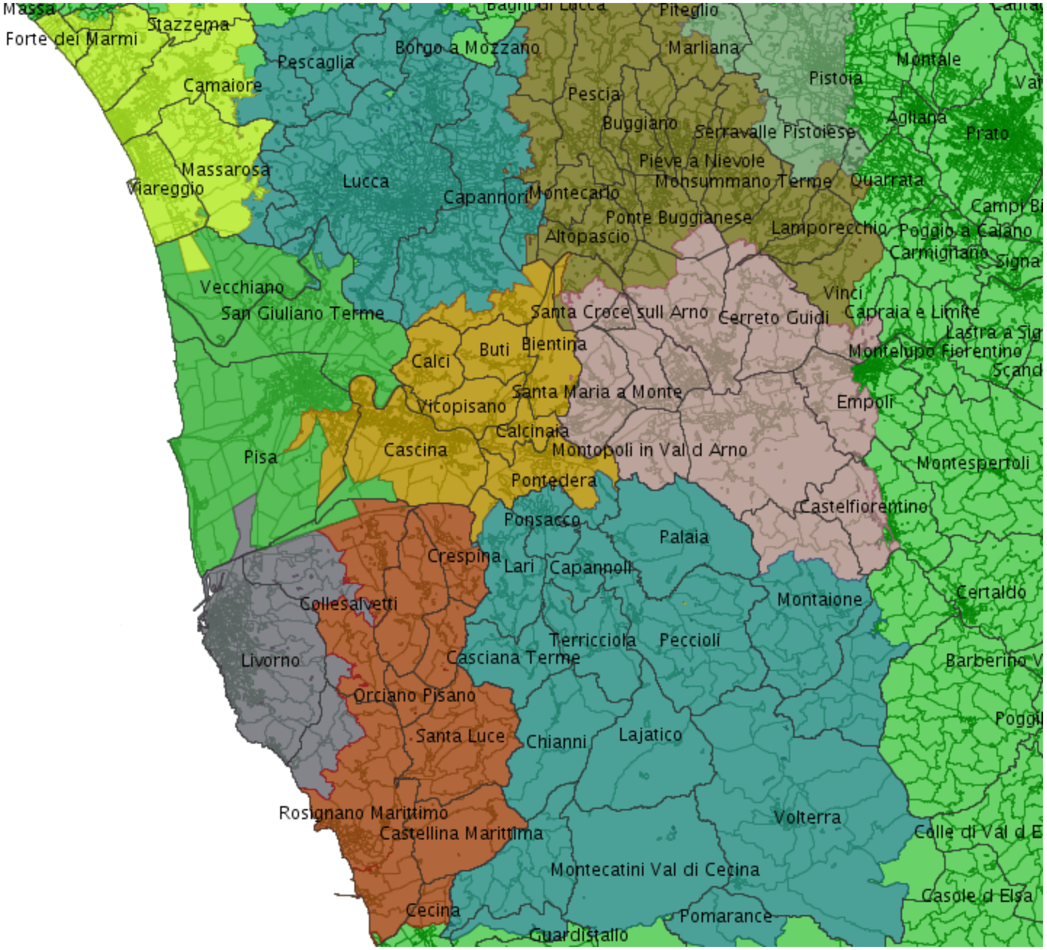
Step 5: map nodes back to geography



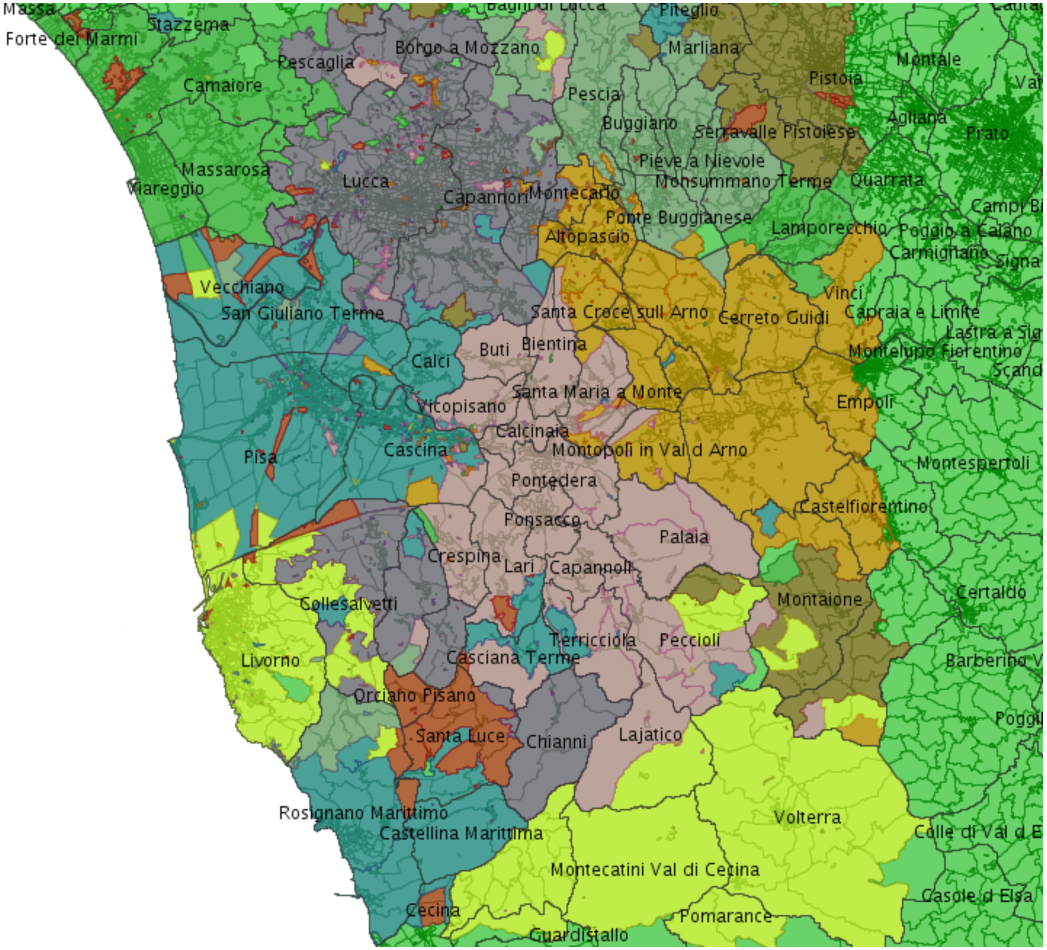
Final result



Final result: comparison

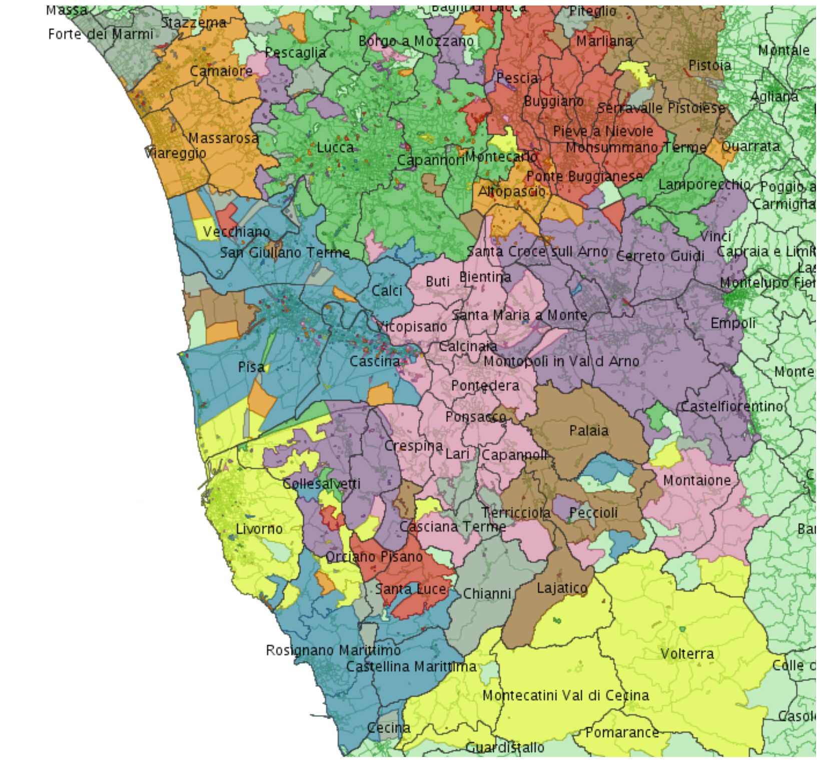


Borders using only OD flows



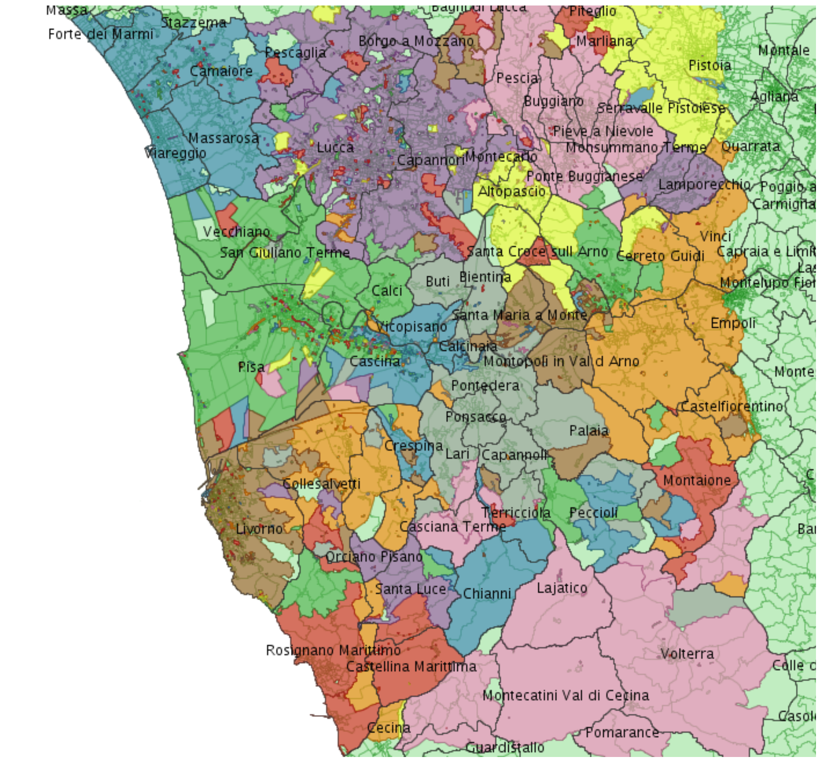
Borders in different time periods

Only weekdays movements



Similar to global clustering: strong influence of systematic movements

Only weekend movements

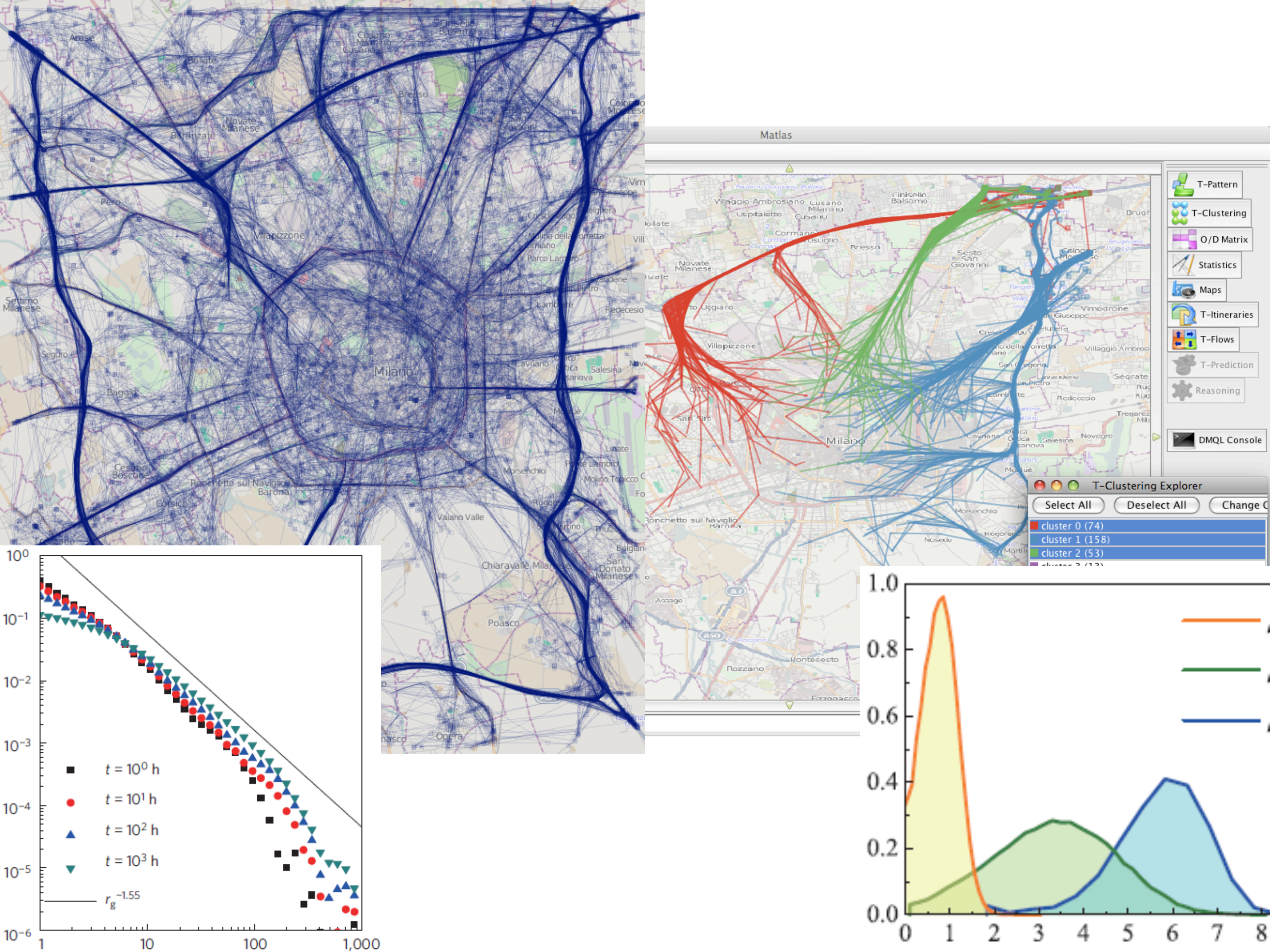


Strong fragmentation: the influence of systematic movements (home-work) is missing

Summarizing: big data push towards converging sciences

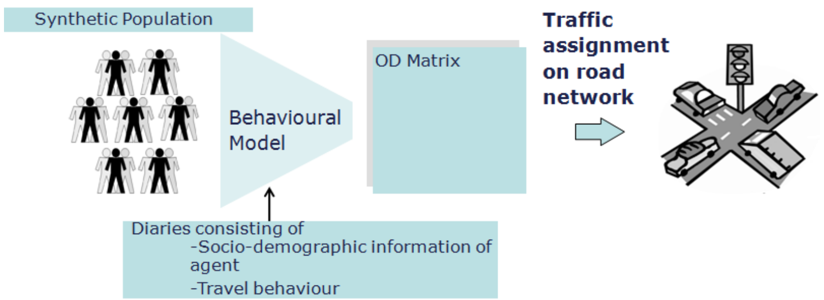
Big data push towards convergence

- Network science
  - **Global models** of complex social phenomena
  - Behavioral **diversity** in society at large
- Data mining
  - **Local patterns** of complex social phenomena
  - Behavioral **similarity** in sub-populations
- Both visions needed to achieve realistic and accurate models for prediction and **simulation**
  - Computational sociology (Lazer et al., Science 2009)
- Both data-driven, each leverage on the other


Matlas
T-Pattern
T-Clustering
O/D Matrix
Statistics
Maps
T-Itineraries
T-Flows
T-Prediction
Reasoning
DMQL Console
T-Clustering Explorer
Select All
Deselect All
cluster 0 (74)
cluster 1 (158)
cluster 2 (53)
Milano
$t = 10^0$ h
$t = 10^1$ h
$t = 10^2$ h
$t = 10^3$ h
$r_g^{-1.55}$

DATA-SIM – Data science for simulating the era of electric vehicles

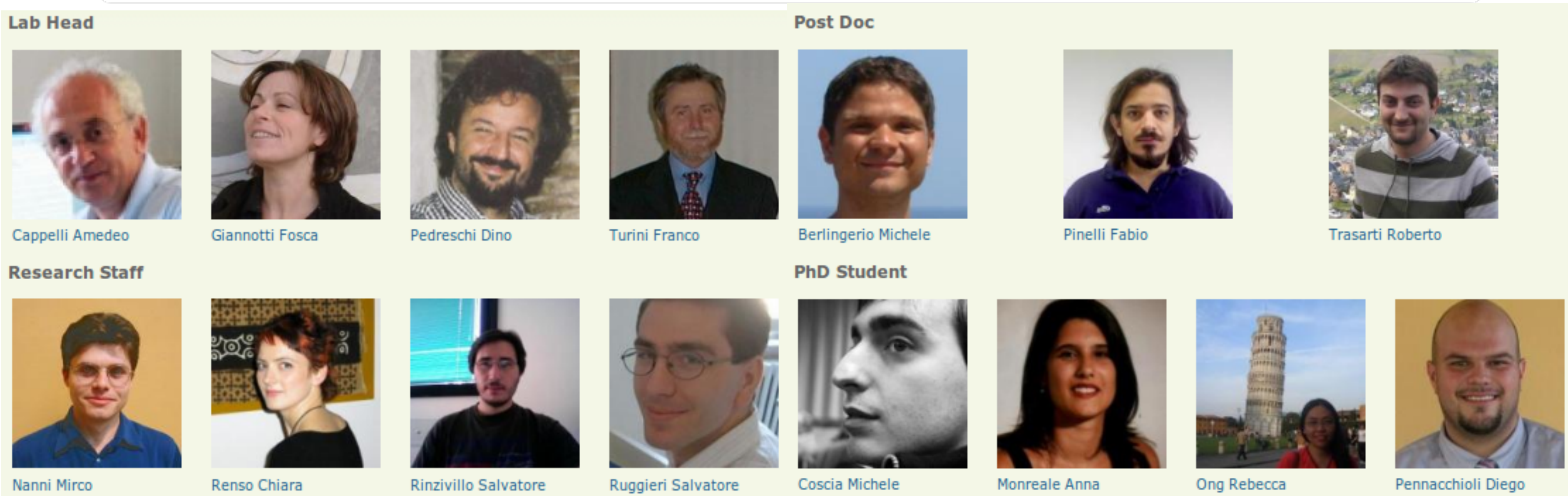
- What's the impact on mobility and energy distribution in the case of a massive switch to electric cars?
- Data mining + network science + agent-based simulation
- FET project started October 2011 www.datasimfet.eu
- KDD LAB Pisa + I-MOB Hasselt + Barabasi Lab Budapest+OCTO



Knowledge Discovery and Data Mining Laboratory

Web Site: http://kdd.isti.cnr.it

Personnel




Caterina D'angelo


Claudio Schifani


Chiara Falchi


Zehui Qu


Barbara Furletti, Andrea Romei, Sergio Barsocchi