Mobility Data Mining

Case Studies

Contents

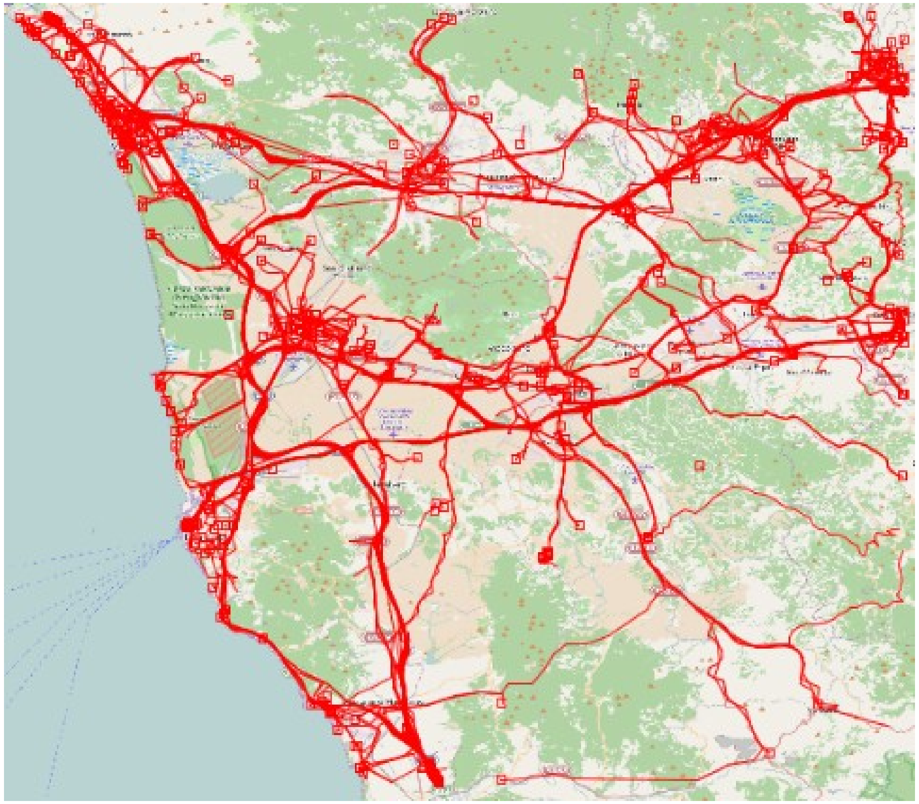
- Corporate Users
  - Geomarketing
  - Monitoring Driving-based Segmentation
- Individual Users
  - Self-awareness
  - Proactive Carpooling
- Public Sector
  - Urban Mobility Atlas
  - Borders

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

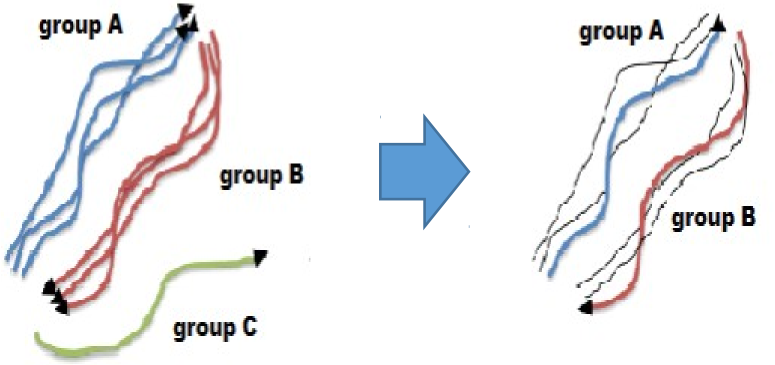
Experts' knowledge: best place to open a mall is where people pass during everyday activities

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.



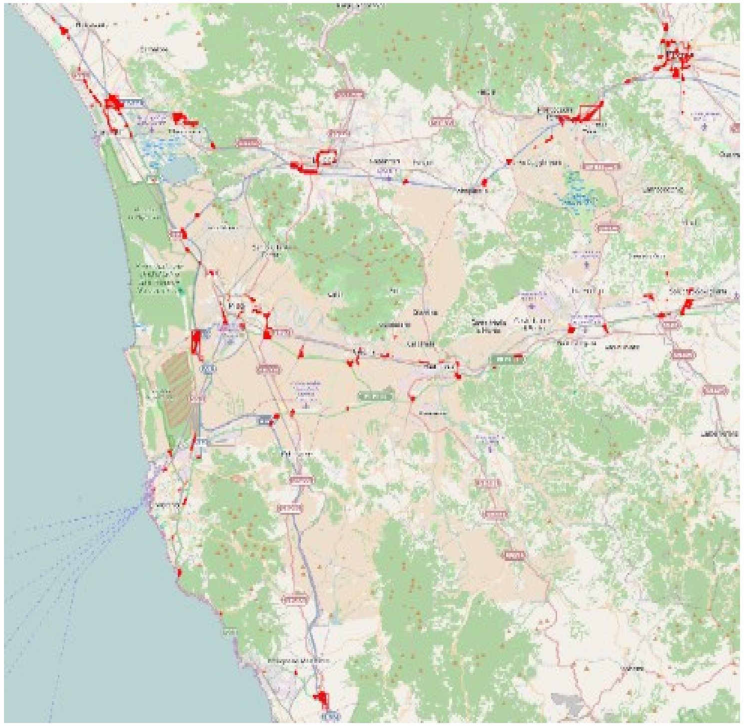
Step 2: Mobility profiles

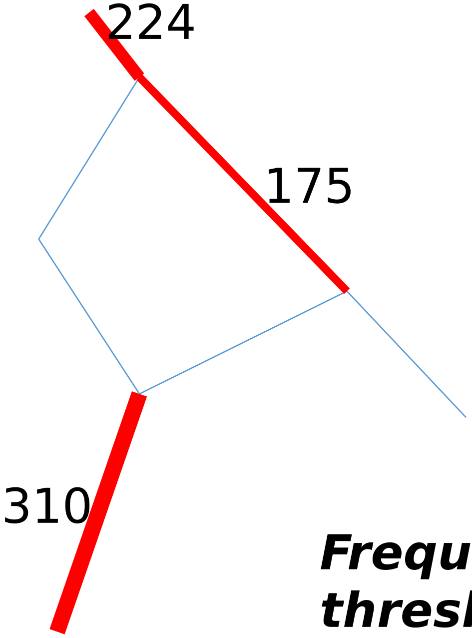
- Select only systematic movements.

User's systematic movement: L1 → L2

Frequently visited road segments

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



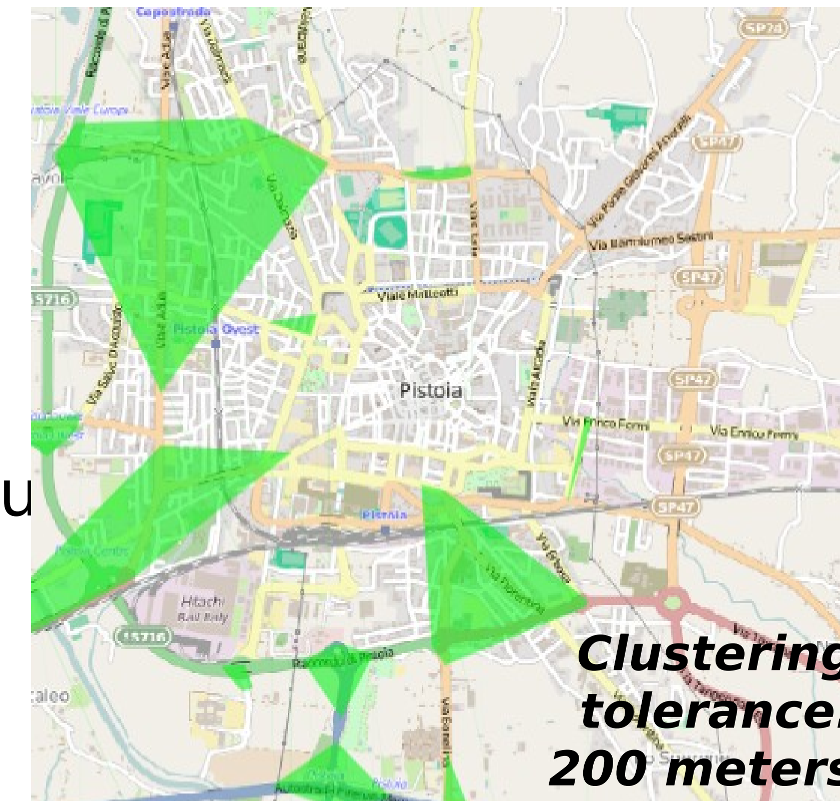


Frequency threshold: 150 users

Candidate areas for a mall

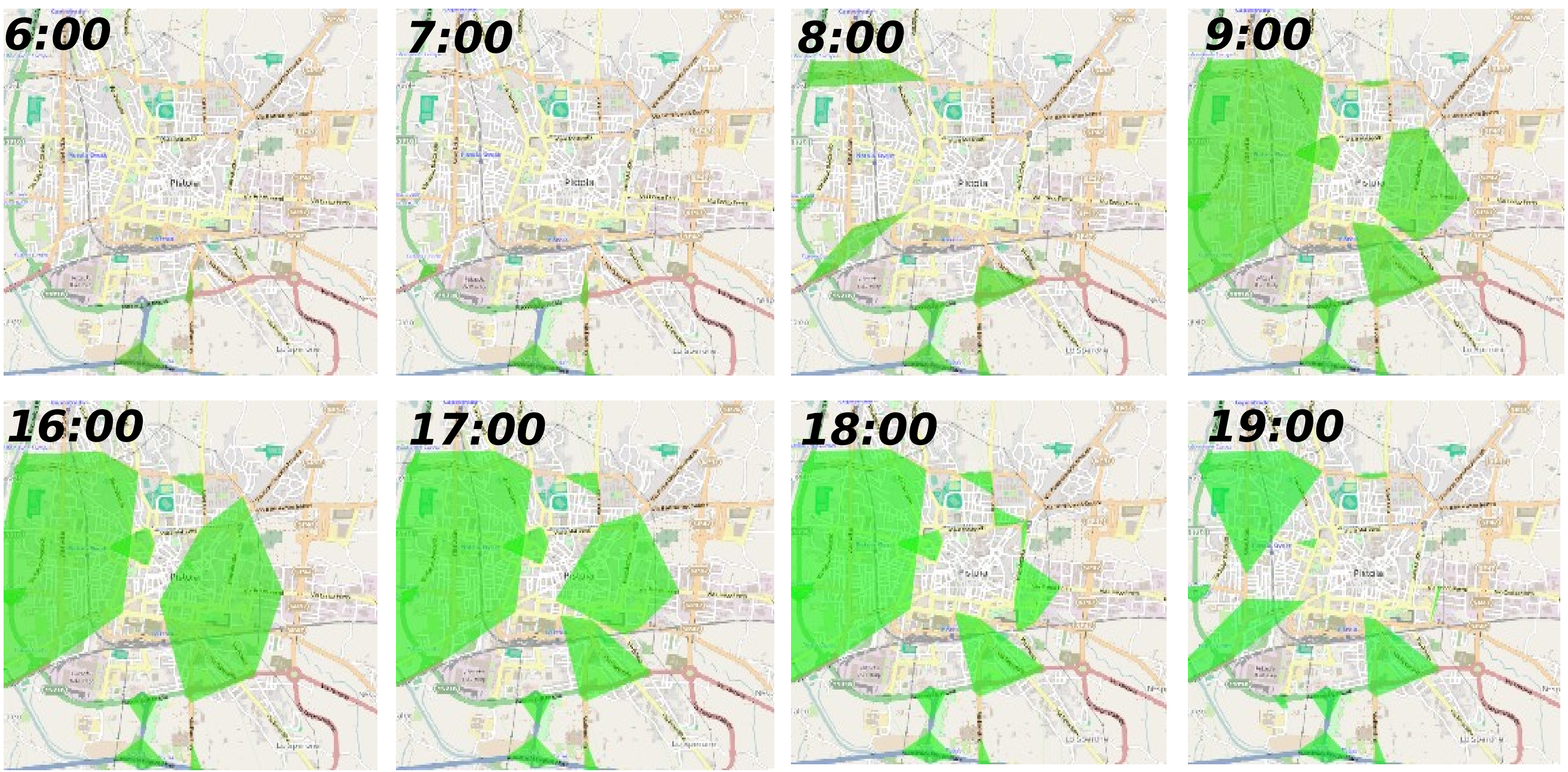
Using a spatial clustering we can extract cluster of frequent road segments which are spatially close each other.

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



Temporal evolution

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



Services Towards Corporate Users

Monitoring Driving-based Segmentation

Segmentation and monitoring

- Mobility application scenario of the LIFT European project

LIFT

USING LOCAL INFERENCE IN MASSIVELY DISTRIBUTED SYSTEMS

SEVENTH FRAMEWORK PROGRAMME

- Focused on distributed monitoring technologies

Scenario context & motivation

- **Customer segmentation**: a marketing strategy that involves dividing a broad target market into subsets of consumers who have common needs

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

- **Needs**: car insurance companies would like to define customer segments that capture different driving profiles
  - Each segment could then be offered suitable contract conditions
- **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

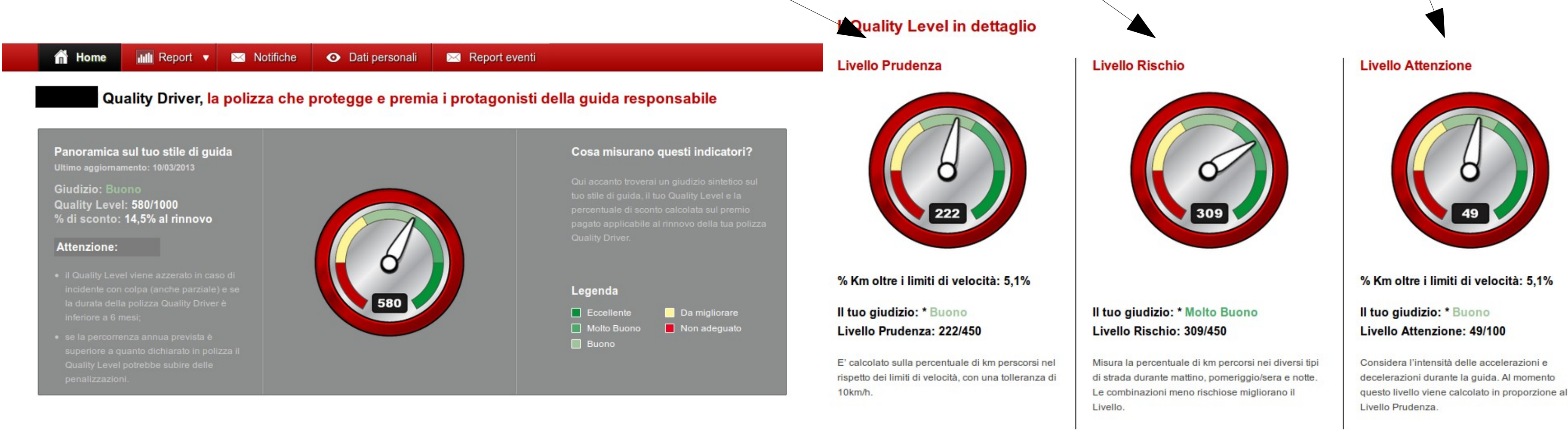
Step 1: Features for individual mobility behaviors

- 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

Basic aggregates

- Phighway = % km on highways
- Pcity = % km inside cities
- Length_arc_crowded = km on 20% most crowded roads
- Pnight = % km in night time

Aggregates on spatial / temporal selection

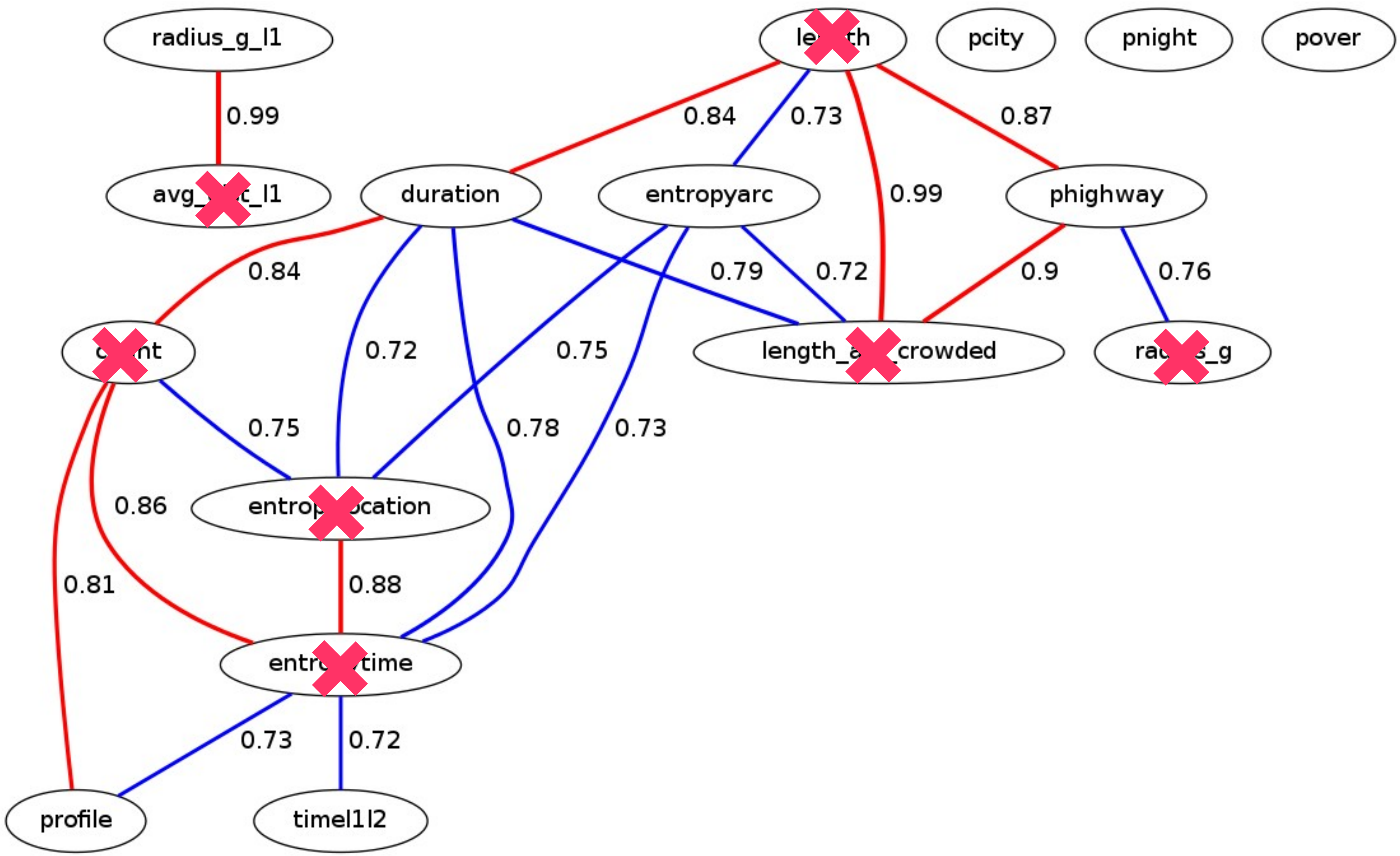
- Pover = % km over speed limit
- Profile = % of km on systematic trips

Count of events

- 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

Spatial/Temporal distribution

Correlation analysis



Features over sliding window

- ~~Length = traveled distance~~
- Duration = time spent driving
- ~~Count = number of trips~~

Basic aggregates

- Phighway = % km on highways
- Pcity = % km inside cities
- ~~Length_arc_crowded = km on 20% most crowded roads~~
- Pnight = % km in night time

Aggregates on spatial / temporal selection

- Pover = % km over speed limit
- Profile = % of km on systematic trips

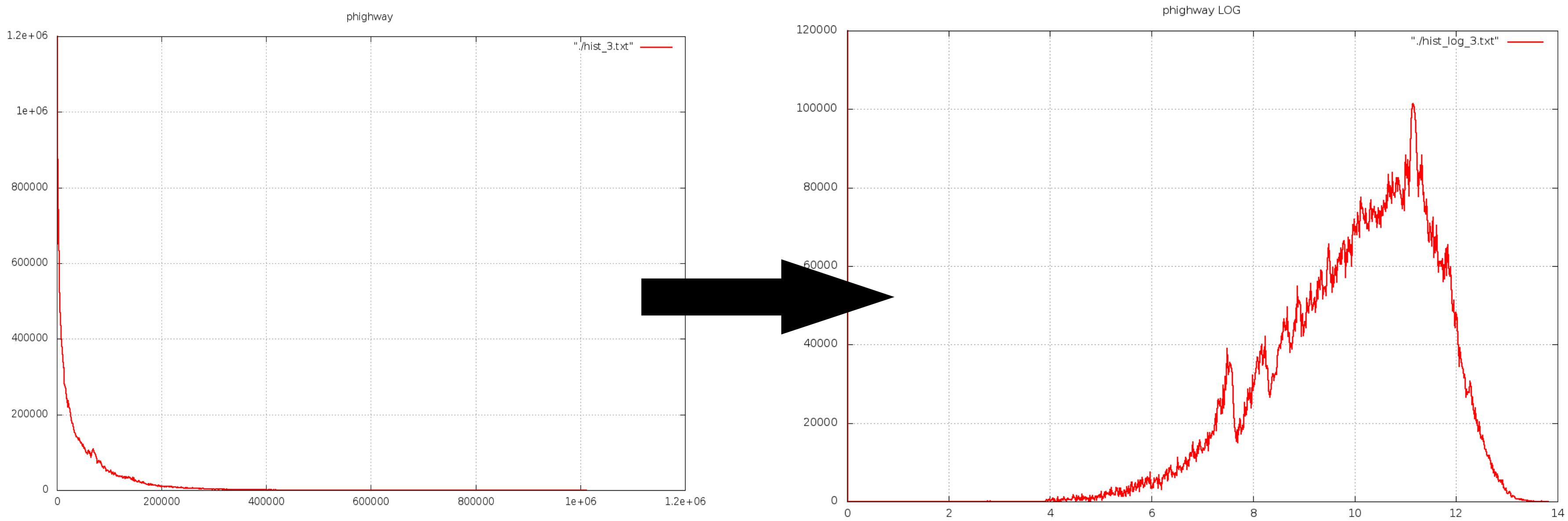
Count of events

- ~~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~~

Spatial/Temporal distribution

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

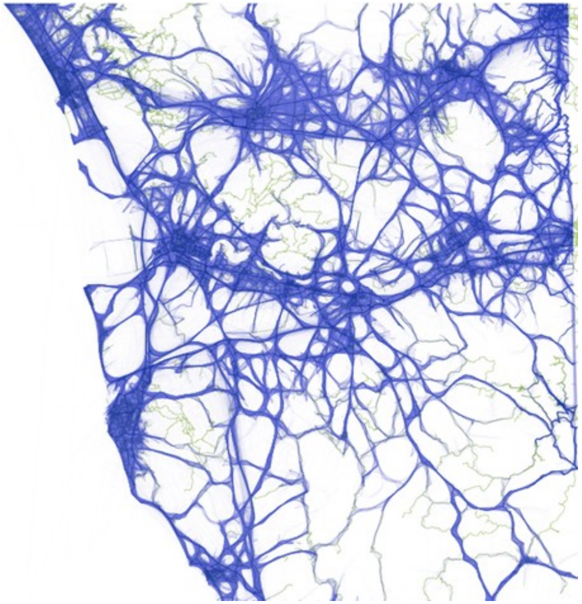
Cluster refinement

- 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

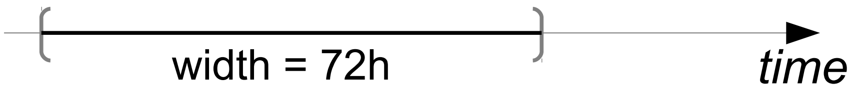
    $$d(p,c) > 2 \text{ median } \{ d(x,c) \mid x \text{ in cluster} \}$$

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

Experimental setting

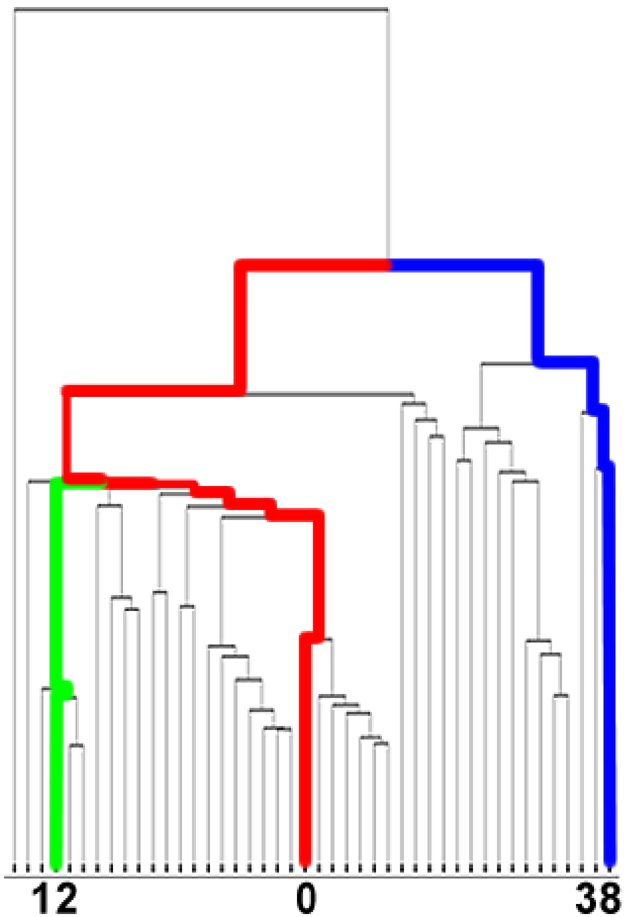


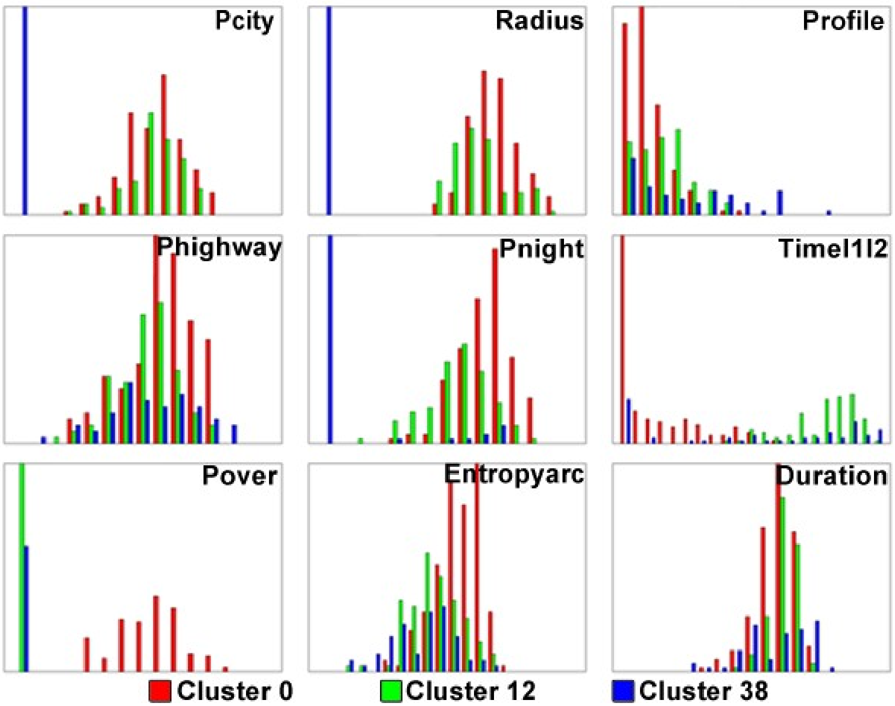
- GSP traces of an insurance company customers
  - 35 days monitoring
- Sample of ~11k vehicles moving in the area
- Short temporal thresholds for testing purposes
  - Compute driving indicators over a sliding window of 3 days



  - Update indicators every 15'
  - Most likely larger in a real application – parameter tuning to be done with domain experts

Experiments: clusters inspection



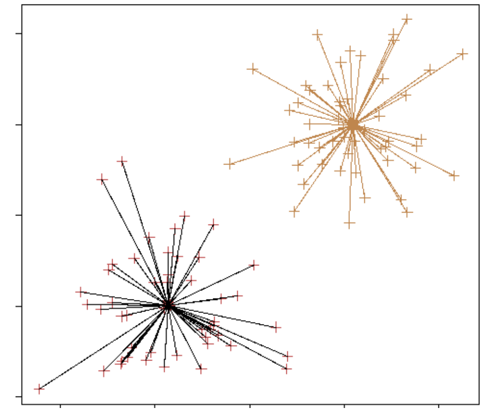


Explorers
Long-range commuters
Sunday drivers

(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

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$



(3) Driving profiles monitoring

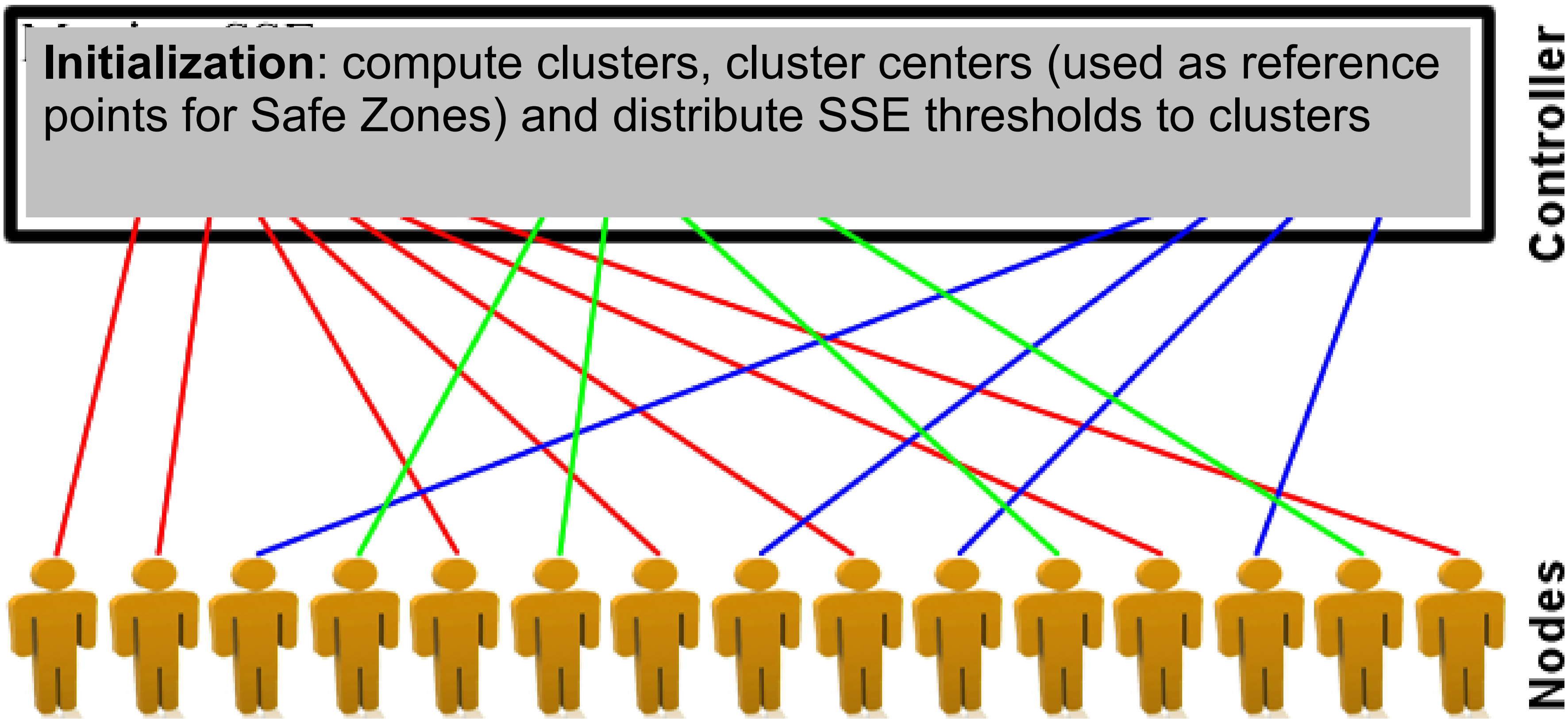
DEFINITION 1 (CLUSTER MONITORING PROBLEM). *Given a clustering $C = \{C_1, \ldots, C_k\}$ having initial SSE equal to SSE_0, and given a tolerance $\alpha \in \mathcal{R}^+$, we require to ensure that at each time instant t the following holds for the SSE of the (dynamic) dataset D_t:*

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

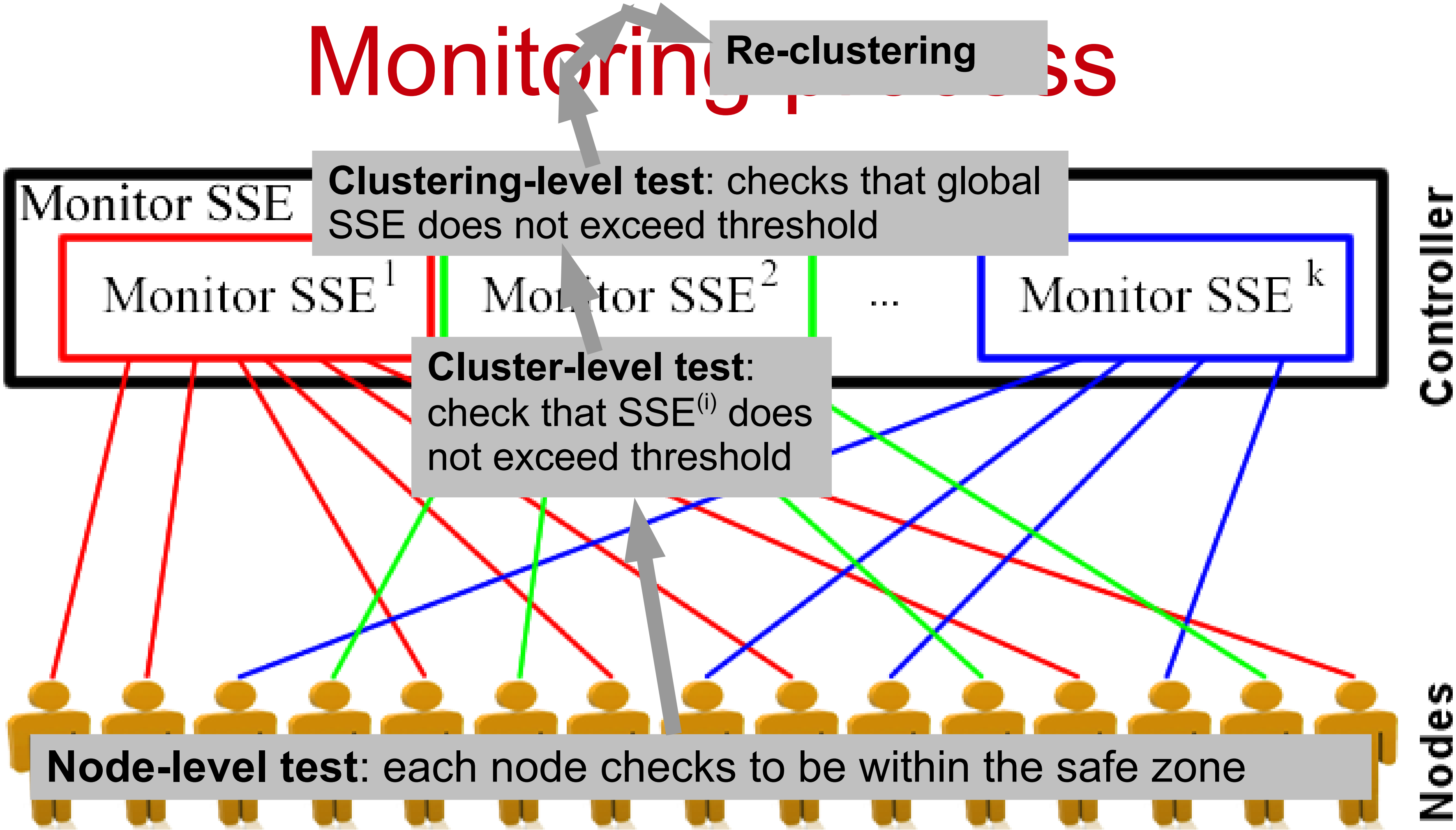
When that does not happen, a recomputation/update of cluster assignments should be performed.

Monitoring process

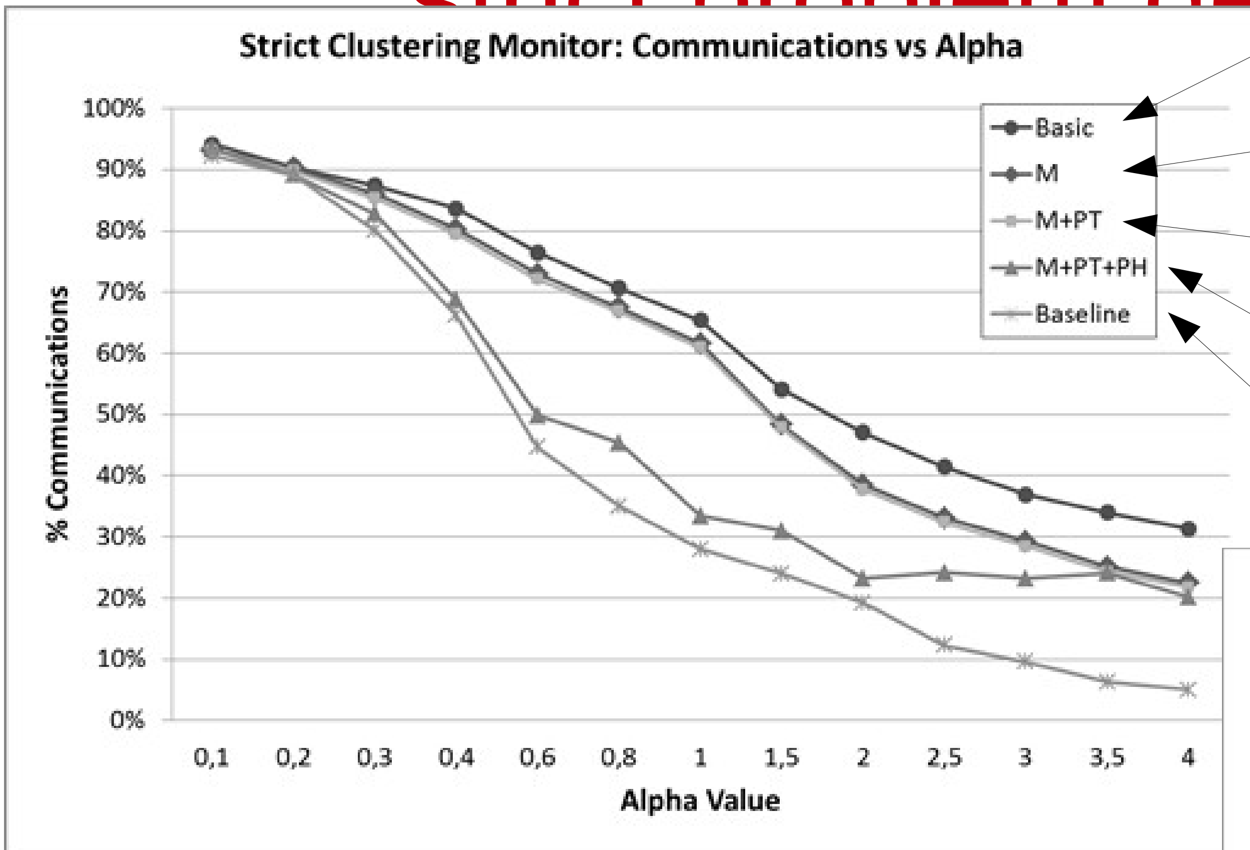
Initialization: compute clusters, cluster centers (used as reference points for Safe Zones) and distribute SSE thresholds to clusters



Monitoring process



Experiments: communications / strict problem def.



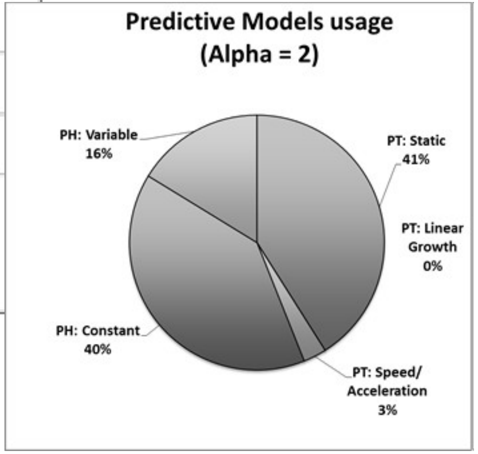
Balancing/memoryless

Balancing/memory

Trend Predictive Ms

History Predictive Ms

Oracle (no false alrms)



Communications from controller w/ broadcasting: between 1.23% and 2.34%, dominated by balancing

Services Towards Individual Users

Self-awareness

Self-awareness services

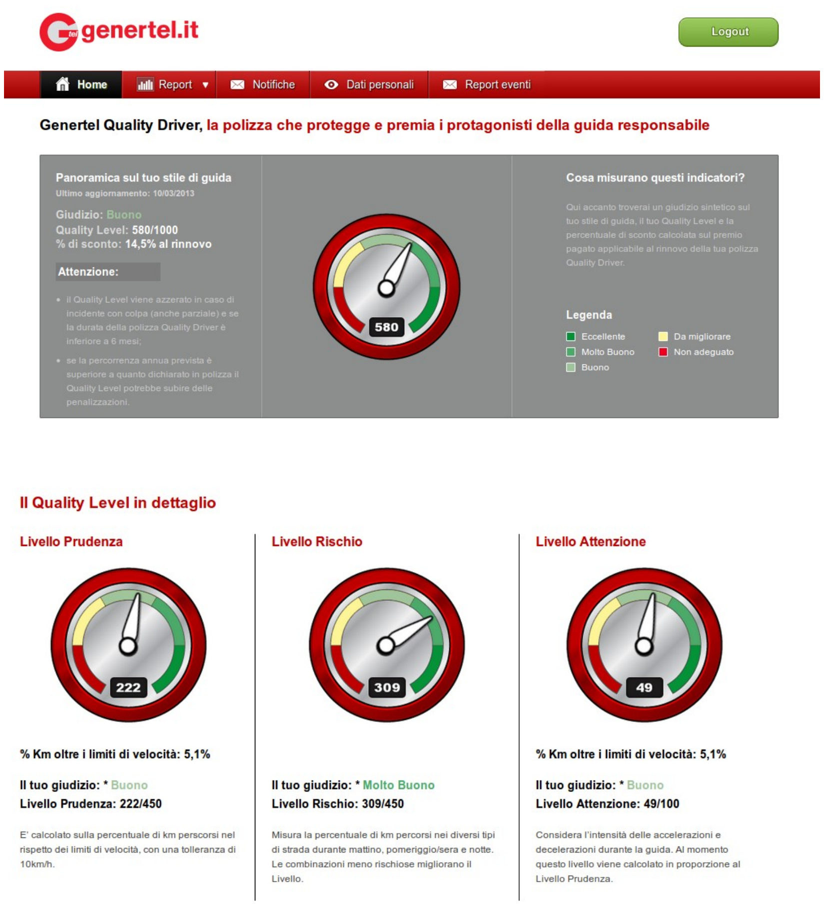
- Mobility-based specialization of self-awareness services for generic users
  - Provide summary of activity of the user
  - Provide comparison against collectivity

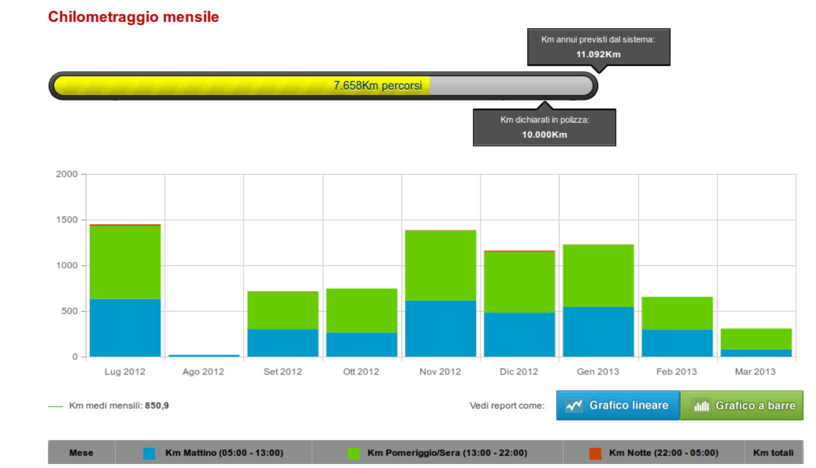
Self-awareness services

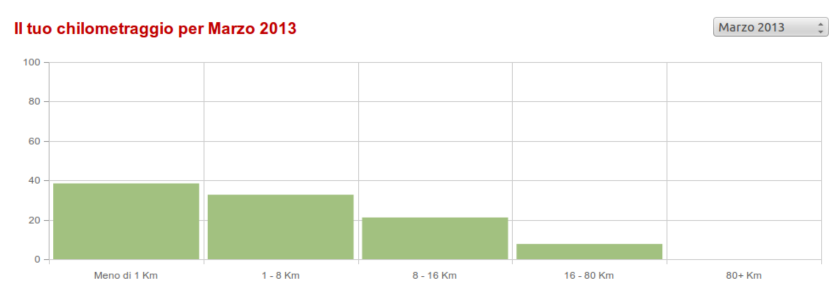
- Summaries based on
  - Temporal statistics
  - Spatial statistics / distributions
  - Movement aggregates

User's activity summaries

- A real example



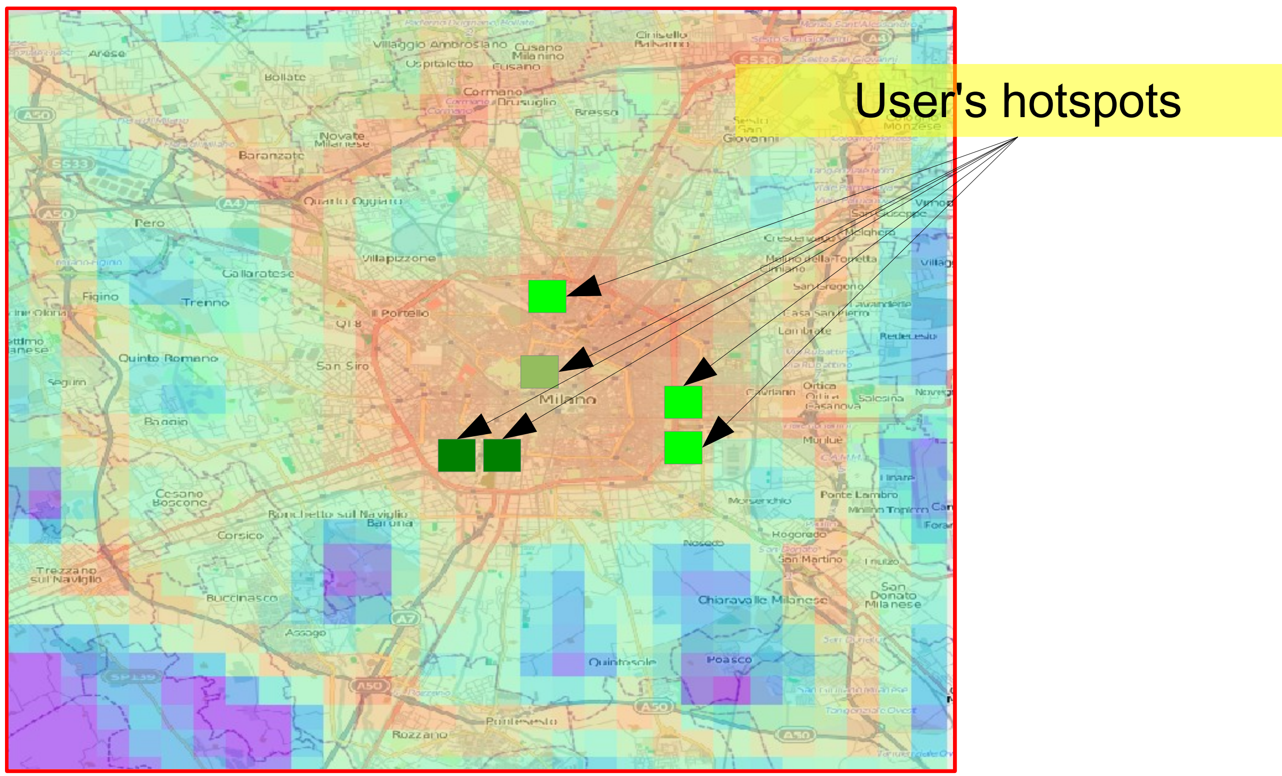




Comparison against collectivity

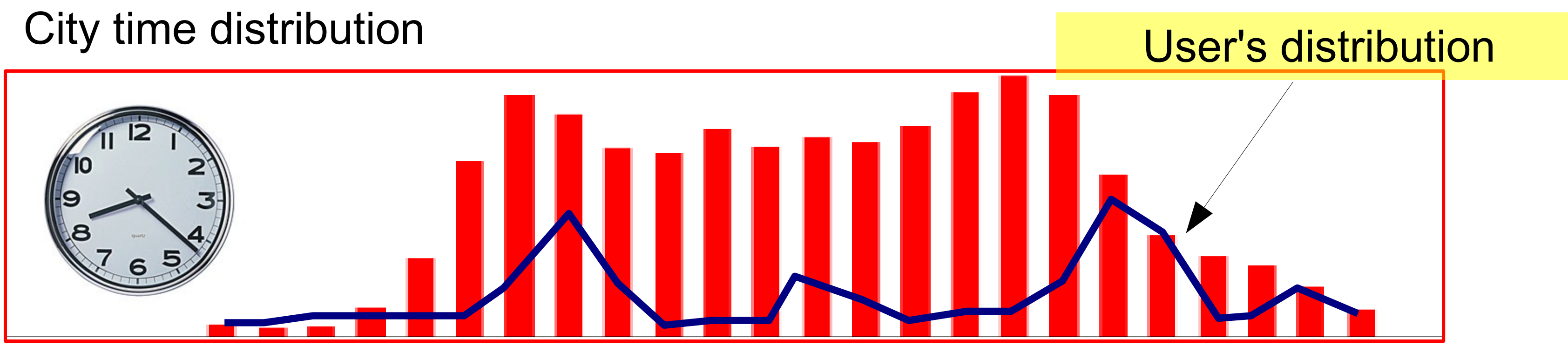
- In space

City hotspots



Comparison against collectivity

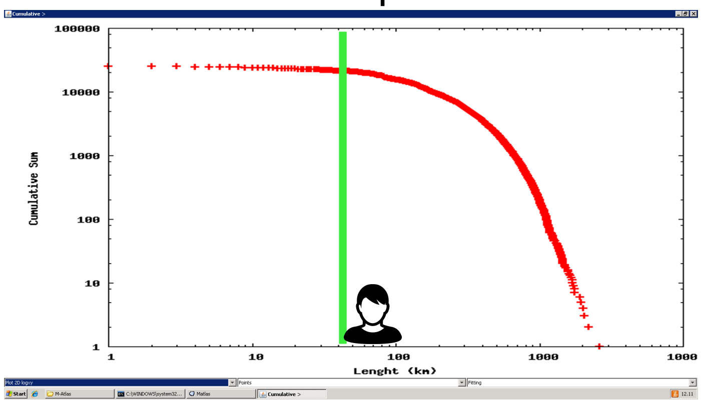
- In time



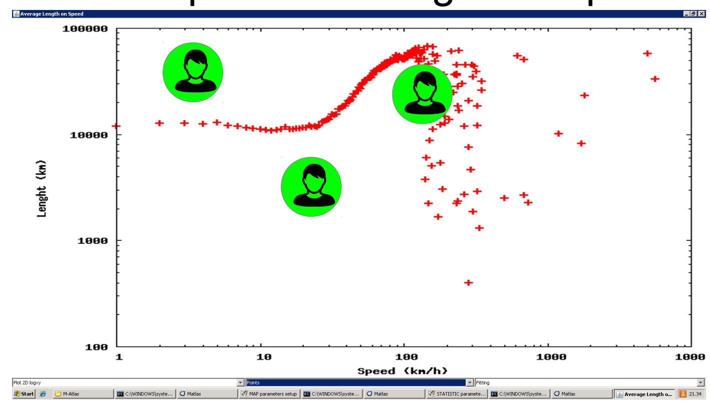
Comparison against collectivity

- On general statistics

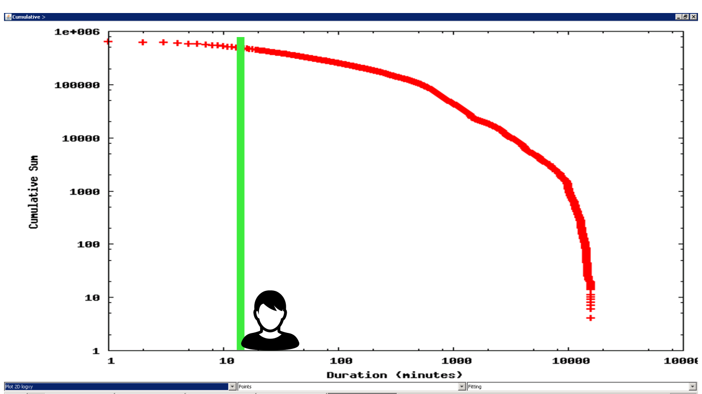
KM traveled per month



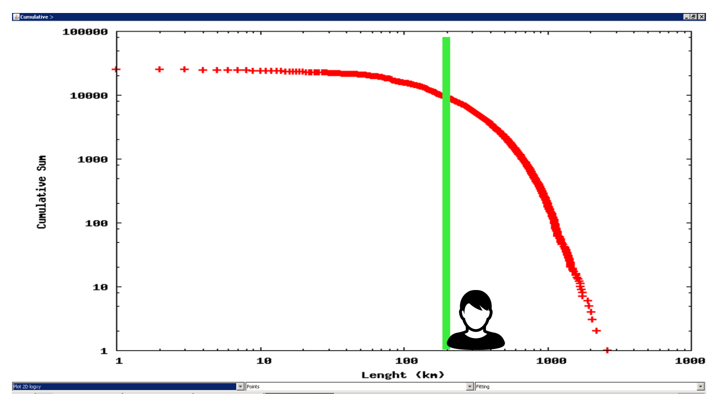
Speed vs. Length of trips



Total duration of travels



Radius of gyration



Services Towards Individual Users

Proactive Carpooling

Proactive car pooling

ICON

Application developed within the EU project ICON

Carpooling cycle
Context

- Several initiatives, especially on the web

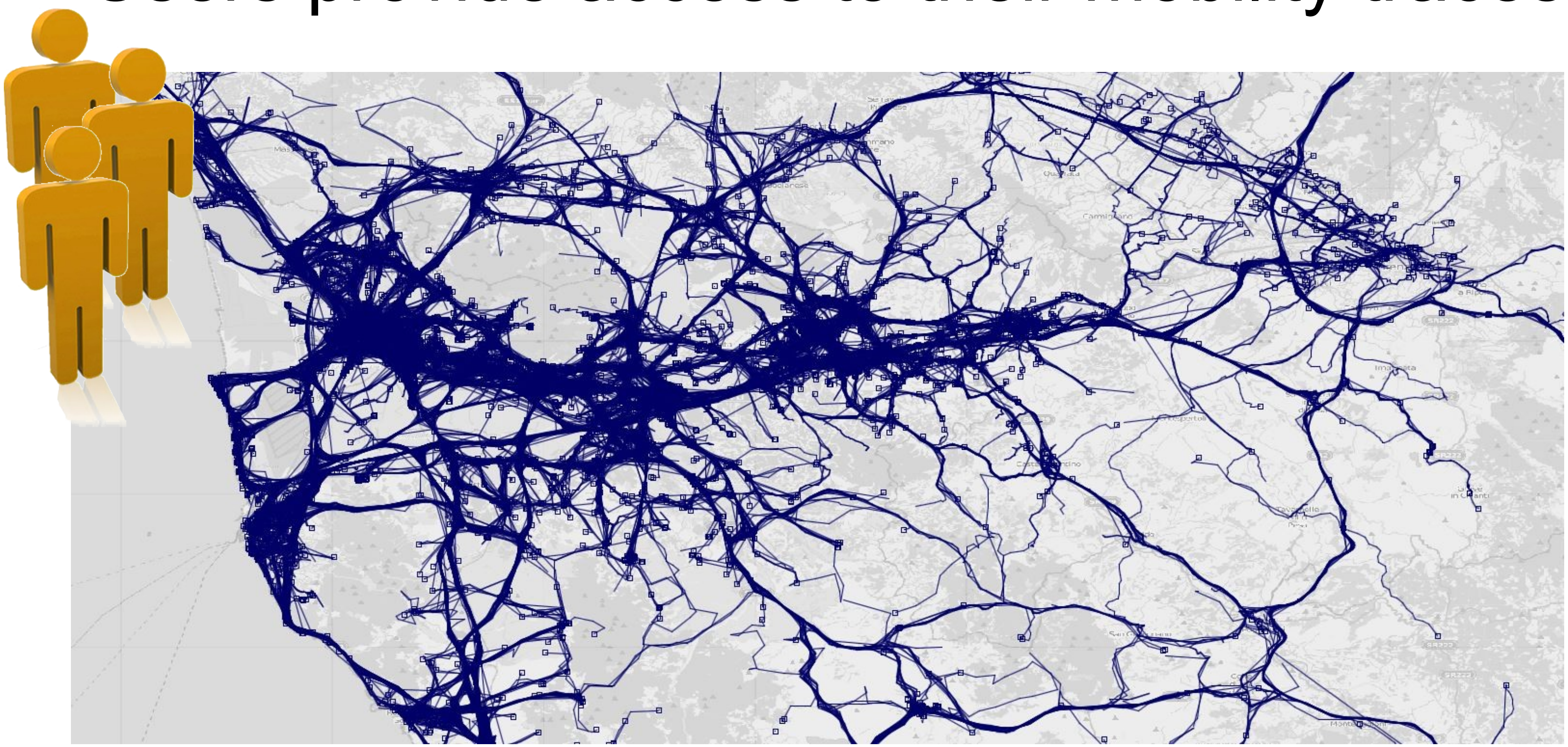
Carpooling cycle
Distinctive features

Traditional approach vs. Data-driven cycle

Traditional approach		Data-driven cycle
• Users manually insert and update their rides	→	• System autonomously detect systematic trips
• Users search and contact candidate pals	→	• System automatically suggest pairings
• Users make individual, “local” choice	→	• System seeks globally optimal allocation

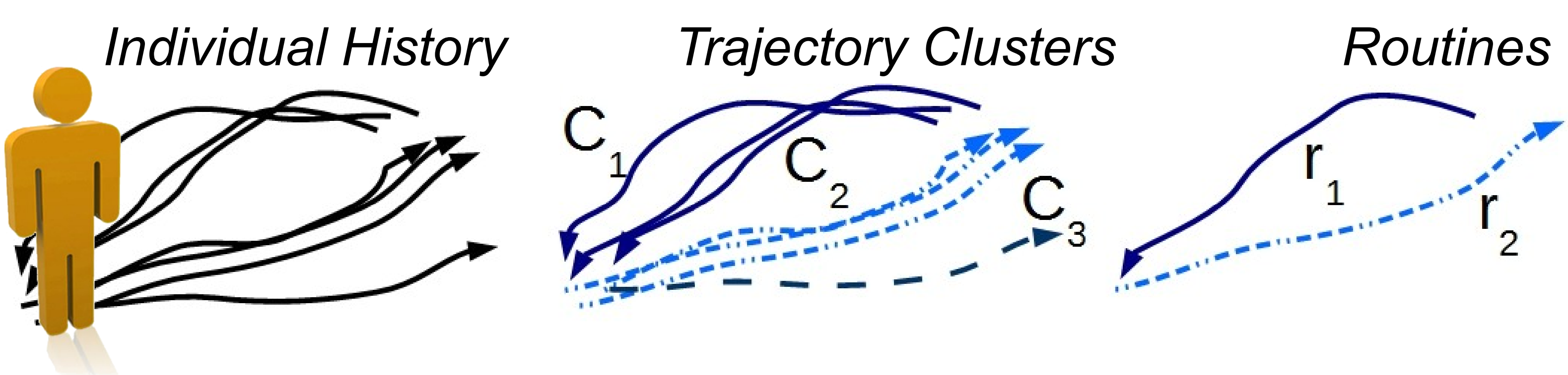
Carpooling cycle
Assumptions

- Users provide access to their mobility traces



Carpooling cycle
Step 1: Inferring Individual Systematic Mobility

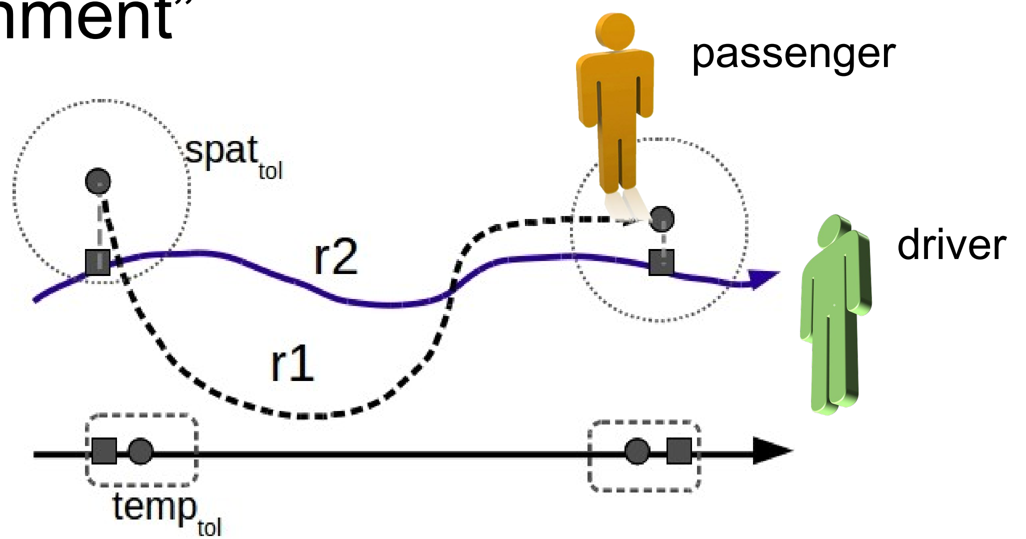
- Extraction of Mobility Profiles
  - Describes an abstraction in space and time of the systematic movements of a user.
  - Exceptional movements are completely ignored.
  - Based on trajectory clustering with noise removal



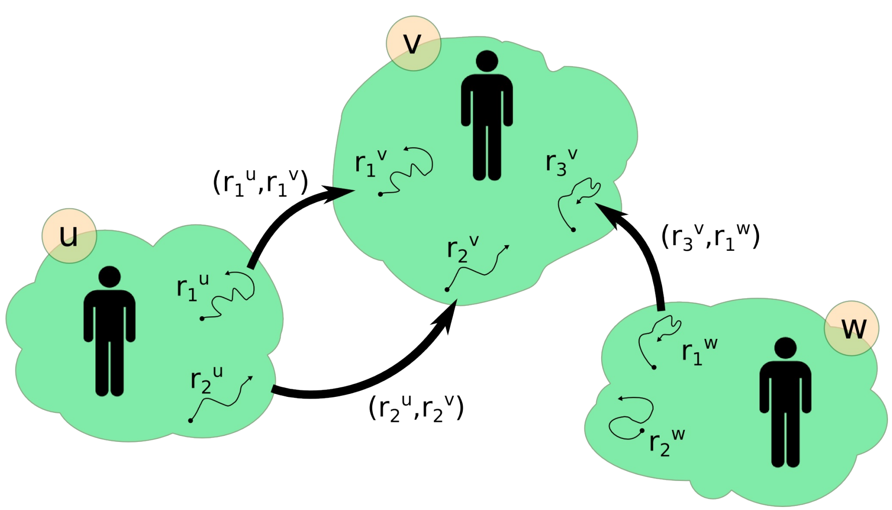
Carpooling cycle

Step 2: Build Network of possible carpool matches

- Based on “routine containment”
  - One user can pick up the other along his trip



- Carpooling network
  - Nodes = users
  - Edges = pairs of users with matching routines

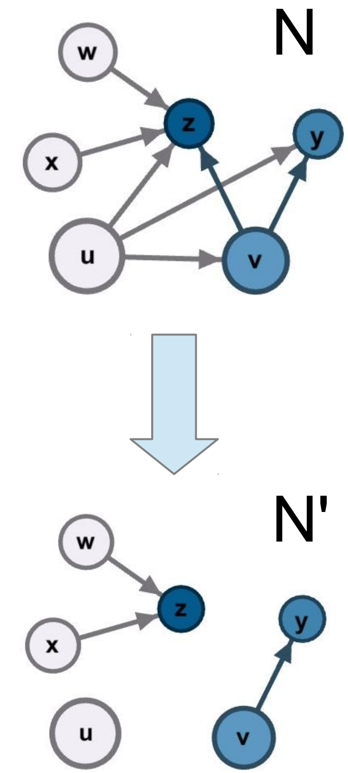


Carpooling cycle
Step 3: Optimal allocation of drivers-passengers

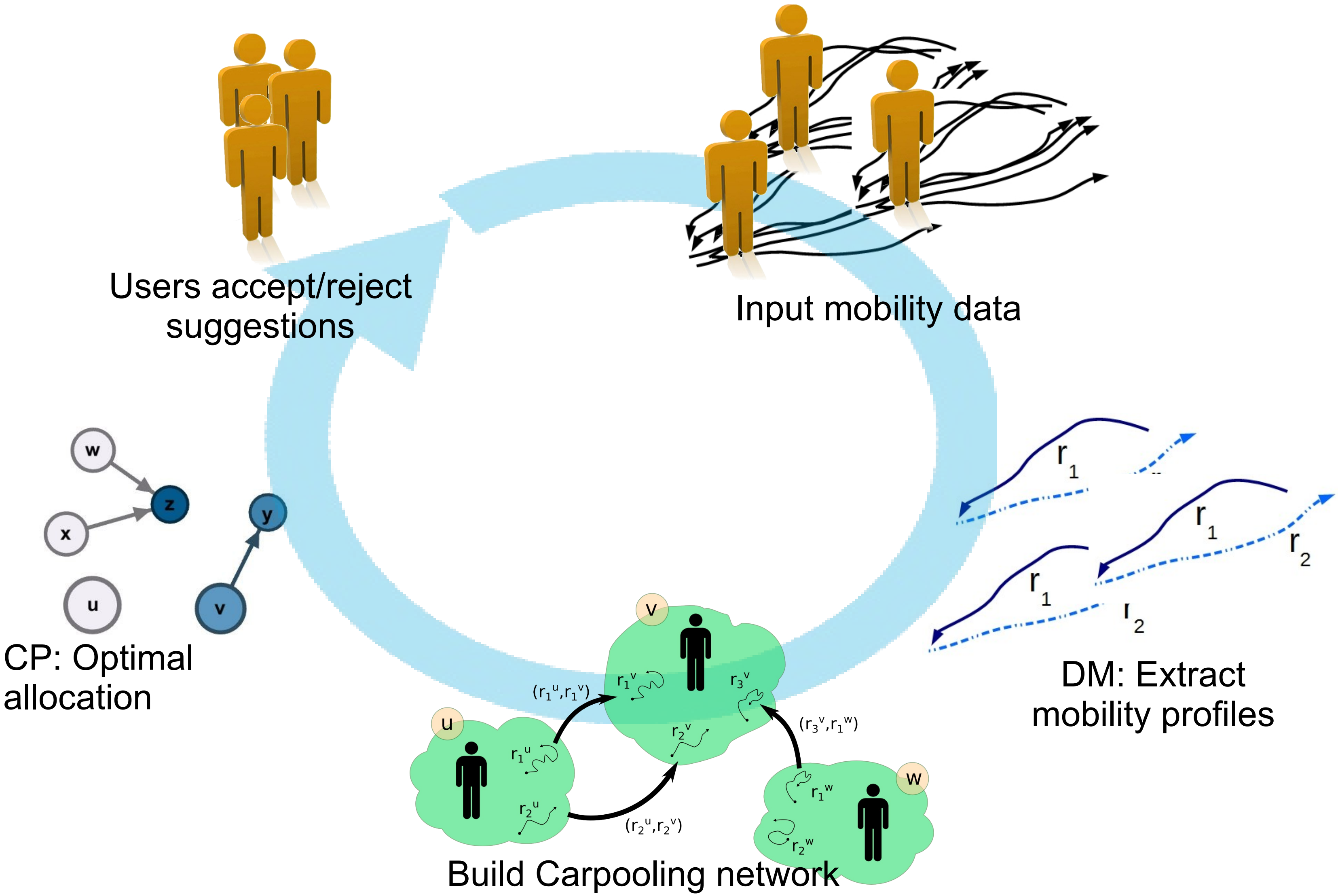
- Given a Carpooling Network N, select a subset of edges that minimizes |S|
  - S = set of circulating vehicles

provided that the edges are coherent, i.e.:

  - indegree(n)=0 OR outdegree(n)=0 (a driver cannot be a passenger)
  - indegree(n) ≤ capacity(n)



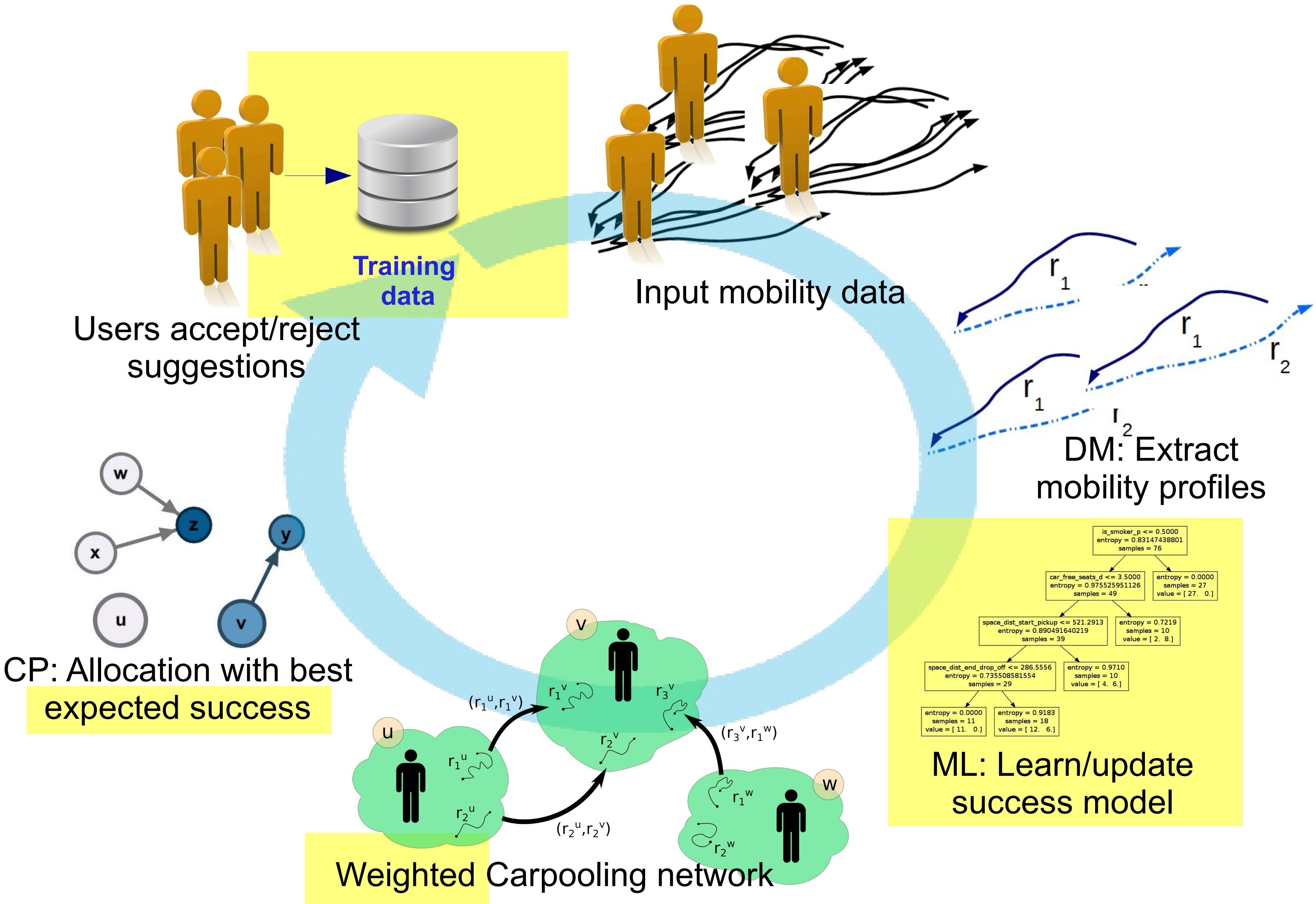
Carpooling cycle



Carpooling cycle
Improvement

- In carpooling (especially if proactive) users might not like the suggested matches
  - Impossible to know who will accept a given match
  - Modeling acceptance might improve results
- Two new components
  - **Learning** mechanism to guess success probability of a carpooling match
  - **Optimization** task exploits it to offer solution with best <u>expected</u> overall success

Carpooling cycle revised

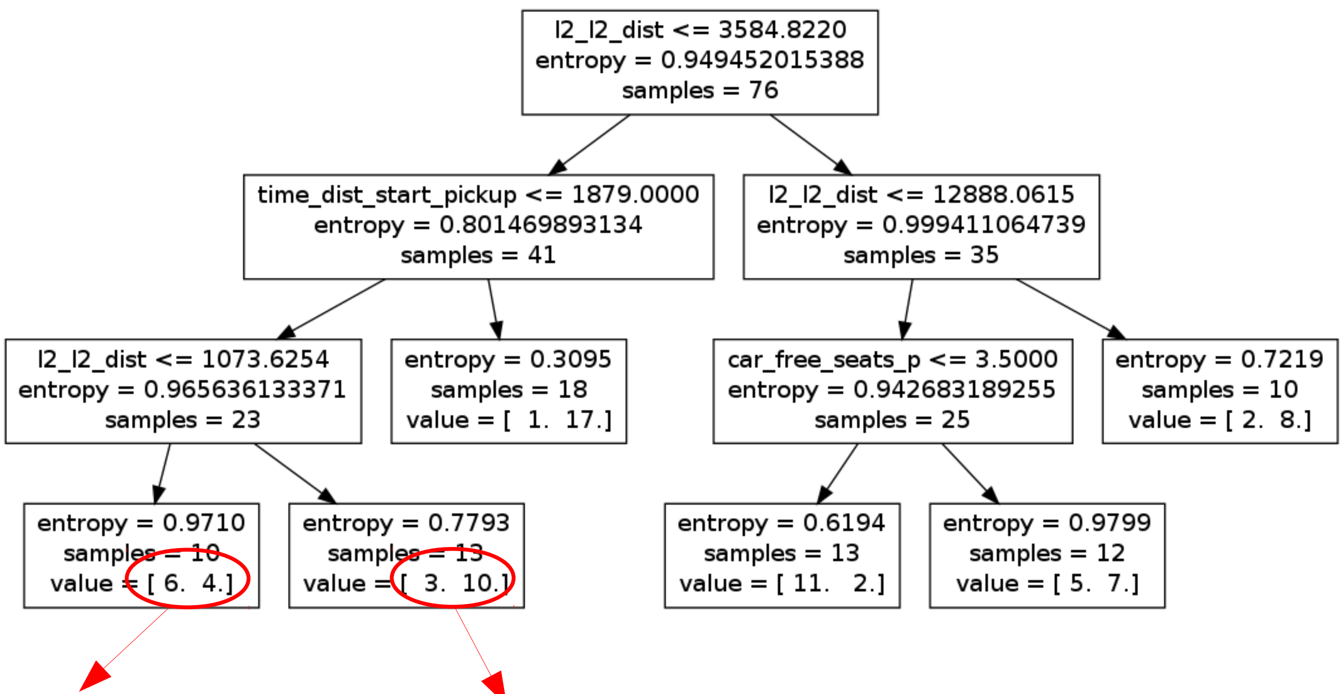


Carpooling cycle
Learning a success model

- **Input**: set of features describing a single carpooling pair
- **Output**: success probability p in [0,1]
- 36 Features adopted
  - **Ease of carpooling**: space_dist_start_pickup, space_dist_end_drop_off, time_dist_start_pickup, time_dist_end_drop_off, time_pick_up_get_off, start_together, end_together, distance_between_homes, dist_between_works
  - **Personal features** (of both driver and passenger): age, gender, marital_status, occupation, is_smoker, has_children, has_animals, car_free_seats → Cannot be inferred, need external data
  - **Past personal history in the service** (of both driver and passenger): last_driver_accepted, last_passenger_accepted, %_acceptance_driver, %_acceptance_passenger
  - **History of the two users together** (if any): last_accepted_pair, last_rejected_pair,%_accepted_pair

Carpooling cycle
Learning a success model

- Model selected: “probability estimation tree”
  → simple decision tree with assigned probabilities of prediction in the leaves



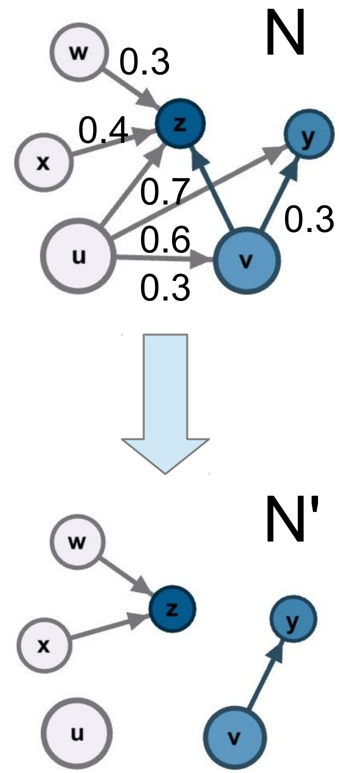
P(Yes) = 6/10 = 60% P(Yes) = 3/13 = 23%

Carpooling cycle
Revised optimization model

- Given a Carpooling Network N, select a subset W of edges that maximize
  - sum p(w) | w in W

provided that the edges are coherent, i.e.:

  - indegree(n)=0 OR outdegree(n)=0 (a driver cannot be a passenger)
  - indegree(n) ≤ capacity(n)



Carpooling cycle
Two usage scenarios

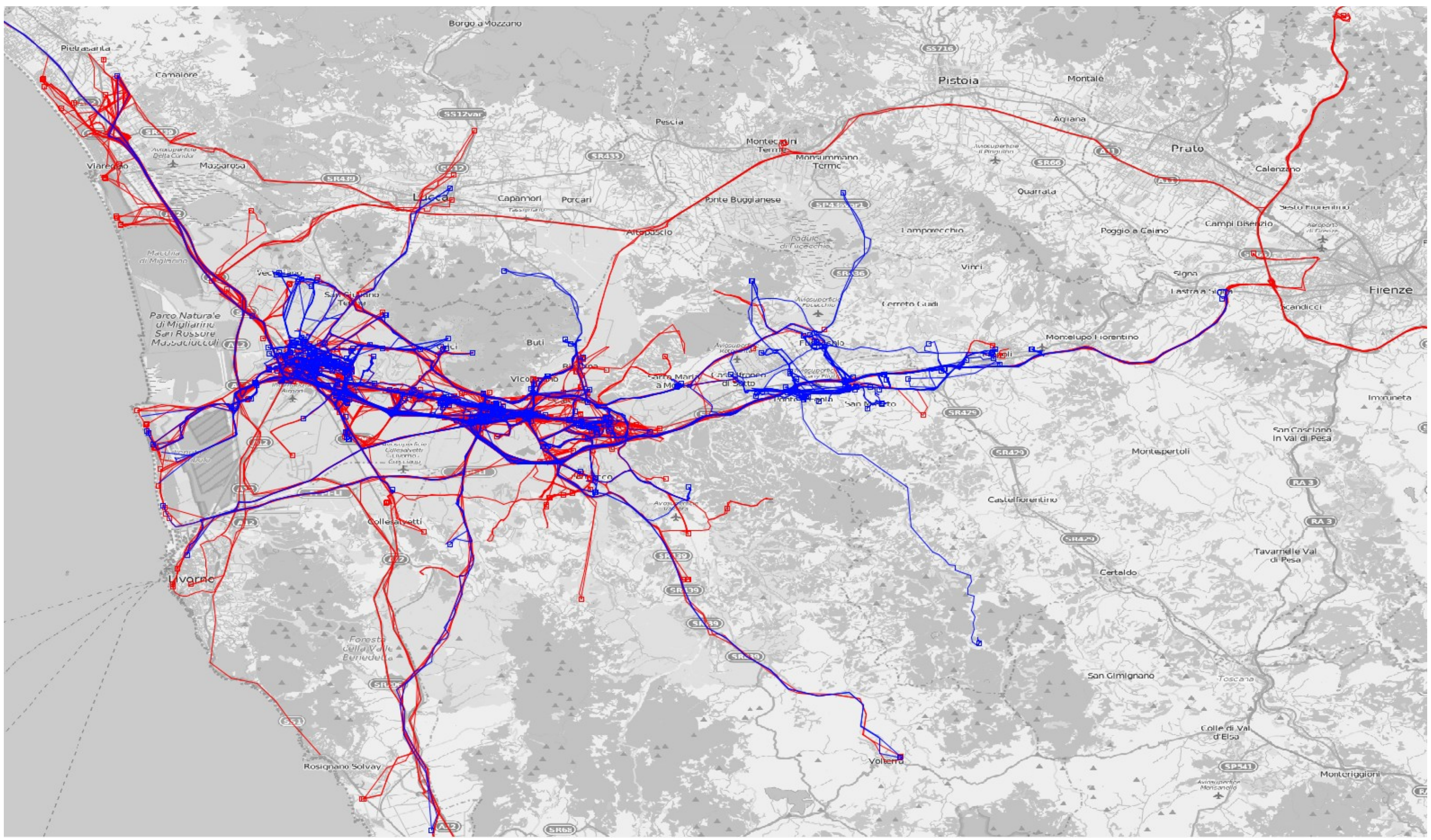
- Scenario 1:

  - Real service is implemented, with real users interacting (accept/reject suggestions)
- Scenario 2:
  - Simulation environment where the users' behaviour is simulated through a model

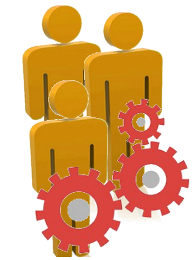
  - Mobility data is taken from historical traces
  - Useful to perform what-if analyses on
    - (i – social) effects of different users' behaviours
    - (ii – performances) effects of different learning strategies

Carpooling cycle
Scenario 2 – sample results

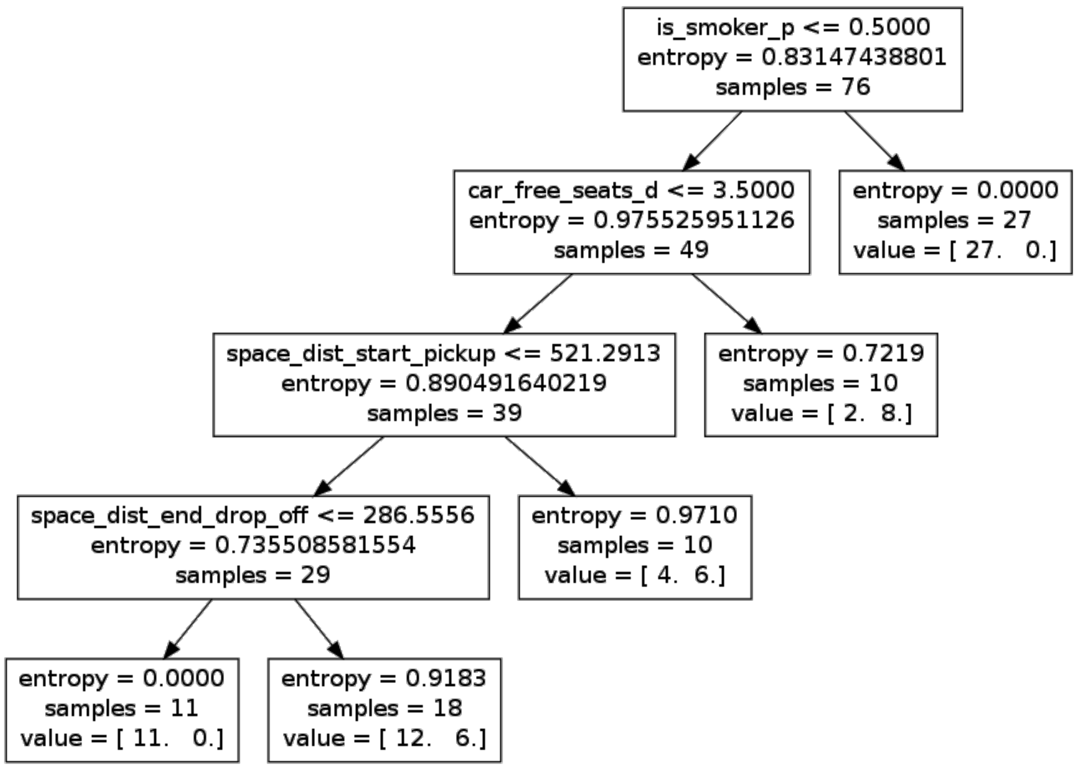
- Profiles involved in carpooling network



Carpooling cycle
Scenario 2 – sample results



- Prediction models



Iteration 0

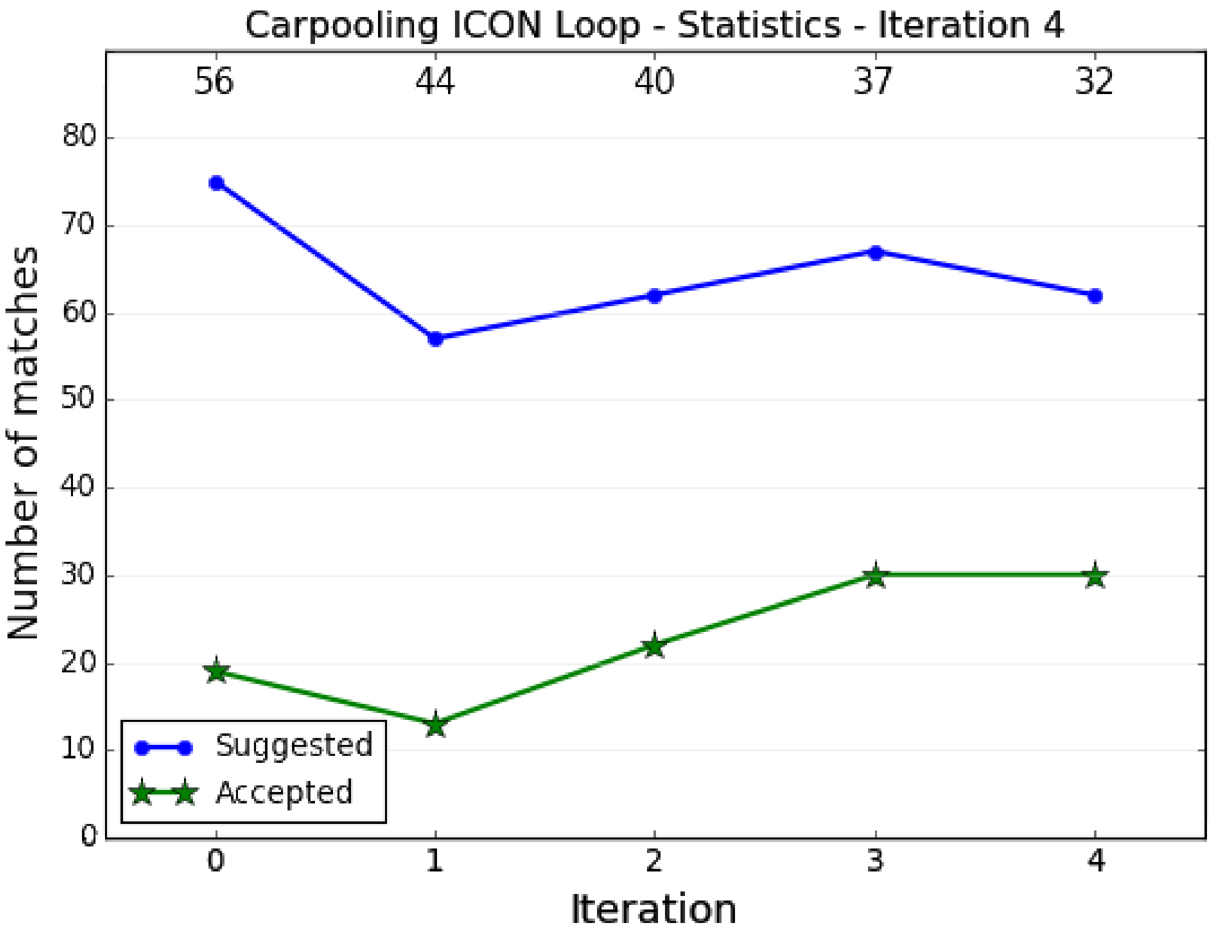
is_smoker_p : 0.51763342041
car_free_seats_d : 0.196822768067
space_dist_end_drop_off : 0.161445930025
space_dist_start_pickup : 0.124097881498
time_dist_start_pickup : 0.0
last_accepted_pair : 0.0
l1_l1_dist : 0.0
age_d : 0.0
gender_p : 0.0
has_children_p : 0.0

Iteration 4

last_accepted_pair : 0.300609683595
%_accepted_pair : 0.18422352604
gender_d : 0.121782490916
is_smoker_d : 0.096830535215
l1_l1_dist : 0.0947711528021
is_smoker_p : 0.0921934235296
age_p : 0.0549409842076
gender_p : 0.0396236591312
time_dist_start_pickup : 0.00874162379163
car_free_seats_d : 0.00628292077177

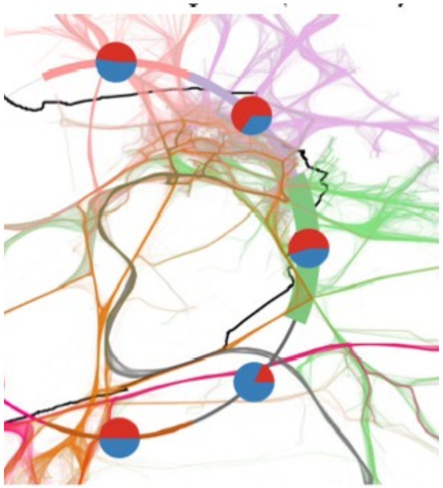
Carpooling cycle
Scenario 2 – sample results

- Performances

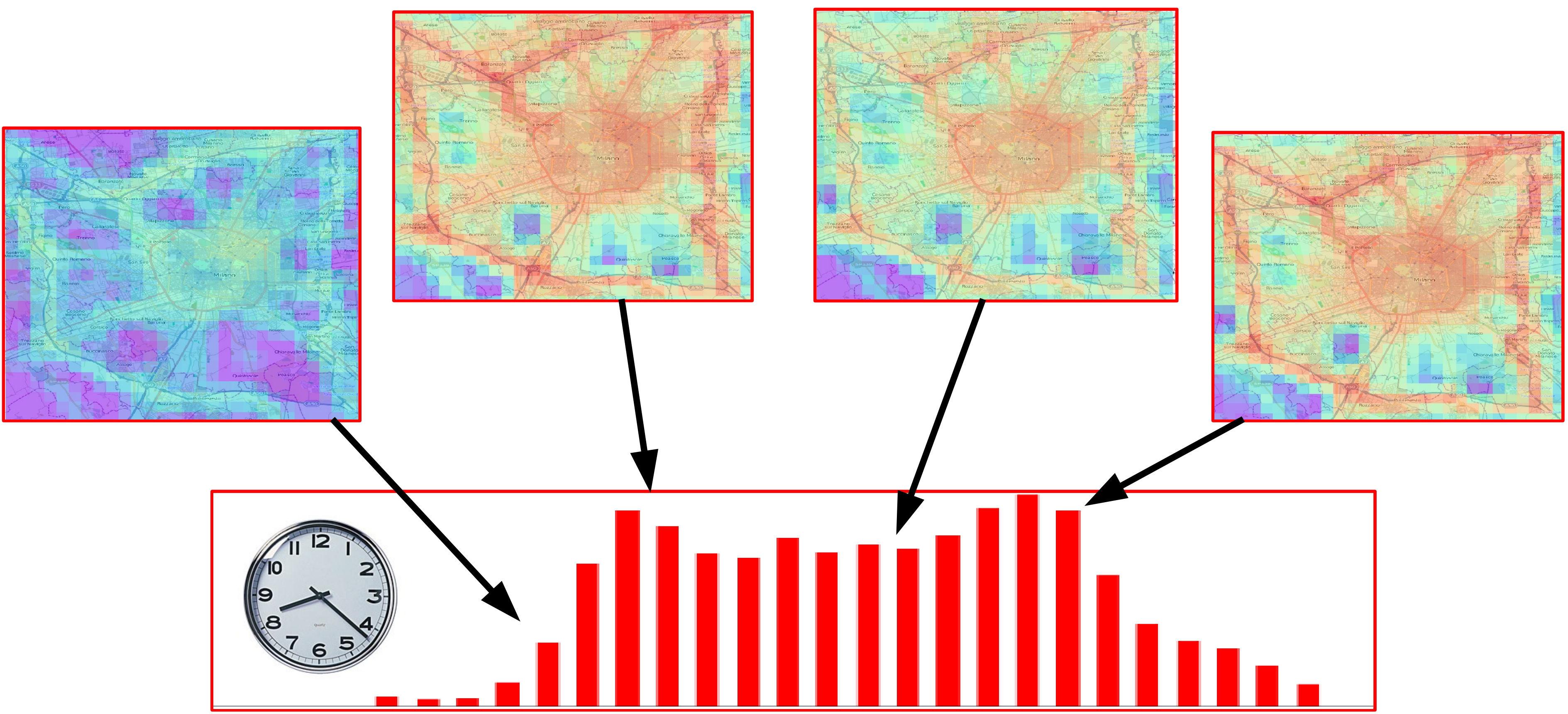


Services Towards Public Sector

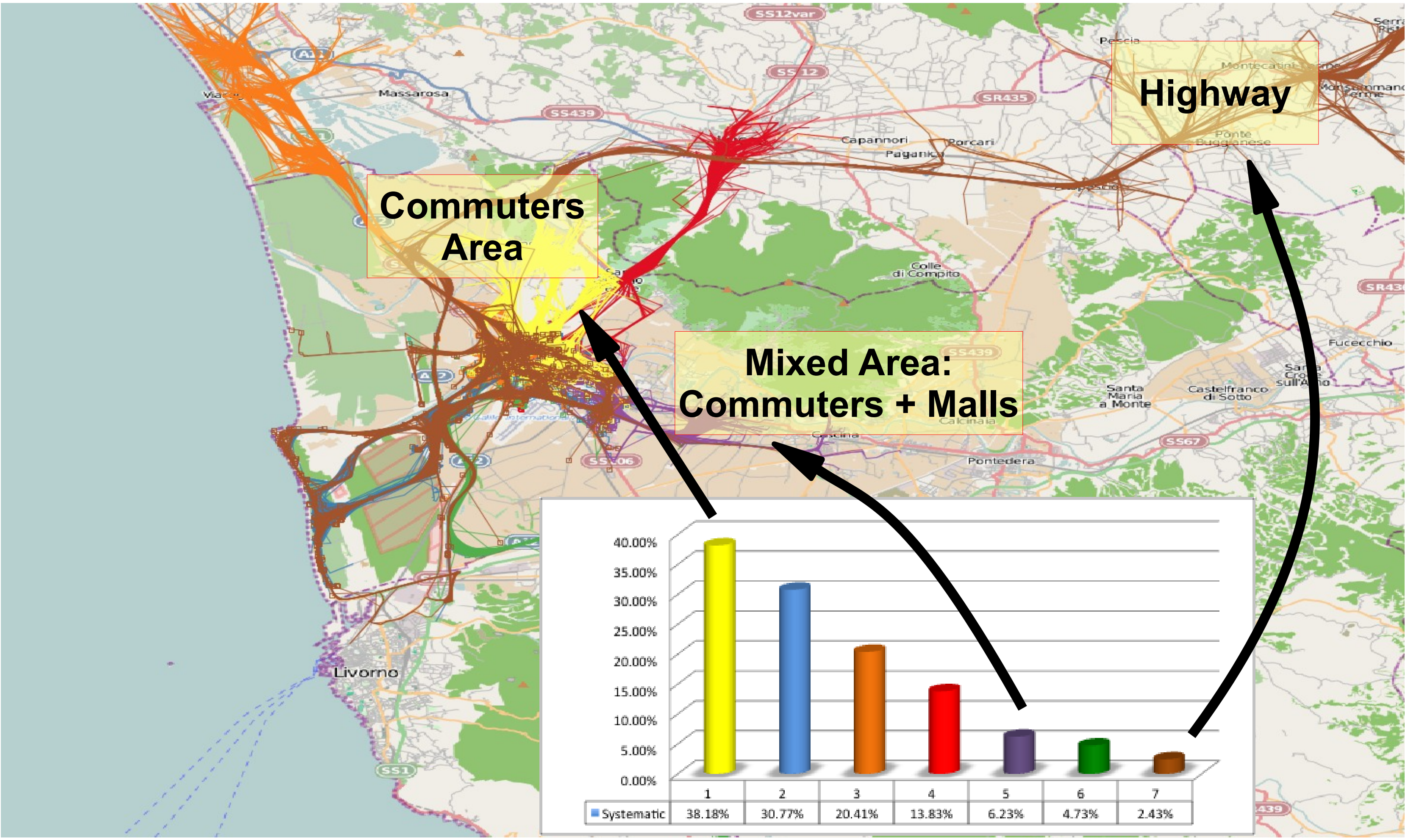
Urban Mobility Atlas



Dynamics of urban mobility



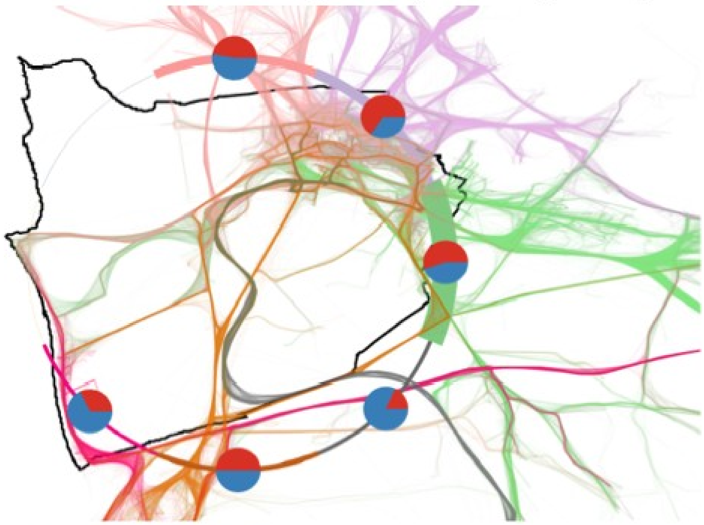
Impact of Systematic Mobility



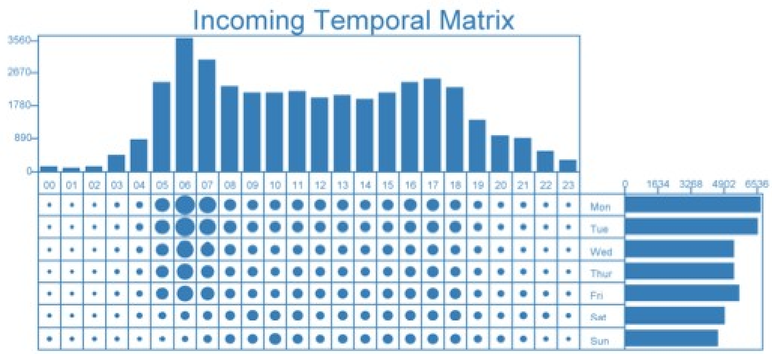
Access Routes
Systematic Mobility (%)

Pisa – Incoming traffic

Incoming Traffic (38.464 Trajectories)

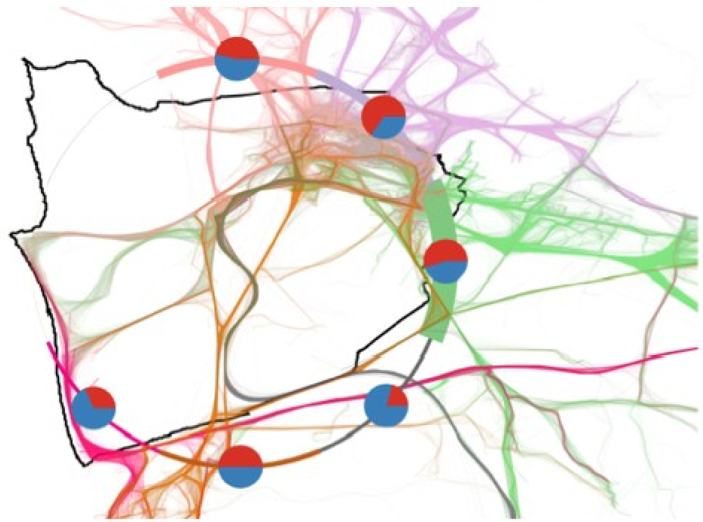


	City	Traj	Perc
NORD 32%	San Giuliano T..	4.816	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
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



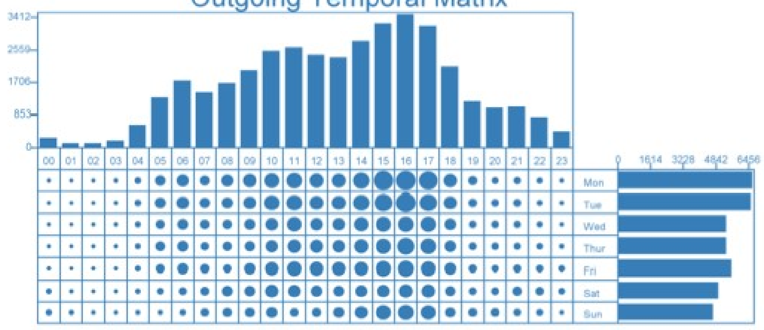
Pisa – Outgoing Traffic

Outgoing Traffic (38.271 Trajectories)

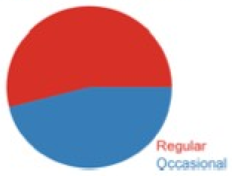


	City	Traj	Perc
NORD 32%	San Giuliano T...	4.842	62%
	Vecchiano	1.418	93%
	Viareggio	1.117	99%
	Lucca	886	67%
	Camaiore	329	96%
OVEST 0%			
SUD 13%	Livorno	2.812	92%
	Collesalvetti	565	51%
	Rosignano Mari...	143	44%
	Fauglia	130	19%
	Cecina	123	45%
EST 54%	Cascina	7.253	97%
	San Giuliano T...	2.860	37%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

Outgoing Temporal Matrix

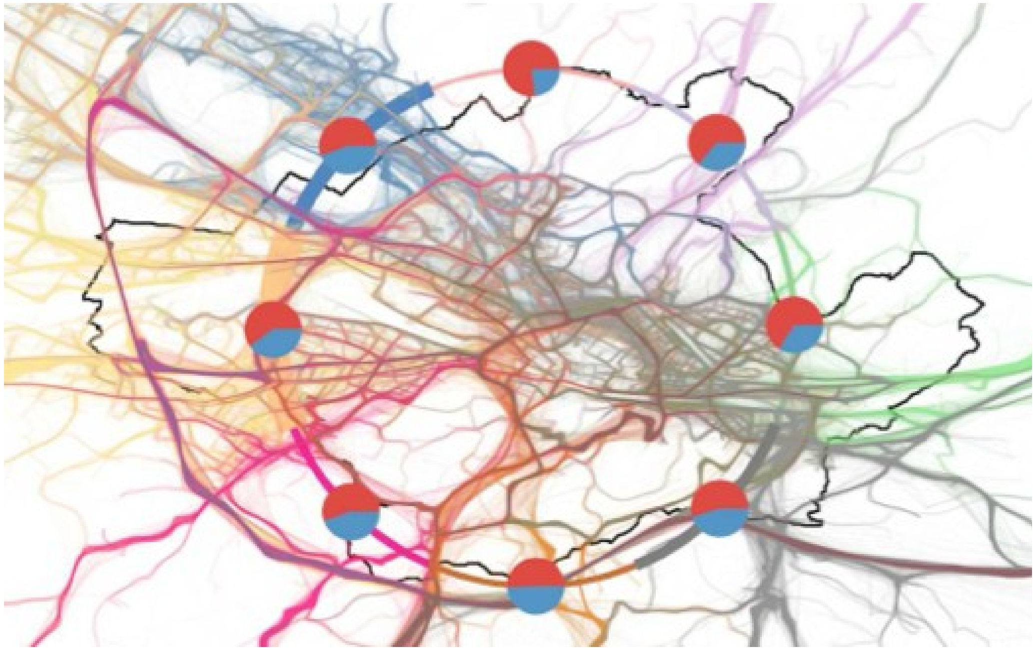


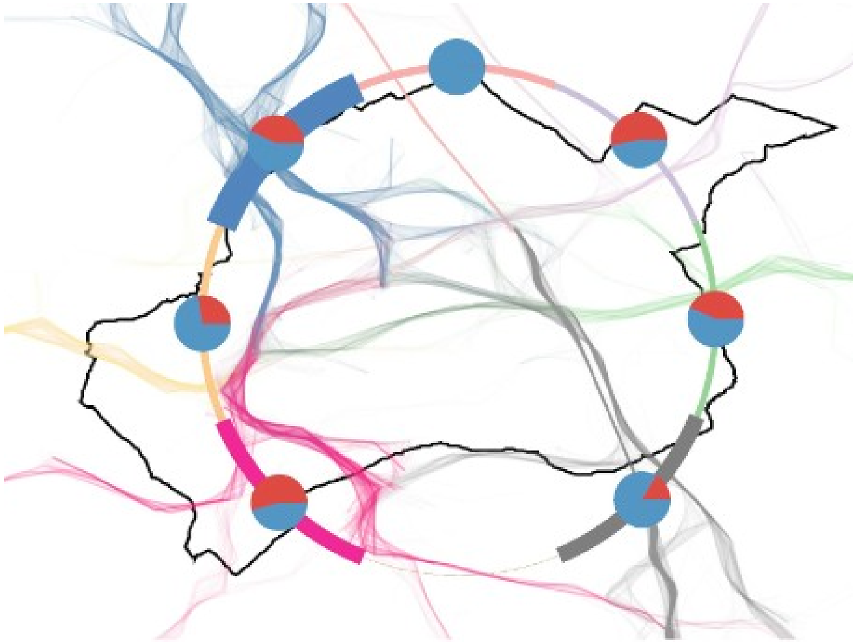
Regular VS Occasional



… and Comparison

Florence	Montepulciano

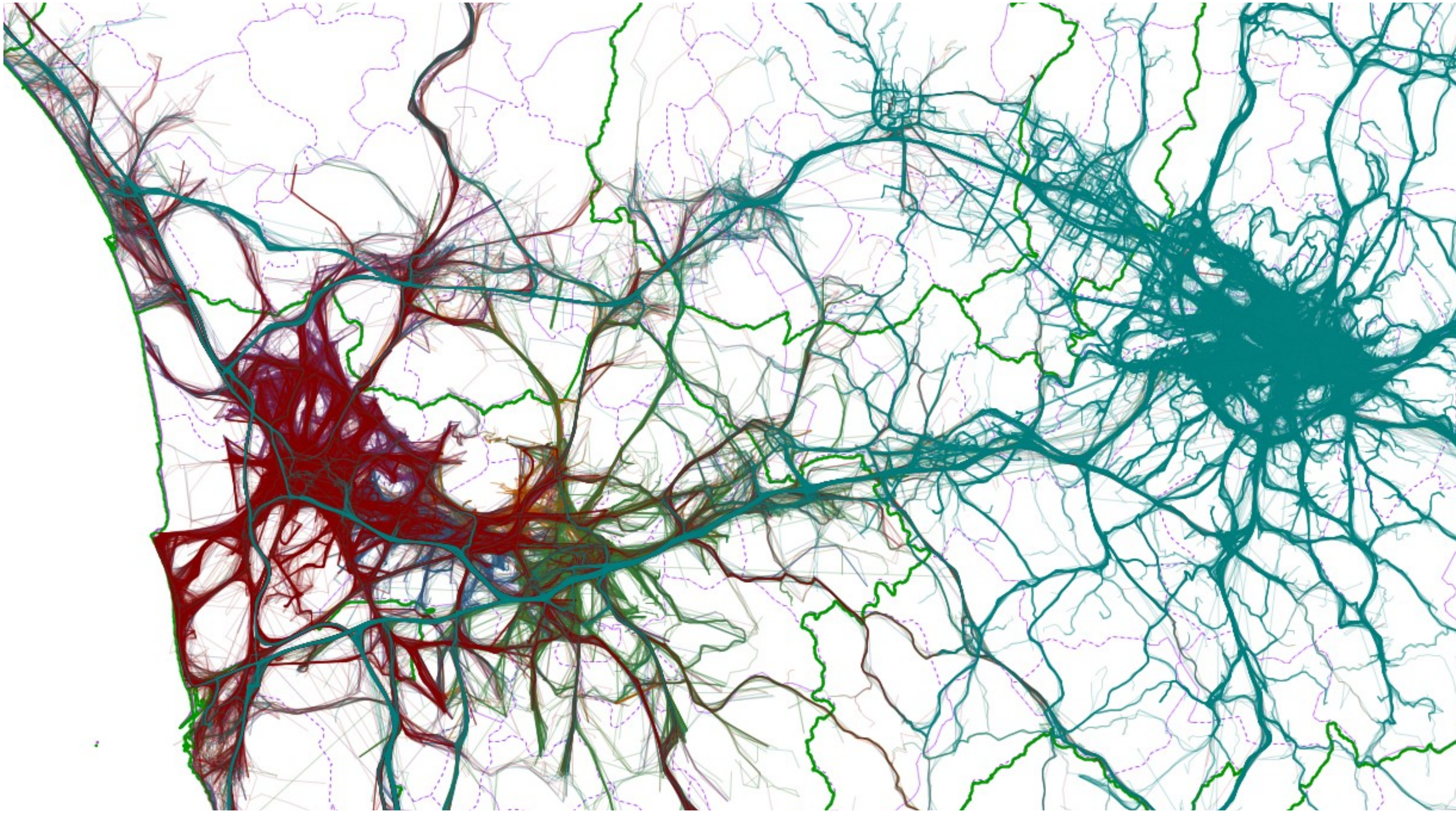




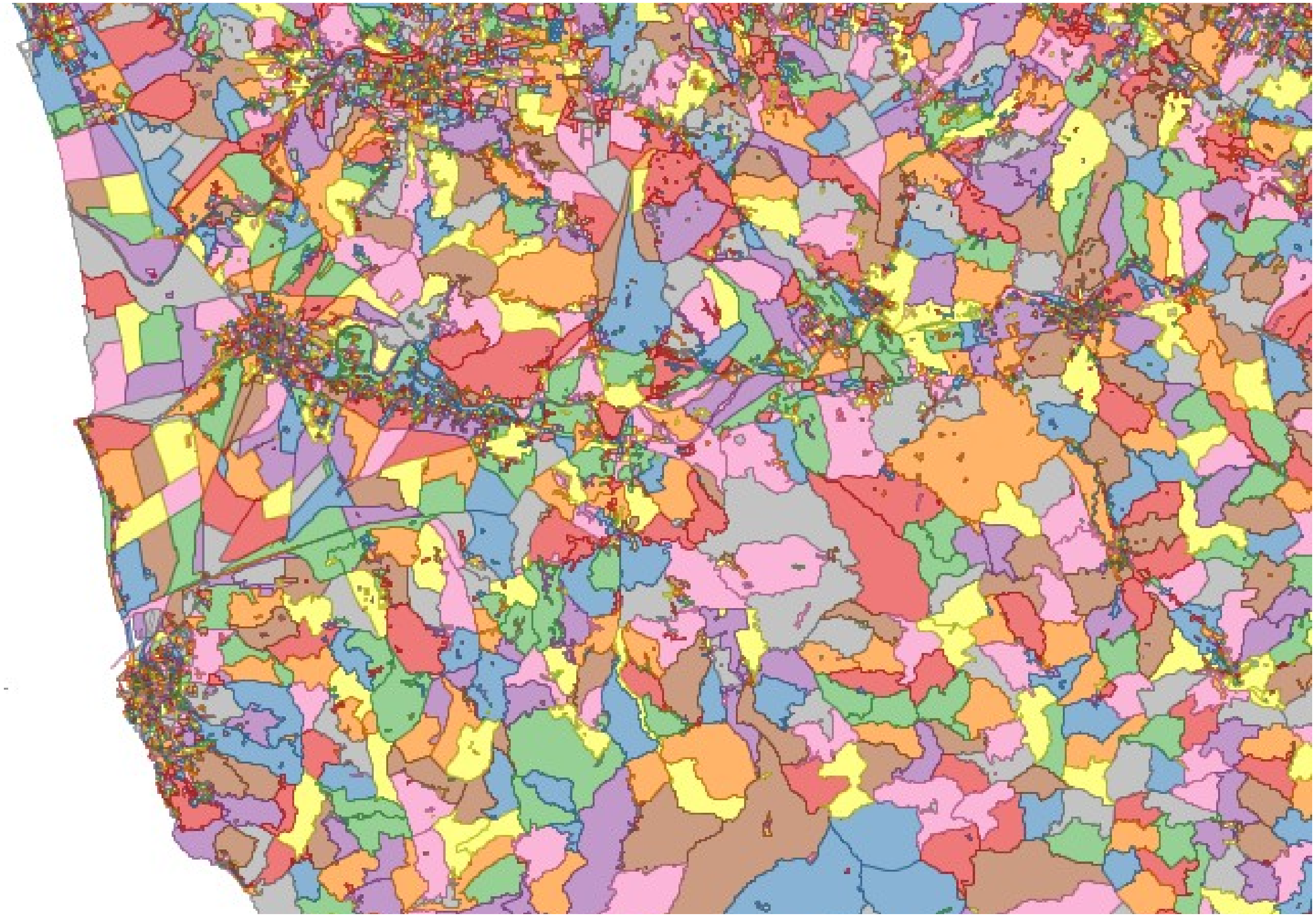
Services Towards Public Sector

Mobility-based Redefinition of Borders

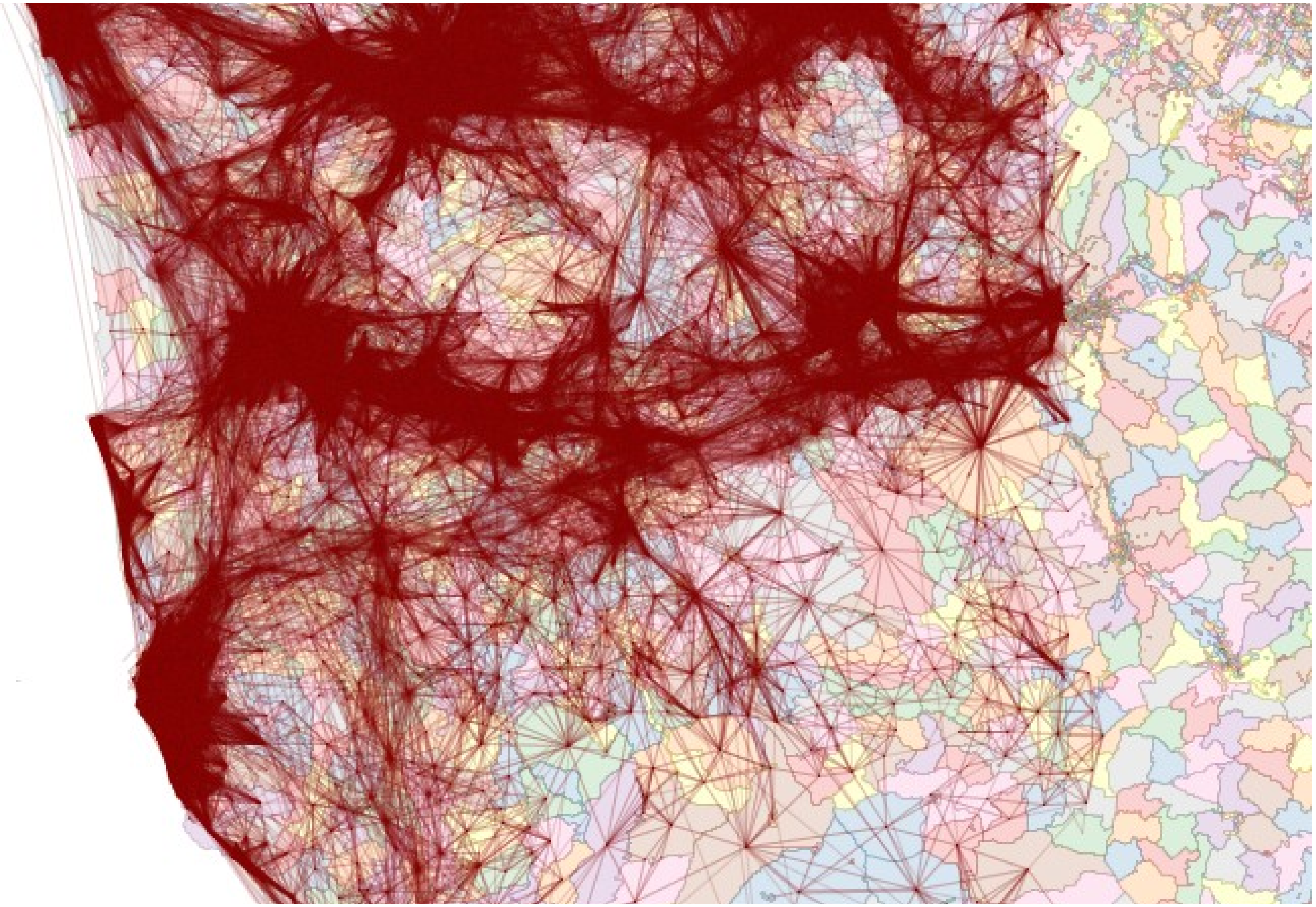
Mobility coverages



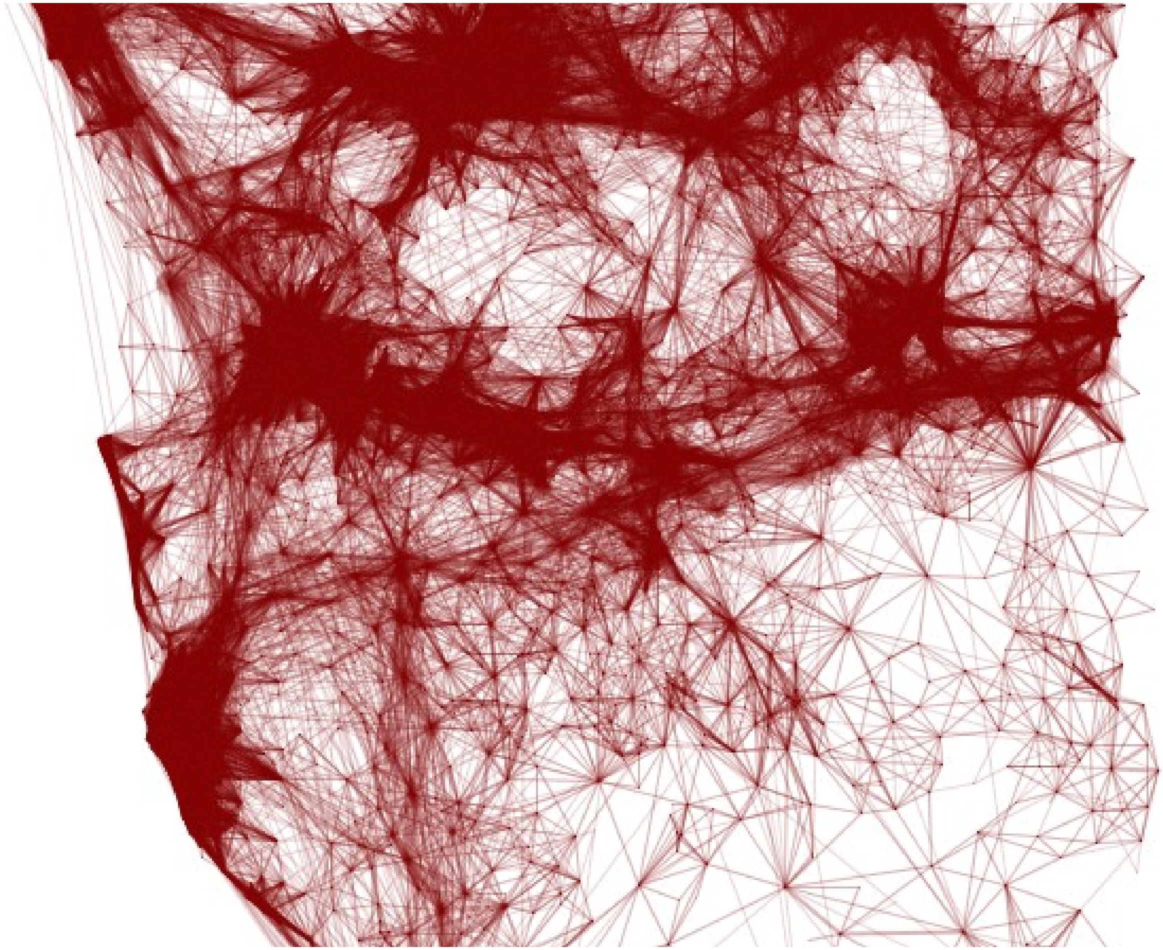
Step 1: spatial regions



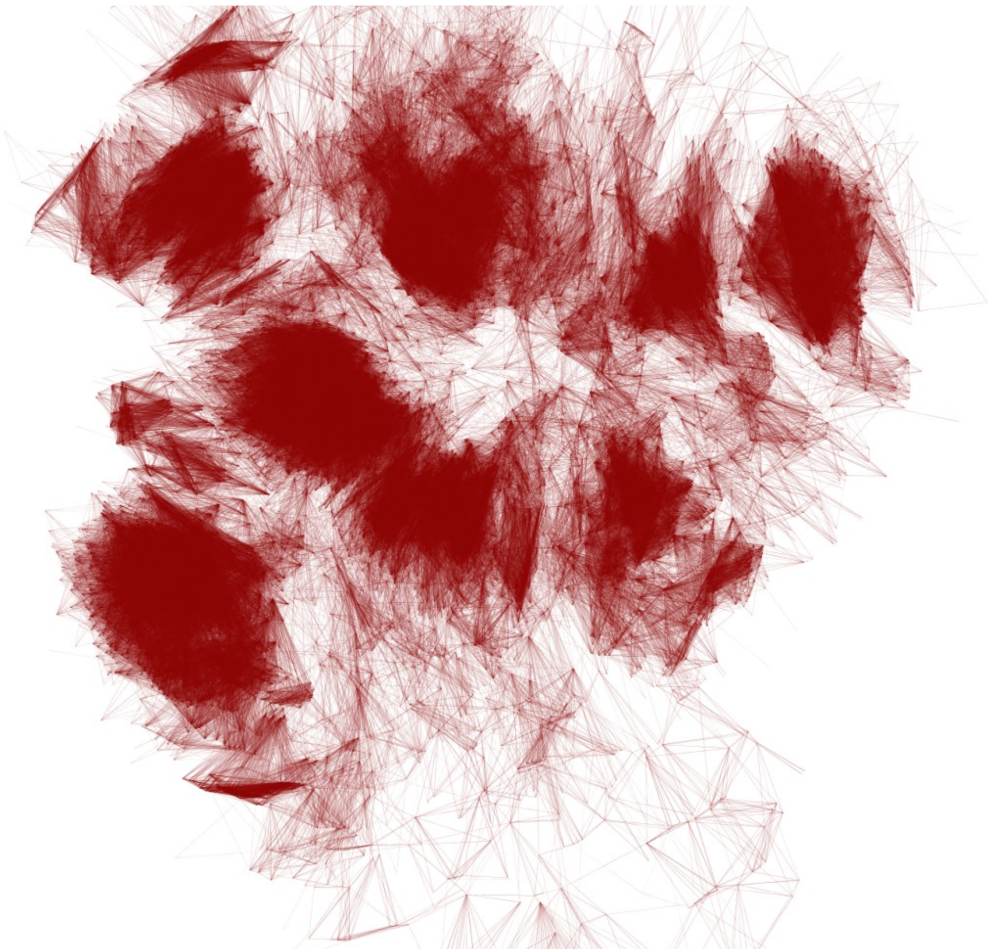
Step 2: evaluate flows among regions



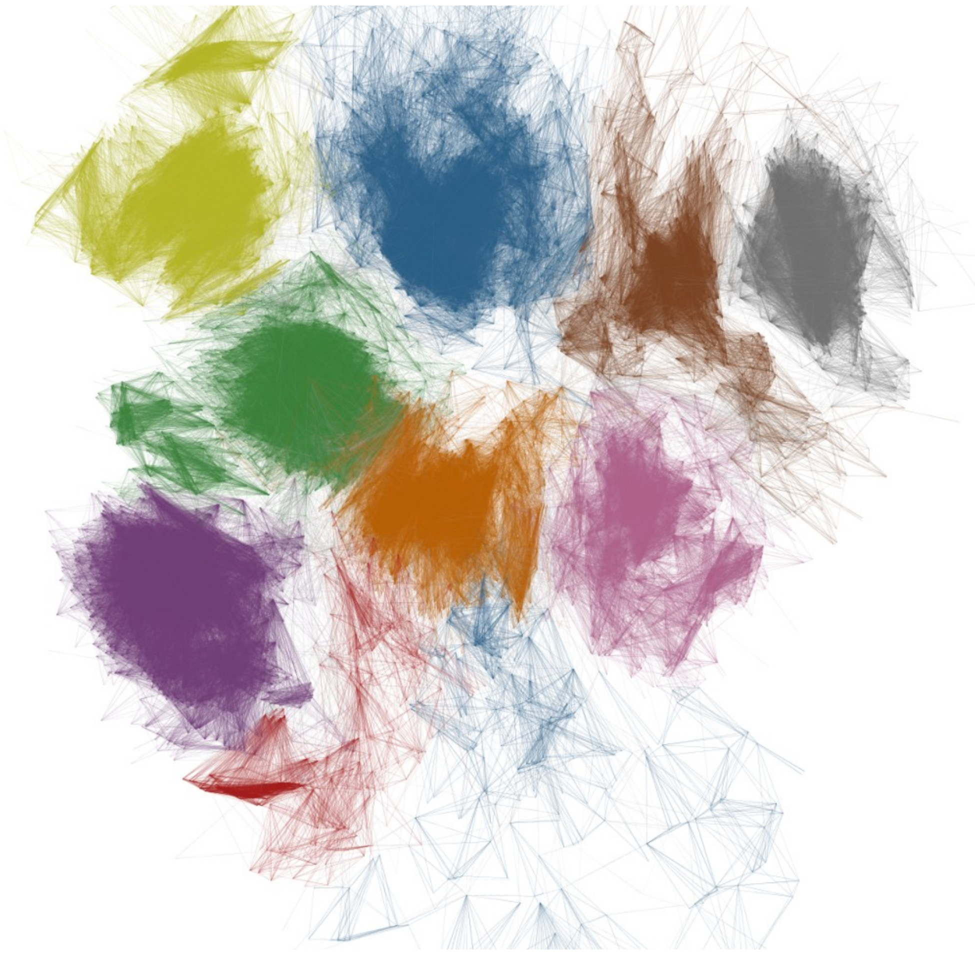
Step 3: forget geography



Step 4: perform community detection

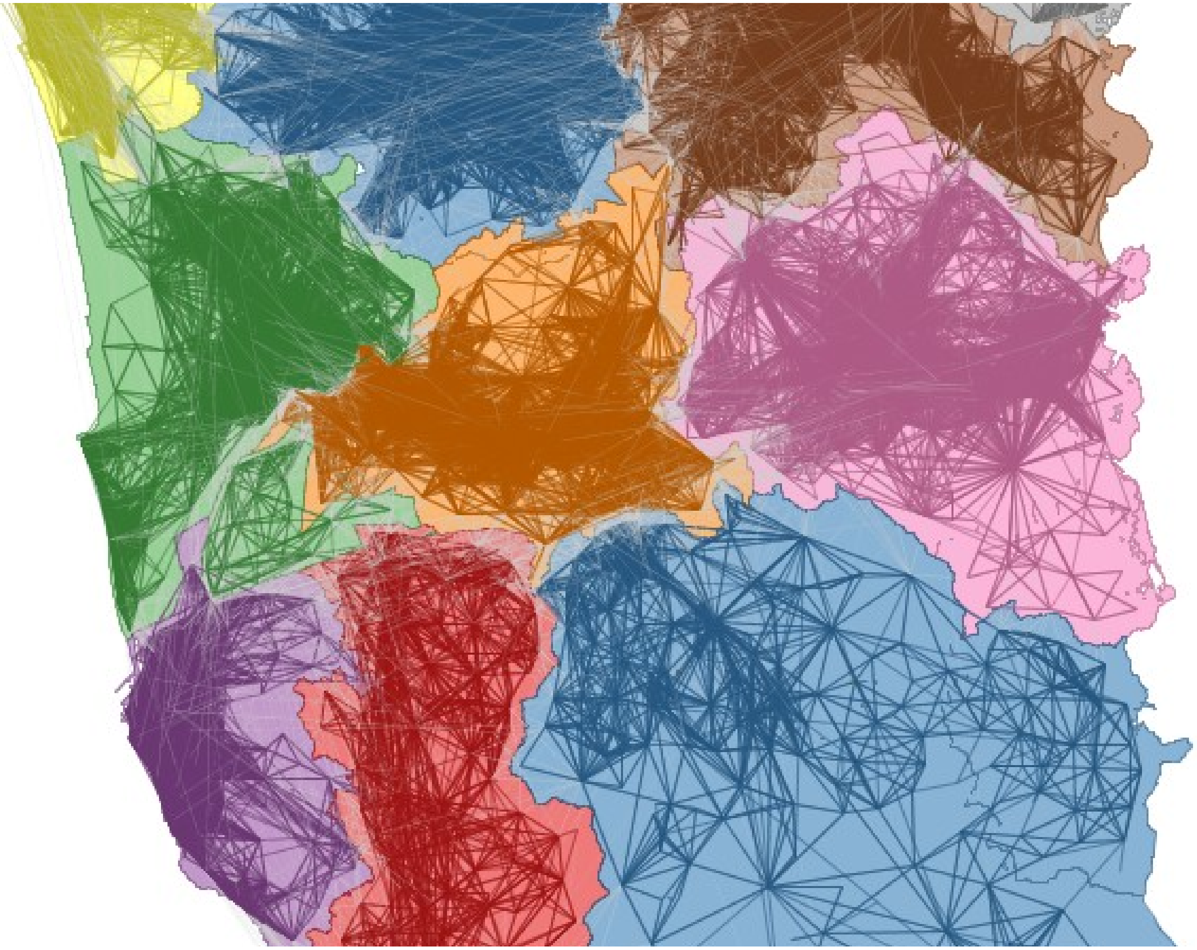


Step 4: perform community detection

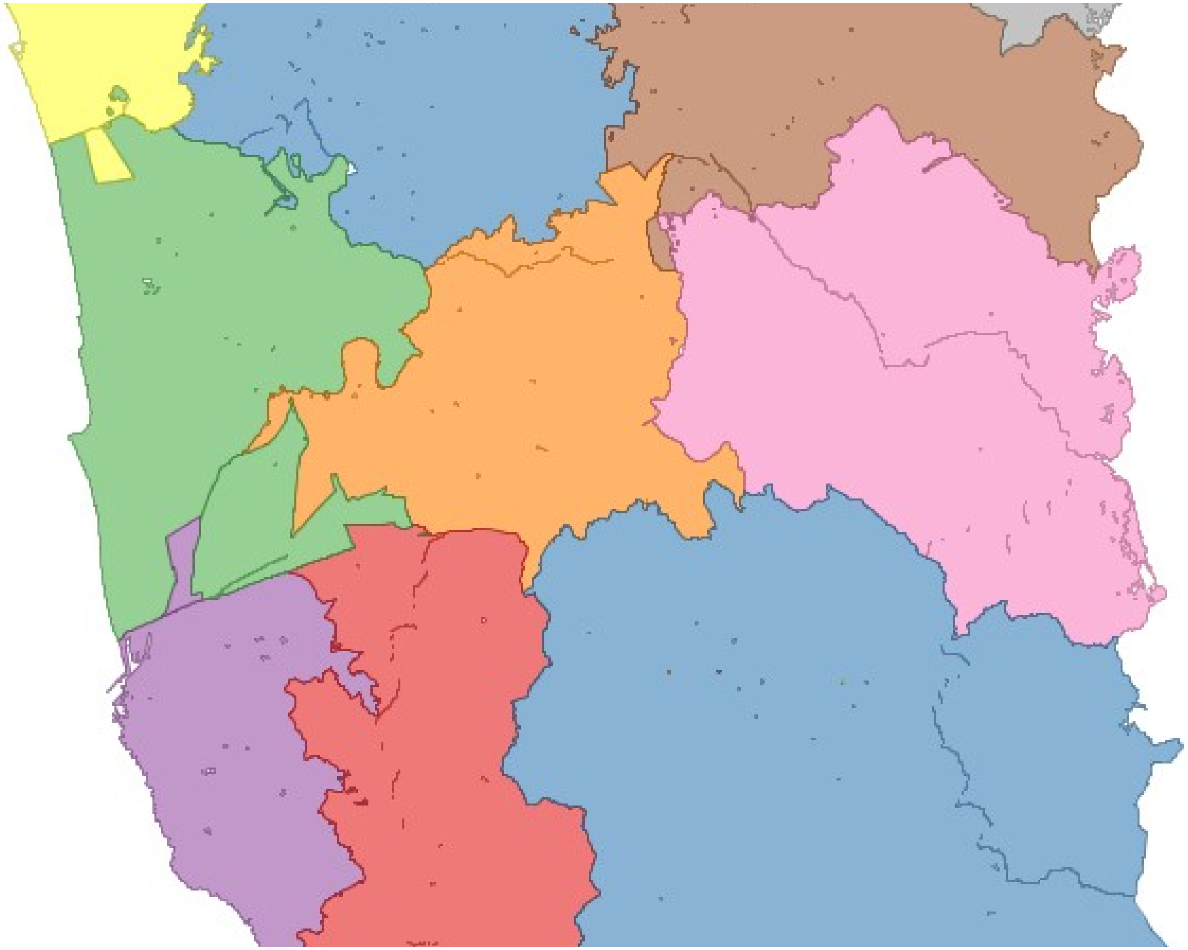


Step 5: map back to geography

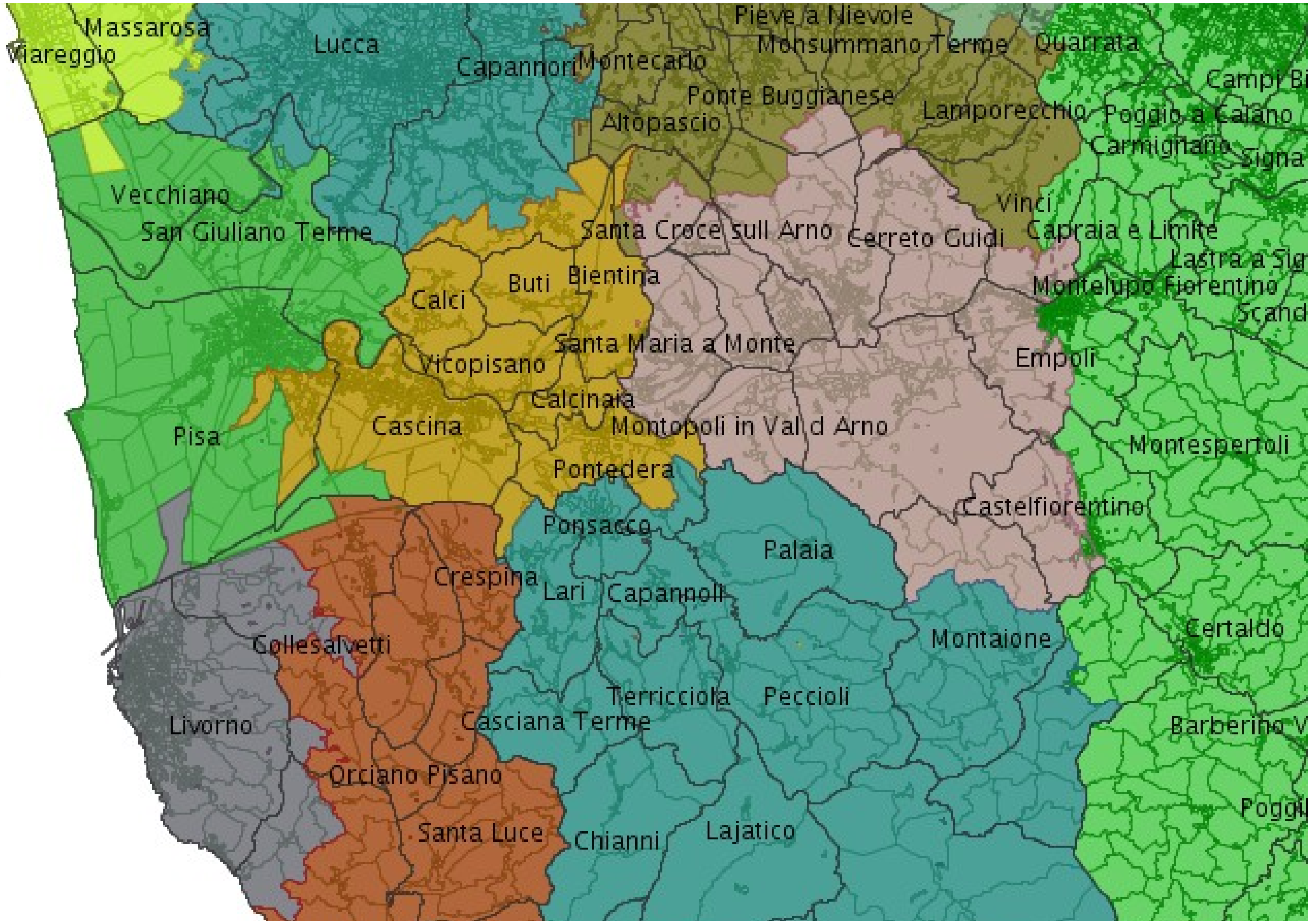
Step 6: draw borders



Final result

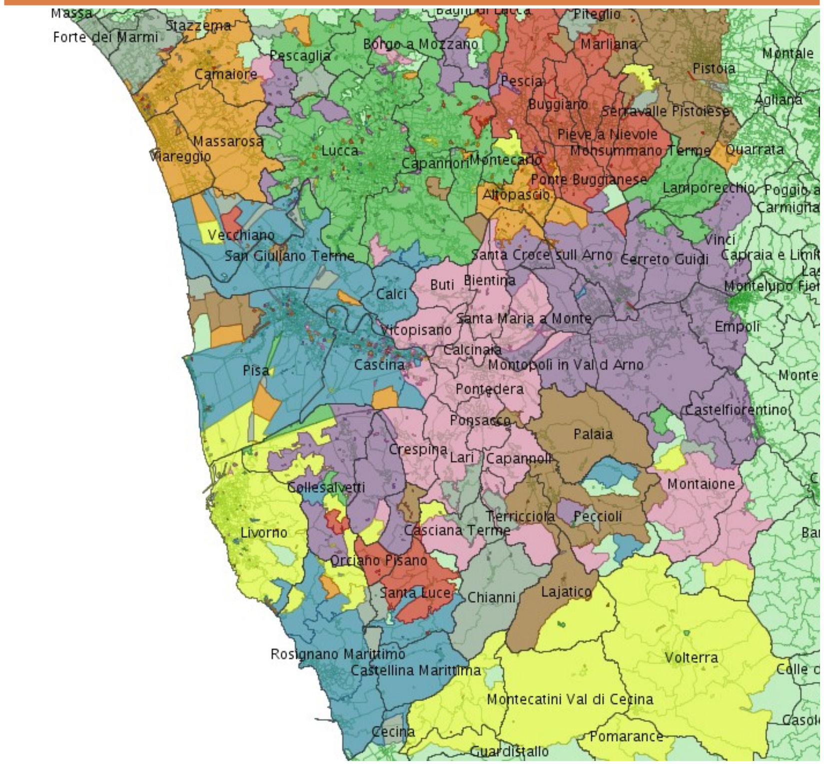


Final result: compare with municipality borders



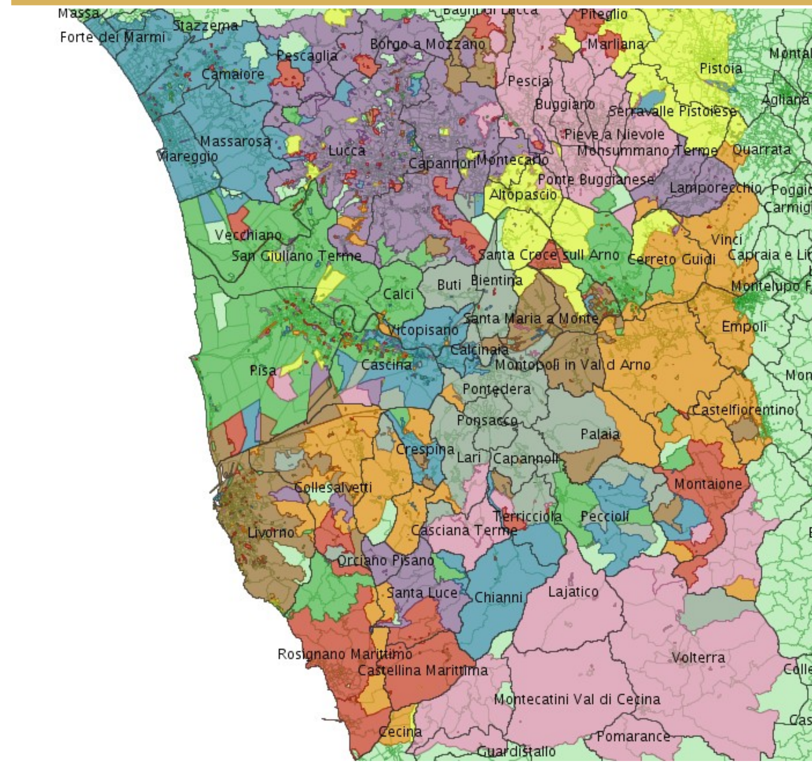
Borders in different time periods

Only weekdays movements



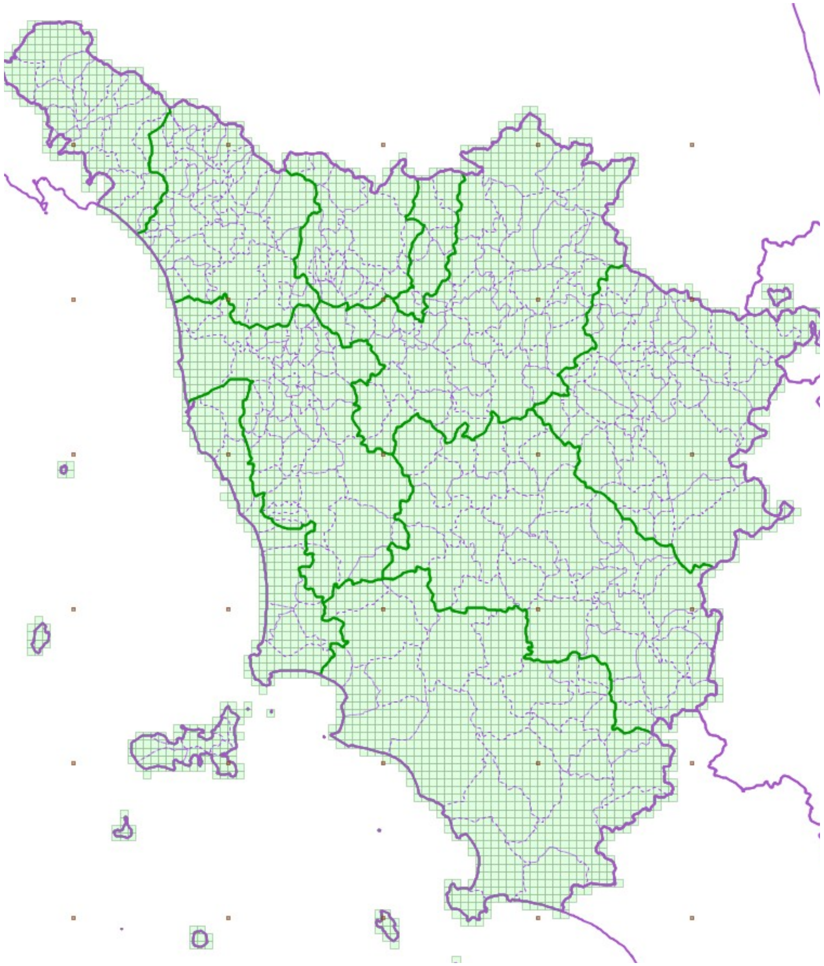
Similar to global clustering: strong influence of systematic movements

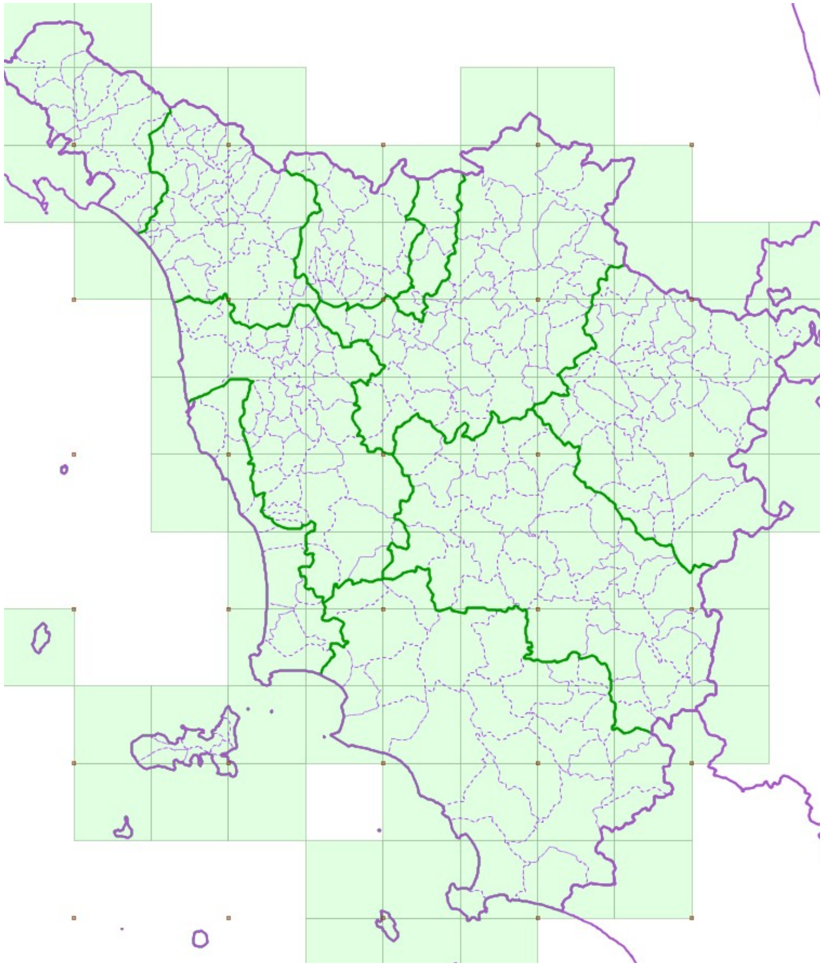
Only weekend movements



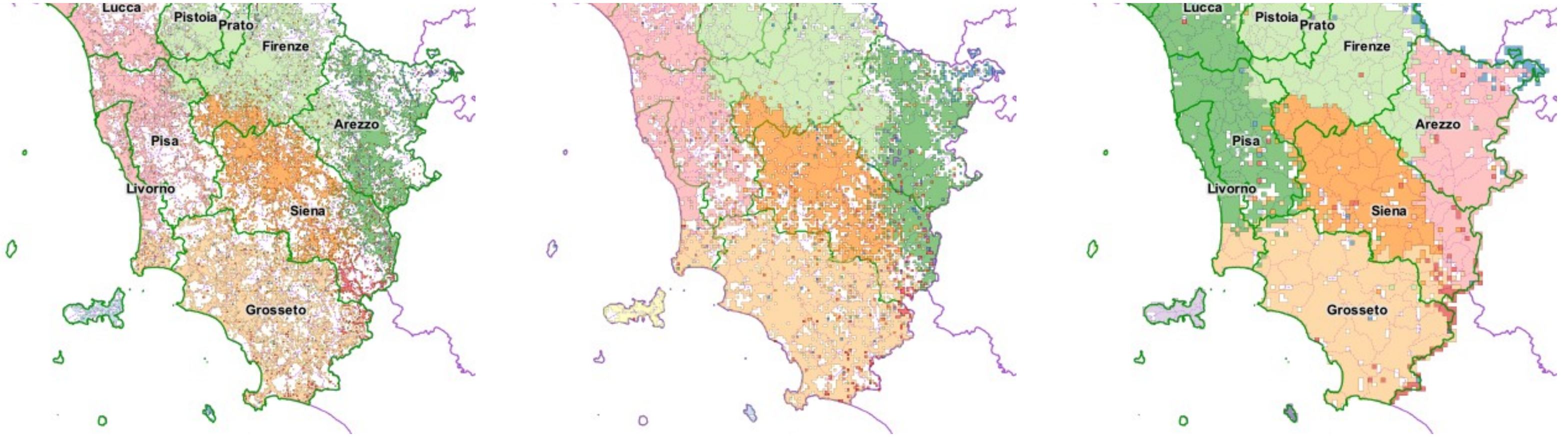
Strong fragmentation: the influence of systematic movements (home-work) is missing

Borders at regional scale

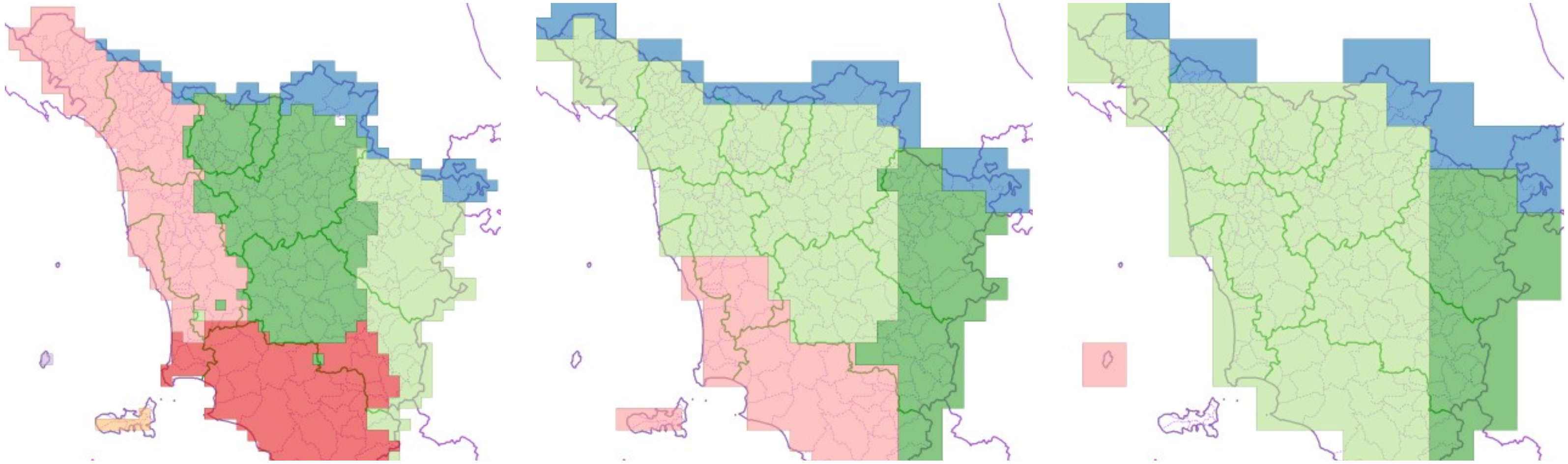




Final results



(a) 500m (b) 1000m (c) 2000m



Comparison with “new provinces”