Visual Analytics of Movement

Fraunhofer Institut Intelligente Analyse- und Informationssysteme IAIS

universität bonn

Dr. Gennady Andrienko
Dr. Natalia Andrienko

http://geoanalytics.net
http://visual-analytics.info

In collaboration with groups of
Prof. D.Pedreschi (Univ.Pisa), Prof. F.Gianotti (CNR)
Prof. D.Keim (Univ.Constance)

October 2009

Movement Data

Movement data is a temporal sequence of position records:

- {id,}x,y,t

Movement data are now collected in rapidly growing amounts owing to the development of tracking technologies:

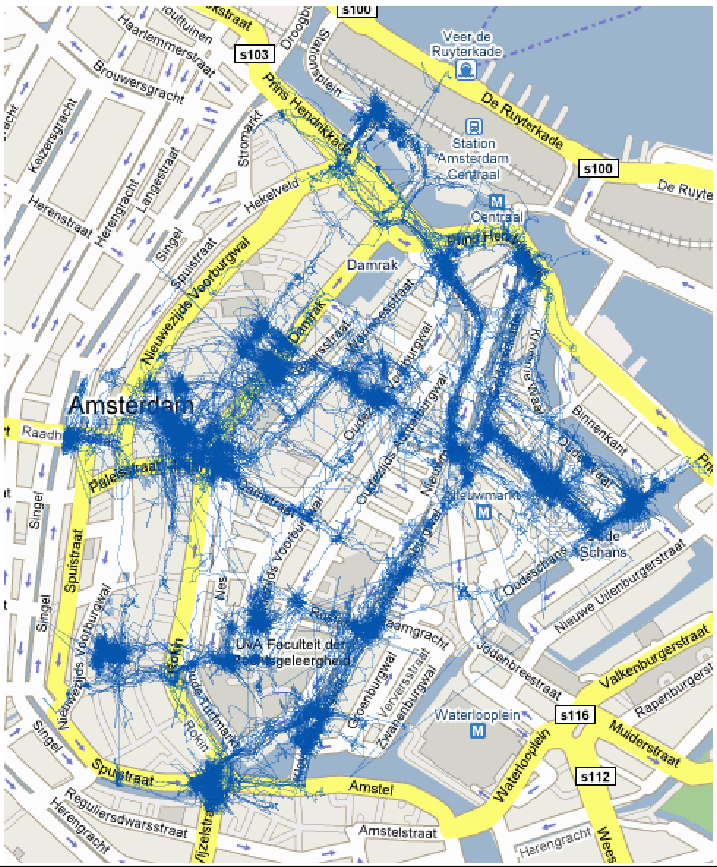
- GPS, RFID, WiFi, …

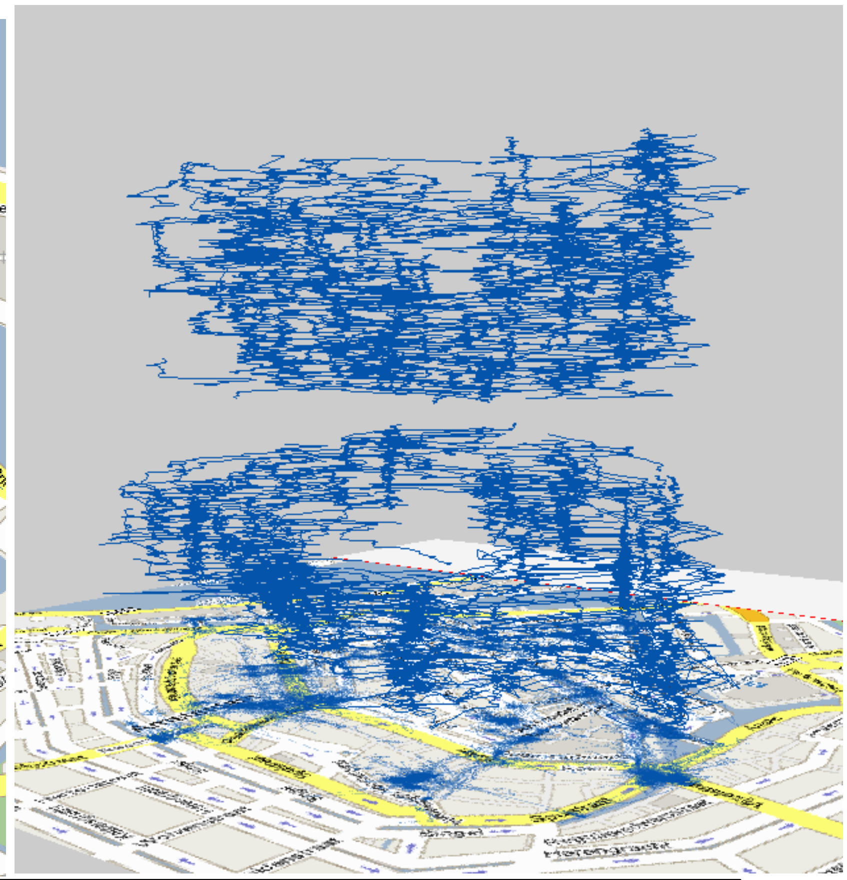
and also

- surveillance video, use of mobile phones, banking transactions, sensor networks, …

Examples of movement data: children in Amsterdam

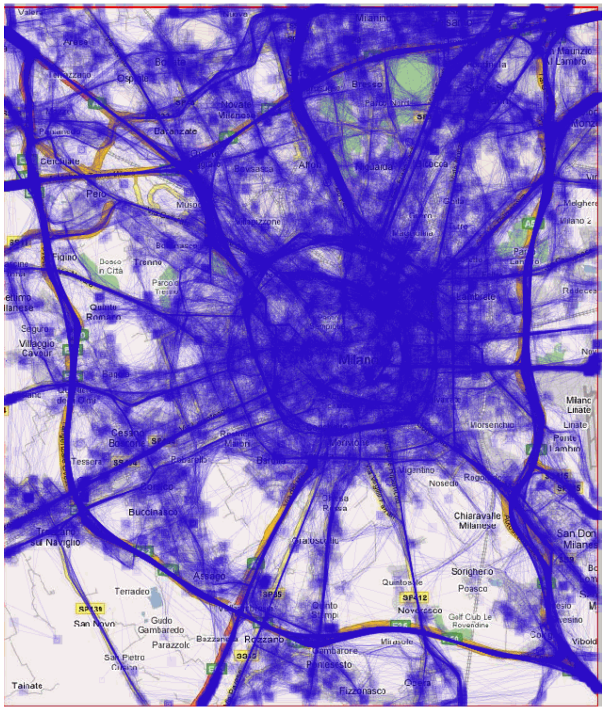
- GPS tracks of 303 schoolchildren playing an educational game in Amsterdam, about 57,000 points

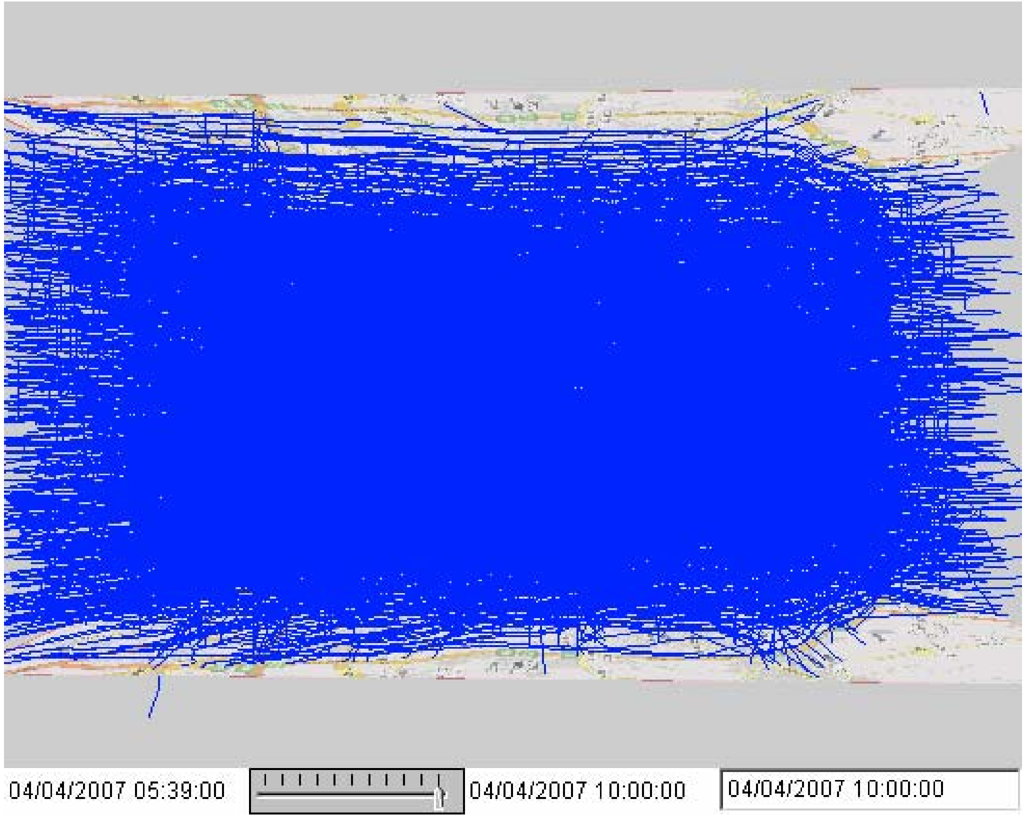




Examples of movement data: cars in Milan

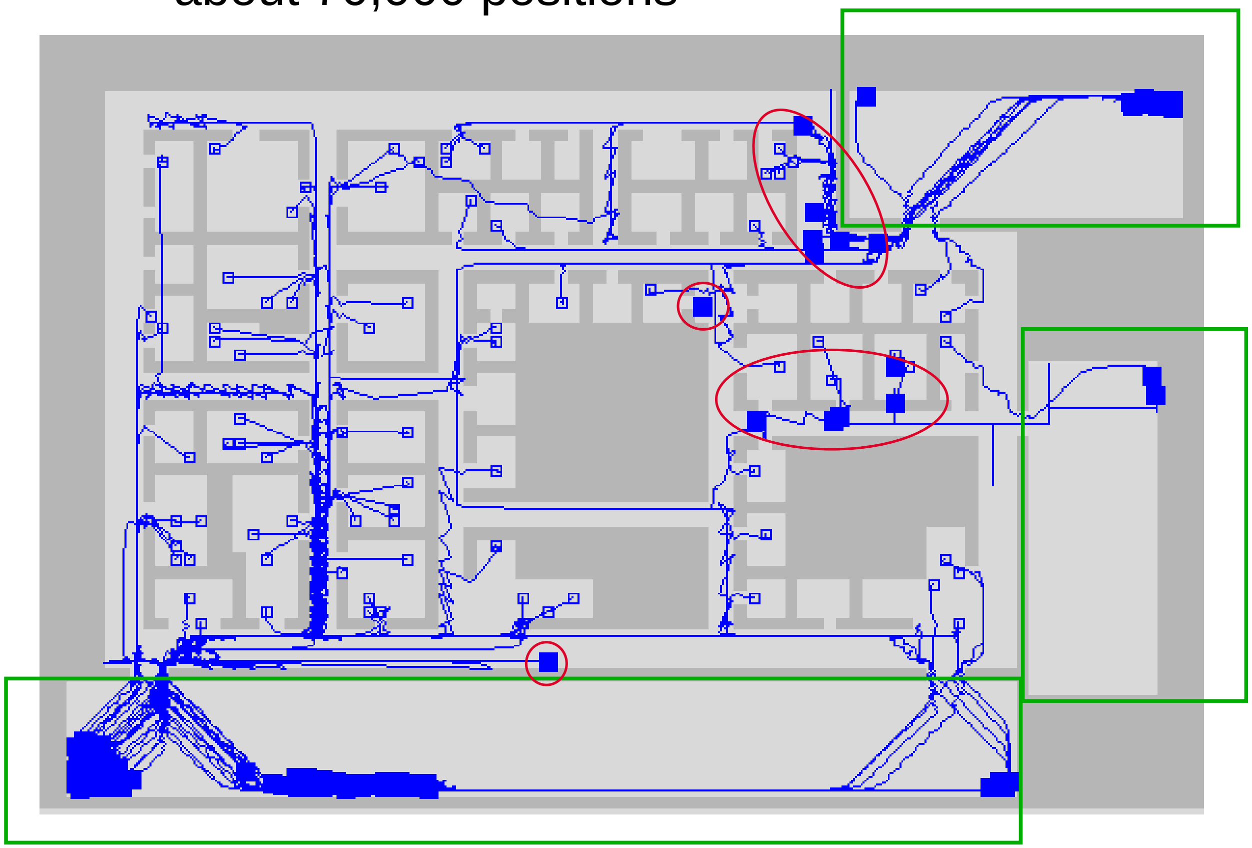
- 2,075,216 position records of 17,241 cars during 1 week

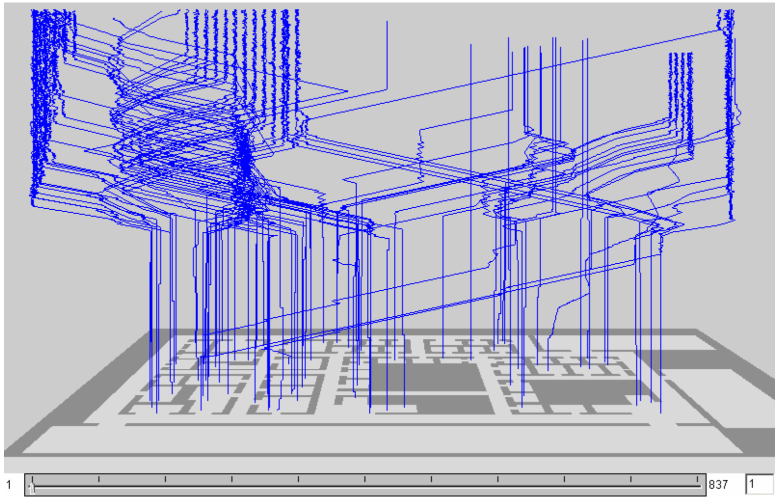




Examples of movement data: evacuation from a building

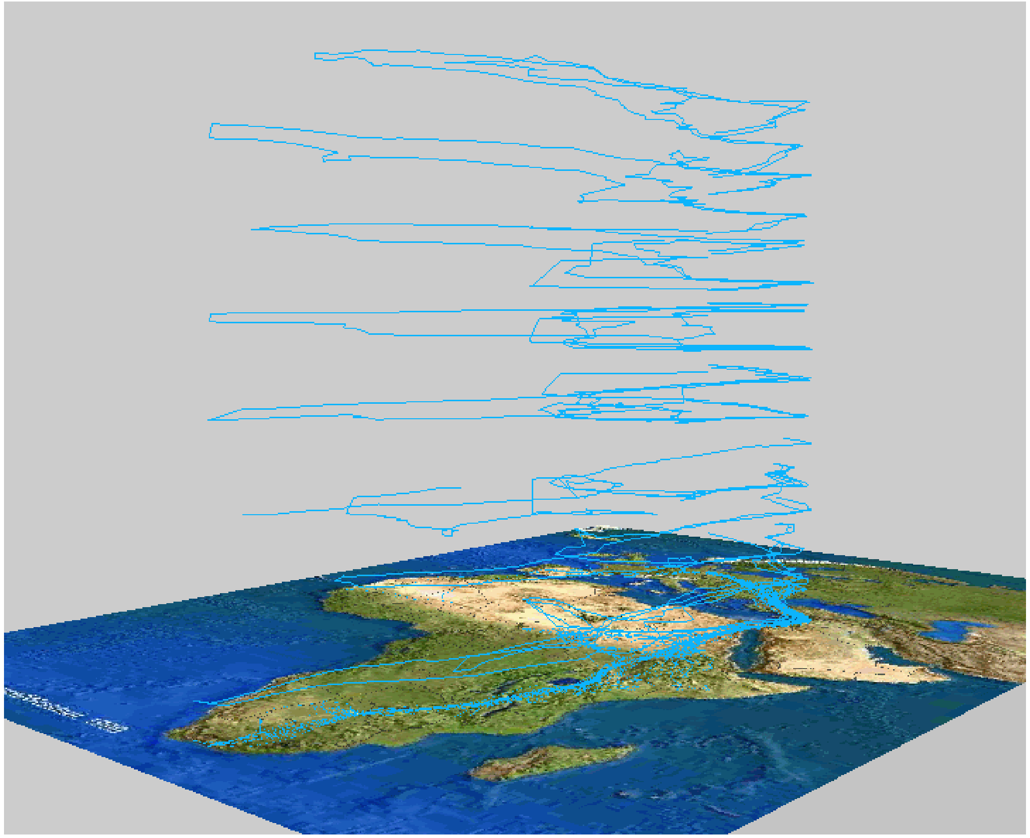
- Tracks of 82 people during evacuation, about 70,000 positions



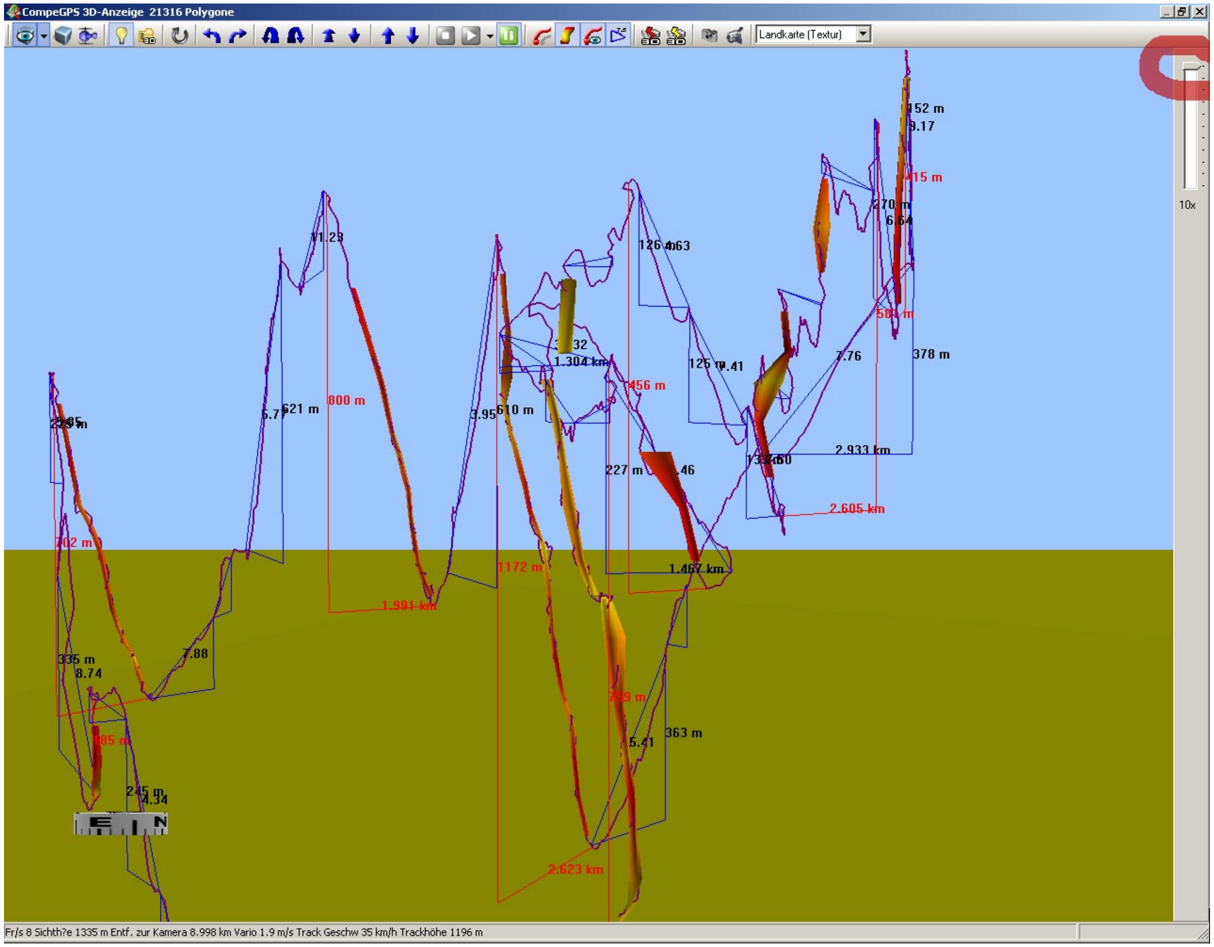


Examples of movement data: migration of white storks

Tracks of 35 storks during 8 years, about 2,000 positions

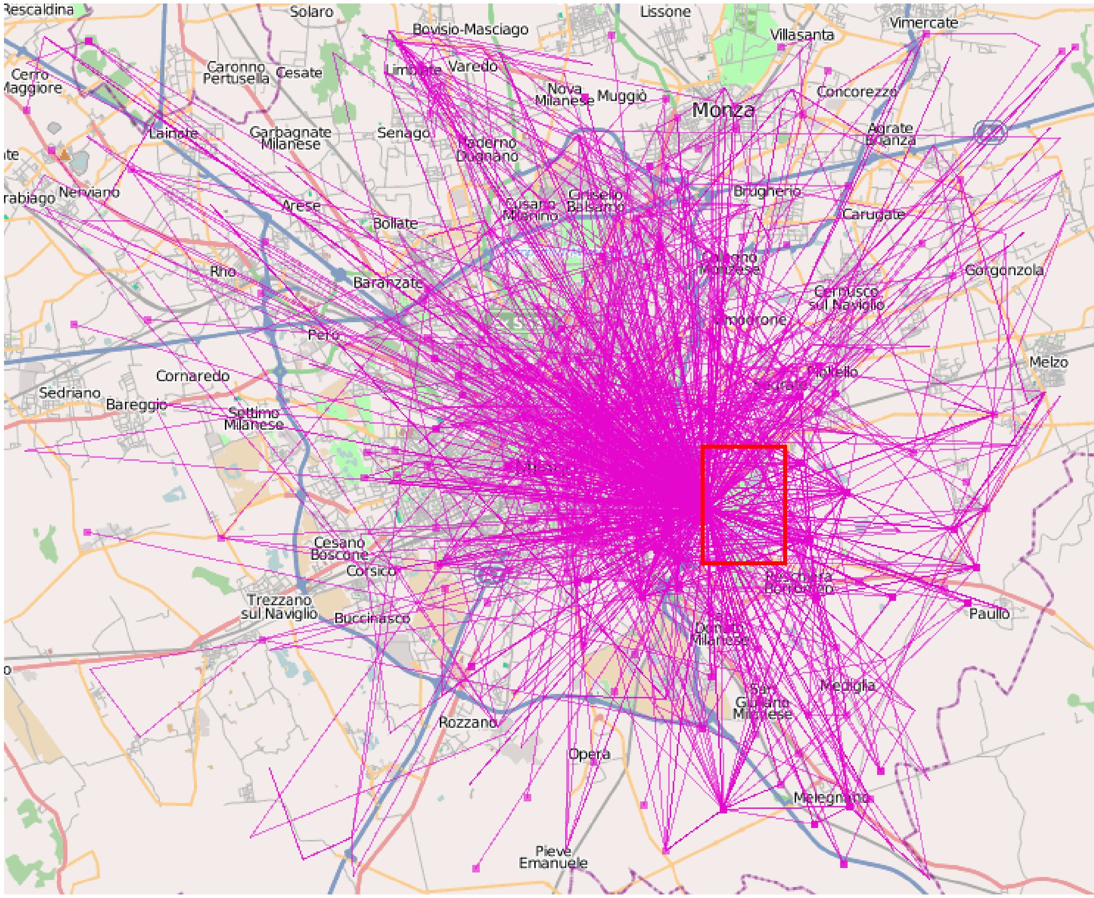


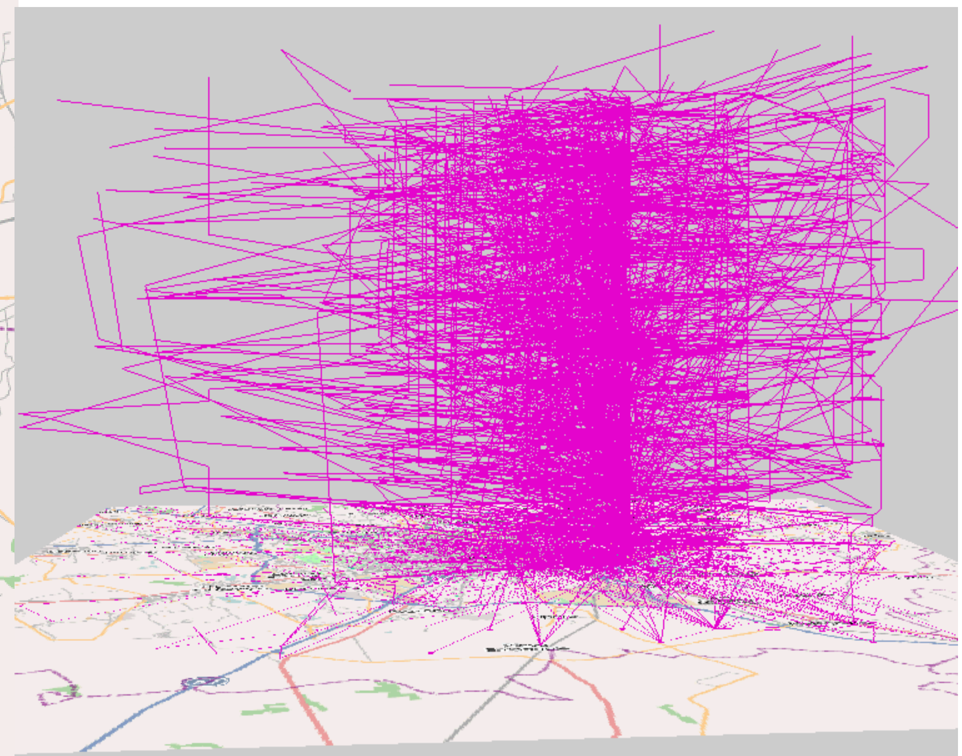
Examples of movement data: paragliding



Examples of movement data: calls of WIND customers in Milan

- 10 days; 550,000 customers; about 5,000,000 calls



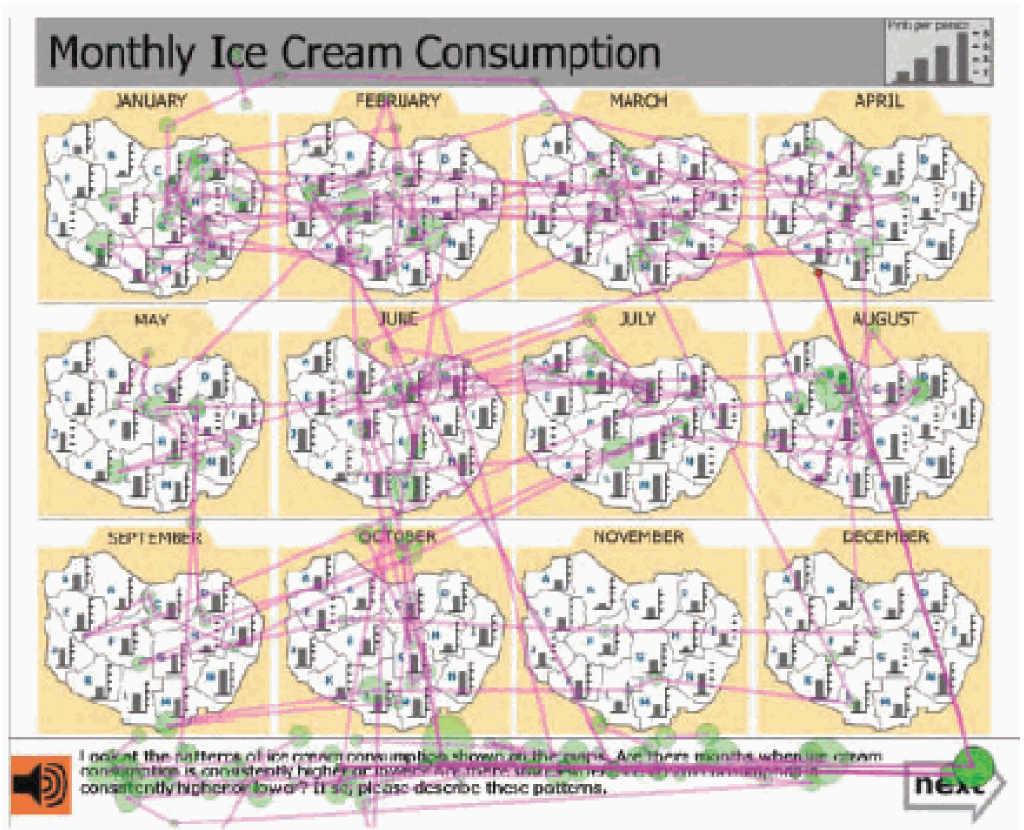


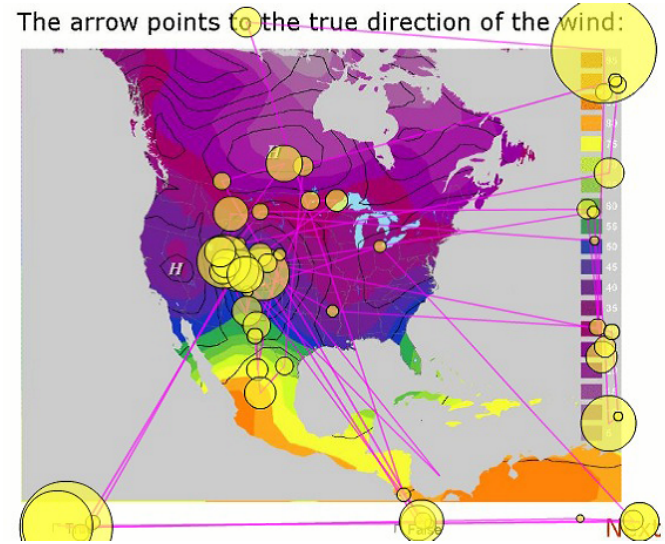
Examples of movement data: movement of flickr and panoramio photographers

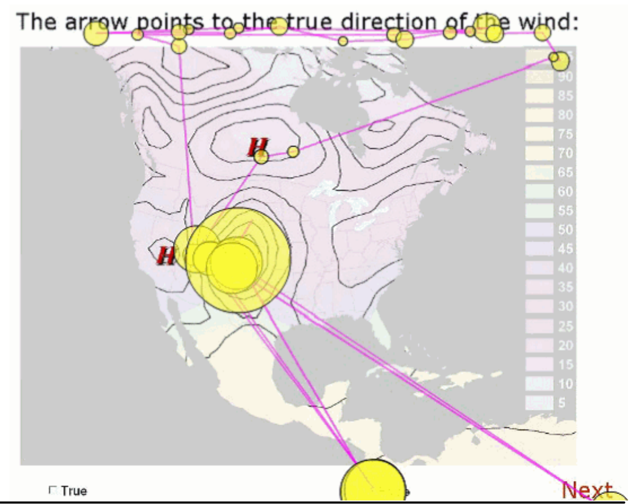
- Millions of georeferenced photos are stored in panoramio.com (displayed in Google Earth); photos of the same user tell about his/her movement!

Examples of movement data: gaze movement

- Eye tracking is used in psychology for studying human attention, perception, and cognition







Movement Data

Movement data: simple structure

- {id,}x,y,t

Complexities:

1. Amount (number of moving entities, number of records)
2. Geographic space with its structure and complexity
3. Time, linear and also multiple nested and overlapping cycles
4. Data properties:
   - imprecision (errors in location, time, attributes)
   - irregular and/or sparse sampling
   - missing values
5. Open-world, ill-defined problems

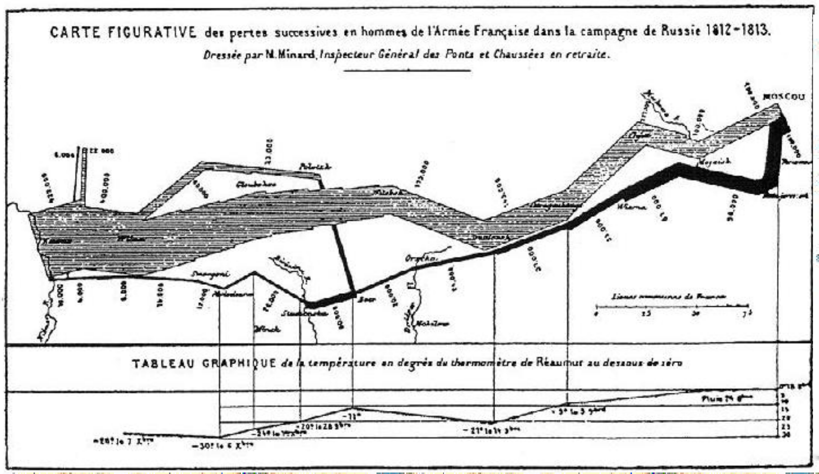
Motivating applications

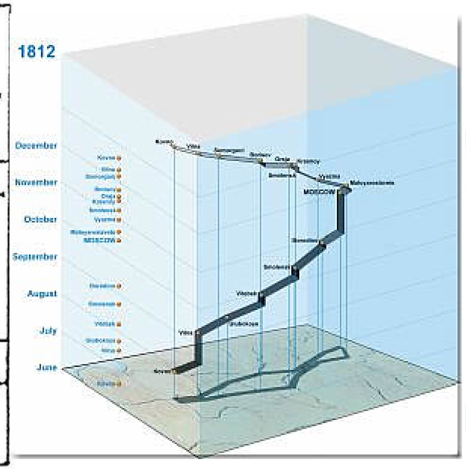
- Health: analyzing and predicting spread of diseases in hospitals (tracking patients) or by migrating birds
- Biology: studying behaviors of animals
- Environment protection and nature preservation: detection of illegal activities
- Social science and history: analyzing individual history, revealing social structures and patterns of interaction
- Business: transportation management, targeting outdoor advertisements, optimizing layout of trade spaces, detecting bottlenecks in logistic systems
- Mobile gaming and education: analyzing involvement of participants and usage of space
- Sport: post-game and online support for team managers, journalists, and general public
- Security and safety: improving layout of public buildings, supporting evacuation from crisis-affected areas, identifying suspicious behaviors or fraud banking transactions

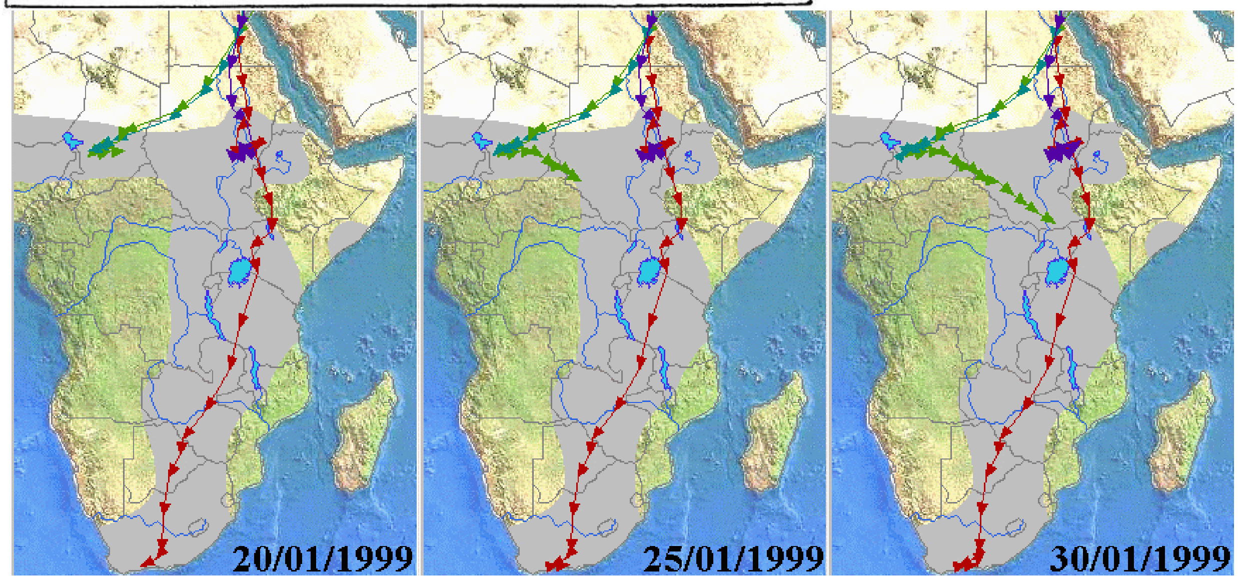
Types of movement data and related analysis tasks

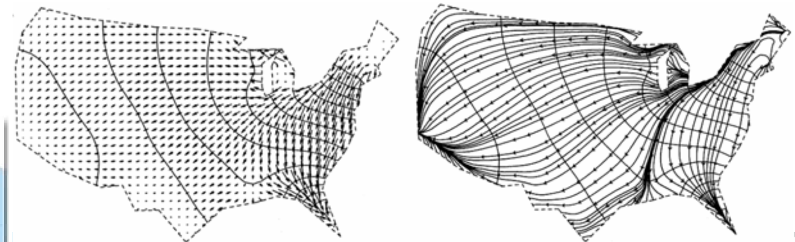
Data	Tasks
movements of a single object	Analysis of the object's behaviour: significant places, times and durations of the visits to different places, typical trips, times and durations of the trips, deviations and their reasons
movements of multiple unrelated objects	1) Studies of space use, accessibility, permeability, connectivity, major flows, typical routes between places.
	2) Studies of emerging patterns of collective movement: concentration/ dispersion, convergence/divergence, propagation of movement characteristics etc.
movements of multiple related objects	Studies of relative movements (approaching, encountering, following, evading, etc.) and interactions between the objects.

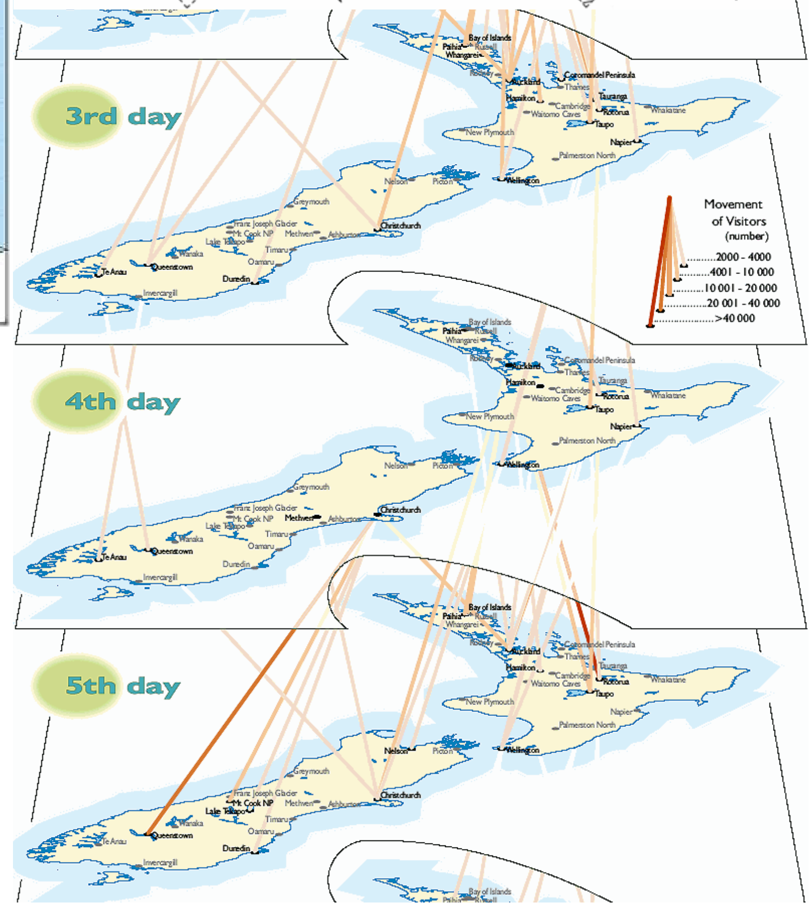
State of the art - visualization









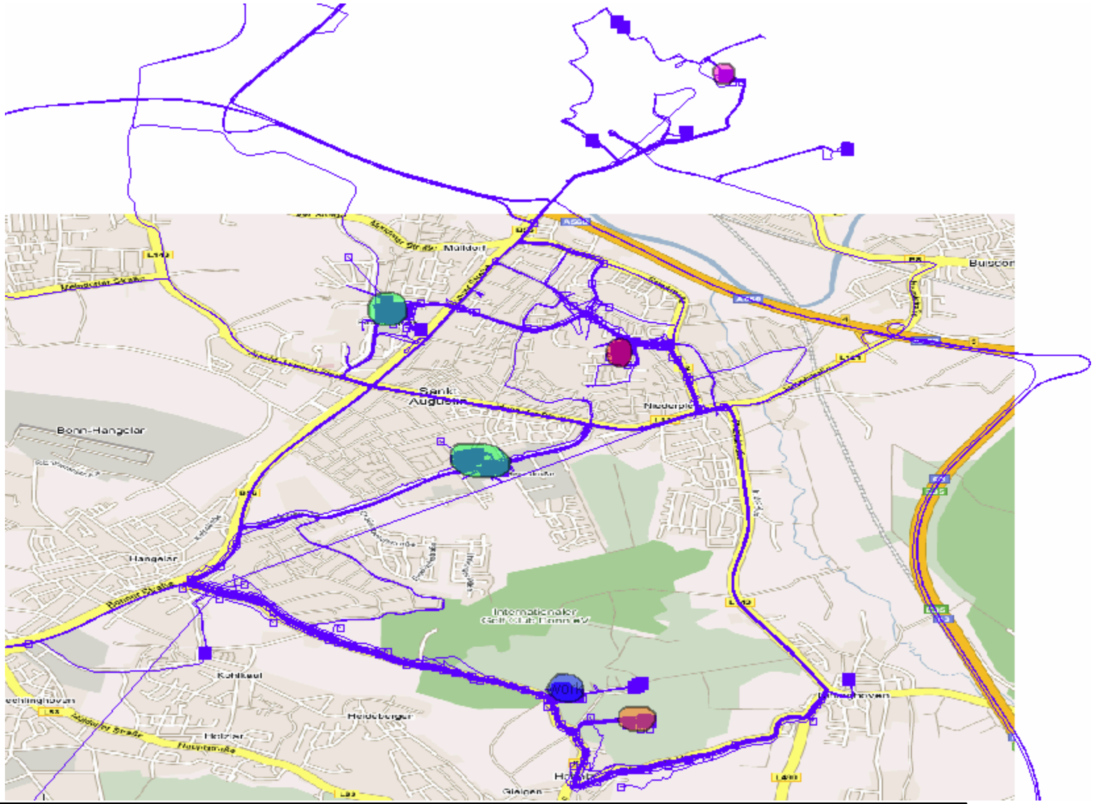


State of the art – data mining

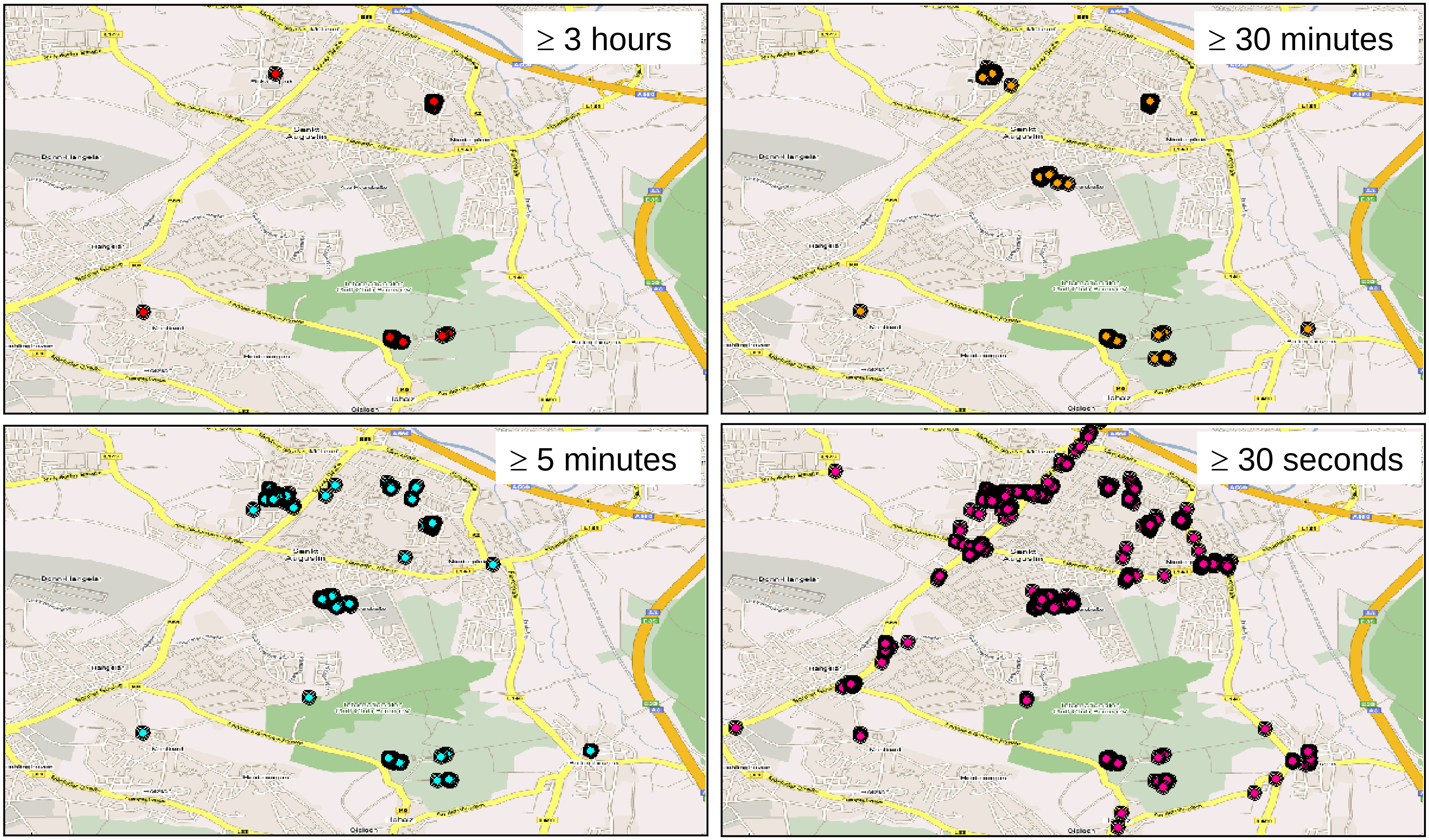
- Definition of various distance functions for trajectories, to be used for similarity search and clustering
- Ad hoc methods for specific kinds of patterns:
  - T-patterns (same sequences of stops with similar transition times),
  - Relative motion patterns (flock, leadership etc.)
- Statistical models for location prediction
- Classification models to identify the mode of movement (walking, public transport, bike, or car)

Example of visual analytics: analysis of a personal movement behaviour

- Data: positions of a personal car tracked over a long time period (about a year)
- Task: investigate the movement behaviour of the car owner:
  - Identify significant places and relate them to person’s activities
  - Detect and interpret typical trips: sources, destinations, routes, intermediate stops, purposes, …
  - Detect different routes between the same places; explain when and why each route is chosen
  - …

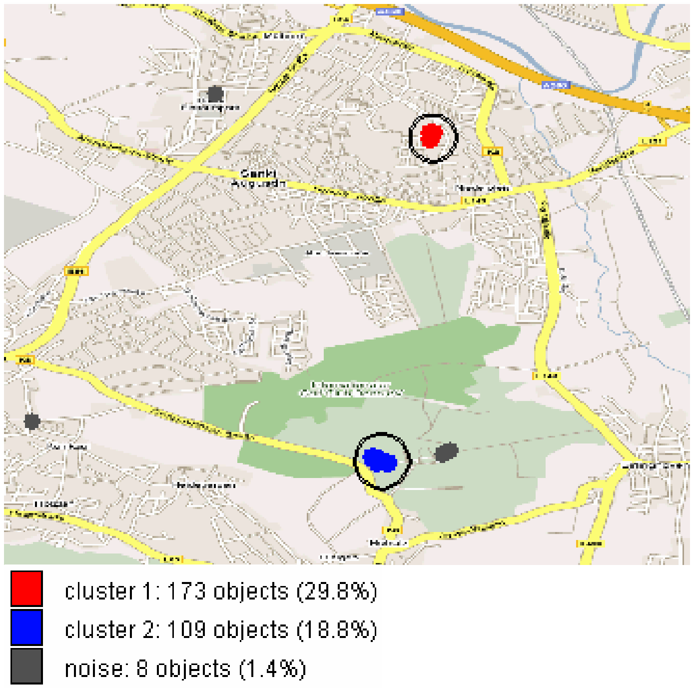


Finding significant places: looking at the positions of the stops

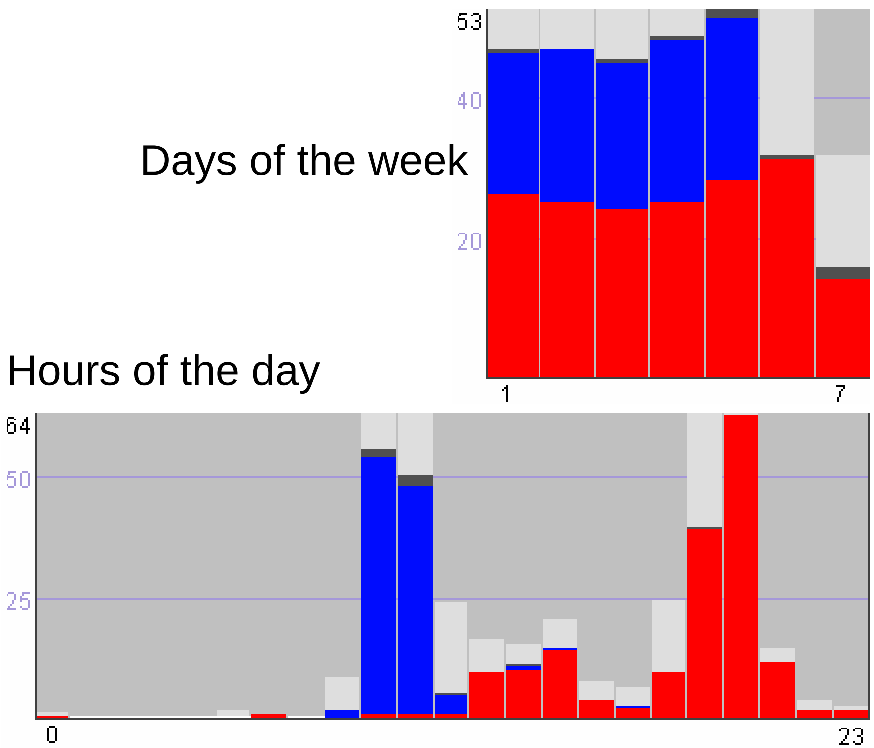


Interpretation of the places of stops
A) Long stops (≥ 3 hours)

1) Spatial clustering: find repeated stops



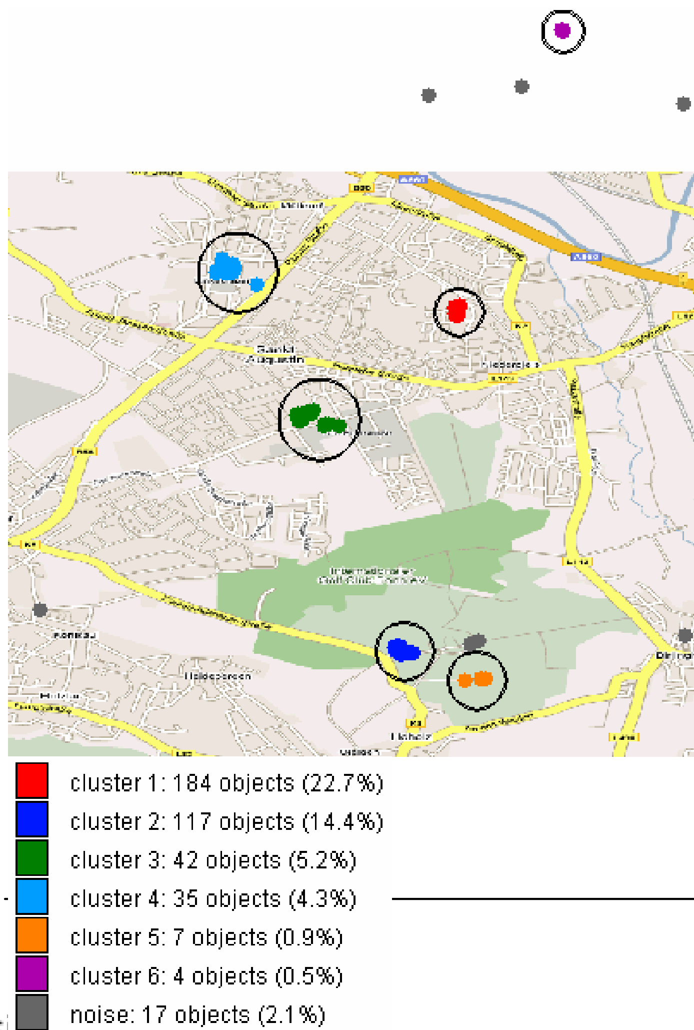
2) Look at the days and times of the occurrence



⇒ **Red: home, blue: work**

Interpretation of the places of stops
B) Medium stops (≥ 30 minutes)

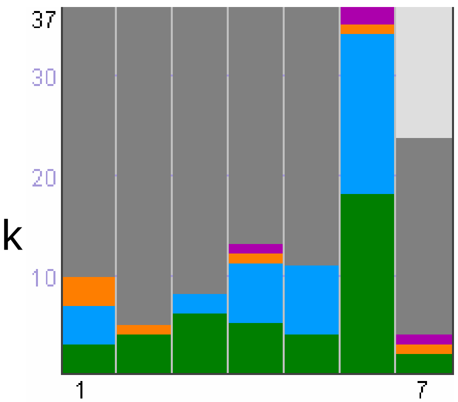
1) Spatial clustering: find repeated stops



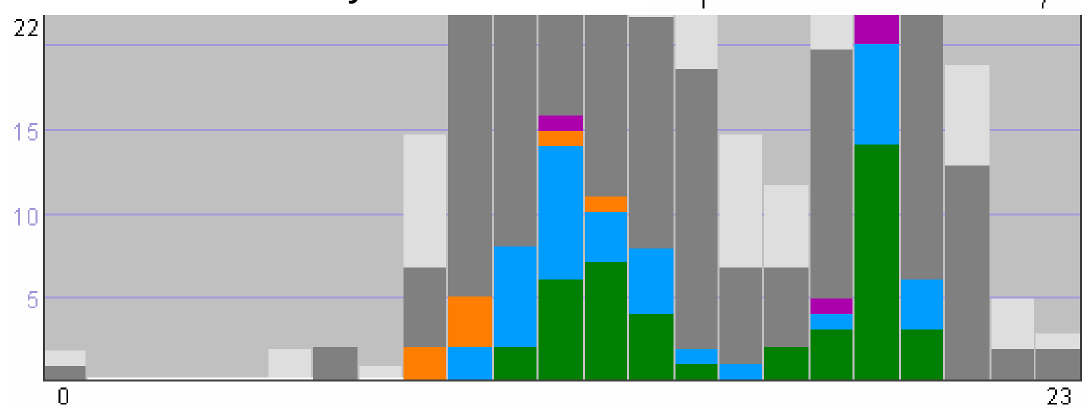
2) Look at the days and times of the occurrence

(clusters 1&2 excluded)

Days of the week



Hours of the day

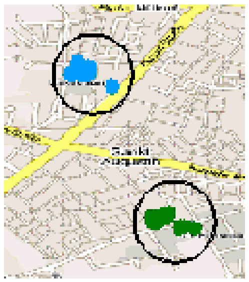


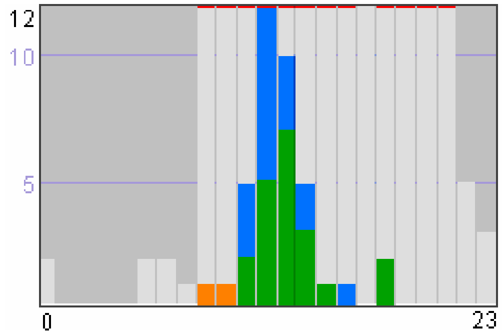
⇒ **Green & light blue: probably shopping**

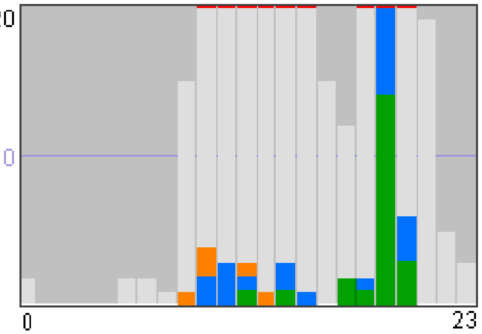
Interpretation of the clusters 3 and 4 (continued)

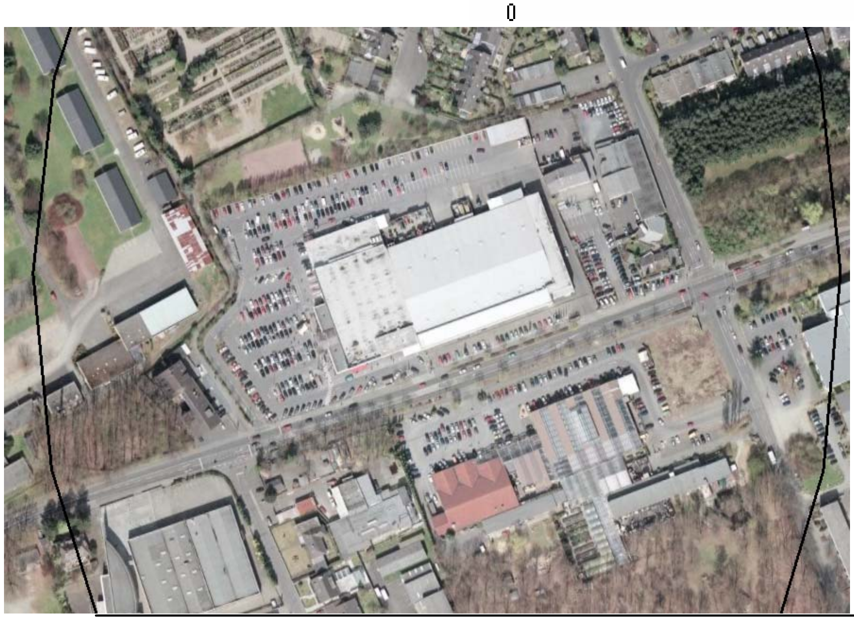
Saturday and Sunday: the stops mostly occur between 10 and 14

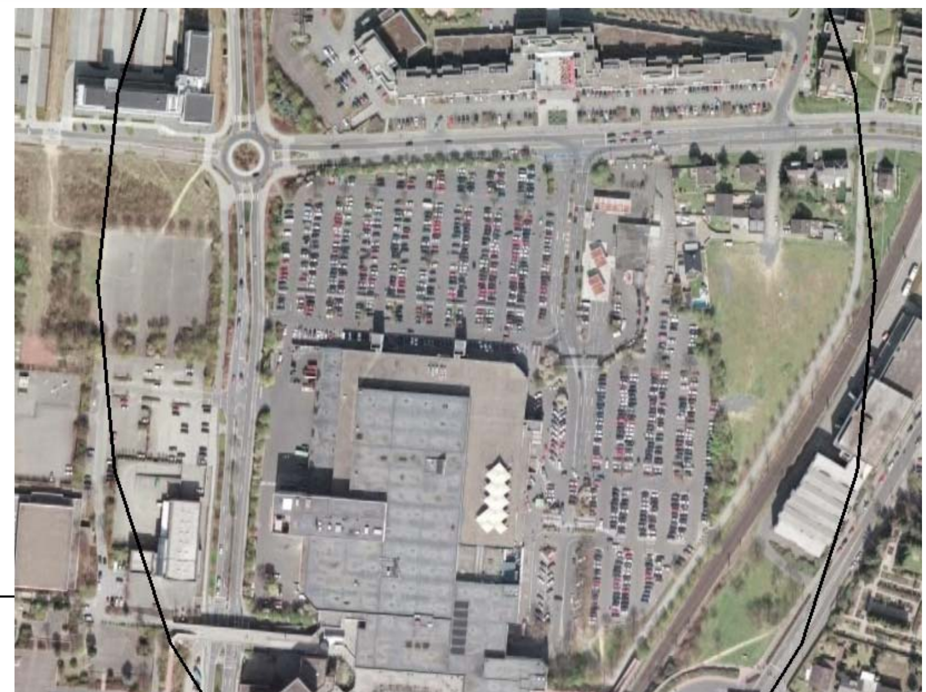
Monday to Friday: the stops mostly occur in the evening hours (max between 18 and 19)







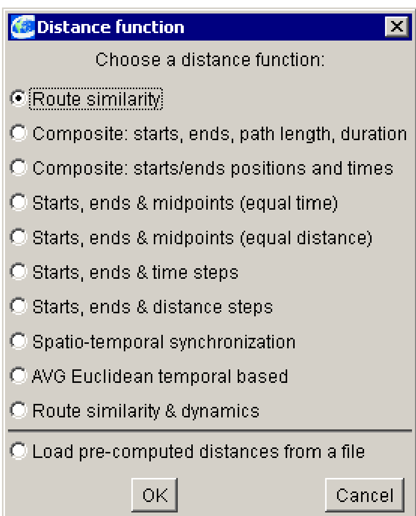


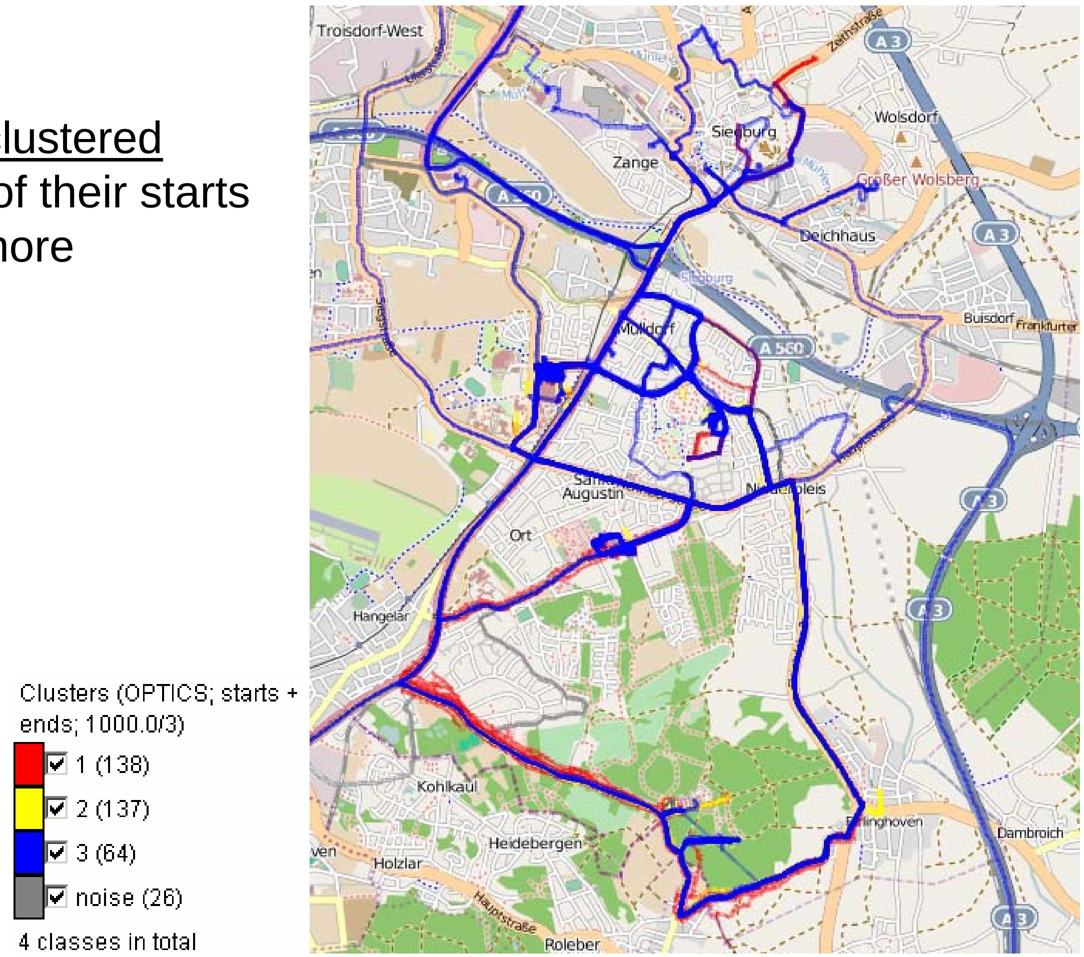


Looking for typical trips: From where and to where?

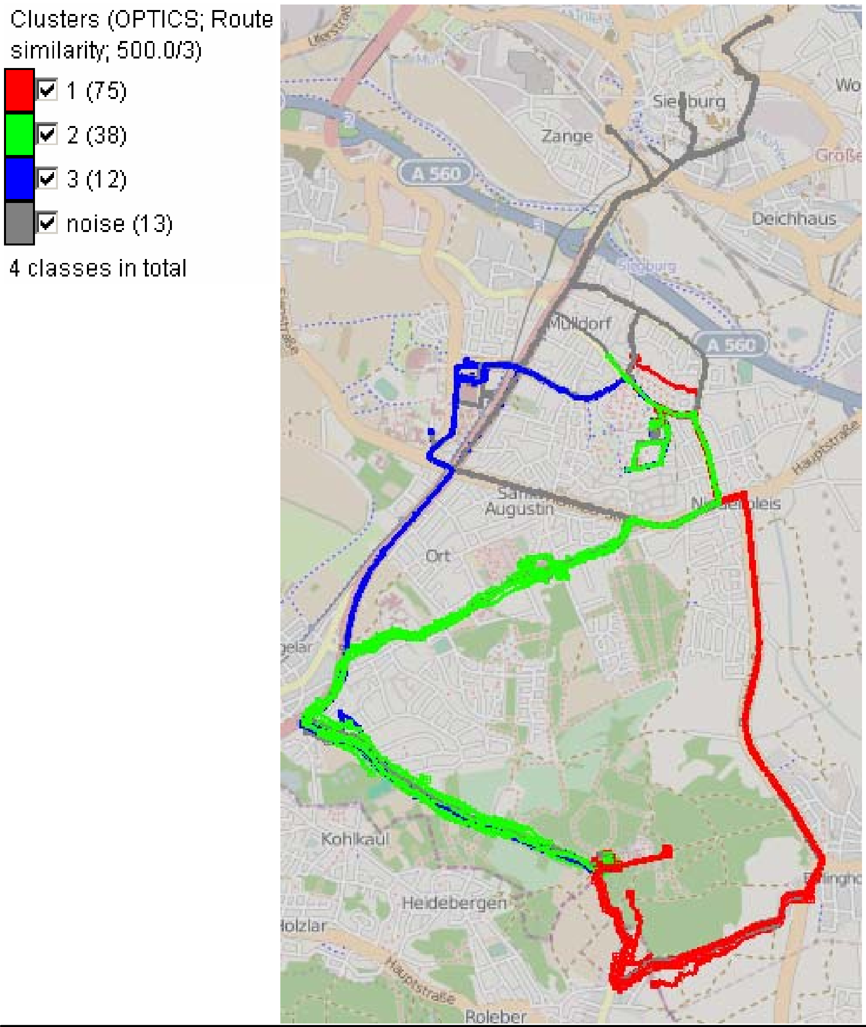
“Progressive clustering”:

The trajectories are very roughly <u>clustered</u> according to the spatial proximity of their starts and ends; further clustering with more sophisticated distance functions

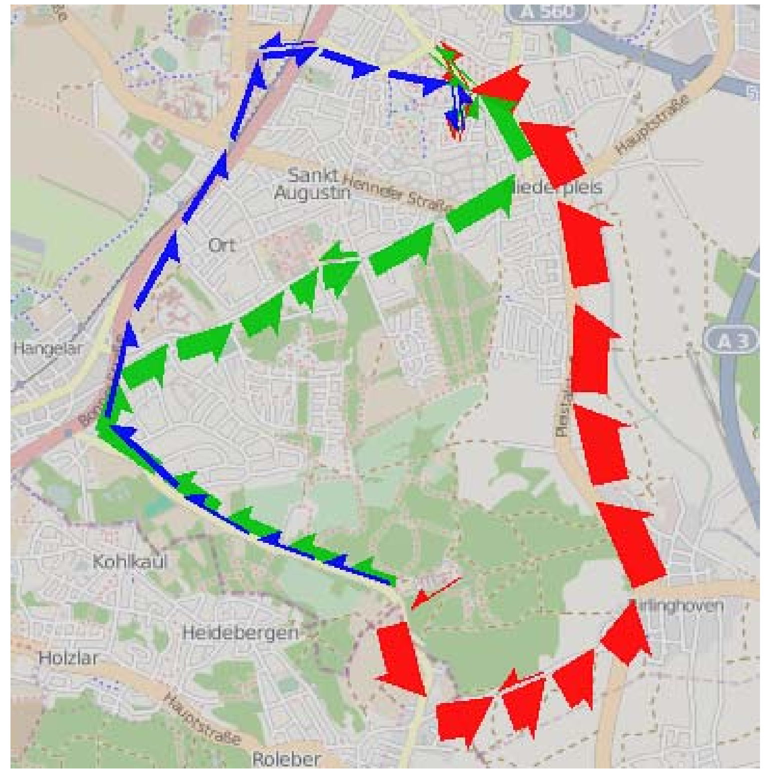




Analysing the trips from work to home (138 trips)



Summarised trajectories excluding the noise

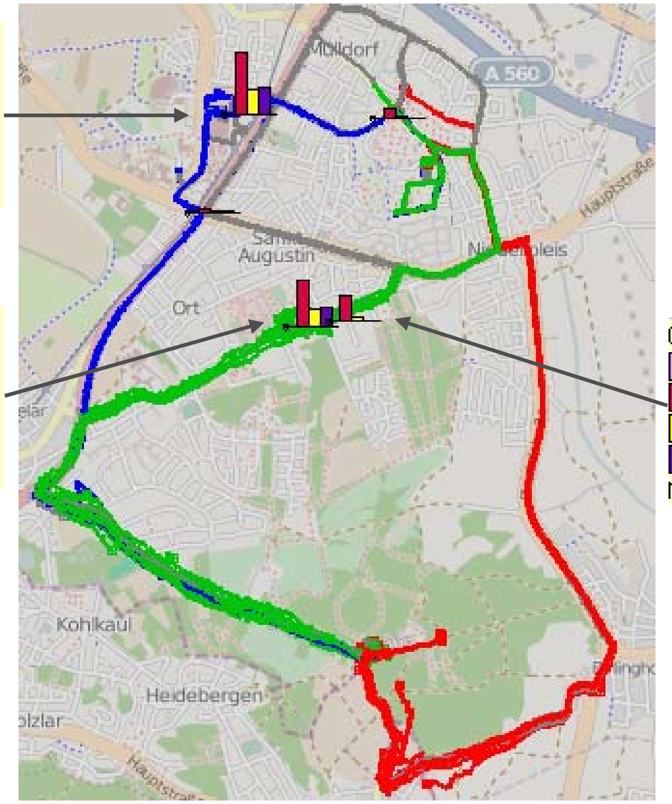


Hypothesis: one of the routes (red) is direct; the others are used for visiting certain places like shops.

Checking the hypothesis by looking at the times spent in different places along the routes

014	ID=014
Min duration of stay (active)	33
Max duration of stay (active)	5135
Mean duration of stay (active)	2116
Median duration of stay (active)	2236
N visits (active)	13

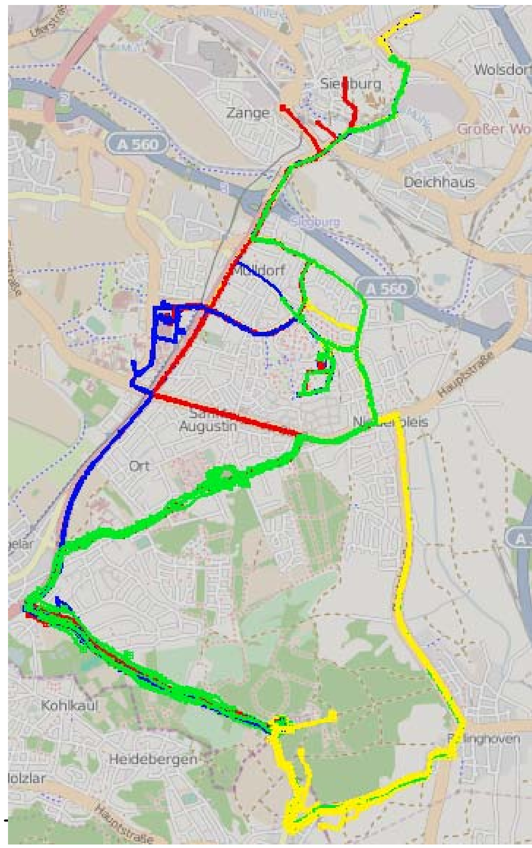
002	ID=002
Min duration of stay (active)	1
Max duration of stay (active)	3861
Mean duration of stay (active)	1479
Median duration of stay (active)	1656
N visits (active)	50



018	ID=018
Min duration of stay (active)	20
Max duration of stay (active)	2122
Mean duration of stay (active)	283
Median duration of stay (active)	47
N visits (active)	50

Is there any regularity in preferring one shopping centre to the other?

On the basis of the clustering results, we have interactively classified the trips according to the visits of the shopping centres.

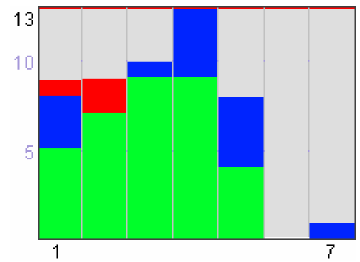


Passing shopping centres

- none: 76 objects (20.8%)
- both: 6 objects (1.6%)
- 1: 14 objects (3.8%)
- 2: 42 objects (11.5%)

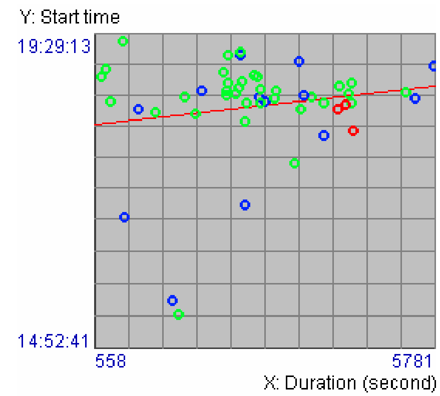
Total: 365 objects; active: 138

Weekly distribution of the trips



Red: trips through both shopping centres

Start times and durations



1) The trips through the shopping centres occur more often on Wednesdays and Thursdays.

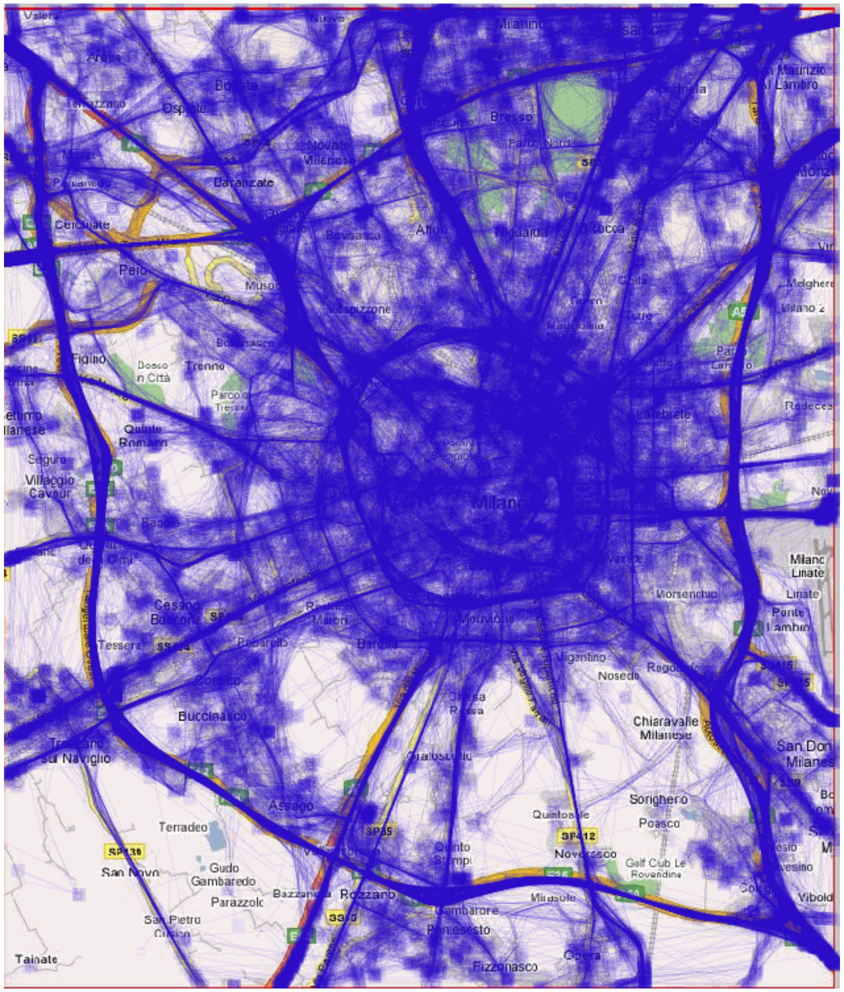
2) Shopping centre 1 is usually visited on Mondays, Thursdays, and Fridays.

3) The starting times and durations of the trips through shopping centres 1 and 2 do not significantly differ.

What we have learned about the car owner:

- The places where the person lives, works, and shops
- The typical routes from home to work, from work to home, to the shopping areas
- The places where the person frequently stops on the way from work to home
- The places where the person may stop on the way from home to work
- The durations of the stops, times spent for visiting the shopping areas
- The times of the trips and of the stops
- How the chosen routes are distributed over the days of the week
- The car owner lives in a small town
- The person has a flexible work schedule
- The person has no small children (concluded from the times of the trips from the work to home)
- The person is often away or sick (judging from the distribution of the trips over the time period, especially in the summer)
  - **A threat to the personal privacy!!!**

Example of visual analytics: analysis of city traffic



E.g. Milan car movement data:

***2,075,216** position records*

of 17,241 cars during 1 week

This is far too much for processing in RAM!

Approach:

1) aggregate the data in the database using the standard DBMS operations;
2) load the aggregates in RAM and visualise for enabling analysis

Spatial aggregation: by compartments (cells) of a territory division

Temporal aggregation: by time intervals

- linear time
- cyclic time

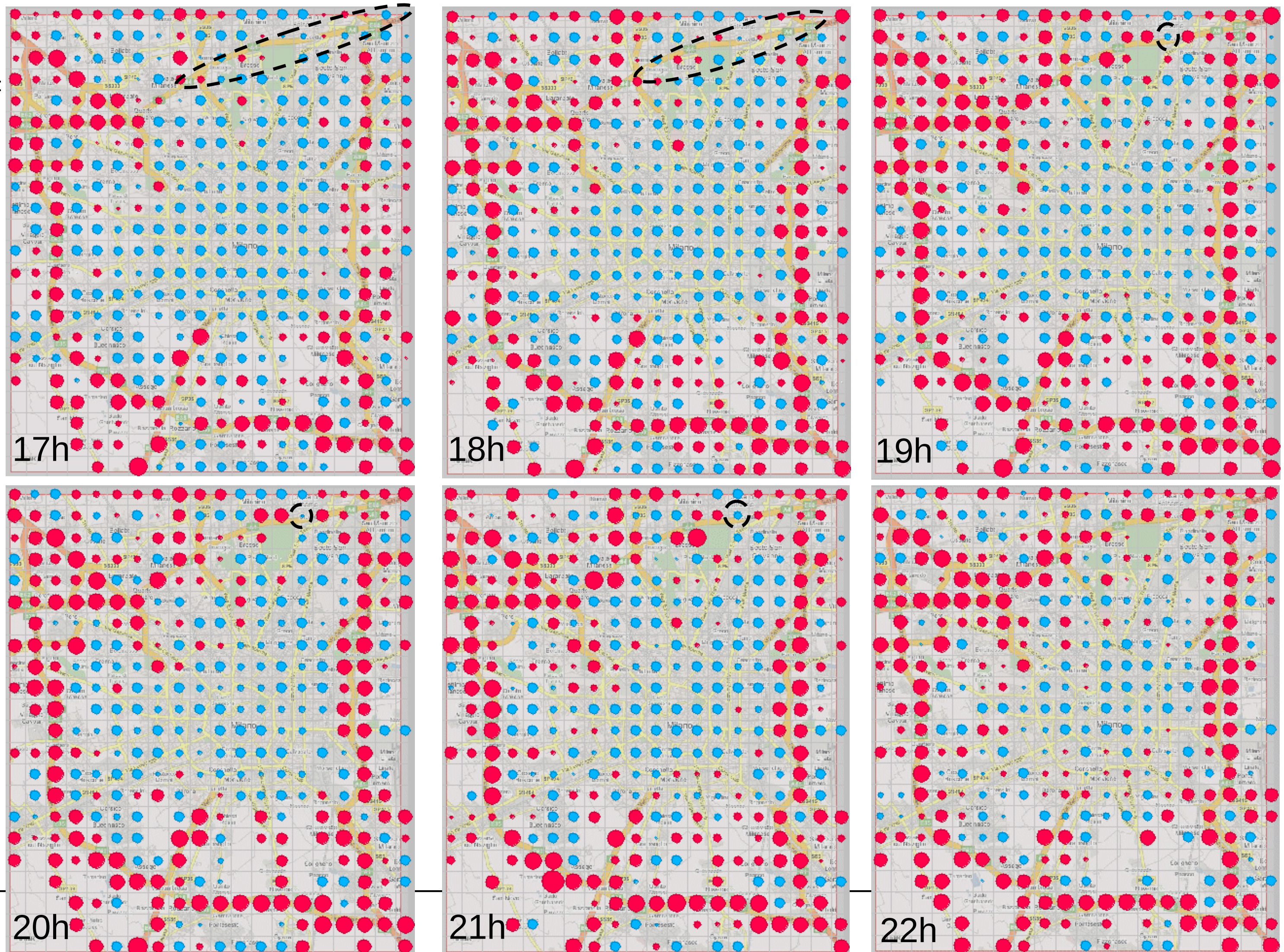
Median speeds by hourly intervals

Car data aggregated by grid
median speed (Day of week=1)

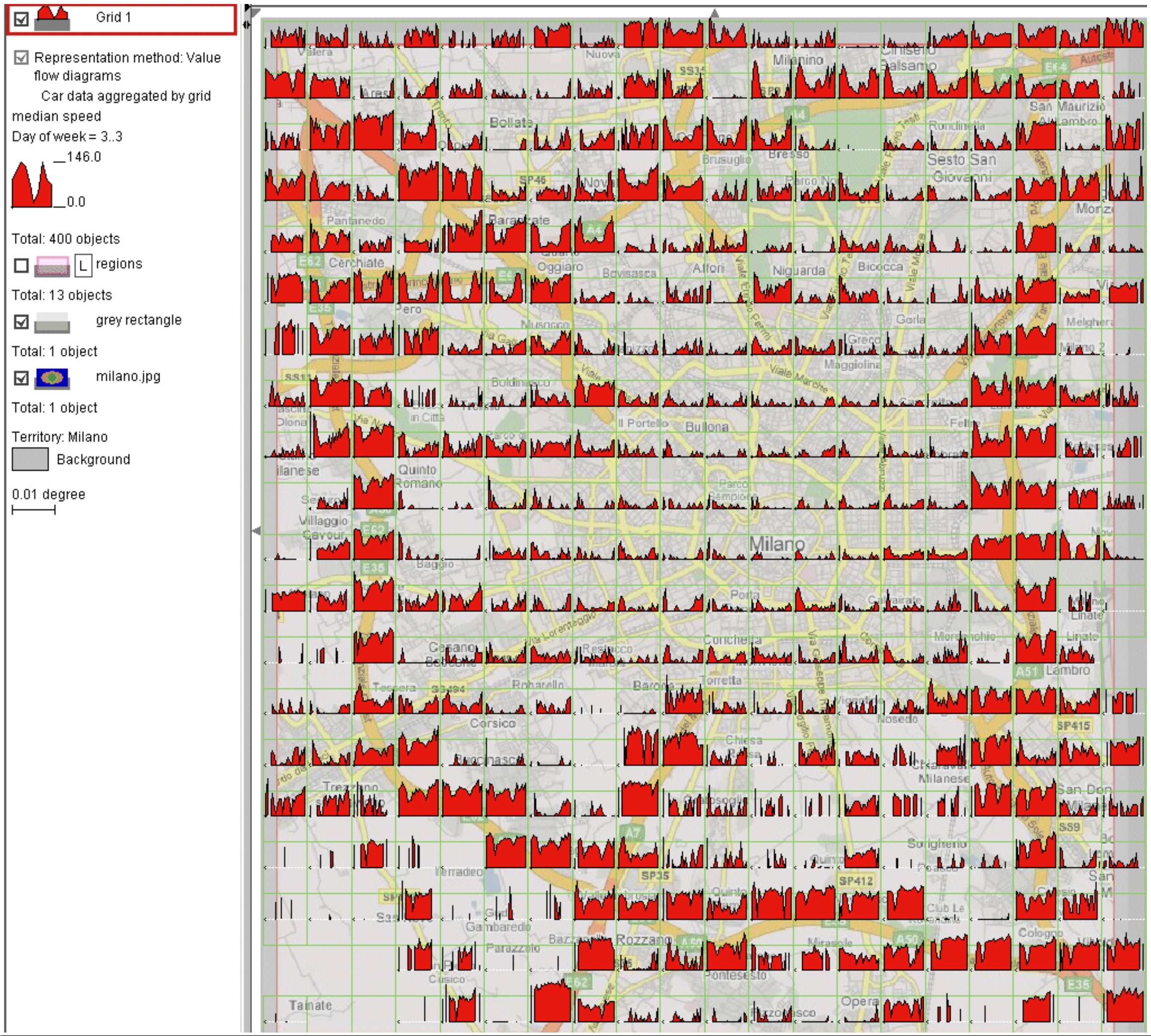
Circle area is proportional to value:

→ 264.0
→ 30.0
→ 0.0

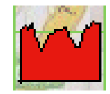
very low speeds on a major belt road



Hourly variation of the median speed in different places



day: Wednesday

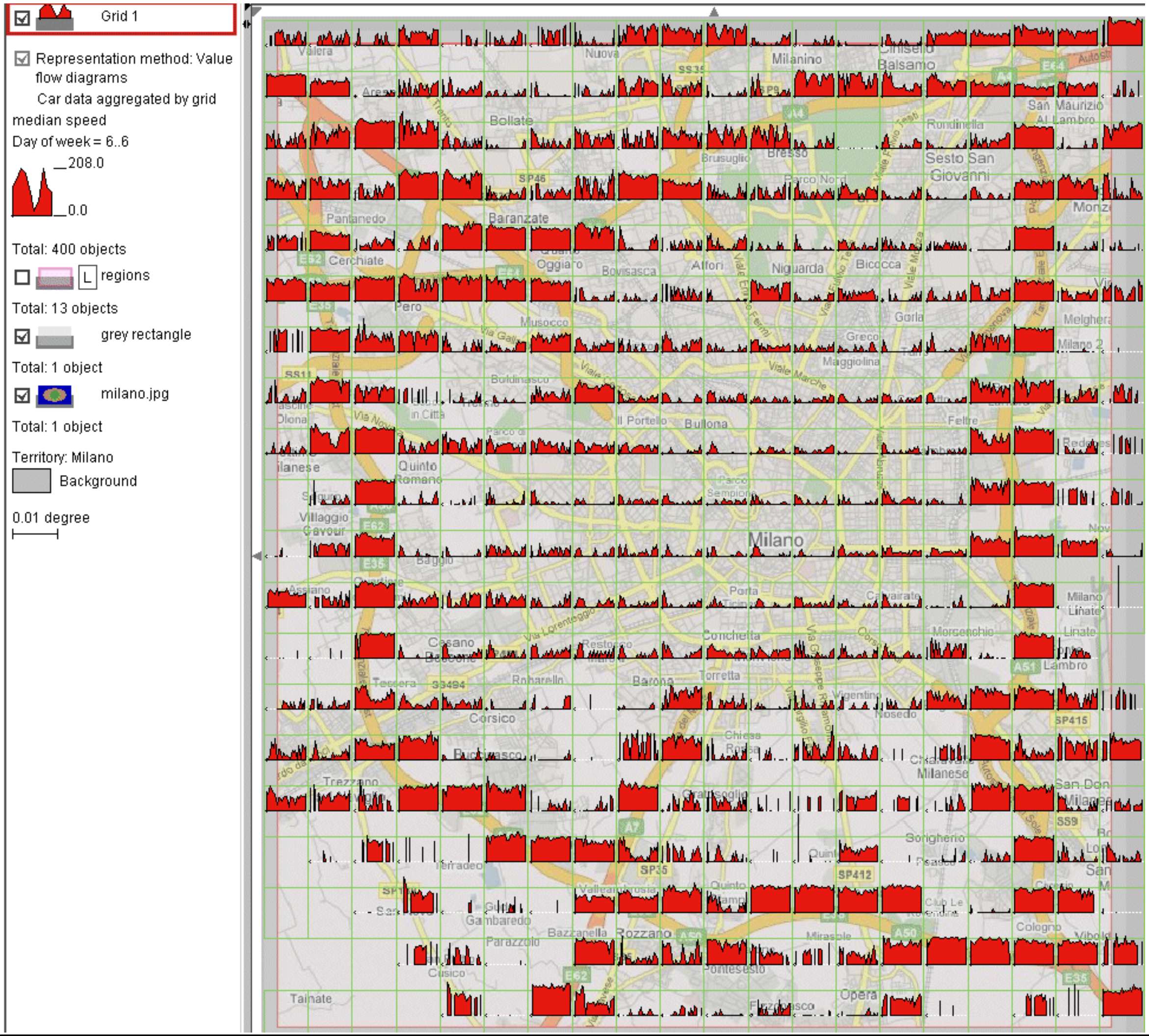


A frequent temporal pattern: significant drop of the speed in the morning and afternoon rush hours

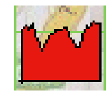
universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

Fraunhofer Institut Intelligente Analyse- und Informationssysteme
IAIS

Hourly variation of the median speed in different places



day: Saturday

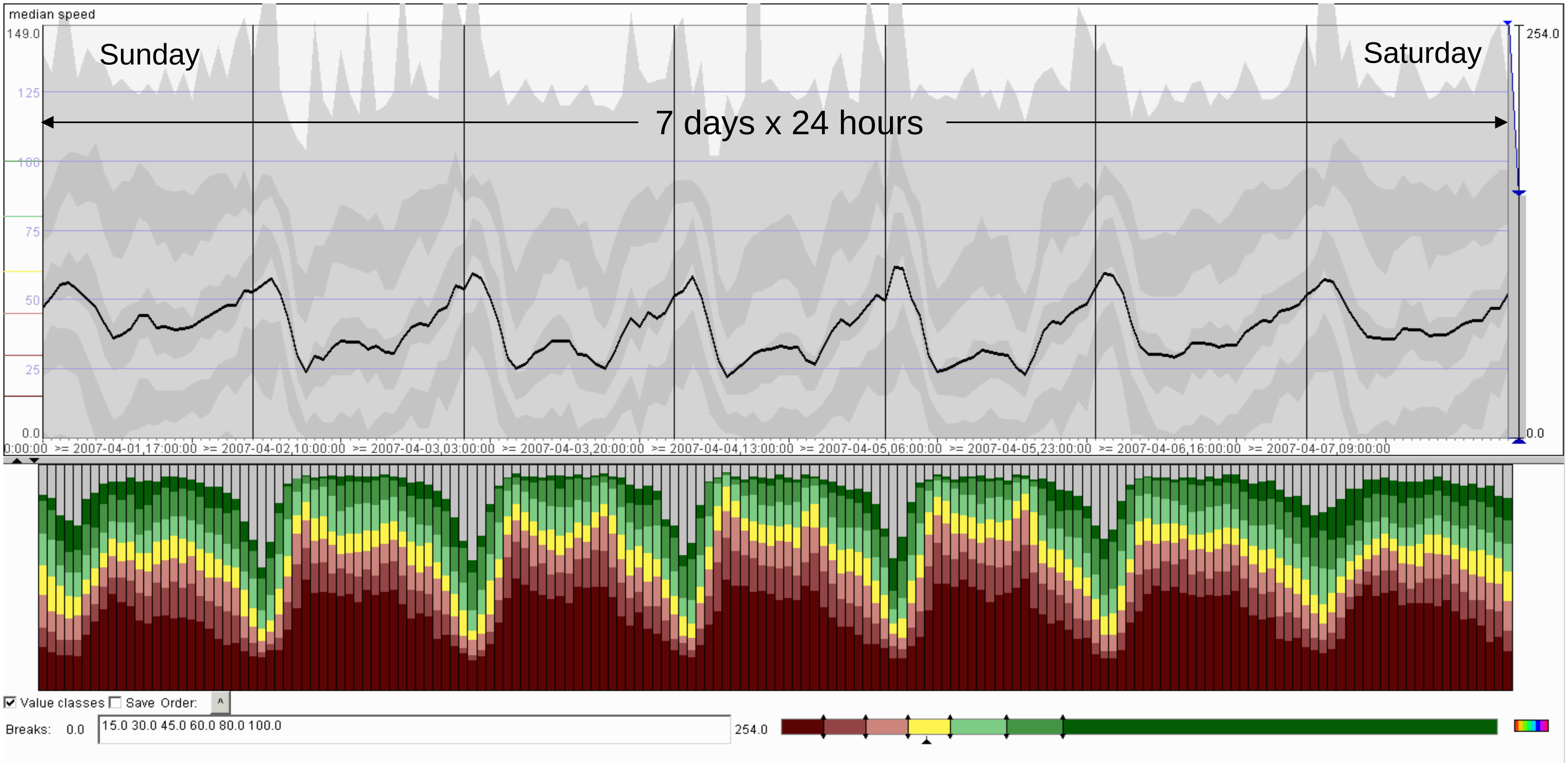


This temporal pattern does not occur on Saturday

universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

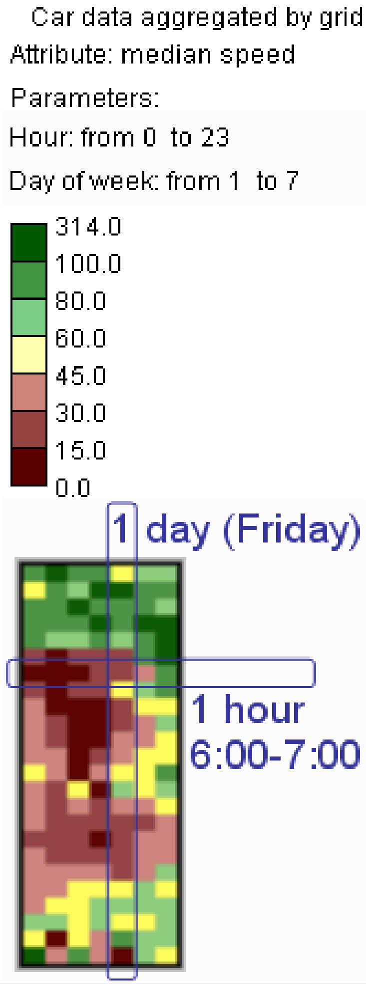
IAIS
Fraunhofer Institut Intelligente Analyse- und Informationssysteme

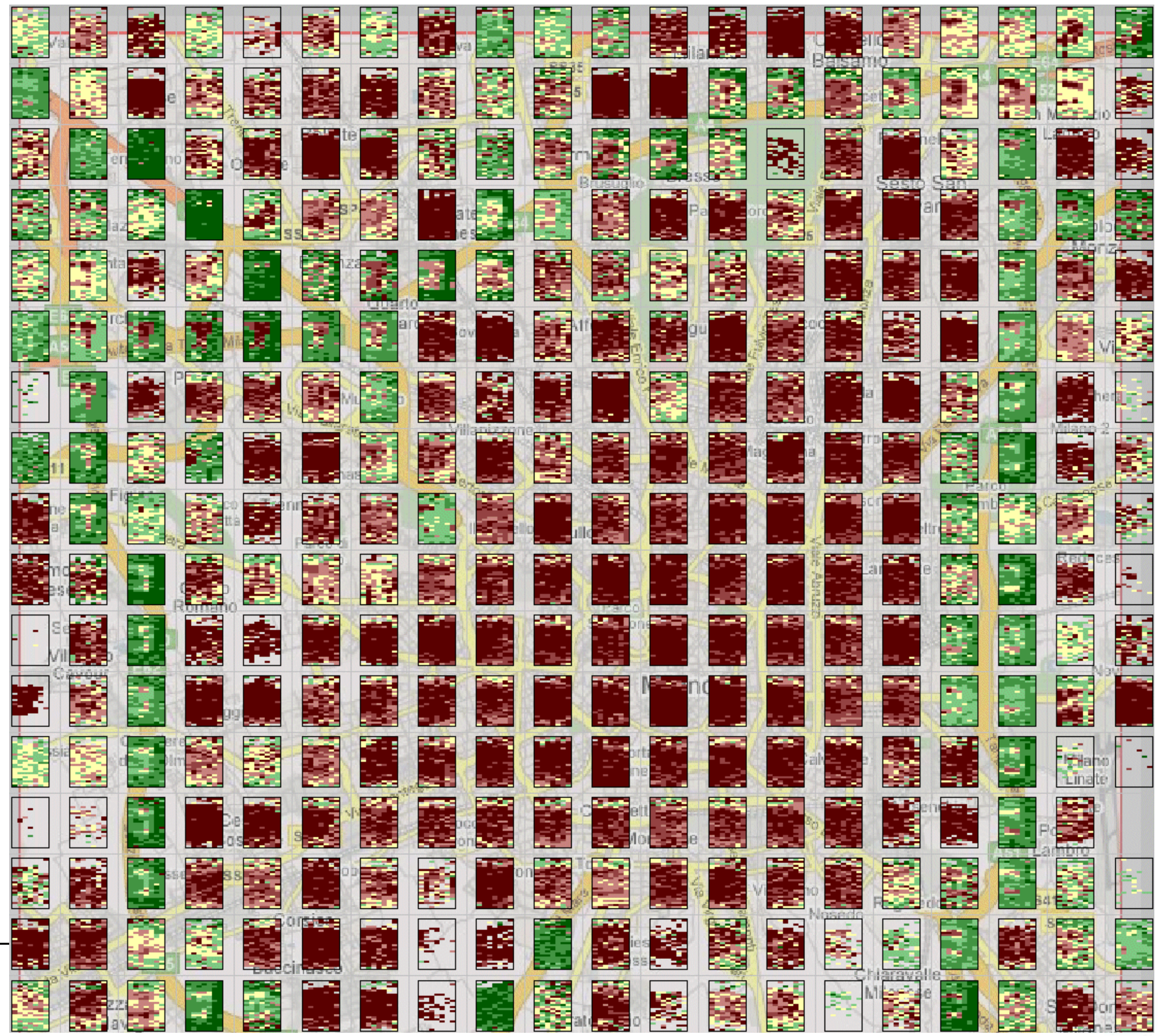
Variation of the median speeds in all cells over the whole time period



Note the similarity of the daily patterns from Monday to Thursday and the difference of the Friday pattern

Daily and weekly cycles in the speed variation over the territory

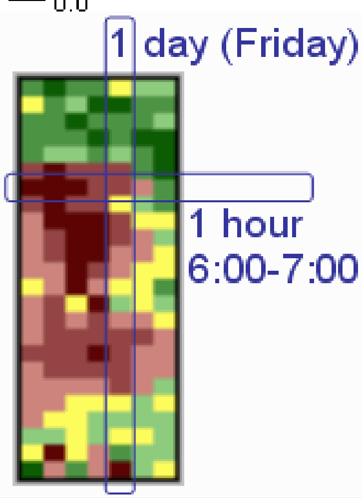


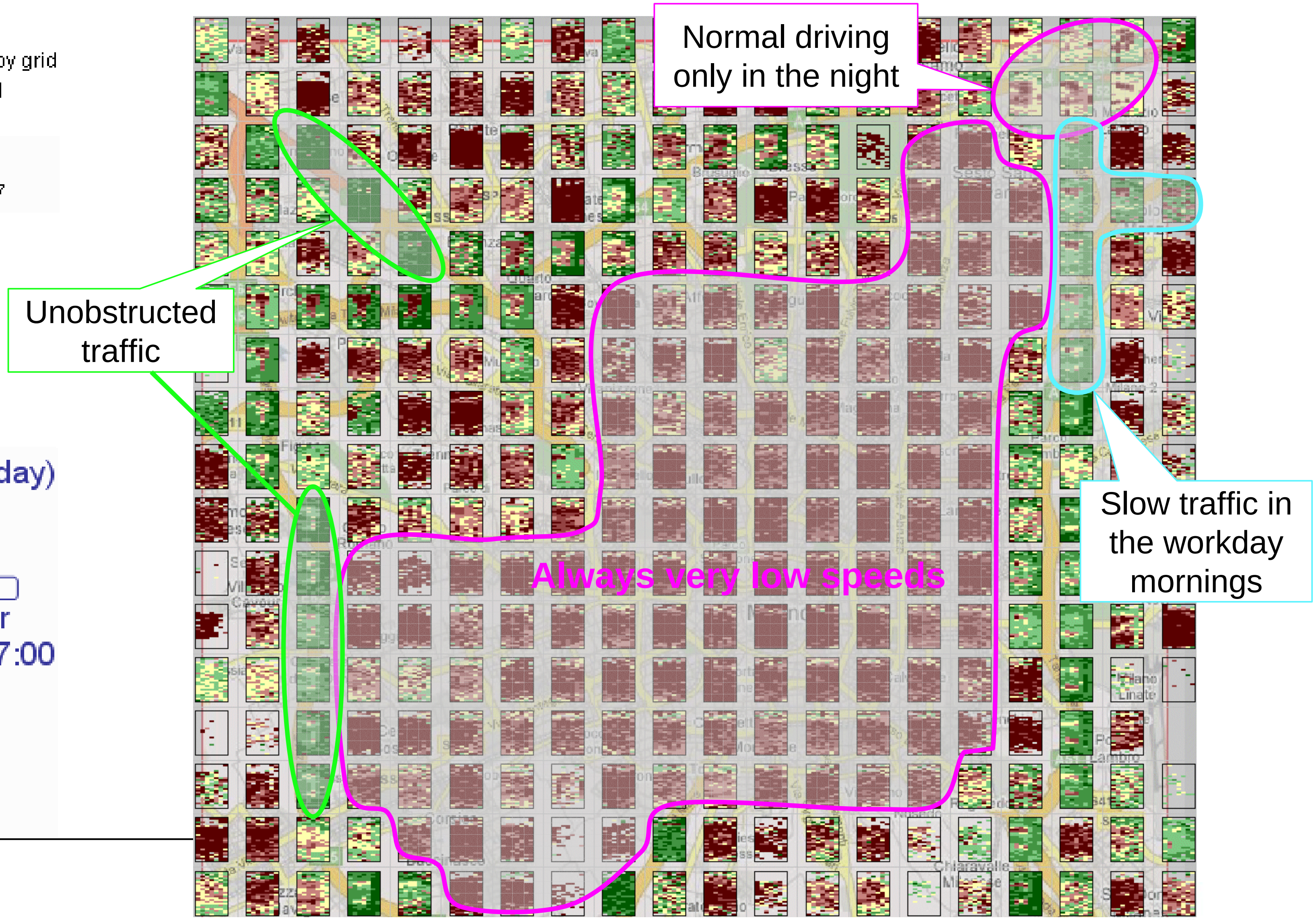


Daily and weekly cycles in the speed variation over the territory

Car data aggregated by grid
Attribute: median speed
Parameters:
Hour: from 0 to 23
Day of week: from 1 to 7

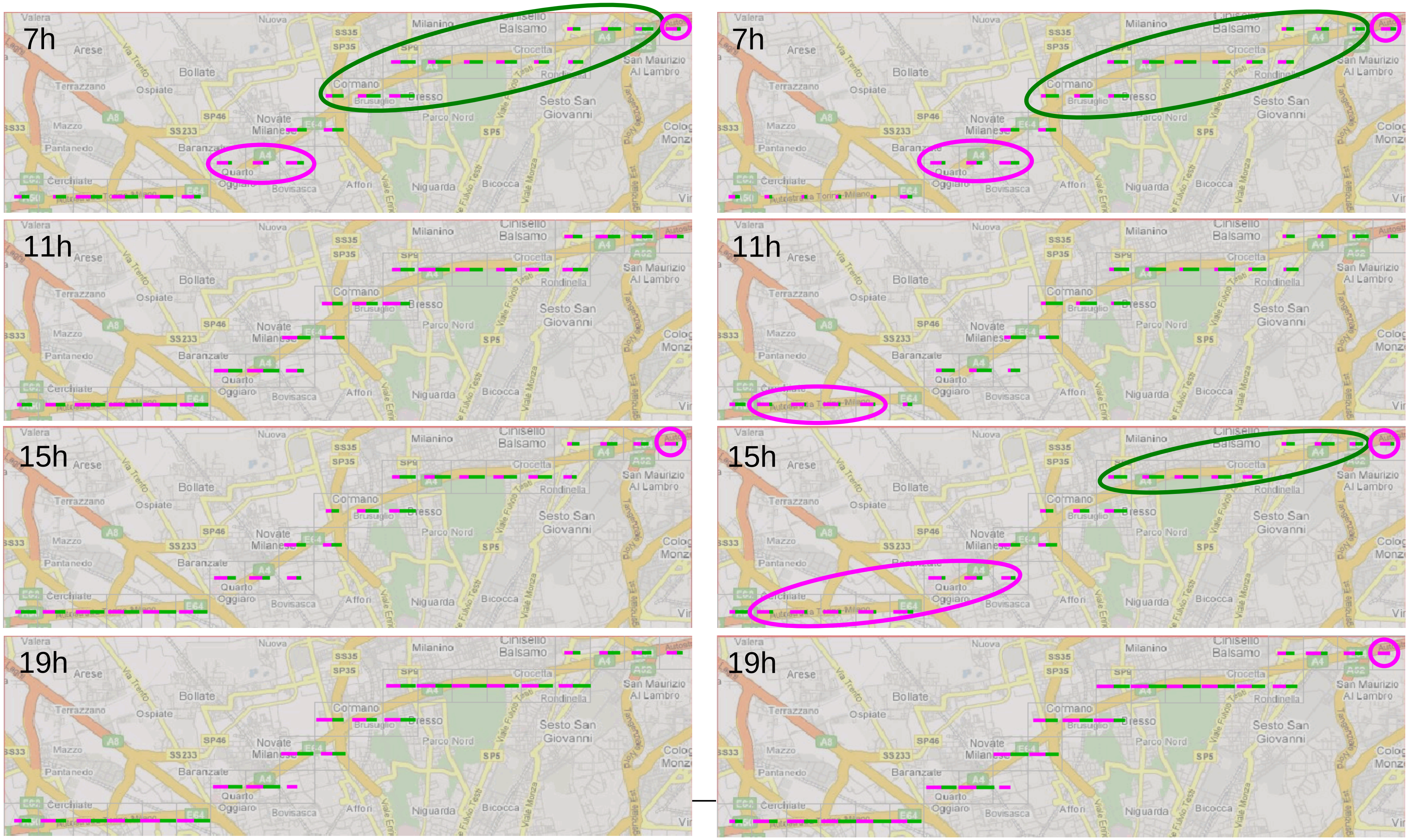
314.0
100.0
80.0
60.0
45.0
30.0
15.0
0.0





Aggregation by movement directions

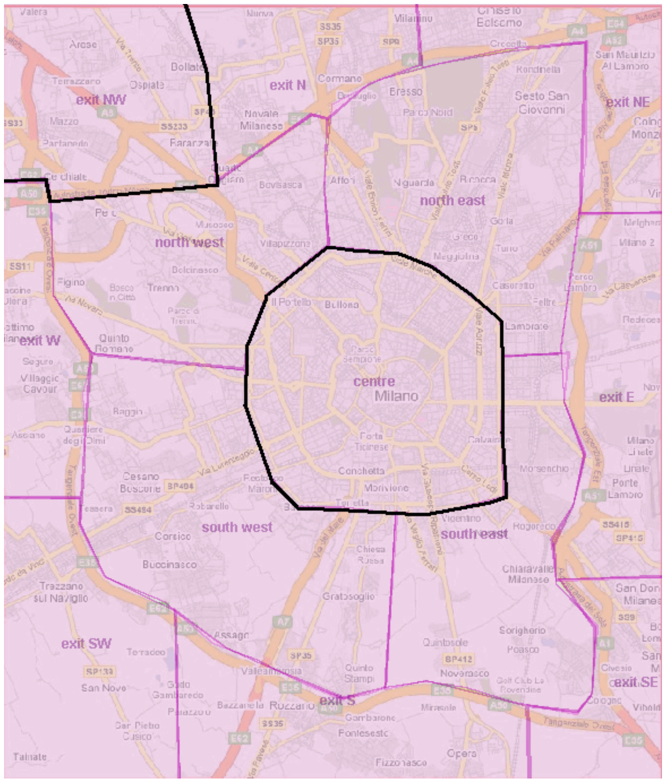
The bar lengths are proportional to the **median speeds** of the cars moving in the respective directions.



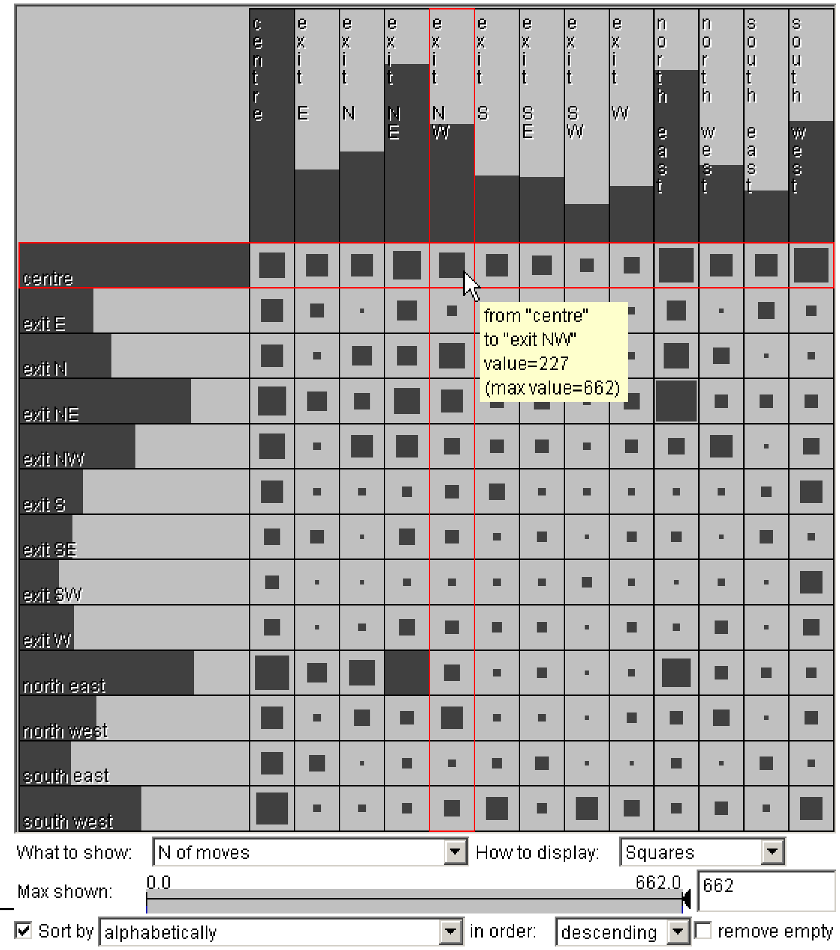
Note the asymmetry of some diagrams signifying different load of the street in two directions

Aggregation by origin-destination

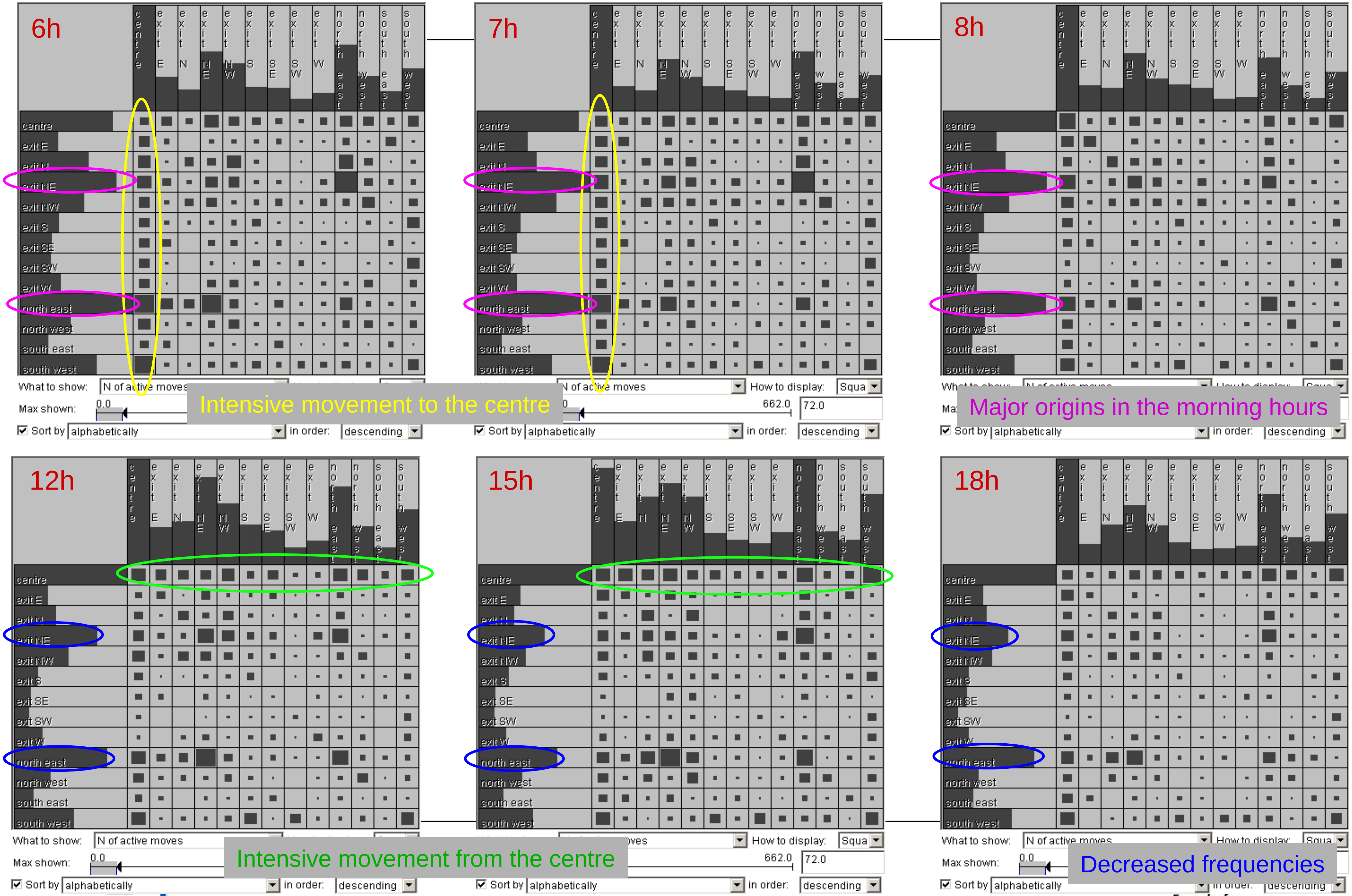
1. The start and end points of the trajectories or segments of trajectories are referred to areas in space



2. For each pair of areas, the number of trajectories and statistics of their numeric properties (duration, distance) are computed



time intervals of the length of 1 hour



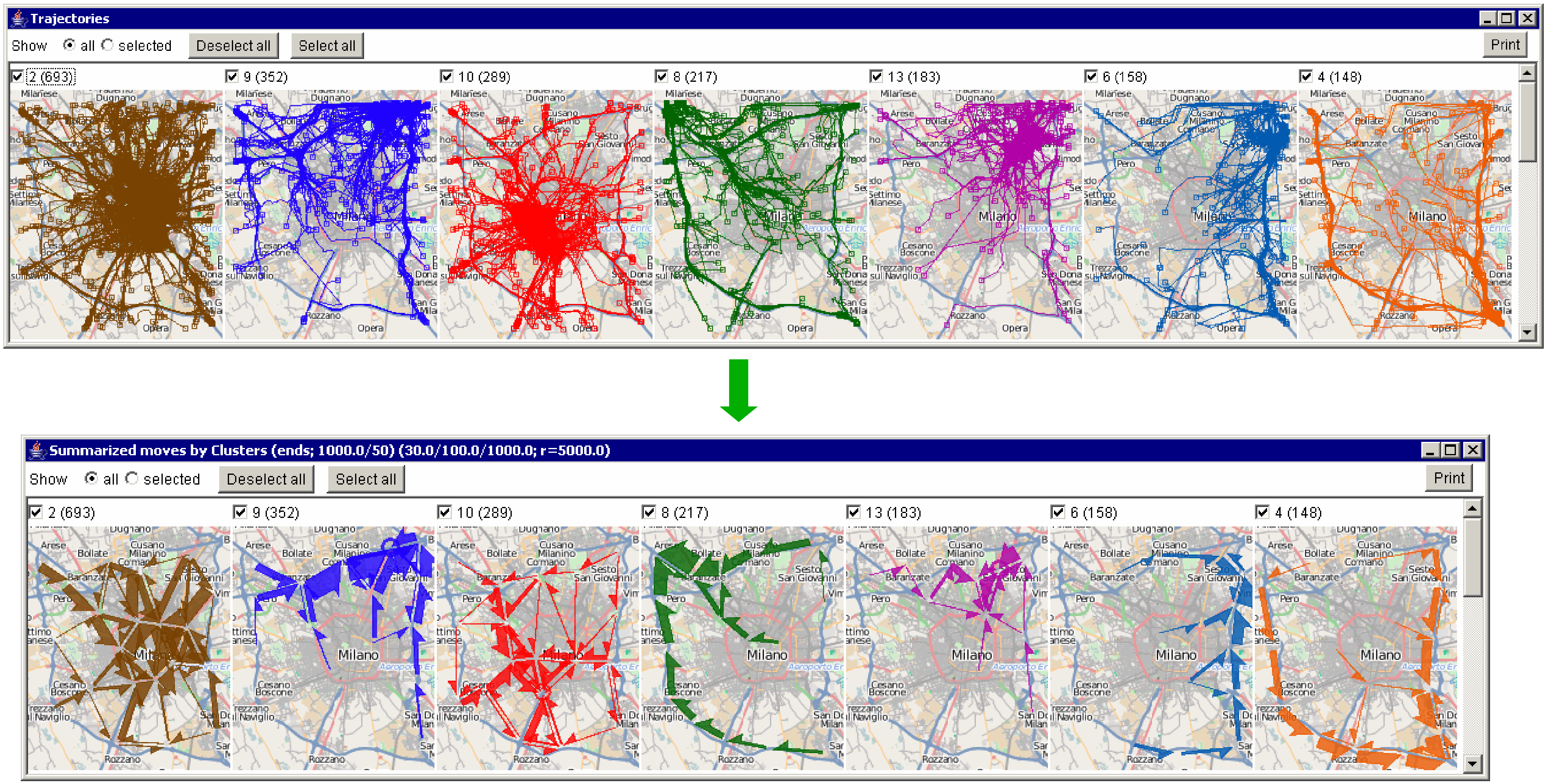
The standard spatio-temporal aggregation is not enough

- The aggregation allows us to deal with very large amounts of data
- It helps us to notice significant patterns and answer a number of questions about collective movements
- … but we do not see the routes of the movement!

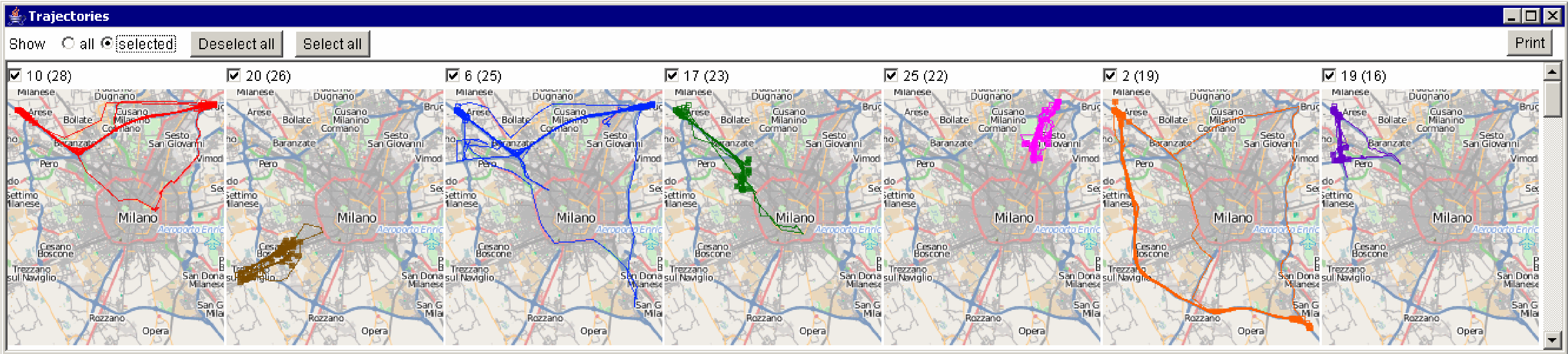
What can be done?

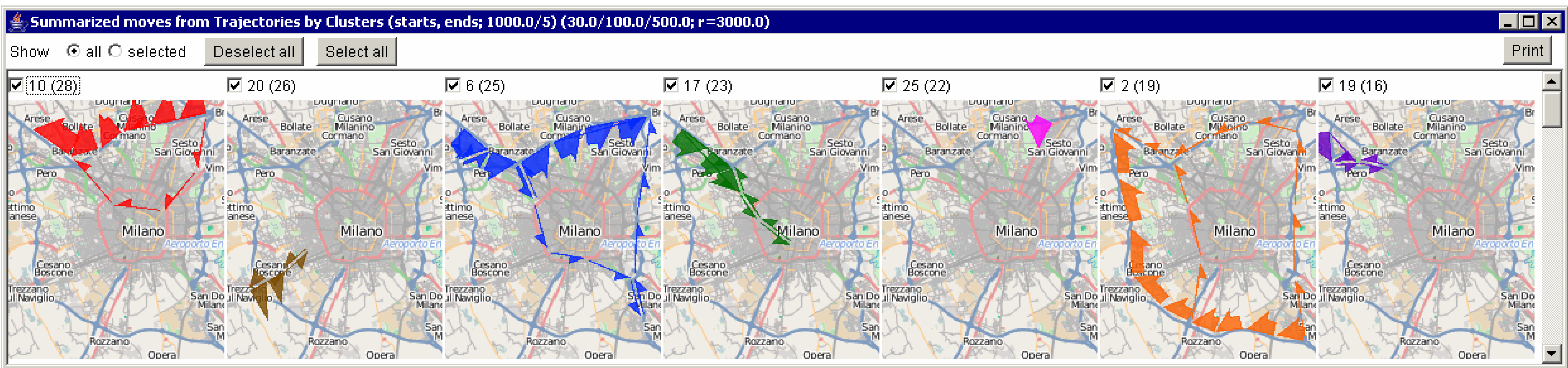
- Clustering techniques allow us to group trajectories by similarity of the routes
- We can examine and compare the groups instead of looking at every trajectory, which is unfeasible
- Groups of similar trajectories need to be represented in such a way that the routes are perceptible

Example: clusters of trajectories by close destinations

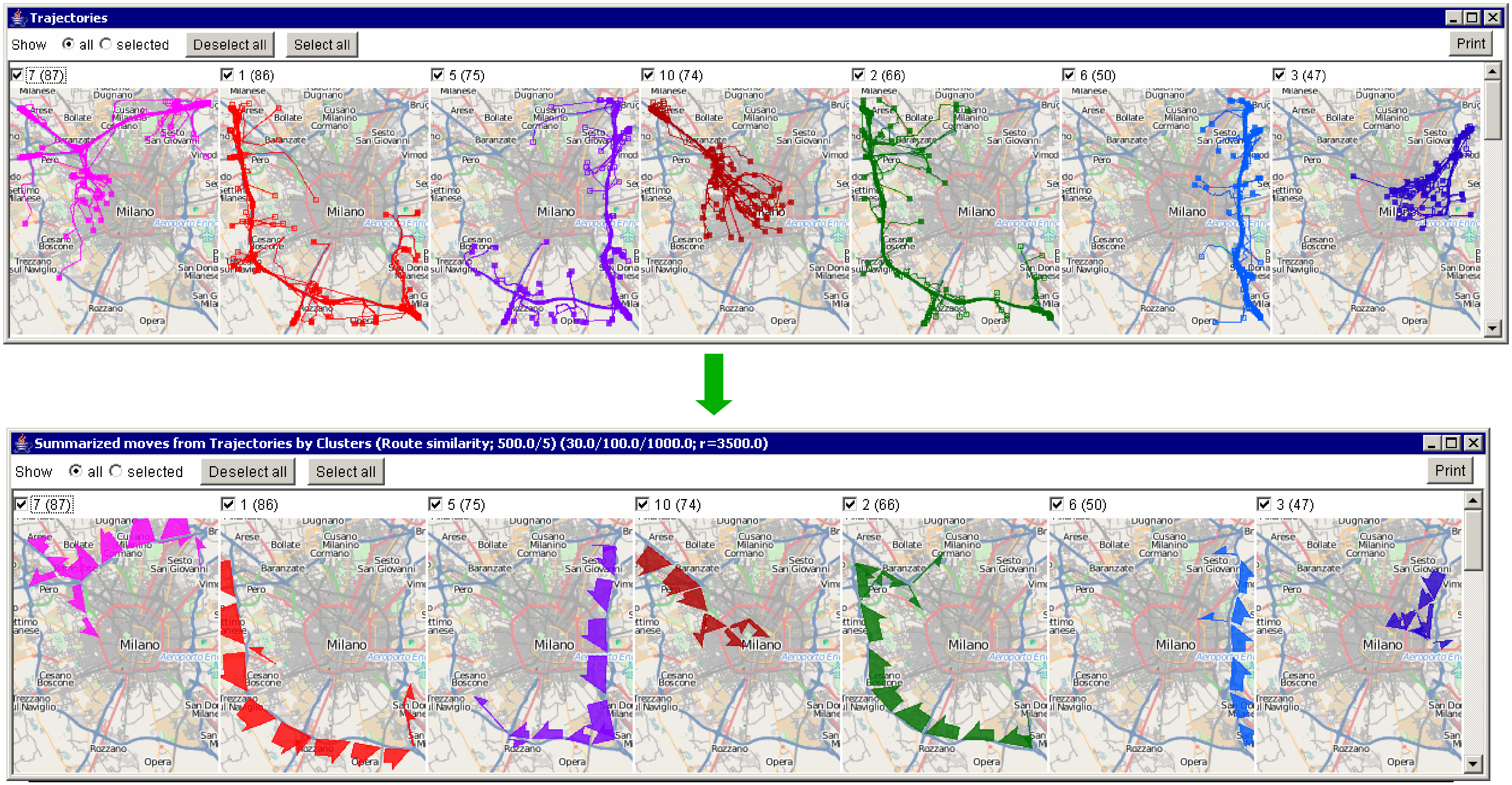


Example: clusters of trajectories by common origins and destinations

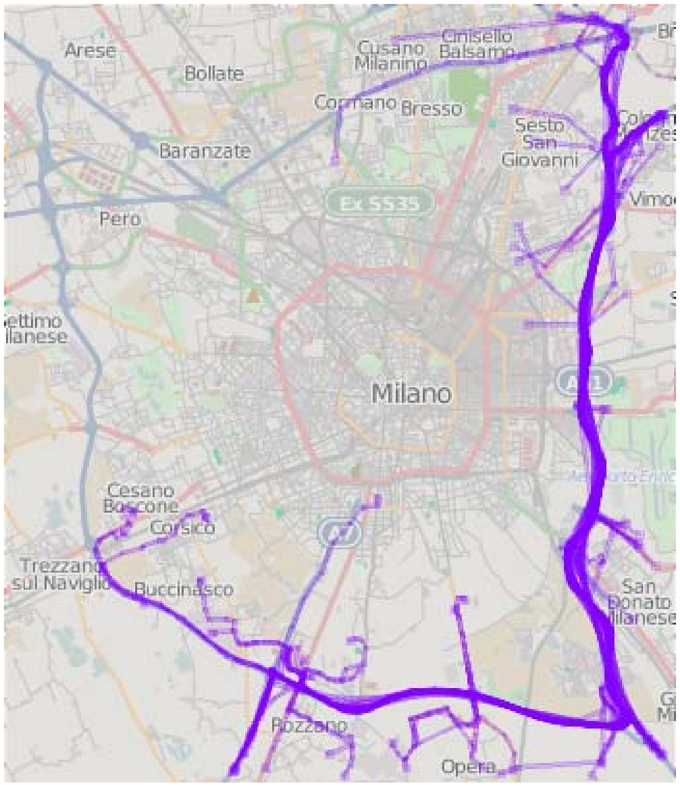


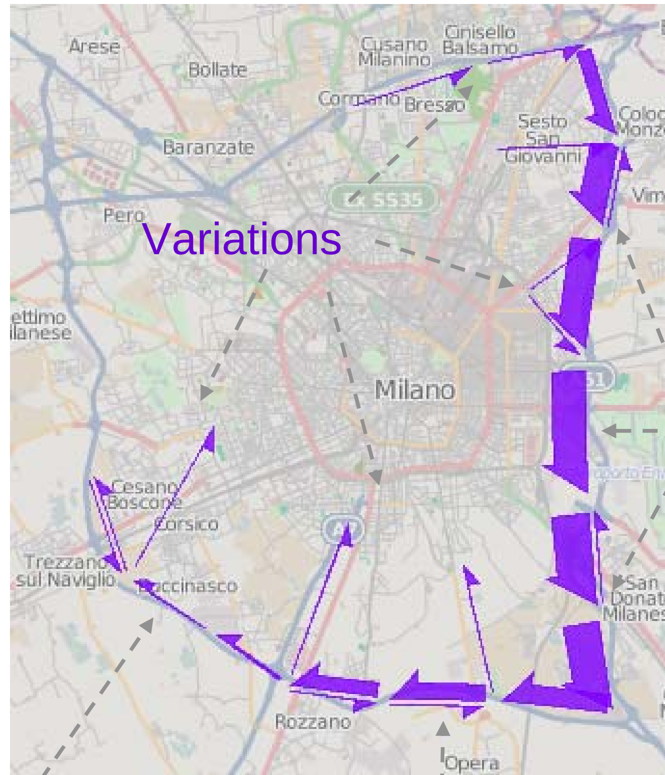


Example: clusters of trajectories by similar routes



A spatial summary of a group of trajectories



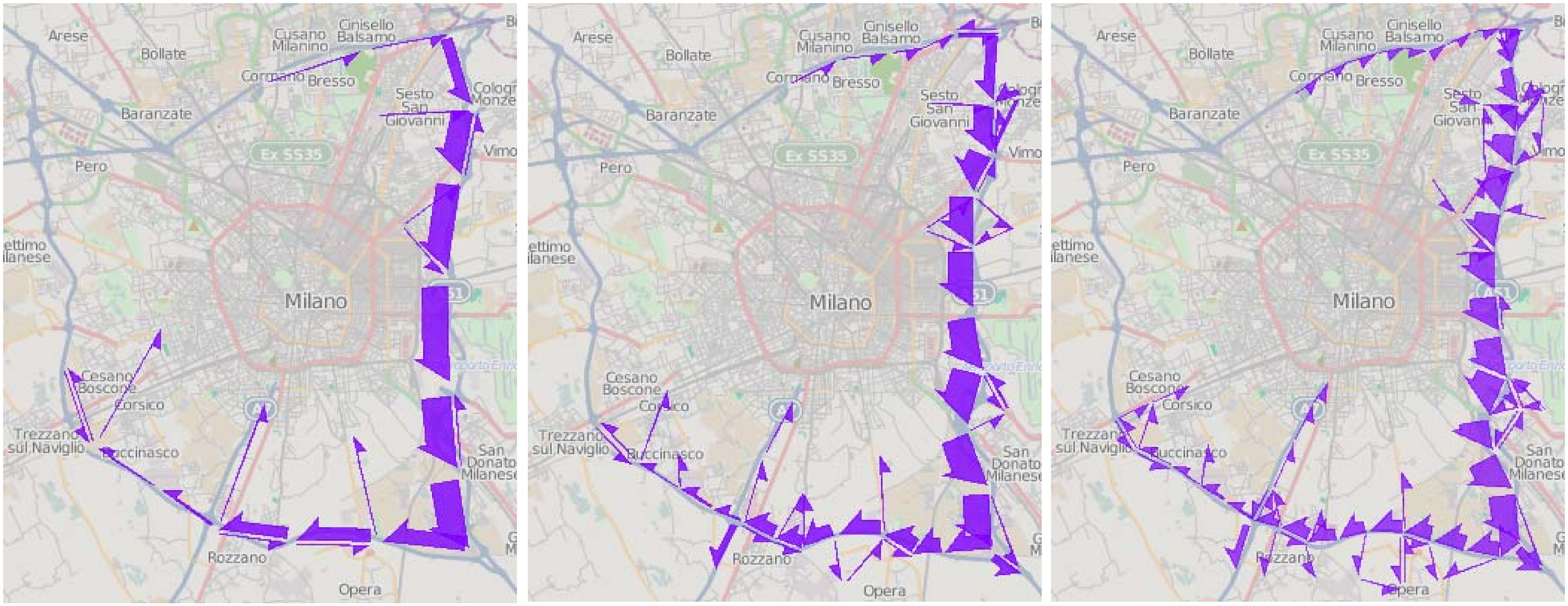


The “core” of the route, common part of the most of the trajectories

Quite a few cars continue this way

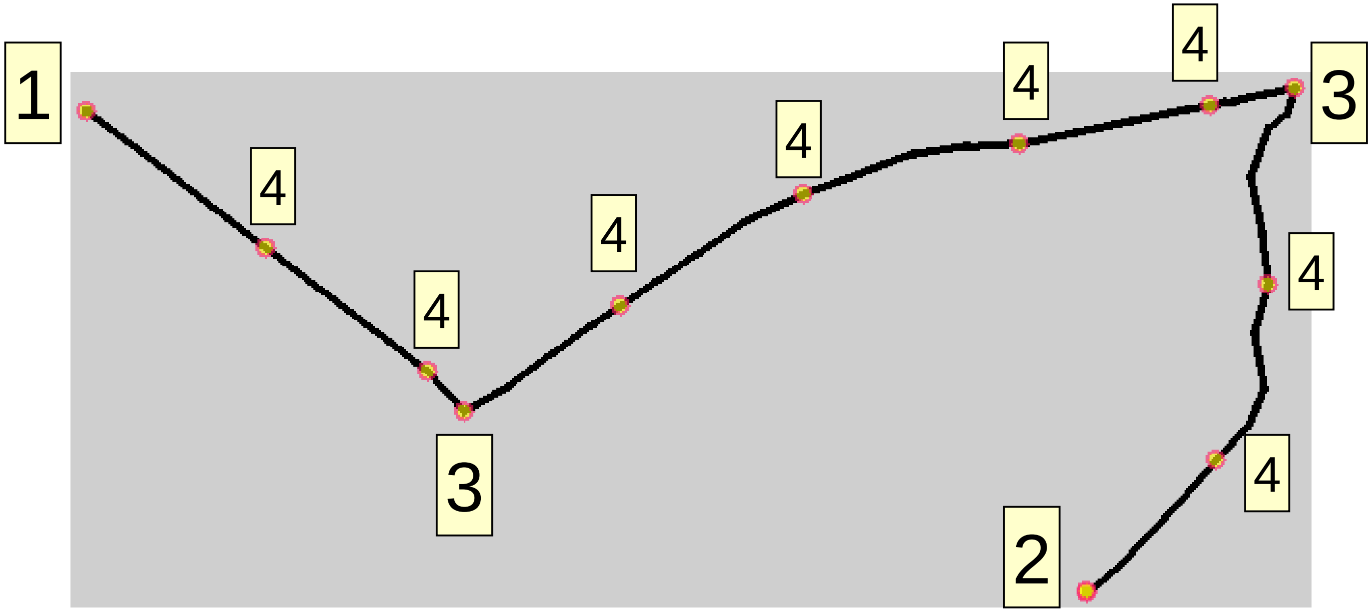
Fewer cars continue this way

Different spatial scale and level of detail



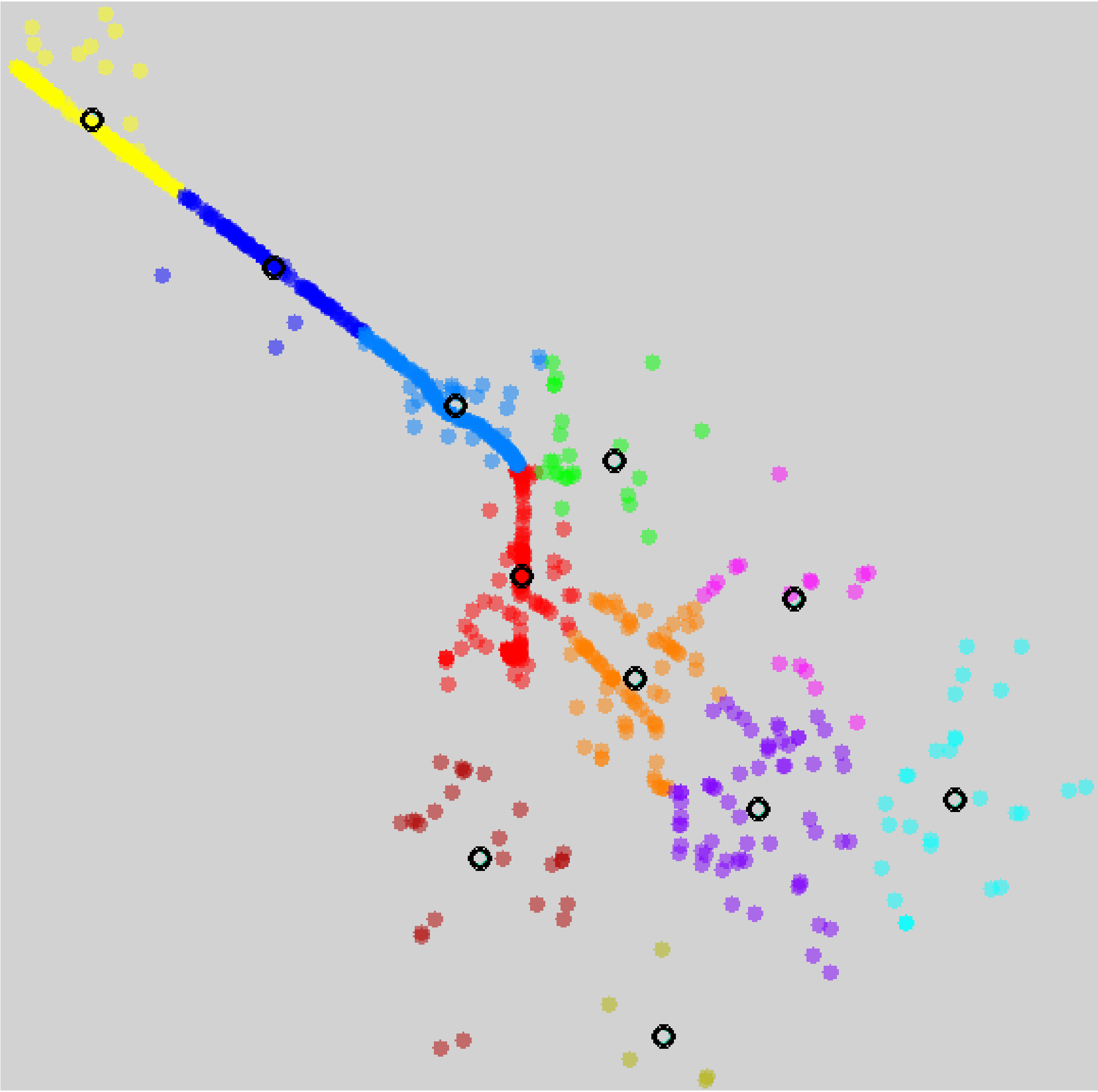
How is it done?

1. Extract characteristic points from the trajectories. Characteristic points include starts (1), ends (2), points of significant turns (3), points of significant stops, and representative points from long straight segments (4).



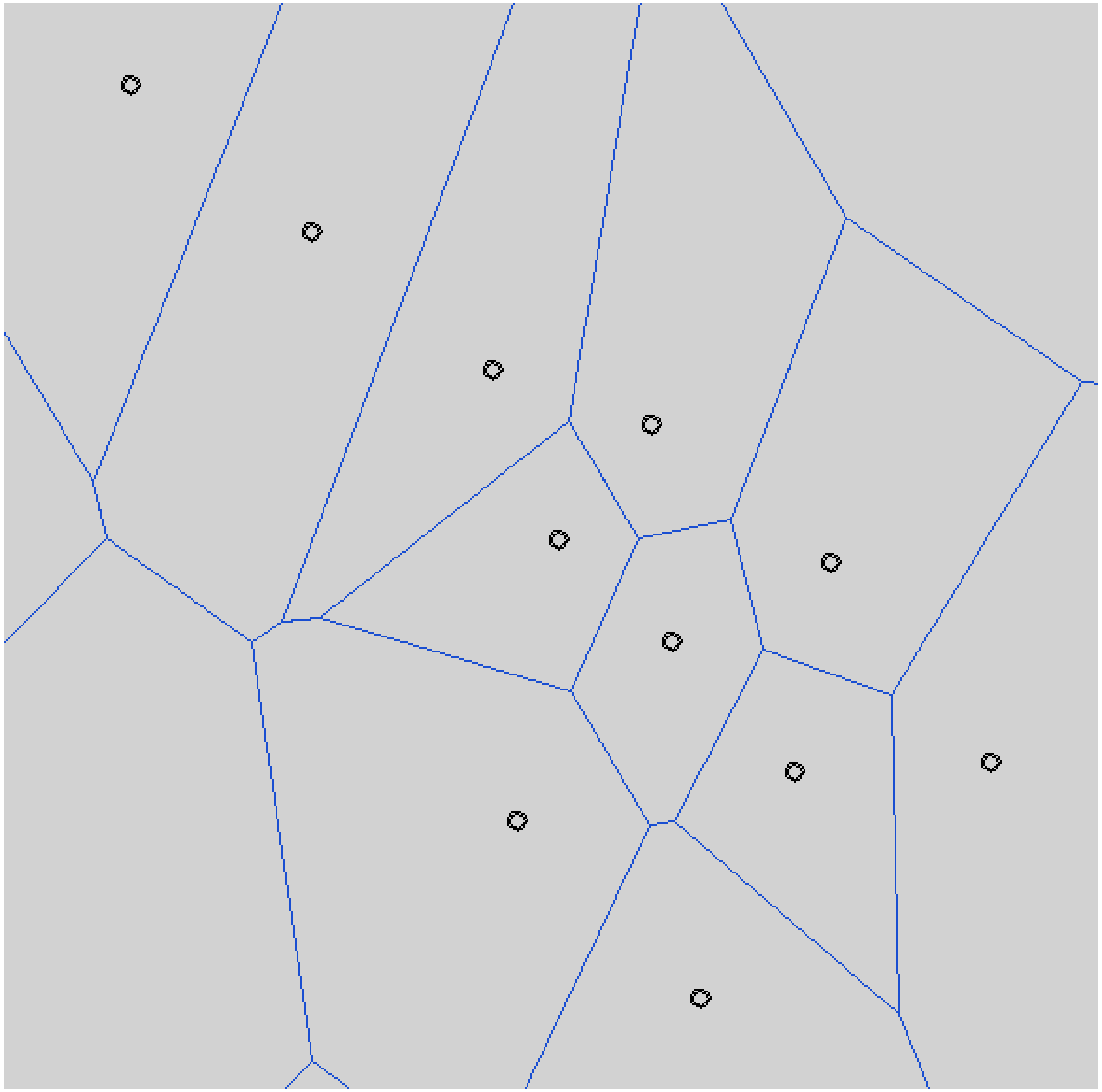
How is it done?

2. Group the extracted points in space so that the groups have a desired spatial extent. The extent is specified by the parameter MaxRadius, which determines the degree of the generalization.



How is it done?

3. Partition the territory into Voronoi cells using the centroids of the groups of points as generating points.



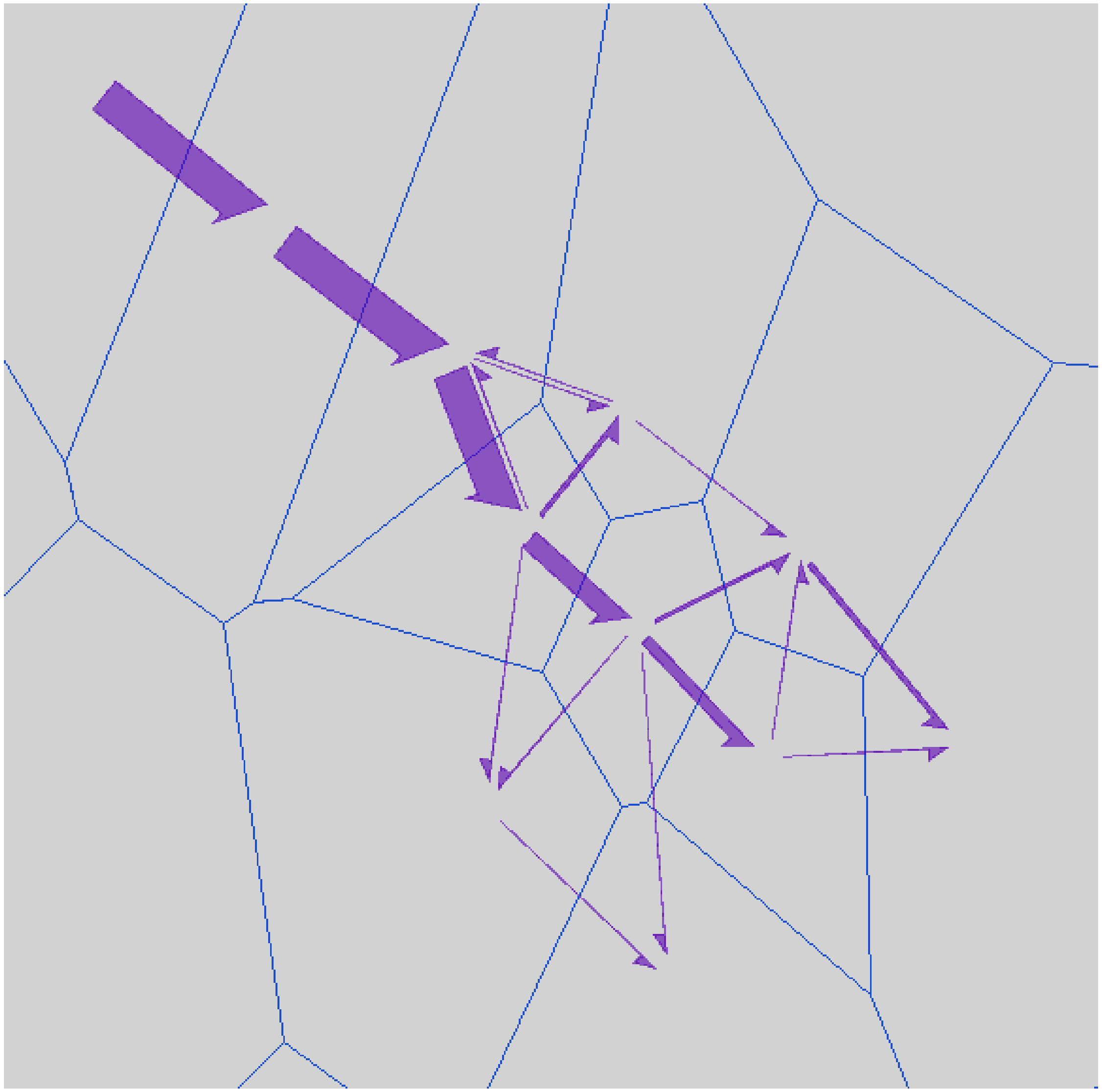
How is it done?

4. Divide the trajectories into segments that link Voronoi cells.

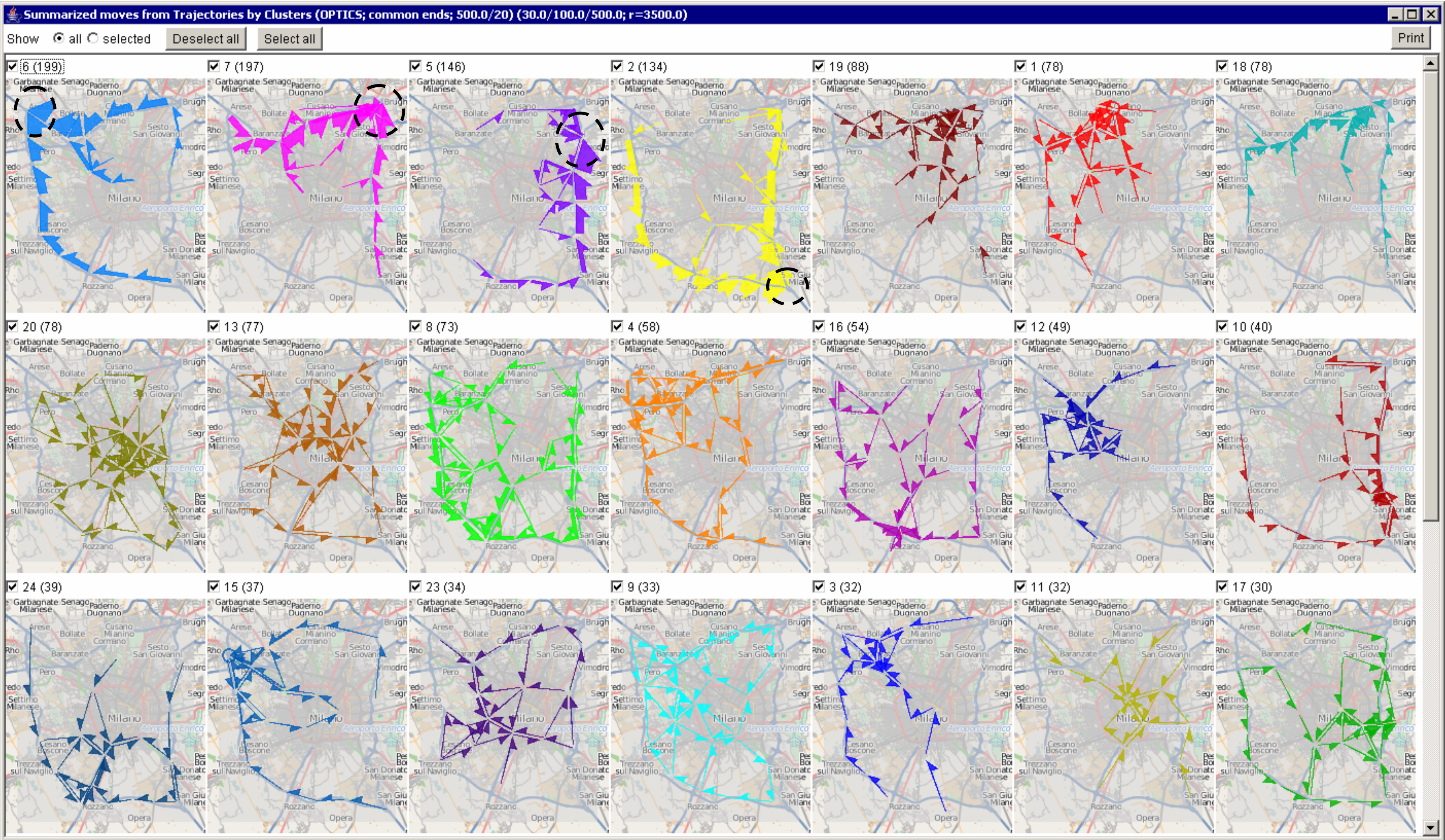
How is it done?

5. For each ordered pair of cells, combine all segments starting in the first cell and ending in the second one into an aggregate move.

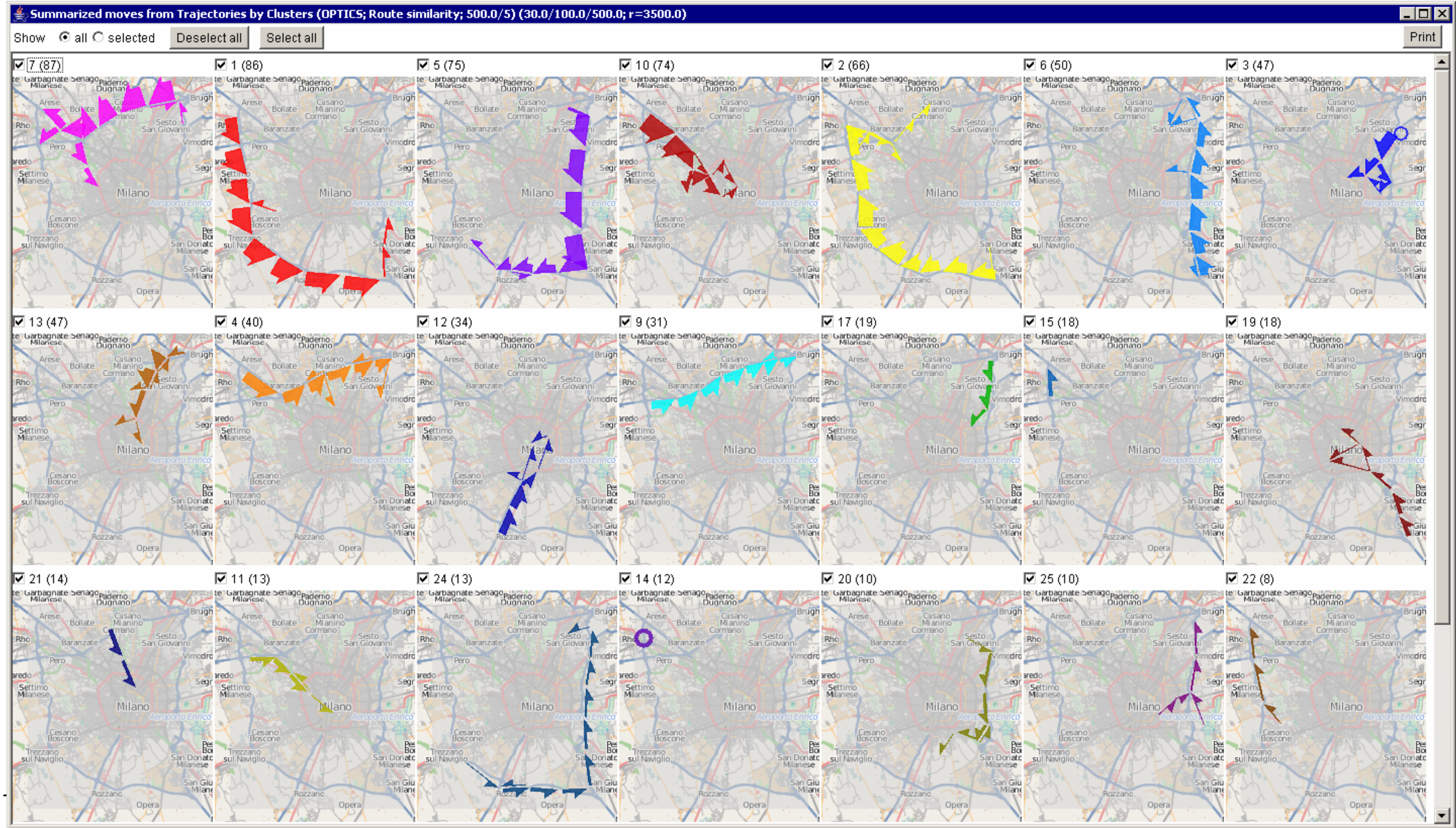
   Represent the aggregate move in a map by a half-arrow symbol connecting the centers of its start and end cells so that the thickness of the symbol is proportional to the number of the trajectory segments combined in the aggregate move.



What are the most popular destinations in Milan on Wednesday morning? From where and how do people reach them?



What are the most frequent car routes in Milan on Wednesday morning?



Can we answer such questions for the whole set of Milan car trajectories?

- Problem: clustering of complex objects (such as trajectories) involving non-trivial distance functions (such as “route similarity”) can only be done in RAM, i.e. for a relatively small dataset

- Our approach:

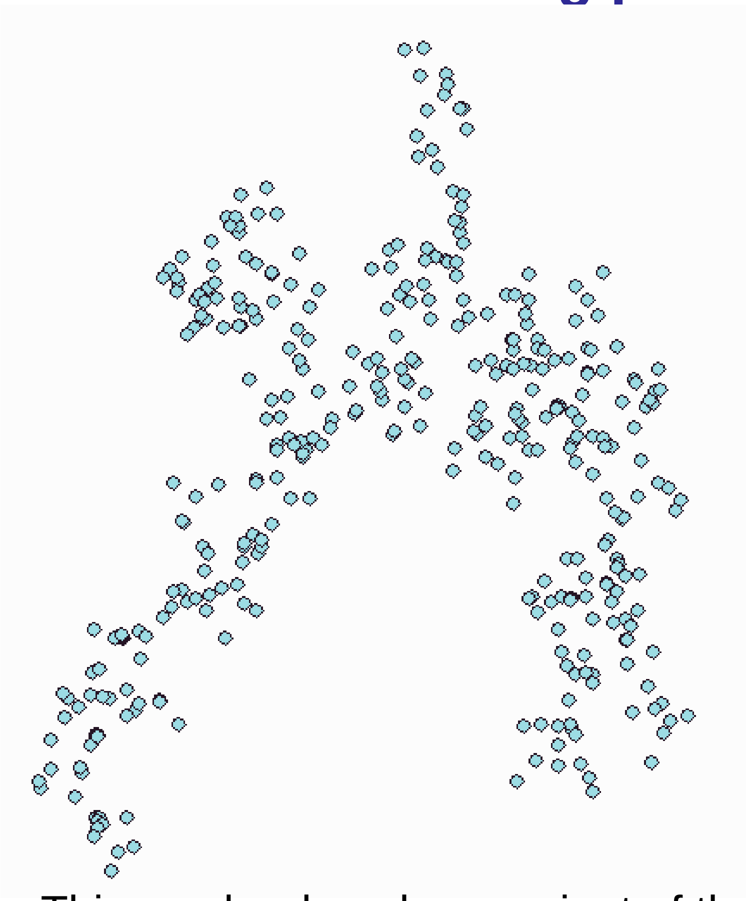
  1. Take a subset (sample) of the objects suitable for processing in RAM.
  2. Discover clusters in the subset.
  3. Load the remaining objects into RAM by portions.
     Classify each object = identify to which of the discovered clusters the object belongs.
     Store the result of the classification in the database.
  4. Take the objects that remained unclassified and apply steps 1 to 3 to them.
     Repeat the procedure until no meaningful new clusters can be discovered.

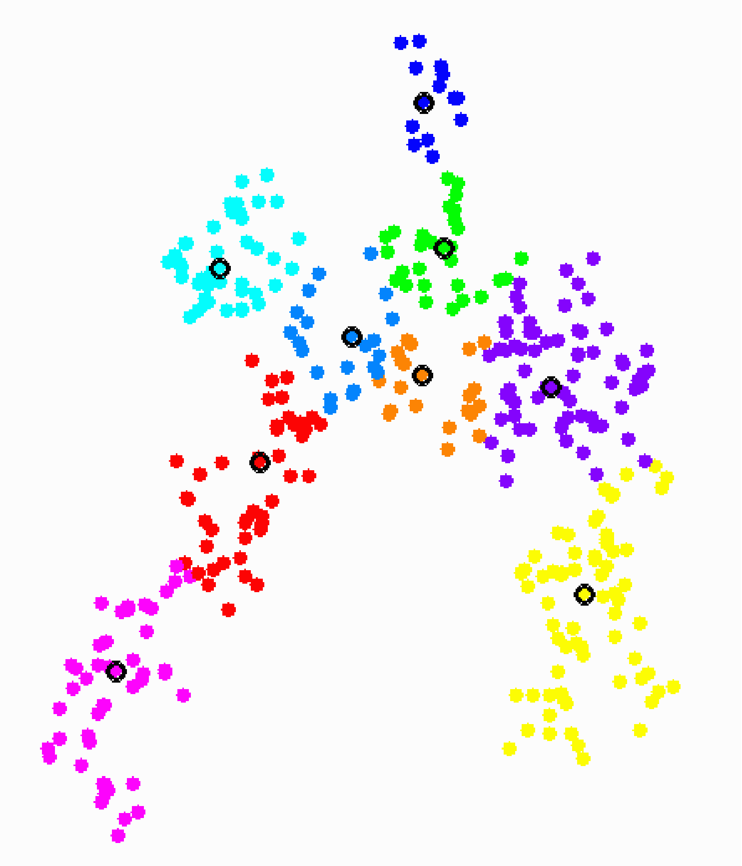
- Question: how to identify the cluster where an object belongs?

Classifier, the main idea

- From each cluster C_i select one or more representative objects (prototypes) and respective distance thresholds:
  $\{ (pt_1, d_1), \ldots, (pt_n, d_n) \}$ such that $\forall o \in C_i\ \exists k, 1 \leq k \leq n$: distance $(o, pt_k) < d_k$
  - The set of all cluster prototypes with the respective distance thresholds defines the classifier
- A new object o′ may be ascribed to the cluster if the same condition holds for it.
  $\Rightarrow$ For each object from a large database:
  - measure the distances to all prototypes;
  - take the closest prototype among those with the distances below the thresholds and ascribe the object to the respective cluster;
  - if no such prototypes found, label the object as unclassified.
- To select prototypes:
  - Divide the cluster into “round” subclusters
  - Take the medoid of each subcluster as one of the prototypes
  - Take the maximum of the distances from the subcluster medoid to the subcluster members as the distance threshold for this prototype

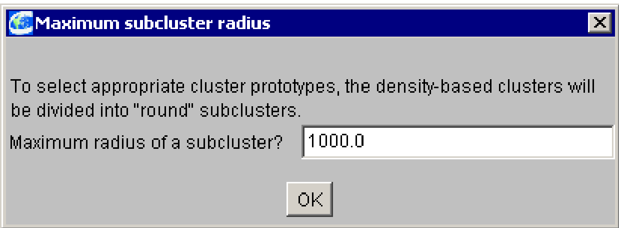
Dividing a cluster into round sub-clusters: an illustration using points

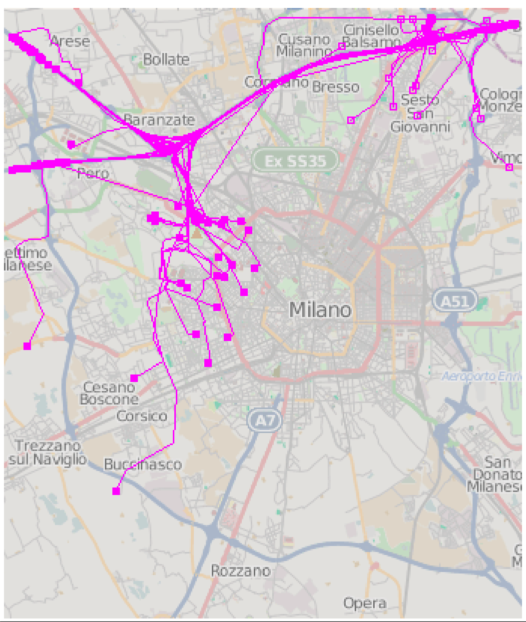




This can be done by a variant of the K-medoids clustering algorithm where the desired maximum radius of a subcluster is a parameter.

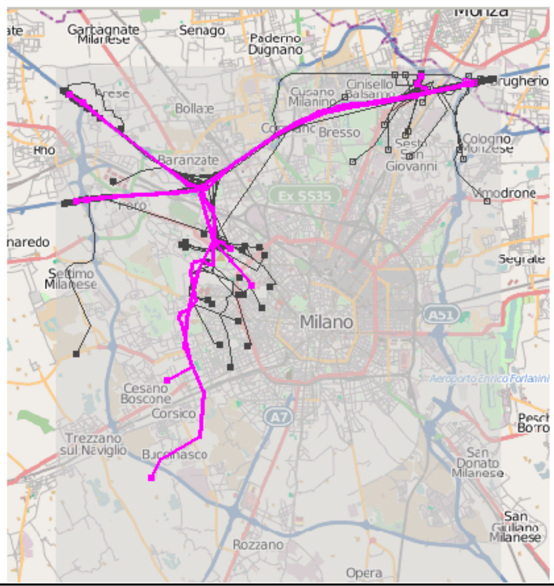
Division of a cluster of trajectories into “round” subclusters

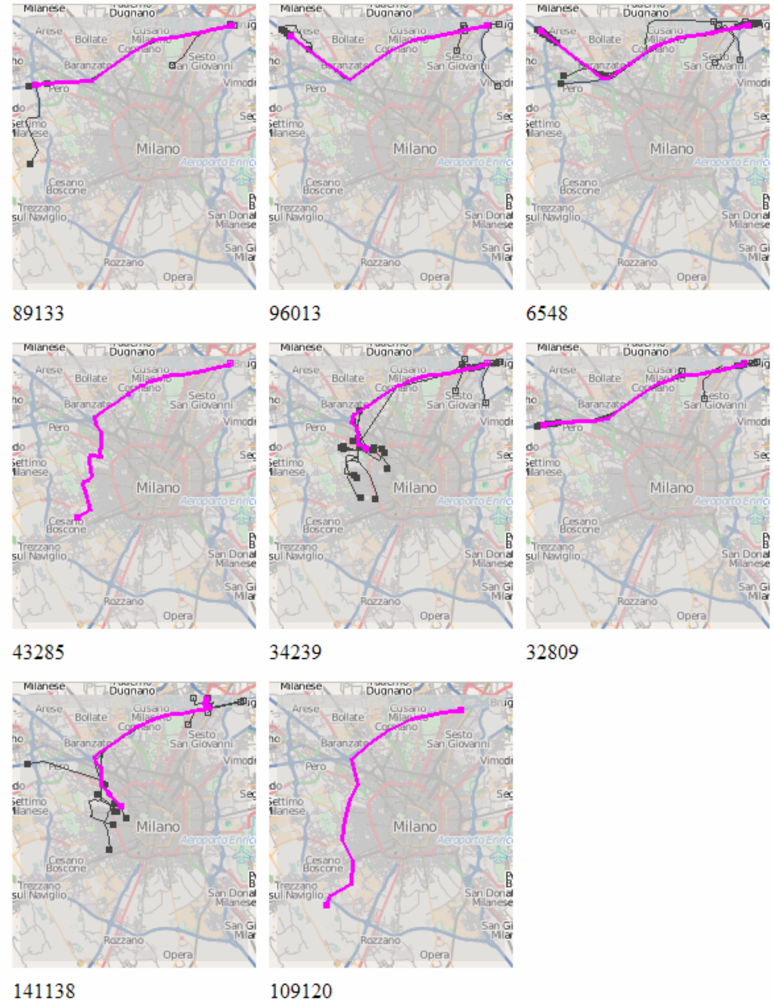




25.09.2009 11:05:24 - Cluster 7

prototype ID	Distance threshold	Original subcluster size	N neighbours found in the test	Mean distance to the original neigbours	Mean distance to the found neigbours
89133	438.2	4	0	240.5	0
96013	200.0	8	0	96.9	0
6548	526.5	29	0	161.1	0
43285	200.0	1	0	0.0	0
34239	414.3	19	0	186.7	0
32809	368.2	15	0	121.2	0
141138	485.0	10	0	271.3	0
109120	200.0	1	0	0.0	0

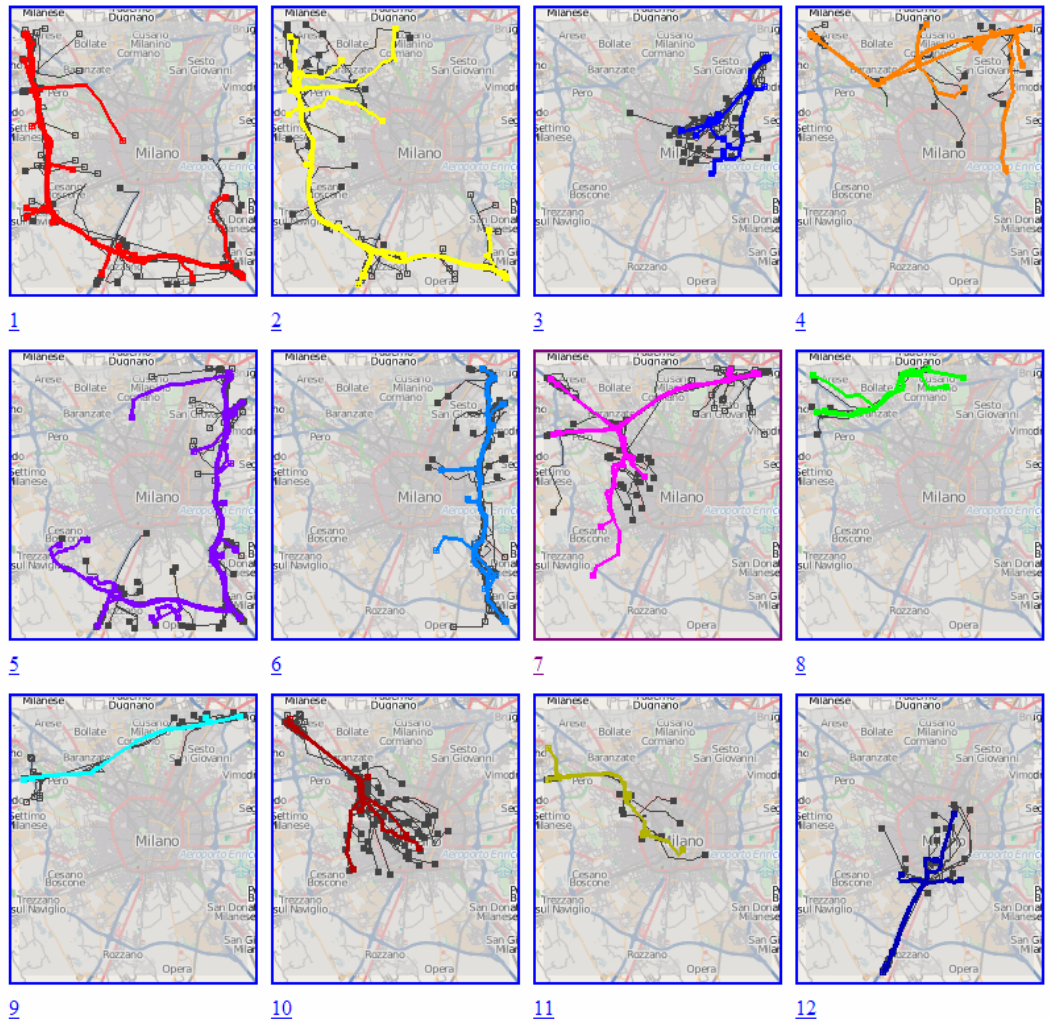




All clusters have been automatically divided; some prototypes and respective distance thresholds have been chosen

25.09.2009 11:05:24 - All clusters

cluster ID	N sub-clusters	max distance threshold	N original member
1	12	545.0	86
2	11	571.2	66
3	8	631.5	47
4	6	422.1	40
5	12	874.6	75
6	7	945.6	50
7	8	526.5	87
8	3	291.4	8
9	2	269.3	31
10	6	938.9	74
11	3	714.7	13
12	5	787.9	34
13	8	694.0	47
14	1	465.6	12
15	3	489.7	18
16	1	411.0	7
17	4	536.8	19
18	1	465.5	8
19	4	783.1	20
20	3	472.9	10
21	1	819.1	15
22	1	341.3	8
23	1	315.9	5
24	4	308.4	13
25	3	593.5	10
26	1	200.0	8
27	1	463.9	3

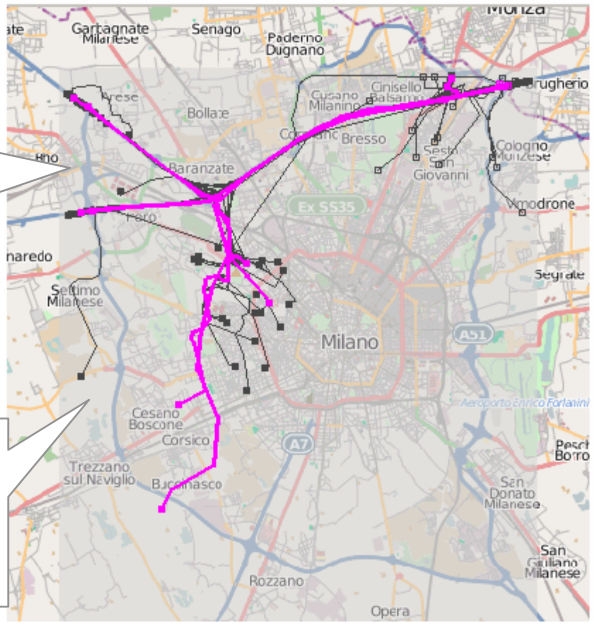


Is this all that is needed for a good analysis?

To obtain meaningful results, the analyst needs to review and, possibly, edit the classifier

25.09.2009 11:05:24 - Cluster 7

prototype ID	Distance threshold	Original subcluster size	N neighbours found in the test	Mean distance to the original neigbours	Mean distance to the found neigbours
89133	438.2	4	0	240.5	0
96013	200.0	8	0	96.9	0
6548	526.5	29	0	161.1	0
43285	200.0	1	0	0.0	0
34239	414.3	19	0	186.7	0
32809	368.2	15	0	121.2	0
141138	485.0	10	0	271.3	0
109120	200.0	1	0	0.0	0



89133 96013 6548

43285 34239 32809

141138 109120

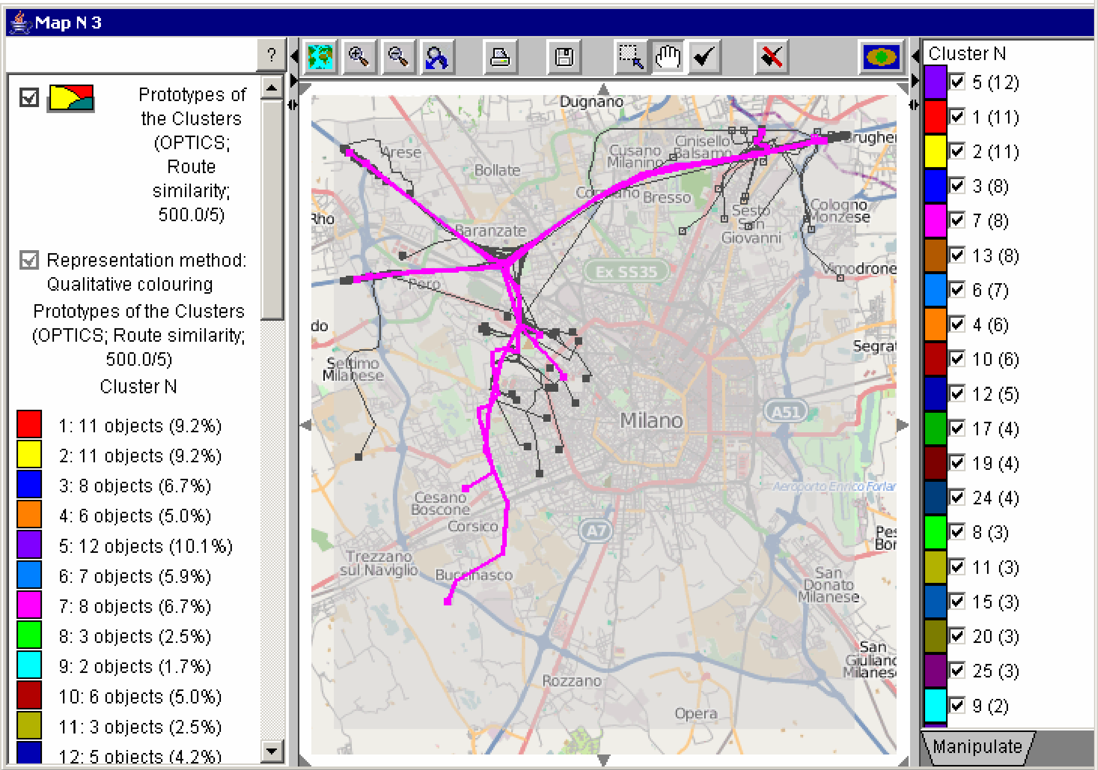
Should I keep the three branches in one cluster?

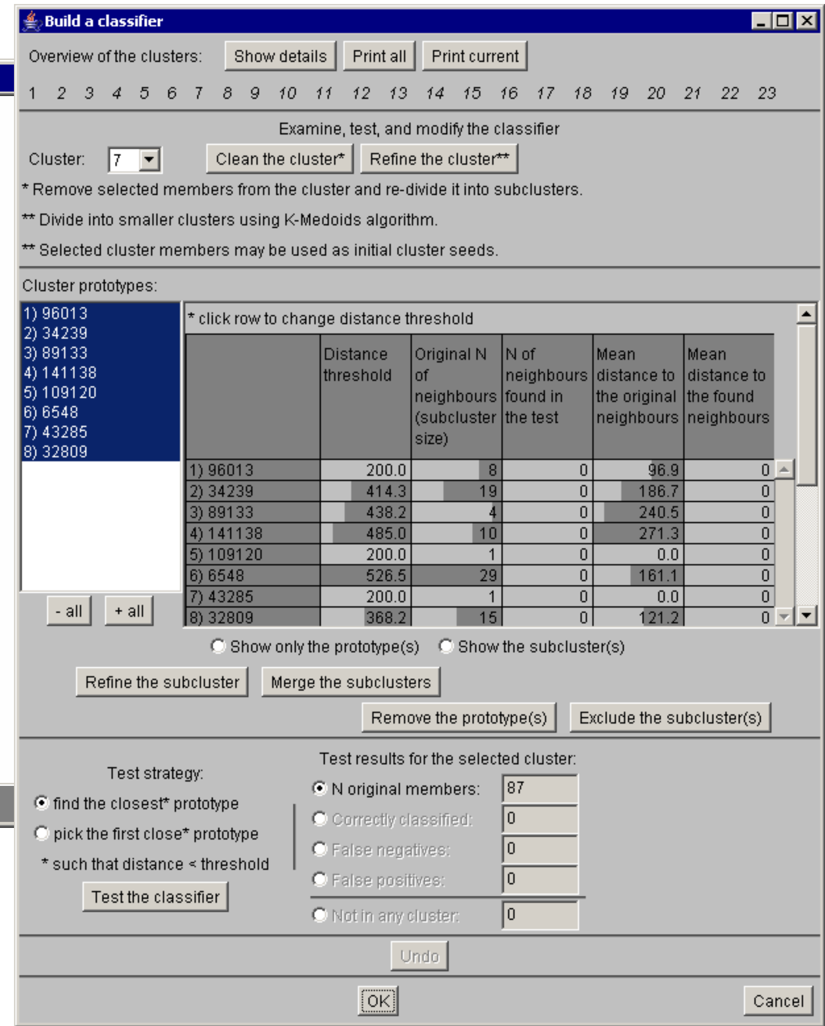
Or should I divide the cluster into two or three clusters?

Some trajectories are not very similar to the others. Should such trajectories be in the cluster?

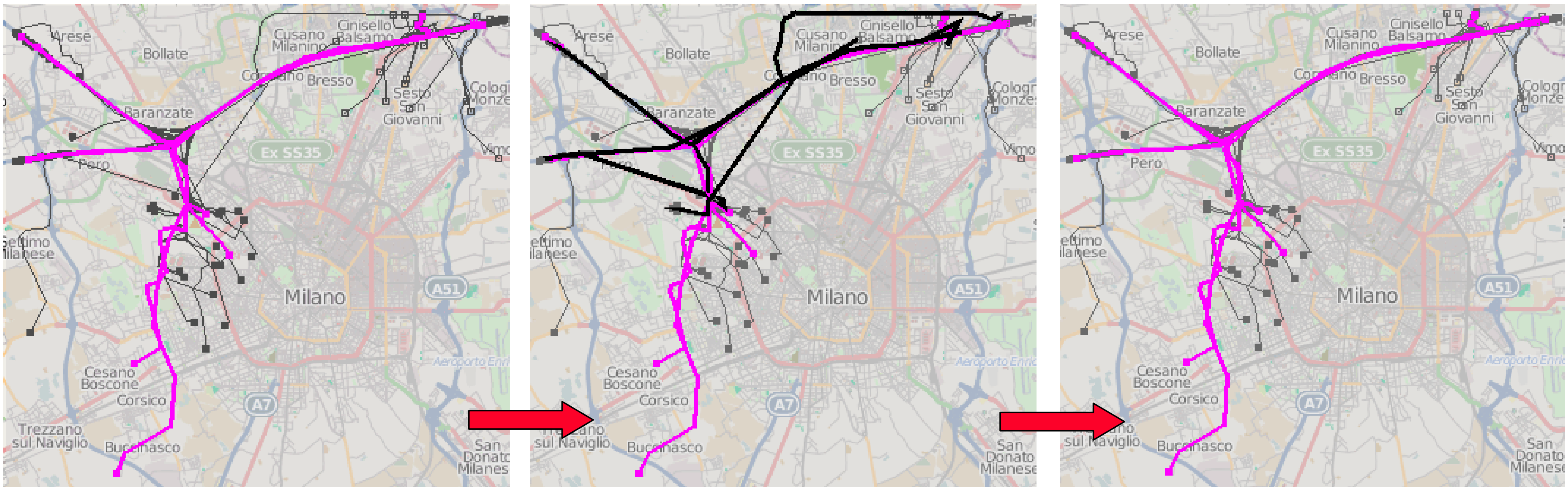
Is it good to have this prototype? This is not a core trajectory of the cluster.

Interactive editing of the classifier

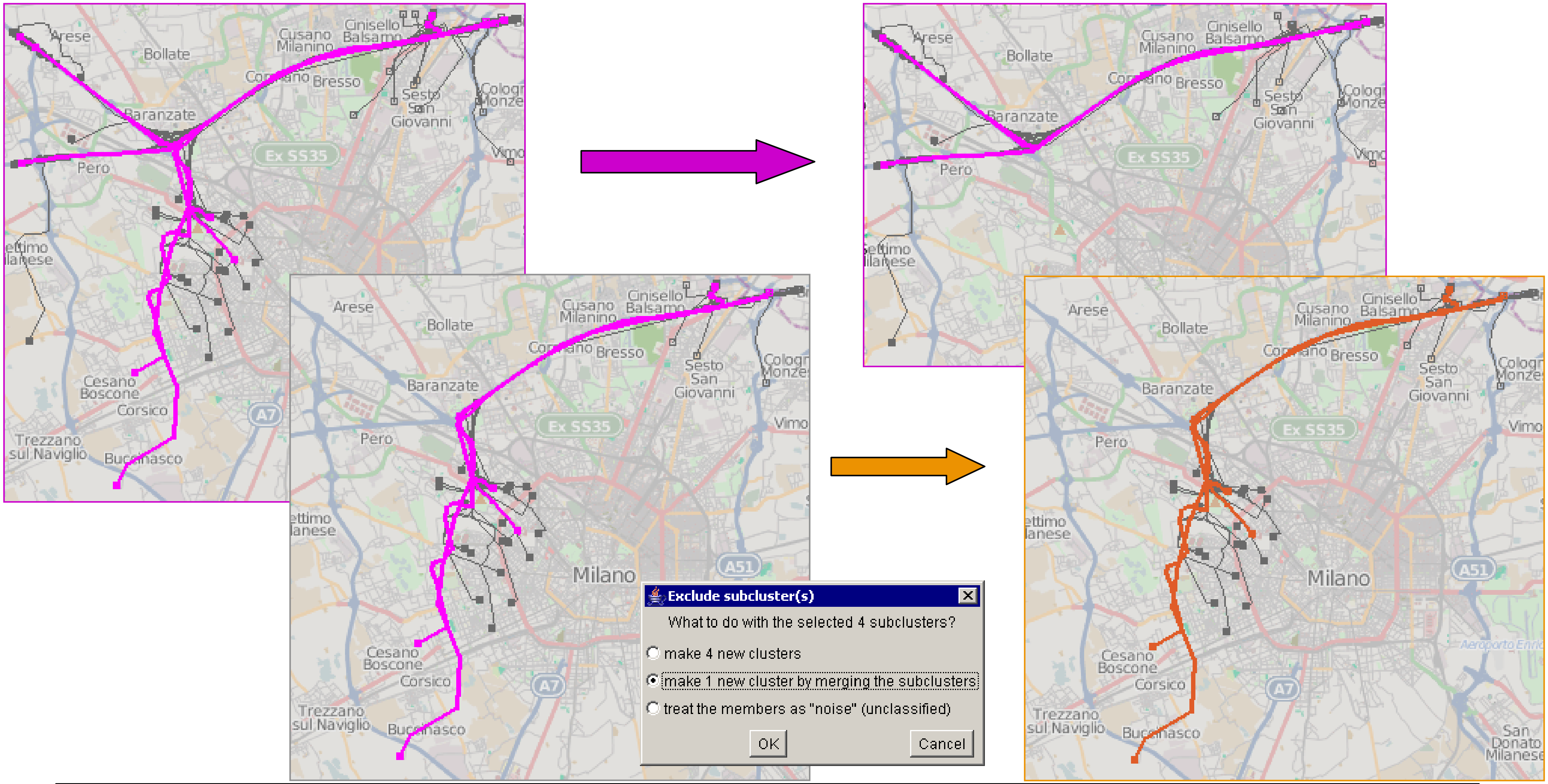




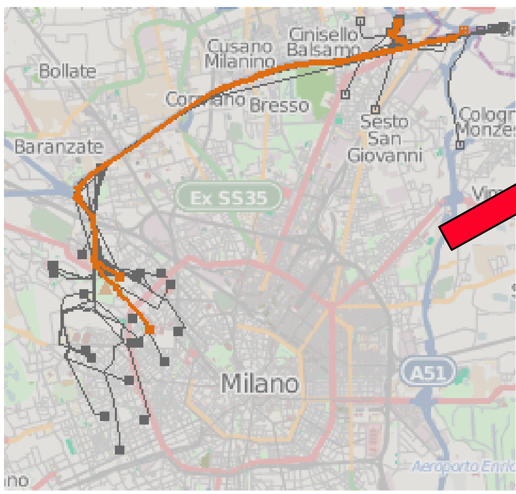
Example of interactive editing: remove selected objects from a cluster

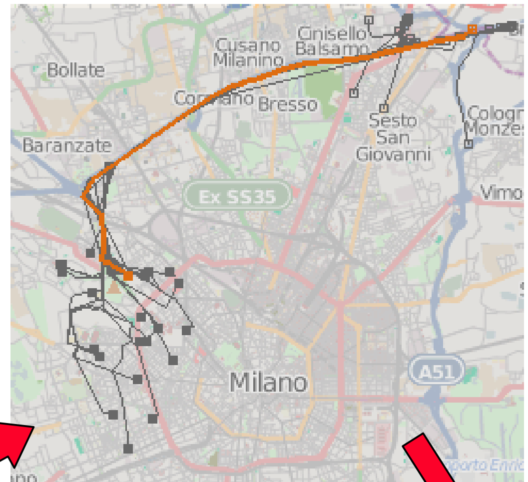


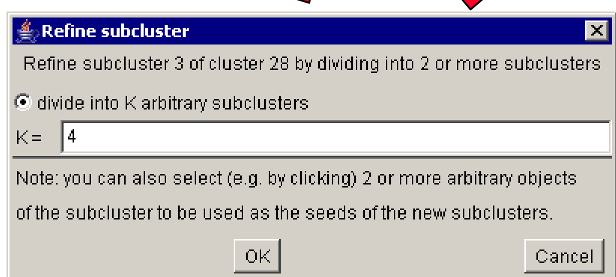
Example of interactive editing: move selected subclusters to a new cluster

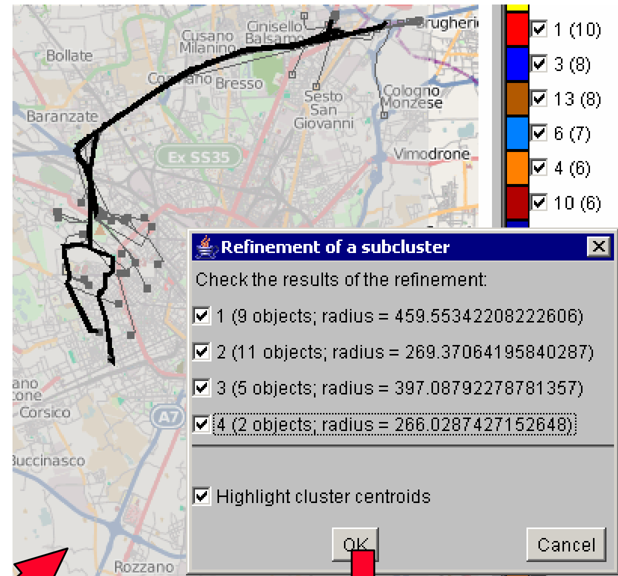


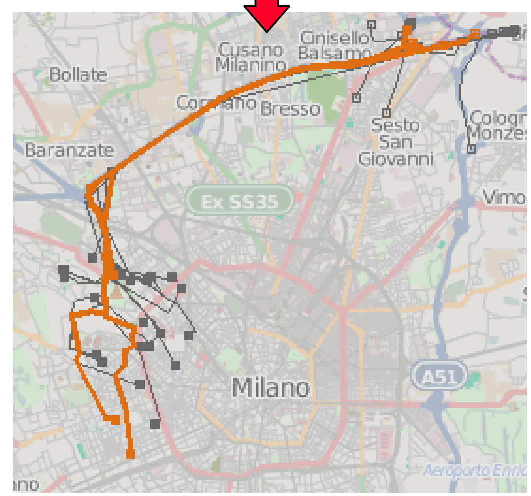
Example of interactive editing: merge and refine subclusters











The finer the division and the smaller the distance thresholds, the fewer “false positives” will be later ascribed to the cluster.

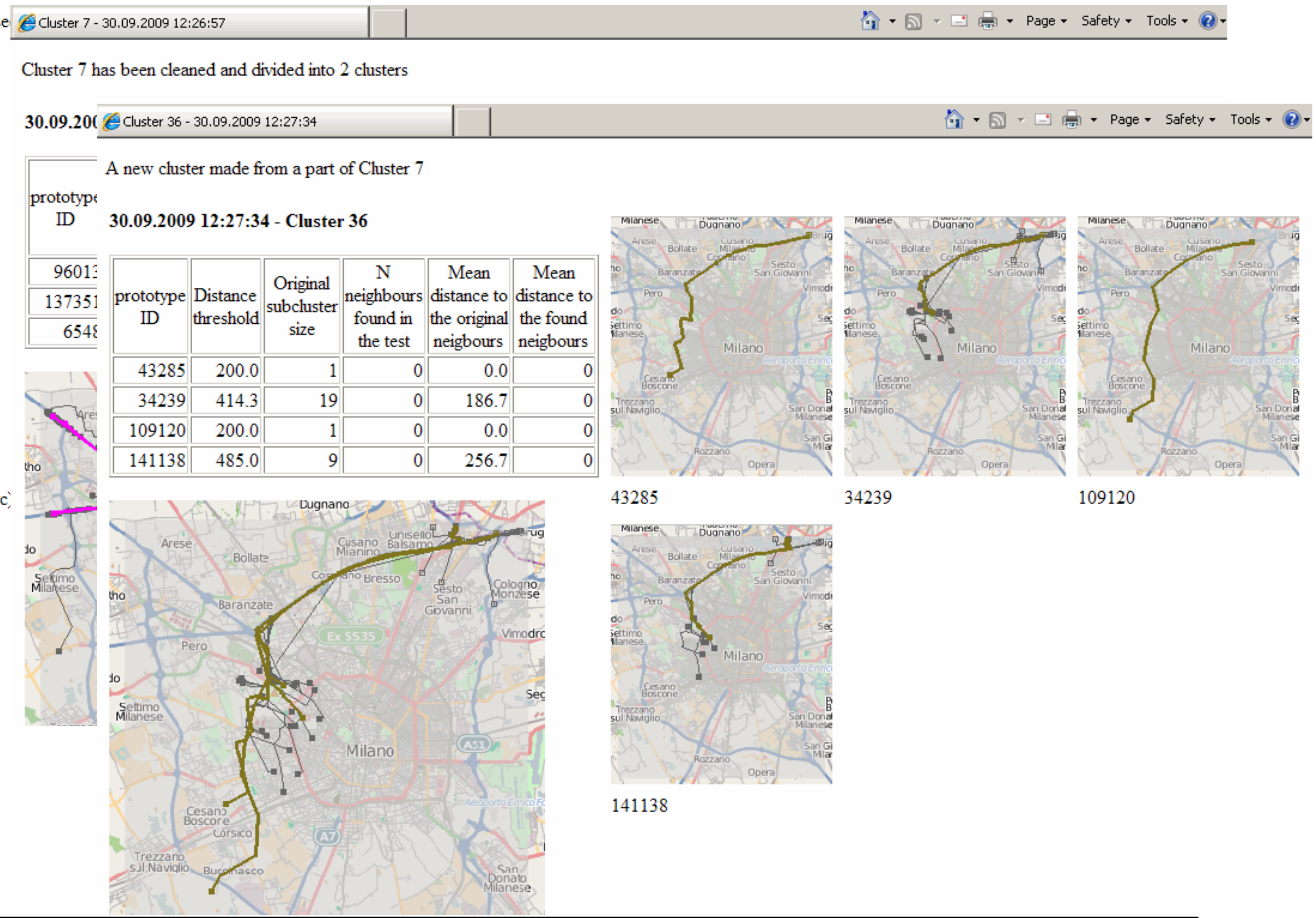
Action logging and report generation

30.09.2009 12:26:22 - 30.09.2009 12:26:24 (duration 2328.0mse
Exclude 3 members from cluster 7
Excluded members = [56267, 4332, 47256]
N remaining members = 84
Max subcluster radius = 1000.0
N subclusters = 7
New prototype 1 id = 43285
New subcluster 1 size = 1
New prototype 2 id = 34239
New subcluster 2 size = 19
New prototype 3 id = 109120
New subcluster 3 size = 1
New prototype 4 id = 141138
New subcluster 4 size = 9
New prototype 5 id = 96013
New subcluster 5 size = 8
New prototype 6 id = 137351
New subcluster 6 size = 19
New prototype 7 id = 6548
New subcluster 7 size = 27

30.09.2009 12:26:51 - 30.09.2009 12:26:51 (duration 625.0msec)
Exclude 4 subclusters from cluster 7
Make new cluster(s) = false
Merge subclusters in one new cluster = true
Excluded subcluster 1 index = 0
Excluded subcluster 1 centroid = 43285
Excluded subcluster 2 index = 1
Excluded subcluster 2 centroid = 34239
Excluded subcluster 3 index = 2
Excluded subcluster 3 centroid = 109120
Excluded subcluster 4 index = 3
Excluded subcluster 4 centroid = 141138
Added cluster = 36

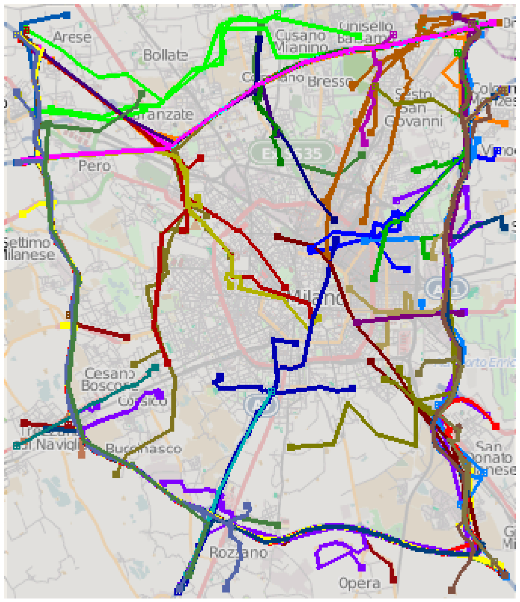
30.09.2009 12:26:57 Classifier for cluster 7

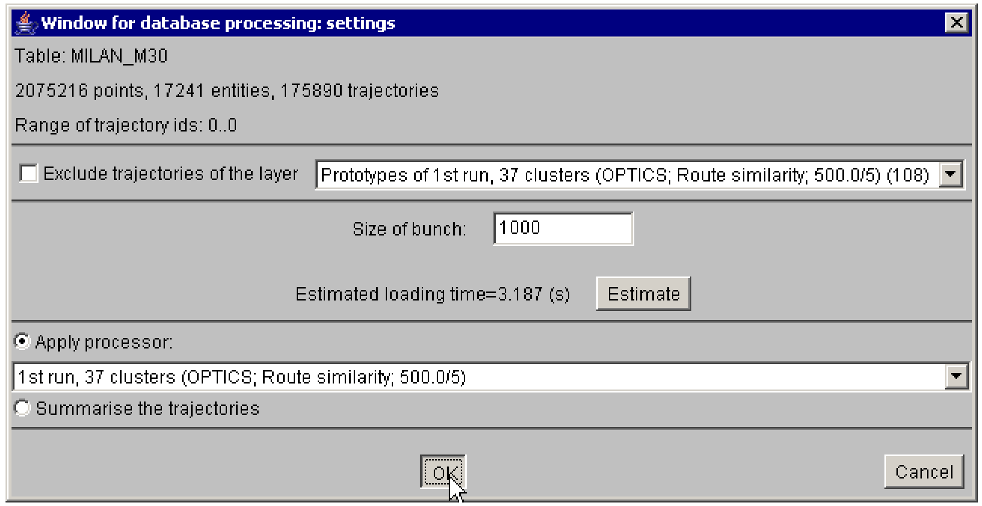
30.09.2009 12:27:34 Classifier for cluster 36

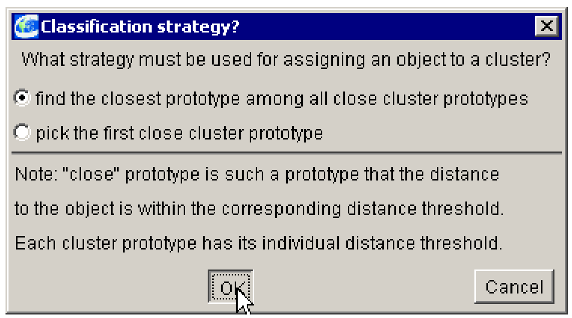


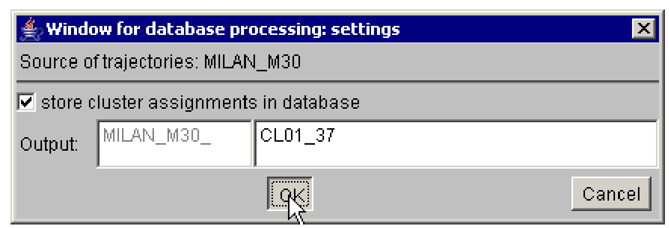
prototype ID	Distance threshold	Original subcluster size	N neighbours found in the test	Mean distance to the original neigbours	Mean distance to the found neigbours
43285	200.0	1	0	0.0	0
34239	414.3	19	0	186.7	0
109120	200.0	1	0	0.0	0
141138	485.0	9	0	256.7	0

Application of the classifier to the whole database

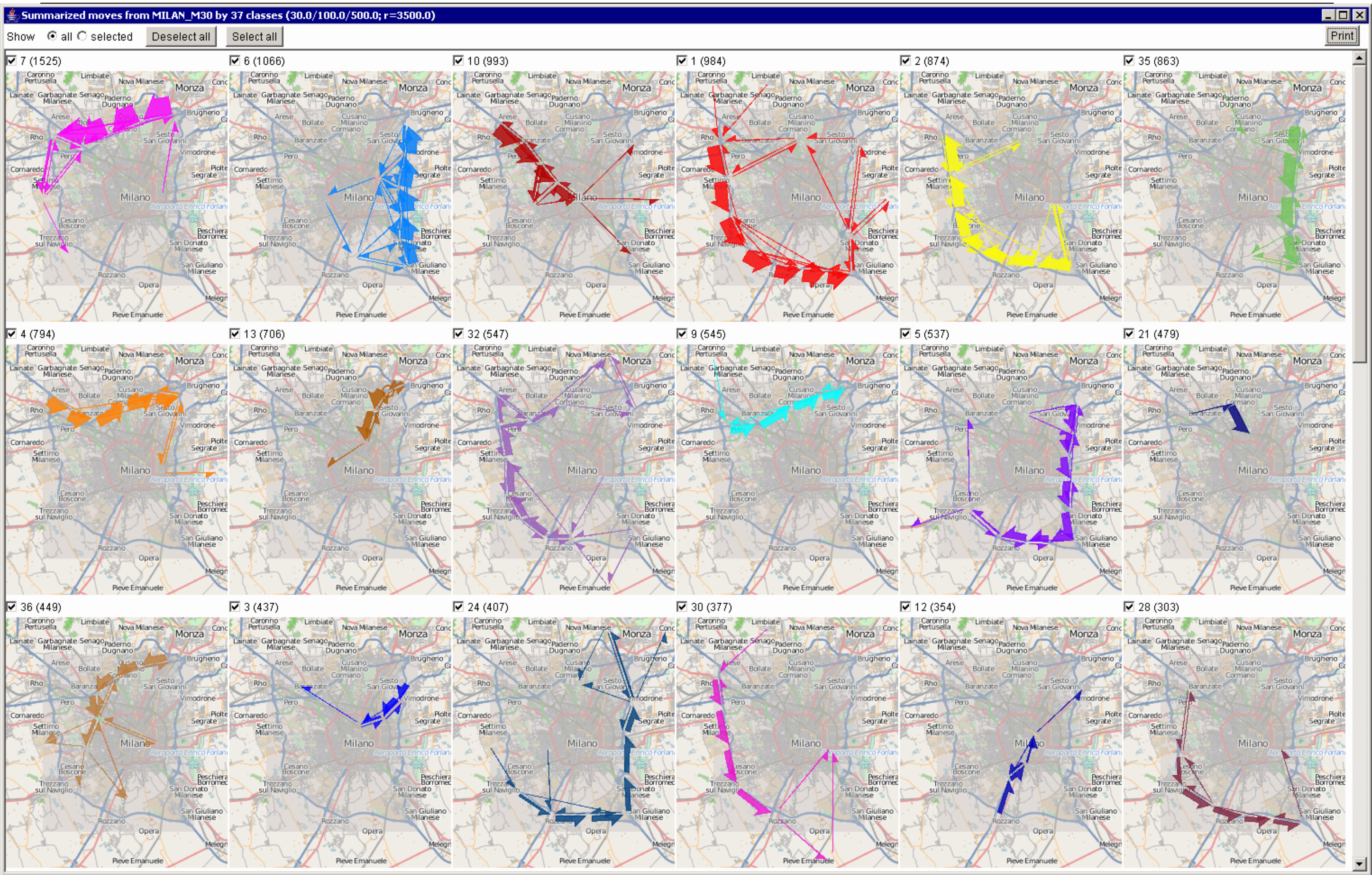




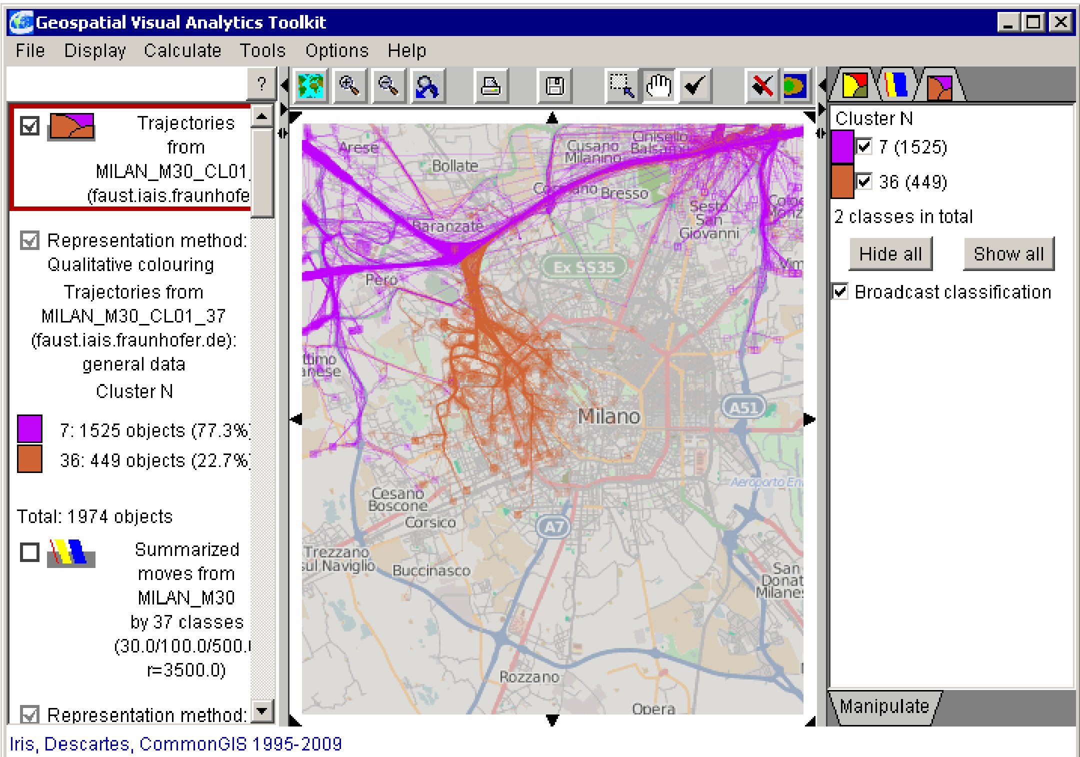


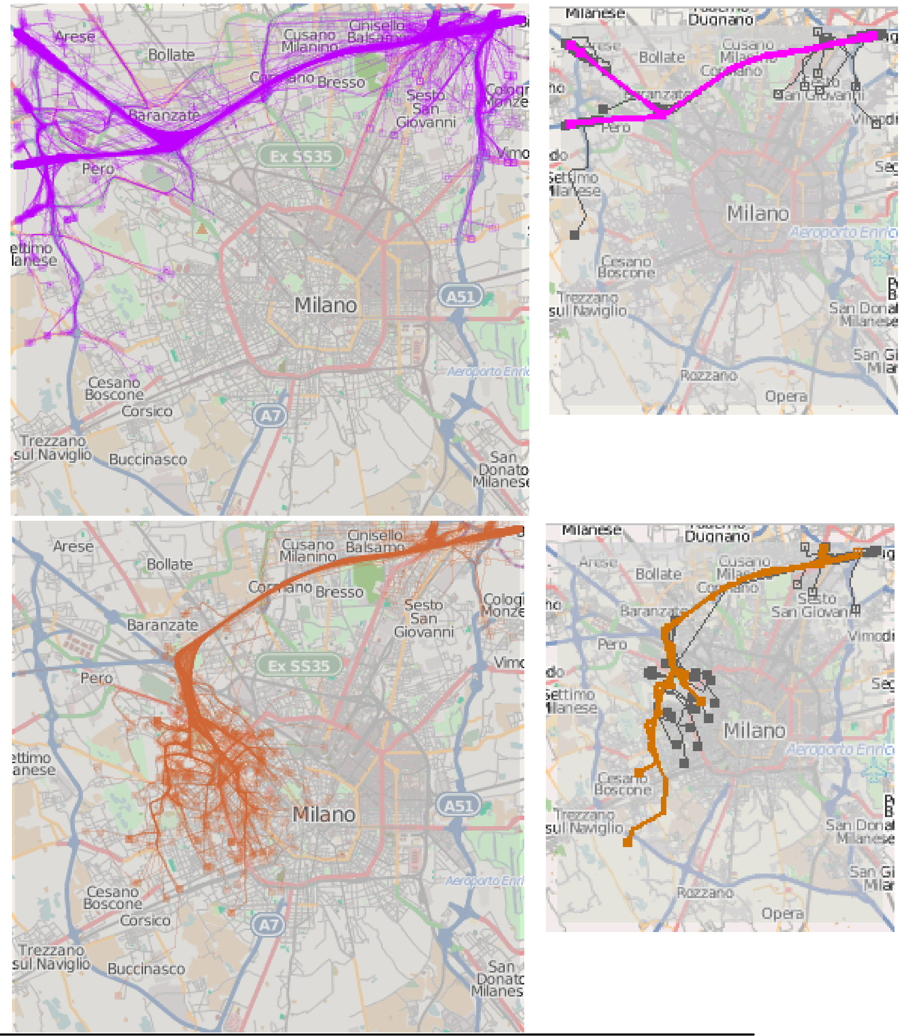


Now we see how frequent these routes are in the whole database



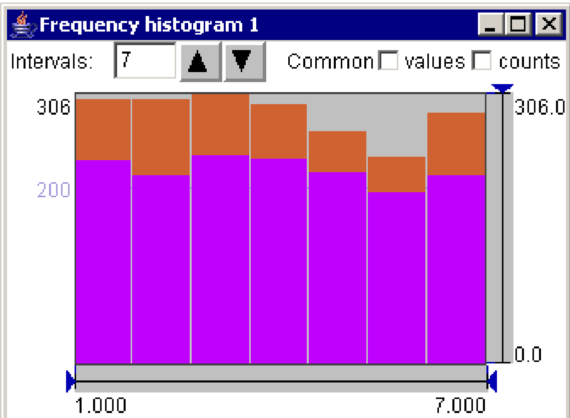
Selected clusters can be loaded from the database and explored in detail



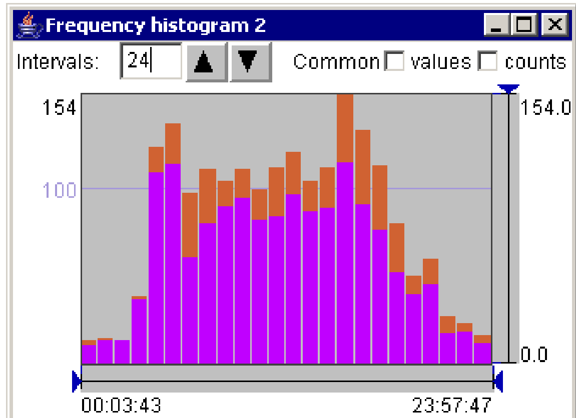


Investigation of the properties of the trajectories in the selected clusters

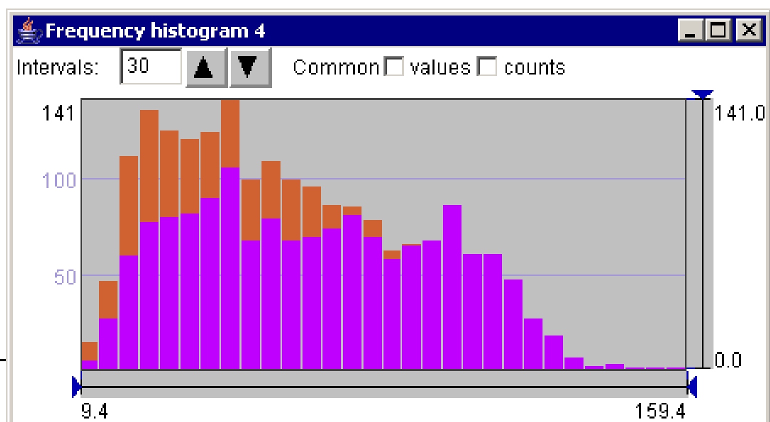
Frequencies by the days of the week

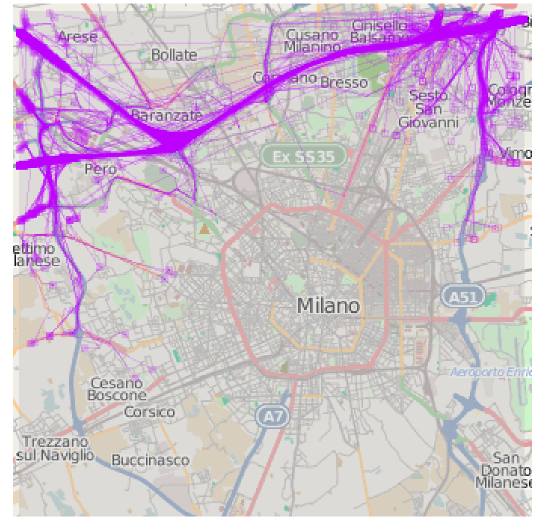


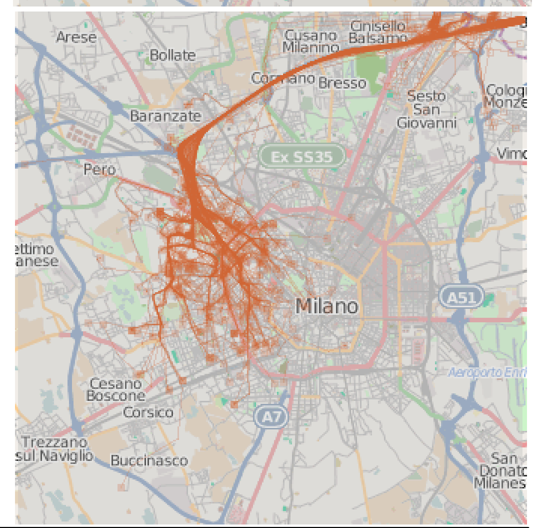
Frequencies by the times of the day



The frequency distribution of the average speeds



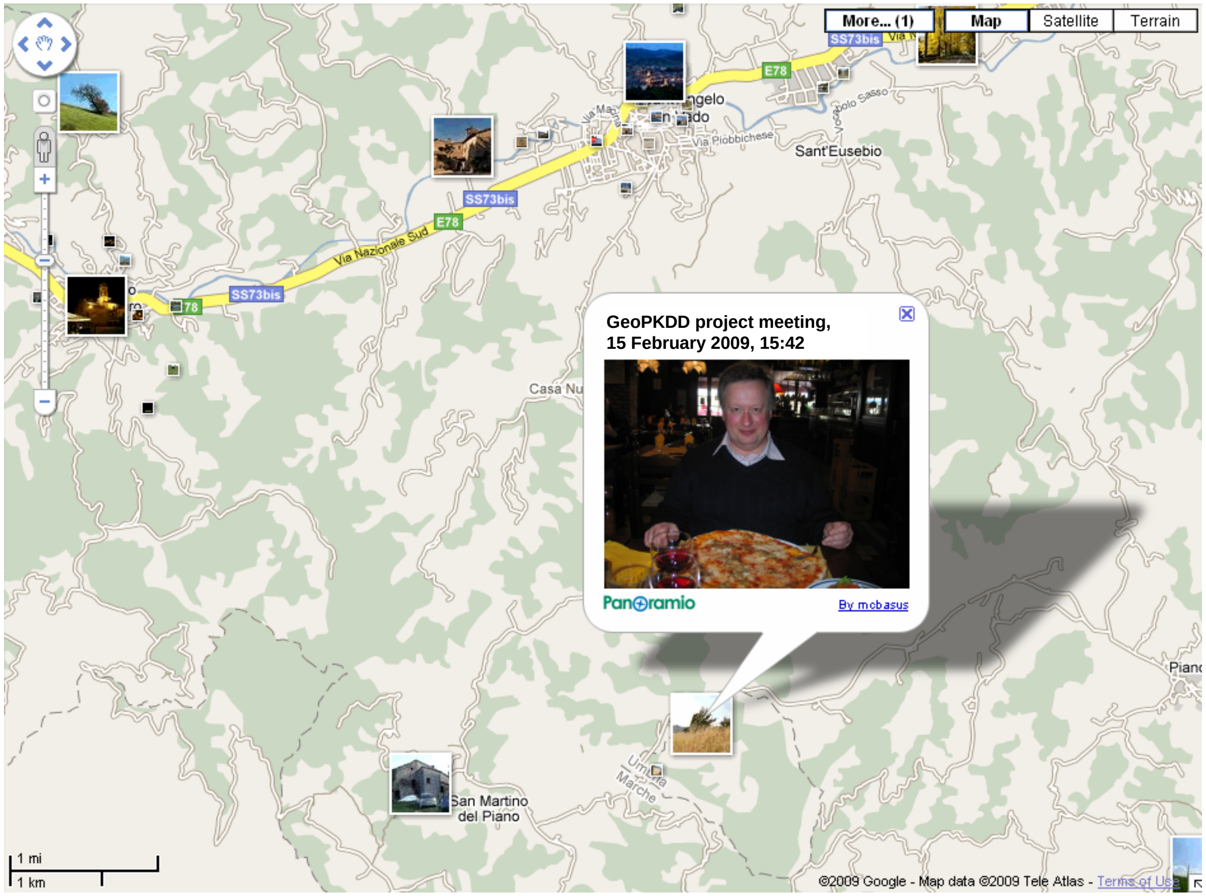




Important properties of the clustering+classification method

- Scalability to very large datasets
- Tangible analysis result: the classifier
  - Describes the clusters without listing all their members
  - May be applied to other data sets and to data streams
- It is possible to adapt this method to other types of data and problems
- Synergistic cooperation between human and computer
  - Computer finds clusters, summarizes, visualizes, makes draft classifiers
  - Human not only takes the results but also directs computer’s work:
    - selects appropriate data subsets, finds suitable clustering parameters;
    - edits the suggested draft classifiers (involves interpretation, evaluation, adaptation to the goals of analysis)

Analysis of temporally sparse movement data (discontinuous trajectories), e.g. positions of photos in Panoramio



Analysis of temporally sparse movement data (discontinuous trajectories), e.g. positions of photos in Panoramio

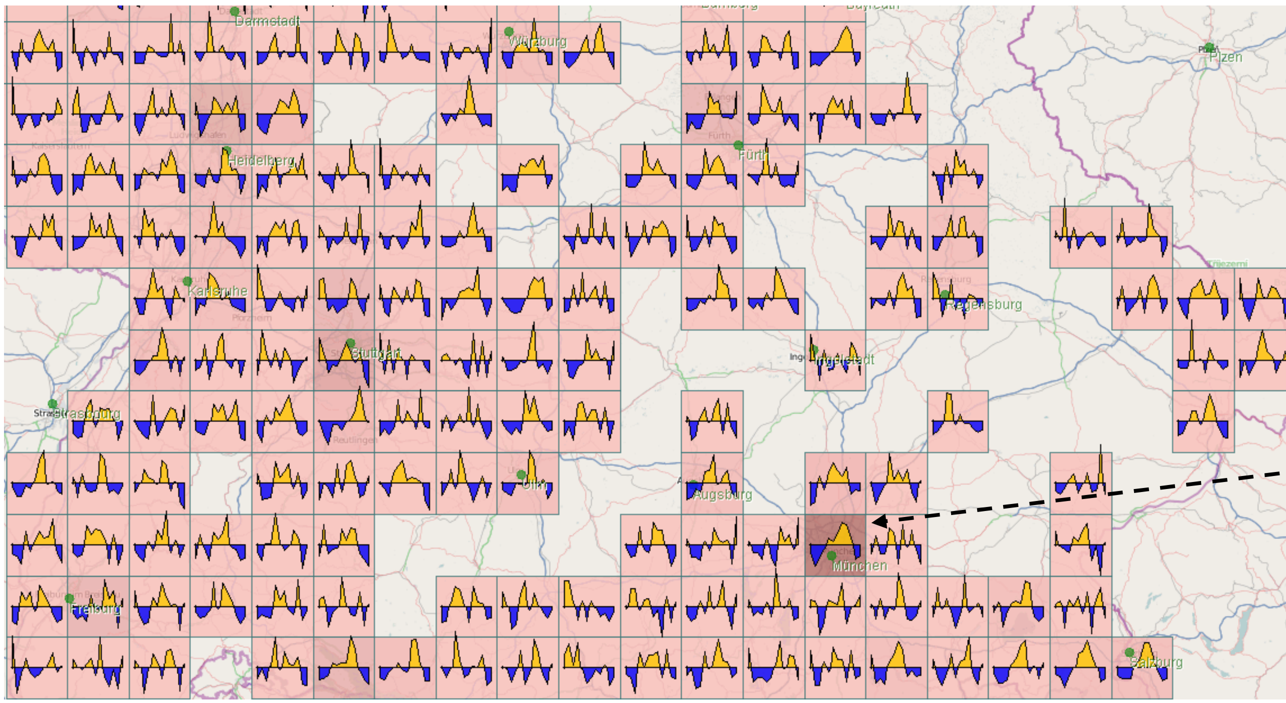
universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

IAIS
Fraunhofer Institut Intelligente Analyse- und Informationssysteme

Possible tasks and methods in analysis of temporally sparse movement data

Focus / Object	Space and Place	Moving agents (people)
Events (places and times of taking the photos)	**Tasks:** Find places of interest, explore temporal patterns of visits to these places **Methods**: Spatial clustering of events, spatio-temporal aggregation of events	**Tasks**: Discover meetings of people, concentrations of people in same place and time **Methods**: spatio-temporal clustering of events, detection of interactions (cases of proximity in space and time)
Trajectories (sequences of positions of the photographers)	**Tasks:** Investigate flows between places **Methods:** Aggregation of moves + flow maps	**Tasks**: Discover joint travels, frequently taken routes **Method**: clustering of trajectories, detection of interactions extended in space and time

Spatio-temporal aggregation of events



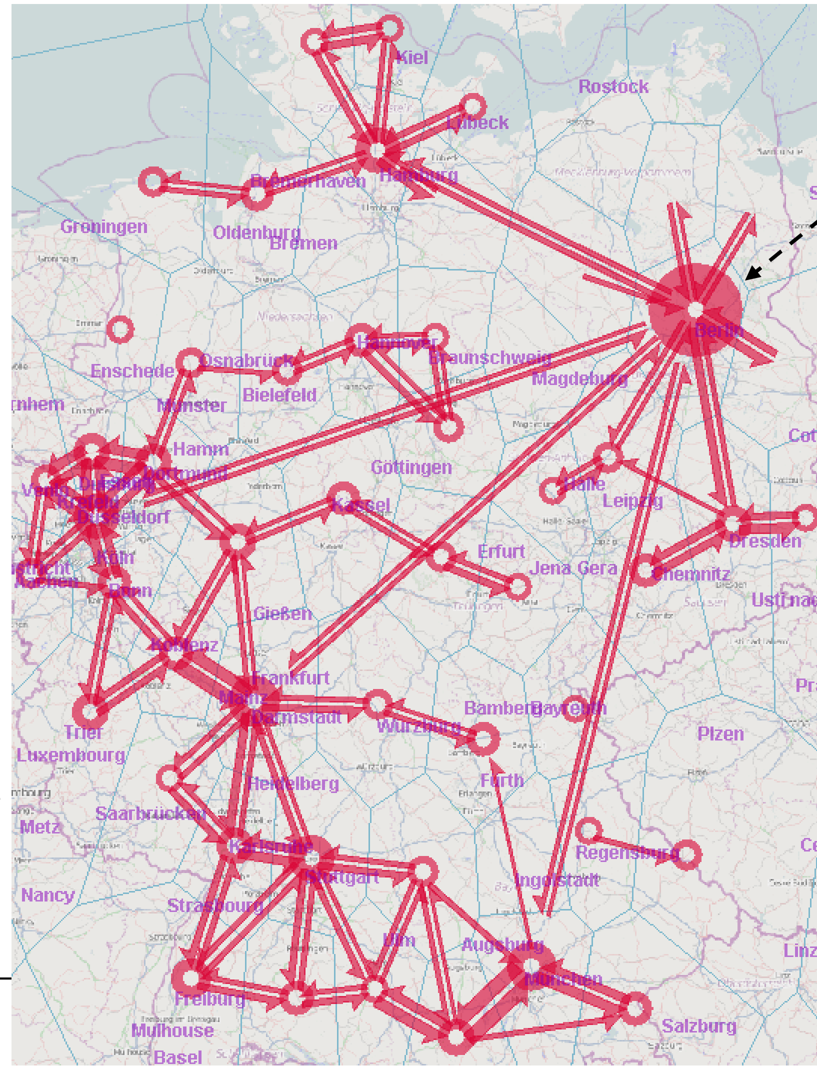
The data have been aggregated spatially by grid cells and temporally by months, irrespective of the years. The shading portrays the total number of different people who made photos in the cells. The cells visited by less than 100 people are hidden. The diagrams represent the yearly variation of the number of photos taken in the cells. The horizontal axis represents 12 months from January to December. The counts are transformed into normalized differences from the local mean values. The map fragment represents the south of Germany.

We can observe a seasonal variation of the frequency of taking photos in most of the places.

The area of Munich is the most frequently visited place on the south of Germany. The highest frequency of taking photos is in autumn, especially in October.

Analysis of movements (flows) at different spatial scales

Most movements occur within areas and between neighbouring areas. However, quite frequent are also distant moves between Berlin and the other biggest cities of Germany: Hamburg, Düsseldorf, Frankfurt, and Munich.



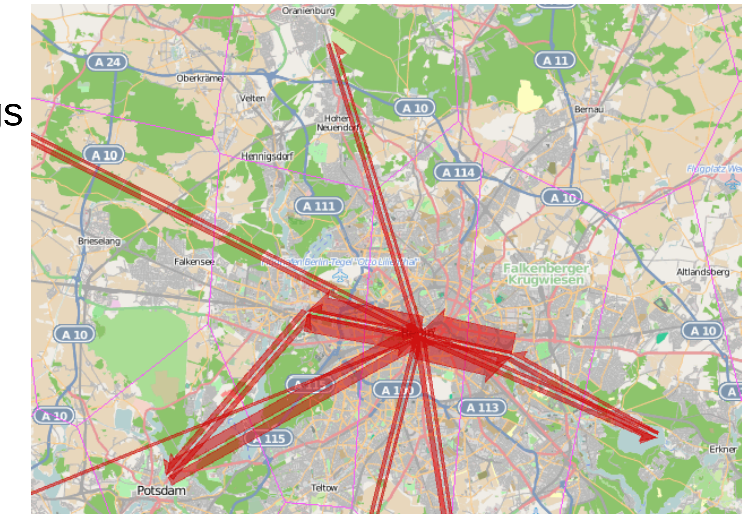
Very many trajectories of the photographers are within the area of Berlin

The data have been aggregated using the approach to movement summarisation on the basis of Voronoi polygons. The aggregate moves have been interactively filtered so that only the moves with the frequency 100 and more are visible.

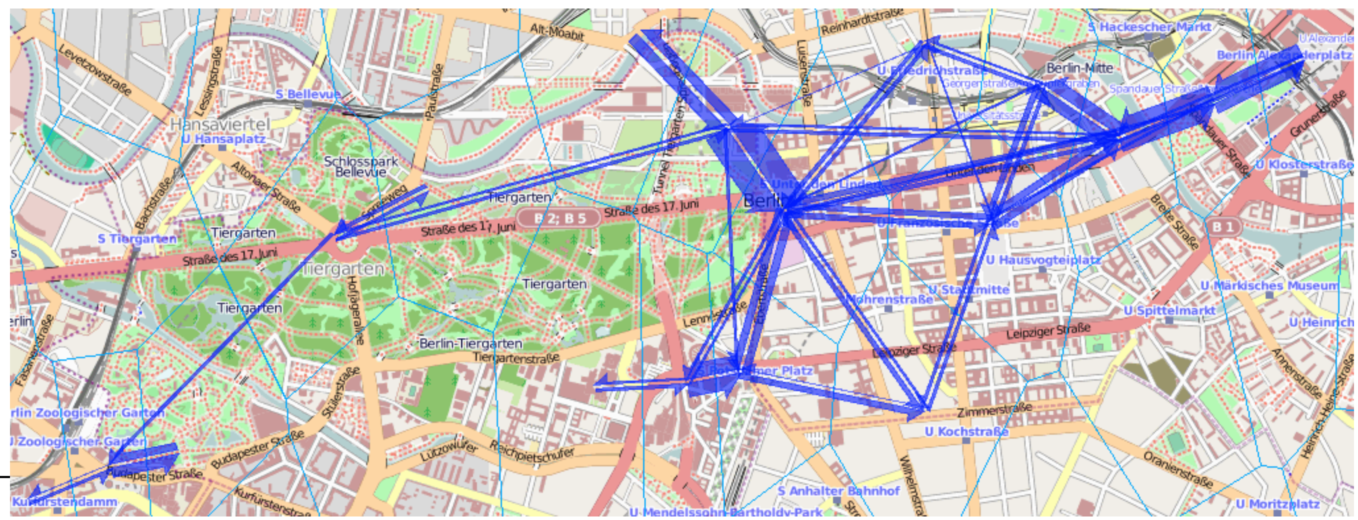
Analysis of movements at different spatial scales

Movements in the western part of Germany

Movements in area of Berlin and surroundings



Movements in Berlin



Challenging research directions

- Movement data require special analysis methods taking into account the spatial and temporal dependency between the positions
- Lack of an appropriate <u>theoretical basis</u> (usual approaches: adopting methods from geographical analysis, time series analysis, sequence mining etc.; designing ad-hoc methods for specific applications)
- <u>Temporally sparse movement data</u> are not properly addressed yet (typical assumption: data represent continuous space-time paths)
- Little has been done on <u>multi-scale (in space and time) analysis</u> of movement
- Little has been done on <u>joint analysis of movement data and multidimensional attributes</u> of the moving entities and of the environment
- Little has been done for <u>analysis of data streams</u>
- Little has been done on analysis of <u>3D movement</u> (e.g. flights, underwater movement)
- <u>Scalability to very large datasets</u>: in most cases analysis is done in RAM

Further reading

1. General framework
   - G.Andrienko, N.Andrienko, S.Wrobel
     Visual Analytics Tools for Analysis of Movement Data
     ACM SIGKDD Explorations, v.9(2), December 2007, pp.38-46
2. Progressive clustering
   - S.Rinzivillo, D.Pedreschi, M.Nanni, F.Giannotti, N.Andrienko, G.Andrienko
     Visually–driven analysis of movement data by progressive clustering
     Information Visualization, 2008, v.7 (3/4), pp. 225-239
3. Aggregation
   - G.Andrienko, N.Andrienko
     Spatio-temporal aggregation for visual analysis of movements
     IEEE Visual Analytics Science and Technology (VAST 2008)
     Proceedings, IEEE Computer Society Press, 2008, pp.51-58
4. Scalable clustering
   - G.Andrienko, N.Andrienko
     Interactive Visual Clustering of Large Collections of Trajectories
     IEEE Visual Analytics Science and Technology (VAST 2009)
     Proceedings, IEEE Computer Society Press, 2009
5. Summarization and generalization
   - G.Andrienko, N.Andrienko
     Spatial Generalisation and Aggregation of Massive Movement Data
     IEEE TVCG, accepted

Visual Analytics of Movement: communities, conferences

- FET-Open GeoPKDD - Geographic Privacy-aware Knowledge Discovery and Delivery, http://www.geopkdd.eu/
- VisMaster - Visual Analytics - Mastering the Information Age (FET-Open Coordination Action), http://www.vismaster.eu
- SPP VA - Scalable Visual Analytics: Interactive Visual Analysis Systems of Complex Information Spaces (DFG Priority Research Program), http://www.visualanalytics.de/
- MODAP - Mobility, Data Mining, and Privacy (FET-Open Coordination Action), http://www.modap.org
- MOVE - Knowledge Discovery from Moving Objects (COST-Action IC0903)

- IEEE VAST & InfoVis
- EuroVAST 2010
- ACM GIS, ACM KDD…

International Cartographic Association
Commission on GeoVisualization

- http://geoanalytics.net

- “Exploring Geovisualization”,
edited by J.Dykes, M.-J.Kraak and A.MacEachren
Elsevier, 2004

- GIScience 2006 workshop outcomes:
Special issue on
“GeoVisual Analytics for Spatial Decision Support”,
including “Setting the Research Agenda” paper
Int.J.GIScience, 2007, v.21 (8)

- AGILE 2008 workshop outcomes:
Special issue on
“GeoVisualization of Dynamics, Movement and Change”
Information Visualization, 2008, v.7 (3/4)

- GeoVA(t): AGILE 2010 workshop + IJGIS spec.issue
http://geoanalytics.net/GeoVA(t)2010

- 80 members

- 150 maillist subscribers