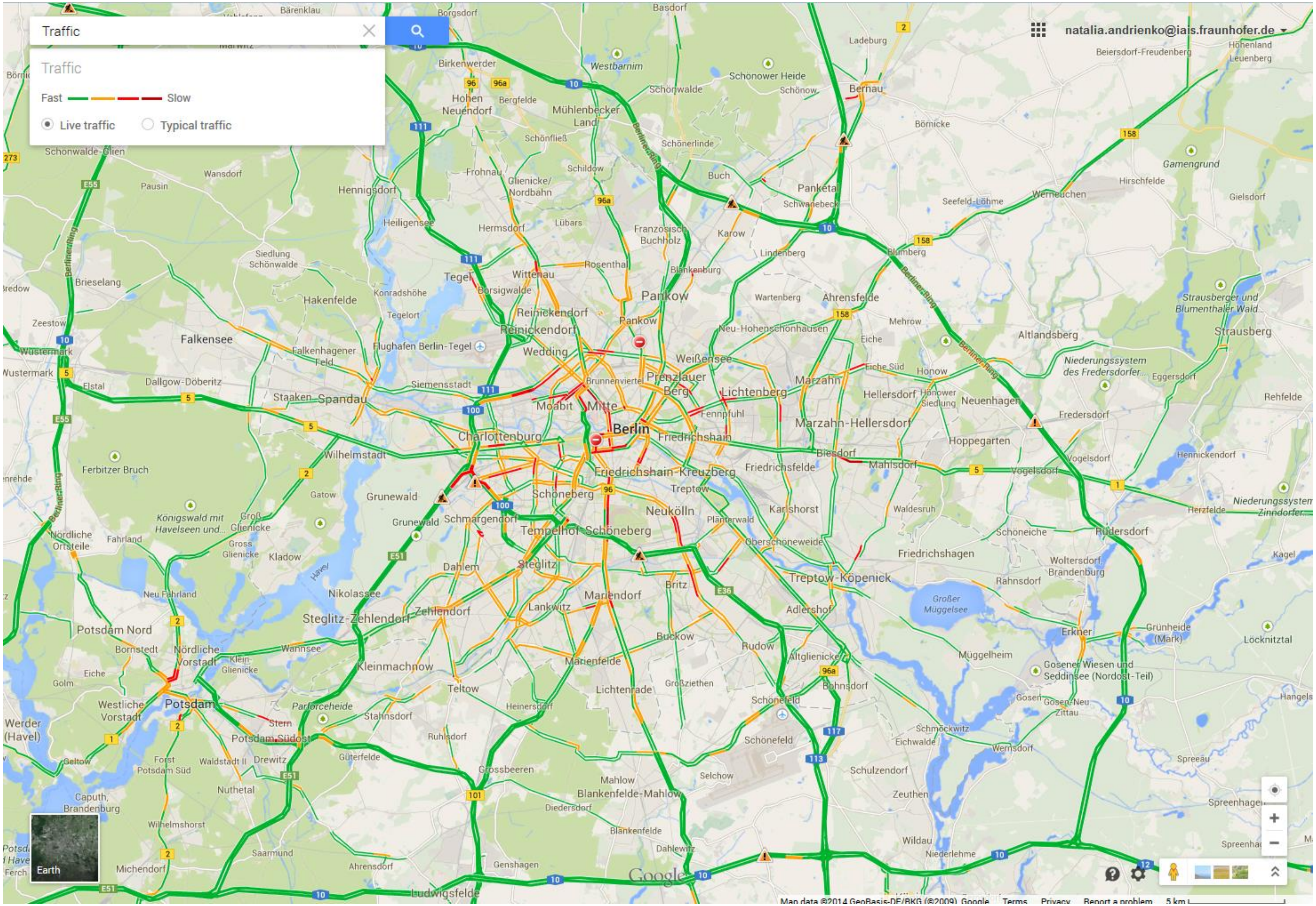
Introduction to GeoVisualization {and Visual Analytics}

Special focus: spatio-temporal and movement data

Gennady Andrienko

Natalia Andrienko

http://geoanalytics.net



CITY UNIVERSITY LONDON

Fraunhofer IAIS

{Geo} Information Visualization, basic principles

- Interactive Maps and Multiple Coordinated Views

Maps: not only for orientation!

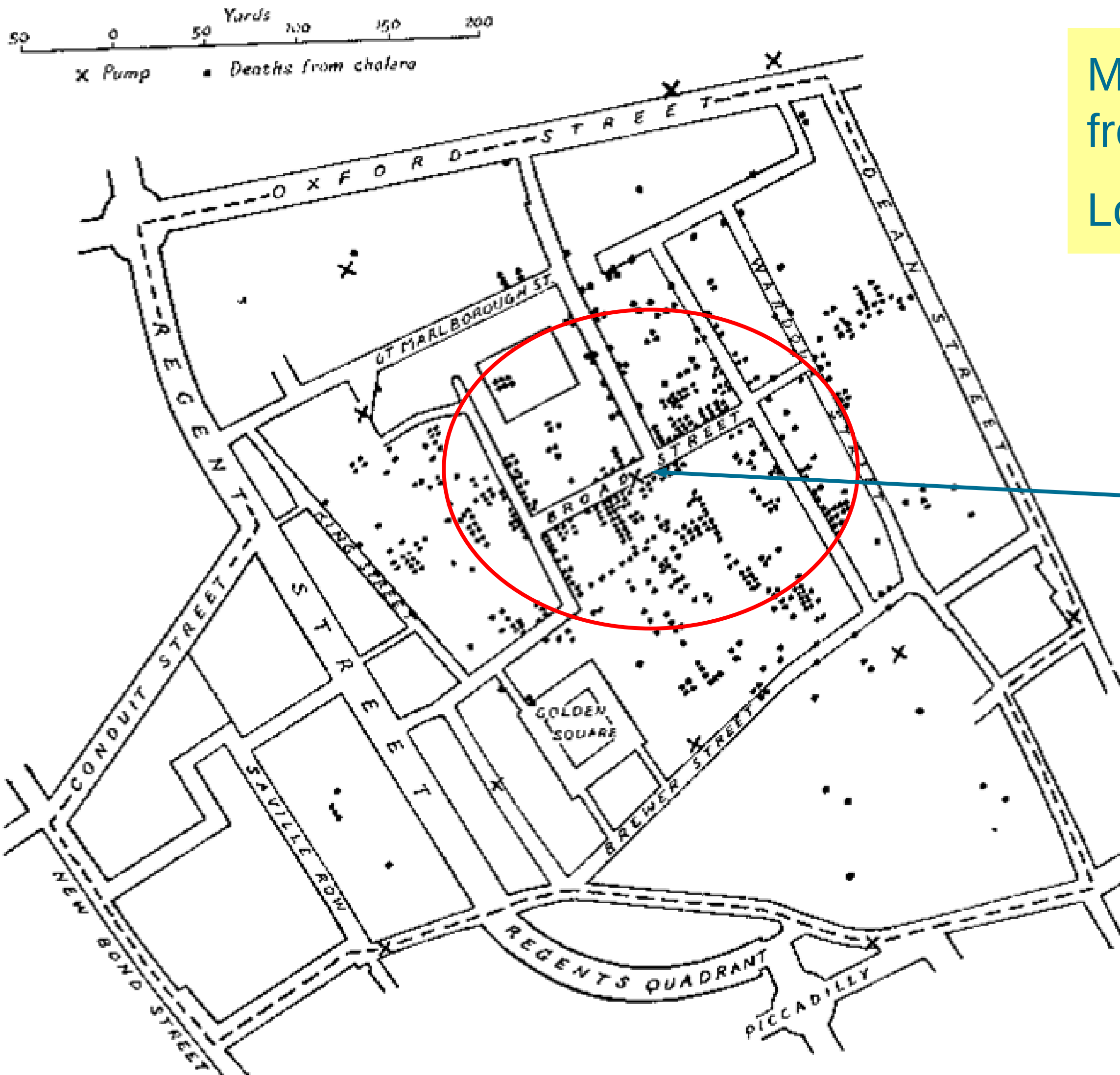
People live in geographical space.
Most of people's decisions and actions depend on

- where the things are;
- how are their locations related.

Maps allow people to perceive the space beyond the directly observable extent.

A map serves as a model of reality and helps to detect patterns existing in the reality

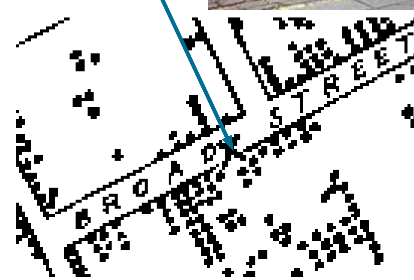
Example: Dr. John Snow's discovery



Map of locations of deaths from cholera

London, September 1854

infected water pump



CITY UNIVERSITY LONDON

Fraunhofer IAIS

Interactive maps

Interactive maps can change in response to user's actions

Many interactive maps are available on the Web, e.g. street maps, tourist maps, election maps, …

Interaction techniques are used to

- compensate for the display deficiencies, e.g. limited size (zoom and pan, showing additional information related to mouse position, …)
- increase the display expressiveness
- enable more sophisticated analyses

Typical interactive operations

Select information layers, e.g. on a tourist map: accommodation, museums, restaurants, nightlife

Select time moments or intervals in displays of time-related information, e.g. election year

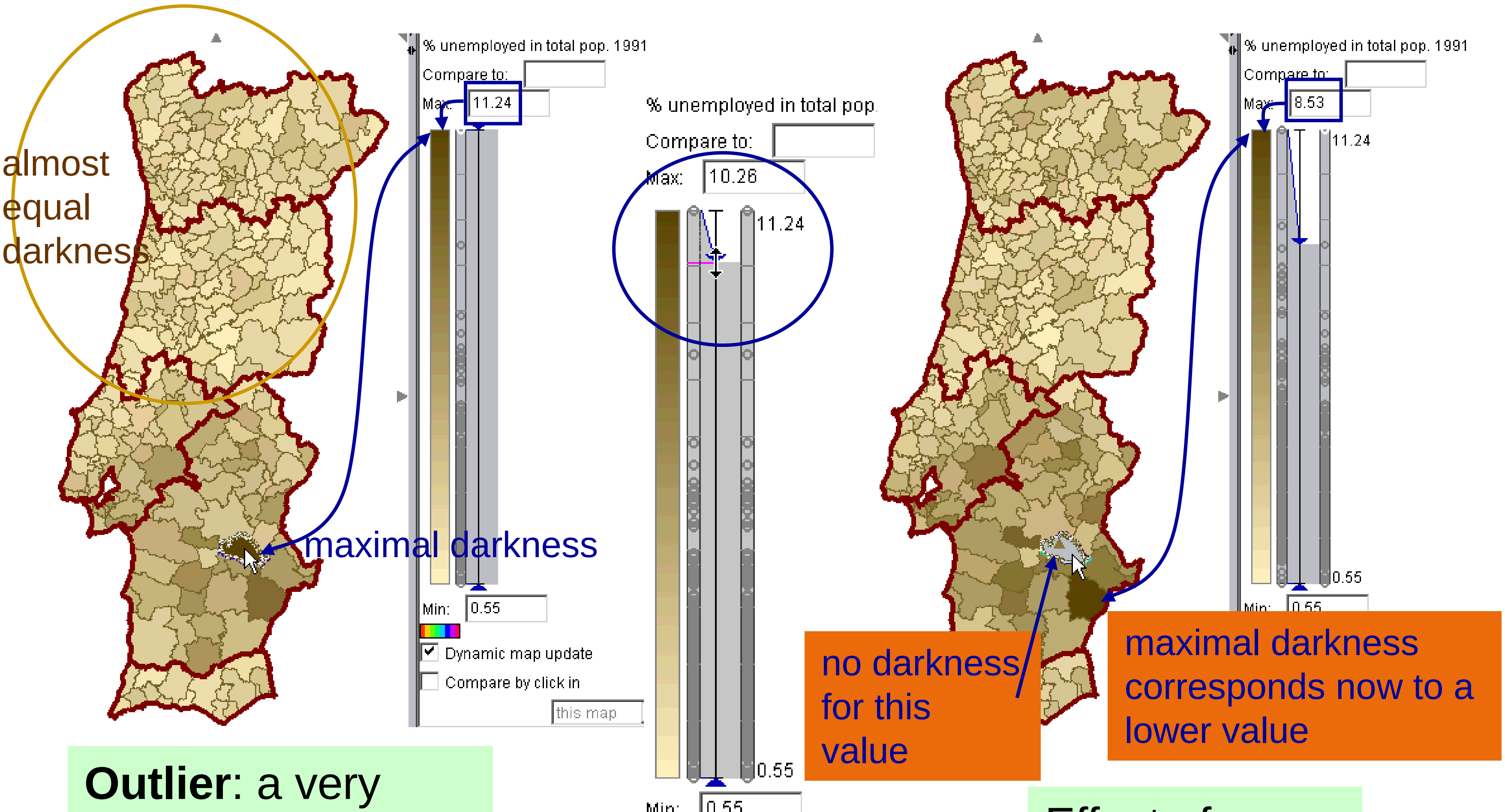
Change the spatial scale, e.g. states or counties

Change the theme, e.g. president elections or governor elections, absolute values or differences in comparison to the previous time

Choose the visualization method, e.g. area painting or proportional symbols

Examples of analytical interactions

Removing Outliers (1)



Outlier: a very high (or very low) value, far apart from others

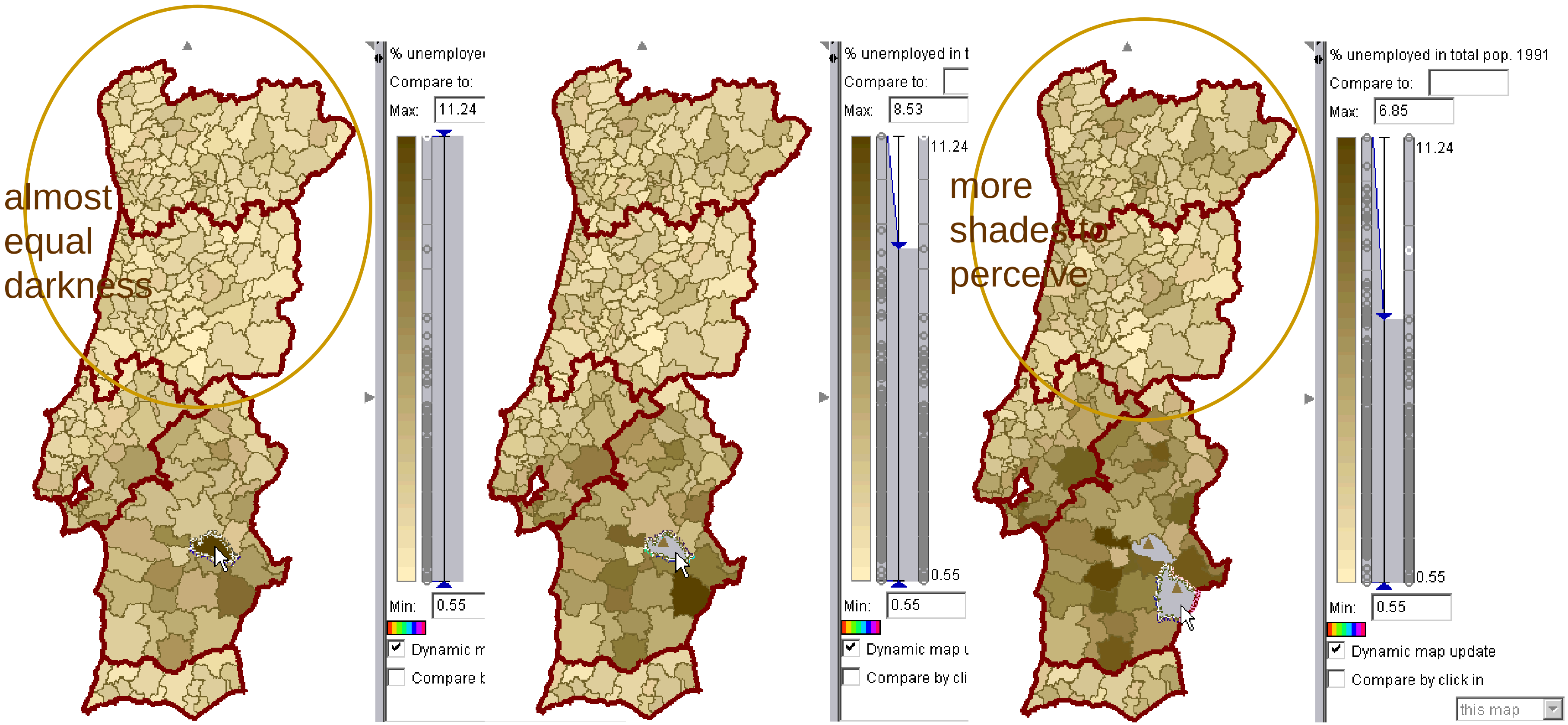
Interactive outlier removal

Effect of outlier removal

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Removing Outliers (2)



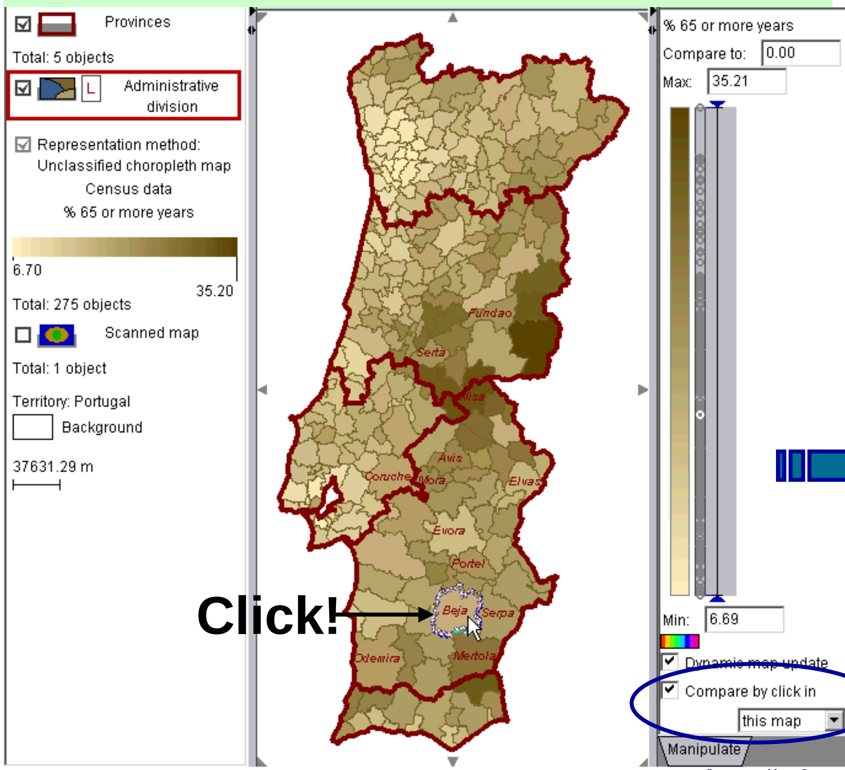
After the removal of two outliers, the differences are better seen

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Object Comparison

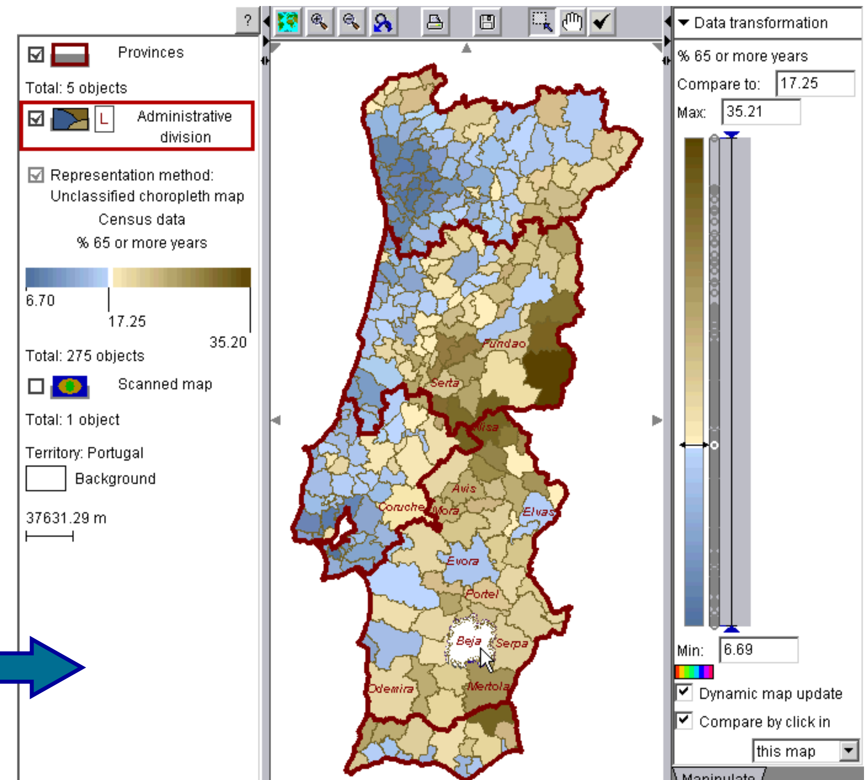
The diverging colour scale allows us to compare an object with all others:



CommonGIS 1998-2003: Portugal census 1981 and 1991

Click!

must be “checked”

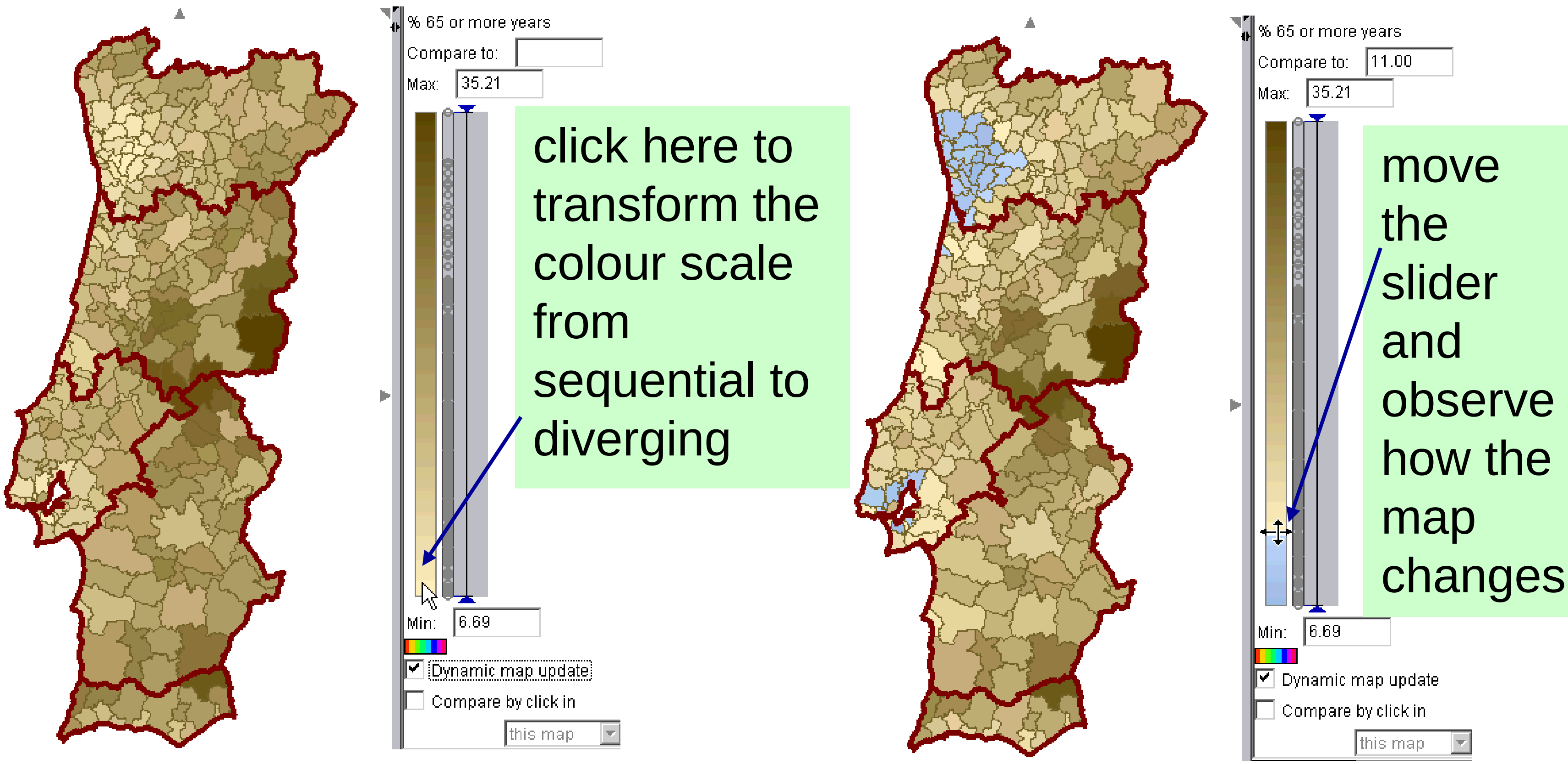


After clicking on Beja:

lower values than in Beja → blue

higher values than in Beja → brown

Pattern Investigation (1)

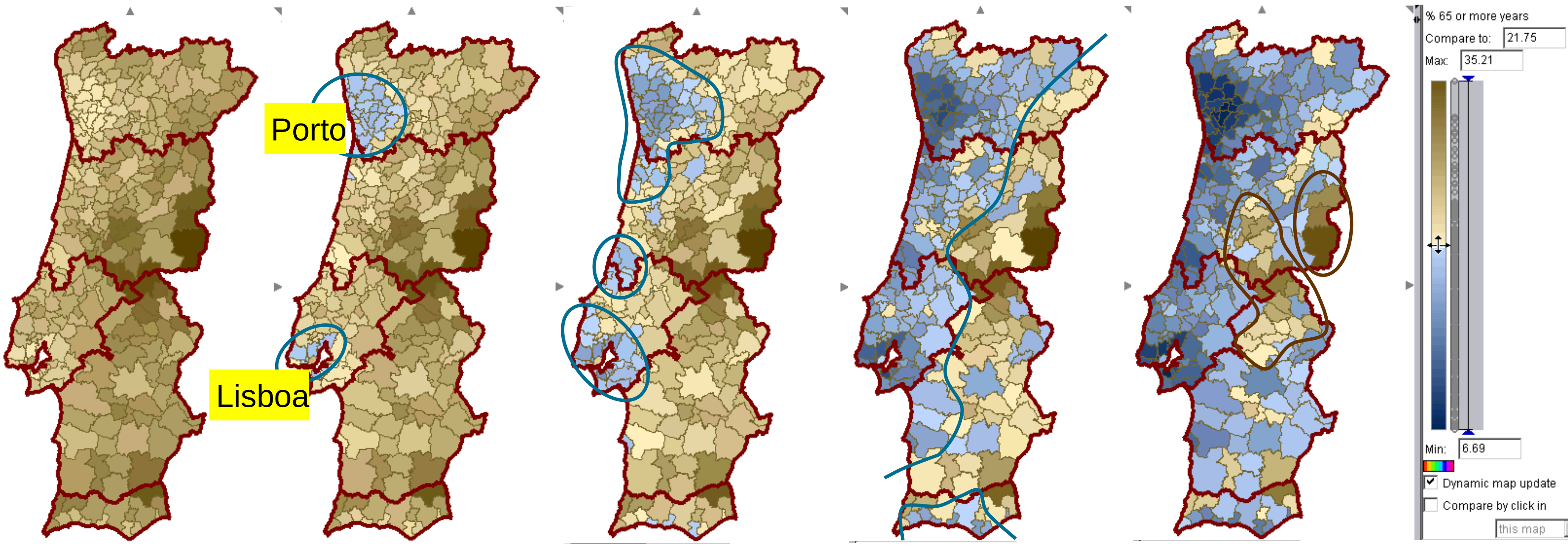


CITY UNIVERSITY LONDON

Fraunhofer IAIS

Pattern Investigation (2)

By moving the slider, we see more patterns and gain more understanding of value distribution



West-to-east increase

Clusters of low values around Porto and Lisboa

One more cluster of low values

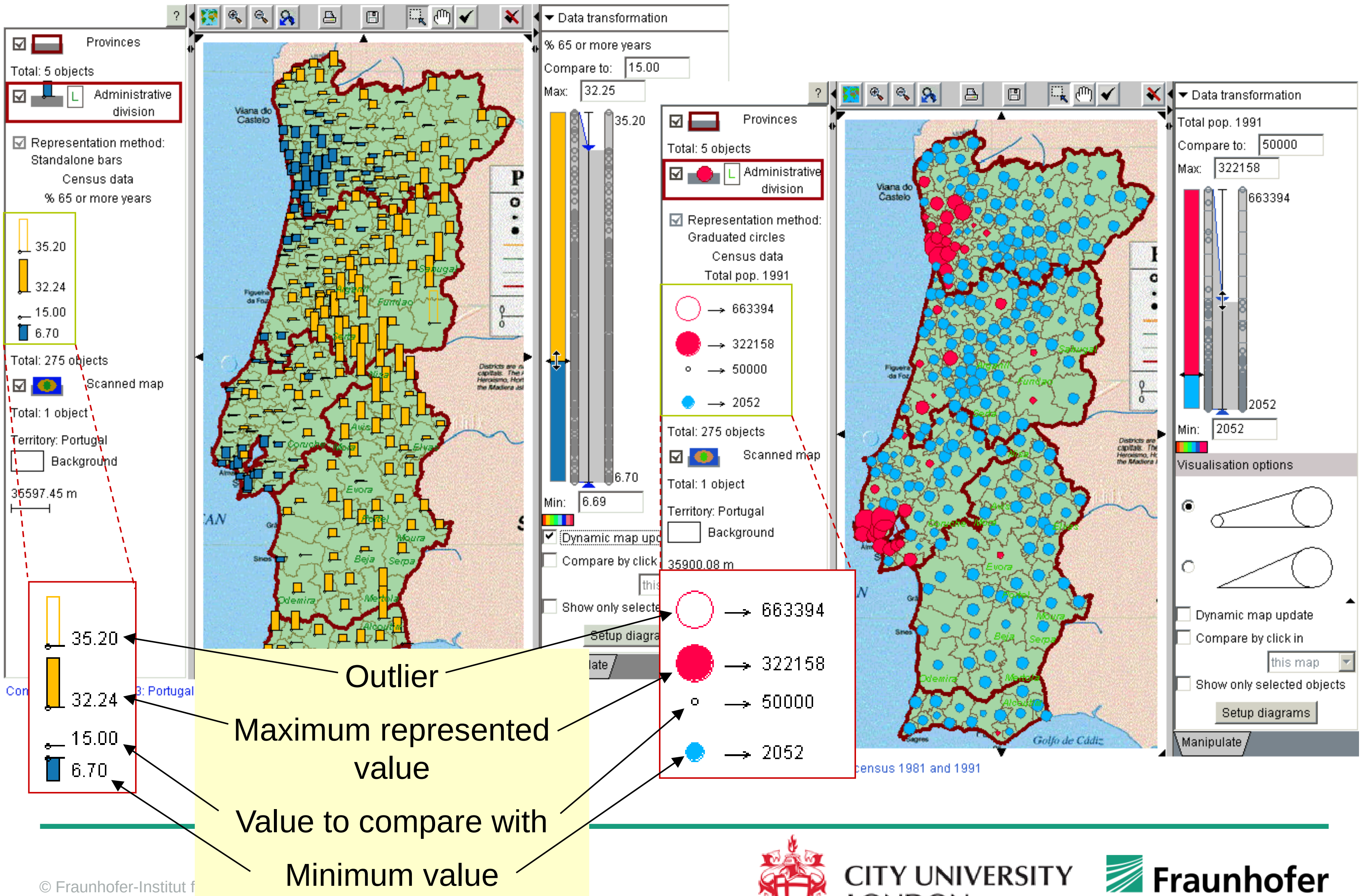
Coast-inland contrast

Clusters of high values in central-east

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Focusing and Visual Comparison on Other Map



CITY UNIVERSITY LONDON

Fraunhofer IAIS

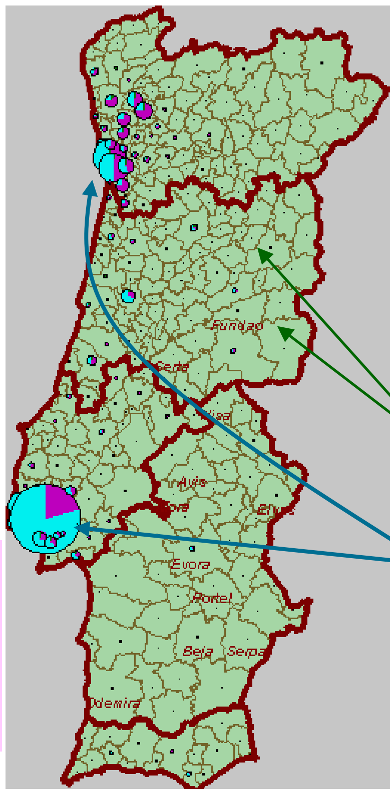
Piechart Map

Applicable to several attributes that together give some meaningful whole

“Pie” size is proportional to the total (sum of the attribute values)

The division into slices shows proportion of each attribute in the total

However, the map often looks like this:

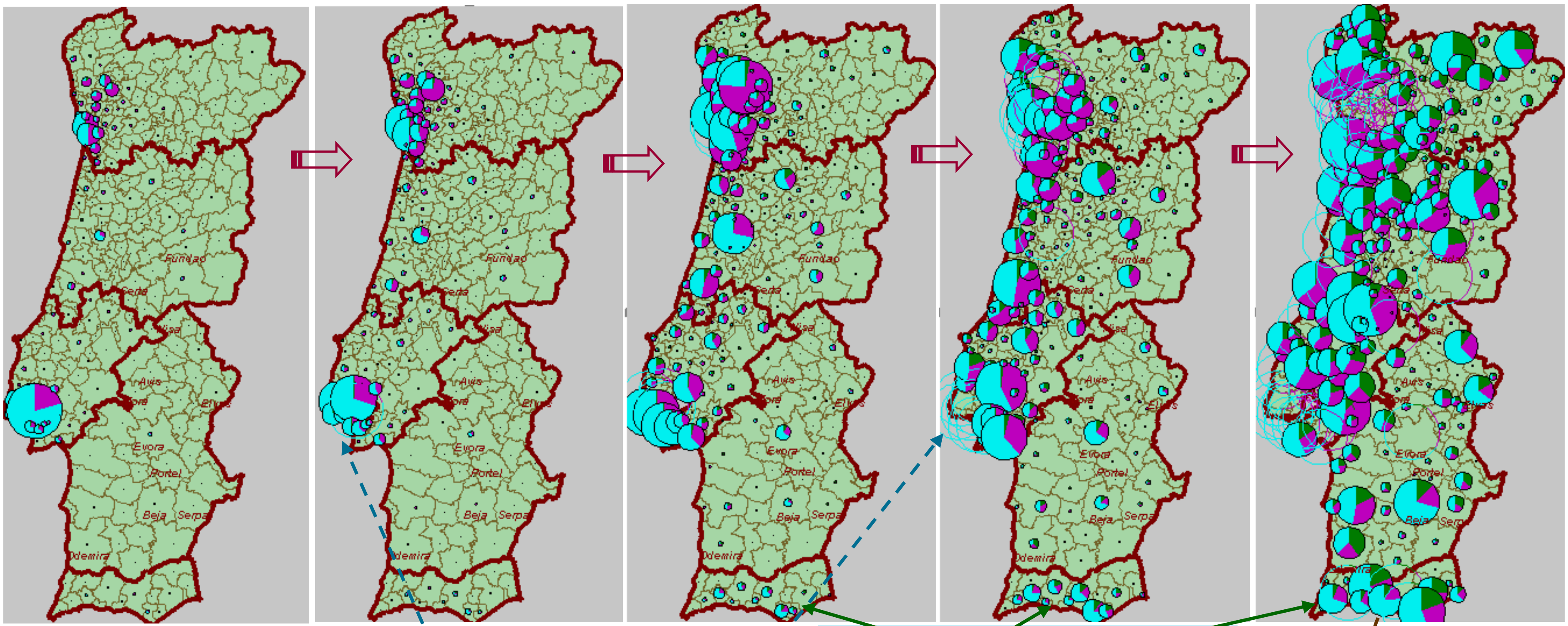


Here the population is very small in comparison to the large cities. Therefore, the pies are too small to be seen

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Piechart Map: Focusing



Move this delimiter to the left

☑ Diameter proportional to sum

566 281078

566 | 281078

The largest pies are gradually removed (replaced by hollow circles)

The remaining pies become larger

Diameter proportional to sum

281078

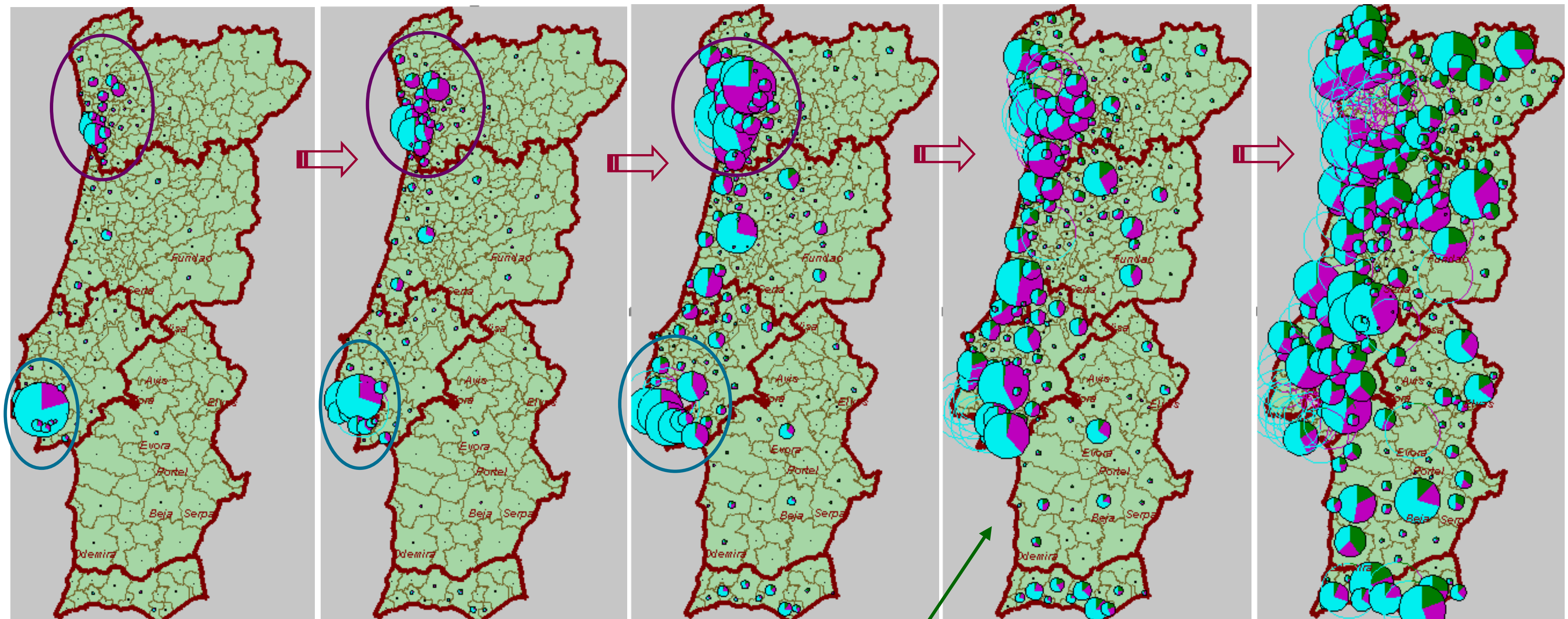
19267

Now the maximum pie size corresponds to this value

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Focusing and Data Investigation



In districts with much population people work in industry (magenta) and services (cyan).

Northwest: more industry

Centre-west: more services

At this stage, the agricultural part (green) becomes visible

In districts with little population considerable proportion of people works in agriculture, but services still prevail

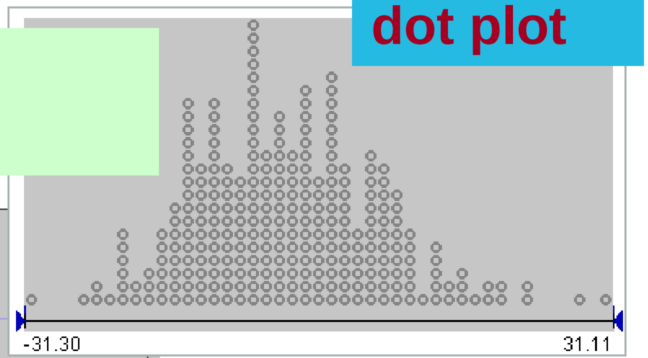
CITY UNIVERSITY LONDON

Fraunhofer IAIS

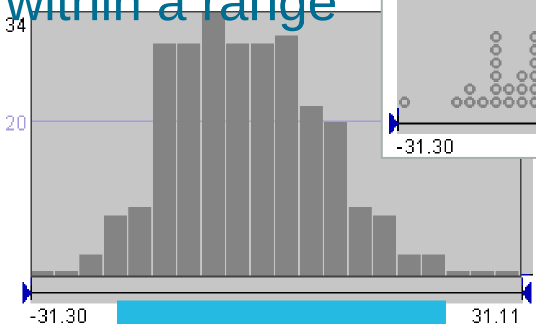
Why to Use Multiple Views?

Dot plot

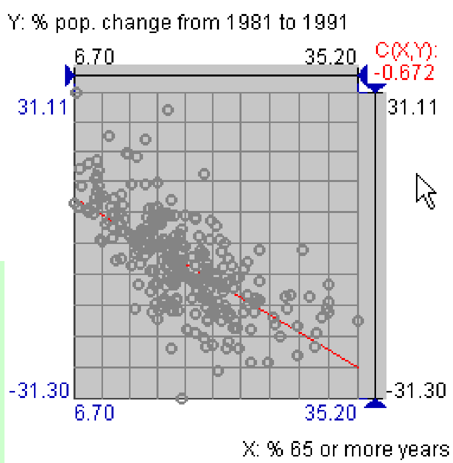
Stacked dot plot



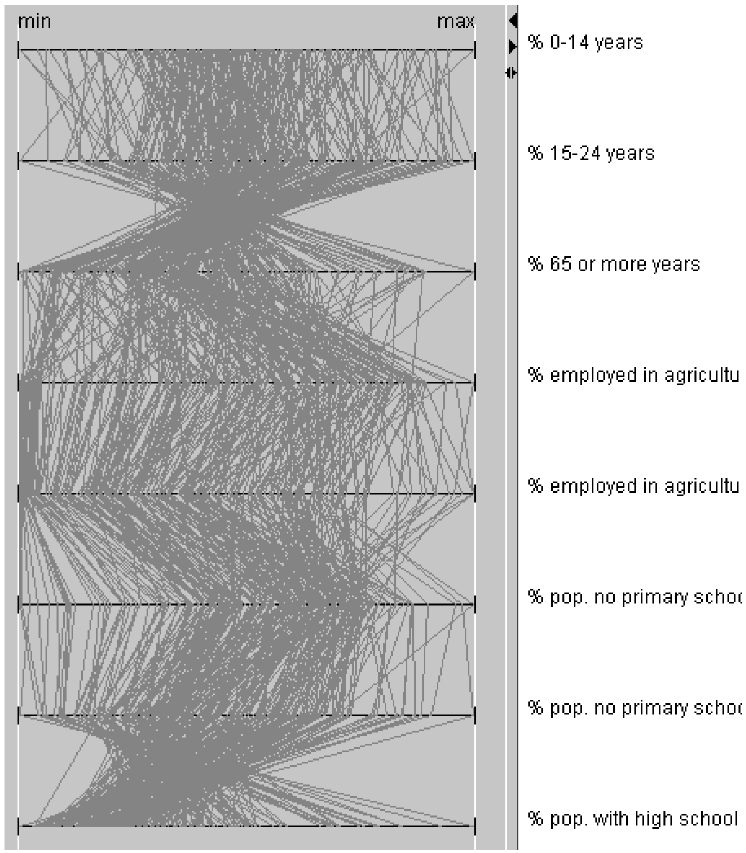
Distribution of attribute values within a range



Frequency histogram



Scatter plot: shows how two attributes are related

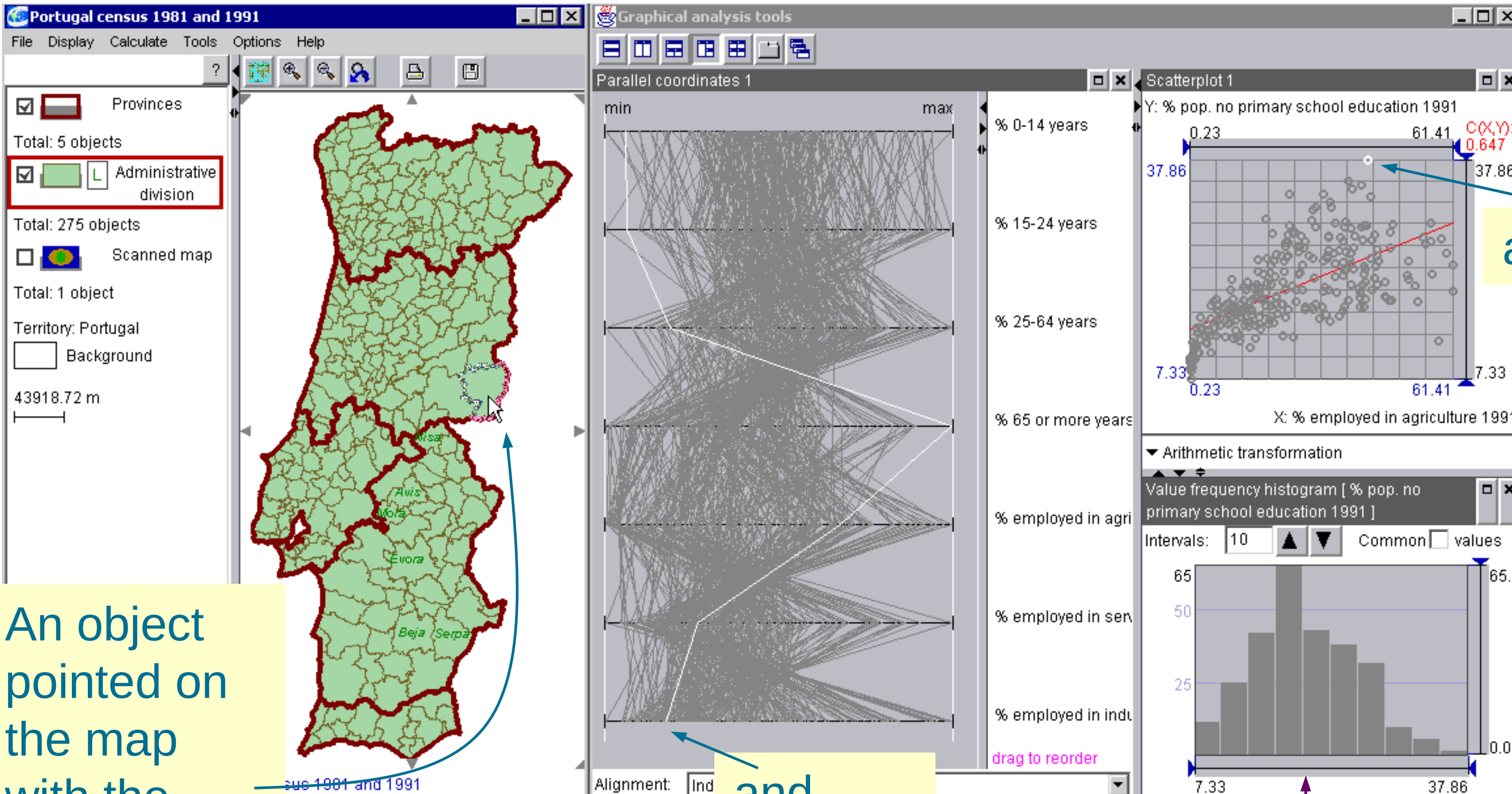


Parallel coordinates: object characteristics profiles; relationships between attributes (look at line slopes)

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Display Linking by Highlighting



An object pointed on the map with the mouse

is simultaneously highlighted here,

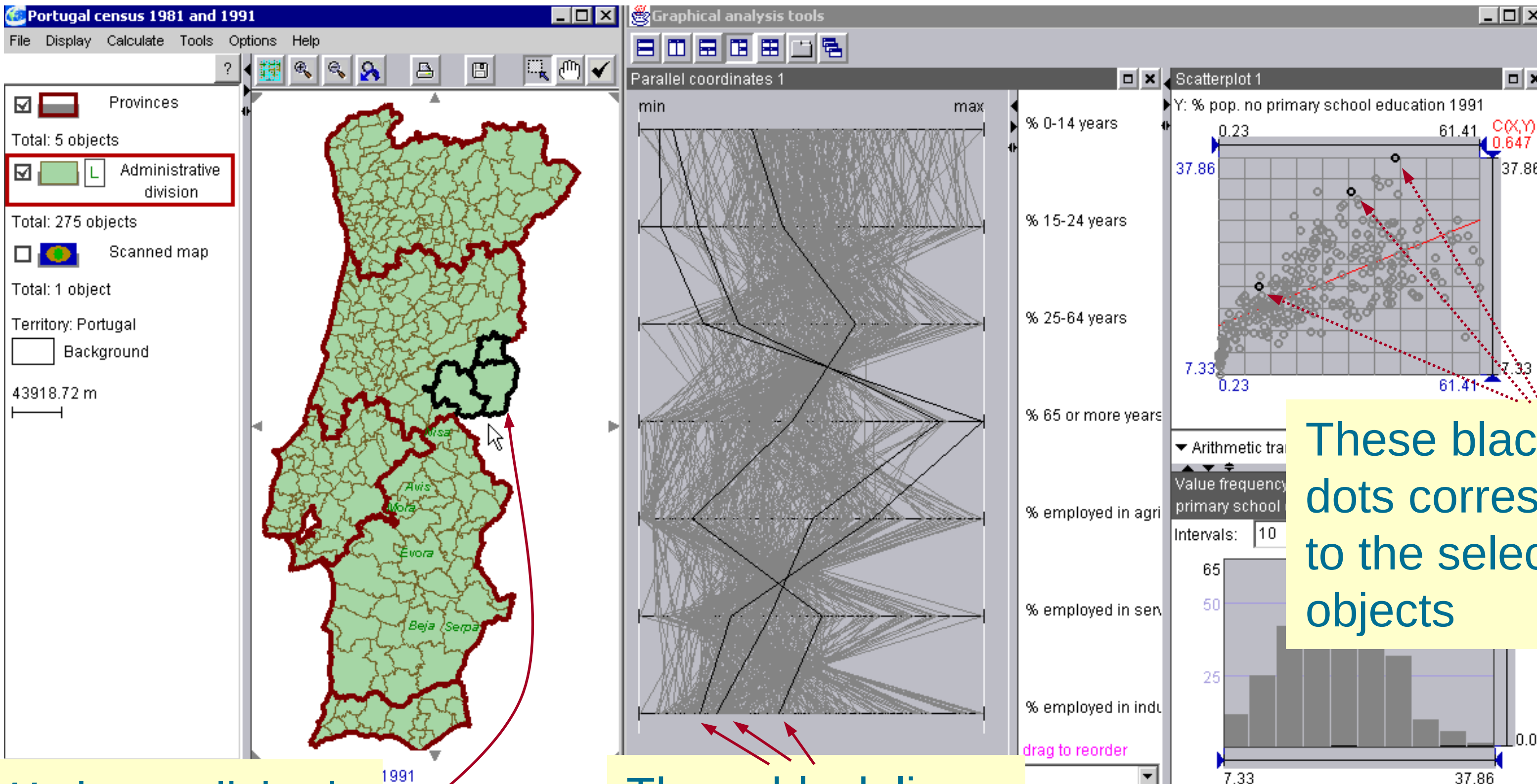
and here,

and here,

but not here: this is an aggregated view that does not show individual objects

Display Linking by Selection

Selection (durable highlighting) does not disappear after the mouse is moved away. One or more objects may be selected e.g. by mouse-clicking on them.



These black dots correspond to the selected objects

We have clicked on each of these 3 objects

These black lines correspond to the selected objects

CITY UNIVERSITY LONDON

Fraunhofer IAIS

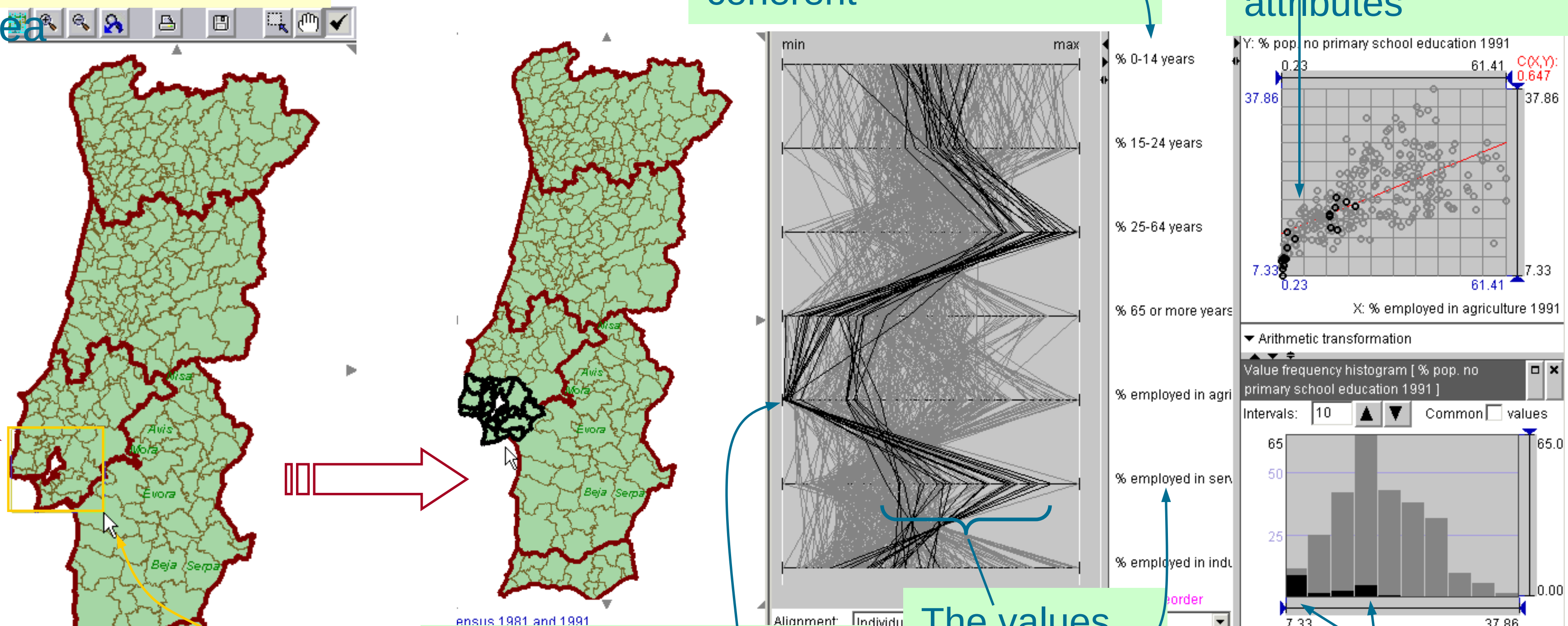
Using Display Linking (1)

Let us examine characteristics of districts in this area

Enclose the area in a frame

The characteristics in terms of the upper 4 attributes are rather coherent

Two distinct clusters in the value space of these two attributes



The values of this attribute are split in two groups with a gap between

The values of this attribute greatly vary

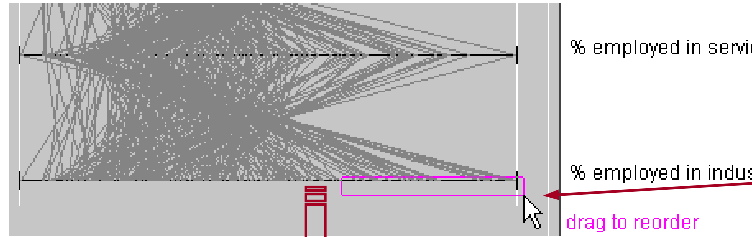
The districts fit in the left half of the histogram, mostly in bars 1 and 4

CITY UNIVERSITY LONDON

Fraunhofer IAIS

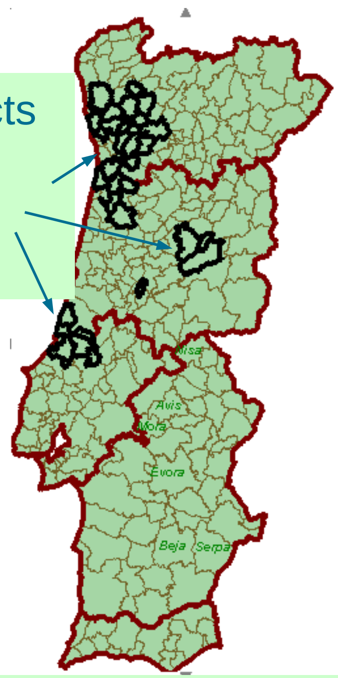
Using Display Linking (2)

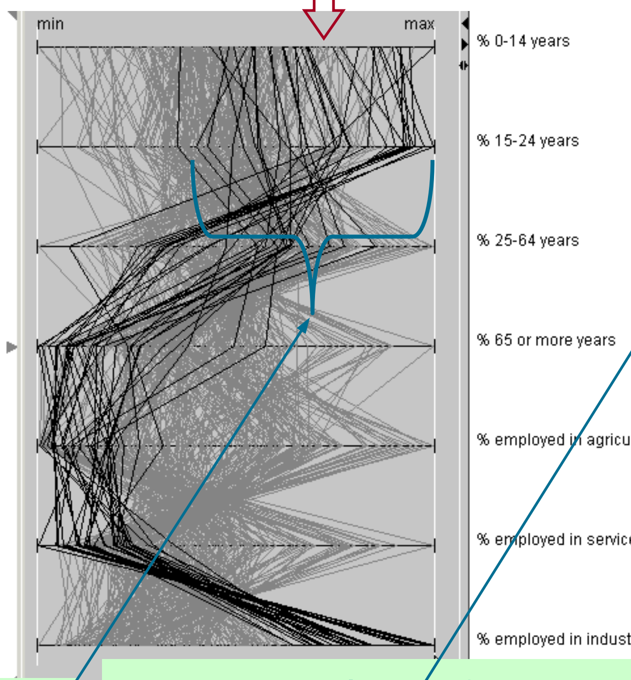
Let us look at the districts with high % employed in industry:

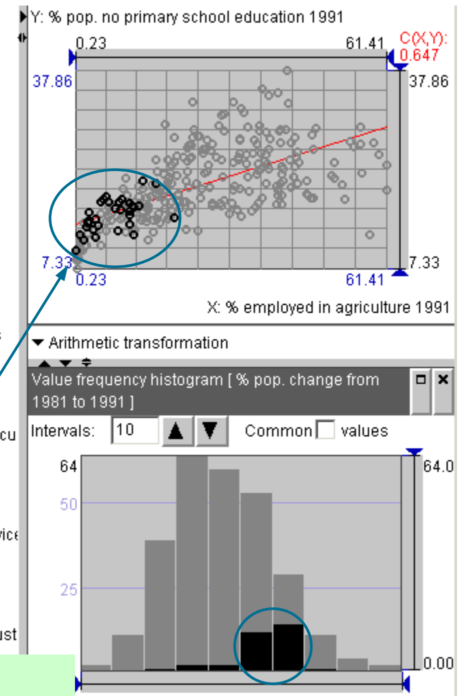


Select high values by drawing a frame

The districts form 3 spatial clusters







The districts have average or high proportions of children and young

Low proportions of agricultural workers and people without primary school education

Population change: mostly between –0.1% and 12.4%

CITY UNIVERSITY LONDON

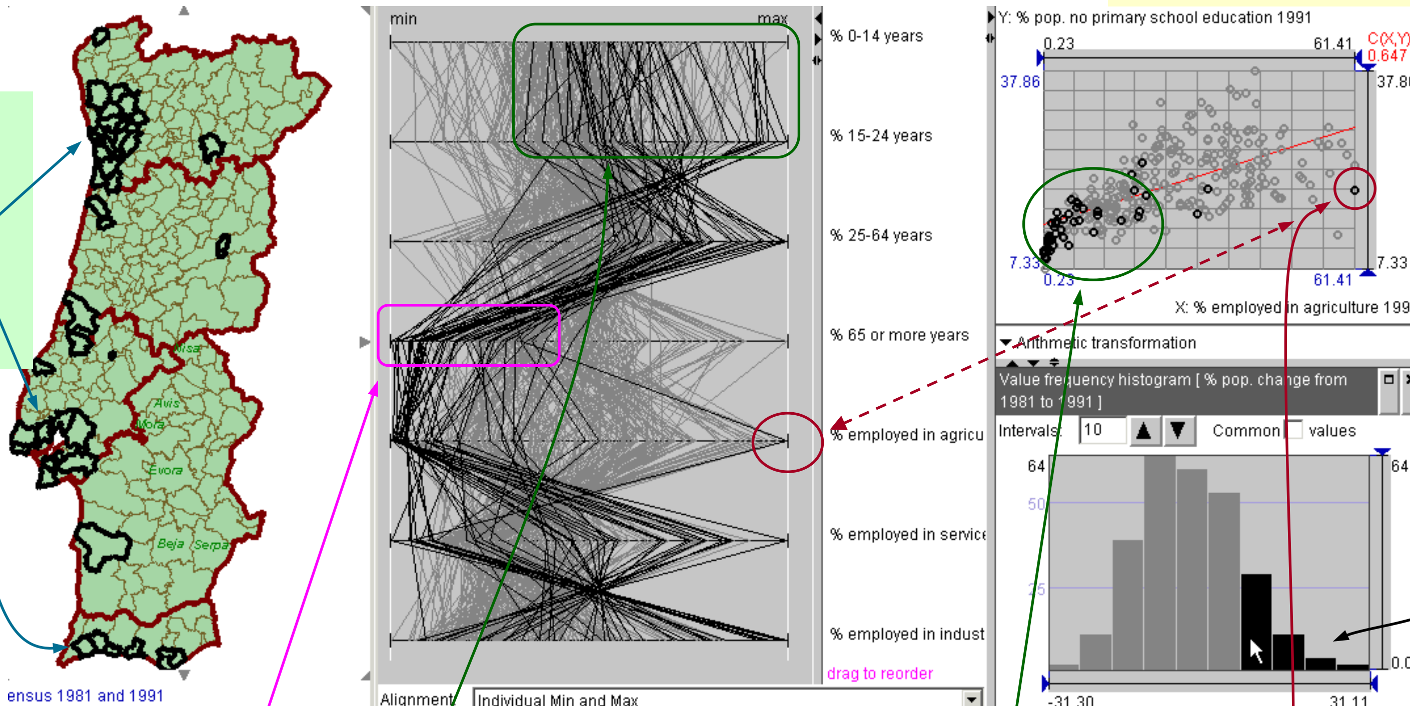
Fraunhofer IAIS

Using Display Linking (3)

Let us look at the districts with the highest population growth:

Click on the rightmost bars in the histogram

The districts form some spatial clusters



The districts have average or high proportions of children and young and low proportions of old people

The proportions of agricultural workers and people without primary school education are mostly low, but there is an outlier

CITY UNIVERSITY LONDON

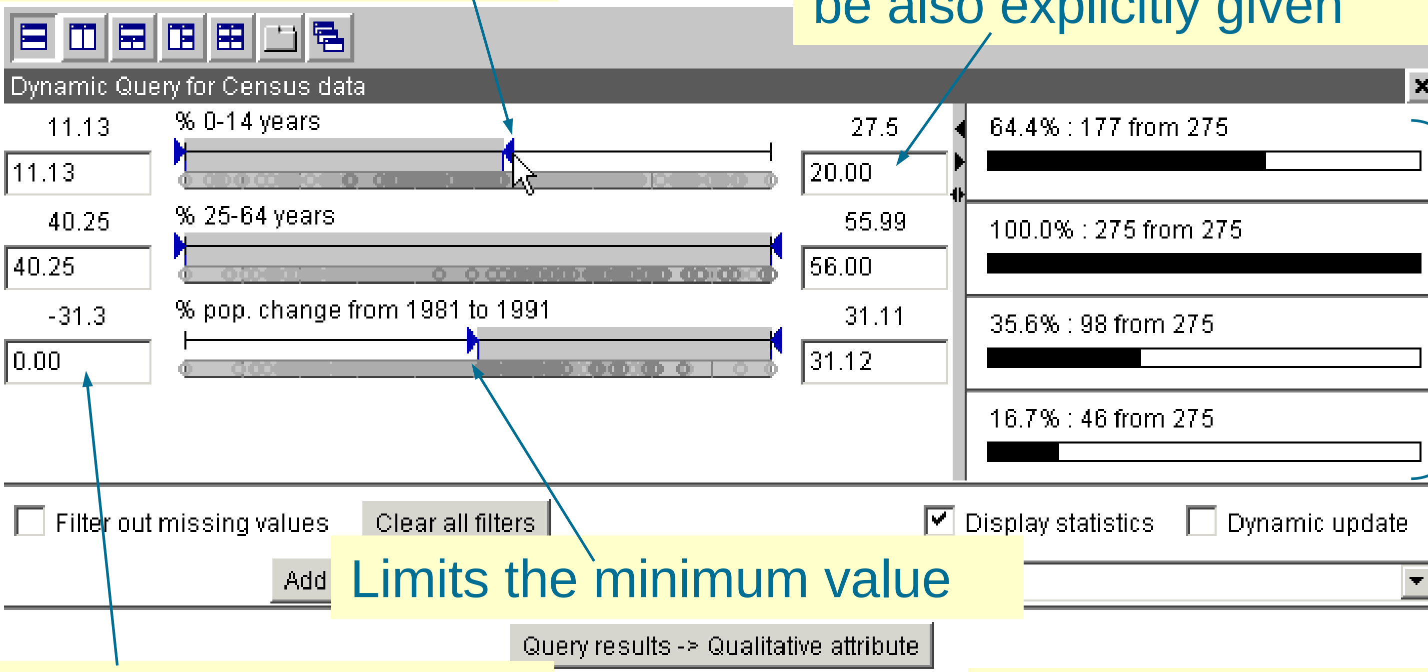
Fraunhofer IAIS

Dynamic Query

Dynamic query allows us to set constraints on attribute values

Limits the maximum value

The maximum limit can be also explicitly given



Limits the minimum value

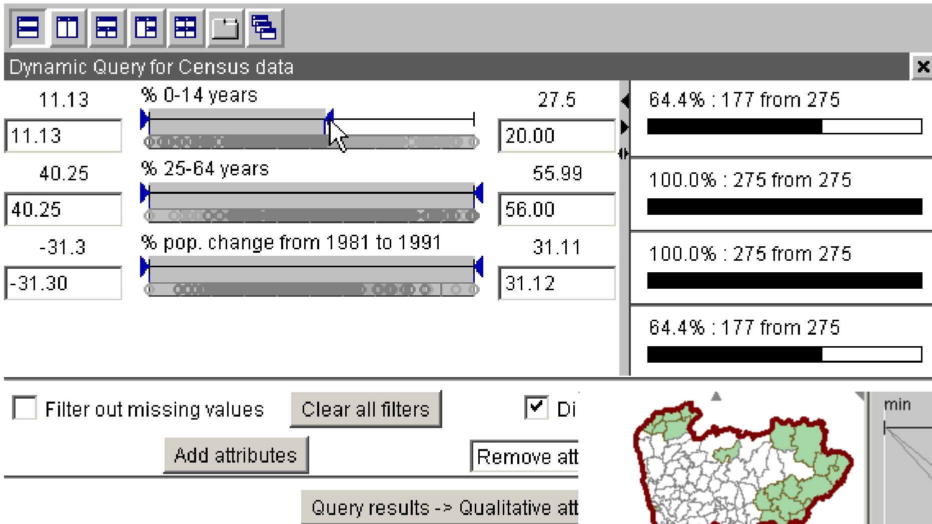
The minimum limit can be also explicitly given

Statistics of constraint satisfaction

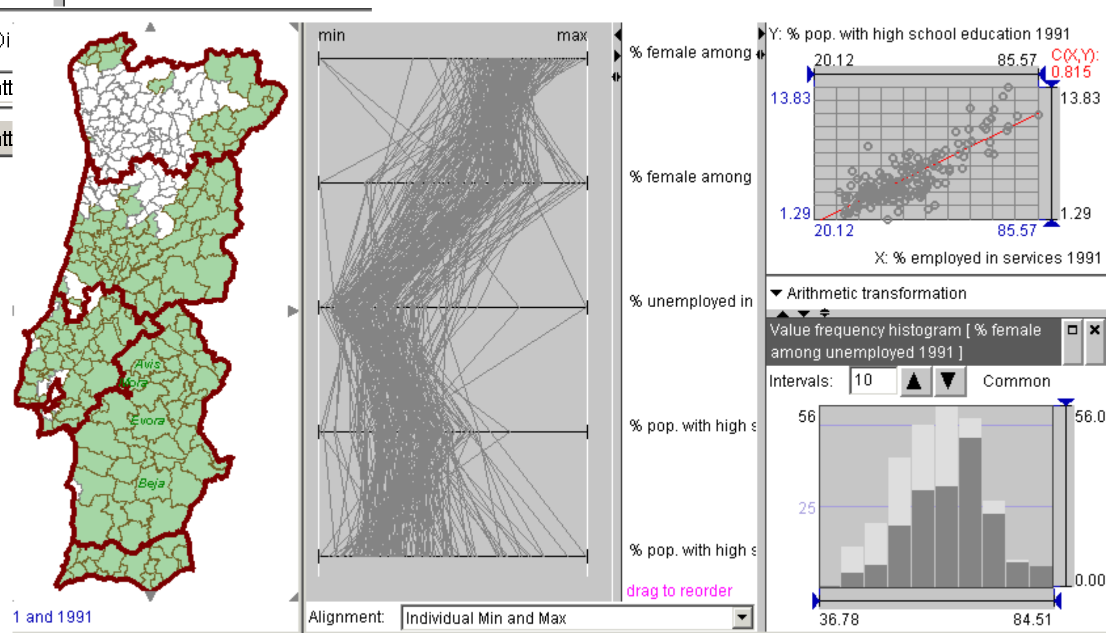
CITY UNIVERSITY LONDON

Fraunhofer IAIS

Dynamic Query in Action (1)



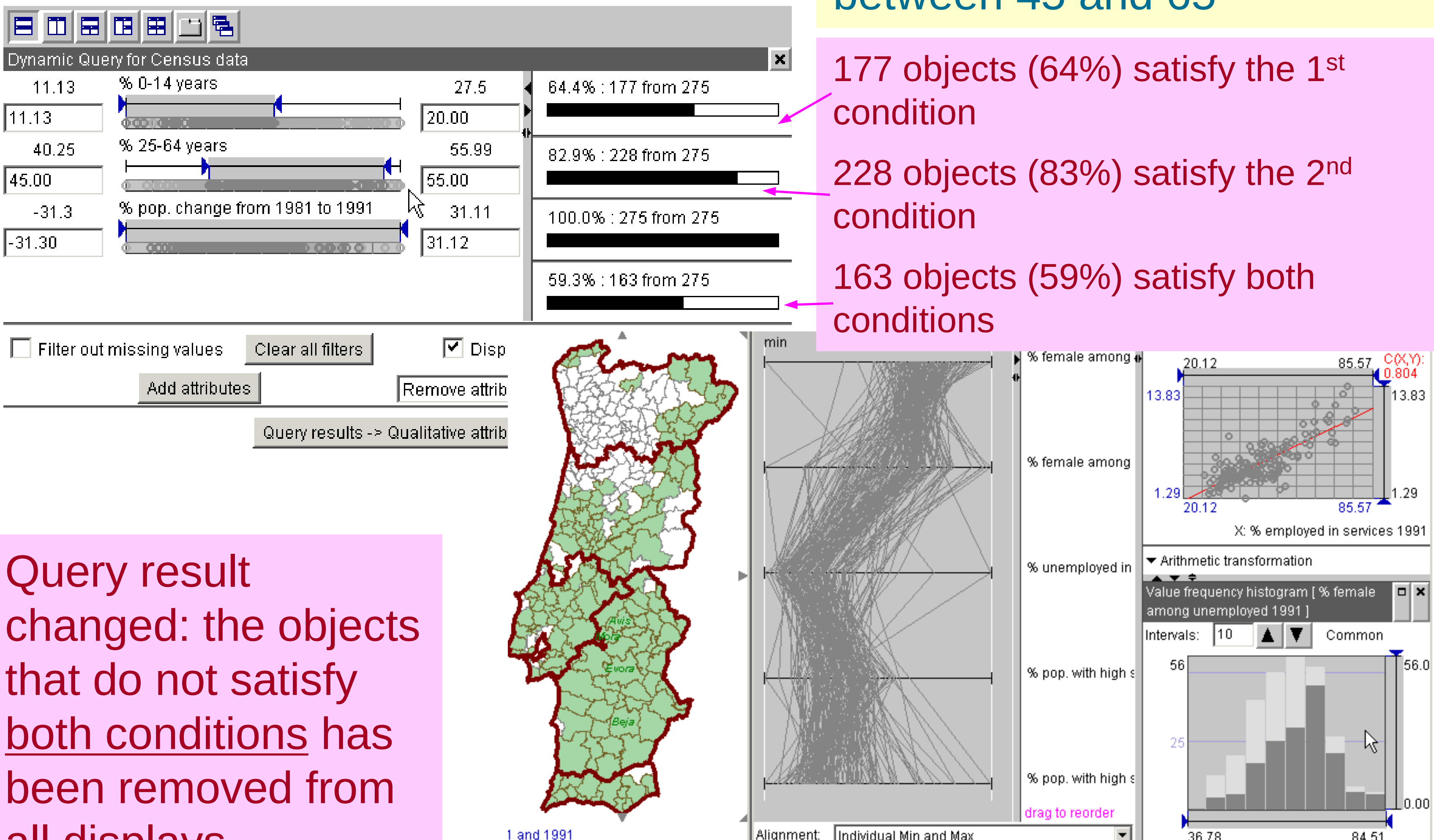
Query condition: % 0-14 years must be below 20



Query result: the objects that do not satisfy the condition has been removed from all displays

Dynamic Query in Action (2)

A second query condition added: % 25-64 years must be between 45 and 65



177 objects (64%) satisfy the 1st condition

228 objects (83%) satisfy the 2nd condition

163 objects (59%) satisfy both conditions

Query result changed: the objects that do not satisfy <u>both conditions</u> has been removed from all displays

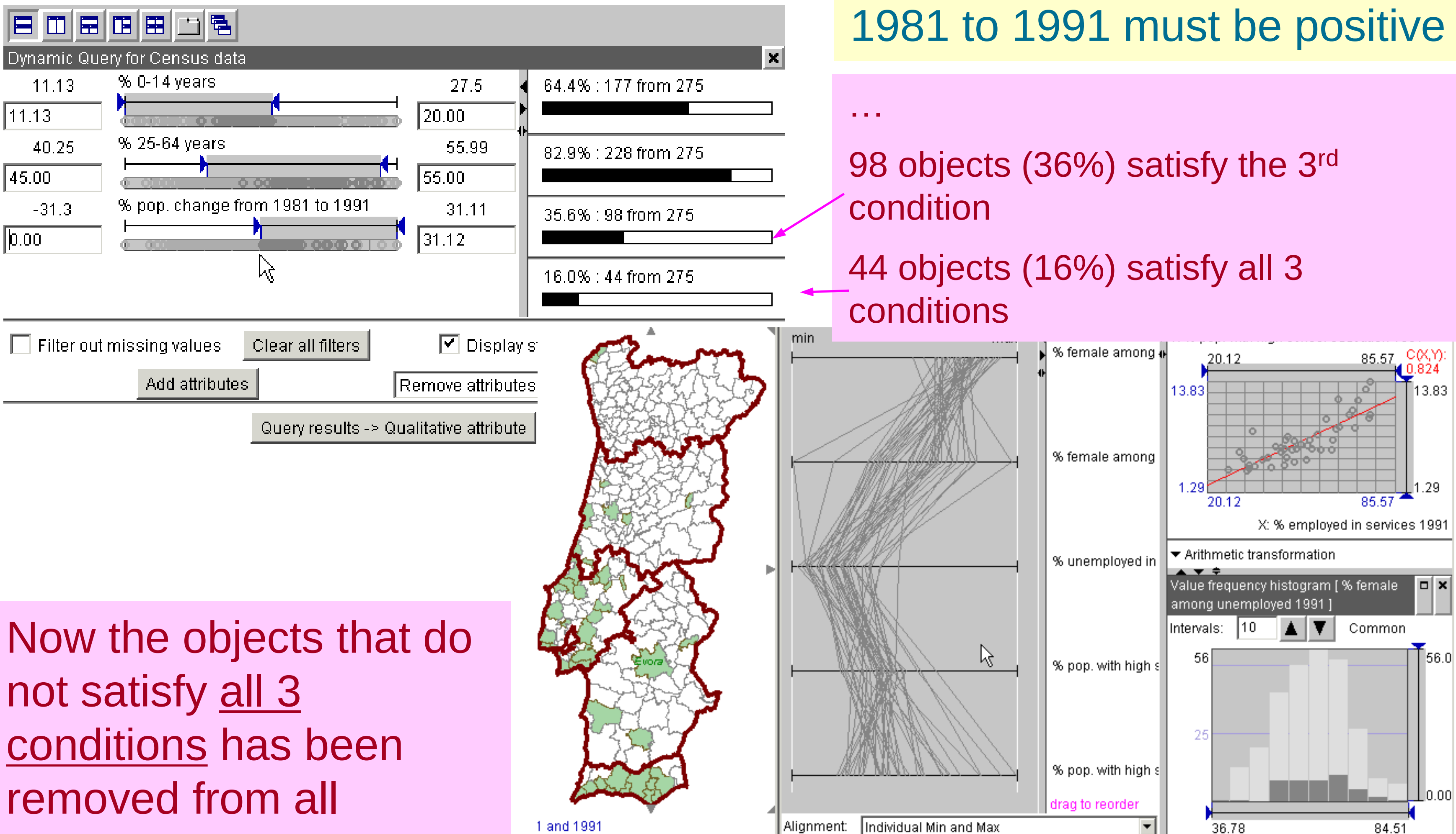
Dynamic Query in Action (3)

One more query condition added: % pop. change from 1981 to 1991 must be positive

…

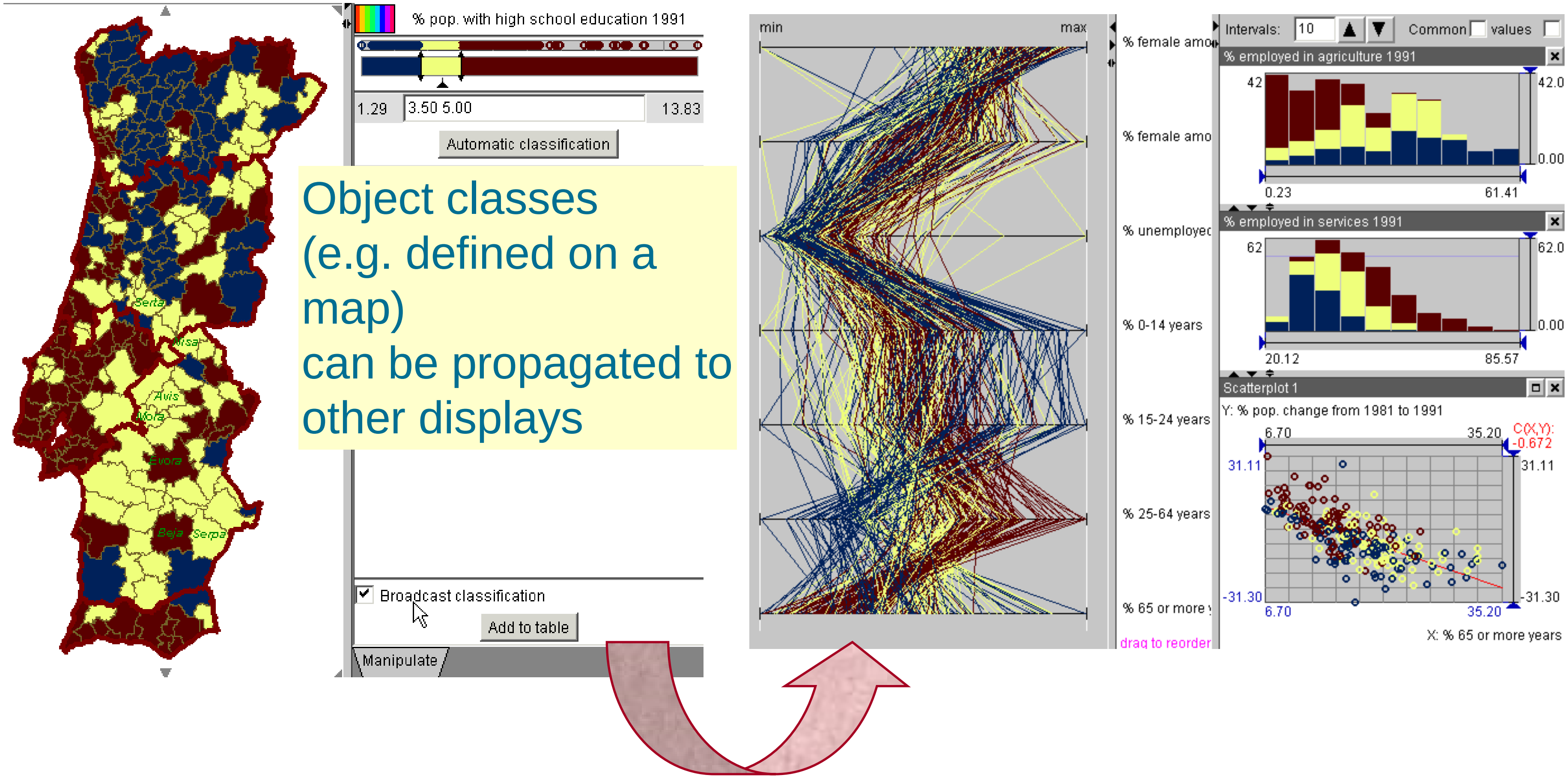
98 objects (36%) satisfy the 3rd condition

44 objects (16%) satisfy all 3 conditions



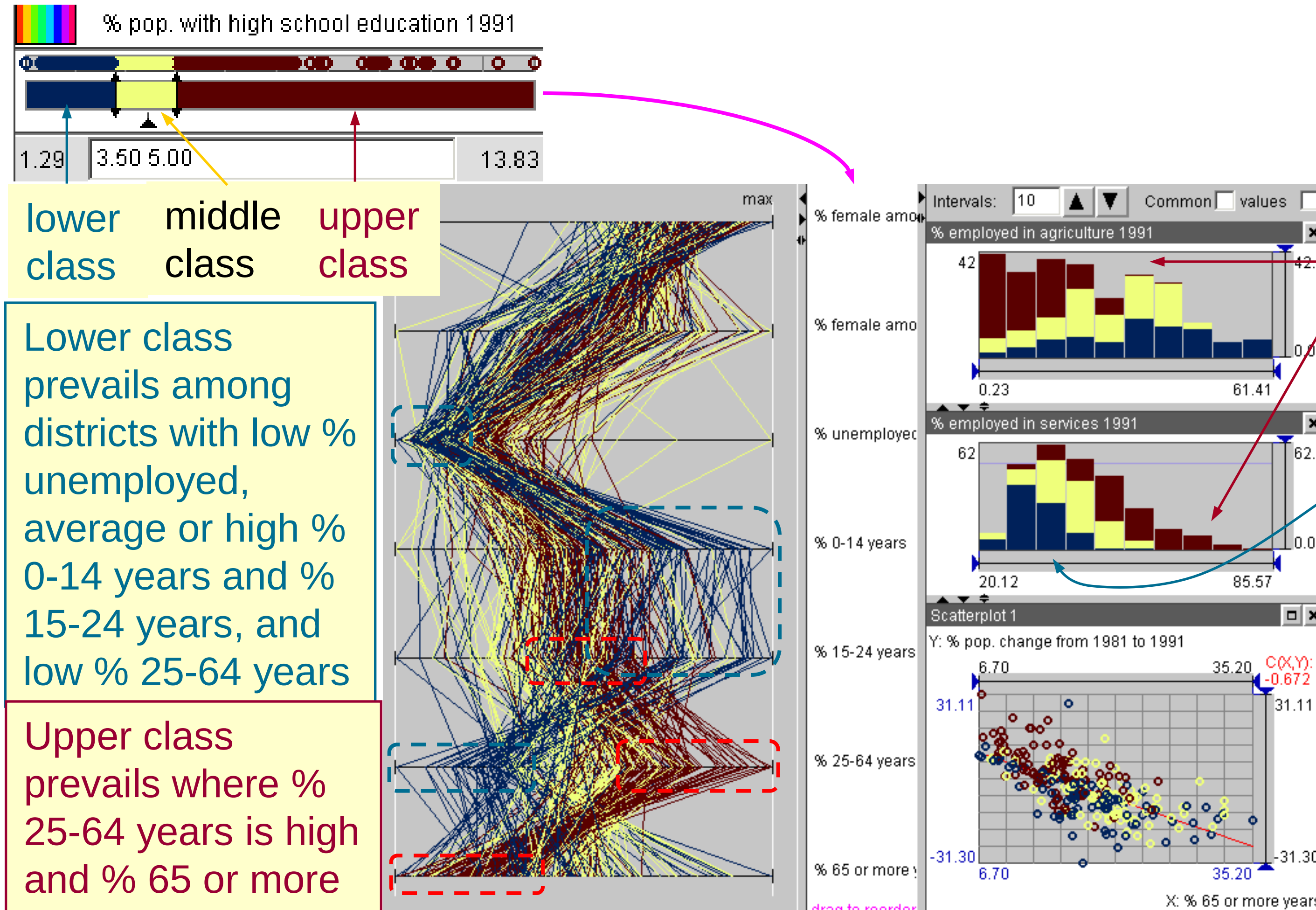
Now the objects that do not satisfy <u>all 3 conditions</u> has been removed from all displays

Propagation of Object Classes



Object classes
(e.g. defined on a map)
can be propagated to other displays

Propagation of Object Classes: Use Example



lower class

middle class

upper class

Lower class prevails among districts with low % unemployed, average or high % 0-14 years and % 15-24 years, and low % 25-64 years

Upper class prevails where % 25-64 years is high and % 65 or more years is low. It occupies mostly the middle part of the axis % 15-24 years.

Characteristics of the middle class are highly variant

Upper class co-occurs with low % employed in agriculture and average to high % employed in services

Lower class co-occurs with low % employed in services

In districts with high % employed in agriculture the proportion of people with high school education is low (mostly below 3.5%). For % employed in services we see the opposite relationship

Table View and Table Lens (1)

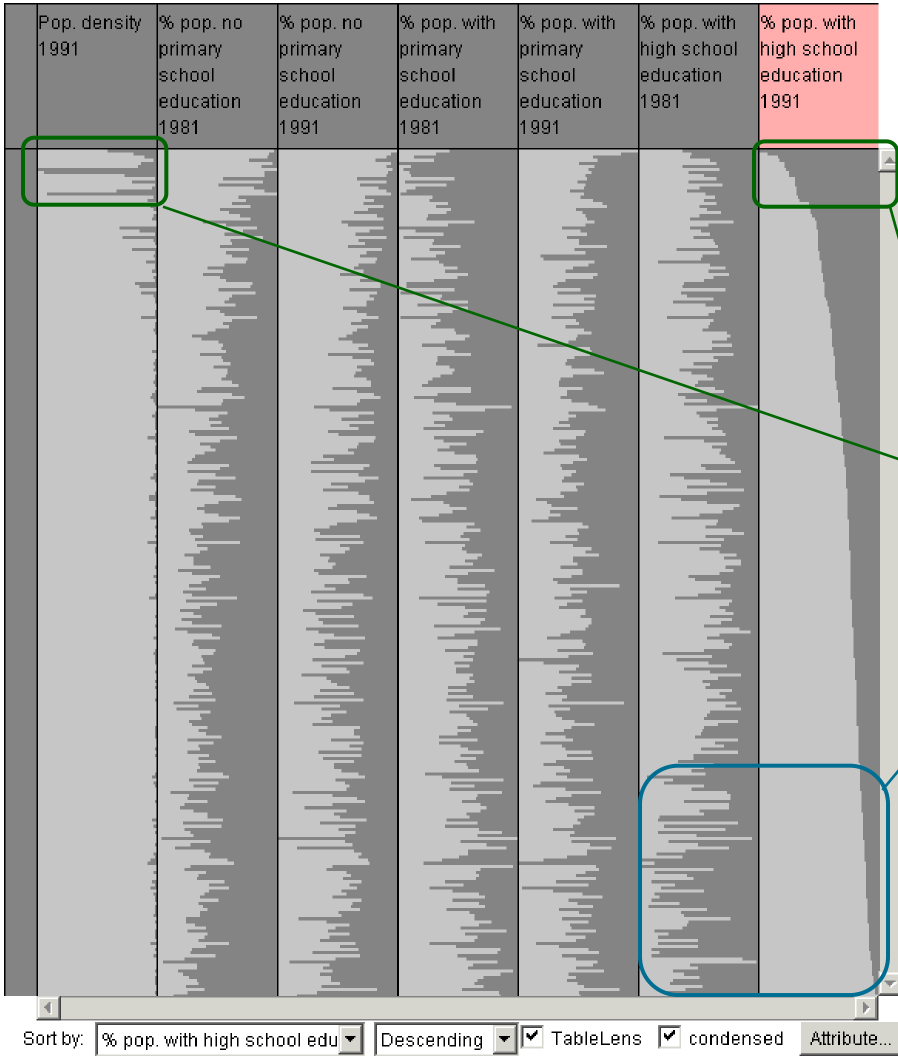
Click for sorting

Table cell shading shows the relative position of the values between the minimum and maximum values of the respective attributes

identifiers	Pop. density 1981	Pop. density 1991	% pop. no primary school education 1981	% pop. no primary school education 1991	% pop. with primary school education 1981	% pop. with primary school education 1991	% pop. with high school education 1981	% pop. with high school education 1991
Lisboa	9636.423	7912.43	37.26	8.51	39.49	25.45	7.675	11.11
Porto	7858.089	7260.49	37.32	8.83	39.97	26.22	8.756	9.84
Amadora	6895.481	7454.641	40.41	9.68	38.66	25.18	9.941	10.80
Oeiras	3257.592	3301.527	37.89	7.33	32.80	18.69	10.444	13.83
Barreiro	2796.012	2723.485	40.64	10.01	39.09	28.16	10.089	9.85
Matosinhos	2190.979	2434.703	42.45	11.01	40.12	29.08	9.857	7.78
Almada	2110.581	2169.072	40.58	10.26	38.49	25.99	8.813	10.66
Sao Joao da Madeira	2027.62	2275.216	41.12	10.21	39.62	28.59	9.658	7.68
Espinho	1513.025	1631.933	44.68	11.37	35.82	27.50	10.919	7.56
Cascais	1457.75	1579.276	39.37	8.92	34.97	21.13	9.718	12.95
Loures	1419.716	1654.349	40.84	10.29	39.32	27.12	9.950	10.09
Vila Nova de Gaia	1324.968	1455.128	43.66	11.61	39.18	29.86	10.352	7.07
Gondomar	981.172	1074.426	43.20	12.31	39.93	29.72	10.493	6.93
Maia	975.854	1112.915	44.45	11.71	39.08	30.47	10.406	7.07
Moita	966.594	1181.663	46.27	14.50	37.64	26.61	9.123	7.71
Seixal	952.844	1249.3	41.18	10.15	38.33	24.60	10.375	11.15
Valongo	880.038	1016.194	42.80	11.90	39.94	30.70	11.492	6.40
Entroncamento	874.161	1038.394	38.51	9.10	40.18	26.73	7.952	11.22

Sort by: Pop. density 1981 | Descending | TableLens | condensed | Attribute...

Table View and Table Lens (2)

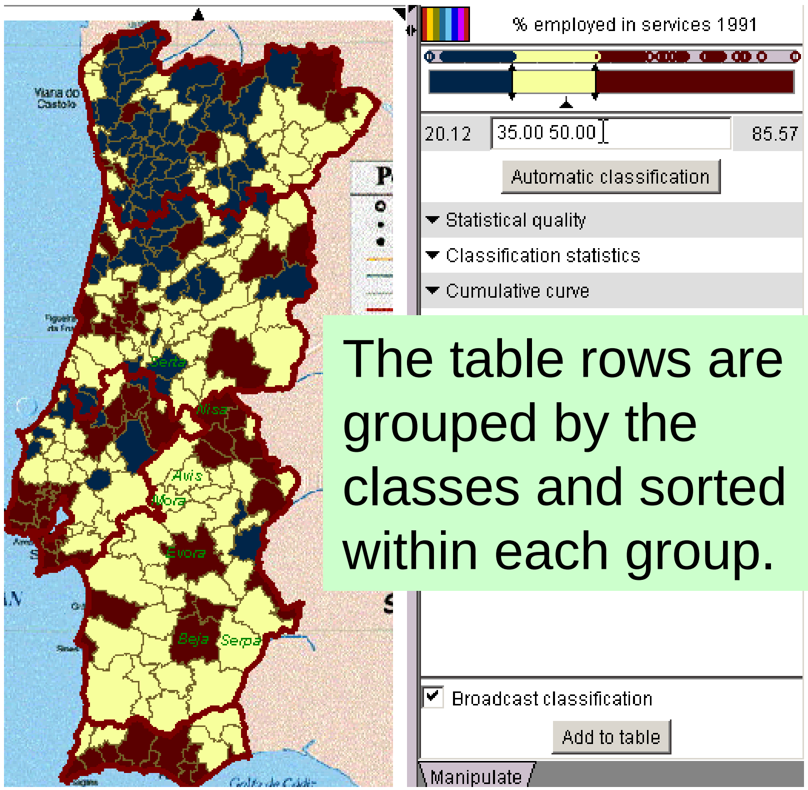


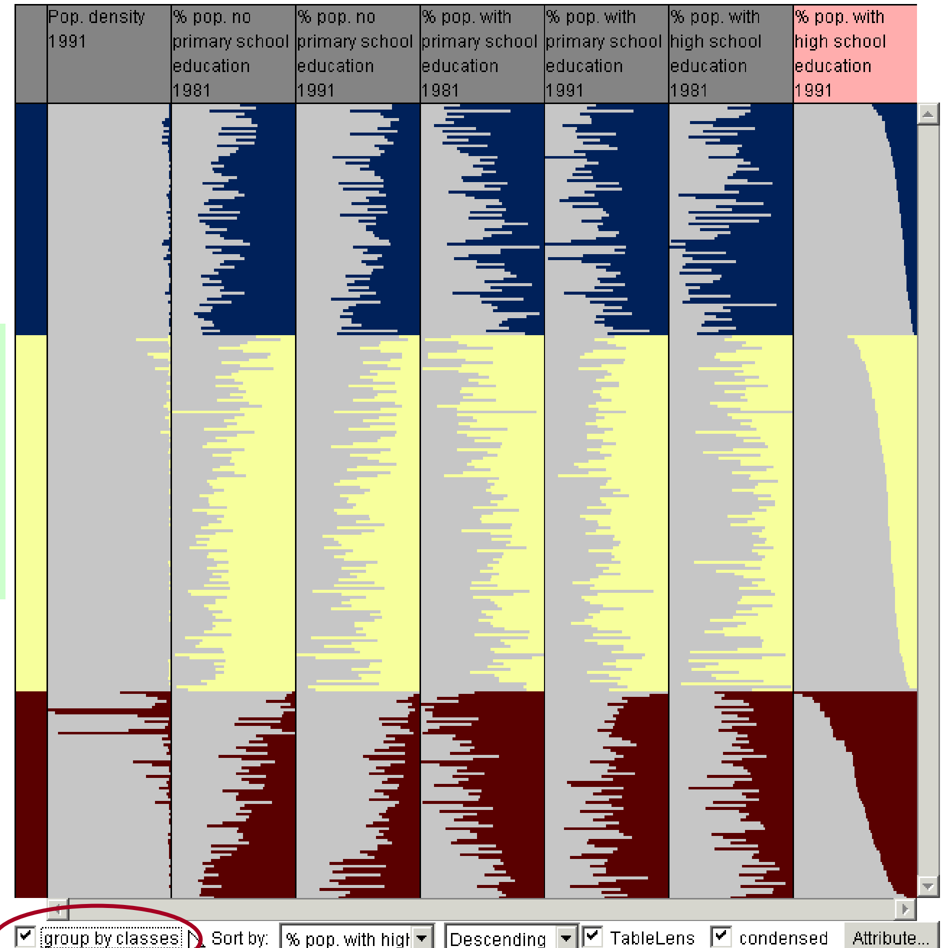
The same information can be represented in a “condensed” form. We do not see the details about particular objects but get an overall impression about value variation and relationships between attributes.

High proportions of people having high school education often co-occur with high population density

Surprisingly, the districts with the lowest proportions of people having high school education in 1991 had much higher proportion of such people in 1981

Class Propagation to Table View (1)

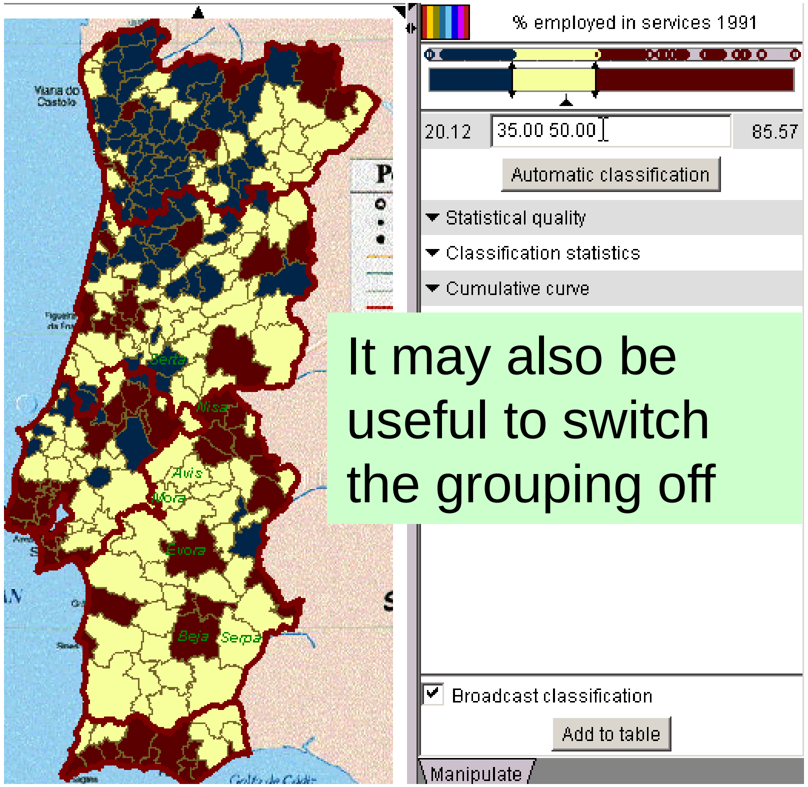




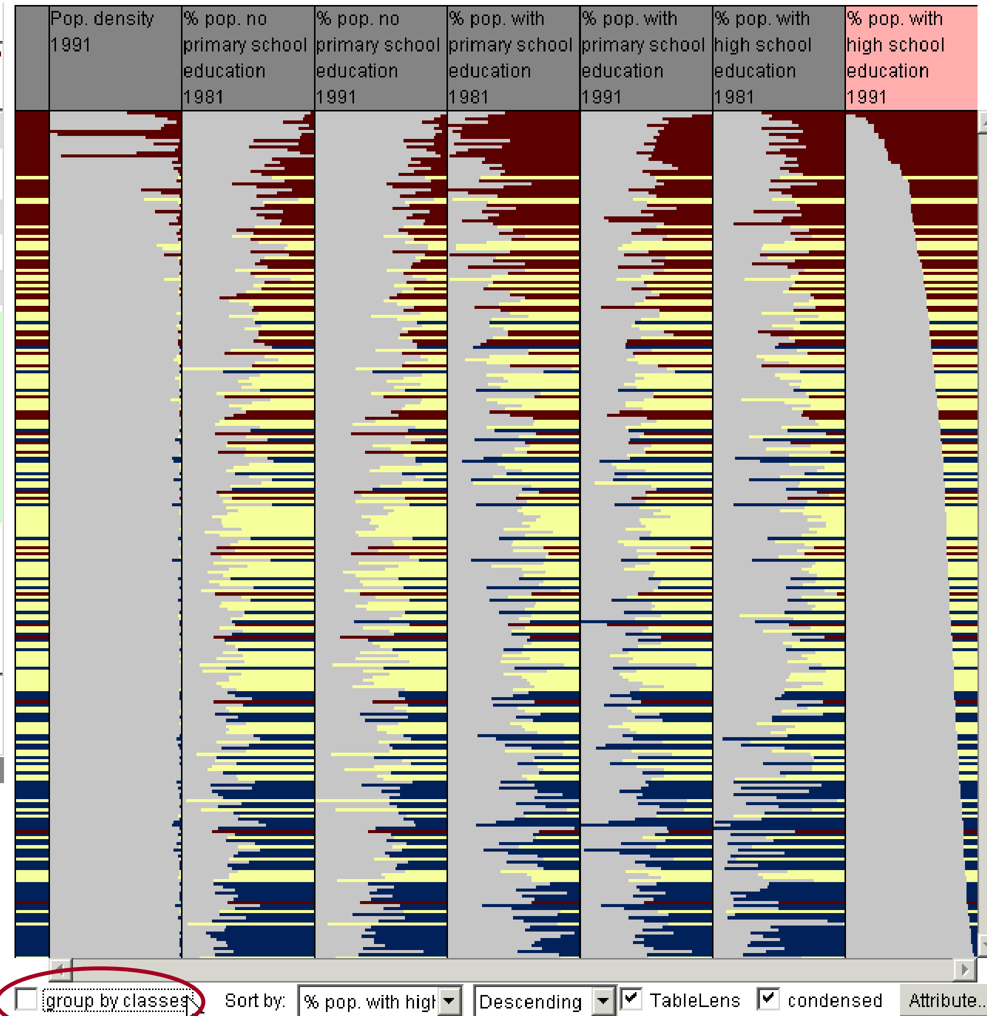
The table rows are grouped by the classes and sorted within each group.

These linked views show us, for example, that the general educational level tends to be higher in districts with high proportion of people employed in services

Class Propagation to Table View (2)



It may also be useful to switch the grouping off



We see that the red rows occur mostly at the top of the table and blue ones at the bottom. Note that the rows are sorted according to % people with high school education in 1991.

CITY UNIVERSITY LONDON
Fraunhofer IAIS

Summary

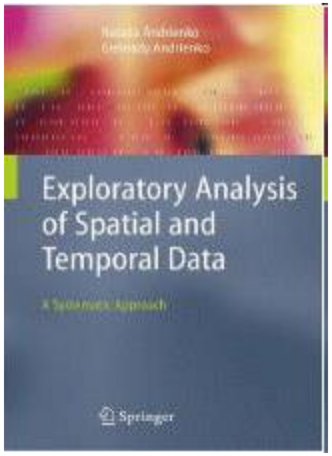
This lecture was supposed to

- introduce the concept of analytical interactive maps
- stress the importance of exploring various aspects of data using multiple views
- demonstrate some types of non-cartographic displays useful in analysis of geodata
- demonstrate various techniques of display linking
- show how to use this in data analysis

See also

- Natalia and Gennady Andrienko
  Exploratory Analysis of Spatial and Temporal Data
  A Systematic Approach
  Springer-Verlag, December 2005

  section 4.8, pp.428-449



Data structure

So far:

- Id, x, y, attribute(s)

Let's add one more special component: **time**

- Id, x, y, **t**, attribute(s)

Overview

Types of Analysis Tasks on Spatio-Temporal Data

Three primary task foci (target information)

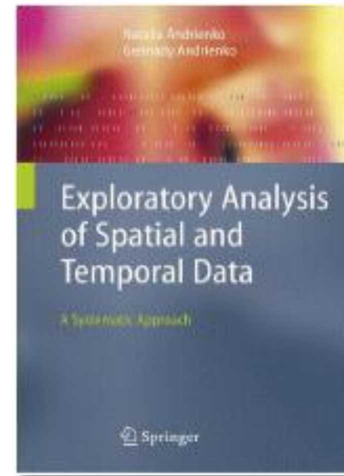
- Focus on *objects*
- Focus on *space*
- Focus on *time*

Two types of task subject

- Characteristics
- Relations

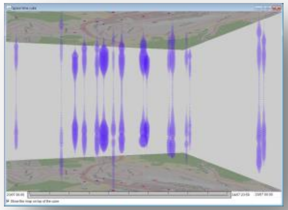
Two levels of analysis

- *Elementary*: focus on one or more elements of a set
- *Synoptic*: focus on a set as a whole, disregard individual elements
- Task may be elementary w.r.t. one subject and synoptic w.r.t. another

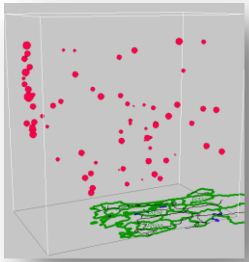


Data Types and Transformations

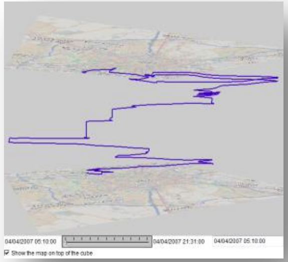
Methods & Techniques for Different Spatio-Temporal Data



Spatial time series



Events



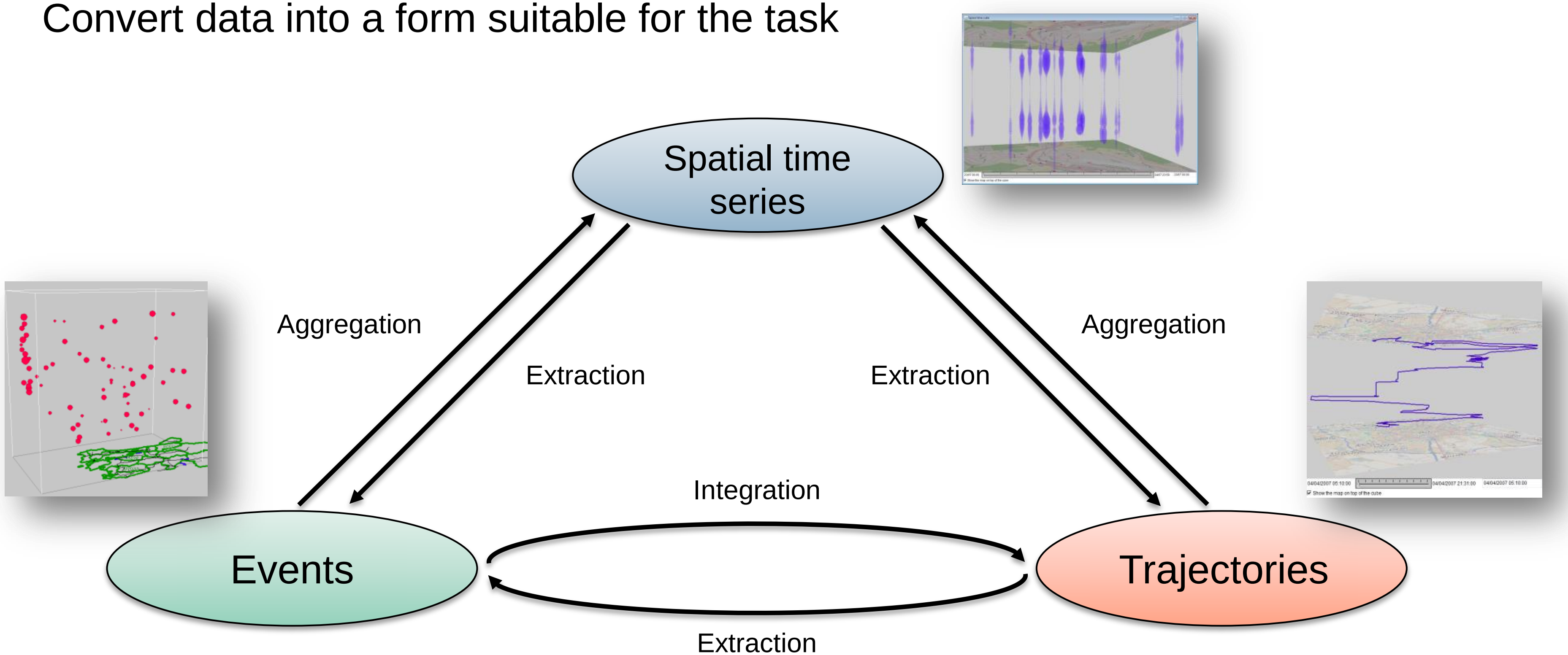
Trajectories

Data Types and Transformations

Methods & Techniques for Different Spatio-Temporal Data

Representation forms offer different conceptual views on ST data

Convert data into a form suitable for the task



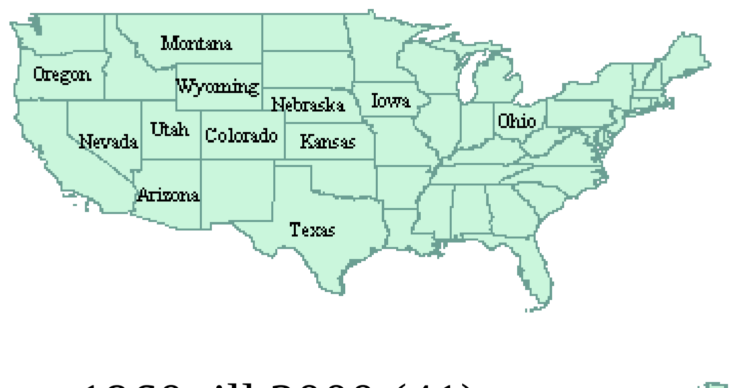
CITY UNIVERSITY LONDON

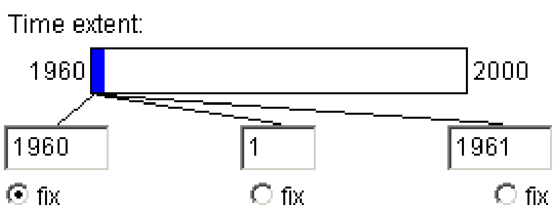
Fraunhofer IAIS

Spatial Time Series

Data structure

Spatial references: states of the USA





Temporal references: years from 1960 till 2000 (41)

Attributes: population + various crime rates

Population	Index offense rate	Violent Crime rate	Murder and nonnegligent manslaughter rate	Forcible rape rate	Robbery rate	Aggravated assault rate	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate

CITY UNIVERSITY LONDON
Fraunhofer IAIS

High level analysis questions

Spatial Time Series (STS)

How are attribute values distributed over the territory at a given time moment?

How do the attribute values at a given place vary over time?

How does the overall spatial pattern of value distribution evolve over time?

How are different behaviour patterns distributed over the territory? Are there spatial clusters of similar behaviours?

Spatial Time Series

Methods & Techniques for Different Spatio-Temporal Data I

Visualization methods

- Animated maps
- „Layman techniques”: (animated) charts embedded in maps (bar charts, pie charts, …)
- “Small multiples” map displays
- Time Graphs and their transformations

Spatial Time Series

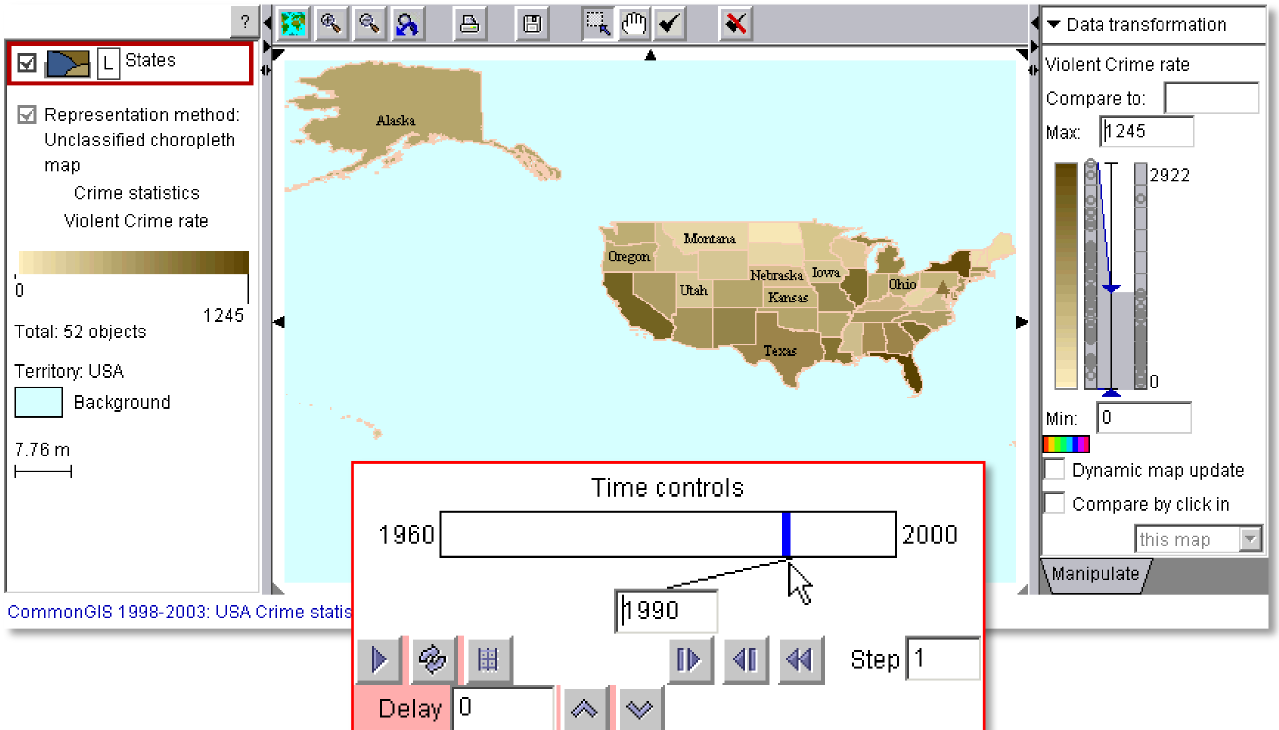
Methods & Techniques for Different Spatio-Temporal Data I

Visualization methods

- Animated maps
- “Small multiples” map displays
- „Layman techniques”: (animated) charts embedded in maps (bar charts, pie charts, …)
- Time Graphs and their transformations

Time Map

Spatial Time Series: Basic Visualization Methods I



For a time map, one can use any representation method suitable for static data. Choropleth maps are good for exploring *spatial distribution patterns.*

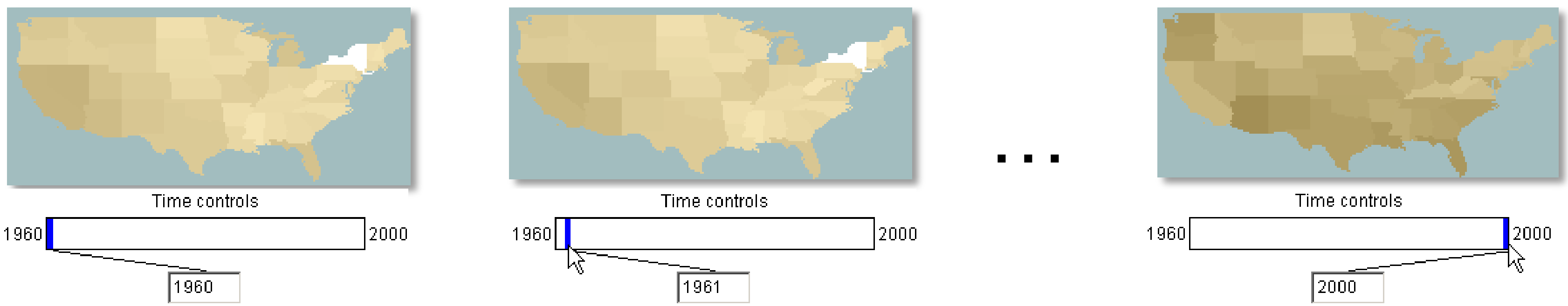
Time-dependent data may be represented on a time map, which is manipulated through time controls and, in particular, allows animation

CITY UNIVERSITY LONDON

Fraunhofer IAIS

When time map is useful

Spatial Time Series: Basic Visualization Methods I



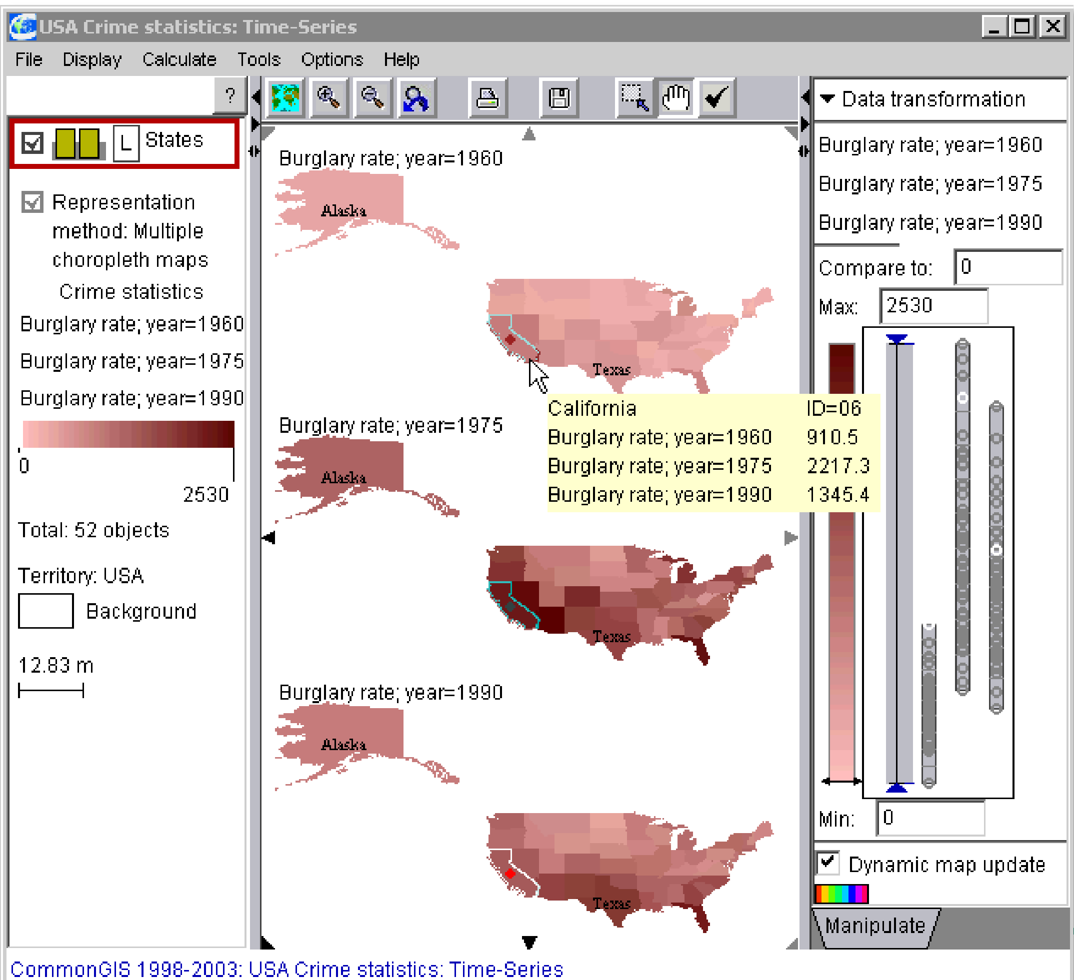
How did the spatial distribution pattern develop over time?

At moment *t*, how were the values distributed over the whole space?

At what time moment were the values distributed over the space in the given manner?

Map Series

Spatial Time Series: Basic Visualization Methods I



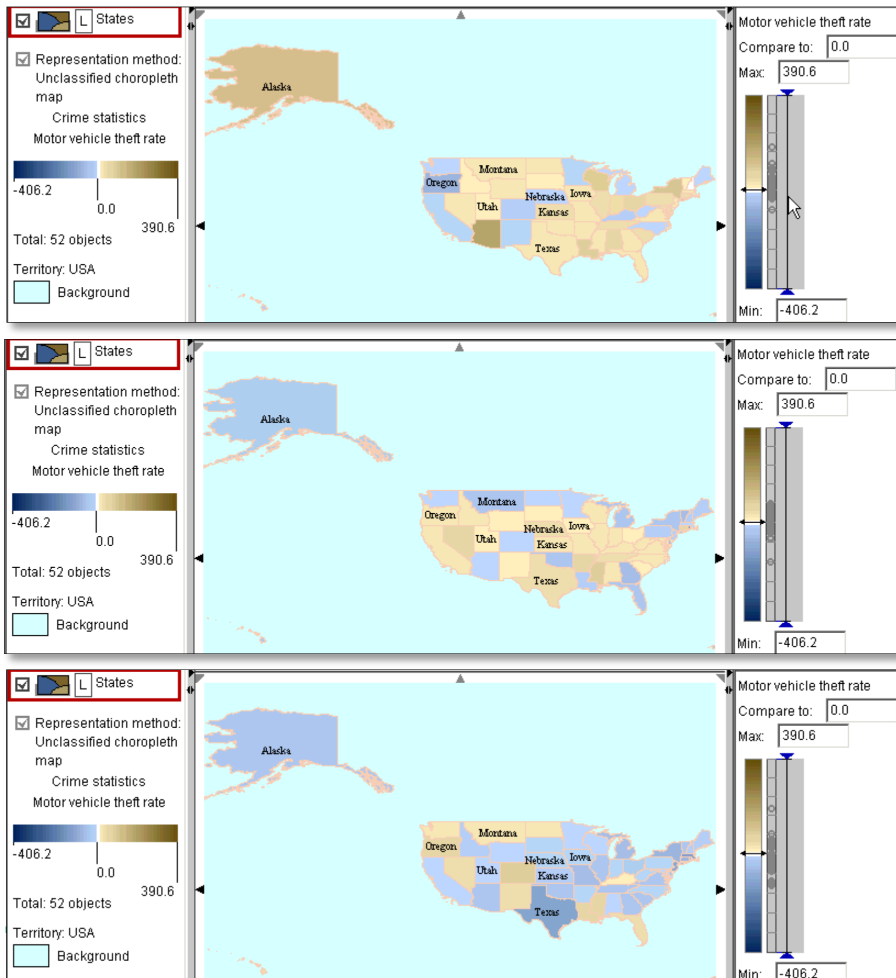
To compare the spatial distributions of attribute values at two or more time moments, we need to see these distributions simultaneously. Best of all is to use multiple maps displayed in a common panel and manipulated through a common set of controls.

CITY UNIVERSITY LONDON

Fraunhofer IAIS

Exploring the Distribution of Changes

Spatial Time Series: Data Transformations



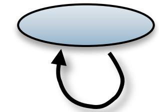
Instead of original attribute values, a time map or map series can represent changes, that is, differences or ratios to the previous moment or to any selected moment

Here the maps correspond to years 1990, 1991, and 1992 and represent differences to the previous years. Positive differences (i.e. increased values) are shown in brown and negative differences (i.e. decreased values) in blue

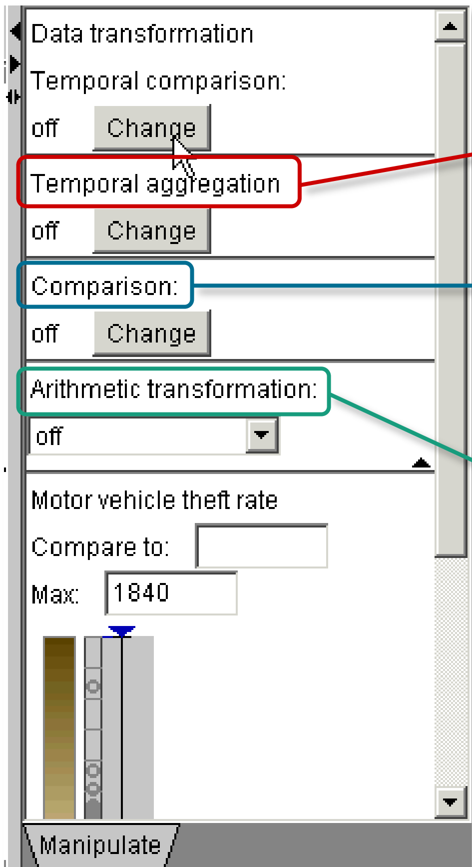
CITY UNIVERSITY LONDON

Fraunhofer IAIS

Map Series: Useful Transformations



Spatial Time Series: Data Transformations



Temporal smoothing and computing of residuals

For each time moment computes differences or ratios to a particular object or value or to the mean or median among all objects at this moment

Value scale transformation, e.g. logarithmic

Comparison to country's median

Spatial Time Series: Data Transformations

Build map series with transformed data:
relative difference to median value

Spatial distribution patterns of attribute values may become more vivid

Value and patterns evolution over time can be seen more easily

Spatial Time Series

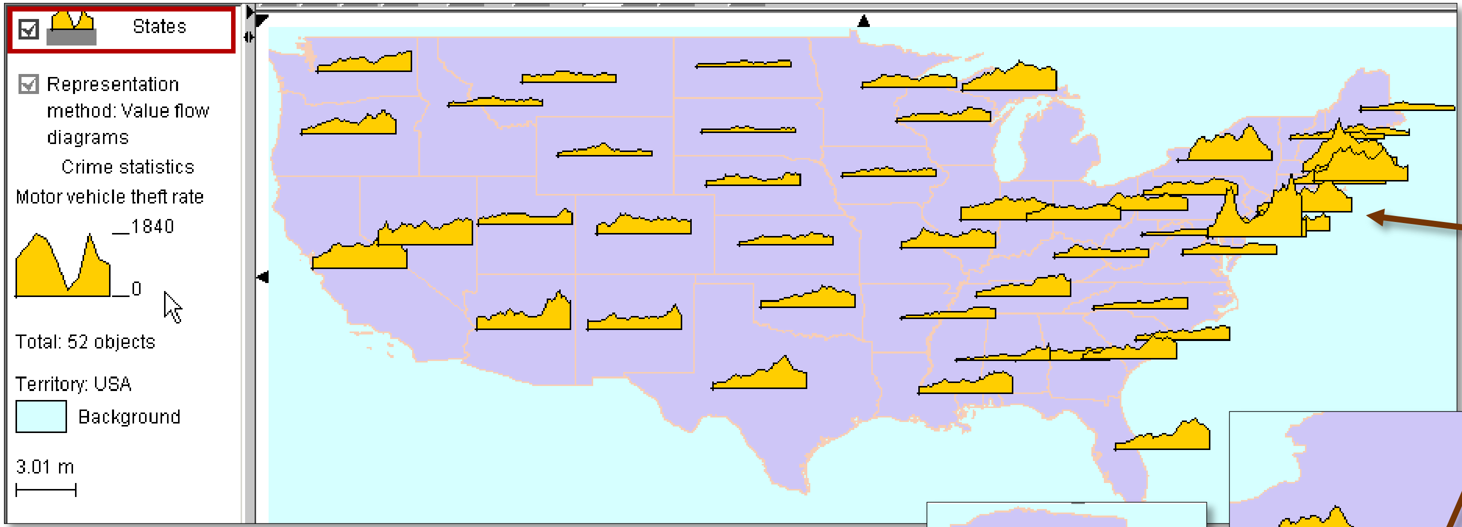
Methods & Techniques for Different Spatio-Temporal Data I

Visualization methods

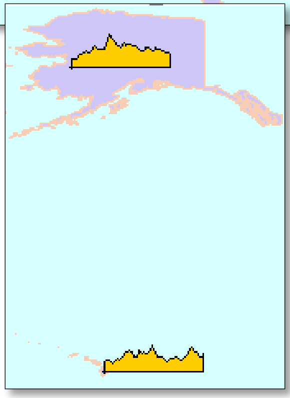
- Animated maps
- „Layman techniques": (animated) charts embedded in maps (bar charts, pie charts, …)
- "Small multiples" map displays
- Time Graphs and their transformations

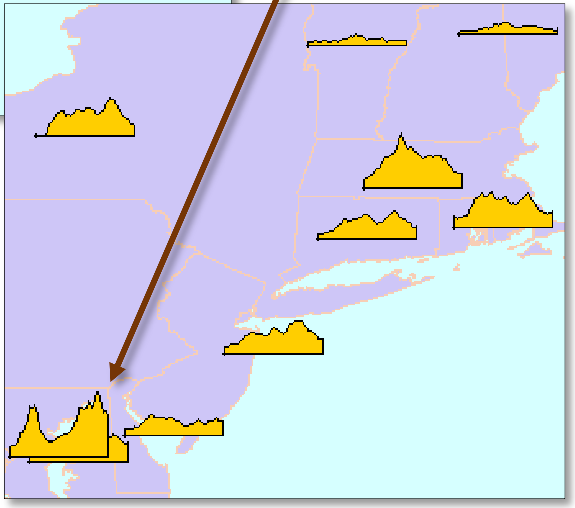
Value Flow Map

Spatial Time Series: Basic Visualization Methods II



Unfortunately, symbol overlapping creates significant inconveniences, and zooming does not always help

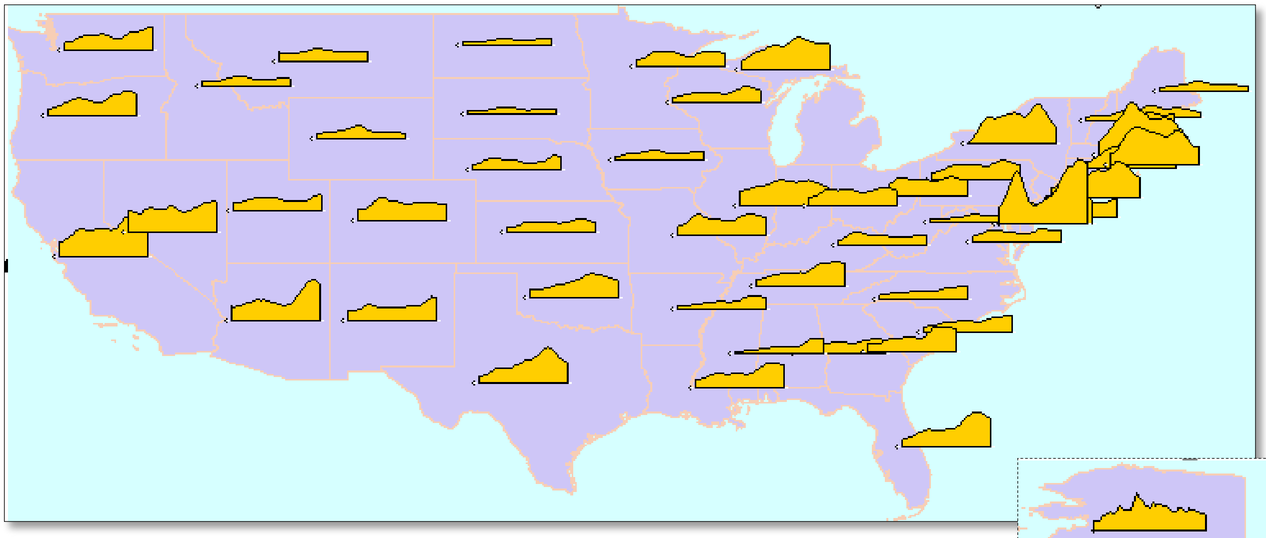




The value flow symbols show us the evolution of attribute values (*temporal behavior*) at each location.

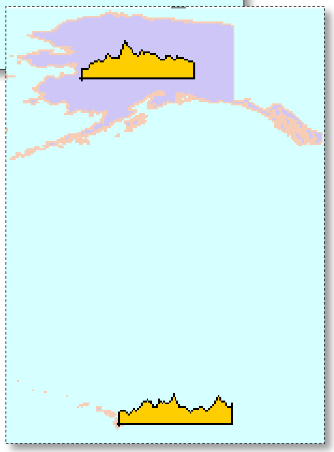
Value Flow Map (2)

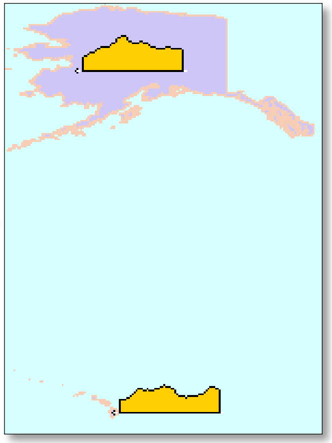
Spatial Time Series: Basic Visualization Methods II



Temporal smoothing allows to disregard small fluctuations and see overall trends.

Here the values for each year have been replaced by 5-year means. You can compare to the previous variant and see the effect of the smoothing.



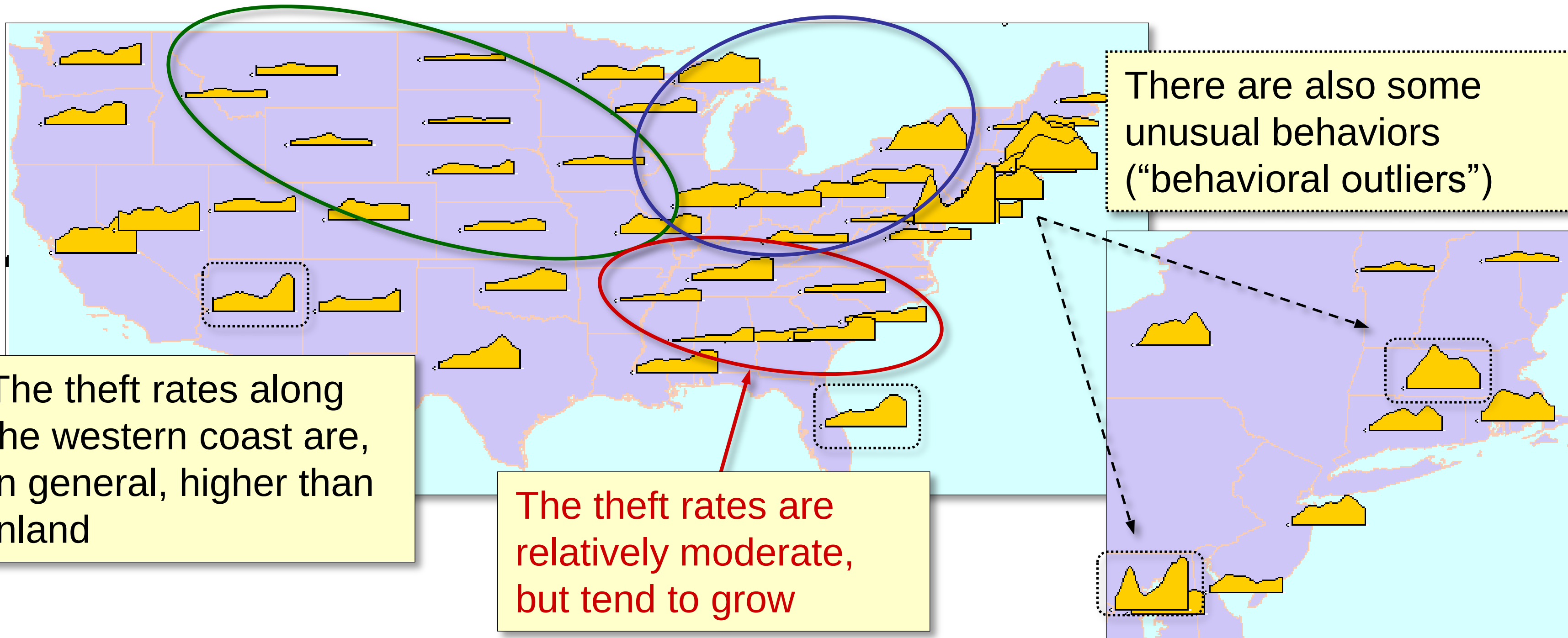


Value Flow Map (2)

Spatial Time Series: Basic Visualization Methods II

This appears to be a spatial cluster of similar behaviors

Around the Great Lakes, the theft rates were high, but tended to decrease in last years



There are also some unusual behaviors ("behavioral outliers")

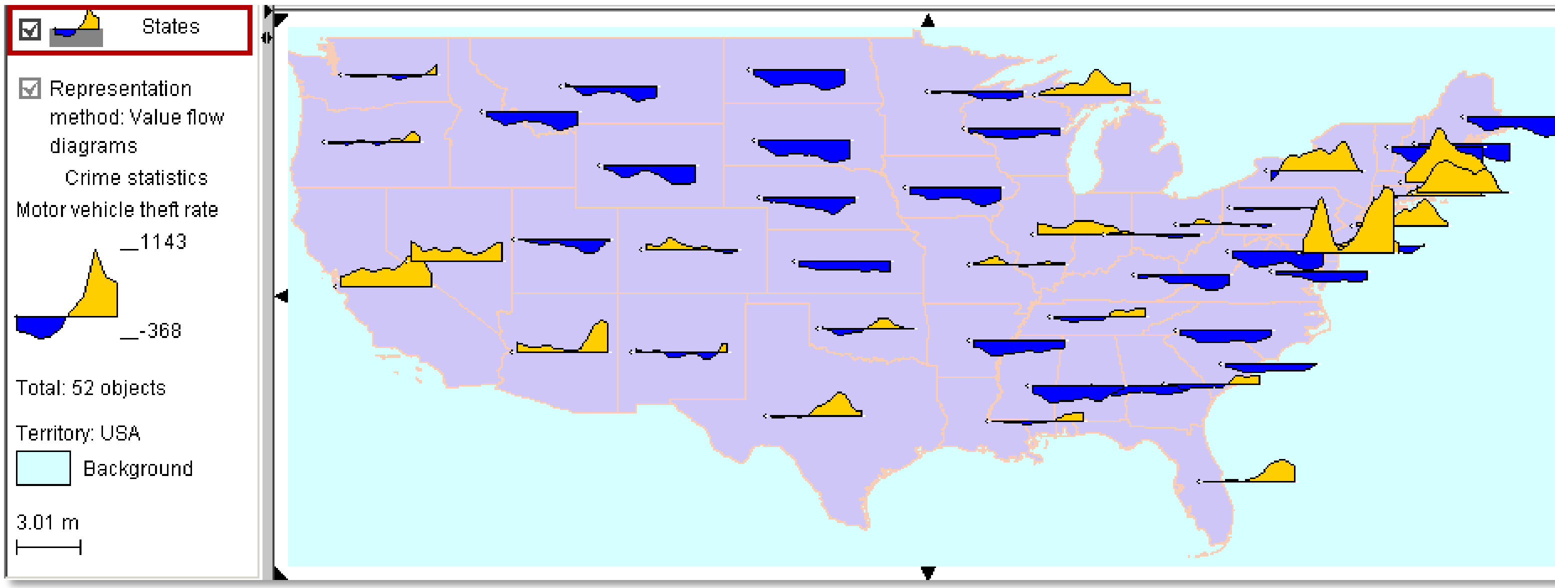
The theft rates along the western coast are, in general, higher than inland

The theft rates are relatively moderate, but tend to grow

Temporal Behavior Exploration

Spatial Time Series: Data Transformations

As with time maps, various transformations can be applied to value flow maps

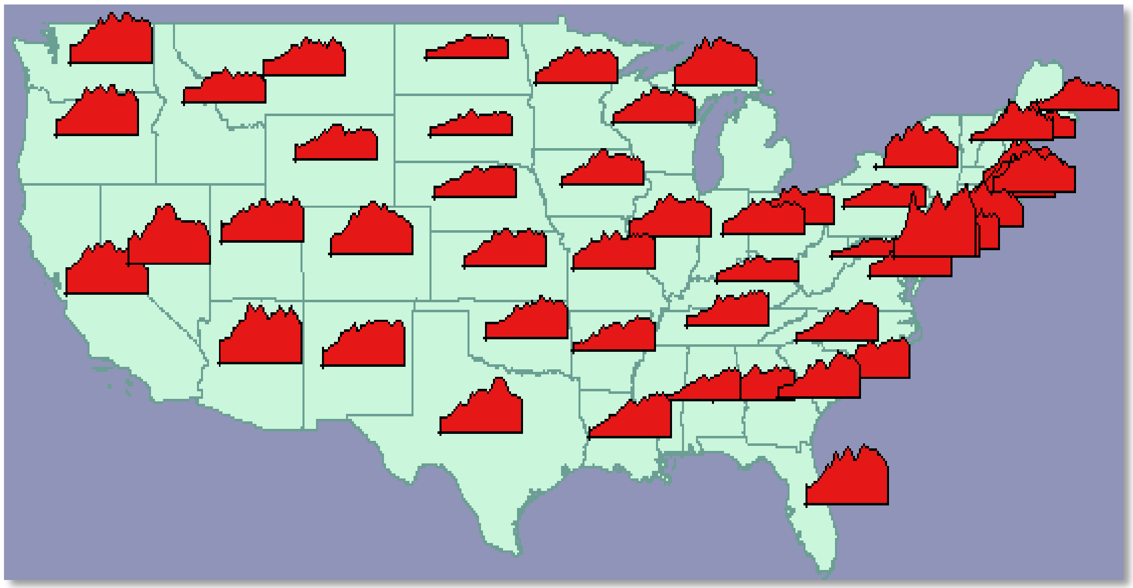


Here: comparison to each country's mean

- For every time moment, each state's values is replaced by its difference to the country's overall mean value at that moment
- Yellow color corresponds to positive differences, and blue – negative

Value Flow Map Disadvantages

Spatial Time Series: Basic Visualization Methods II



- ✓ seeing the temporal behaviors in their spatial context
- × seeing all behaviors at once
- × detecting behaviors with particular features
- × noticing what sorts of features exist in the data

The diagrams are perceived as separate entities → the map must be scanned and cannot be grasped as a single image

Absence of ordering complicates seeking for specific behaviour patterns

Diagram overlapping is a serious problem

Spatial Time Series

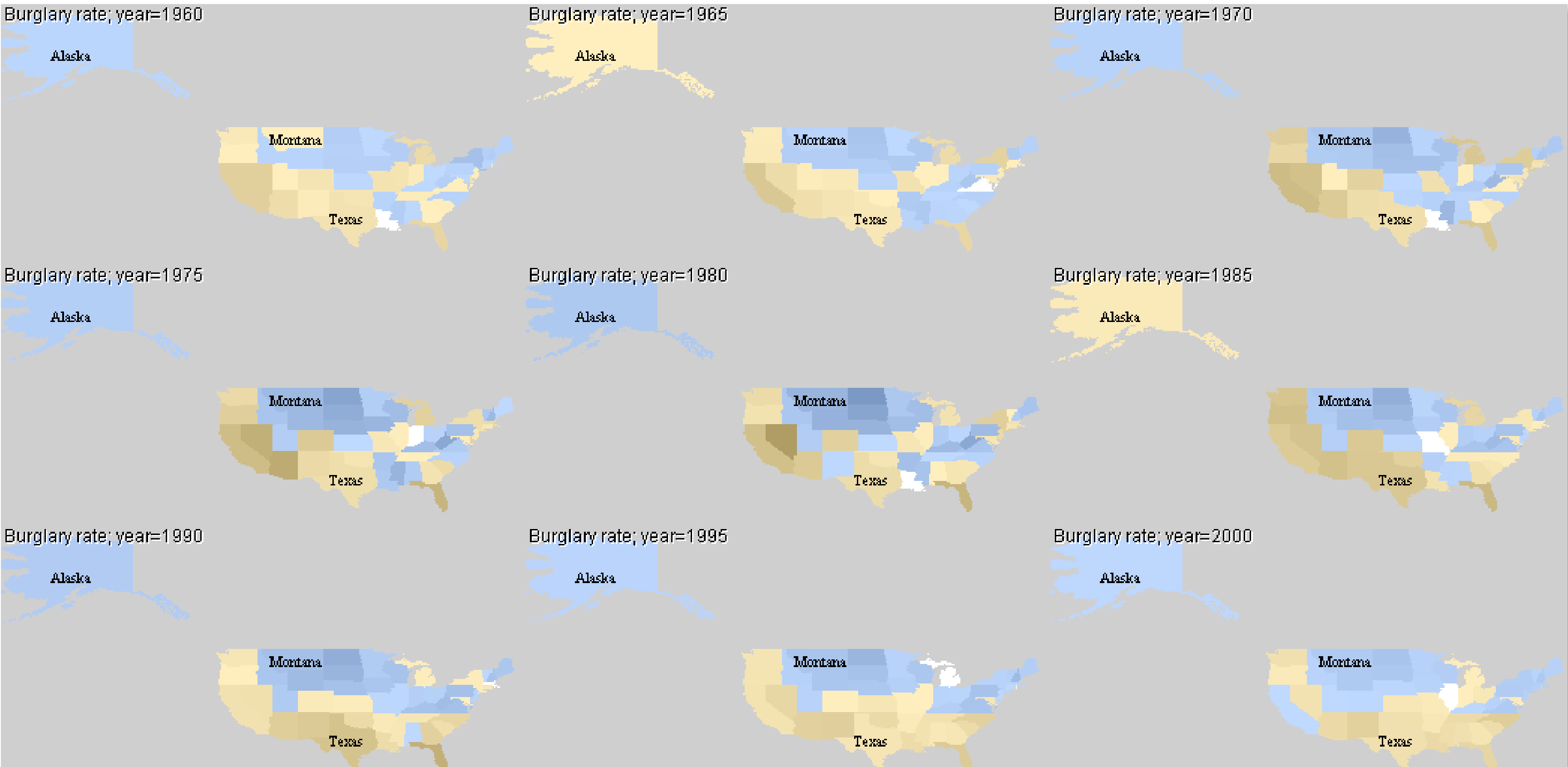
Methods & Techniques for Different Spatio-Temporal Data I

Visualization methods

- Animated maps
- „Layman techniques”: (animated) charts embedded in maps (bar charts, pie charts, …)
- “Small multiples” map displays
- Time Graphs and their transformations

Comparison to country’s median

Spatial Time Series: Data Transformations

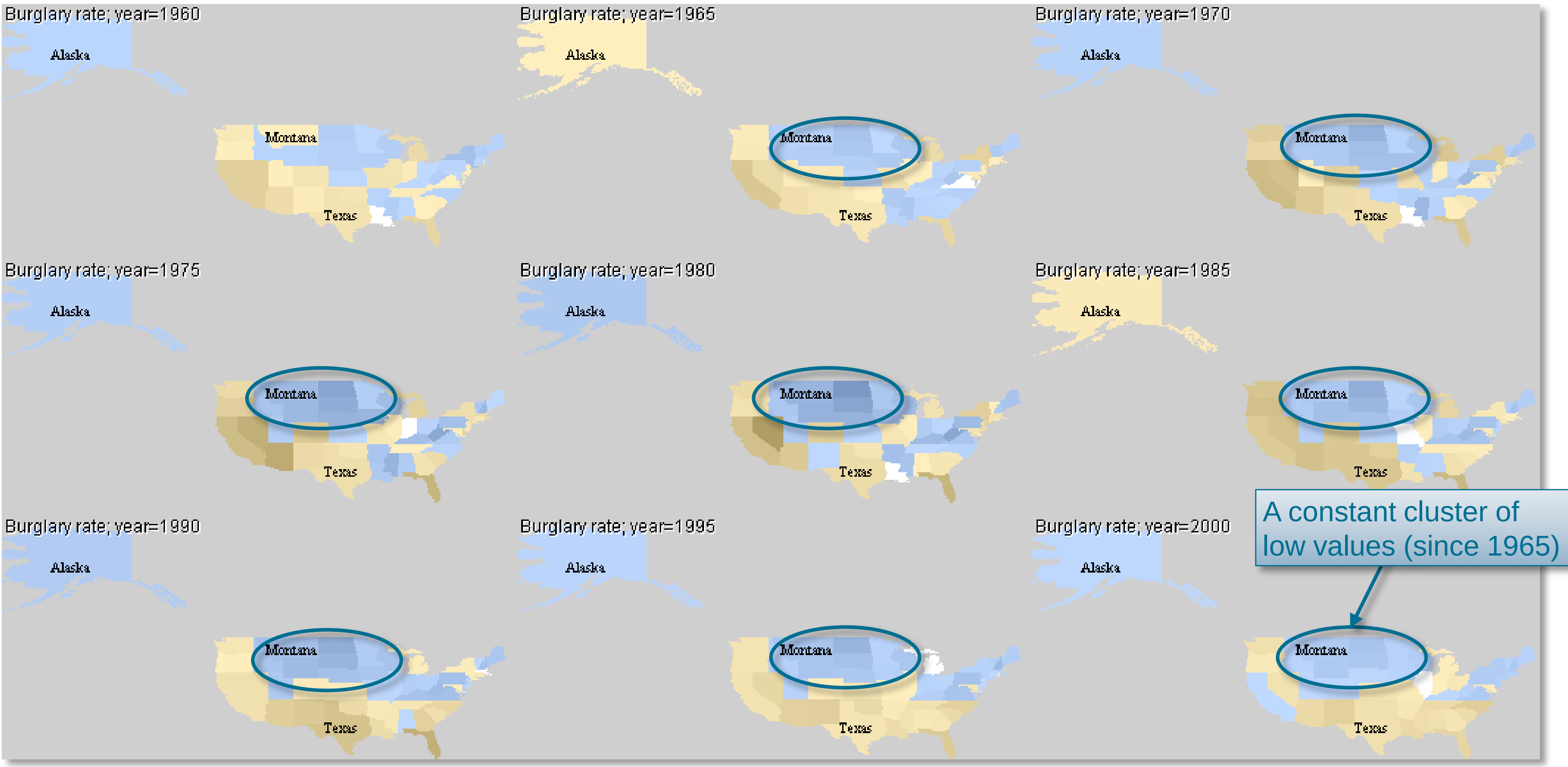


CITY UNIVERSITY LONDON

Fraunhofer IAIS

Comparison to country's median

Spatial Time Series: Data Transformations

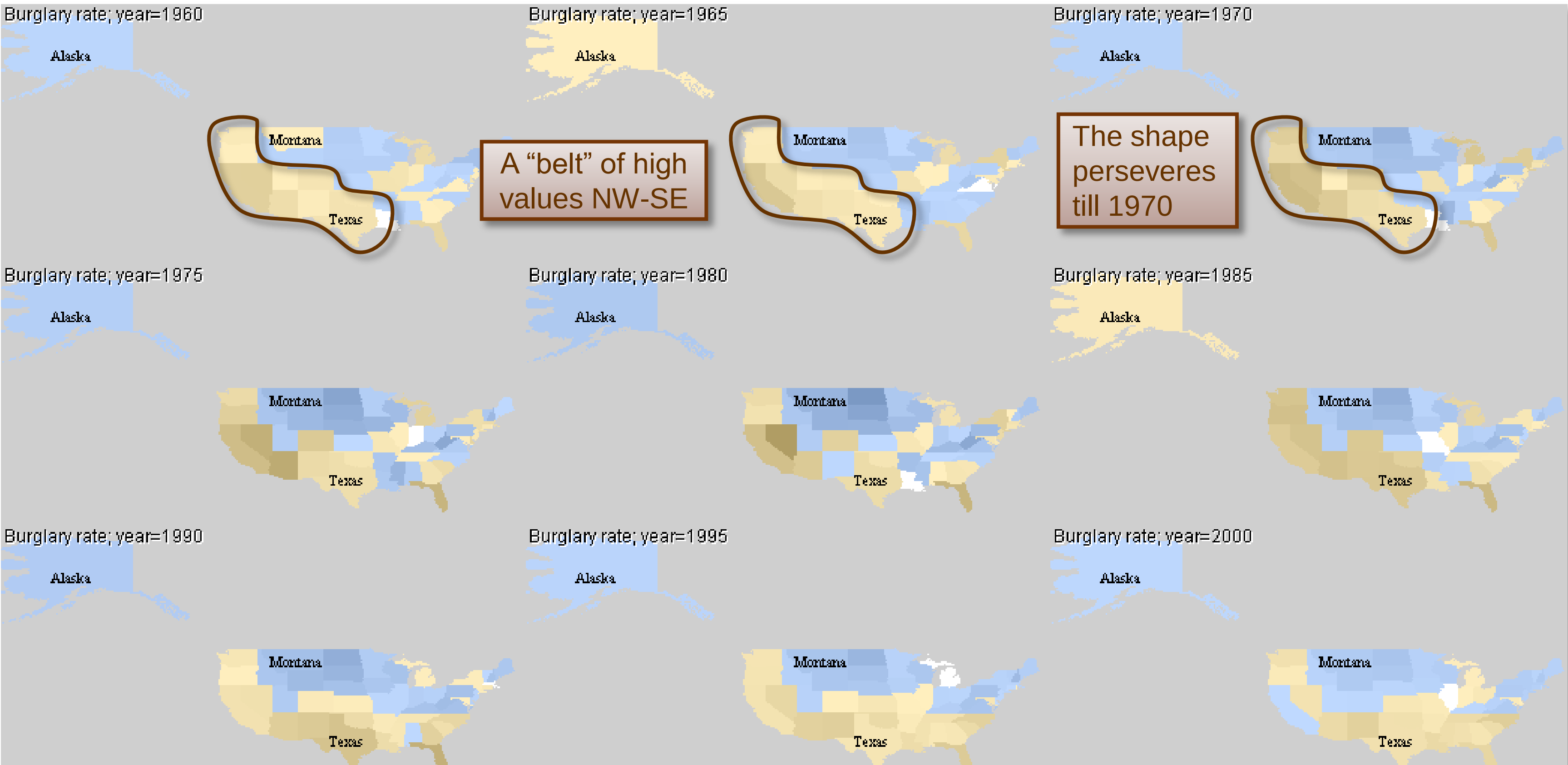


CITY UNIVERSITY LONDON

Fraunhofer IAIS

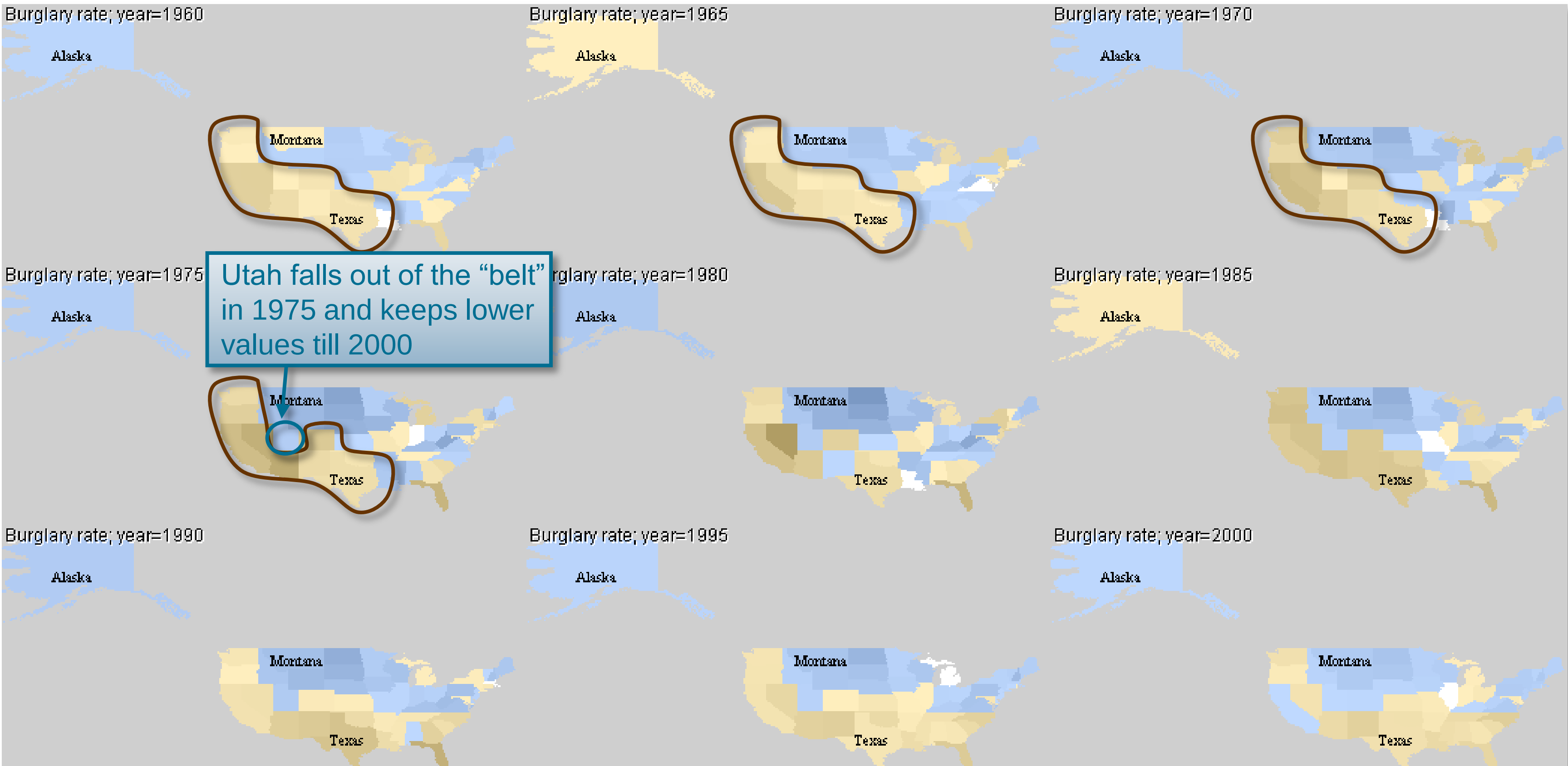
Comparison to country's median

Spatial Time Series: Data Transformations



Comparison to country's median

Spatial Time Series: Data Transformations

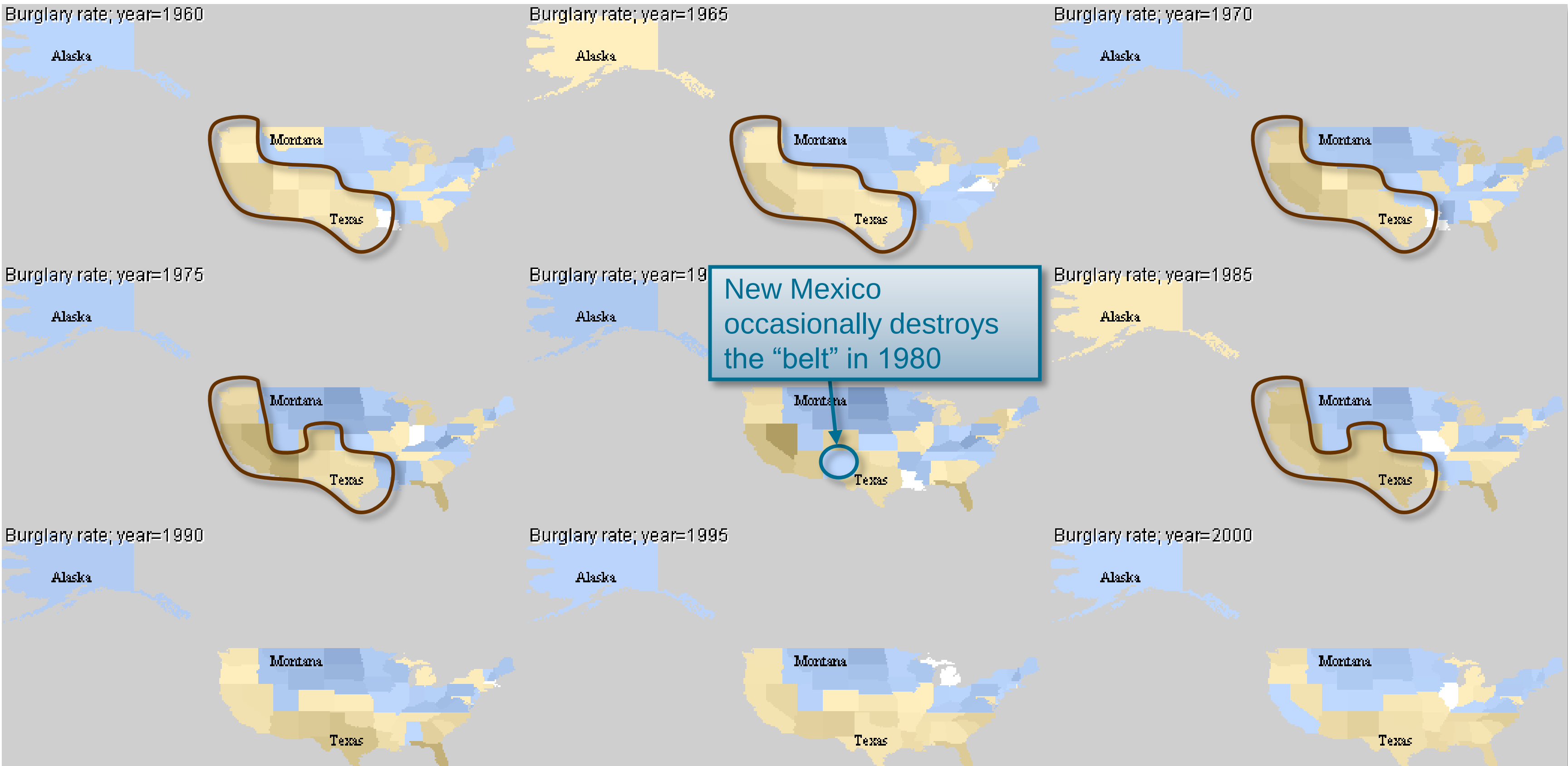


CITY UNIVERSITY LONDON

Fraunhofer IAIS

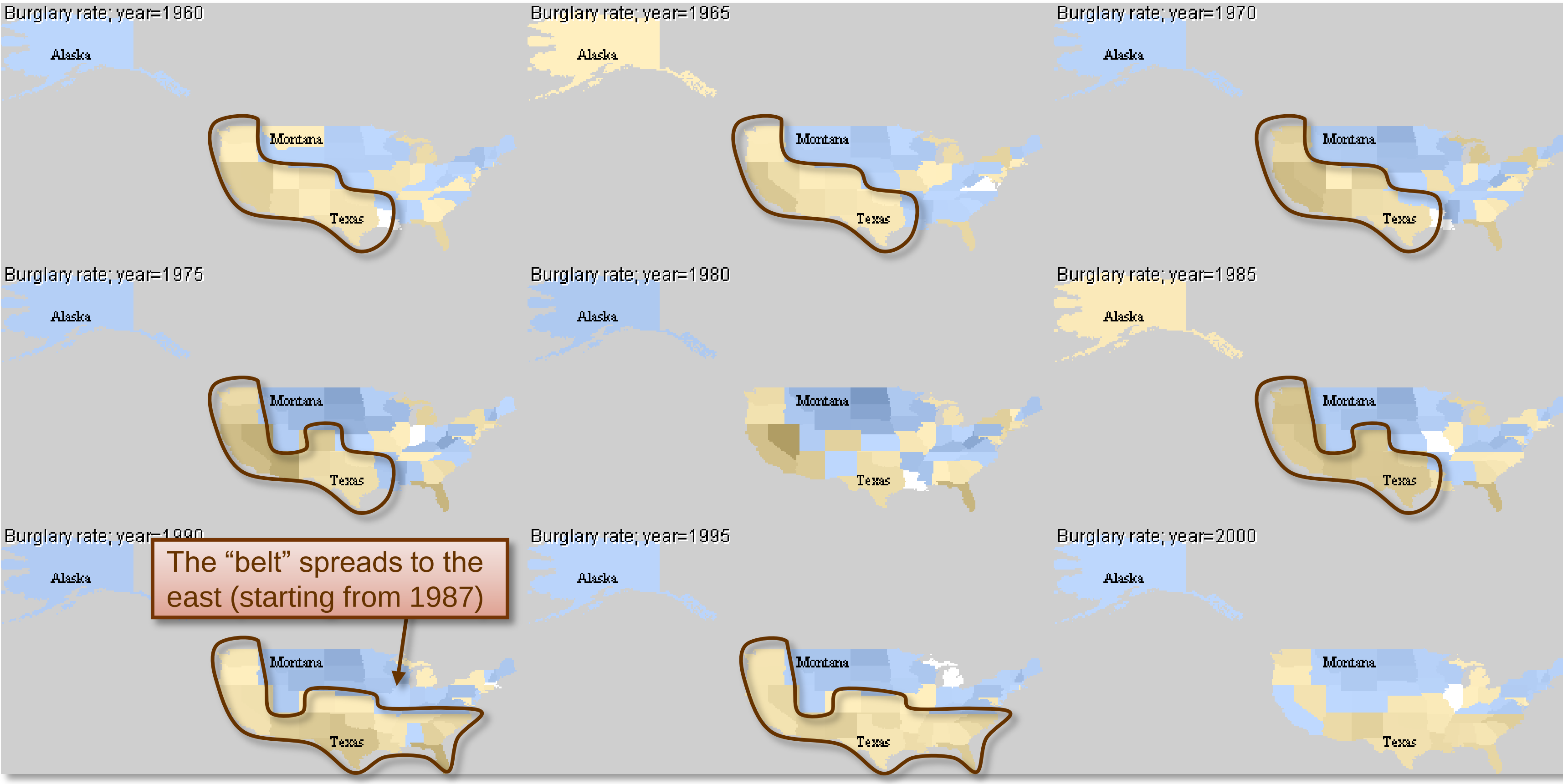
Comparison to country’s median

Spatial Time Series: Data Transformations



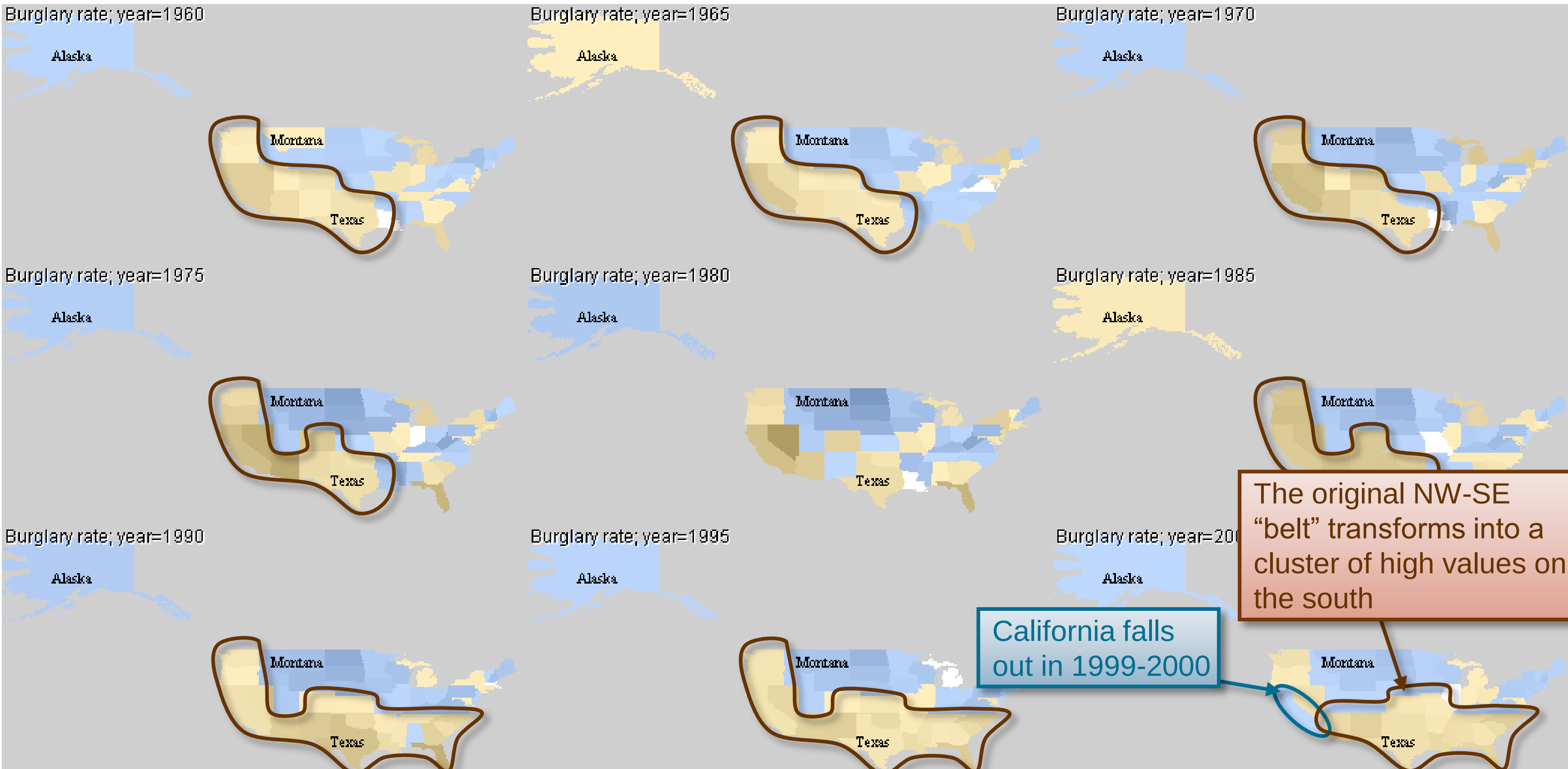
Comparison to country's median

Spatial Time Series: Data Transformations



Comparison to country's median

Spatial Time Series: Data Transformations

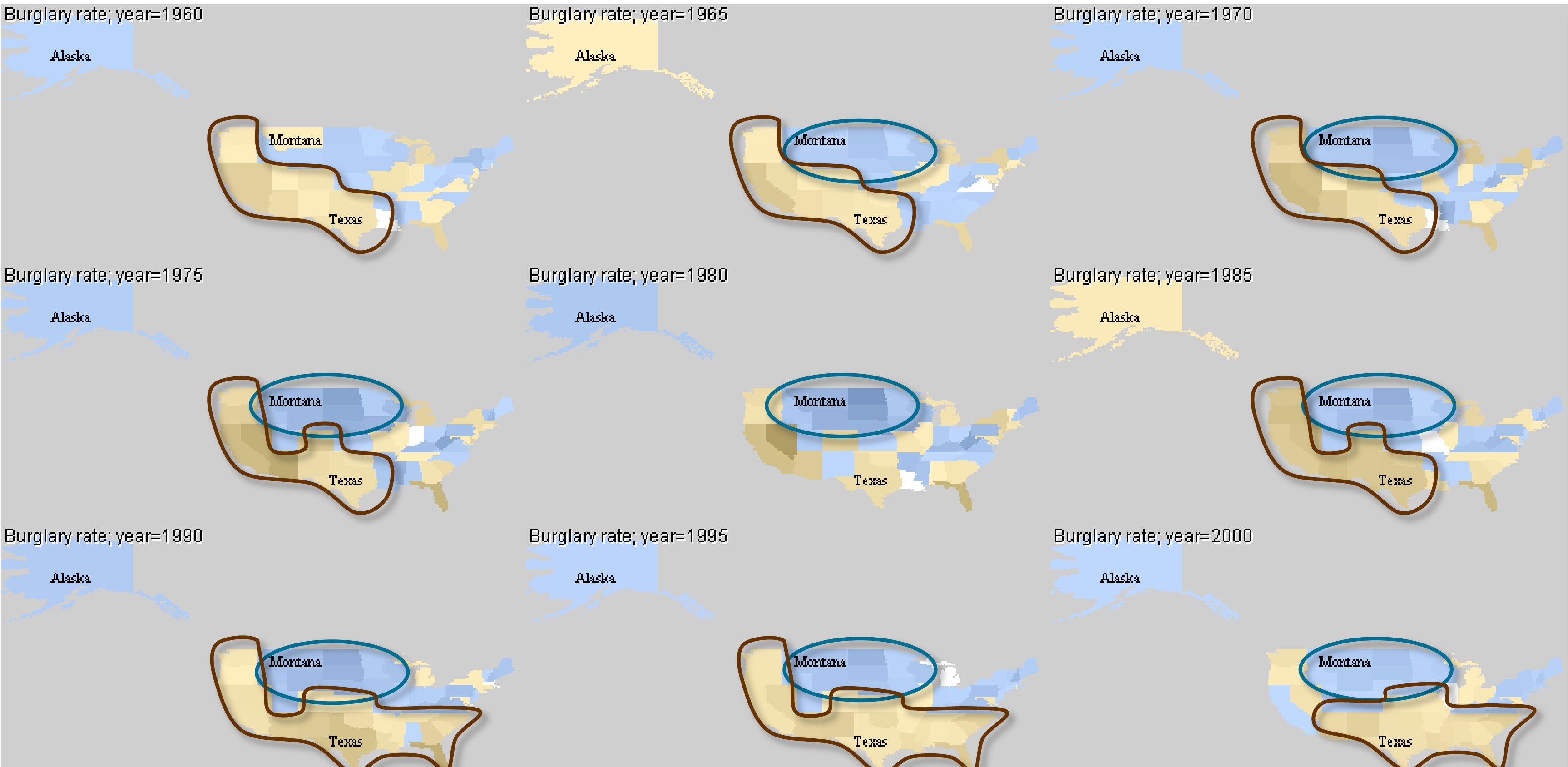


CITY UNIVERSITY LONDON

Fraunhofer IAIS

Comparison to country's median

Spatial Time Series: Data Transformations



CITY UNIVERSITY LONDON

Fraunhofer IAIS

Spatial Time Series

Methods & Techniques for Different Spatio-Temporal Data I

Visualization methods

- Animated maps
- „Layman techniques": (animated) charts embedded in maps (bar charts, pie charts, …)
- "Small multiples" map displays
- Time Graphs and their transformations

Time Graph

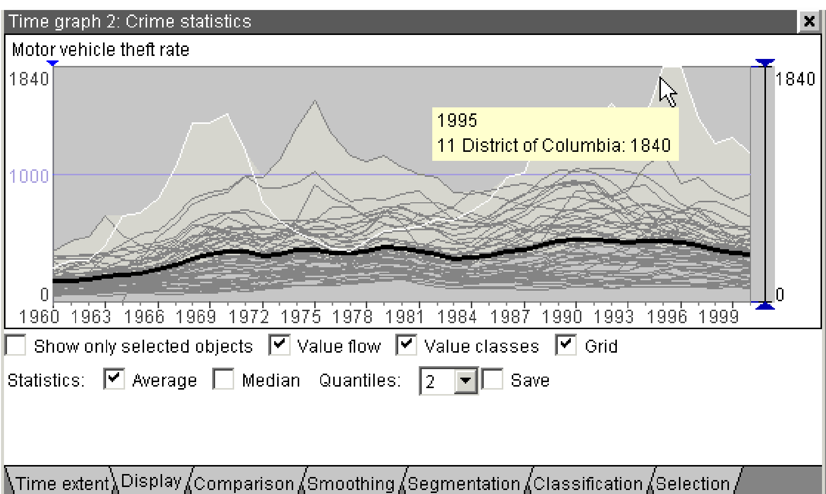
Spatial Time Series: Basic Visualization Methods III

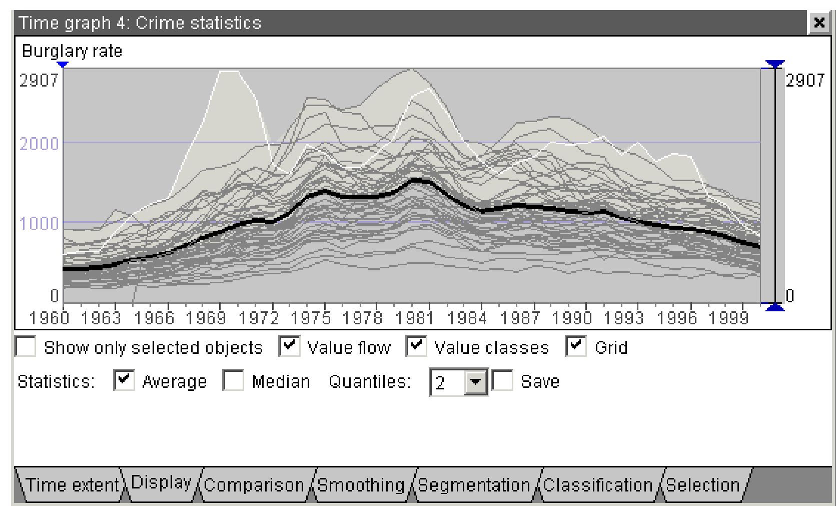
Putting all behaviors together makes their comparison more convenient

- Numeric details on-demand by pointing on an object's line

Plus, to better understand general development trends :

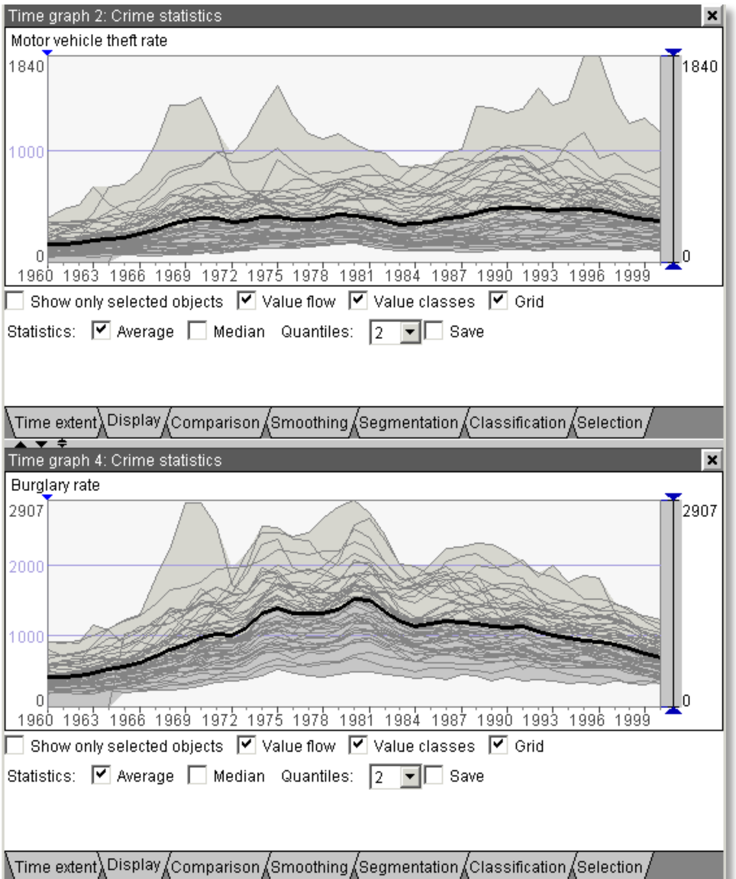
- "mean behavior line" connecting means of each year
- "median behavior" connecting the year medians





Time Graph: Multiple View Comparisons

Spatial Time Series: Basic Visualization Methods III

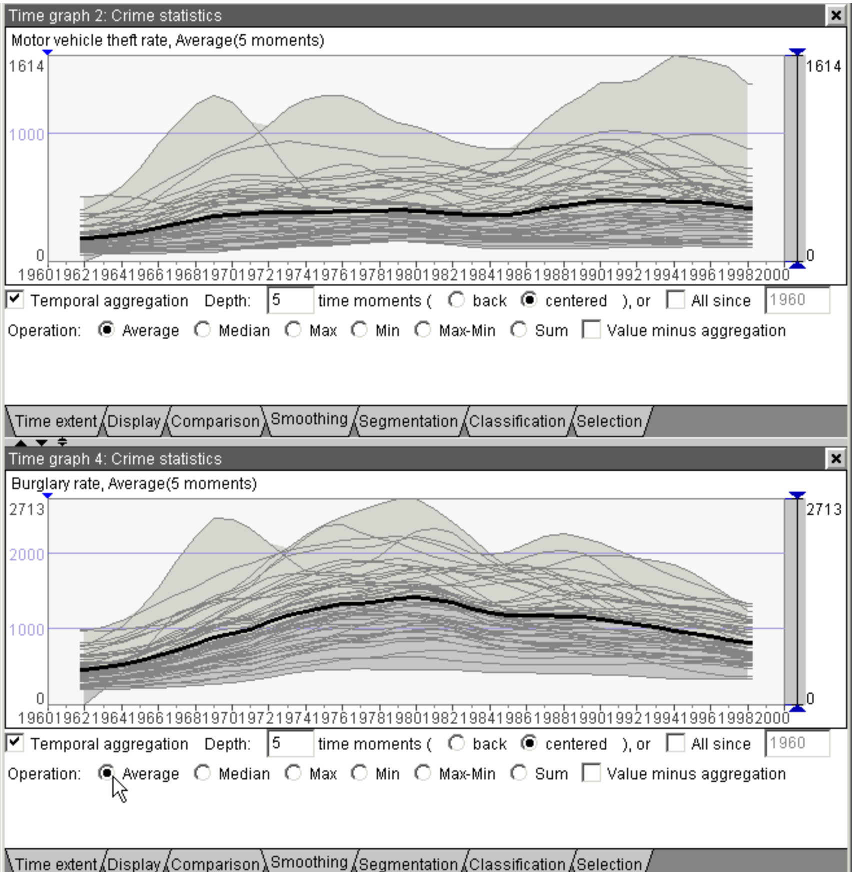


Juxtaposed time graphs are suitable for comparing trends and temporal variations of two or more attributes

This example shows that attributes "Motor vehicle theft rate" and "Burglary rate" have quite different trends of general development

Time Graph: Multiple View Comparisons

Spatial Time Series: Basic Visualization Methods III



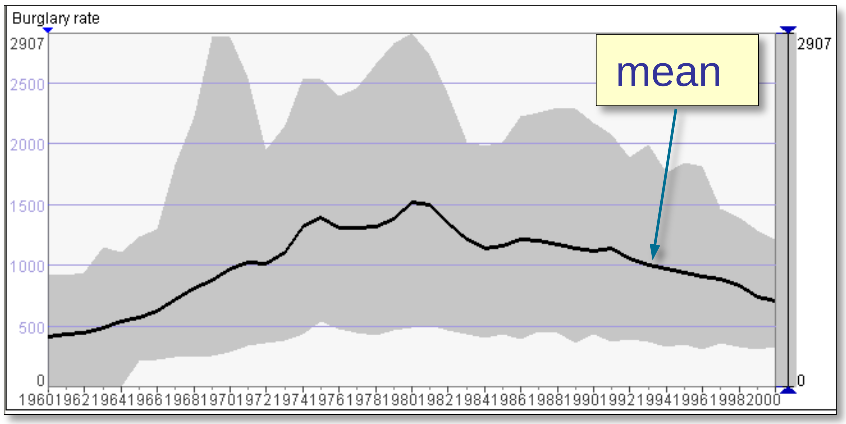
Additionally employ <u>smoothing</u> (value averaging over intervals)

Mitigates small fluctuations

Exposes trends more clearly

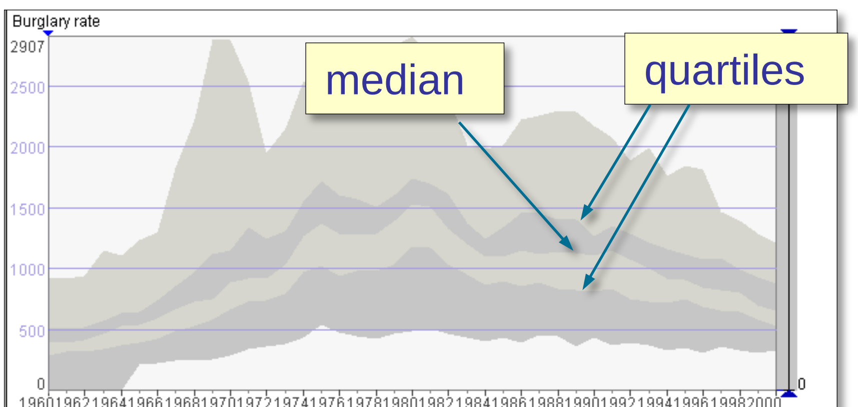
Time Graph: Level of Detail

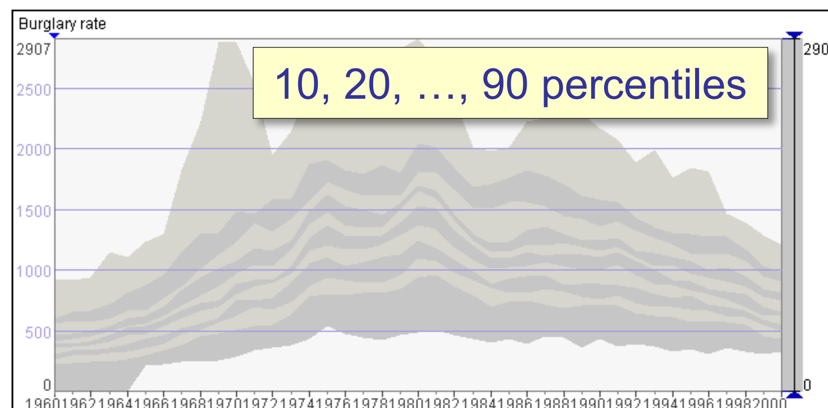
Spatial Time Series: Basic Visualization Methods III



Mean and median lines only give very coarse picture of the general value variation properties

For a finer analysis, may also look at the quartiles or even smaller percentiles.

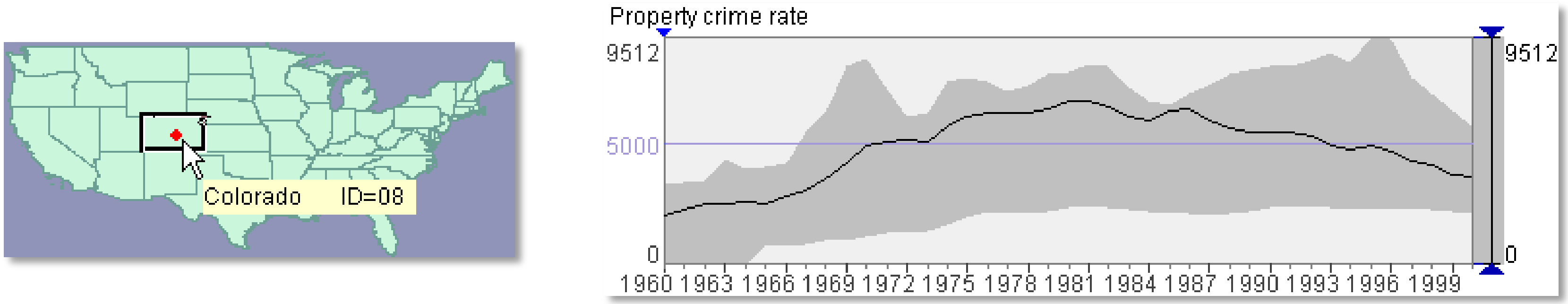




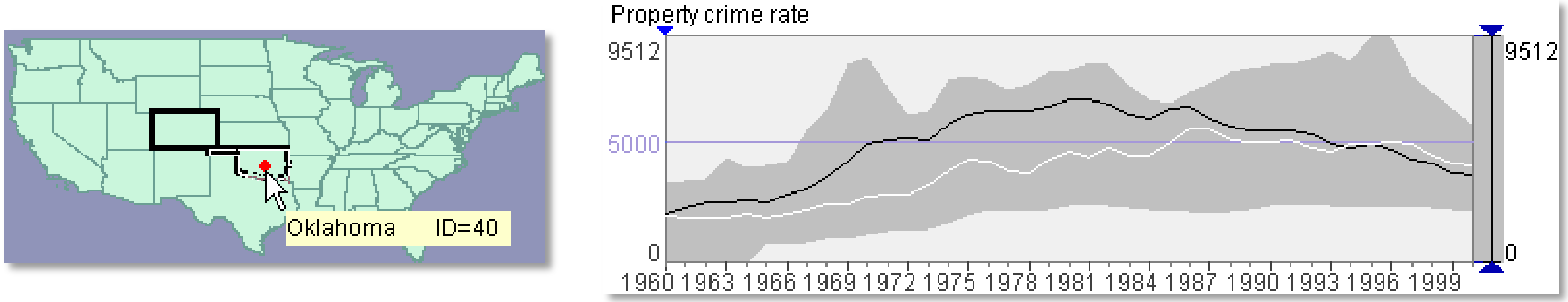
Linked Views: Map + Time Graph

Spatial Time Series: Basic Visualization Methods IV

At place L_1, how did the values behave over the entire time period?

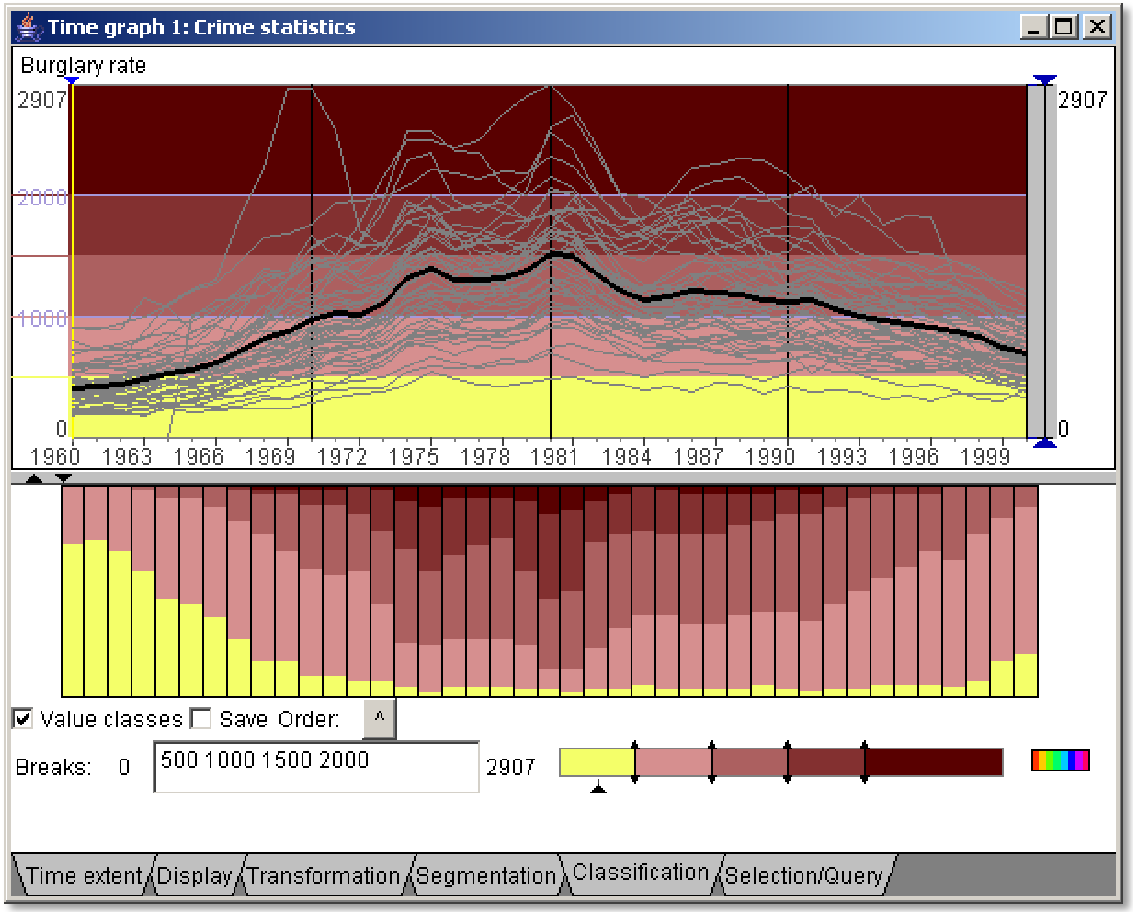


Compare the temporal behavior at places L_1 and L_2



Time Graph++: Time Histogram

Spatial Time Series: Basic Visualization Methods V



Divide value range of the attribute into intervals

Choose specific color/ shade for each interval

Size of colored segments encodes relative frequencies of values from corresponding interval, for each time moment

This example:

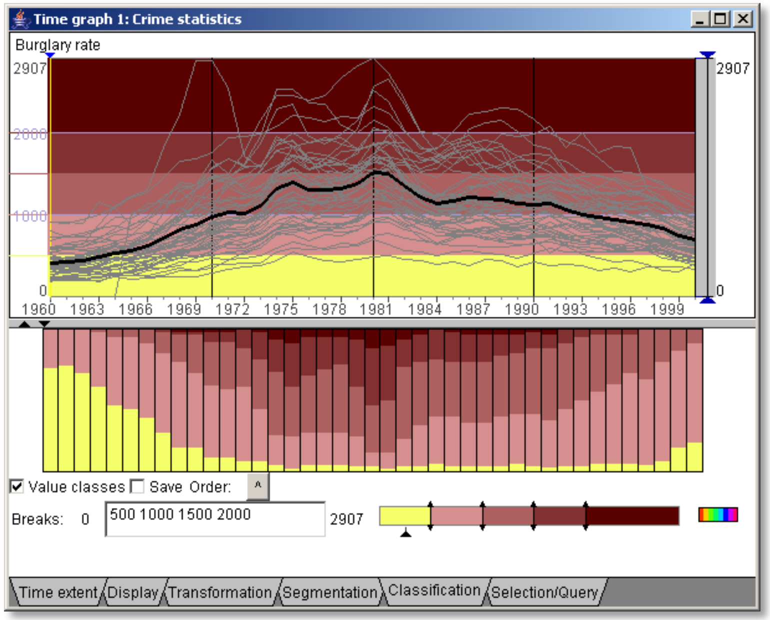
- Each bar: one year
- Shows increase of the crime rates over the country in 70's & early 80's
- Note two peaks in 1975 and 1981-82, followed by gradual decrease

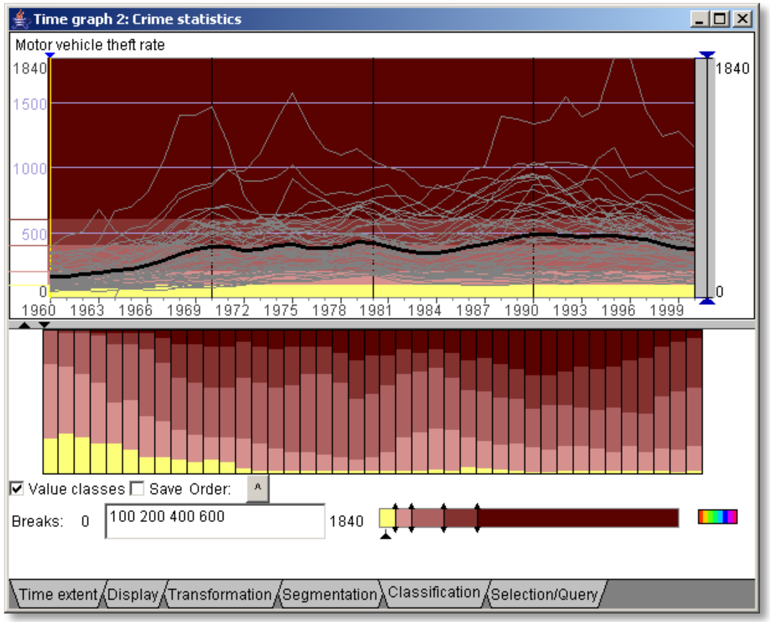
Time Graph++: Time Histogram

Spatial Time Series: Basic Visualization Methods V

Time histograms facilitate comparison of 2 (or more) attributes

- Compare temporal trends *despite differences in value ranges*





- Here: attributes “Burglary rate” and “Motor vehicle theft rate”

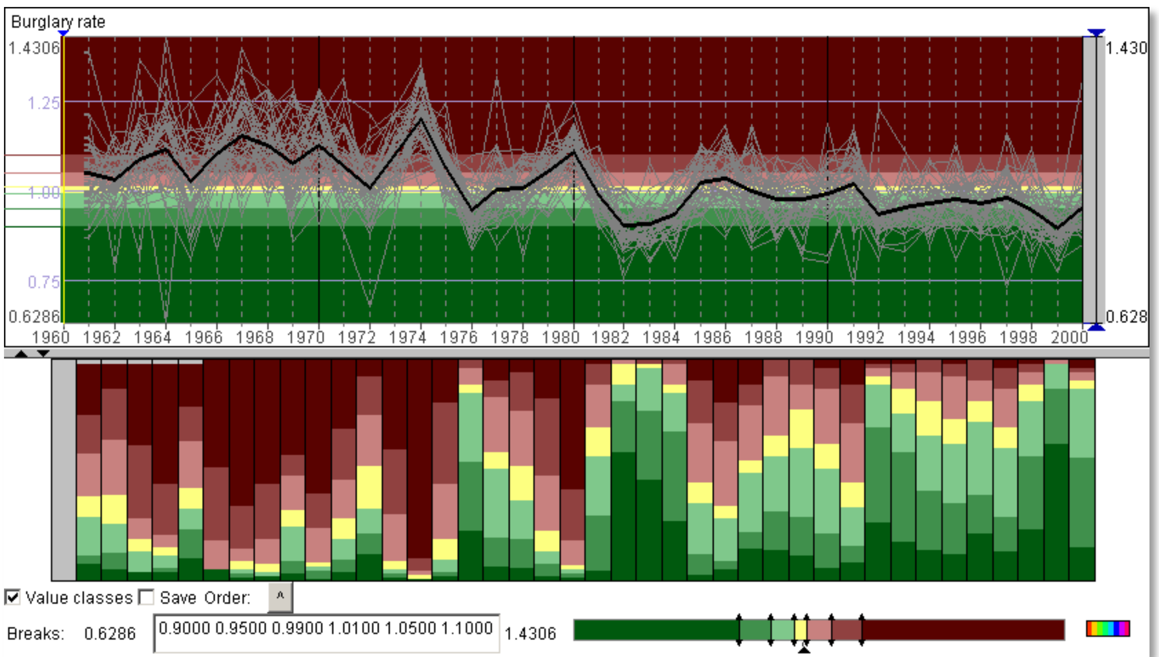
Time Histogram of Changes

Spatial Time Series: Basic Visualization Methods VI

Transform attribute values into relative differences

Apply aggregation by value intervals to the transformed values

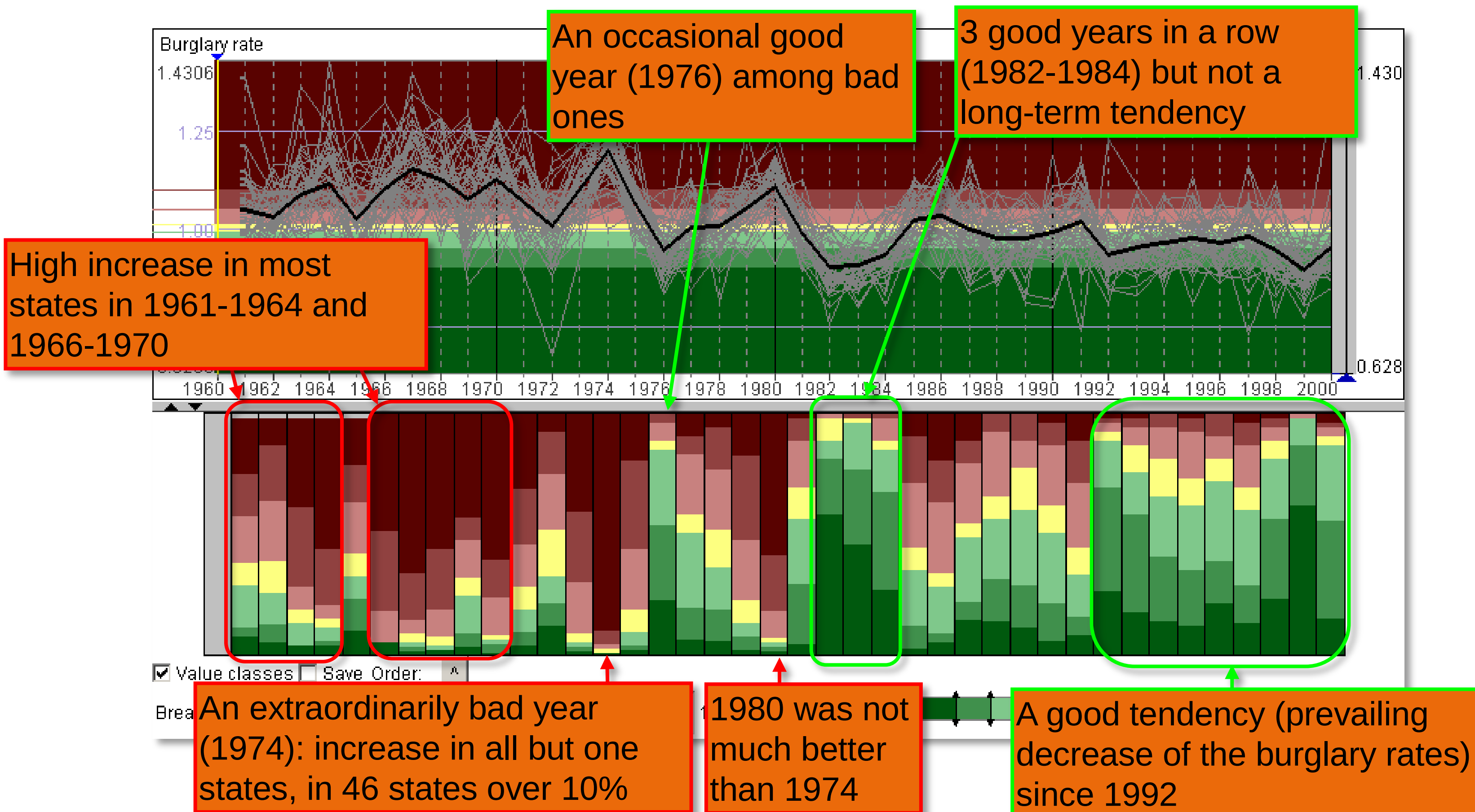
Encode values using partitioned color scale



- "Burglary rate" values, transformed into relative difference to previous year
- Three color segments: red for increase (< +5%, +5-10%, > +10%), yellow for no change (±1%), green for decrease (< -5%, -5-10%, > -10%)

Time Histogram of Changes

Spatial Time Series: Basic Visualization Methods VI



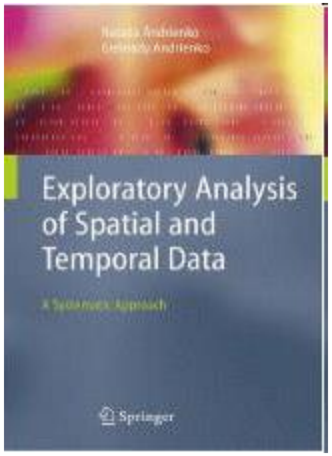
CITY UNIVERSITY LONDON

Fraunhofer IAIS

See also

- Natalia and Gennady Andrienko
  Exploratory Analysis of Spatial and Temporal Data
  A Systematic Approach
  Springer-Verlag, December 2005

  Chapter 4



An Introduction to Visual Analytics

Special focus: movement data

Gennady Andrienko

Natalia Andrienko

http://geoanalytics.net

CITY UNIVERSITY LONDON

EST 1894

Fraunhofer

IAIS

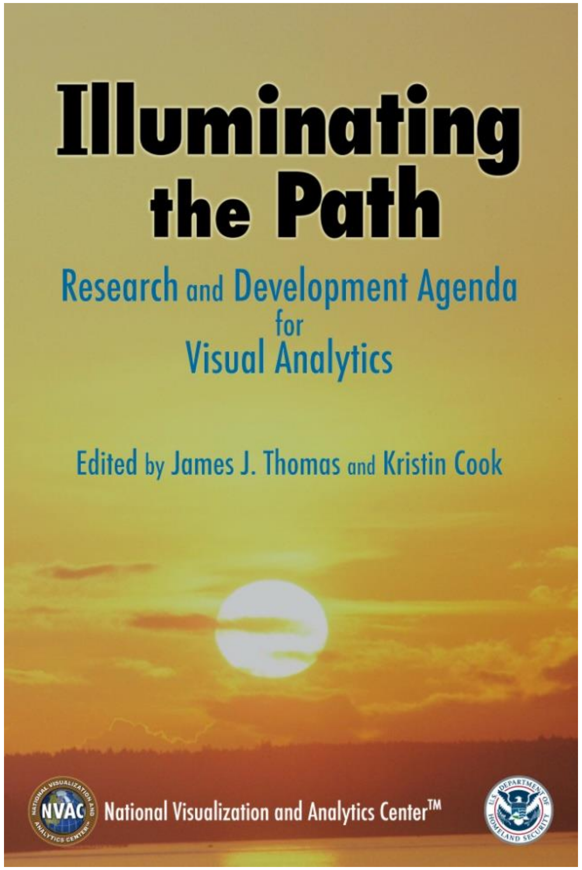
Definition of Visual Analytics

Visual Analytics:

the science of analytical reasoning facilitated by interactive visual interfaces

People use visual analytics tools and techniques to

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, and understandable assessments
- Communicate assessment effectively for action



The book (IEEE Computer Society 2005) is available at http://nvac.pnl.gov/

Analytical reasoning =

data → information → knowledge → explanation

(interpreted data) (for myself) (for others)

CITY UNIVERSITY LONDON Fraunhofer IAIS

Visual Analytics:

divide the labor between the computers and humans to use the best of each

Computers

- can store and process great amounts of information
- are very fast in searching information
- are very fast in processing data
- can extend their capacities by linking with other computers
- can efficiently render high quality graphics, both static and dynamic

Humans

- are flexible and inventive, can deal with new situations and problems
- can solve problems that are hard to formalise
- can reasonably act in cases of incomplete and/or inconsistent information
- can simply ***see*** things that are hard to compute
- can employ their previous knowledge and experience

Visual Analytics:

the importance of visualisation

- Visualise = make perceptible to human's mind
- "An estimated 50 percent of the brain's neurons are associated with vision. Visualisation <…> aims to put that neurological machinery to work."
  - B. McCormick, T. DeFanti, and M. Brown. Definition of Visualization. *ACM SIGGRAPH Computer Graphics*, 21(6), November 1987, p.3
- "An **abstractive grasp of structural features** is the very basis of perception and the beginning of all cognition."
  - R. Arnheim. *Visual Thinking*. University of California Press, Berkeley 1969, renewed 1997, p. 161

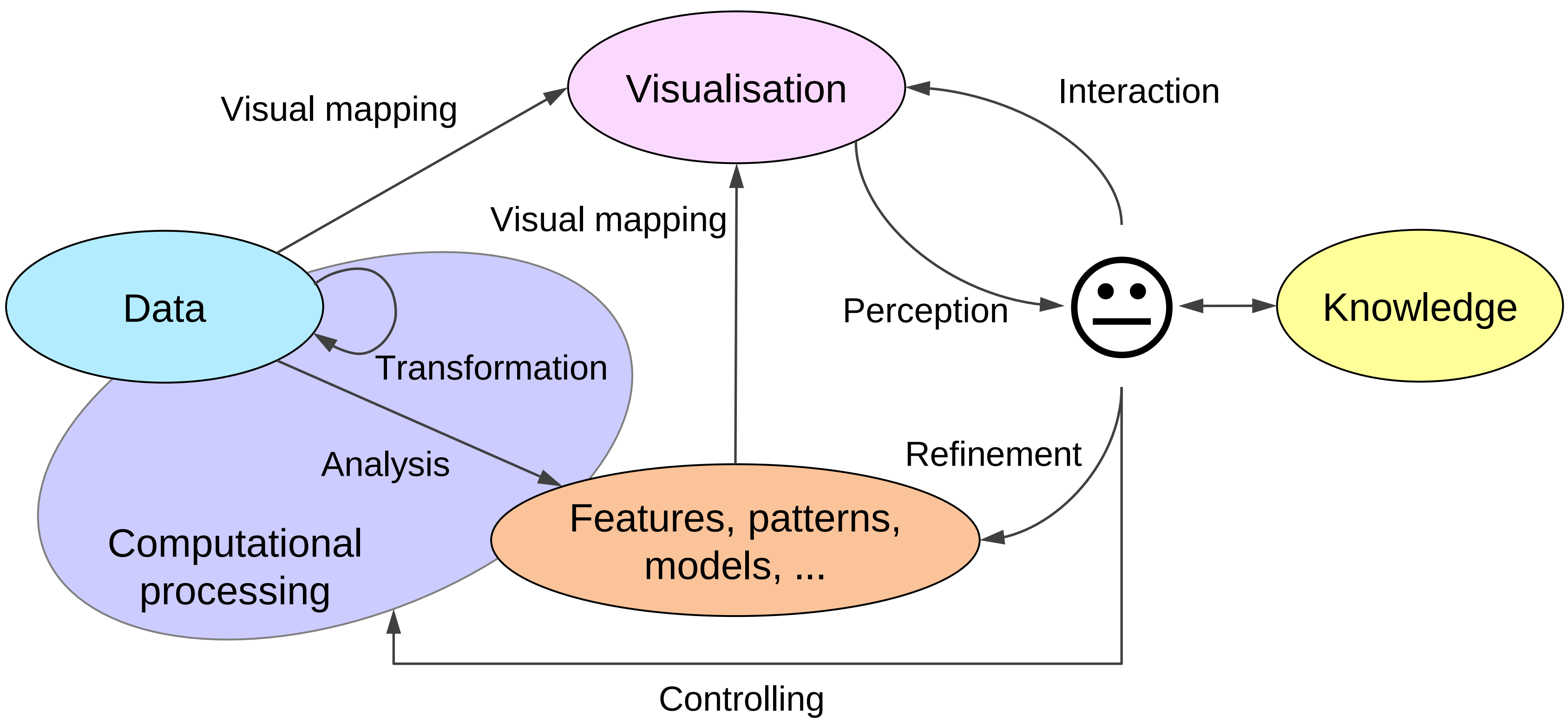
(in other words: seeing already includes analyzing)

⇒ Visualisation is essential for enabling human analysts to use their inherent cognitive capabilities

Visual Analytics technology:

combining methods for visual and computational analysis

Goal: enable synergistic work of humans and computers



Visual Analytics:
a summary

- Defined as *the science of analytical reasoning facilitated by interactive visual interfaces*

Analytical reasoning ≈ data →	information →	knowledge →	explanation
	(interpreted data)	*(for myself)*	*(for others)*

- The challenge of huge and complex data: distil *relevant information* and *connections* between them; gain insight from data!
- **Visual Analytics** (VA) combines interactive visualisations with computational processing
  - database processing, data mining algorithms, statistics, geographical analysis methods, ...
- VA focuses on the division of labour between humans and machines*:*
  - *Computational power* amplifies *human perceptual and cognitive capabilities*
  - *Visual representations* are the *most effective means* to convey information to human's mind and prompt human cognition and reasoning
- Hence, **VA** may be more broadly defined as ***the science of human-computer data analysis, knowledge building, and problem solving***

Data Types and Structures

Everything begins with data

Analytical reasoning =

data → information → knowledge → explanation

- **Data**: factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation *(Merriam-Webster dictionary)*
- **Structured data**: collection of items (records) consisting of components of the same kinds; can be represented in a tabular form

Name	Birth date	School grade	Address	Distance to school, m	Getting to school
Peter	17/05/2005	3	12, Pine street	850	by bus
Julia	23/08/2004	4	9, Oak avenue	400	on foot
Paul	10/12/2005	2	56, Maple road	1500	by car
Mary	06/10/2003	5	71, Linden lane	900	on foot

Types of components in data

* Stevens, S.S. (1946). "On the Theory of Scales of Measurement". *Science* **103** (2684): 677–680.

Types of values:

- Numeric
- Textual
  - Predefined values (e.g., codes)
  - Free text
- Spatial
  - Coordinates
  - Place names
  - Addresses
- Temporal
- Other (image, video, audio, ...)

Scales of measurement*:

- Nominal *(¬ order, ¬ distances)*
  - gender, nationality, ...
- Ordinal *(✓order, ¬ distances)*
  - evaluations: bad, fair, good, excellent
- Interval *(✓order, ✓distances, ¬ ratios, ¬ meaningful zero)*
  - temperature, time, ...
- Ratio *(✓order, ✓distances, ✓ratios, ✓meaningful zero)*
  - quantities, distances, durations, ...

Name	Birth date	School grade	Address	Distance to school, m	Getting to school
Peter	17/05/2005	3	12, Pine street	850	by bus
Julia	23/08/2004	4	9, Oak avenue	400	on foot
Paul	10/12/2005	2	56, Maple road	1500	by car
Mary	06/10/2003	5	71, Linden lane	900	on foot

Semantic roles of data components

- **Reference**: What is described?
  - Object (physical or abstract)
  - Place
  - Time unit
  - Object × time unit
  - Place × time unit
  - Generally: anything specified as a single element or combination
- **Characteristic**, or **attribute**: What is known about it?

In our example:

the data describe children denoted by their names

Name	Birth date	School grade	Address	Distance to school, m	Getting to school
Peter	17/05/2005	3	12, Pine street	850	by bus
Julia	23/08/2004	4	9, Oak avenue	400	on foot
Paul	10/12/2005	2	56, Maple road	1500	by car
Mary	06/10/2003	5	71, Linden lane	900	on foot

Data may have two or more references

Reference: time

Reference: place

Attributes

year	id	State	Population	Index offenses	Violent crime	Murder	Forcible rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft
1960	1	Alabama	3266740	39920	6097	406	281	898	4512	33823	11626	19344	2853
1960	2	Alaska	226167	3730	236	23	47	64	102	3494	751	2195	548
1960	4	Arizona	1302161	39243	2704	78	209	706	1711	36539	8926	23207	4406
1960	5	Arkansas	1786272	18472	1924	152	159	443	1170	16548	5399	10250	899
1960	6	California	15717204	546069	37558	616	2859	15287	18796	508511	143102	311956	53453
1960	8	Colorado	1753947	38103	2408	73	229	1362	744	35695	9996	21949	3750
1960	9	Connecticut	2535234	29321	928	41	103	236	548	28393	8452	16653	3288
1960	10	Delaware	446292	9642	375	33	41	157	144	9267	2661	5867	739
1960	11	District of Co	763956	20725	4230	81	111	1072	2966	16495	4587	9905	2003
1960	12	Florida	4951560	133919	11061	527	403	4005	6126	122858	39966	73603	9289
...													
1972	54	West Virgini	1781000	25584	2299	109	146	562	1482	23285	7356	13976	1953
1972	55	Wisconsin	4520000	133382	4358	126	376	1661	2195	129024	28862	89642	10520
1972	56	Wyoming	345000	10461	511	14	48	117	332	9950	2057	7190	703
1973	1	Alabama	3539000	91389	12390	468	751	2809	8362	78999	31754	39206	8039
1973	2	Alaska	330000	16313	1269	33	147	221	868	15044	3852	9456	1736
1973	4	Arizona	2058000	137966	9877	167	637	3031	6042	128089	40301	76560	11228
1973	5	Arkansas	2037000	56149	5905	180	398	1456	3871	50244	18088	29204	2952
1973	6	California	20601000	1298872	116563	1862	8357	49531	56813	1182309	407824	643488	130997
1973	8	Colorado	2437000	133933	10088	193	944	3970	4981	123845	38963	70931	13951
1973	9	Connecticut	3076000	112717	6421	102	342	2589	3388	106296	31661	58742	15893
...													
2000	44	Rhode Island	1048319	36444	3121	45	412	922	1742	33323	6620	22038	4665
2000	45	South Caroli	4012012	209482	32293	233	1511	5883	24666	177189	38888	123094	15207
2000	46	South Dakot	754844	17511	1259	7	305	131	816	16252	2896	12558	798
2000	47	Tennessee	5689283	278218	40233	410	2186	9465	28172	237985	56344	154111	27530
2000	48	Texas	20851820	1033311	113653	1238	7856	30257	74302	919658	188975	637522	93161
2000	49	Utah	2233169	99958	5711	43	863	1242	3563	94247	14348	73438	6461
2000	50	Vermont	608827	18185	691	9	140	117	425	17494	3501	13184	809
2000	51	Virginia	7078515	214348	19943	401	1616	6295	11631	194405	30434	146158	17813
2000	53	Washington	5894121	300932	21788	196	2737	5812	13043	279144	53476	190650	35018
2000	54	West Virgini	1808344	47067	5723	46	331	749	4597	41344	9890	28139	3315
2000	55	Wisconsin	5363675	172124	12700	169	1165	4537	6829	159424	25183	119605	14636
2000	56	Wyoming	493782	16285	1316	12	160	70	1074	14969	2078	12318	573

Common classes of data structures

according to the types of the references

- **Object-referenced data**: attributes of objects
  - **Events**: attributes include time of existence (moment or interval)
  - **Spatial objects**: attributes include spatial location (point, area, or volume)
  - **Spatial events**: attributes include existence time and location
- **Time-referenced data**, a.k.a. *time series*: attributes observed in different times (moments or intervals)
- **Space-referenced data**, a.k.a. *spatial data*: attributes observed in different places
- **Object time series**: attributes of objects observed in different times
  - **Trajectories** of moving objects: time series of spatial locations
- **Spatial time series**: attributes observed in different places and times

Multidimensional data

- Data including multiple attributes
- May have various types of references: objects, places, times, combinations ...
- Multiple attributes referring to times: *multidimensional time series*
- Multiple attributes referring to places: *multidimensional spatial data*
- Multiple attributes referring to places + times: *multidimensional spatial time series*

Classes of spatio-temporal data

Include space and time as references or characteristics

- *Reference*: objects; *attributes*: existence time + location + ...
  → ***spatial events***
  - Earthquakes, mobile phone calls, public events, ...
- *References*: places + times; *attributes*: ...
  → ***spatial time series***
  - Weather, population census data for different years, election results, ...
- *References*: objects + times; *attributes*: location + ...
  → ***trajectories***
  - Trajectories of people, animals, vehicles, icebergs, hurricanes, ...

Running example dataset: trajectories of cars in Milan

The reliable way

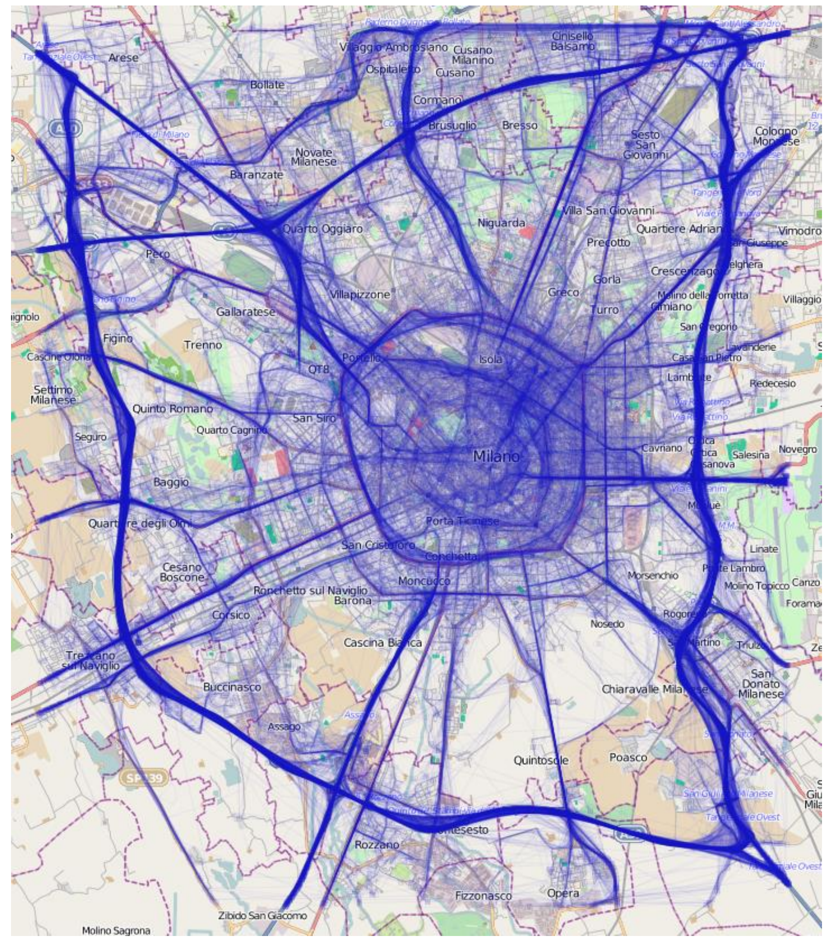
GPS-tracks of 17,241 cars in Milan, Italy

Time period: from Sunday, the 1st of April, to Saturday, the 7th of April, 2007

Received from Octo Telematics
www.octotelematics.com
special thanks to Tina Martino

Data structure:

- Anonymized car identifier
- Date and time
- Geographic coordinates
- Speed



The trajectories from one day are drawn on a map with 5% opacity

CITY UNIVERSITY LONDON

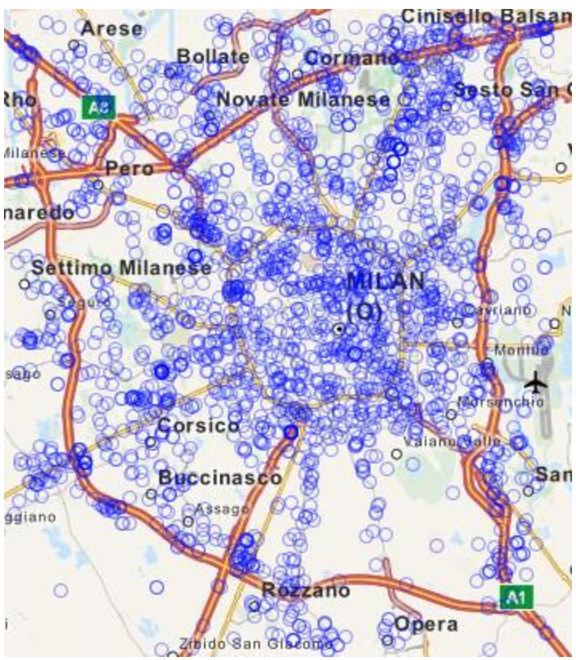
Fraunhofer IAIS

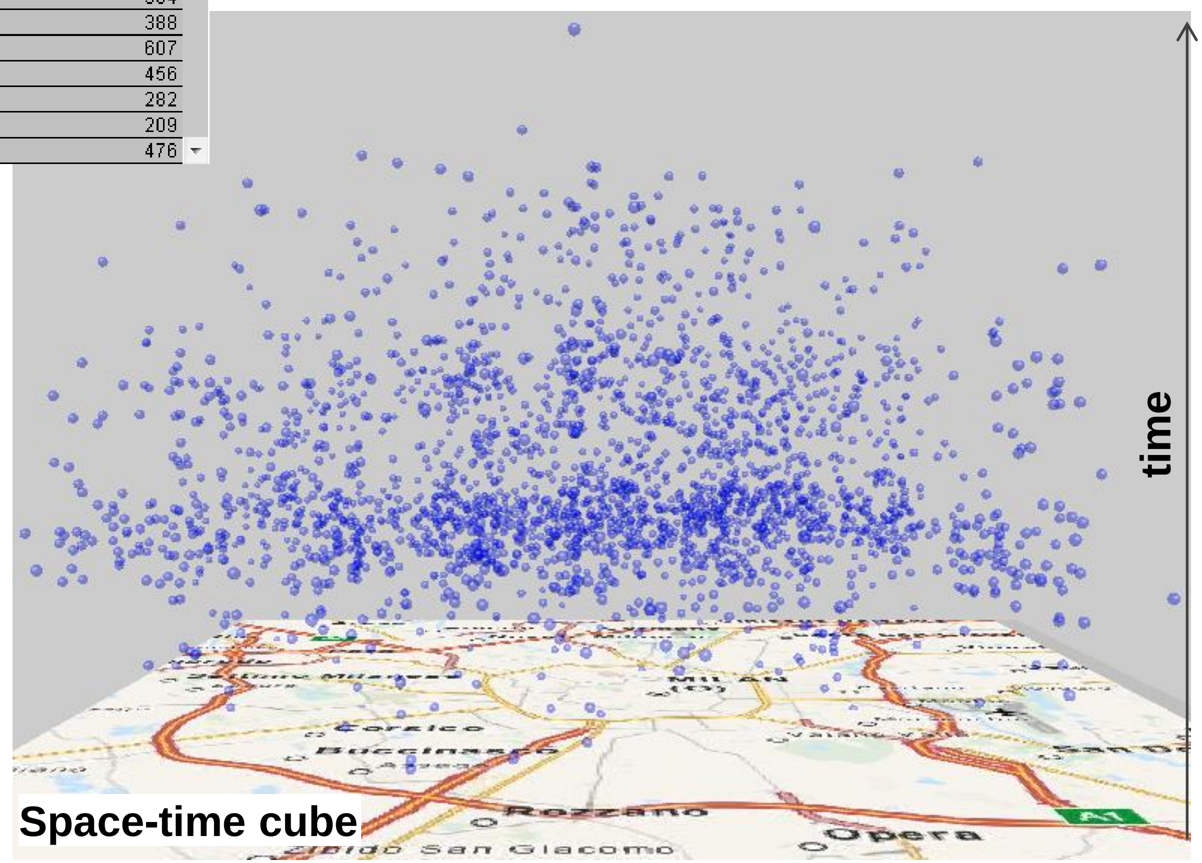
Spatial events

ident	Longitude	Latitude	Start time	Duration (minutes)
1	9.149877	45.435764	04/04/2007 12:18:59	404
2	9.118851	45.405876	04/04/2007 04:30:18	305
3	9.153637	45.384422	04/04/2007 05:50:05	539
4	9.088079	45.526165	04/04/2007 06:38:20	241
5	9.088102	45.52614	04/04/2007 12:07:45	331
6	9.224027	45.529514	04/04/2007 06:47:07	186
7	9.230214	45.545048	04/04/2007 11:03:16	318
8	9.27865	45.46803	04/04/2007 06:24:51	385
9	9.088983	45.533787	04/04/2007 06:00:37	248
10	9.08896	45.53379	04/04/2007 11:27:47	304
11	9.225346	45.528656	04/04/2007 12:29:03	388
12	9.135795	45.486305	04/04/2007 05:14:31	607
13	9.080329	45.539093	04/04/2007 07:19:02	456
14	9.208277	45.46646	04/04/2007 13:45:37	282
15	9.188722	45.473045	04/04/2007 18:42:53	209
16	9.172471	45.38386	04/04/2007 04:34:31	476

References: objects
Attributes: existence time + location + ...

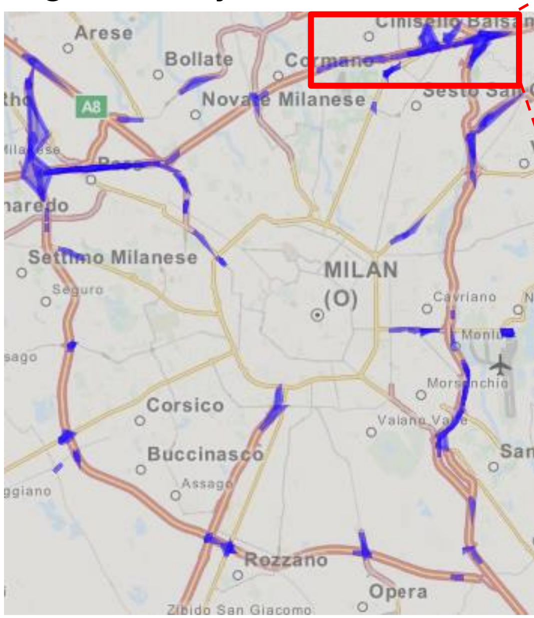
Examples: earthquakes, mobile phone calls, public events, car stops, traffic jams, ...

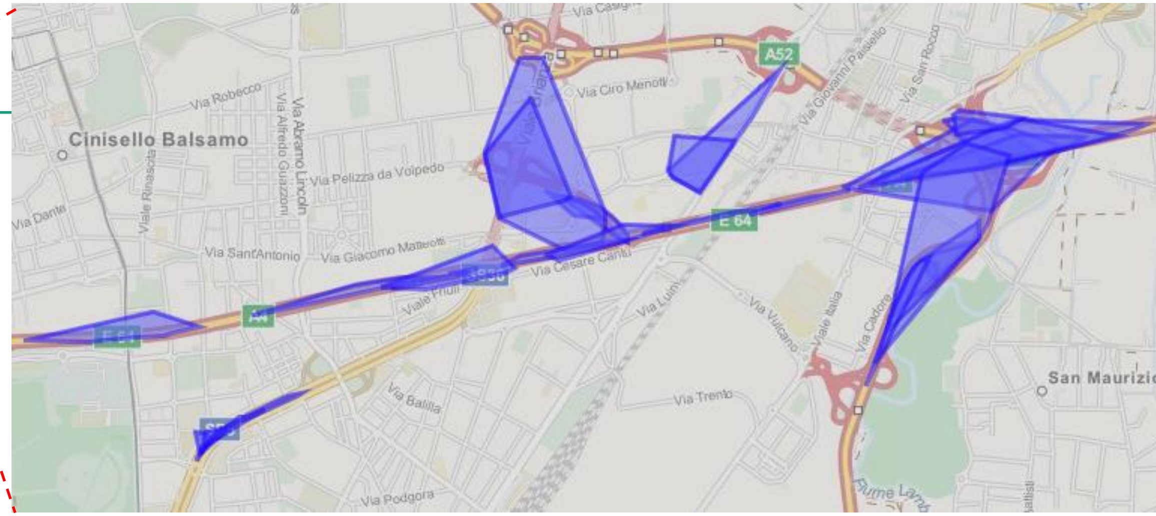




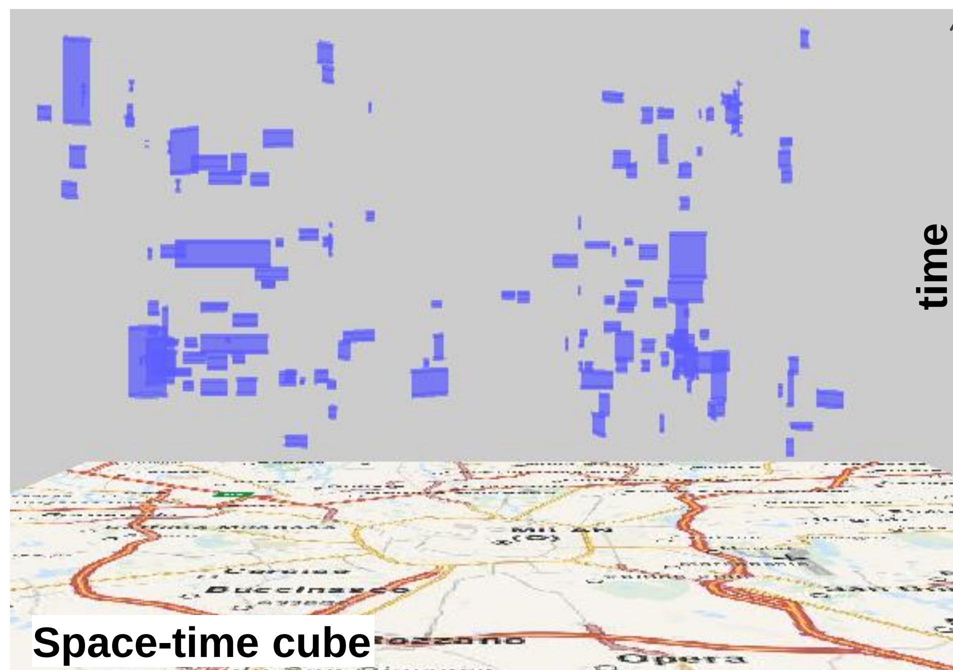
Spatial events may be extended in time and/or space

e.g., traffic jams





	extent, m	Begin time	End time	Duration, minutes
1	1909.82	04/04/2007 05:04:56	04/04/2007 08:08:19	183.4
2	291.75	04/04/2007 05:22:46	04/04/2007 05:39:20	16.6
3	734.22	04/04/2007 05:06:35	04/04/2007 05:41:35	35.0
4	576.55	04/04/2007 05:08:45	04/04/2007 05:46:38	37.9
5	491.55	04/04/2007 05:33:43	04/04/2007 06:03:21	29.6
6	143.53	04/04/2007 05:35:34	04/04/2007 05:45:41	10.1
7	459.94	04/04/2007 05:38:42	04/04/2007 06:07:08	28.4
8	473.88	04/04/2007 15:11:05	04/04/2007 15:56:18	45.2
9	358.59	04/04/2007 15:41:36	04/04/2007 16:16:29	34.9
10	1256.72	04/04/2007 10:32:13	04/04/2007 12:45:12	133.0
11	270.41	04/04/2007 12:14:09	04/04/2007 12:23:01	8.9
12	371.49	04/04/2007 06:00:20	04/04/2007 06:06:14	5.9



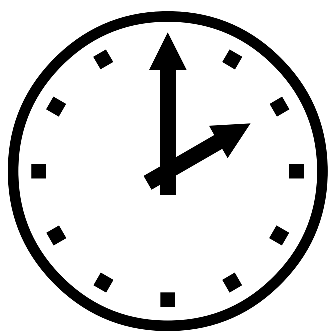
Spatial time series

References: places + times; attributes: any

Example: number of cars that visited the regions of Milan in different hours of a day

identifie	hour=00: N visitors by hours	hour=01: N visitors by hours	hour=02: N visitors by hours	hour=03: N visitors by hours	hour=04: N visitors by hours	hour=05: N visitors by hours	hour=06: N visitors by hours	hour=07: N visitors by hours	hour=08: N visitors by hours	hour=09: N visitors by hours	hour=10: N visitors by hours	hour=11: N visitors by hours
Centre	68	45	46	116	272	630	935	928	852	878	890	791
North	24	25	20	65	203	394	478	420	362	384	423	420
South	11	7	22	47	102	191	250	265	207	218	222	200
West	18	14	17	42	98	219	314	263	230	211	240	215
East	12	15	28	68	144	262	314	279	257	229	227	261
Northeast	29	27	28	78	233	437	538	512	464	492	462	446
Northwest	24	19	21	72	197	397	506	425	341	371	382	361
Southwest	18	14	19	37	111	208	327	304	247	236	265	245
Southeast	12	8	27	47	102	185	234	228	168	184	172	190

identifie	hour=12: N visitors by hours	hour=13: N visitors by hours	hour=14: N visitors by hours	hour=15: N visitors by hours	hour=16: N visitors by hours	hour=17: N visitors by hours	hour=18: N visitors by hours	hour=19: N visitors by hours	hour=20: N visitors by hours	hour=21: N visitors by hours	hour=22: N visitors by hours	hour=23: N visitors by hours
Centre	812	885	883	926	1007	956	725	479	407	413	292	177
North	431	395	419	490	521	477	330	212	157	145	106	61
South	216	236	240	261	248	260	178	97	85	74	57	38
West	259	231	249	258	304	295	162	107	80	88	49	33
East	275	264	311	300	330	281	202	126	102	99	64	46
Northeast	504	445	497	515	540	523	359	222	203	169	130	62
Northwest	427	367	425	493	499	449	261	187	154	127	88	50
Southwest	240	276	264	338	335	301	198	108	82	72	50	35
Southeast	186	191	207	181	218	205	128	95	69	65	32	34

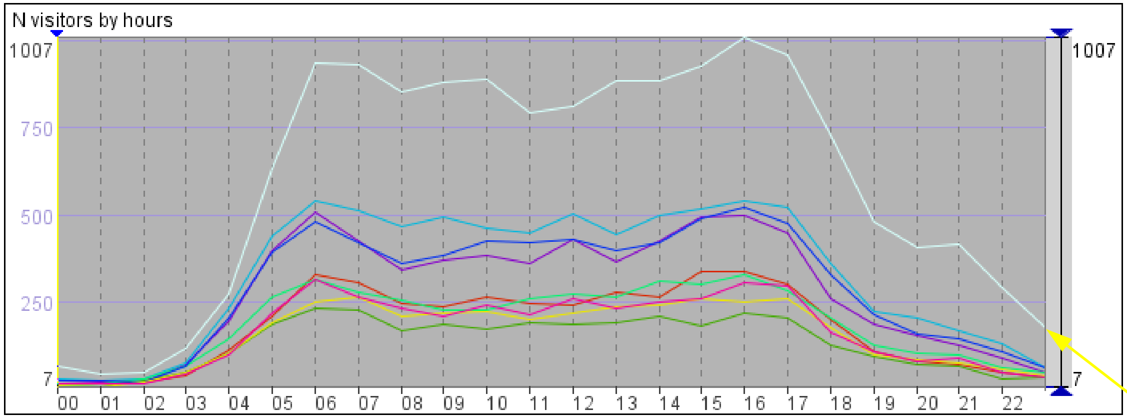


Data structure: (region, hour) → number of cars

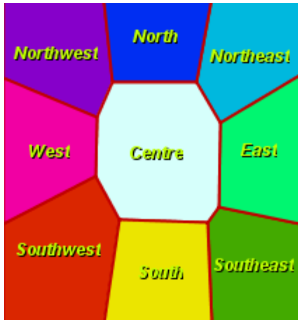
Spatial time series

viewed as spatially distributed local (location-associated) time series

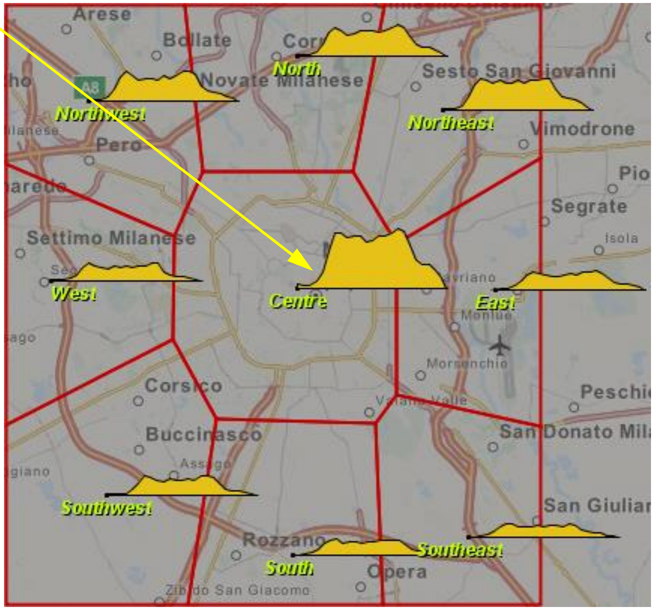
Time graph

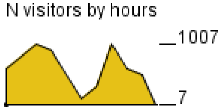


Each line represents the temporal variation of the attribute values in one place



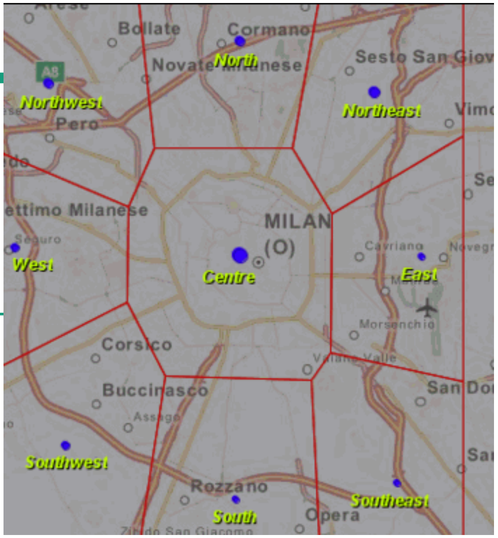
Map with time diagrams



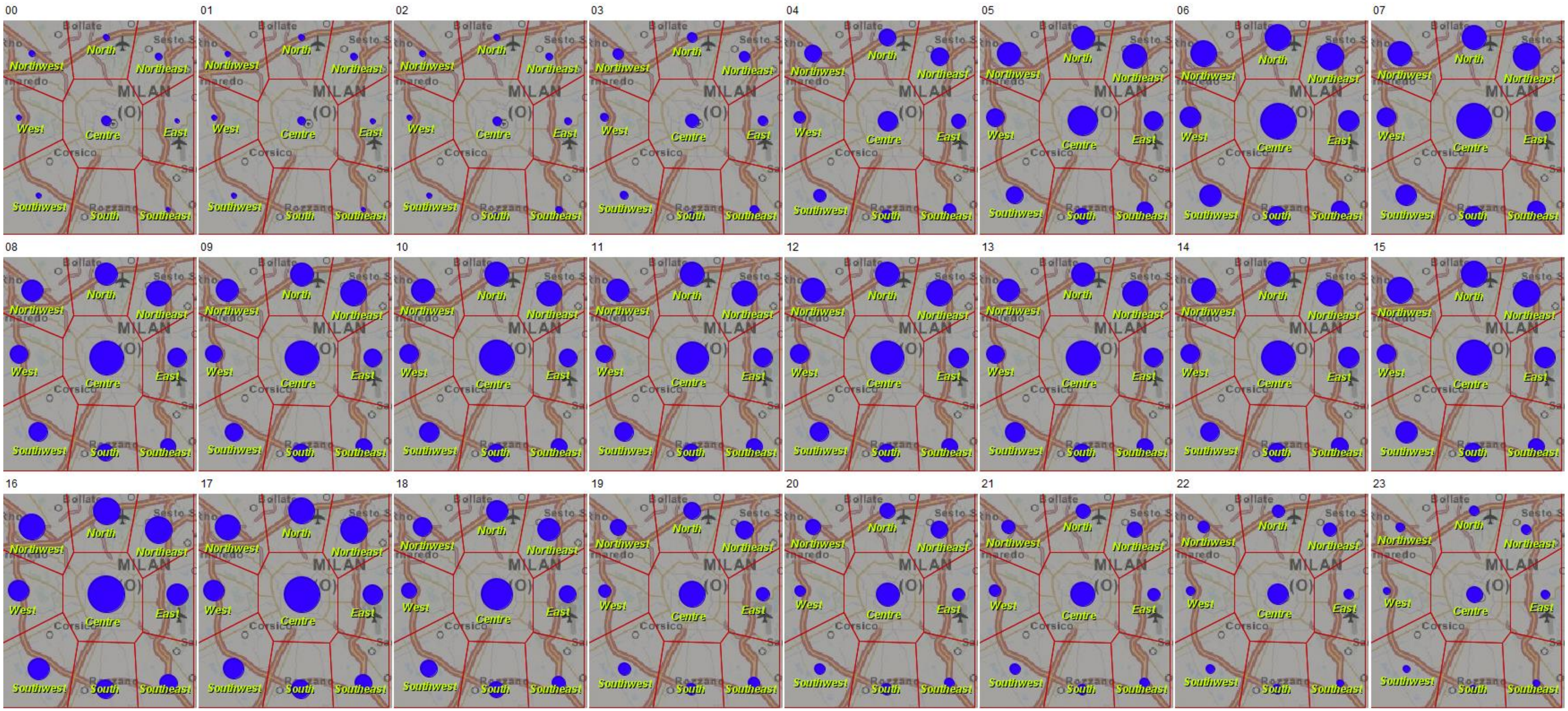


Spatial time series

viewed as a temporal sequence of spatial situations



Numbers of cars in the regions in different hours are represented by proportional sizes of the circle symbols. A sequence of spatial situations may be visualised using an animated map (right) or multiple maps each showing one time moment or interval (below).

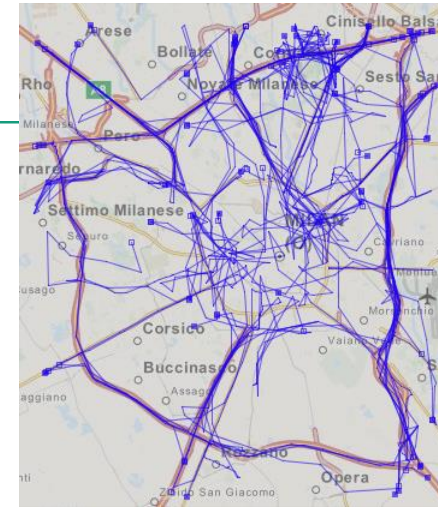


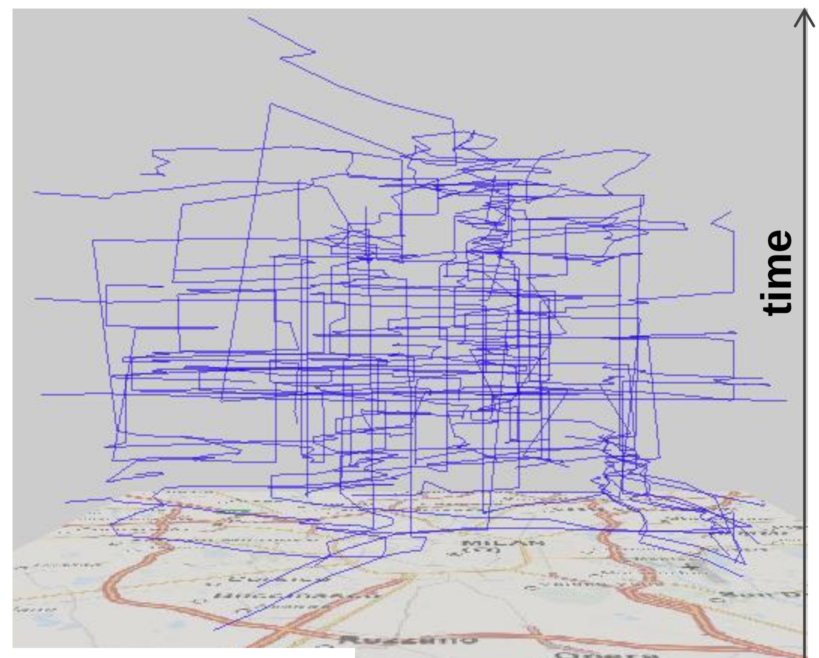
Trajectories

temporal sequences of spatial positions of discrete objects

	Car id	point N	longitude	latitude	time
1	104876	1	9.119127	45.558304	04/04/2007 06:45:15
2	104876	2	9.142448	45.559753	04/04/2007 06:48:49
3	104876	3	9.156955	45.554962	04/04/2007 06:54:54
4	104876	4	9.156504	45.55017	04/04/2007 07:00:12
5	104876	5	9.156504	45.55017	04/04/2007 07:11:08
6	104876	6	9.156844	45.547703	04/04/2007 07:13:26
7	104876	7	9.156909	45.547688	04/04/2007 07:19:23
8	104876	8	9.162037	45.554867	04/04/2007 07:40:02
9	104876	9	9.167628	45.55907	04/04/2007 08:02:32
10	104876	10	9.172845	45.555725	04/04/2007 08:05:38
11	104876	11	9.172696	45.555492	04/04/2007 10:03:31
12	104876	12	9.166886	45.54498	04/04/2007 10:09:38
13	104876	13	9.163299	45.557983	04/04/2007 10:12:05
14	104876	14	9.162168	45.554855	04/04/2007 10:13:51
15	104876	15	9.162158	45.55487	04/04/2007 11:36:23
16	104876	16	9.162622	45.557976	04/04/2007 12:08:17
17	104876	17	9.16232	45.55496	04/04/2007 12:09:19
18	104876	18	9.162361	45.554943	04/04/2007 15:30:22
19	104876	19	9.122161	45.55825	04/04/2007 15:38:51
20	110800	1	9.266509	45.386322	04/04/2007 05:21:45
21	110800	2	9.261211	45.40307	04/04/2007 05:22:57
22	110800	3	9.247442	45.418125	04/04/2007 05:24:13
23	110800	4	9.254333	45.43362	04/04/2007 05:29:45
24	110800	5	9.257282	45.451492	04/04/2007 05:32:44
25	110800	6	9.252168	45.468708	04/04/2007 05:34:21
26	110800	7	9.251433	45.48671	04/04/2007 05:35:48
27	110800	8	9.258238	45.504066	04/04/2007 05:37:05
28	110800	9	9.260647	45.522255	04/04/2007 05:38:26
29	110800	10	9.278728	45.53516	04/04/2007 05:39:48
30	110800	11	9.274316	45.533176	04/04/2007 11:57:53
31	110800	12	9.261258	45.519493	04/04/2007 11:59:21
32	110800	13	9.256271	45.502003	04/04/2007 12:00:51
33	1				12:02:11
34	1				12:03:32
35	1				12:04:53
36	1				12:06:14
37	1				12:07:39
38	1				12:08:54
39	110800	20	9.270314	45.382904	04/04/2007 12:10:10
40	116291	1	9.234817	45.508648	04/04/2007 18:01:18
41	116291	2	9.257177	45.51305	04/04/2007 18:06:57
42	116291	3	9.255232	45.498352	04/04/2007 18:09:05

References: objects + times;
attributes: spatial location + ...



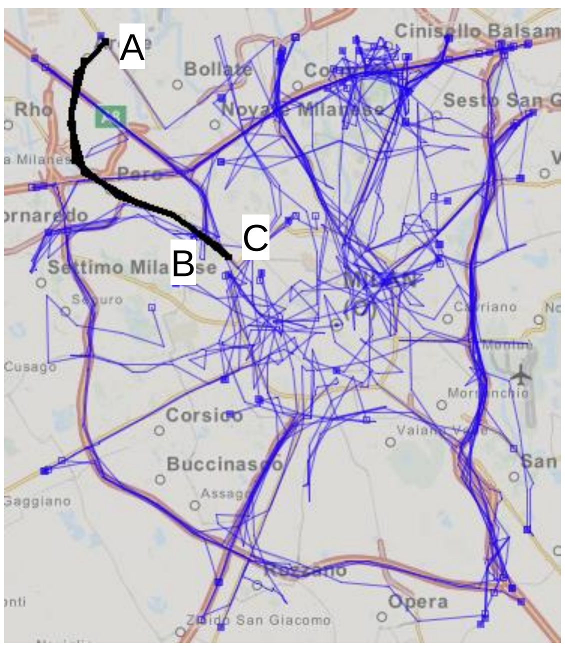


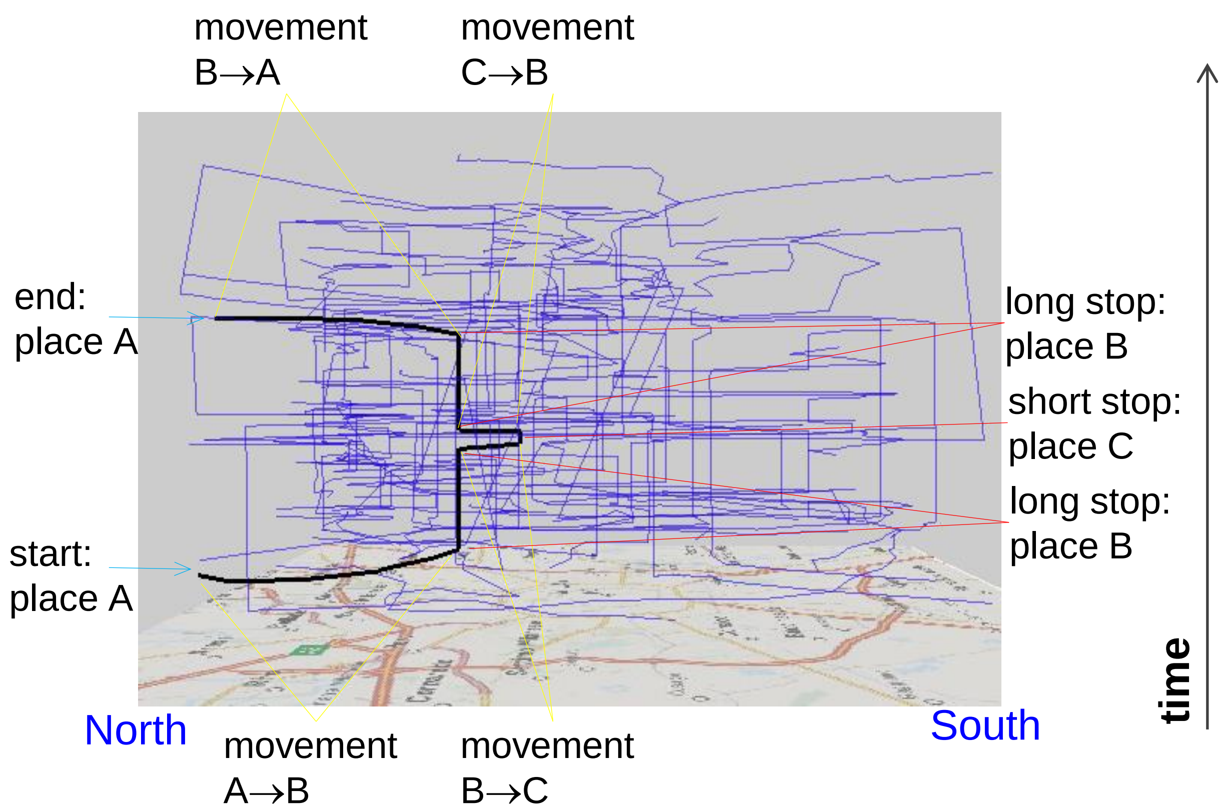
Space-time cube

Trajectories

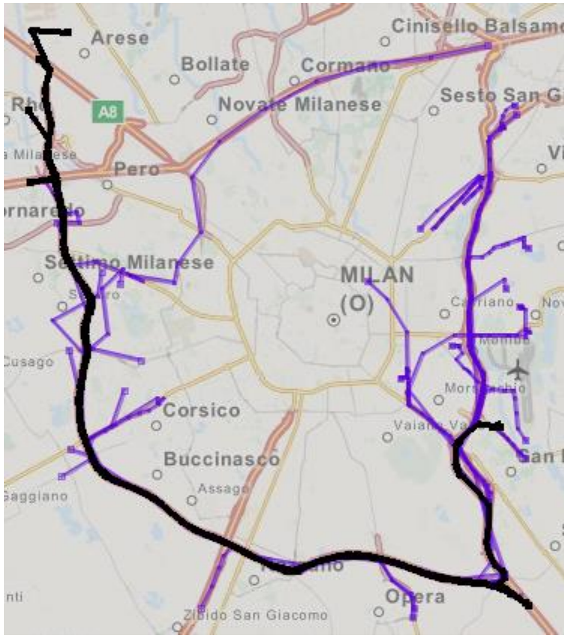
Spatio-temporal view

spatial footprint





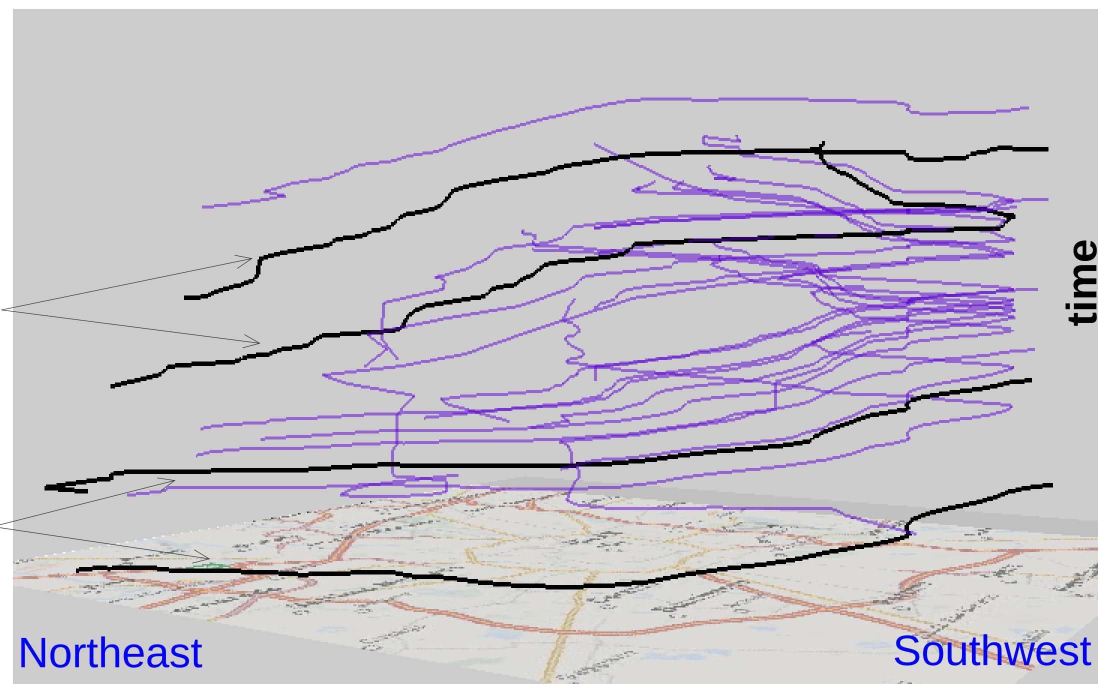
Trajectories



Spatio-temporal view

slow movement

fast movement



Exercise: visual exploration of trajectories

Data: a small sample of daily trajectories

- Visualise the trajectories on a map and in a space-time cube.
- Find stops in the space-time cube.
- Select some trajectories on the map (by clicking) and examine their shapes on the map and in the space-time cube. Rotate the cube when needed for better seeing the trajectory shapes.

Interactive Techniques for Exploration of Spatio-Temporal Data

Focus: Interactive Filtering

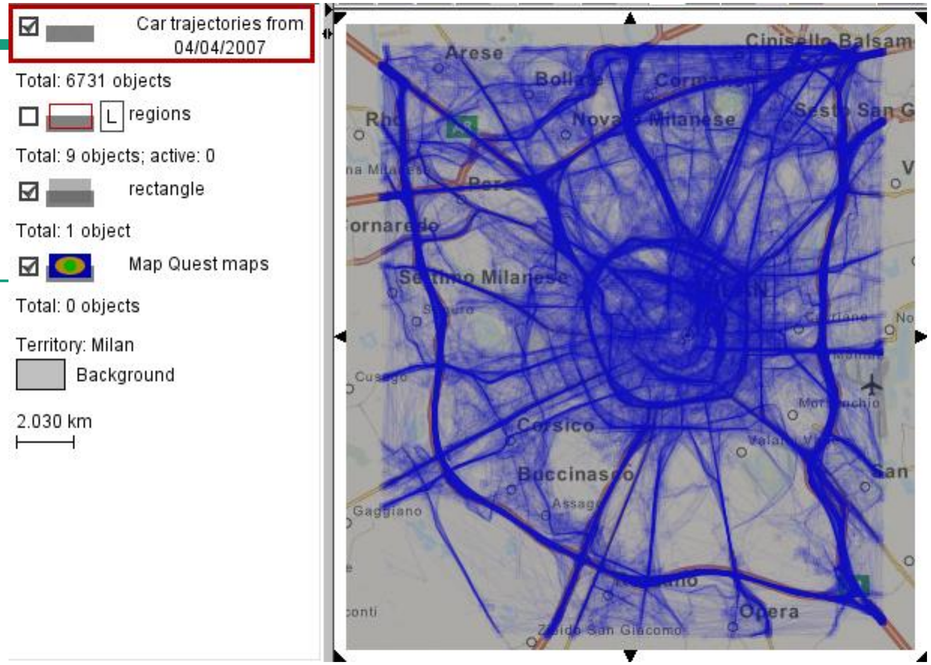
CITY UNIVERSITY LONDON

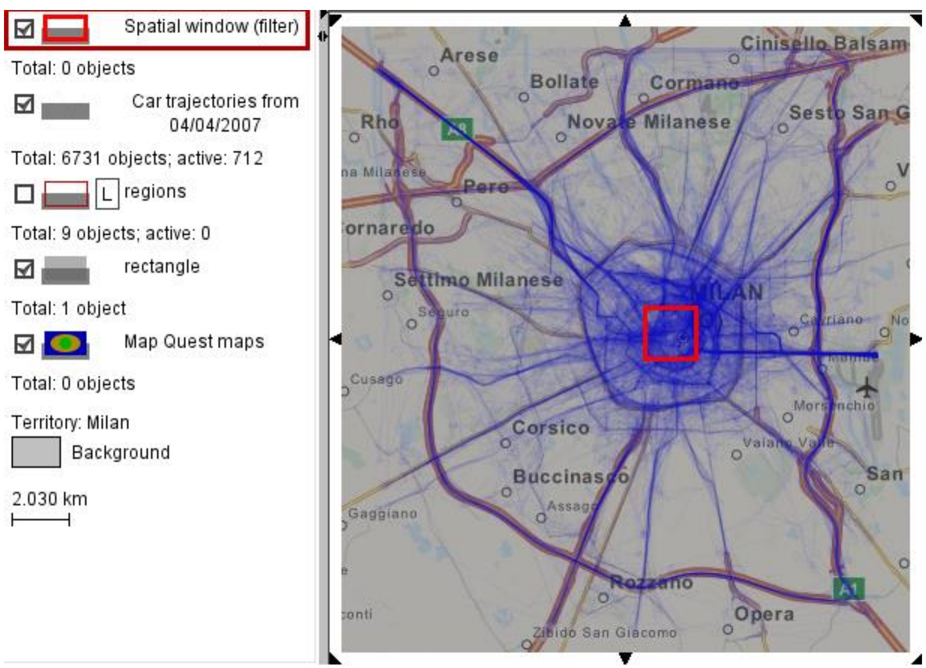
EST 1894

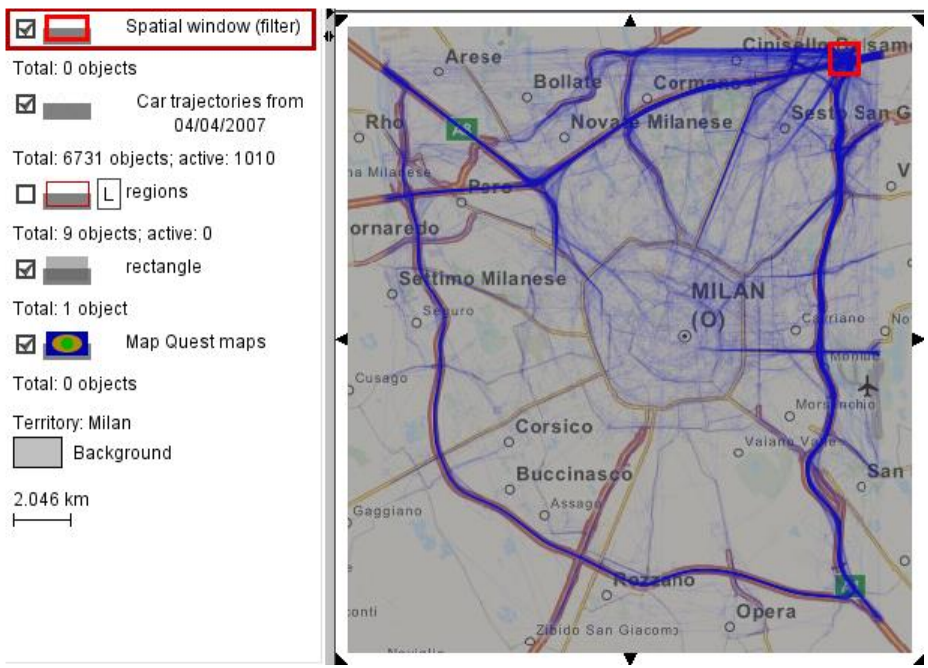
Fraunhofer IAIS

Spatial filtering

by a rectangular "spatial window"

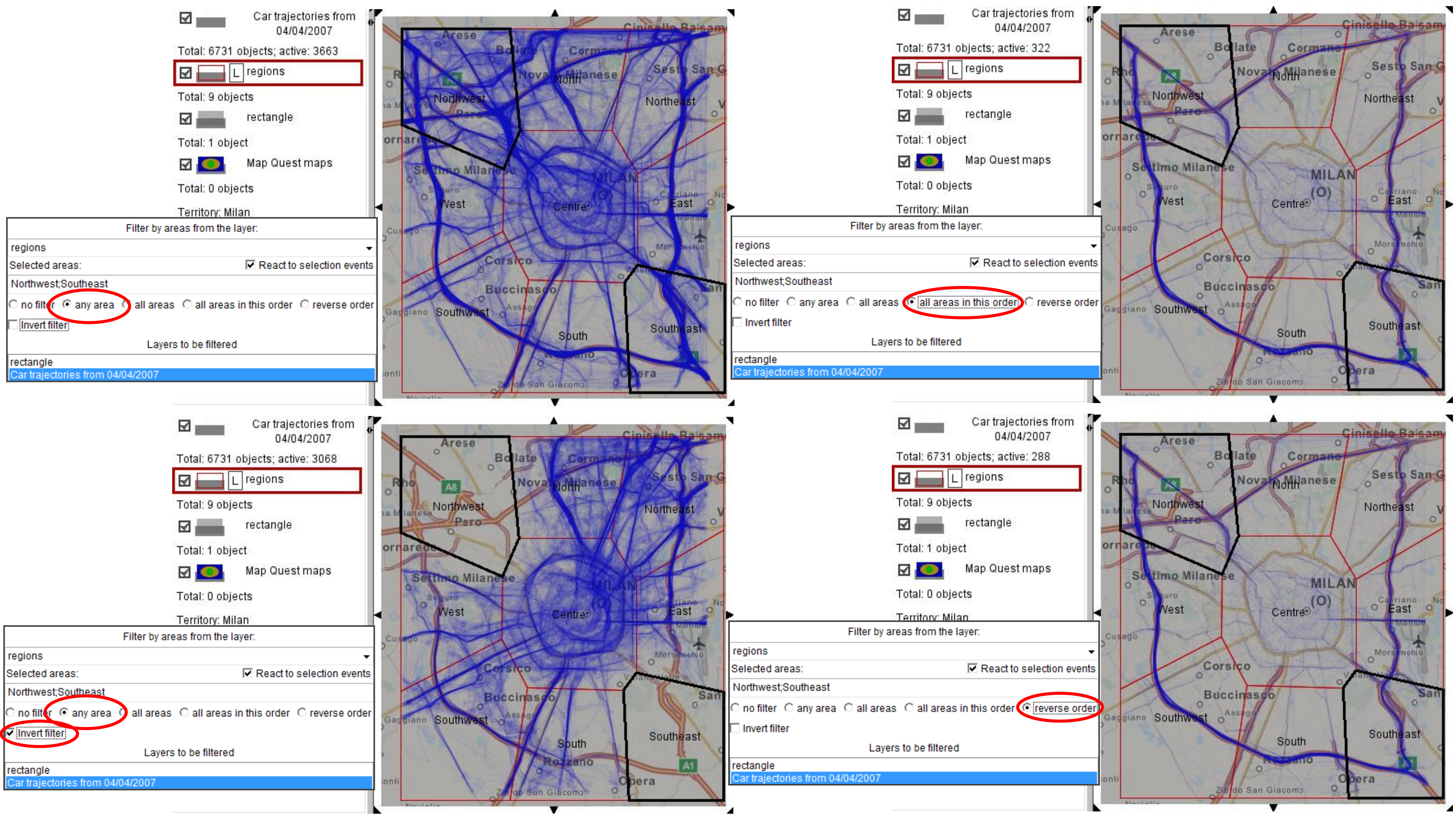




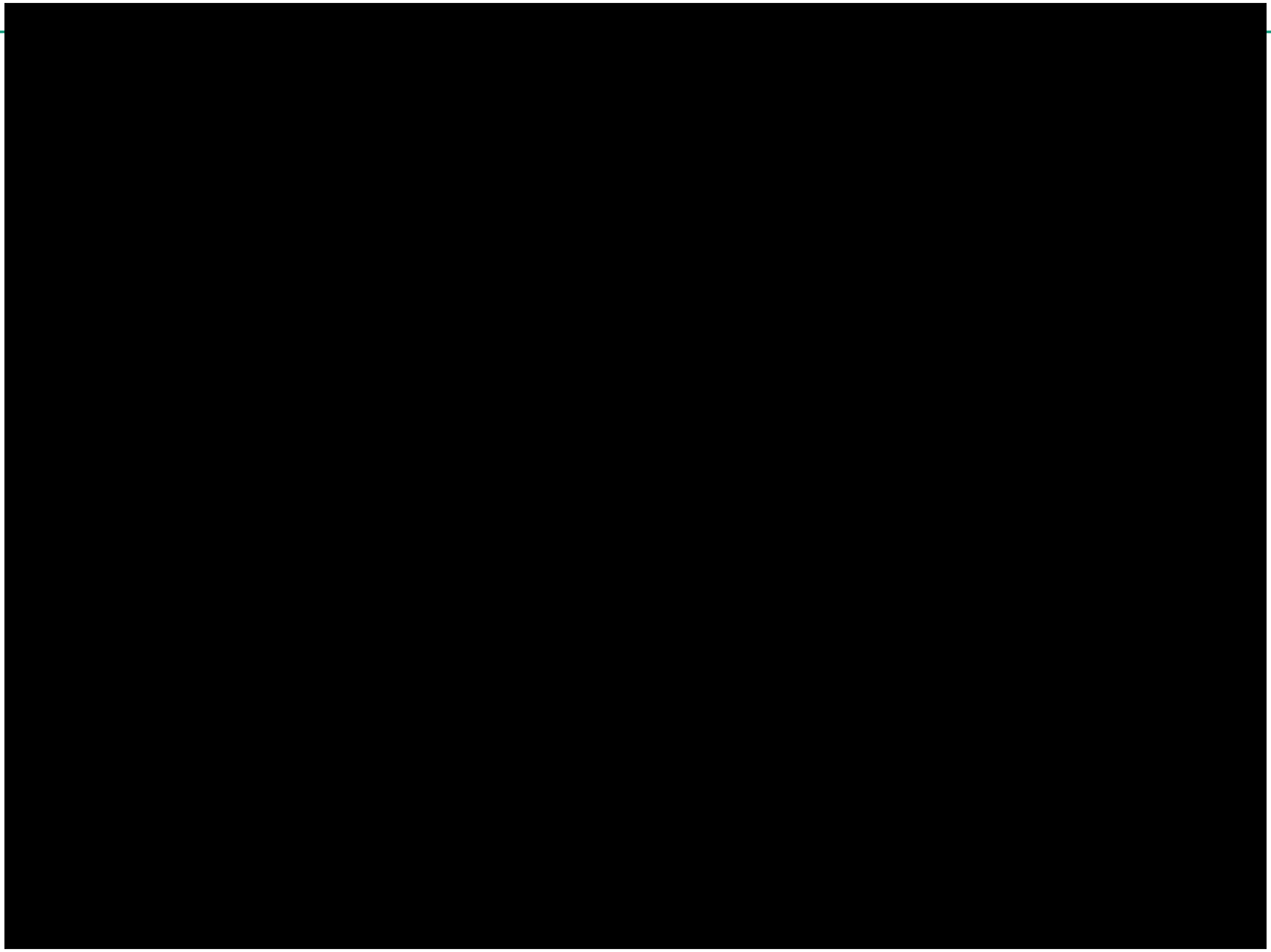


Spatial filtering

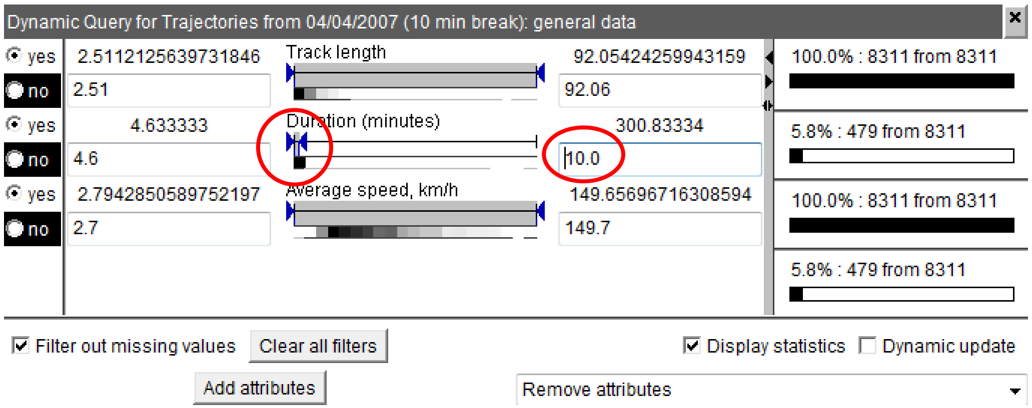
by areas from a map layer

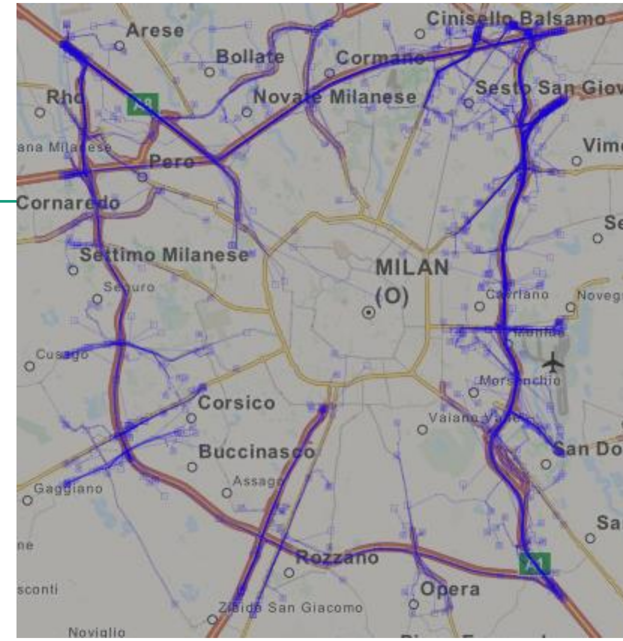


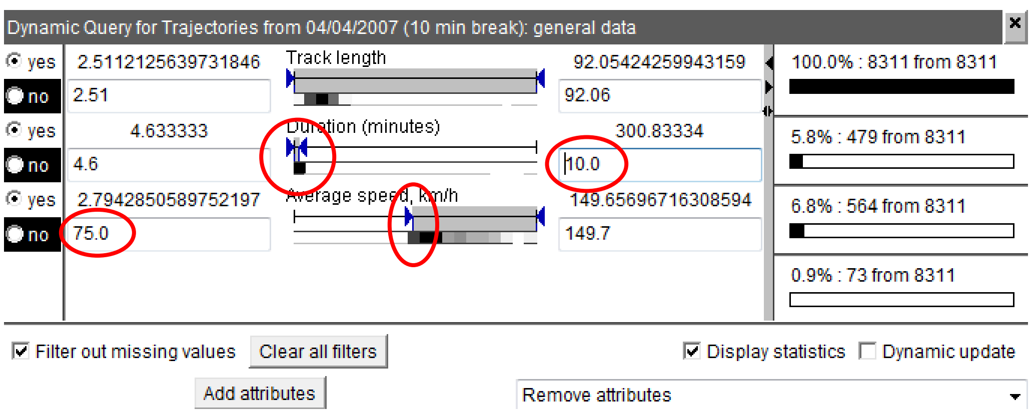
Temporal filtering

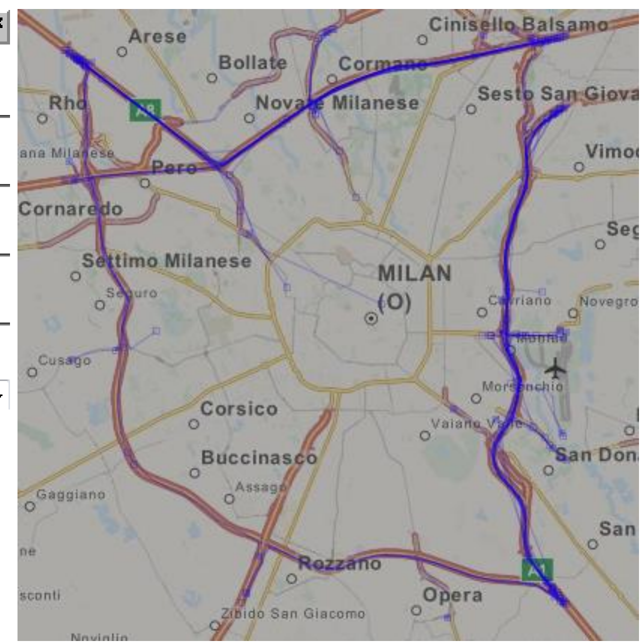


Filtering by attributes

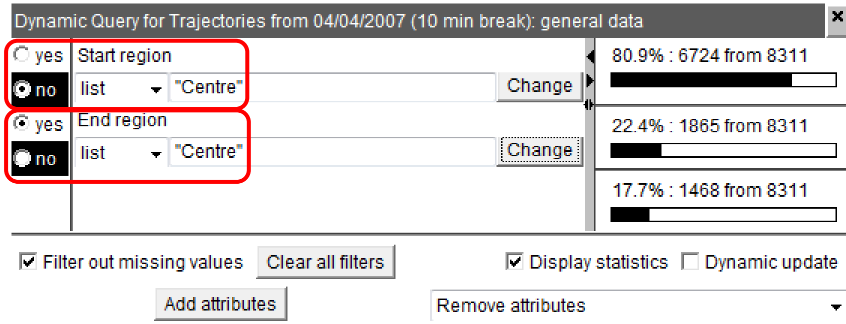


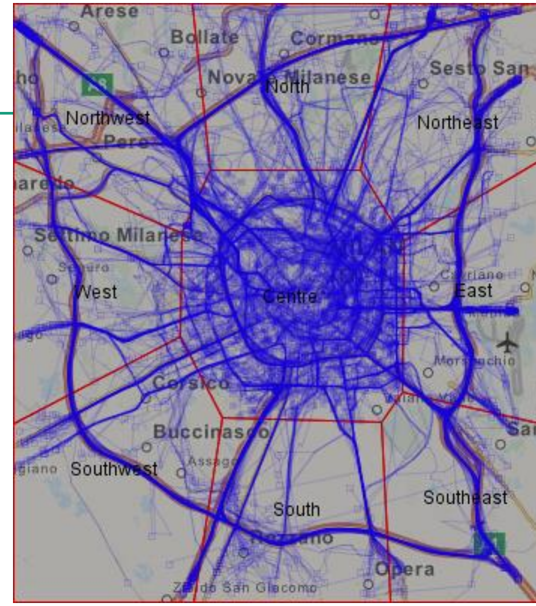






Filtering by attributes



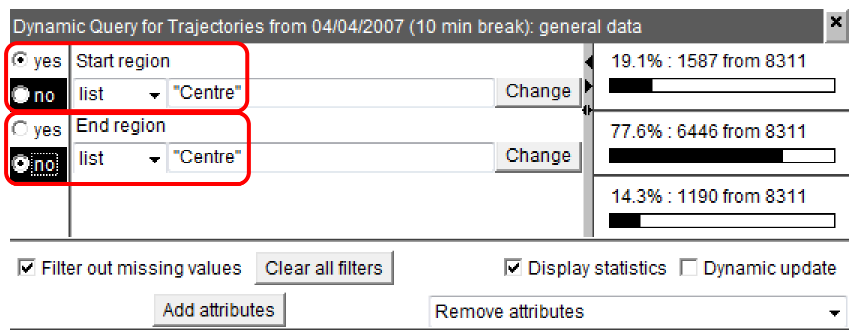


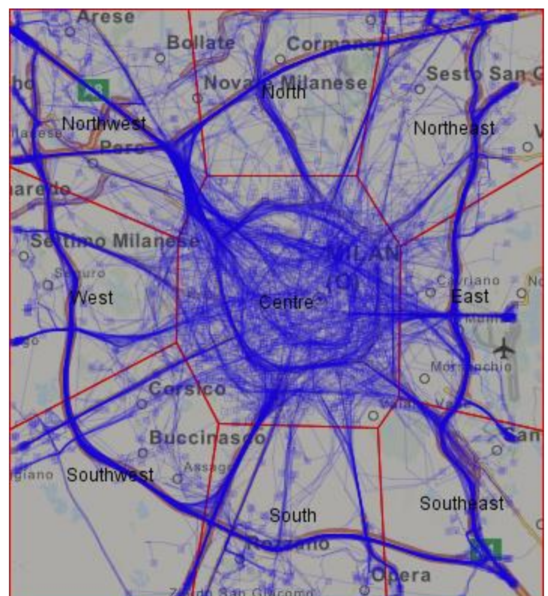
Start region

identifiers	Overall frequency	Frequency after filtering	Ratio (%)
Northeast	1722	355	20.62
Centre	1587	0	0.00
Northwest	1354	322	23.78
Southwest	808	170	21.04
East	737	182	24.69
Southeast	711	133	18.71
North	537	105	19.55
West	480	97	20.21
South	375	104	27.73

End region

identifiers	Overall frequency	Frequency after filtering	Ratio (%)
Centre	1865	1468	78.71
Northeast	1520	0	0.00
Northwest	1347	0	0.00
Southeast	826	0	0.00
Southwest	754	0	0.00
East	580	0	0.00
North	546	0	0.00
West	479	0	0.00
South	394	0	0.00





Start region

identifiers	Overall frequency	Frequency after filtering	Ratio (%)
Northeast	1722	0	0.00
Centre	1587	1190	74.98
Northwest	1354	0	0.00
Southwest	808	0	0.00
East	737	0	0.00
Southeast	711	0	0.00
North	537	0	0.00
West	480	0	0.00
South	375	0	0.00

End region

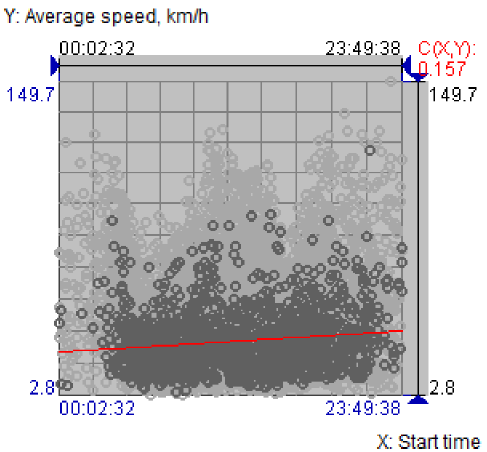
identifiers	Overall frequency	Frequency after filtering	Ratio (%)
Centre	1865	0	0.00
Northeast	1520	229	15.07
Northwest	1347	305	22.64
Southeast	826	157	19.01
Southwest	754	134	17.77
East	580	117	20.17
North	546	70	12.82
West	479	88	18.37
South	394	90	22.84

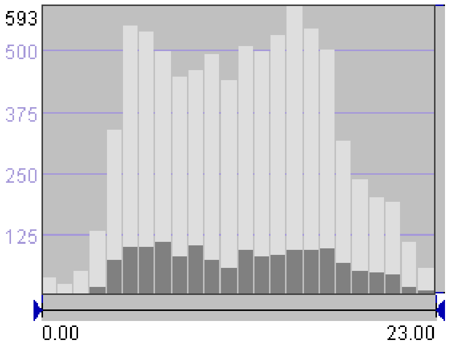
Notes concerning interactive filtering

- Works not only for trajectories but for all types of data, including spatial events, places and place-related data, and flows between places
- Not only maps but all data displays react to filtering and show its results

⇒ Hence, filtering can be used to explore data by portions using multiple displays that show different aspects of the data

- There are also other types of interactive filters
- There are many other types of interactive exploratory techniques





Exercises on filtering

Data: trajectories divided into trips by 10 minutes break

- Select trajectories passing near the Linate airport on the east of the city. How many such trajectories exist?
- Select trajectories visiting the regions Northwest, North, and Northeast. Select trajectories visiting these regions in the given order, then in the opposite order. Select trajectories that did not visit any of these regions.
- Find how many cars were under way in the time intervals 03:00-05:00, 05:00-07:00, 12:00-14:00, 18.00-20:00
- Select trajectories with the length of at least 25 km. How many of them visit the region Centre?

Data Transformations

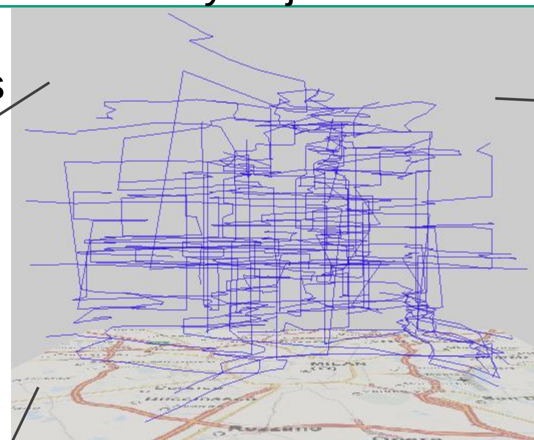
Variety of data transformations

for a variety of purposes

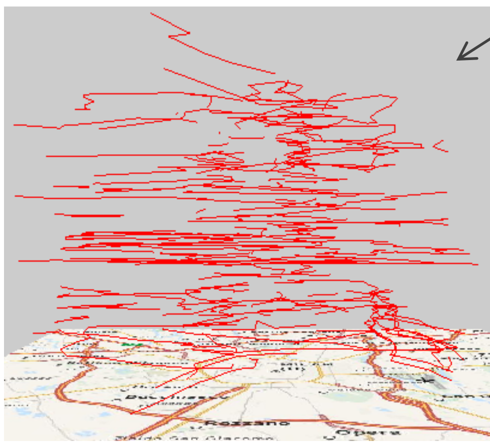
- For supporting abstraction
  - **Simplification**: reduce excessive detail and high-resolution fluctuations
  - **Generalization**: transform to larger units
    - E.g., time moments → intervals; points → areas; individual nominal values → categories
  - **Grouping** of similar and/or close items: elements → subsets
  - **Aggregation**: summarize values by larger units or by groups of items
- For managing large data volumes
  - **Sampling; generalization; grouping; aggregation**
- For obtaining task-relevant information
  - **Computation of derived attributes**
  - **Feature extraction**
    - E.g., trajectories → movement events, time series → trends, peaks, pits, ...

Some examples of data transformations

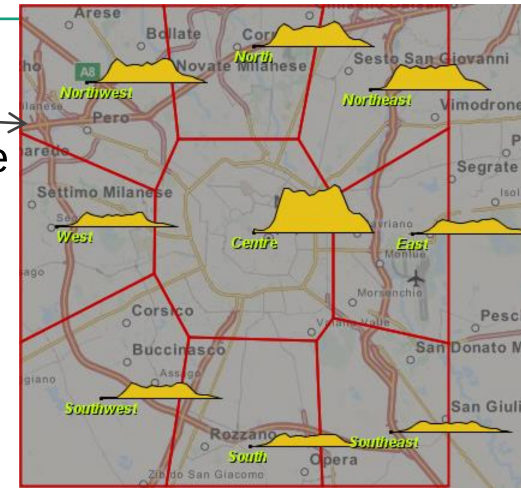
Daily trajectories



divide into trips

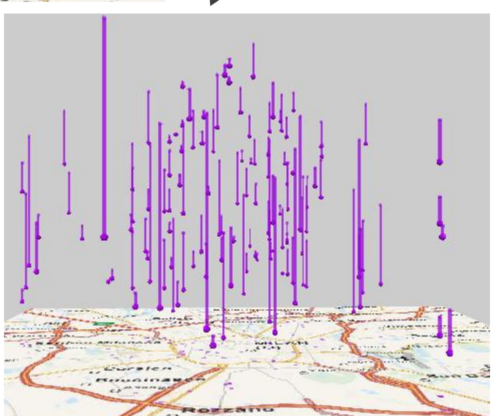


aggregate



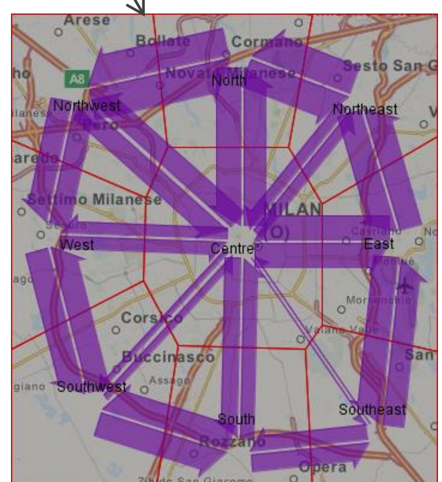
presence of cars in regions by time intervals

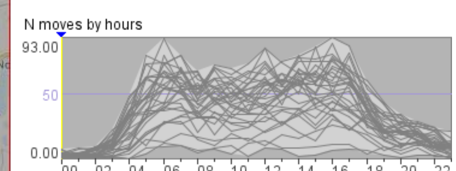
extract events



e.g., stops

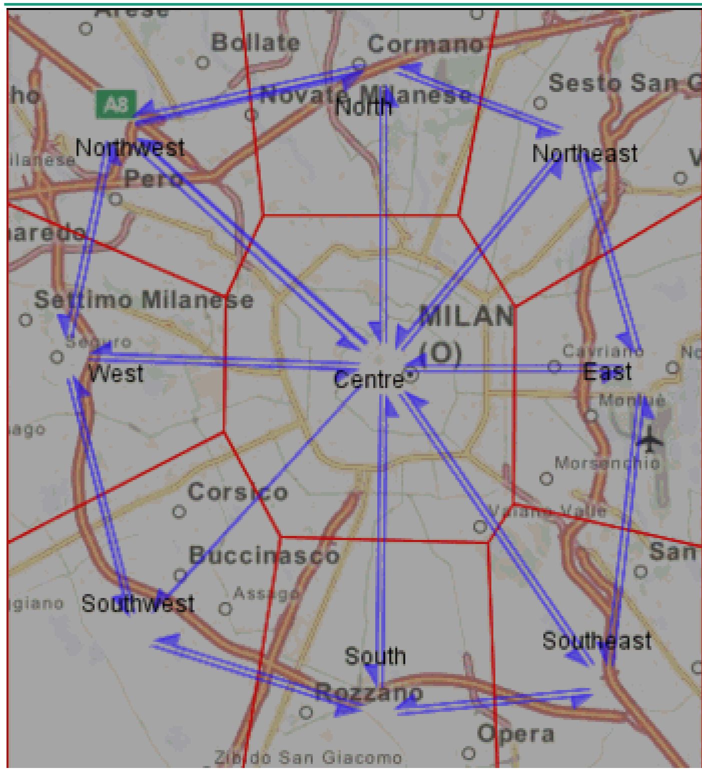
aggregate





flows (aggregate moves) of cars between regions by time intervals

(Aggregate) moves, or flows



Link *(short for 'spatial link')*::= a spatial object representing directed relation, such as movement, between 2 locations.

A link is specified by a pair (origin place, destination place).

Links may have attributes such as number of moving objects, number of transitions, movement speed, ...

Links with attributes describing collective movements are called **flows**.

equal flows in two opposite directions

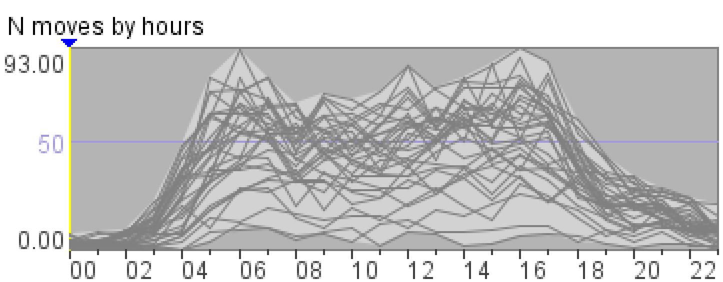
Flows may be represented on a map by half-arrow symbols with widths proportional to numeric attribute values.

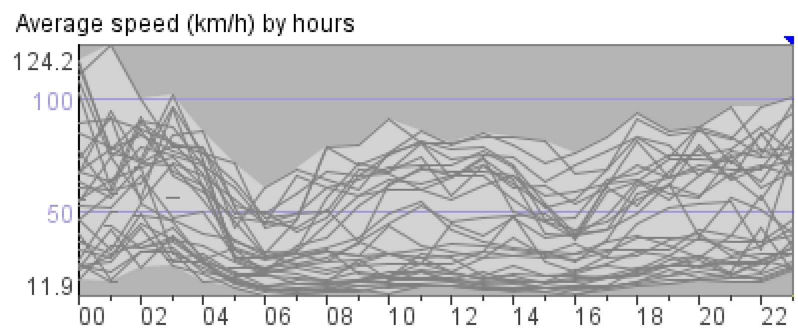
unequal flows in two opposite directions

Spatial time series of flows::= attribute values of the links in different time moments or intervals:

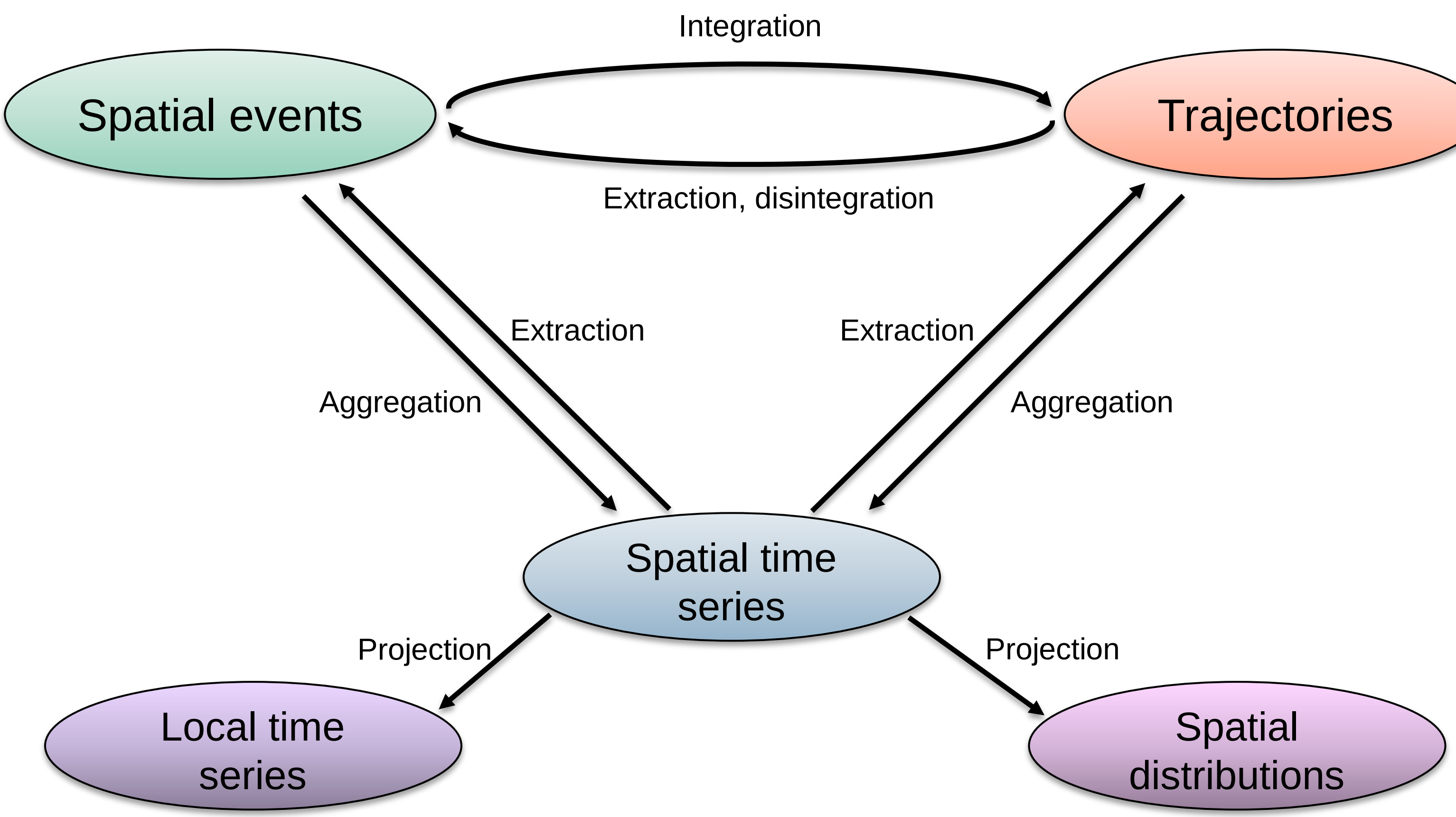
((origin, destination), time) $\rightarrow$

object count, transition count, speed, ...



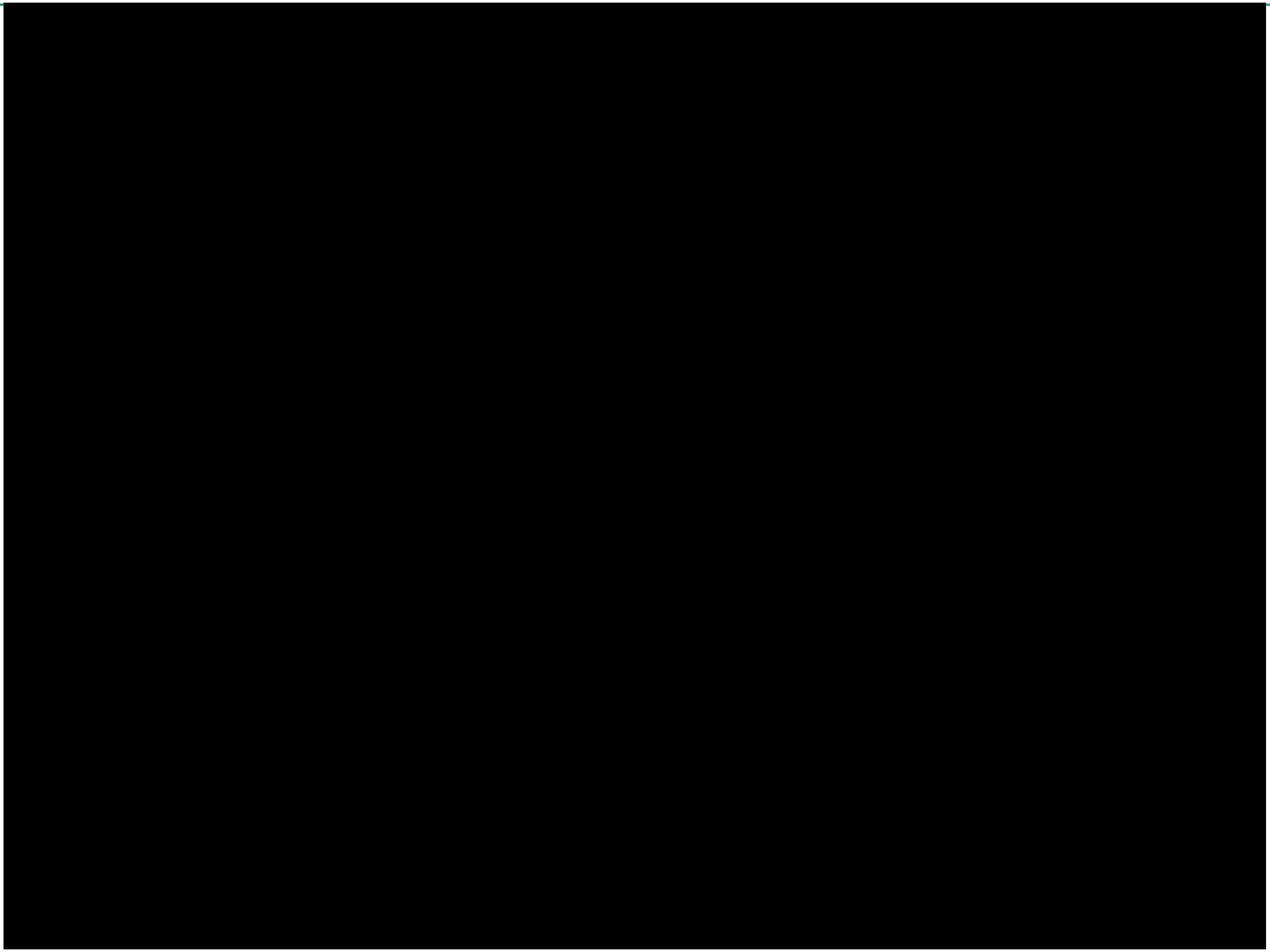


Transformations of spatio-temporal data structures



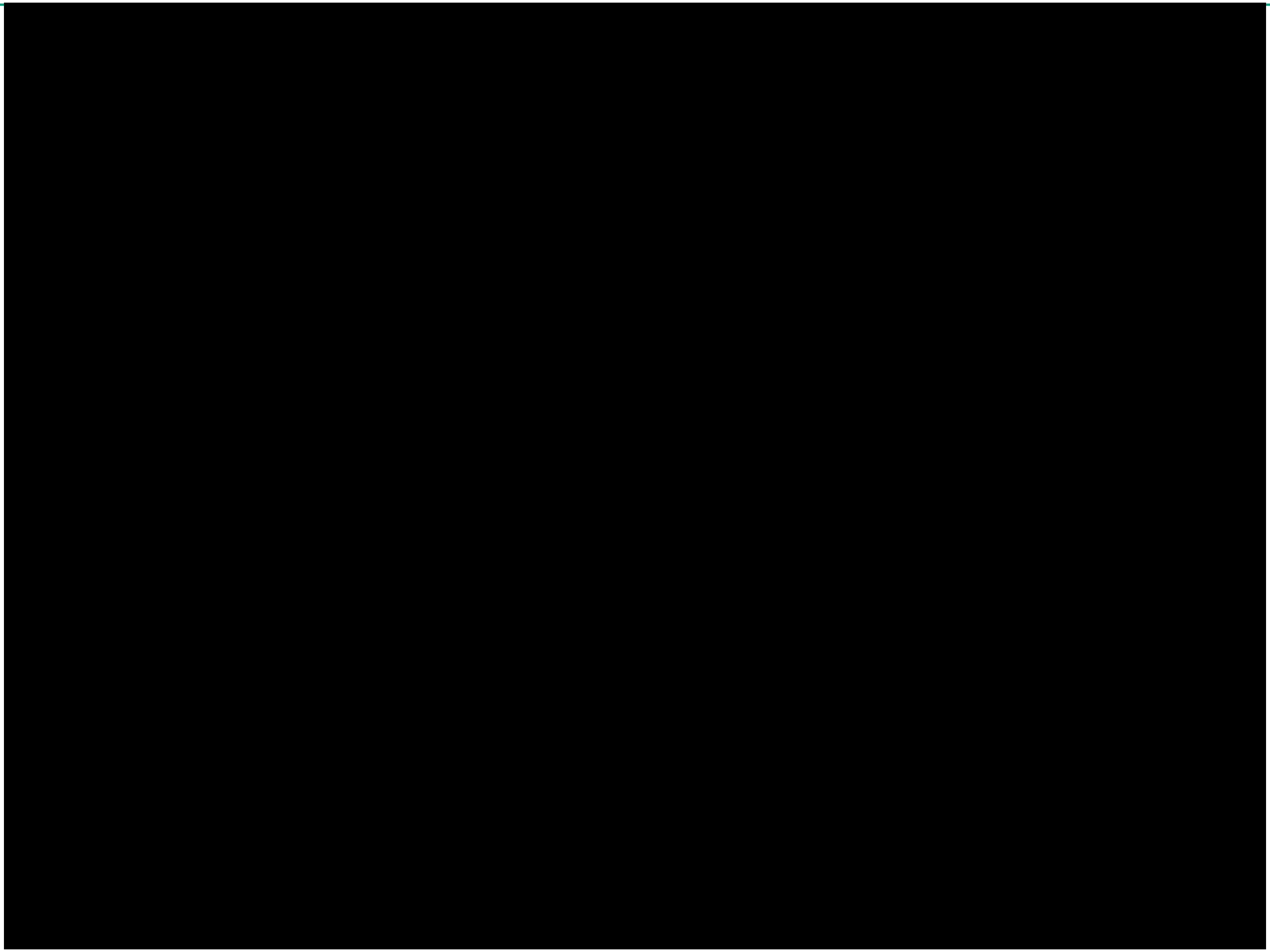
VA support to data transformations

Example: generalization and spatio-temporal aggregation of trajectories



VA support to data transformations

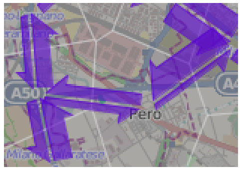
Example: extraction of spatial events from trajectories



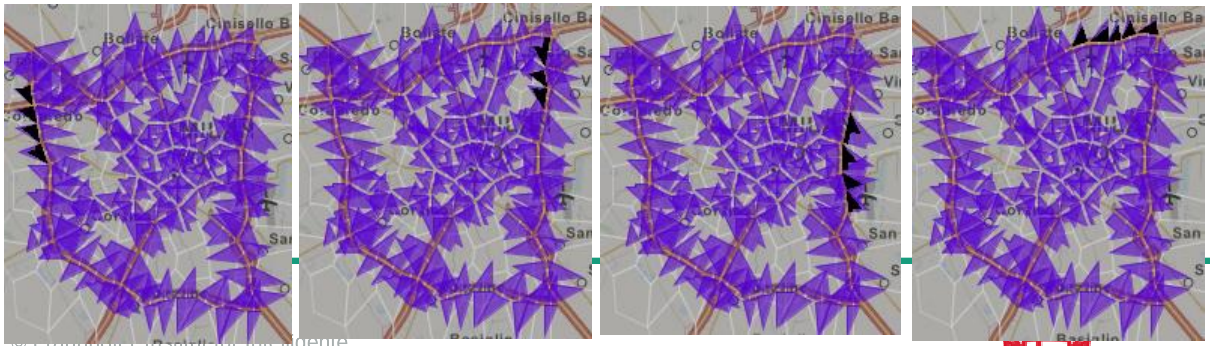
Exercises with transformed data (1)

Data: spatial time series of flows obtained by aggregating trajectories

- Visualise the total numbers of moves between places by line thicknesses (widths of flow symbols)
- Select (by filtering) the links where the total N of moves is not less than 200
- Visualise the dynamics of the average speeds by line thicknesses on an animated map display
- By animating the map, find cases when movement speeds between two places in two opposite directions substantially differ



- Visualize the speed dynamics on a time graph. Select the links marked on the images and describe the respective speed dynamics over the day.



Exercises with transformed data (2)

Data: spatial point events obtained by extracting trip ends from trajectories

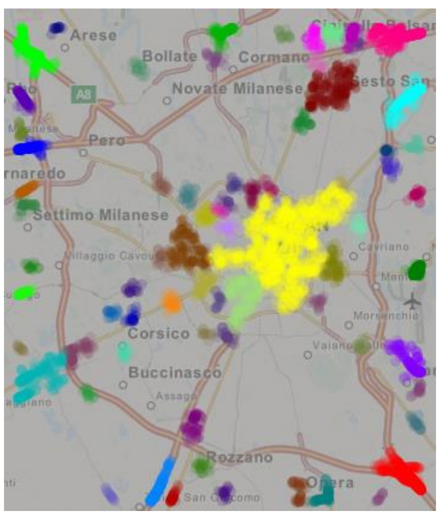
- Observe the spatial distribution of the events on a map. Where do you see event concentrations?
- Observe the spatio-temporal distribution of the events in a space-time cube (STC). What patterns in the STC correspond to the point concentrations on the map?
- Using “spatial window” filter, select the point concentration near the Linate airport on the east. Describe the corresponding pattern in the STC. Are the events distributed evenly in time or there are temporal concentrations and temporal gaps?

- Aggregate the events by the regions and hourly time intervals. Visualize the time series of event counts on a map by temporal diagrams (“value flow diagrams”). What regions have the largest numbers of events? In what regions there are peaks in the morning, in the afternoon, both?

Clustering

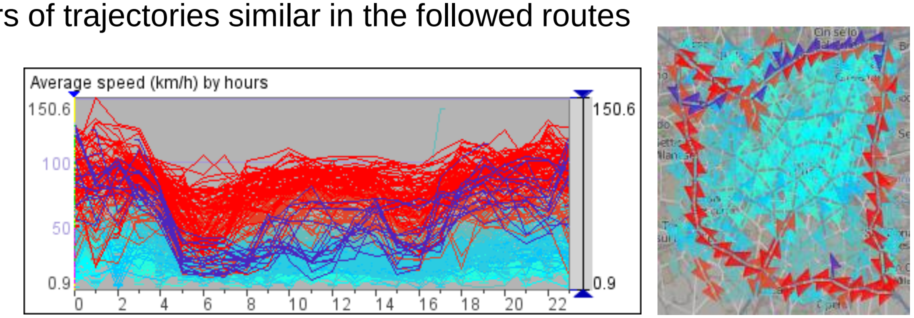
What is clustering?

- Loose definition: clustering is the process of organising objects into groups whose members are close or similar in some way.
- A cluster is a group of objects which are "similar" or "close" between them and are "dissimilar" or "distant" to the objects belonging to other clusters.



Example: clusters of spatially close points

Example: clusters of trajectories similar in the followed routes



Example: clusters of similar time series and clusters of links similar in their time series

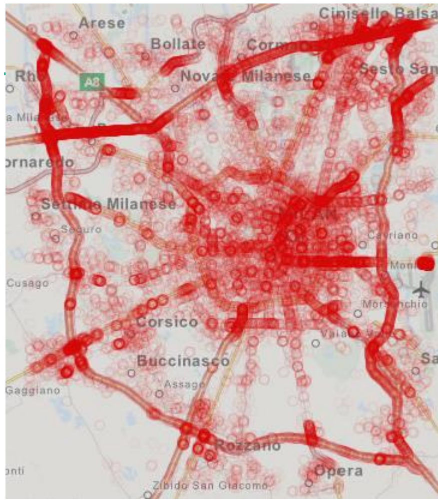
Role of clustering in VA

- Grouping of similar or close items plays an essential role in VA
  - as a tool supporting <u>abstraction</u>: elements $\rightarrow$ subsets; the subsets may be considered as wholes
  - as a tool to manage <u>large data volumes</u>
  - as a tool to <u>study the behaviour</u> of attributes (i.e., distribution of attribute values) over the set of references, particularly,
    - multiple attributes
    - spatial and spatio-temporal positions of objects (events)
    - dynamic (time-variant) attributes, such as
      - time series of numeric values
      - trajectories of moving objects

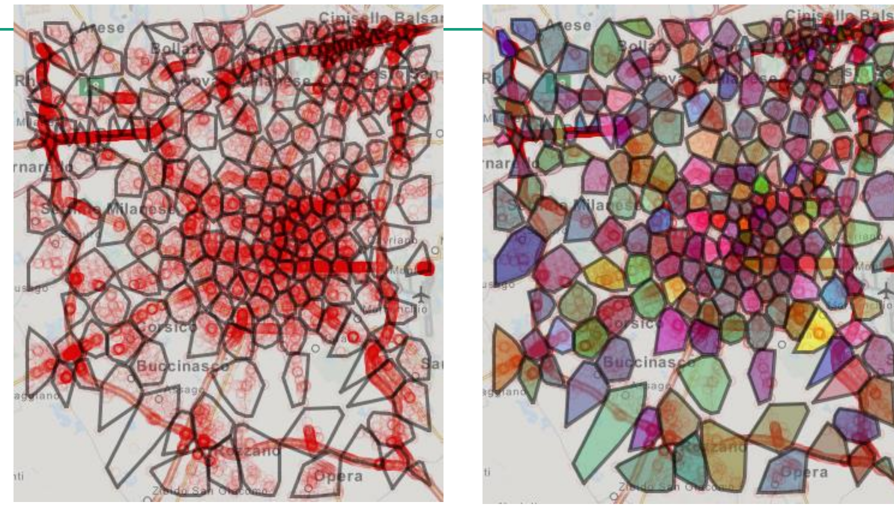
Two major types of clustering

- **Partition-based clustering**: divide items into groups so that items within a group are similar (close) and items from different groups are less similar (more distant)
  - Examples: k-means, self-organizing map
  - Property of the result: each item belongs to some group
- **Density-based clustering**: find groups of highly similar (close) items and separate from them items that are less similar (more distant) to others
  - Examples: DBScan, OPTICS
  - Properties of the results: some items belong to groups, other items remain ungrouped and are treated as “noise”

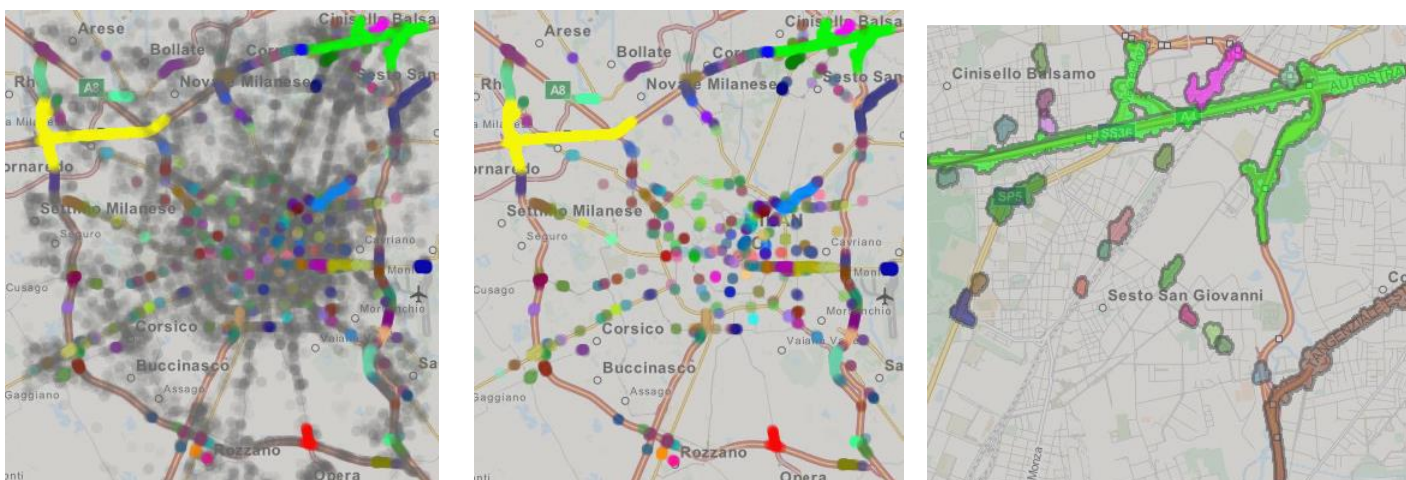
Two major types of clustering: an example



Partition-based: convex clusters including all objects



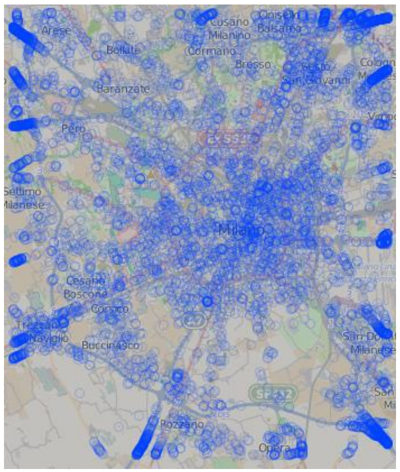
Density-based: dense clusters of arbitrary shapes; many objects are treated as “noise” (gray)



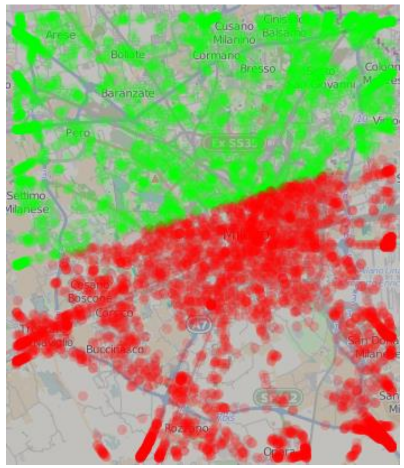
Partition-based clustering

k-means: partitions data into k groups (k is a parameter specified by the user)

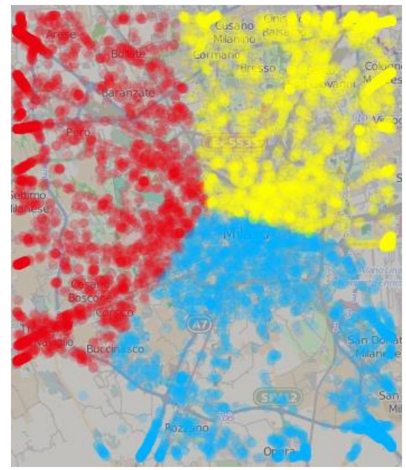
Data: 2D points (X,Y)



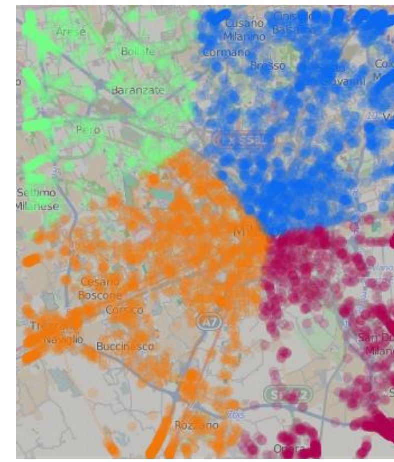
K=2



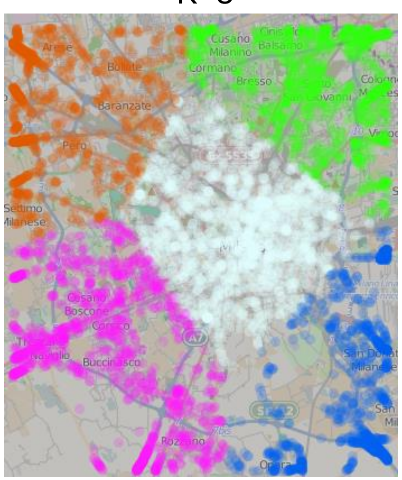
K=3



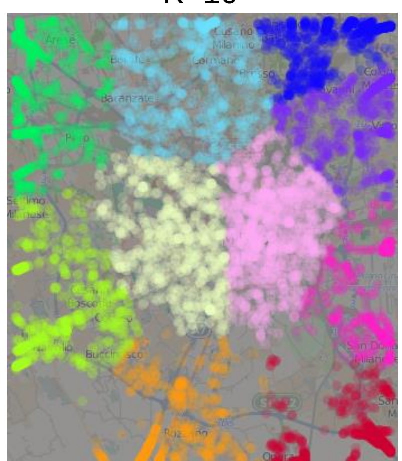
K=4



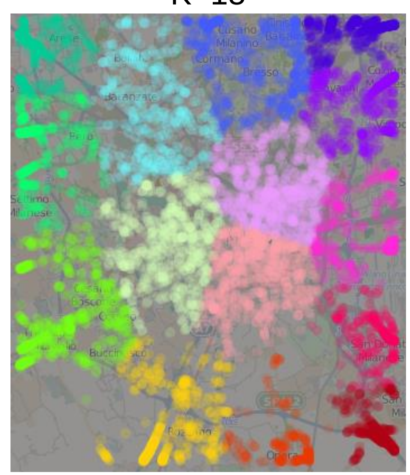
K=5



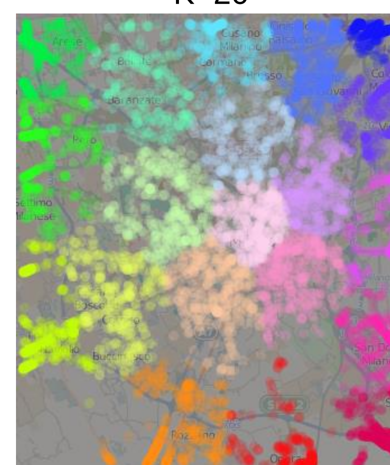
K=10



K=15



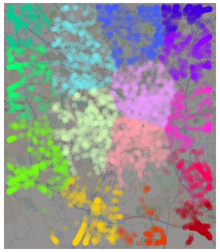
K=20

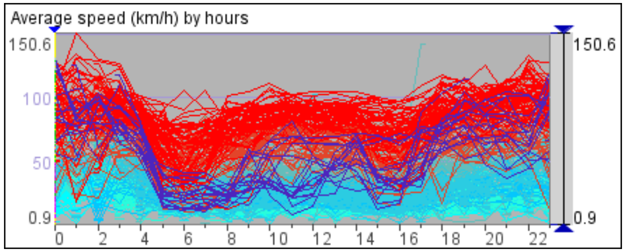


Problem: what value of k to choose?

In general: for any computational technique, what parameter settings to choose?

- Typically not known in advance
- Computation results (in particular, clusters) need to be properly visualised, to allow examination by the user
- The user needs to run the tool with different settings and see how the results change
- The user then selects the settings bringing the “best” results:
  - easy to interpret (e.g., understandable spatial patterns)
  - internal variance within the clusters is sufficiently low
  - fit to the purpose (e.g., the intended analysis scale may require coarser or finer division)
- Different visualizations are needed for different types and structures of data
  - Clustering results are often represented by colour-coding, which is applied to different visual objects, depending on the structure of the input data



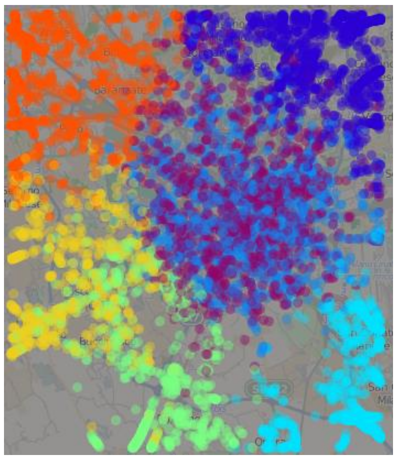


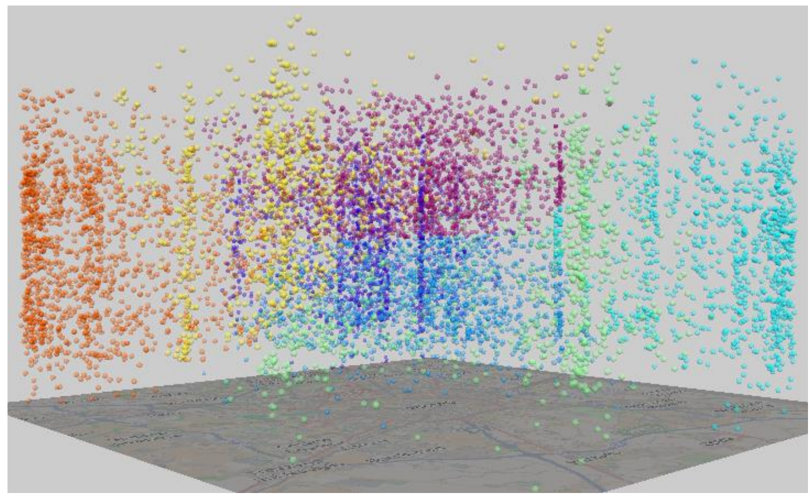
Visualization of clustering results

A single display may be not enough

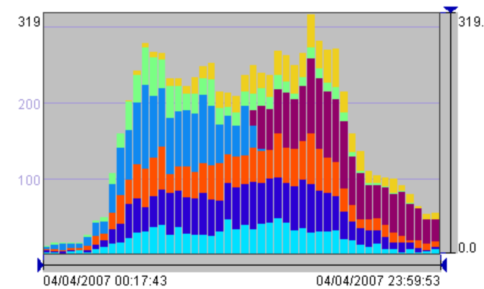
K-means clustering of spatial events according to the spatial and temporal positions (x, y, time)

K=7

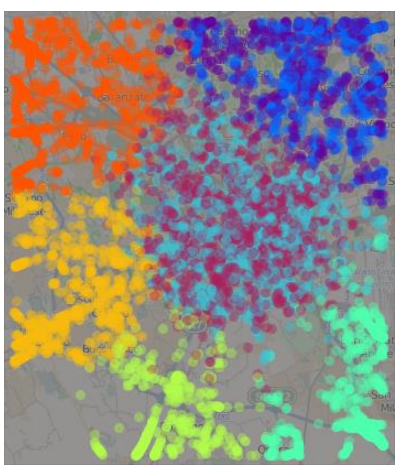


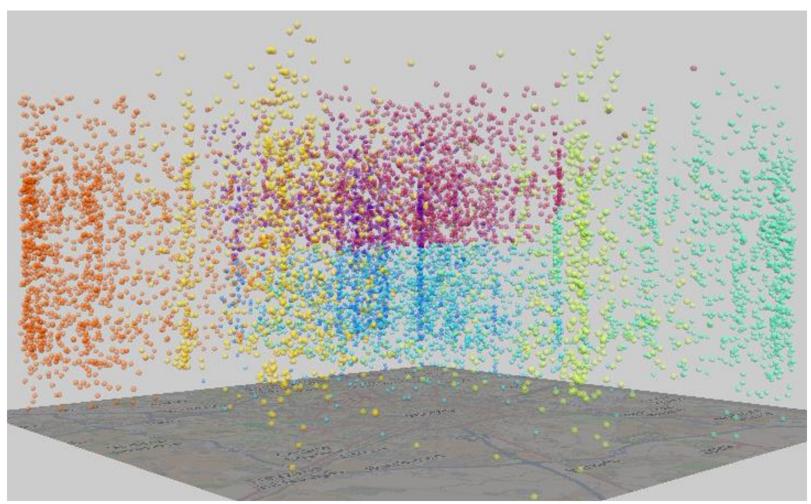


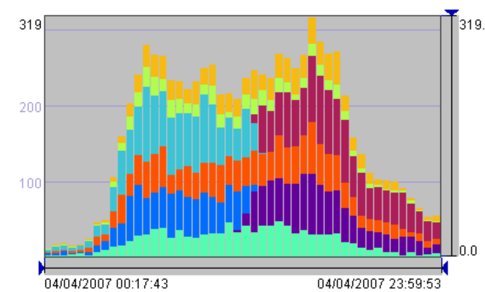
Temporal frequency histograms show the temporal relations between the clusters in a clearer way than the space-time cube



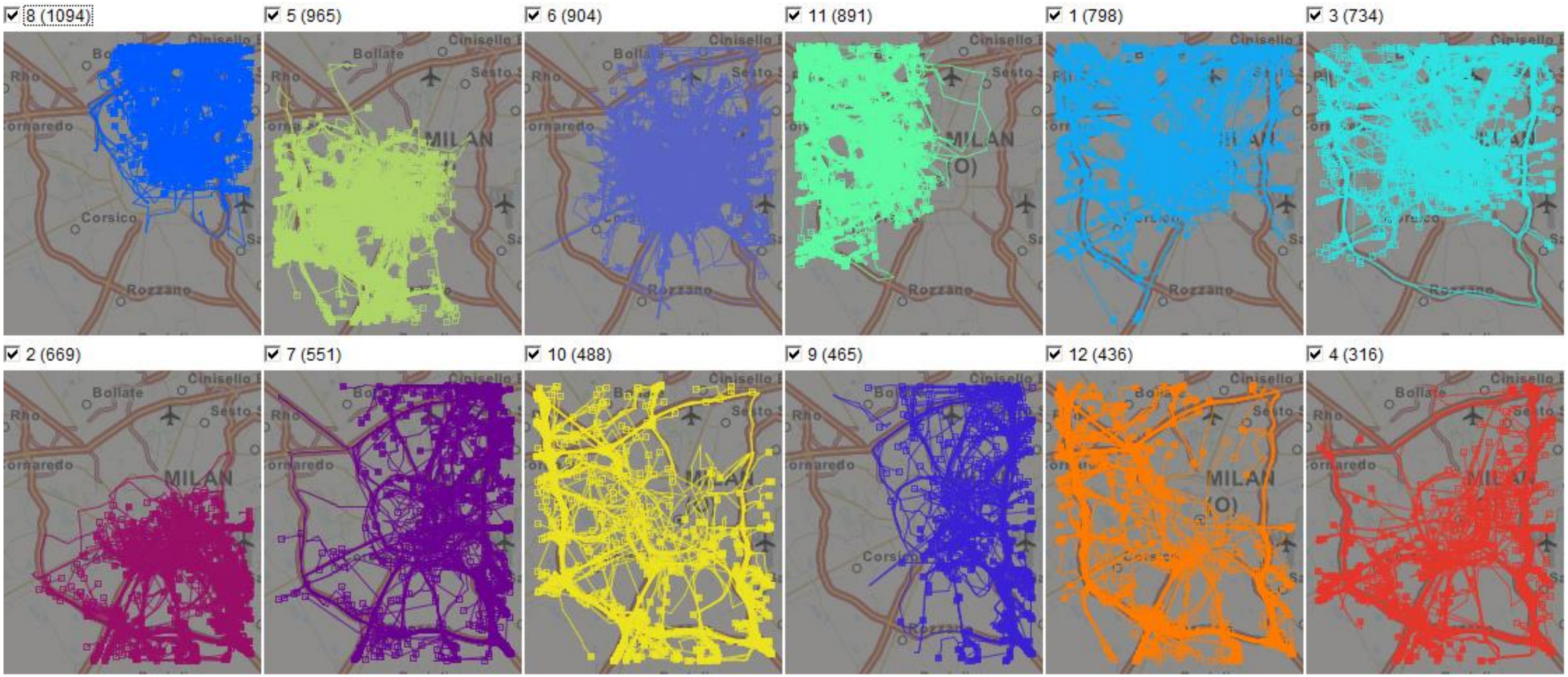
K=8





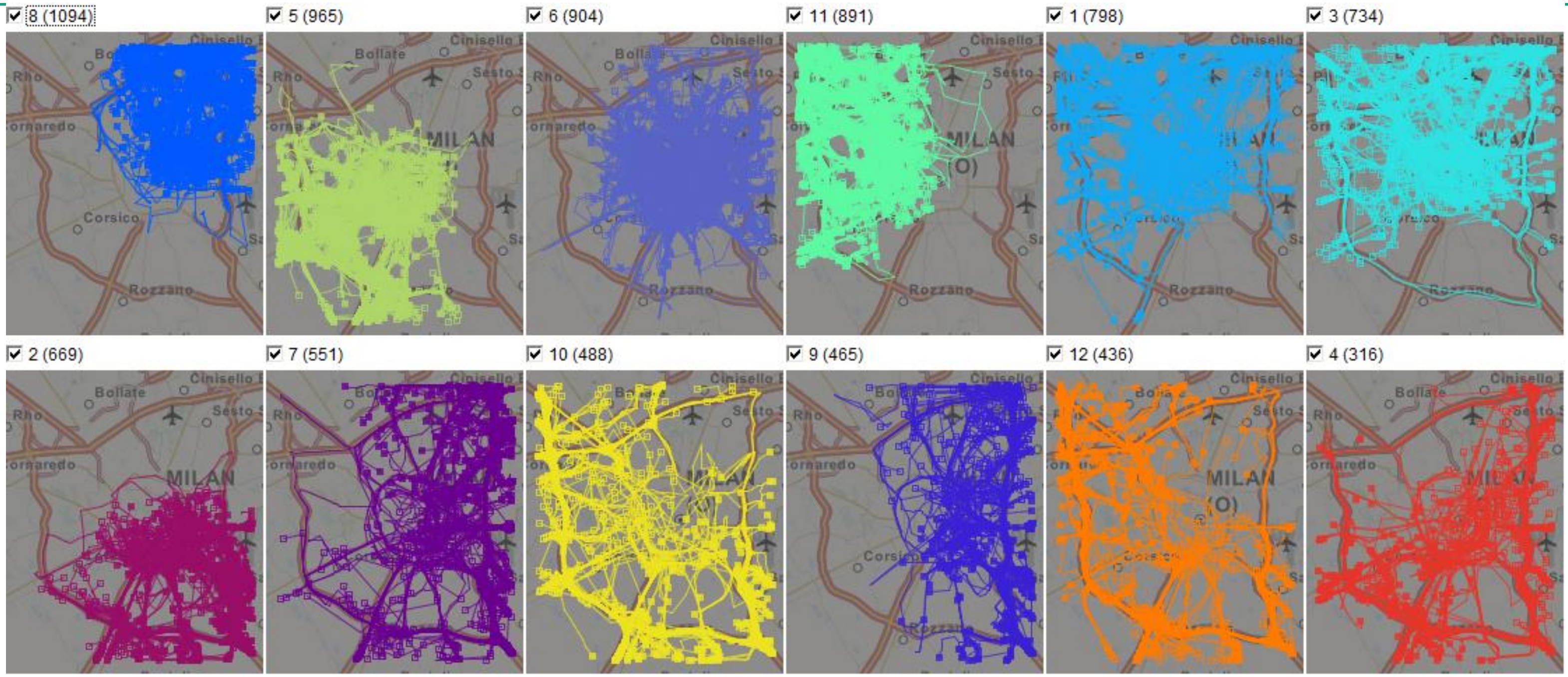


Clustering may be applied to multiple and diverse attributes

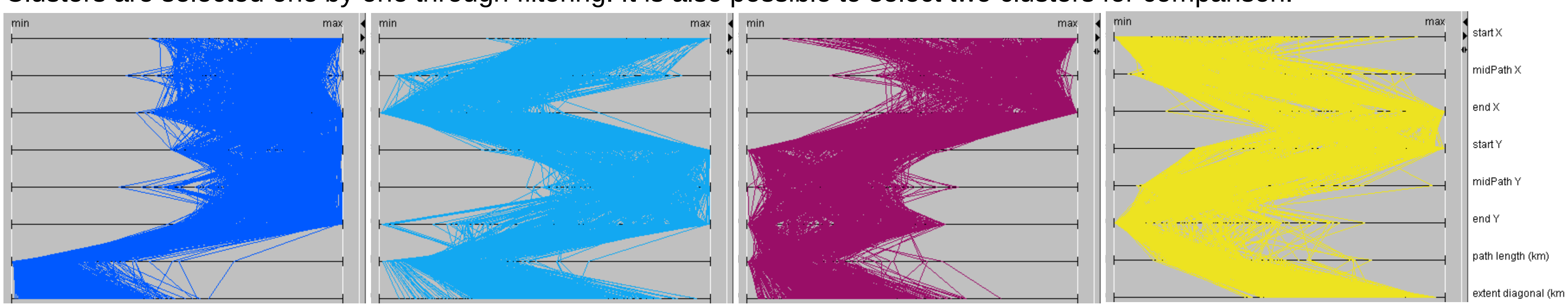


Example: 12 k-means clusters of trajectories grouped according to the x- and y-positions of the start points, end points, and points in the middle of the paths, plus path lengths and spatial extents (i.e., lengths of the bounding rectangle diagonals).

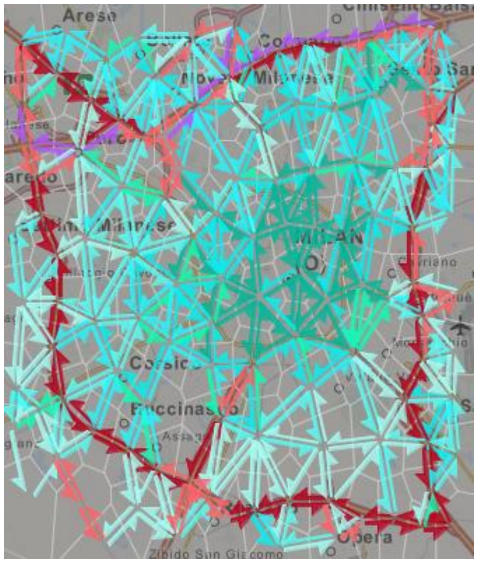
Clusters by multiple attributes may be explored and interpreted using parallel coordinates plot

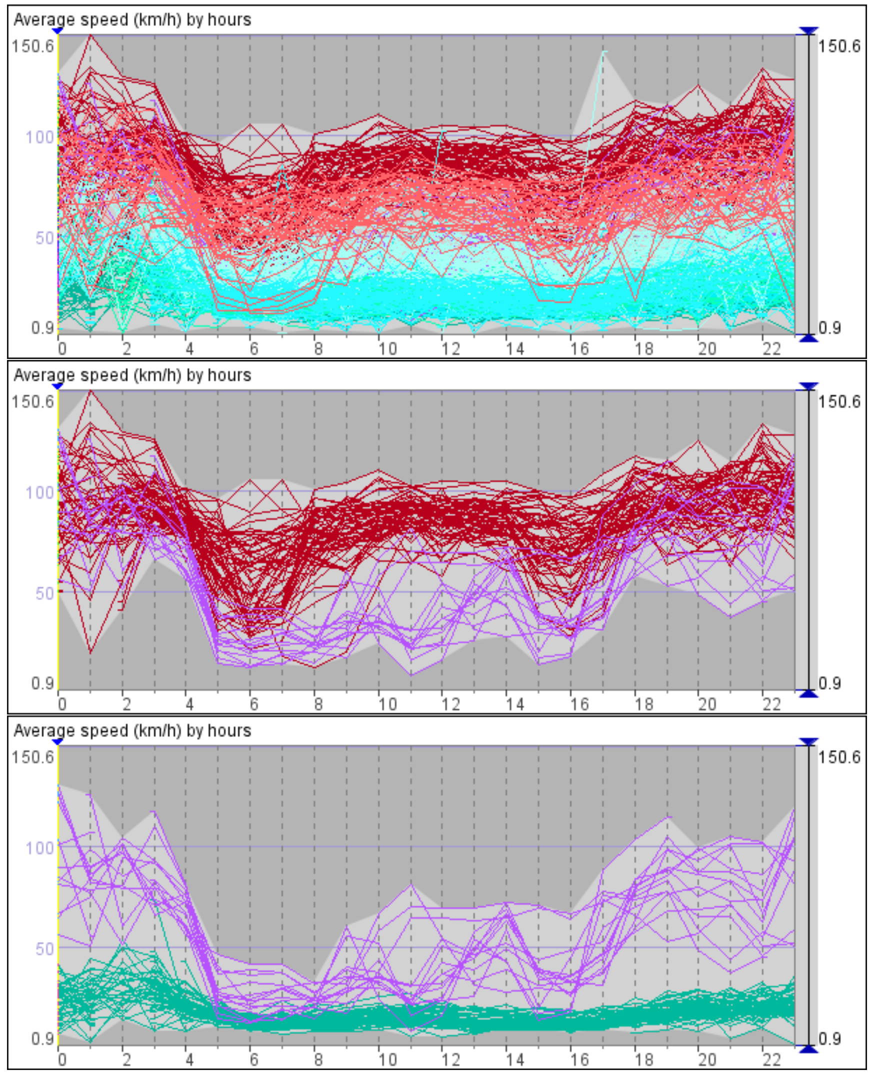


Clusters are selected one by one through filtering. It is also possible to select two clusters for comparison.

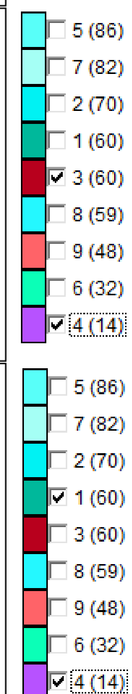


Clusters by time series of numeric attribute values can be explored using a time graph





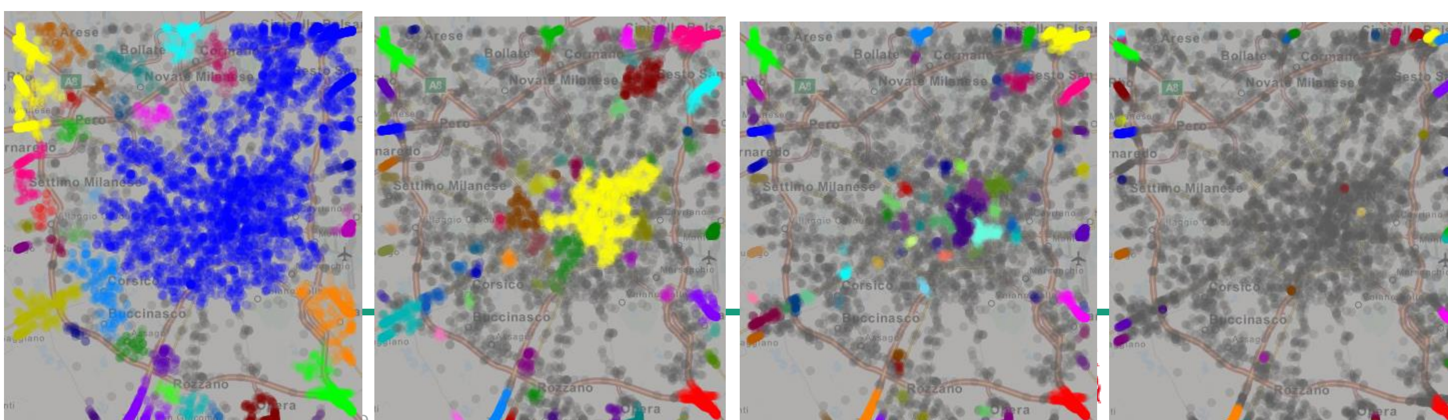
Clusters can be selected one by one for exploration and interpretation or in pairs for comparison.



Density-based clustering (DBC)

Goal: find dense groups of close or similar objects

- An object is treated as a **core** object of a cluster if there are at least N objects within the distance (radius) R around it. These objects are called **neighbours**.
- To make a cluster, (1) some core object with all its neighbours is taken; (2) for each core object already included in the cluster, all its neighbours are also added to the cluster (if not added yet).
- Some objects may remain out of any cluster (when they have not enough neighbours and do not belong to the neighbourhood of any core object). These objects are treated as “noise”.
- For DBC, the user needs to specify the neighbourhood radius R and the minimum number of neighbours N. Therefore, the use of DBC requires an understandable definition of **distance** between objects, e.g., spatial distance or spatio-temporal distance. It may be hard to choose R for a more abstract “distance” between combinations of values of multiple diverse attributes.
- Results of DBC greatly depend on the parameter choice. Visualisation and interactive exploration help to find suitable values for R and N that lead to good results.

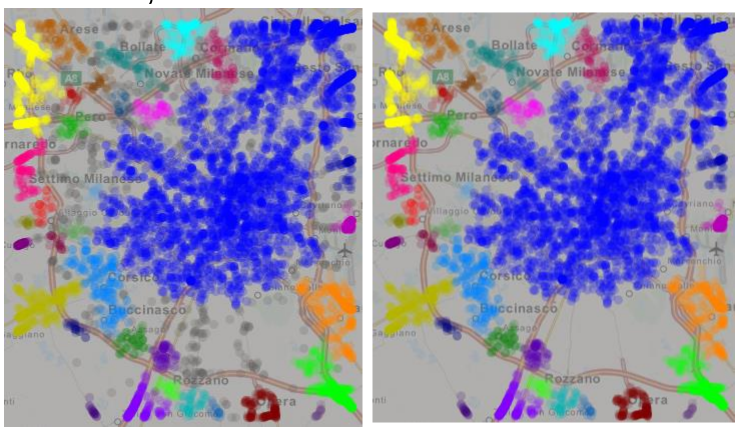


Grey: “noise”

Exploring the impact of the DBC parameters

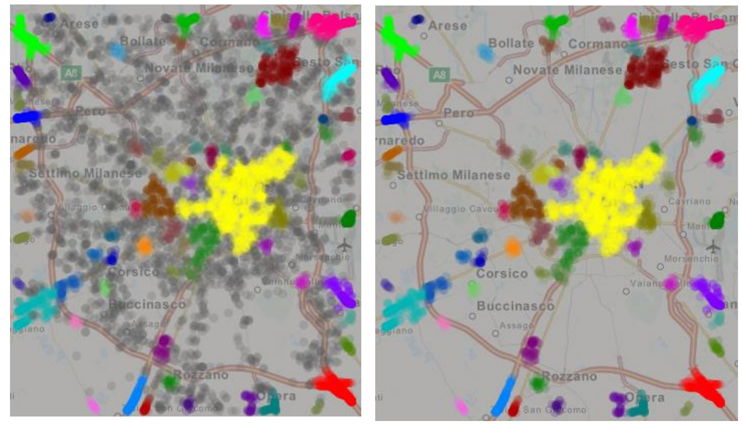
Example: DBC according to the spatial distances between points (trip ends)

R=500m, N=10



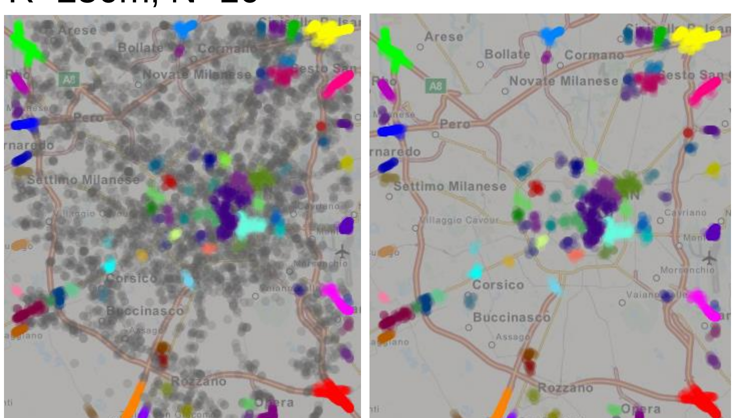
The clusters are too loose and too extended in space.

R=300m, N=10



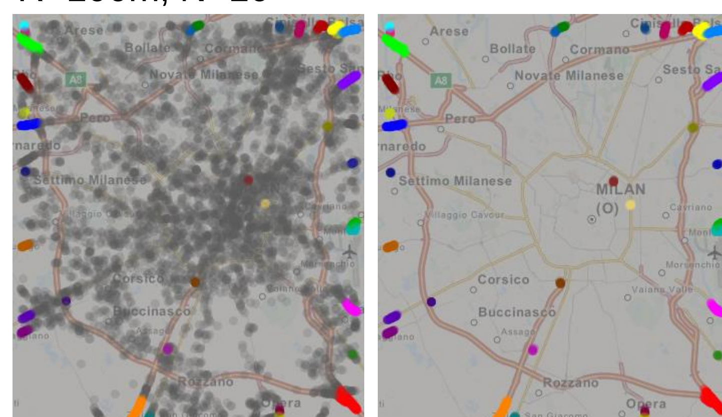
Some clusters are still too loose.

R=250m, N=10



The clusters are more or less OK.

R=100m, N=10



The clusters are nicely compact but, possibly, too small and too few.

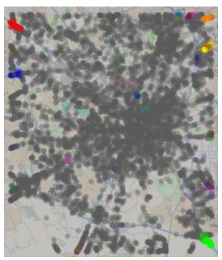
Spatio-temporal distance in DBC

Example: clustering of trip ends according to distances in space and time

For any two objects, there is a distance in space d_{space} and a distance in time d_{time}.

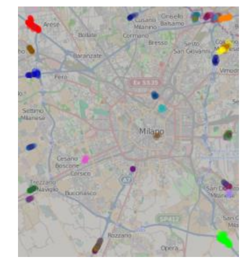
The user specifies two neighbourhood radii R_{space} and R_{time}, e.g., R_{space}= 300 m and R_{time} = 30 minutes.

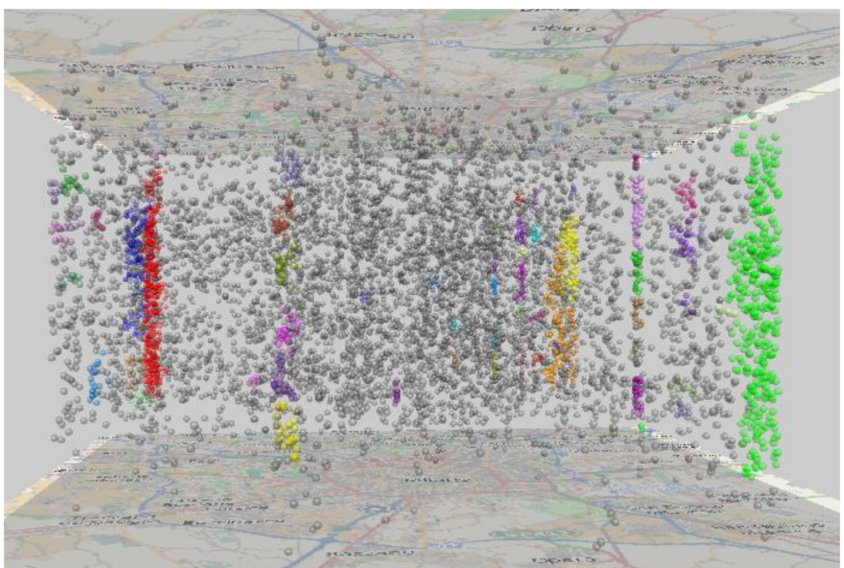
The clustering algorithm requires a single distance and a single radius; therefore, spatial and temporal distances need to be combined together, for example, as $d = \max(d_{space}/R_{space}, d_{time}/R_{time}) * R_{space}$.

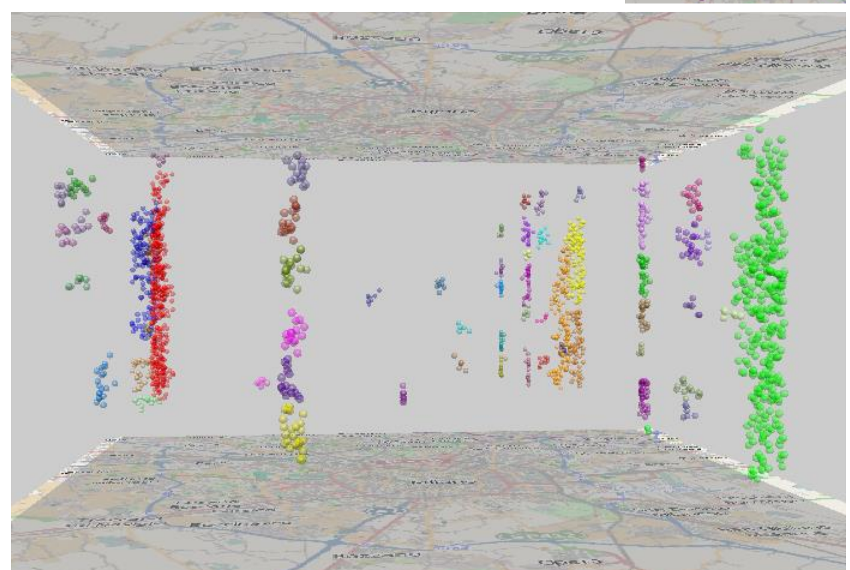


Spatio-temporal clusters of trip ends have been obtained with R_{space}= 300 m, R_{time} = 30 minutes, and N = 5. That is, two events are treated as neighbours if the distance in space between them is not more than 300m and the distance in time is not more than 30 minutes. A core object of a cluster must have at least 5 neighbours.

Some clusters last in time for almost the whole day, others have shorter life times. There are clusters sharing the same area in space but disjoint in time.







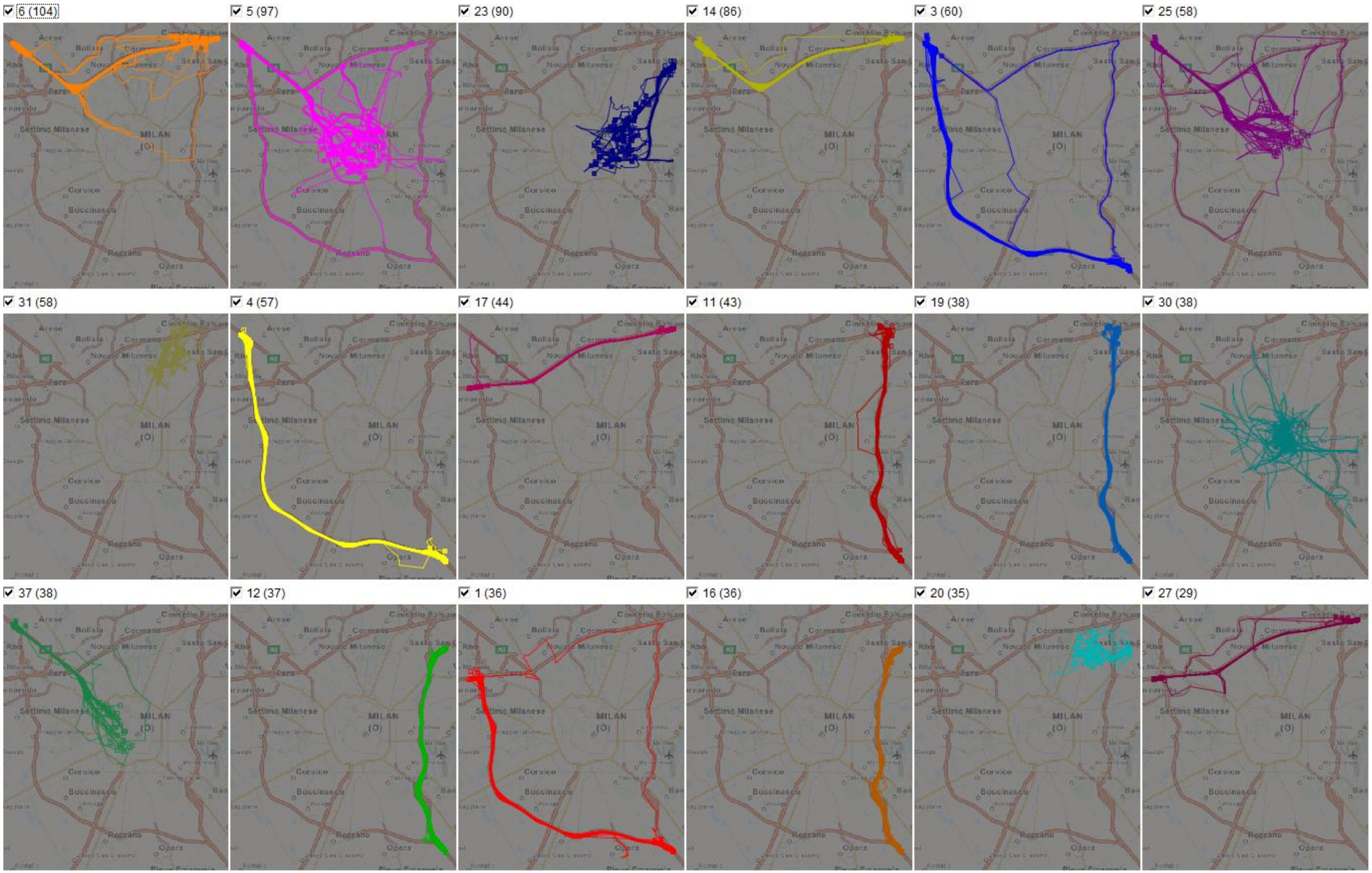
Distances between trajectories

- Trajectories are complex objects consisting of multiple spatio-temporal points, having origins and destinations, particular shapes, lengths, durations, and dynamically changing movement directions and speeds.
- It is hardly possible to define a distance measure that accounts for all these properties. Even if it could be defined, it would be hard to understand. Hence, it would be quite difficult to choose a meaningful value of R for clustering (as in the case of multiple diverse attributes).
- It is more feasible to create a library of simple distance measures (a.k.a. distance functions) addressing different properties. For example,
  - spatial distance between origins and/or between destinations,
  - average spatial distance between corresponding points along the routes,
  - average spatial distance between points reached at the same times, ...
- Different aspects of trajectories are studied using different distance functions.

DB clusters of trajectories (example 1)

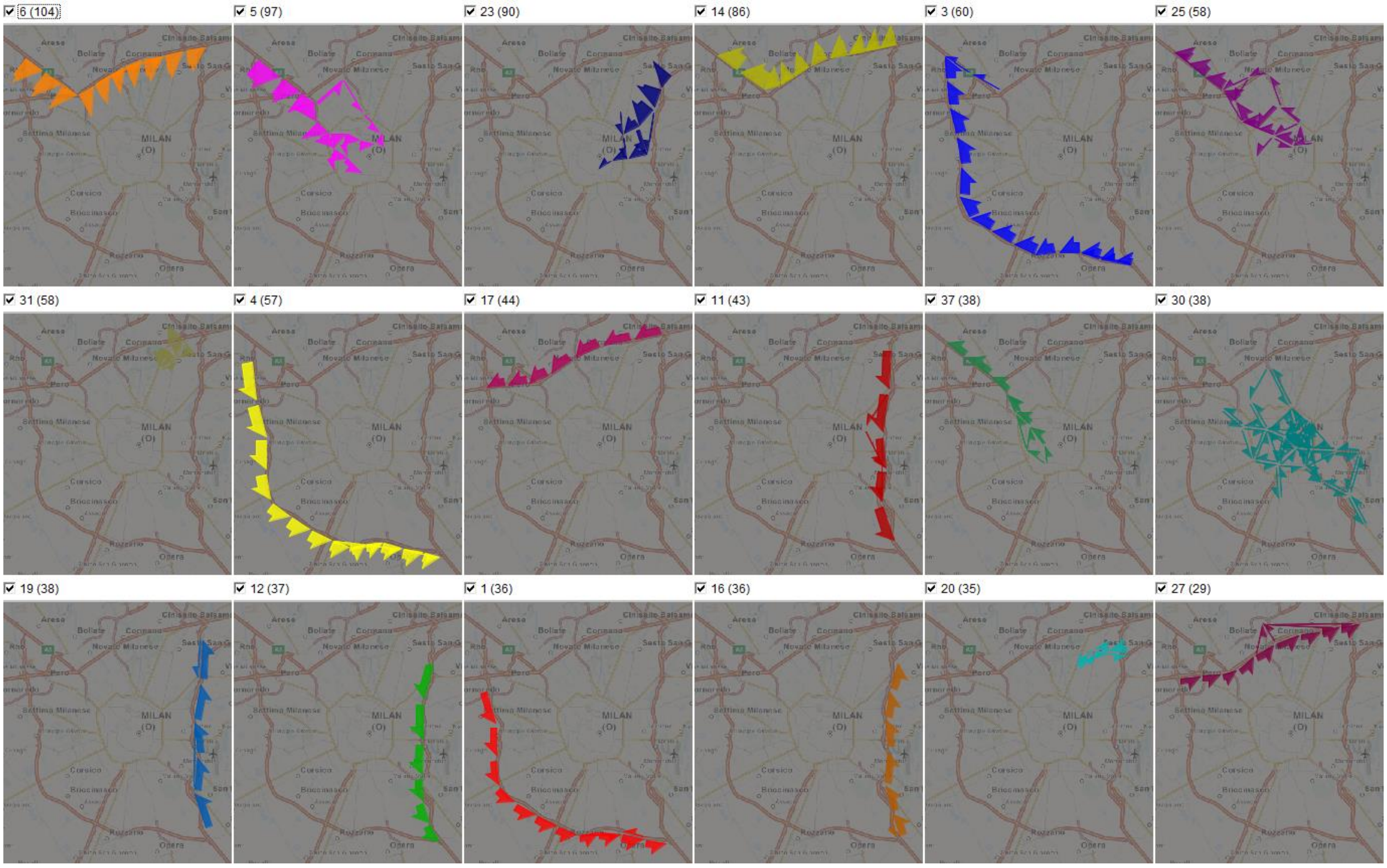
Distance function: the average spatial distance between the origins and between the destinations; R=750m, N=5

Only 18 largest clusters are shown.



Summarised representation of clusters of trajectories

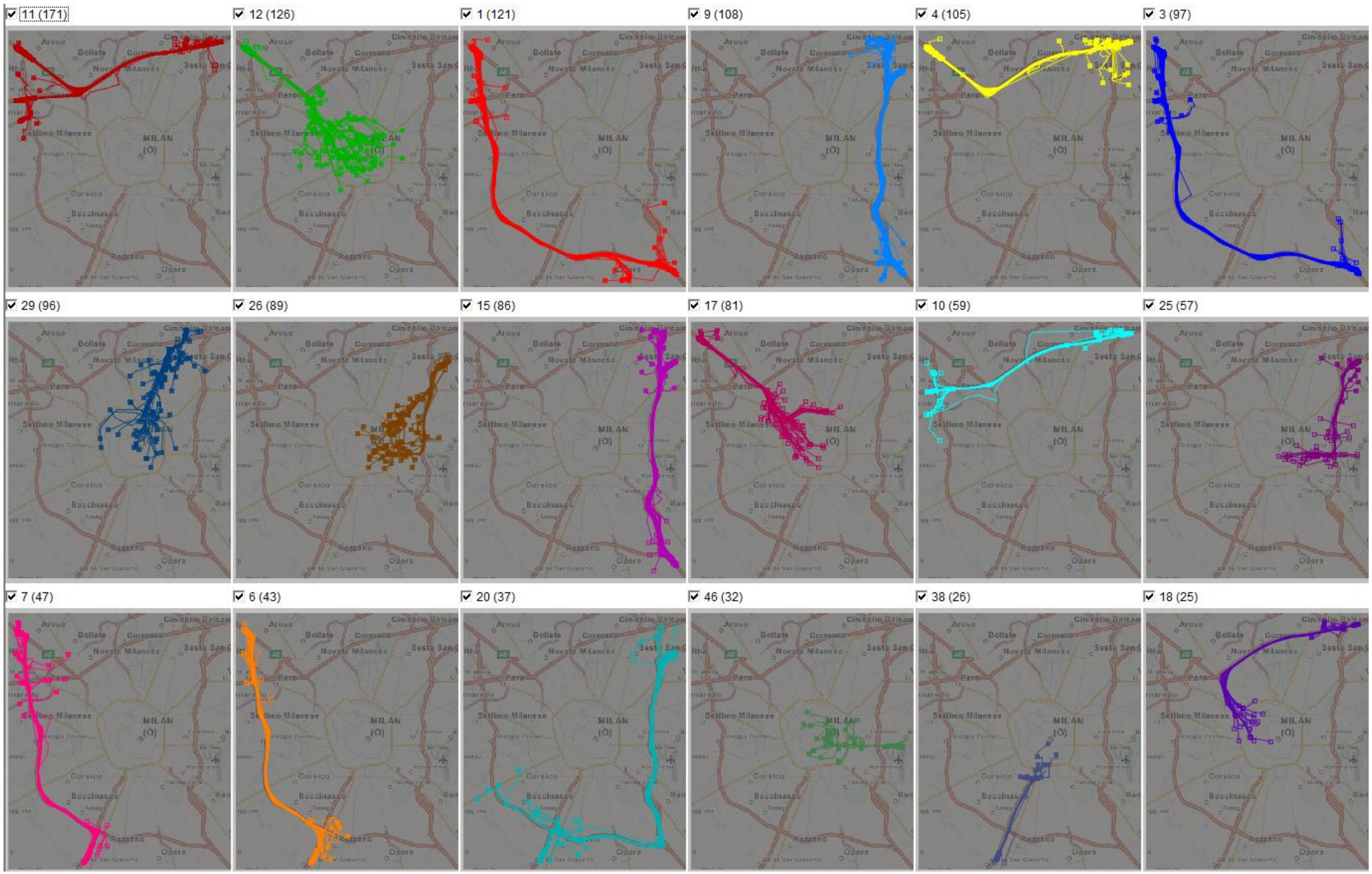
Minor flows are omitted for a clearer view.



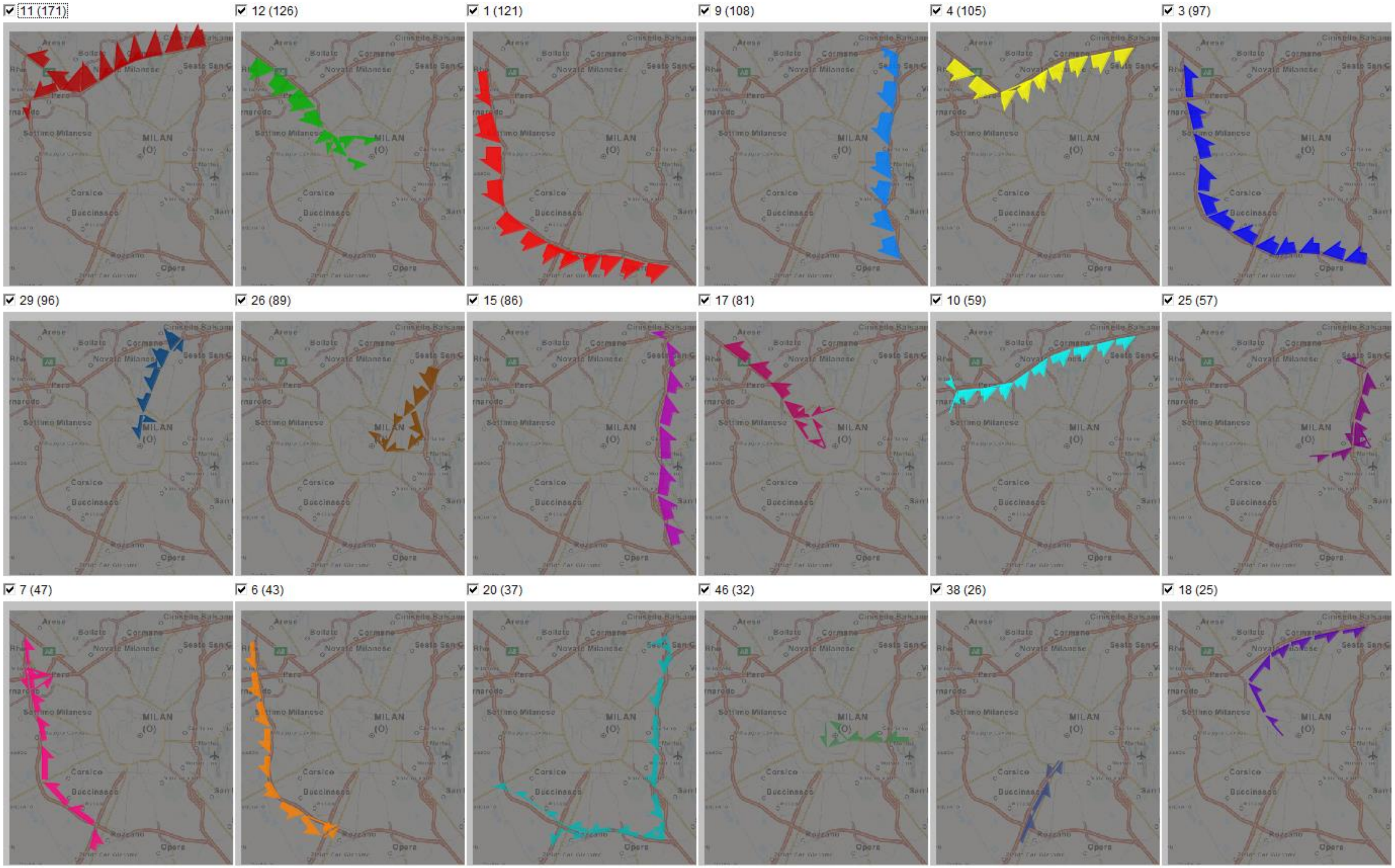
DB clusters of trajectories (example 2)

Distance function: "route similarity", i.e., the average spatial distance between the corresponding points along the route; R=750m, N=5

Only 18 largest clusters are shown.



The same clusters represented in a summarised form



Two major types of clustering: a reminder

- **Partition-based clustering**: divide items into groups so that items within a group are similar (close) and items from different groups are less similar (more distant)
  - Examples: k-means, self-organizing map
  - Property of the result: each item belongs to some group
- **Density-based clustering**: find groups of highly similar (close) items and separate from them items that are less similar (more distant) to others
  - Examples: DBScan, OPTICS
  - Properties of the results: some items belong to groups, other items remain ungrouped and are treated as "noise"

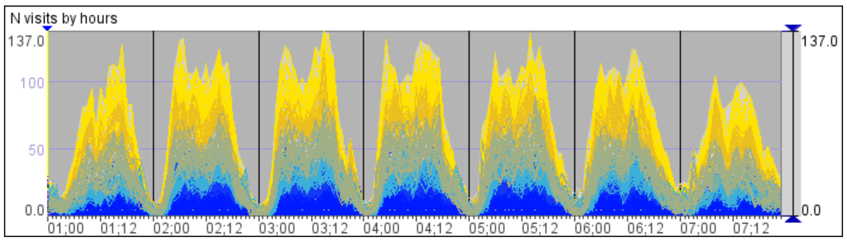
Use of the two types of clustering

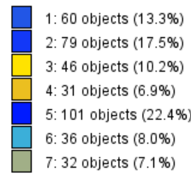
- **Partition-based:**
  - unites elements into subsets for
    - abstraction
    - decomposition of the analysis task and reduction of analytical workload
    - data aggregation
  - integrates multiple attributes, allows comparison of items in terms of multiple attributes
- **Density-based:**
  - separates what is common, frequent from what is specific, infrequent
    - may be a tool for studying attribute behaviours (distributions)
    - concentrations of close/similar objects may have special meanings
      - e.g., spatio-temporal cluster of low speed events $\Rightarrow$ traffic jam

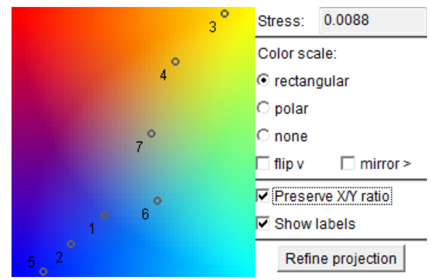
Interactive visual support to clustering

Trying various parameter setting; studying parameter impact

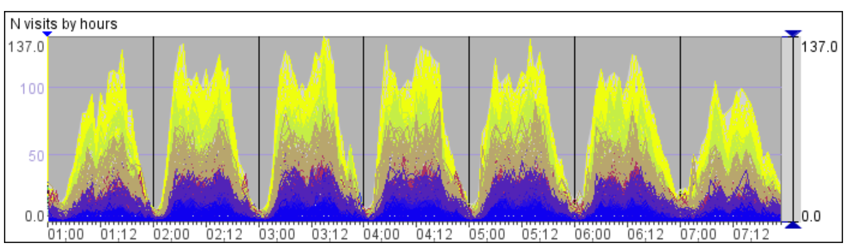
Clustering by k-means; k=7

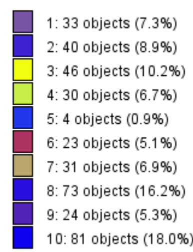


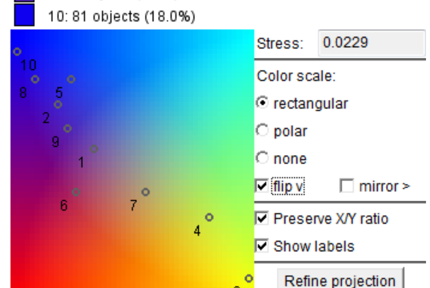




Clustering by k-means; k=10







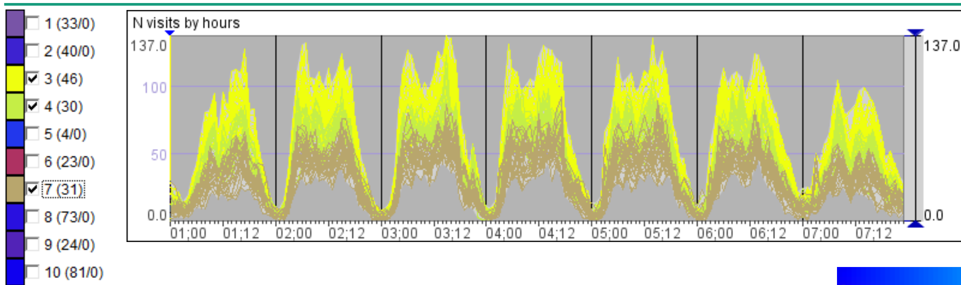
Interactive visual support to clustering

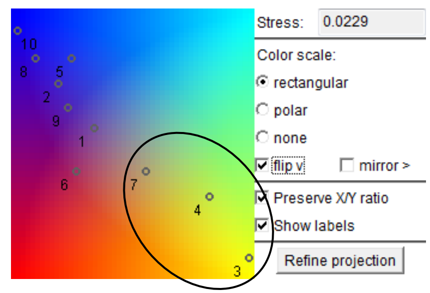
Trying various parameter setting; studying parameter impact

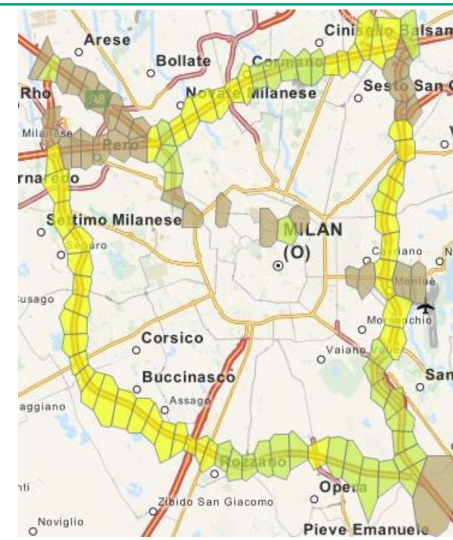
- Interactive interface to a clustering tool
- Immediate visualization of clustering results
- Selection of clusters for close inspection and comparison
- Visual displays of the components used for the clustering allow the analyst to assess the internal variation in the clusters
  - In our example: numeric time series
- Visual displays of other data components supports interpretation of the clusters
  - In our example: map
- Positions of cluster centres can be projected onto a coloured plane
  - Shows distances (amounts of difference) between the clusters
  - Cognitively beneficial colour assignment: close clusters receive similar colours

Interactive progressive clustering

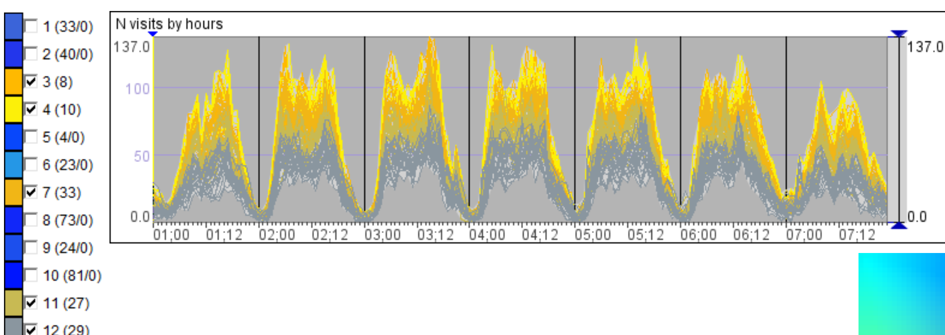
Division of selected clusters with high internal variation

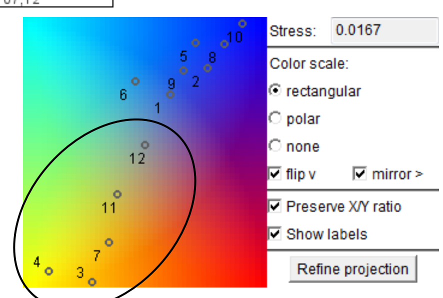


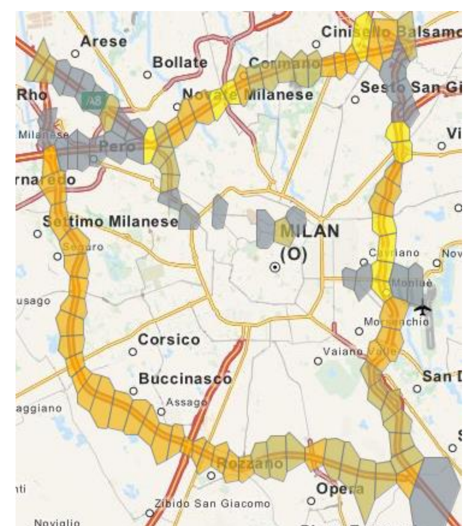




3 clusters are selected;
clustering is applied to the subset of objects from these 3 clusters



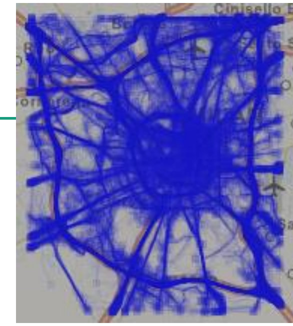




5 clusters have been obtained from the 3 selected clusters

Interactive progressive clustering

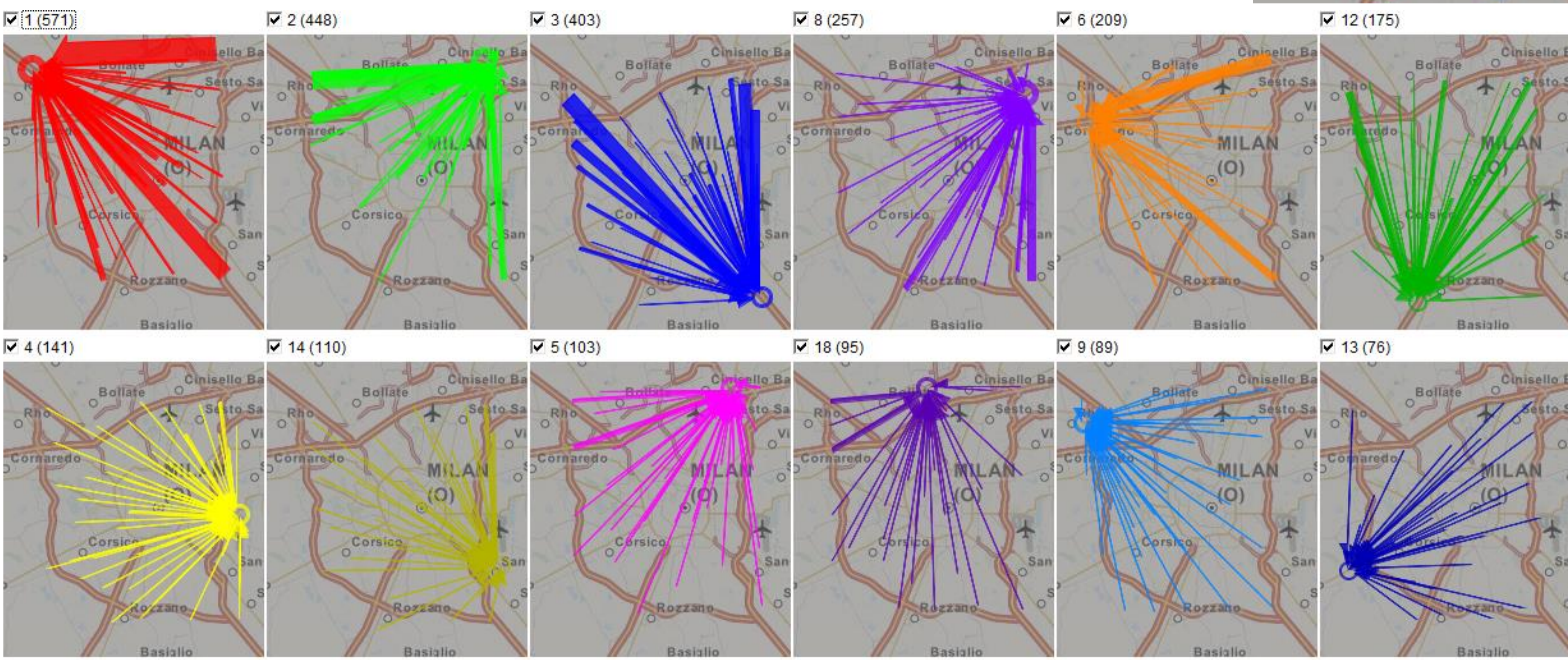
Applying different distance measures



Data: trajectories of cars in Milan

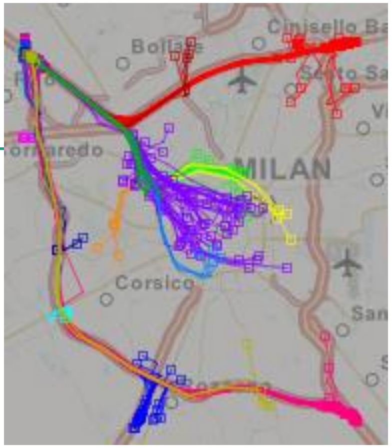
Step 1: clustering according to the spatial proximity of the end points

Question: what are the most frequent destinations of car trips?



Interactive progressive clustering

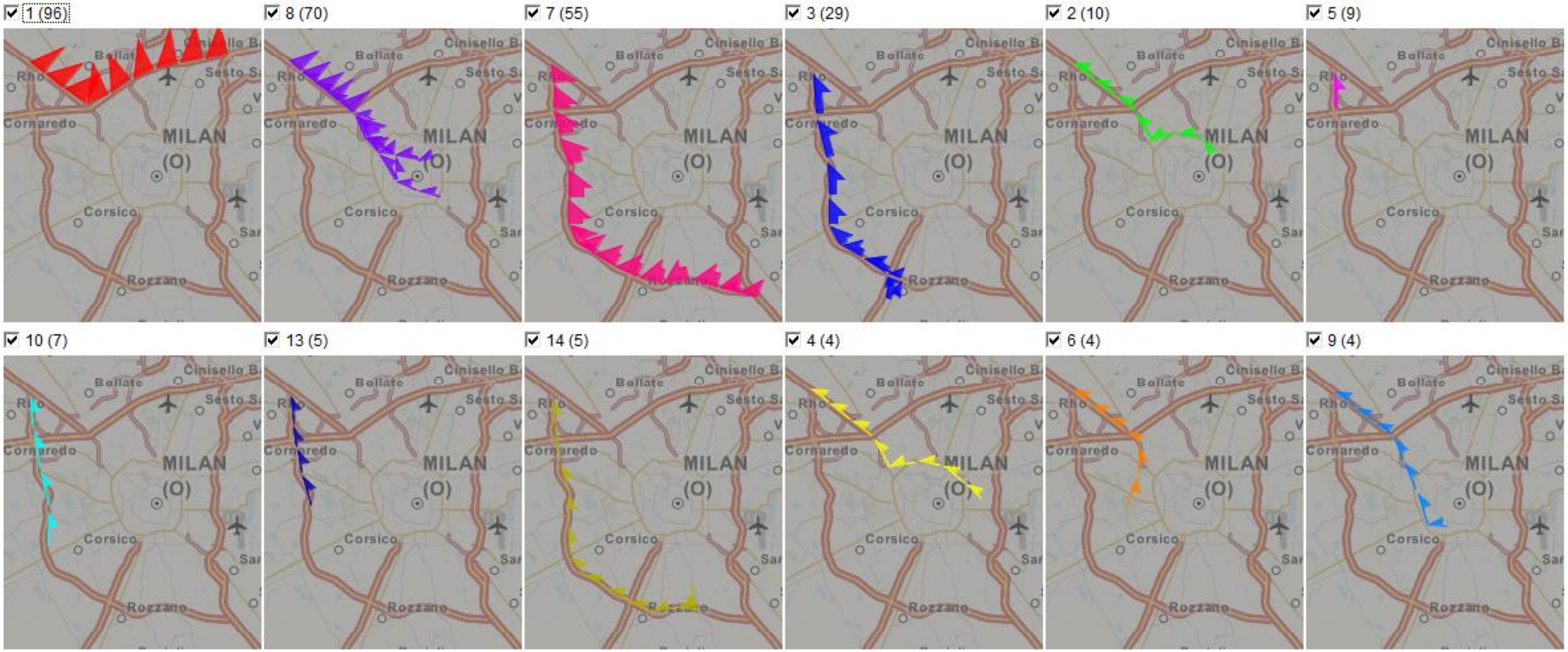
Applying different distance measures (2)



Data: one (or more) selected cluster(s) from the previous step

Step 2: clustering according to the similarity of the routes (shapes)

Question: what routes are usually taken to get to the selected destination?



Interactive progressive clustering

Purposes

- Controlled refinement of previously obtained clusters for
  - reducing internal variation
  - more detailed investigation of data subsets of interest
- Study of a set of complex objects with heterogeneous properties
  - application of diverse distance measures addressing different properties
    - a single distance measure would be hard to implement and results would be hard to interpret
  - incremental construction of multifaceted knowledge by progressively considering different properties

Exercises on applying clustering (k-means)

Data: spatial time series obtained by aggregating trajectories

- Visualise the time series of the average speeds of the traffic flows on a time graph. Apply k-means to the time series. Take k=7; observe on the map how the clustering separates the flows along the motorways from the flows in other parts of the city.
- Try k=8. Have the spatial patterns changed? Have the flows along the motorways been affected by the change of k?
- By selecting clusters one by one (through filtering), observe on the time graph which clusters have low internal variation (i.e., the lines are close to each other) and in which clusters the internal variation is higher. What can be said about the variation in the motorway-related clusters? What are the differences in the speed variation between these clusters?
- Try to refine the motorway-related clusters by progressive clustering: select only these clusters (let the number be m) and apply the k-means tool with setting k=m+2. Observe on the map and in the time graph how the clusters have been divided.
- On what motorways and in what directions are the average movement speeds the lowest?

Exercises on applying density-based clustering

Data: trajectories divided into trips by 10 minutes break

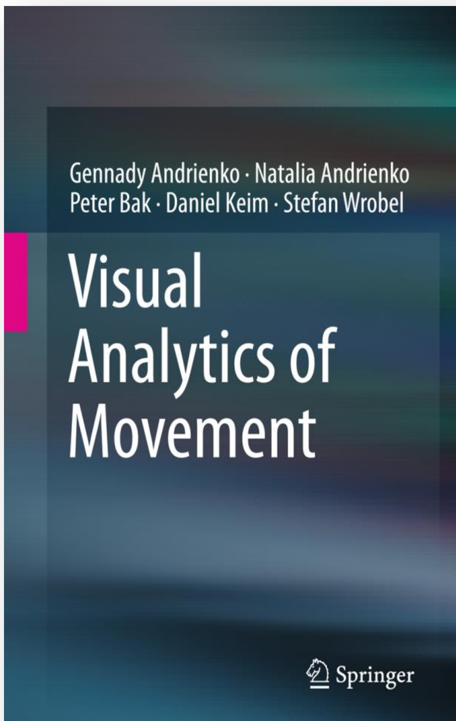
- Apply OPTICS clustering by trip destinations (i.e., end points of the trajectories) with R=250 m and N=10.
- Represent the resulting clusters in a summarized form. Try summarisation using all trajectory points and summarisation using only start and end points.
- What are the three most frequent trip destinations? Is there a cluster of trips ending in the centre? How many trajectories does it include?
- Progressive clustering: apply (in groups) clustering with the distance function “route similarity” to the four largest clusters obtained before (one cluster per group). Take R=750 m and N=5. Describe the most frequently taken routes to the respective destination areas.

Where to read more about visual analytics and about analysis of movement data

Springer, June 2013

ISBN 978-3-642-37582-8

397 p. 200 illus., 178 in colour



Ch.1. **Introduction**

Ch.2. **Conceptual framework**

Ch.3. **Transformations of movement data**

Ch.4. **Visual analytics infrastructure**

Ch.5. **Visual analytics focusing on movers**

Ch.6. **Visual analytics focusing on spatial events**

Ch.7. **Visual analytics focusing on space**

Ch.8. **Visual analytics focusing on time**

Ch.9. **Discussion and outlook**

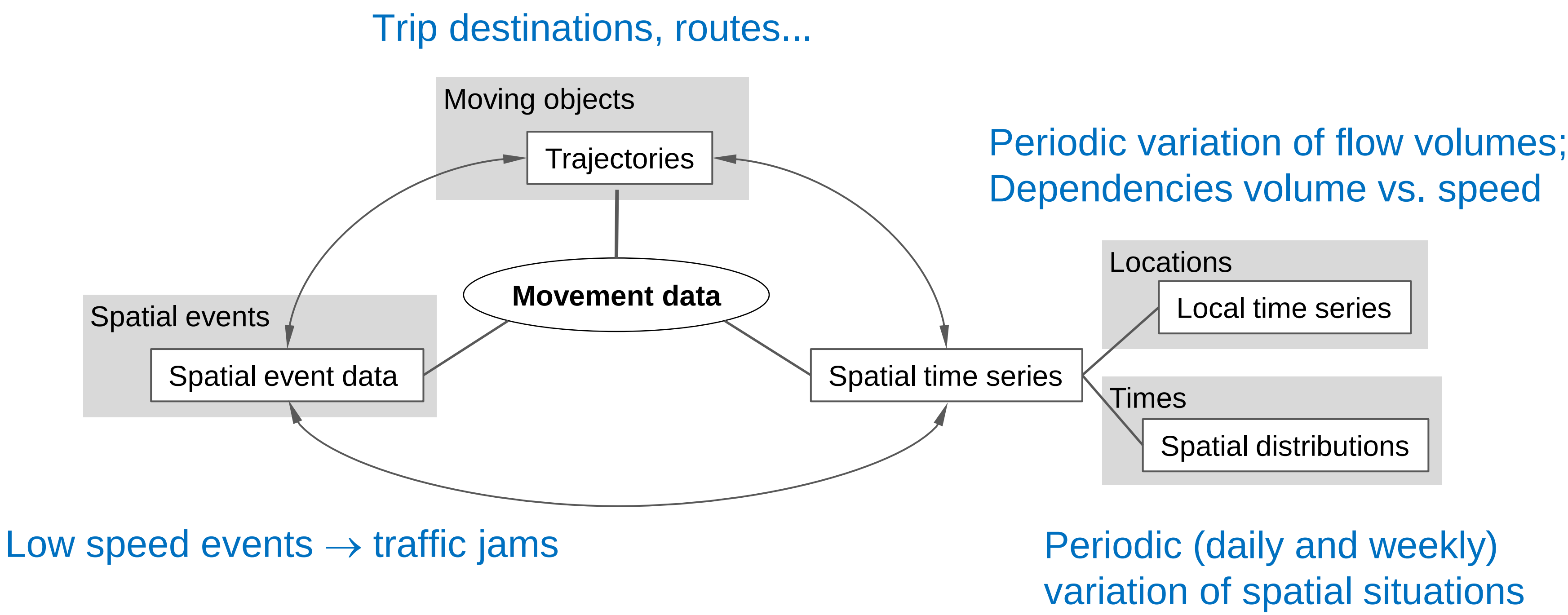
CITY UNIVERSITY LONDON

Fraunhofer IAIS

Multi-perspective analysis of movement

Movement data can be viewed from multiple complementary perspectives.

Multi-perspective analysis allows deeper and more comprehensive understanding of the studied phenomenon, e.g., city traffic.



CITY UNIVERSITY LONDON
Fraunhofer IAIS

Concluding summary

- Visual analytics tools and techniques support human analysts in performing data analysis: Data $\rightarrow$ Information $\rightarrow$ Knowledge $\rightarrow$ Explanation
- VA tools and techniques enable analysts to exploit effectively their vision-based cognitive capabilities
  - Abstraction, grasping general, characteristic features, pattern detection and interpretation, ...
- VA tools and techniques divide the analytical labour between humans and computers
  - Use computer processing where human judgement is not needed
  - Use computers to prepare data to human analysis
  - Use computers to present data to analysts in the most suitable form

Multi-perspective Analysis of Movement Data with Visual Analytics

Gennady Andrienko

Natalia Andrienko

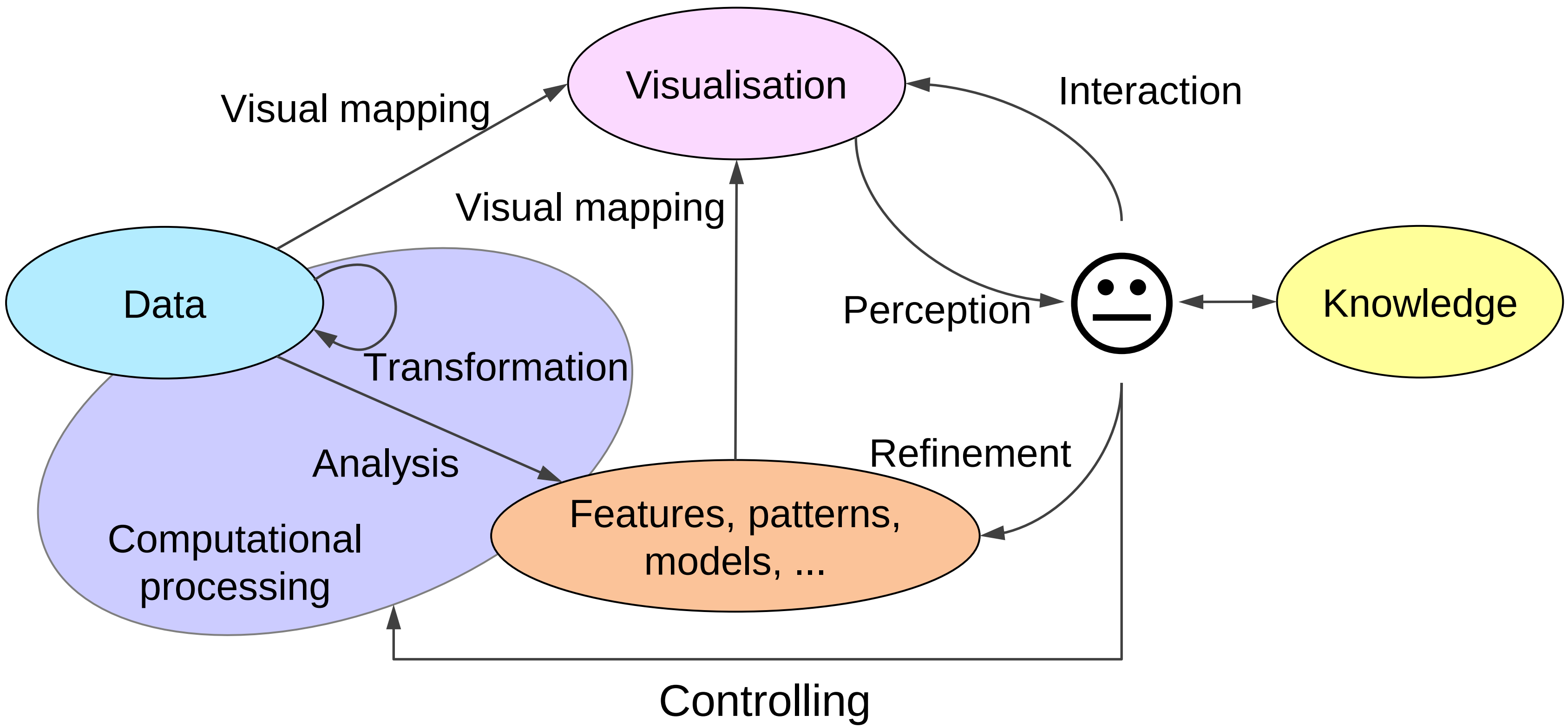
http://geoanalytics.net

CITY UNIVERSITY LONDON
EST 1894

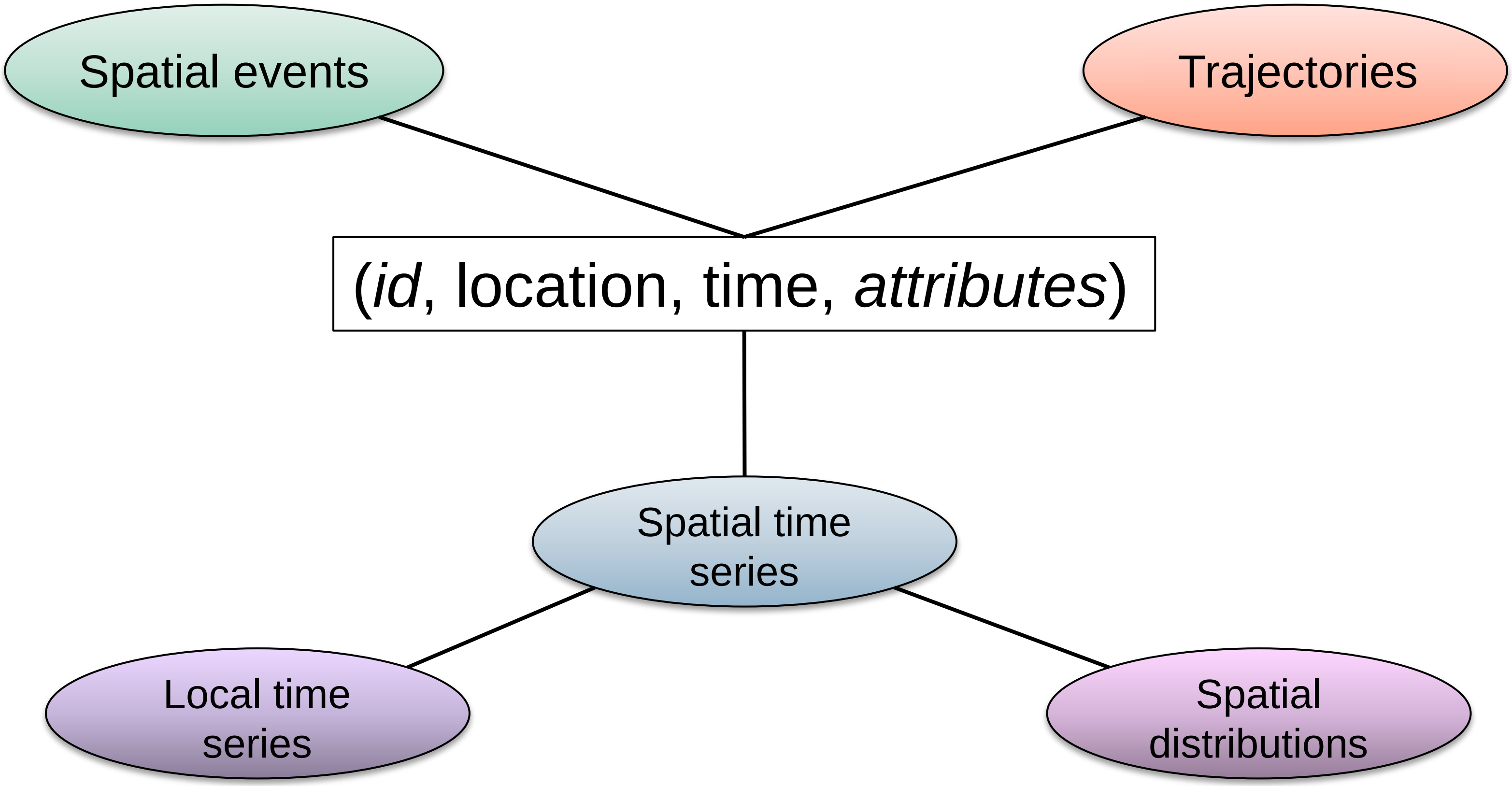
Fraunhofer
IAIS

Visual Analytics

Enabling synergetic work of humans and computers

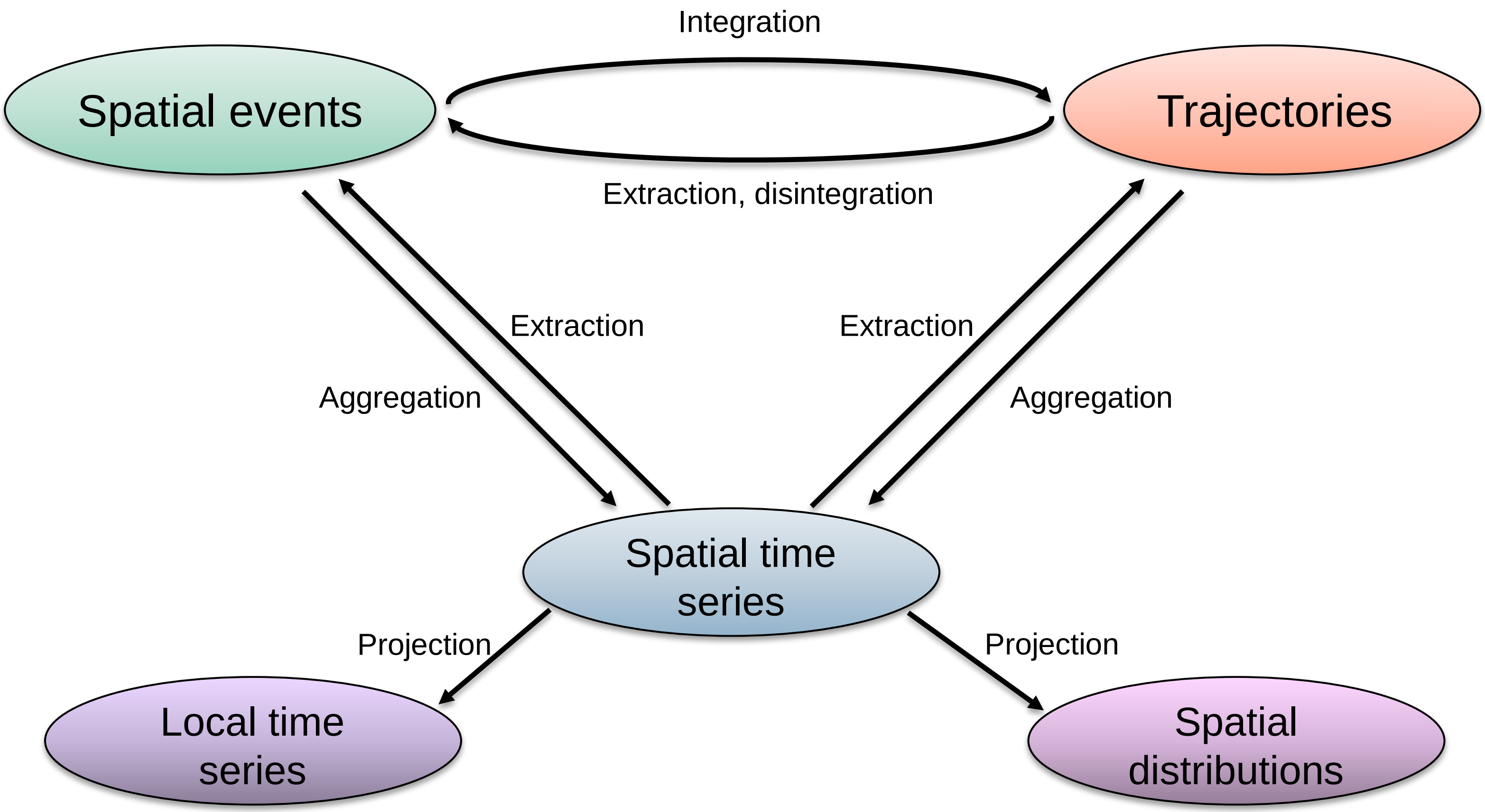


Types of spatio-temporal data



CITY UNIVERSITY LONDON
Fraunhofer IAIS

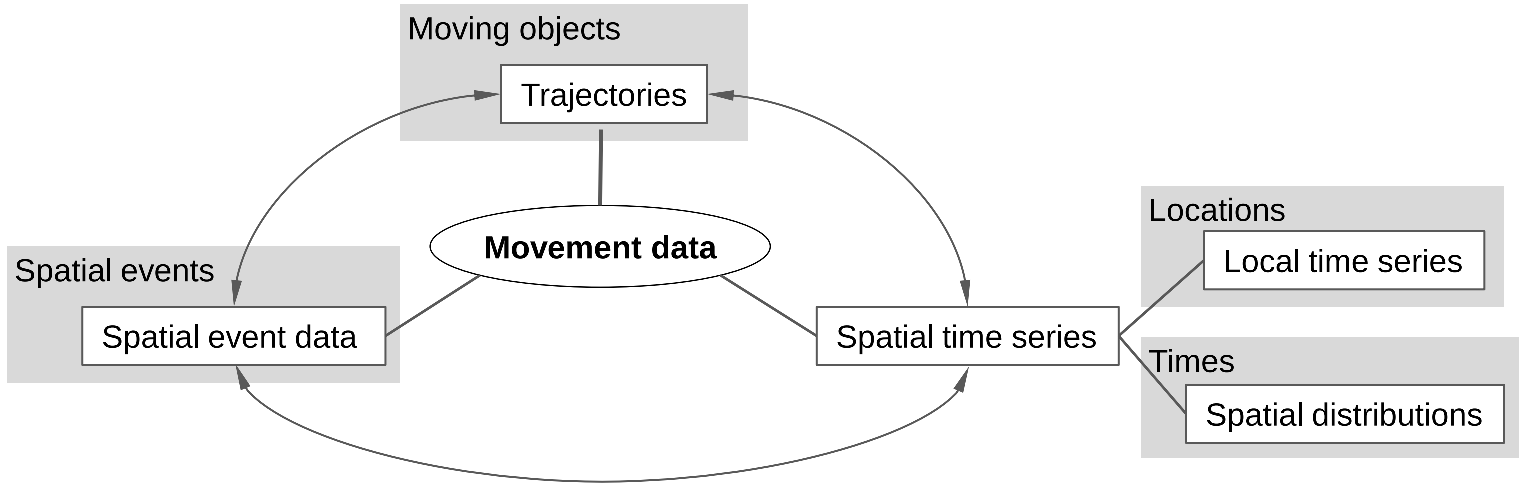
Transformations of spatio-temporal data structures



CITY UNIVERSITY LONDON

Fraunhofer IAIS

Transformations enable multi-perspective analysis of movement data



Running example dataset: trajectories of cars in Milan

OCTO
The reliable way

GPS-tracks of 17,241 cars in Milan, Italy

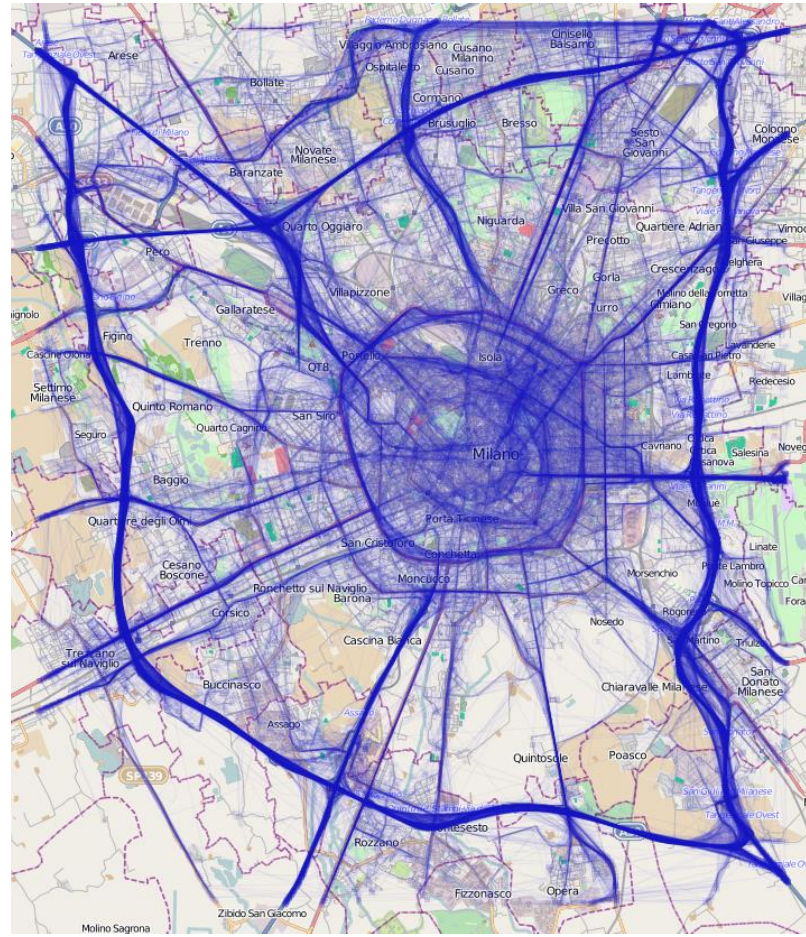
Time period: from Sunday, the 1st of April, to Saturday, the 7th of April, 2007

Received from Octo Telematics
www.octotelematics.com
special thanks to T.Martino

Data structure:

- Anonymized car identifier
- Date and time
- Geographic coordinates
- Speed

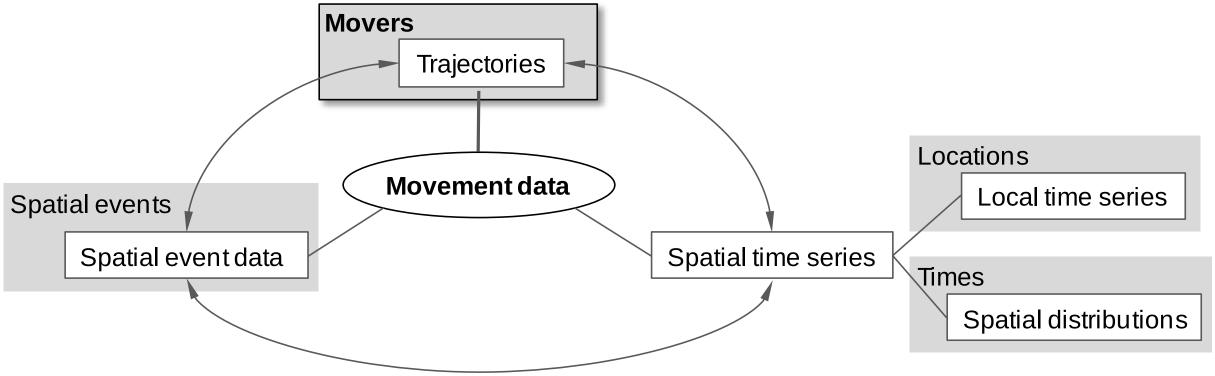
The trajectories from one day are drawn on a map with 5% opacity



CITY UNIVERSITY LONDON

Fraunhofer IAIS

Perspective 1:
Movement data in the form of trajectories



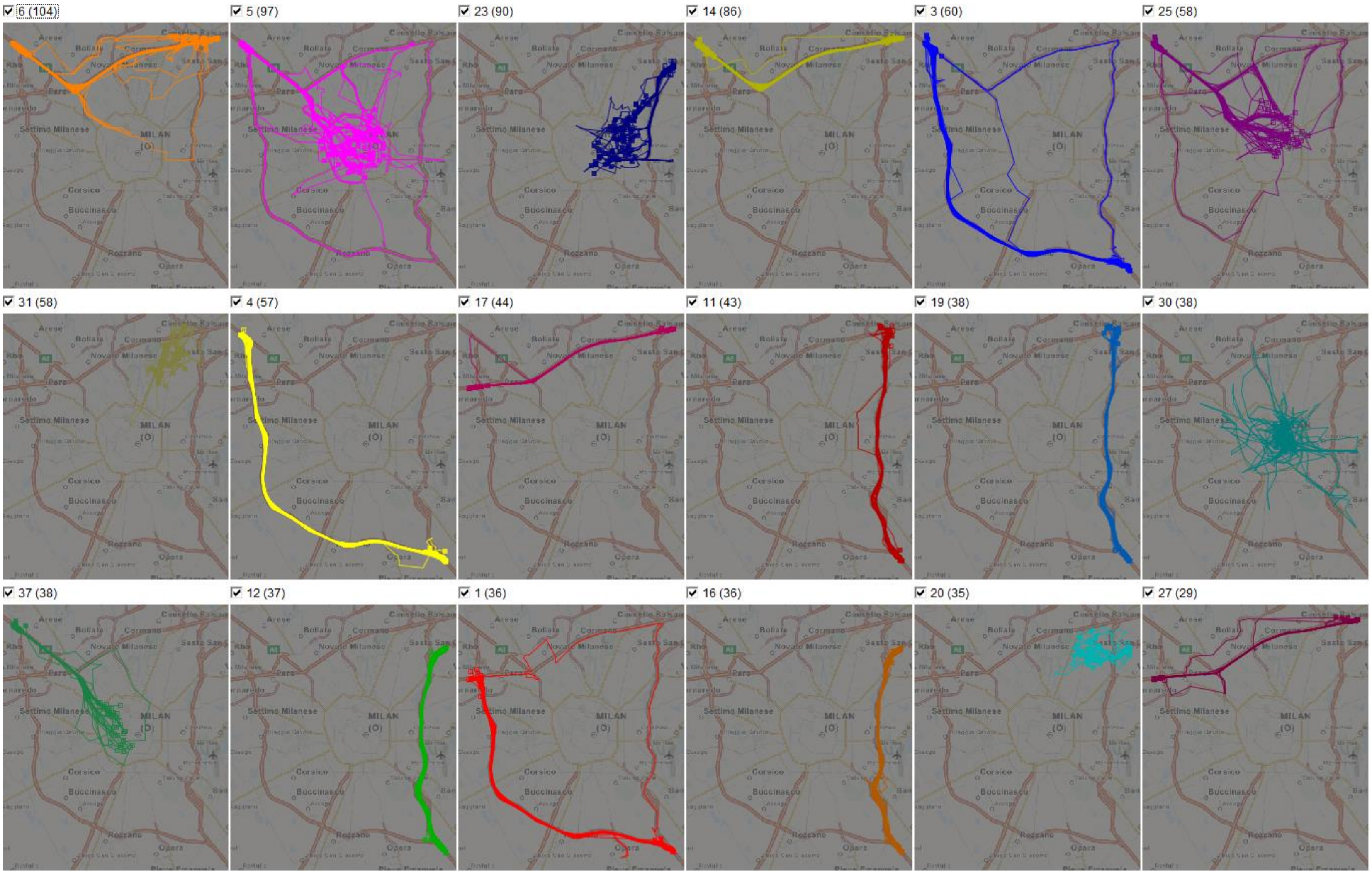
Density-based clustering of trajectories: What distance measure to use?

- Trajectories are time series of spatial positions and other movement attributes
- Trajectories are complex objects with heterogeneous properties: positions in space and in time, shape, dynamics of speed, ...
- A single distance measure accounting for all properties would be hard to implement and results would be hard to interpret
- It is more feasible to create a library of simple distance measures (a.k.a. distance functions) addressing different properties. For example,
  - spatial distance between origins and/or between destinations,
  - average spatial distance between corresponding points along the routes,
  - average spatial distance between points reached at the same times, ...
- Different aspects of trajectories are studied using different distance functions.

DB clusters of trajectories (example 1)

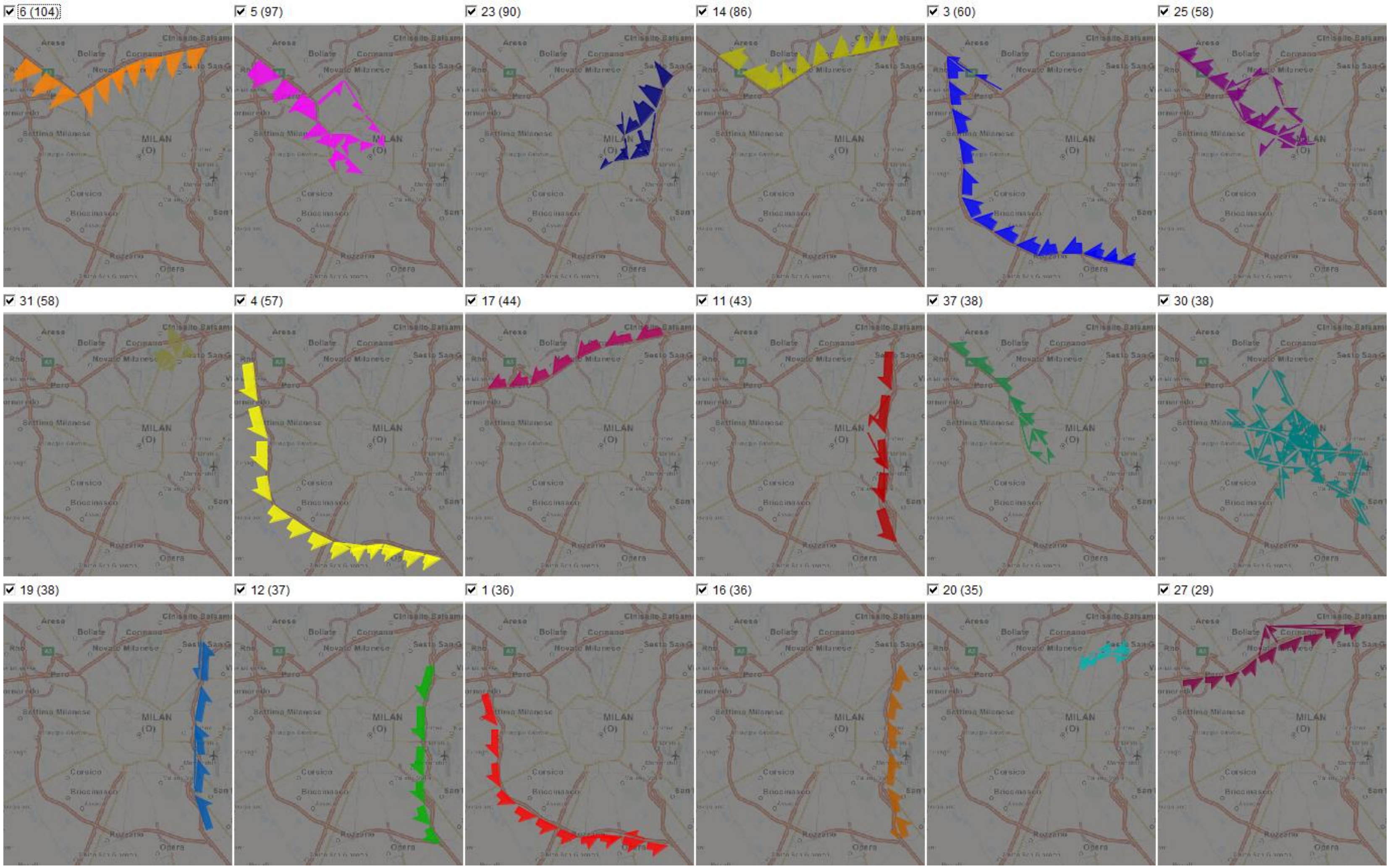
Distance function: the average spatial distance between the origins and between the destinations; R=750m, N=5

Only 18 largest clusters are shown.



Summarised representation of clusters of trajectories

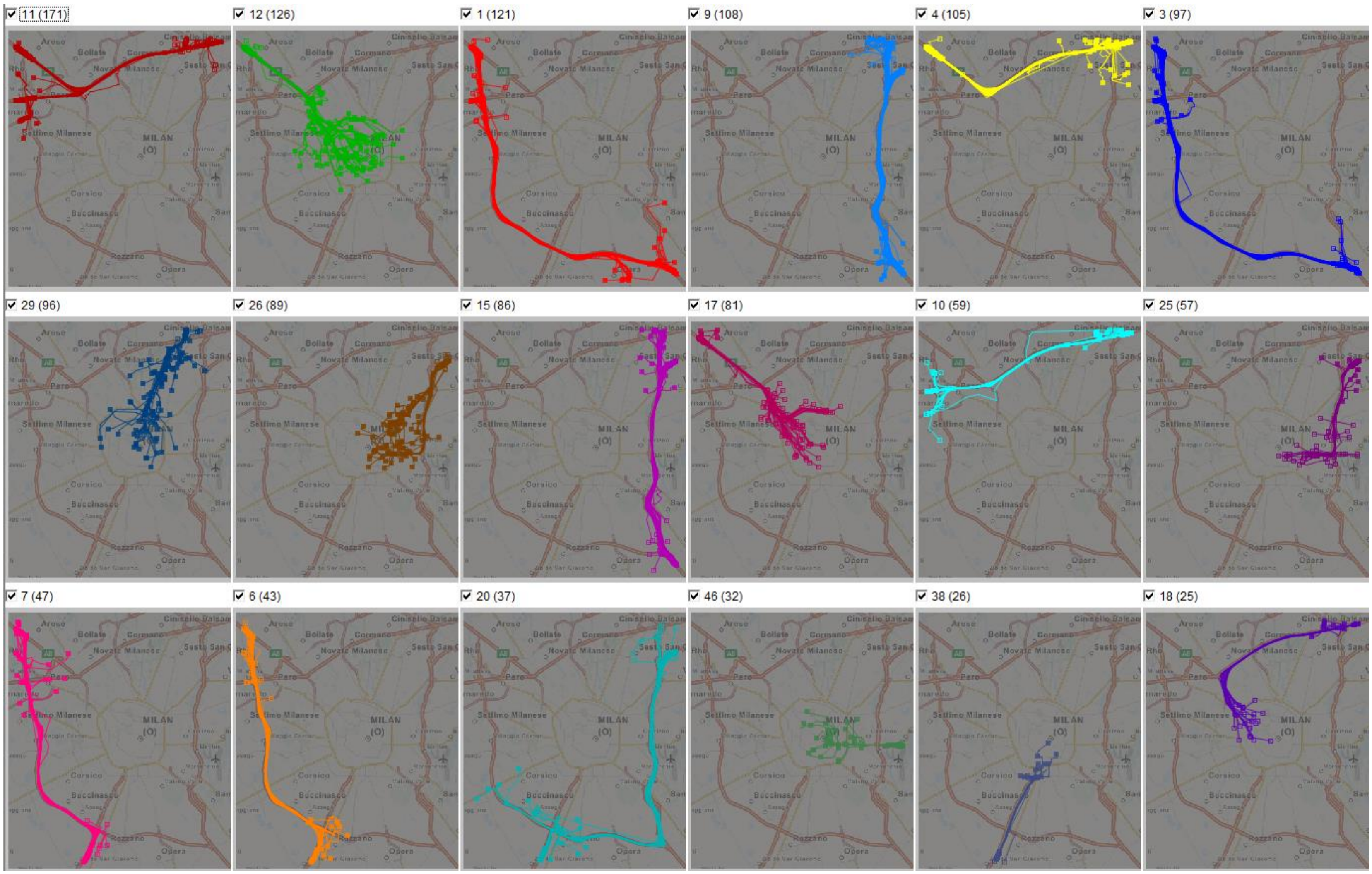
Minor flows are omitted for a clearer view.



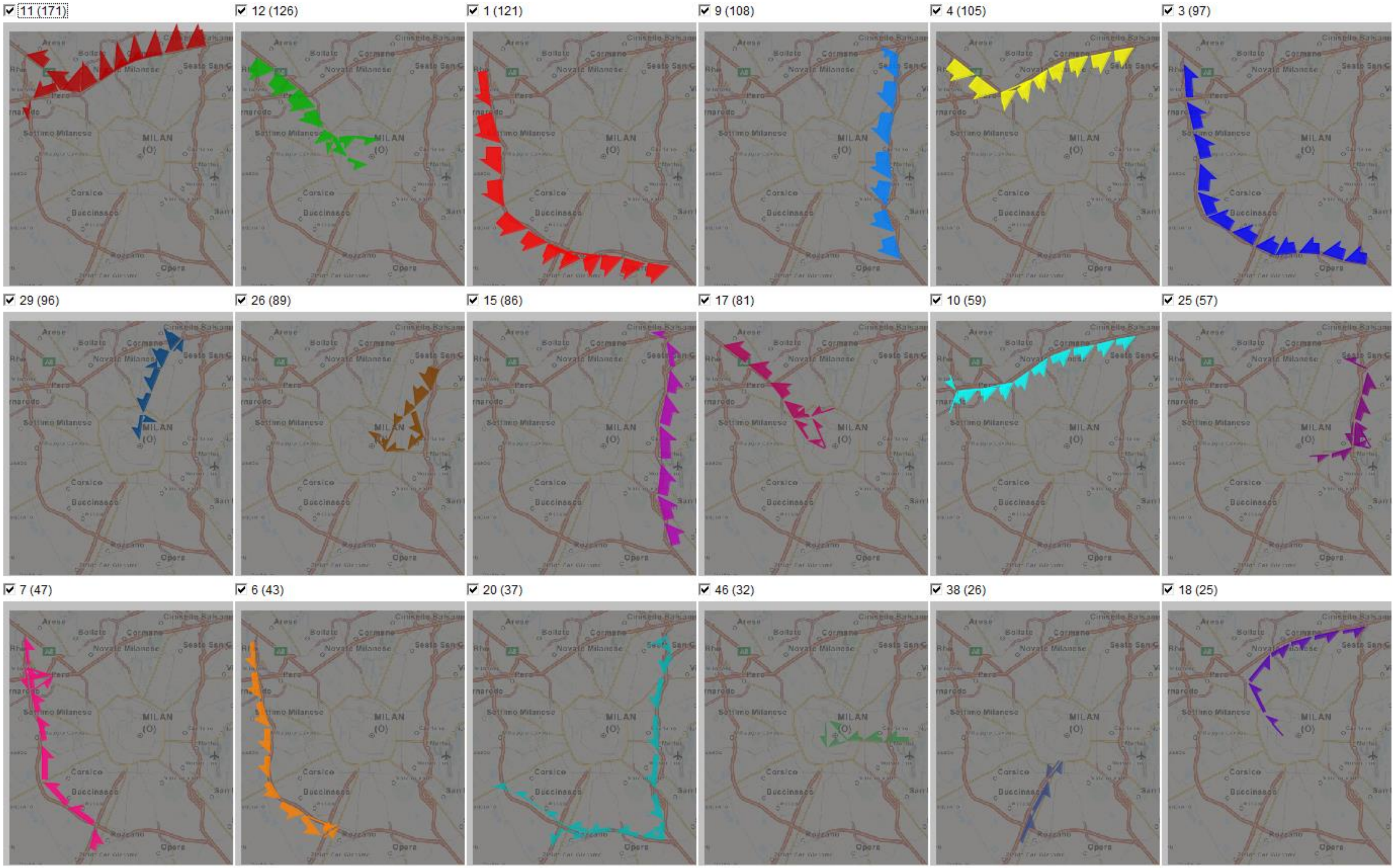
DB clusters of trajectories (example 2)

Distance function: "route similarity", i.e., the average spatial distance between the corresponding points along the route; R=750m, N=5

Only 18 largest clusters are shown.

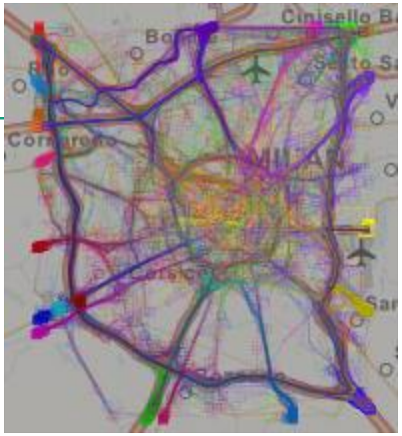


The same clusters represented in a summarised form



Interactive progressive clustering

Applying different distance measures at different steps

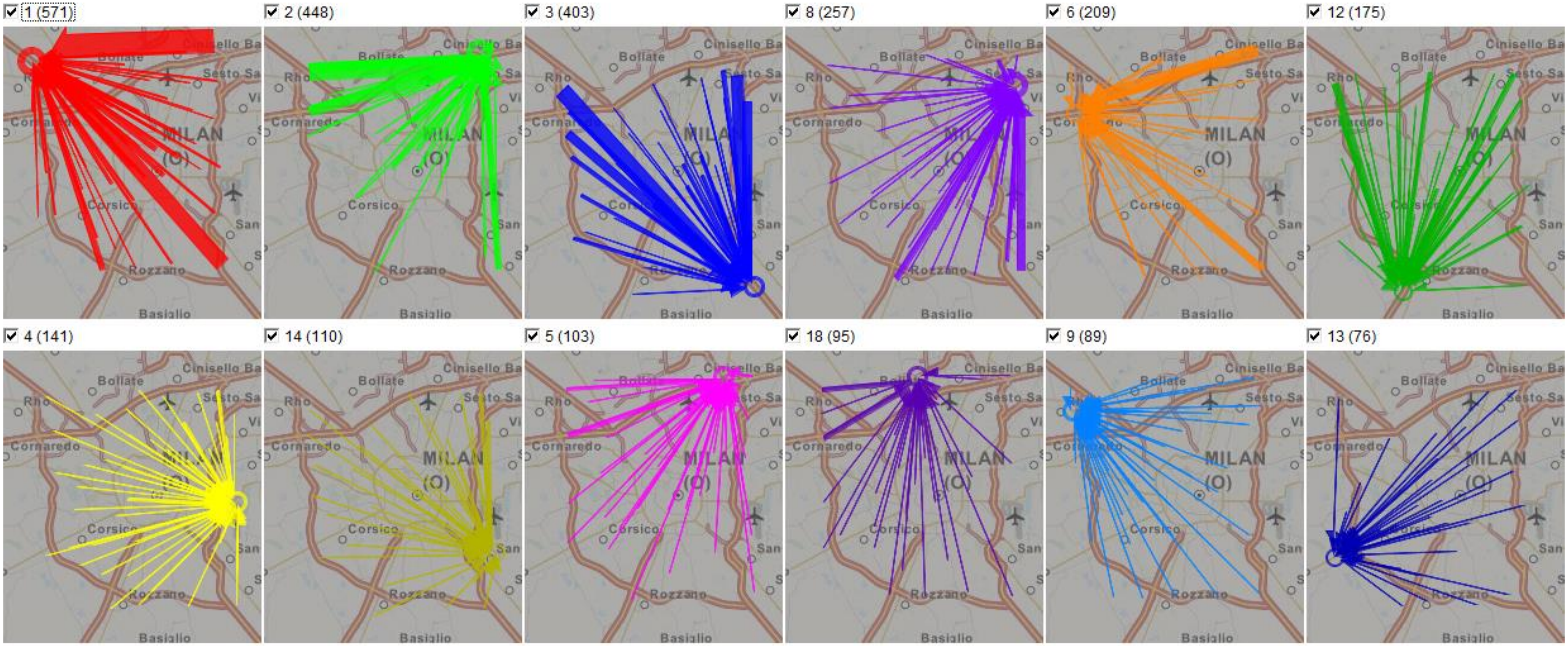


Data: trajectories of cars in Milan

Step 1: clustering according to the spatial proximity of the end points

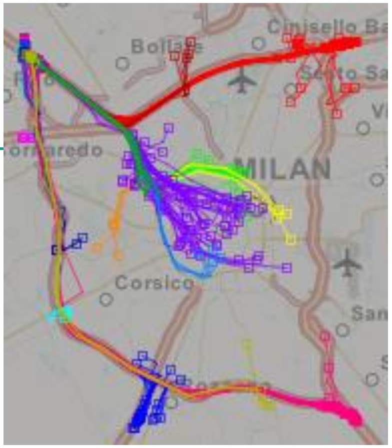
Distance function: “common ends”

<u>Question</u>: what are the most frequent destinations of car trips?



Interactive progressive clustering

Applying different distance measures (2)

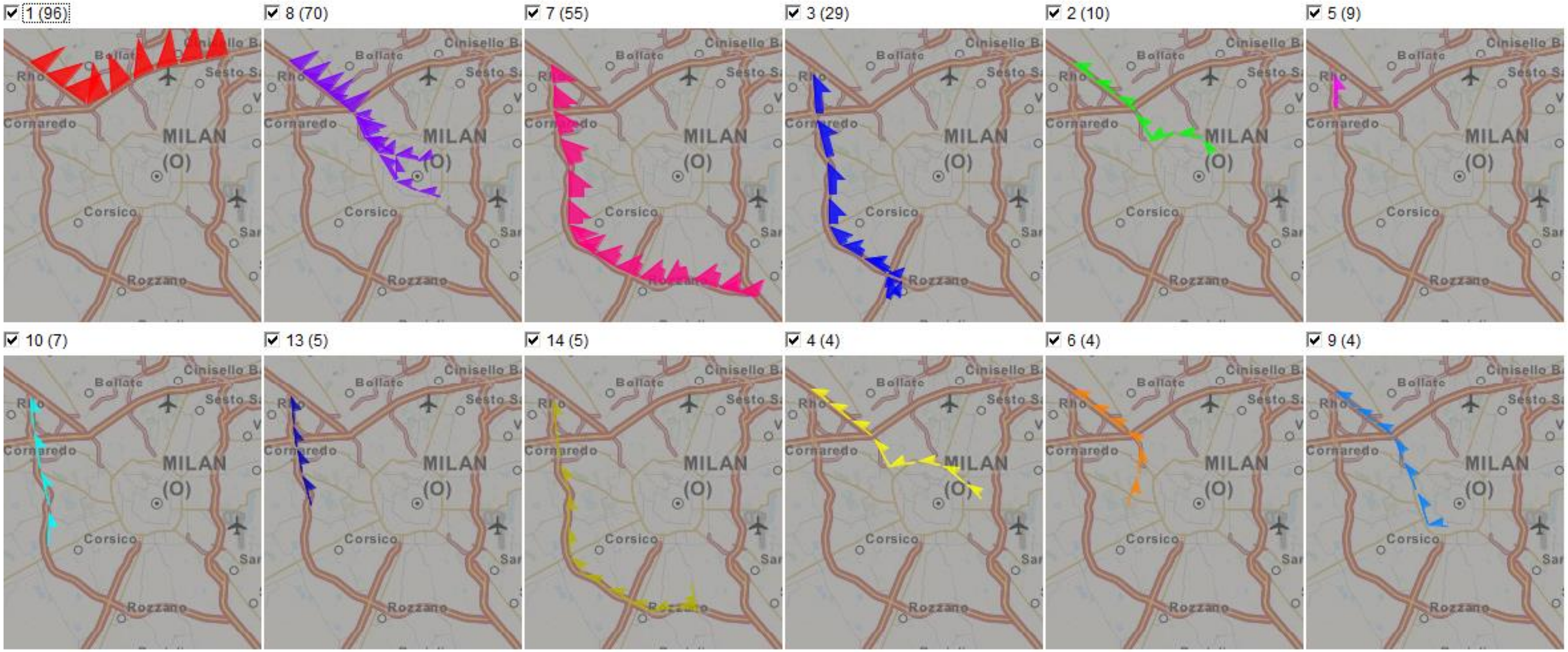


Data: one (or more) selected cluster(s) from the previous step

Step 2: clustering according to the similarity of the routes (shapes)

Distance function: “route similarity”

<u>Question</u>: what routes are usually taken to get to the selected destination?



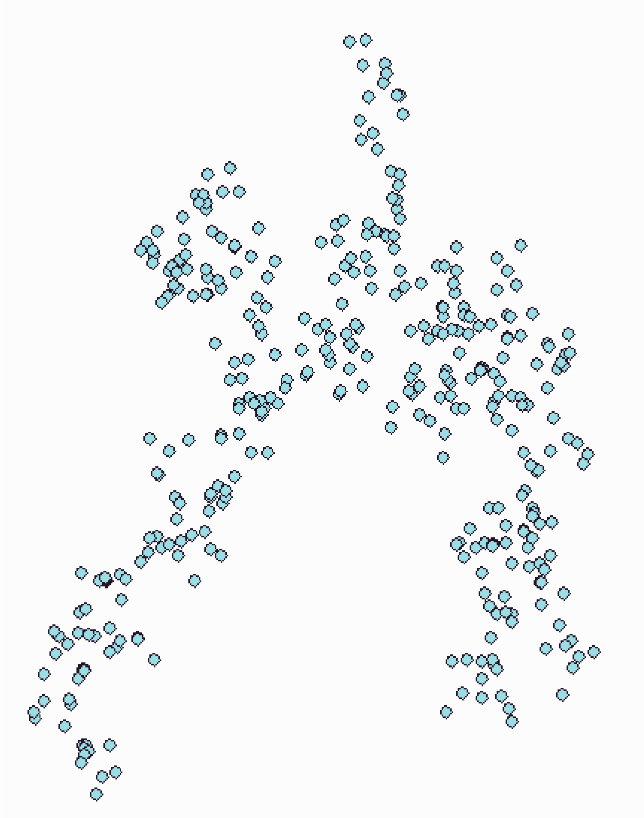
Clustering of very large sets of trajectories

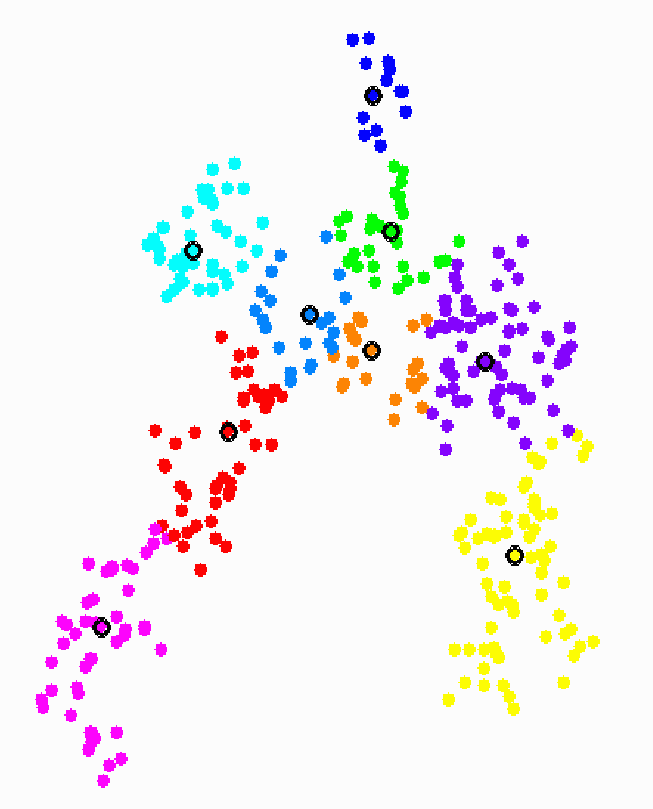
- Problem: clustering of complex objects (such as trajectories) involving non-trivial distance functions (such as "route similarity") can only be done in RAM, i.e. for a relatively small dataset
- Our approach:
  1. Take a subset (sample) of the objects suitable for processing in RAM.
  2. Discover clusters 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

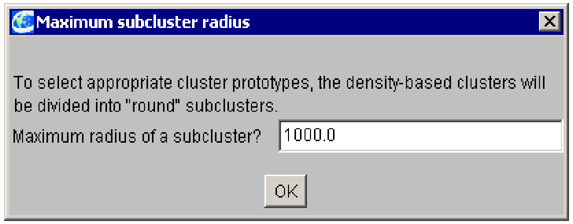
Dividing a cluster into round sub-clusters: an illustration using points

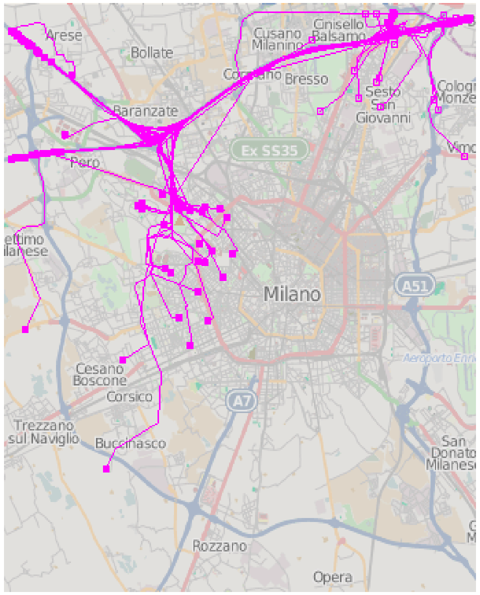




This can be done by a variant of the K-medoids clustering algorithm where the desired maximum radius of a subcluster is a parameter.

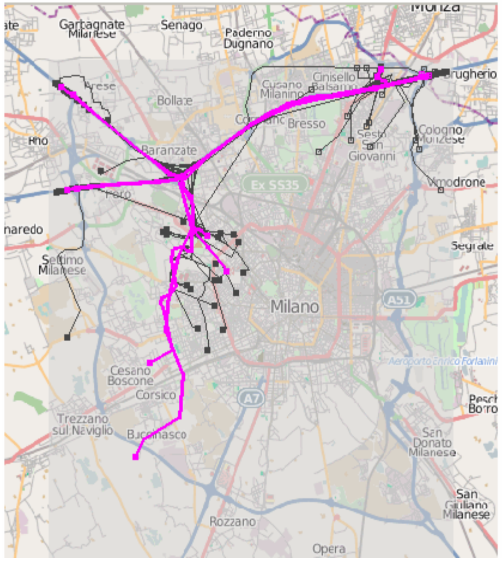
Division of a cluster of trajectories into “round” subclusters

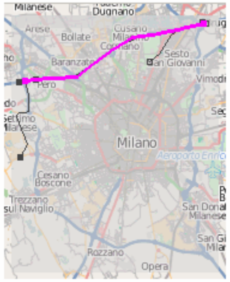




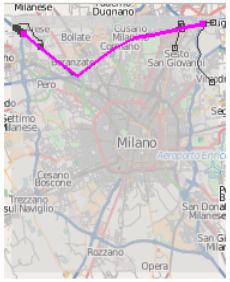
25.09.2009 11:05:24 - Cluster 7

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

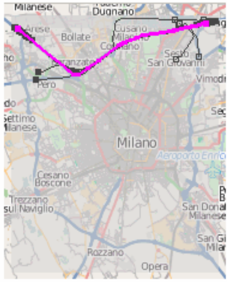




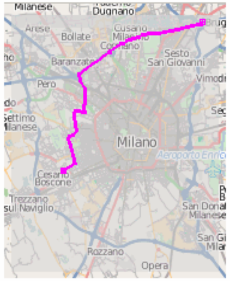
89133



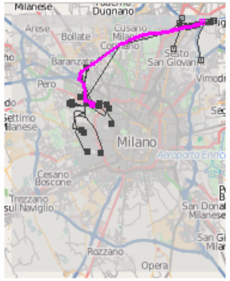
96013



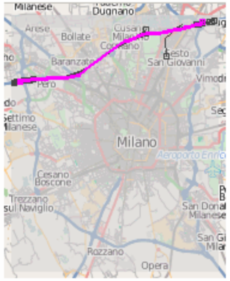
6548



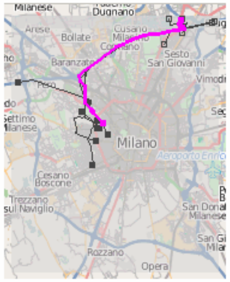
43285



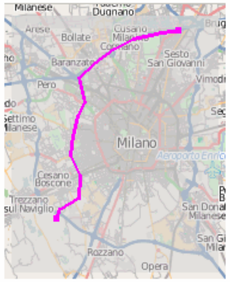
34239



32809



141138

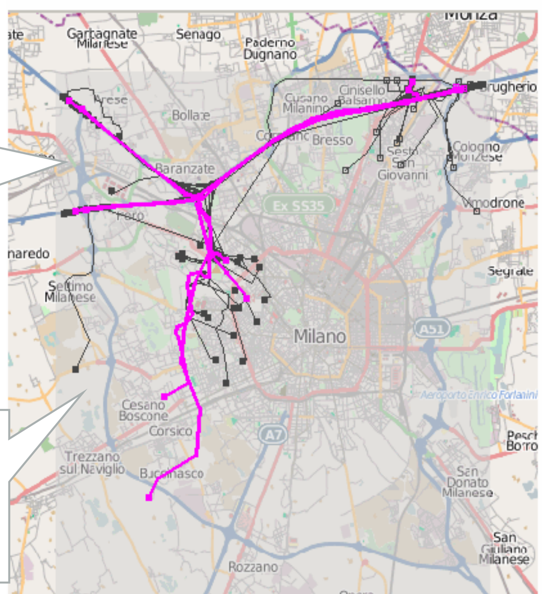


109120

To obtain meaningful results, the analyst may 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



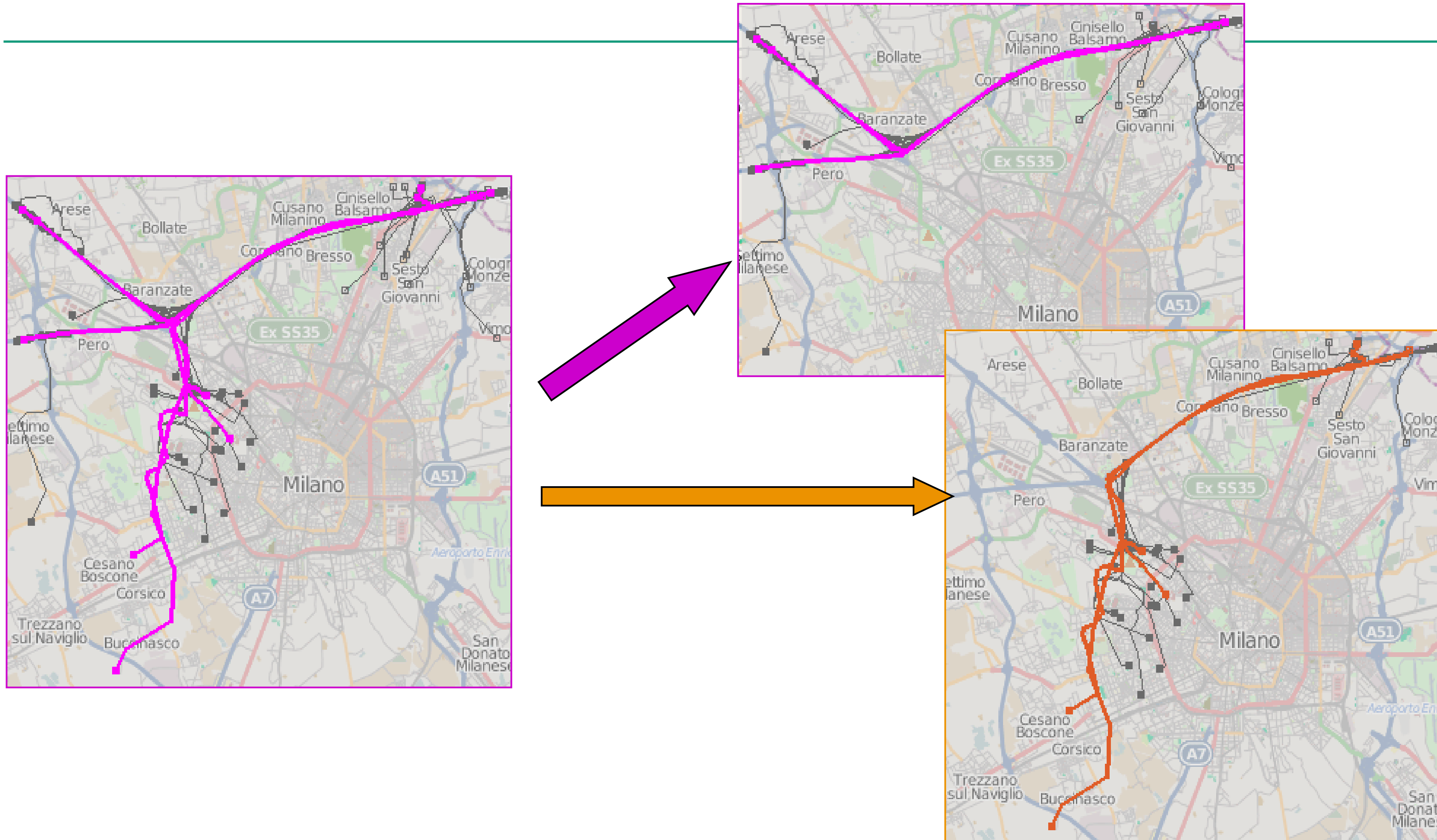
Should I keep the three branches in one cluster?

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?

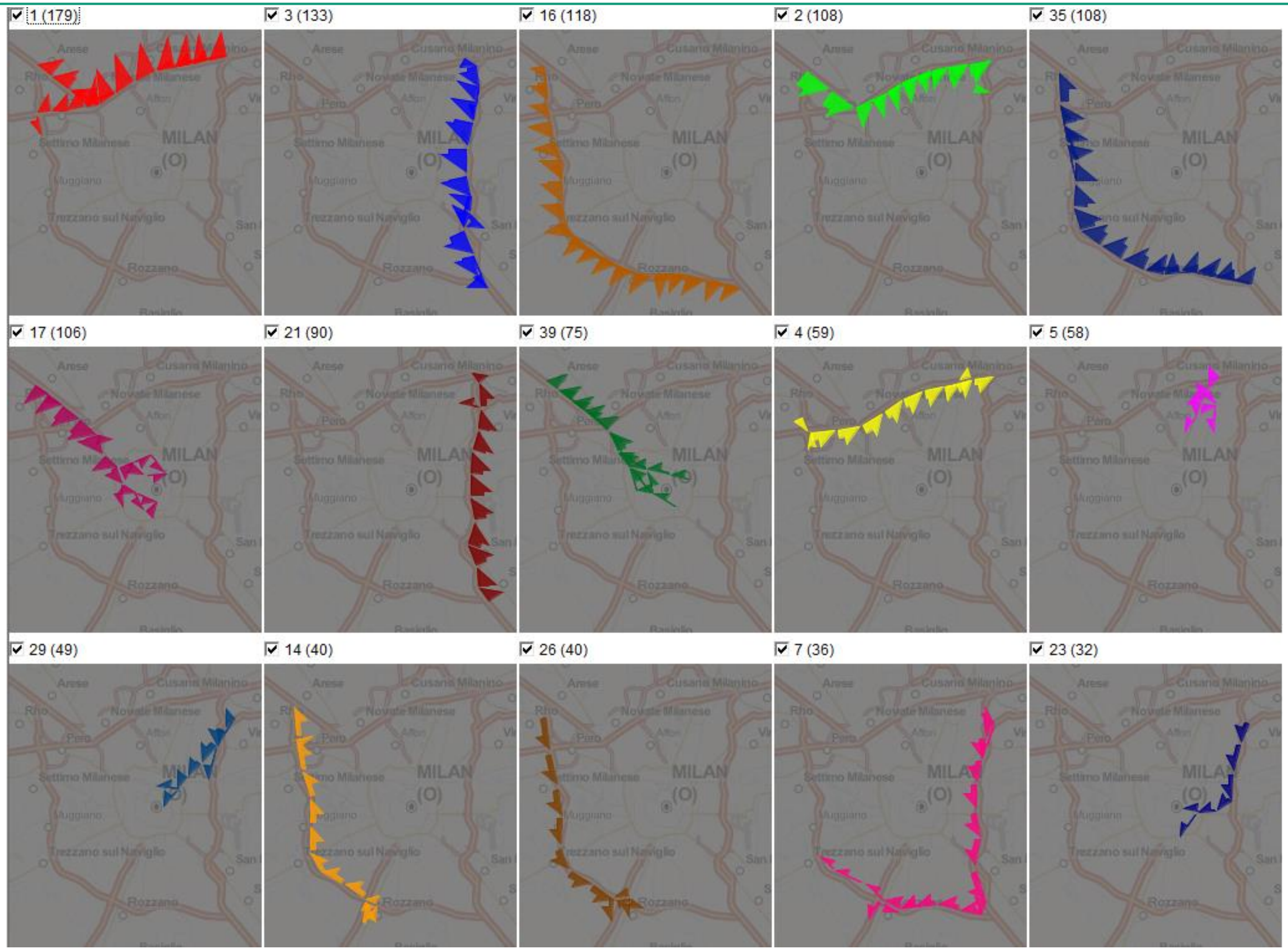
Is it good to have this prototype? This is not a core trajectory of the cluster.

Example of interactive editing



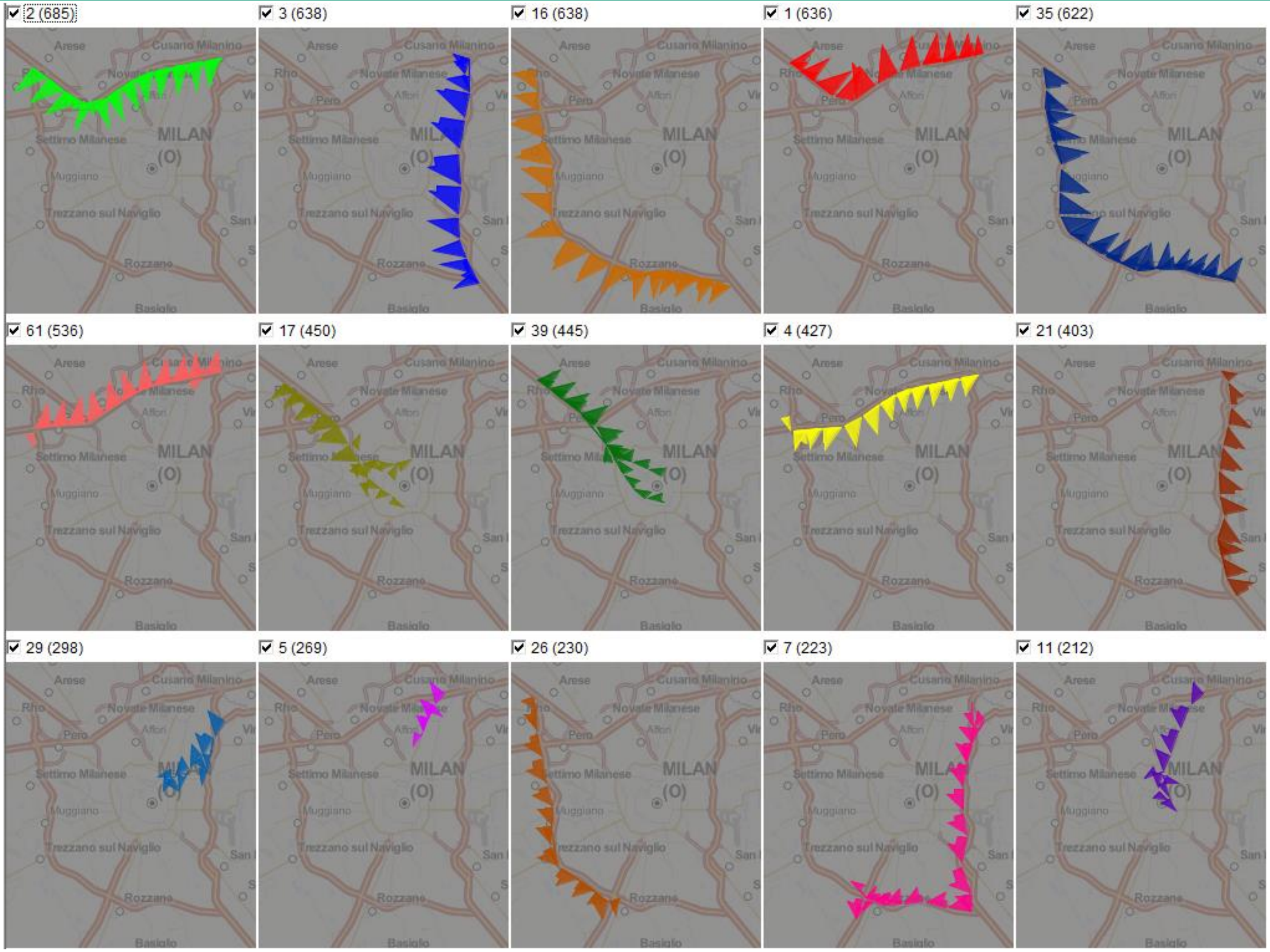
What are the most frequent routes on Wednesday?

Result of clustering of single-day trajectories by route similarity



How frequent are these routes during the whole week?

Result of building a classifier and applying it to the whole set of trajectories



Further analysis of the trajectories

- The analysis is continued by loading a subset of the unclassified trajectories ("noise") to RAM, applying clustering to it, building a new classifier, and applying the classifier to the whole set of unclassified trajectories.
- Empirical experience:
  - With each new iteration step, the number and the sizes of discovered clusters substantially decrease in comparison to the previous step.
    - After 4-5 steps of the procedure, only very small clusters can be discovered.
  - The analyst's effort needed for editing of the classifier also decreases.
    - The editing effort is high for big clusters with high internal variation, which mostly appear in the first step; the following clusters are smaller and "cleaner".
- Unfortunately, no formal criterion for terminating the procedure.

Where to read more

G.Andrienko, N.Andrienko, S.Rinzivillo, M.Nanni, D.Pedreschi, F.Giannotti

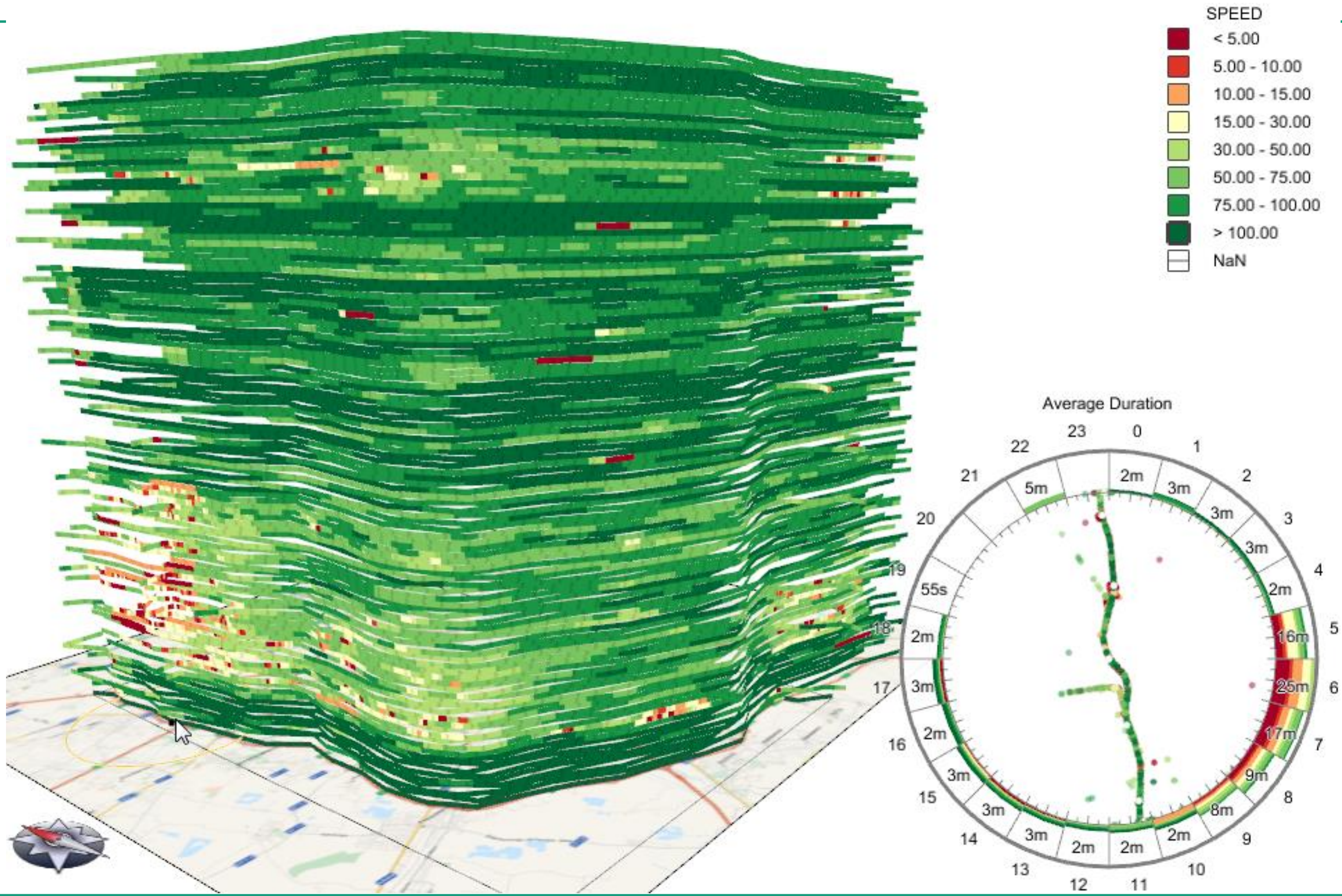
Interactive Visual Clustering of Large Collections of Trajectories

IEEE Visual Analytics Science and Technology (VAST 2009)

Proceedings, IEEE Computer Society Press, 2009, pp.3-10

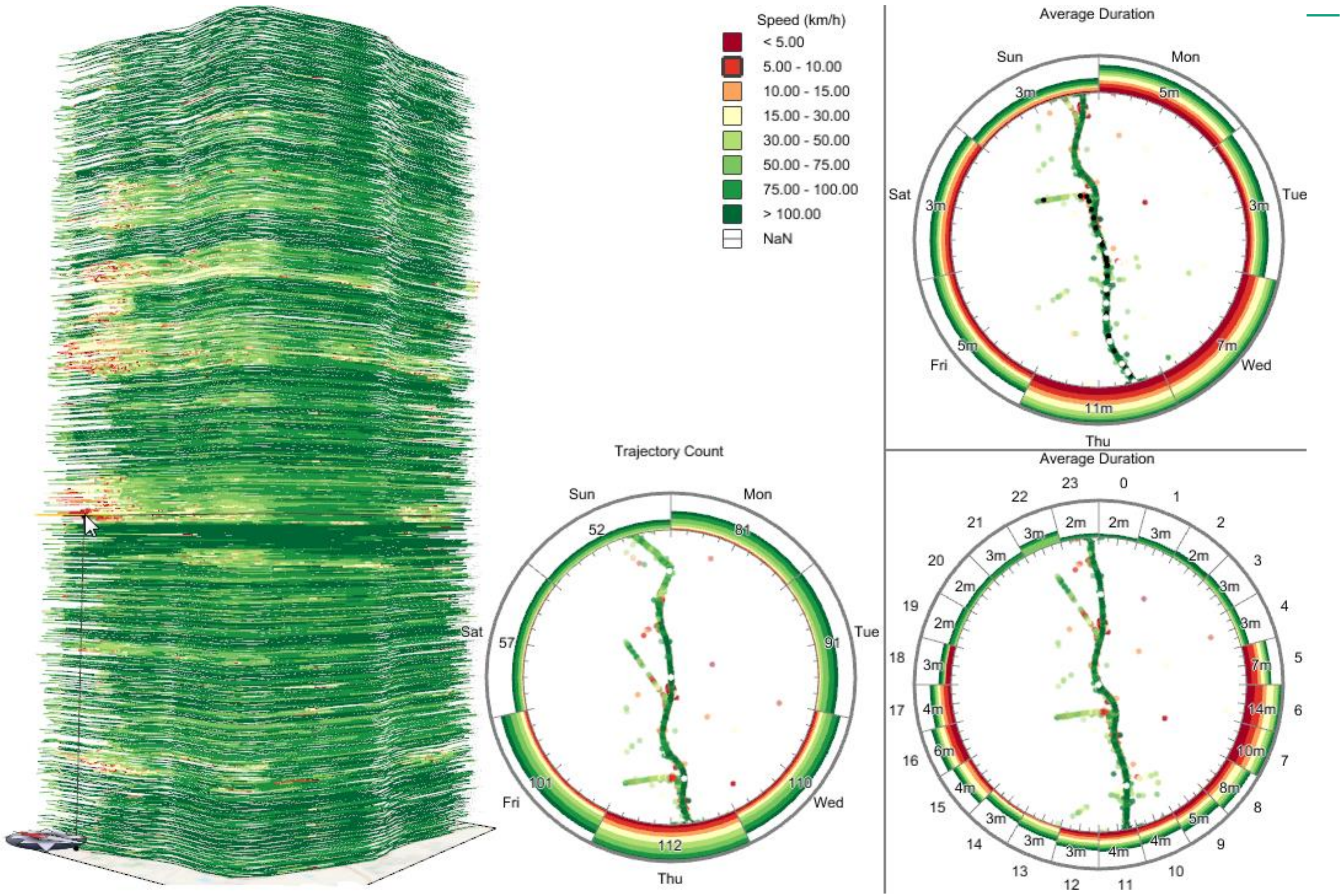
Analysis of movement attributes

Investigate speed variation along a selected route: single day

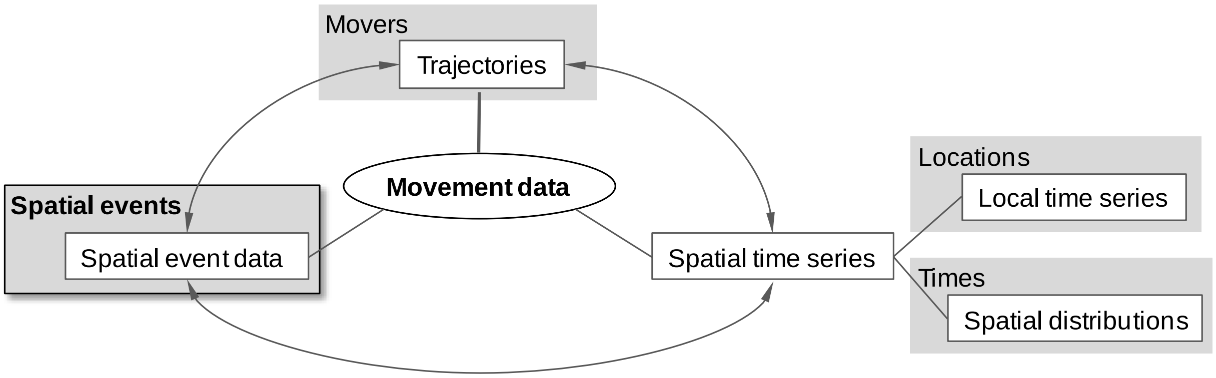


Analysis of movement attributes

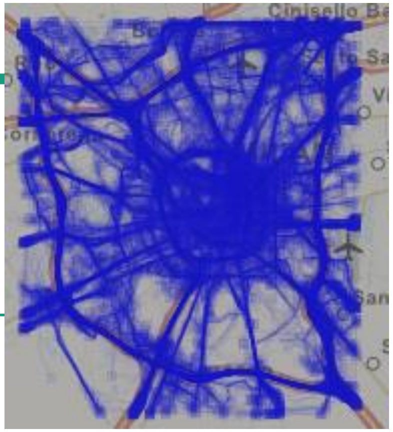
Investigate speed variation along a selected route: whole week



Perspective 2:
Movement data in the form of spatial events



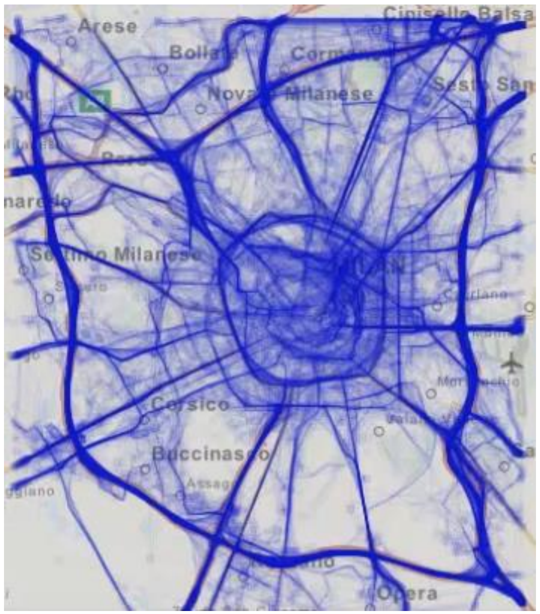
Example of analysis focusing on movement events

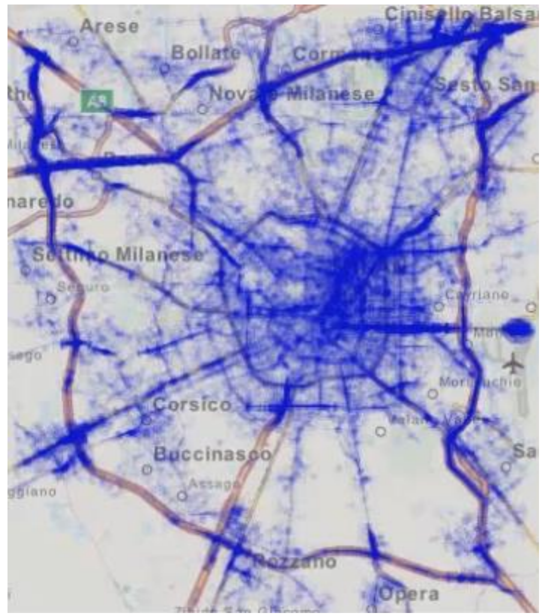


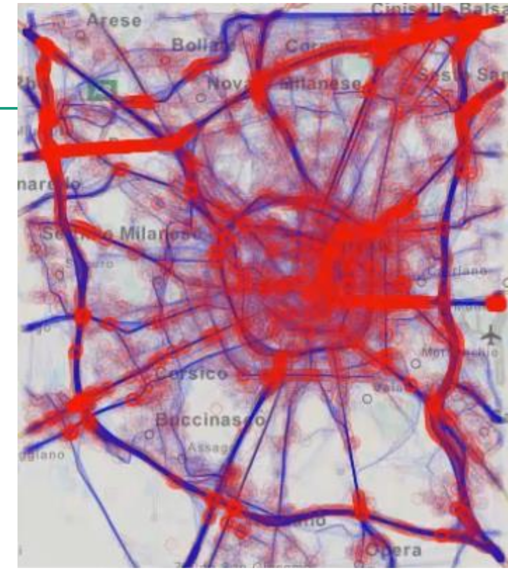
- Data: trajectories of cars in Milan
- Task: find places of traffic congestions and determine their characteristics (times of the congestions, durations, numbers of cars involved, ...)
- Traffic congestion ≈ dense spatio-temporal cluster of low speed movement events
  - Movement direction must be taken into account
- Places of interest: areas where at least one traffic congestion occurred ≈ areas containing the clusters
- Characteristics of places: time series of event counts, vehicle counts, ...
- Data transformations:

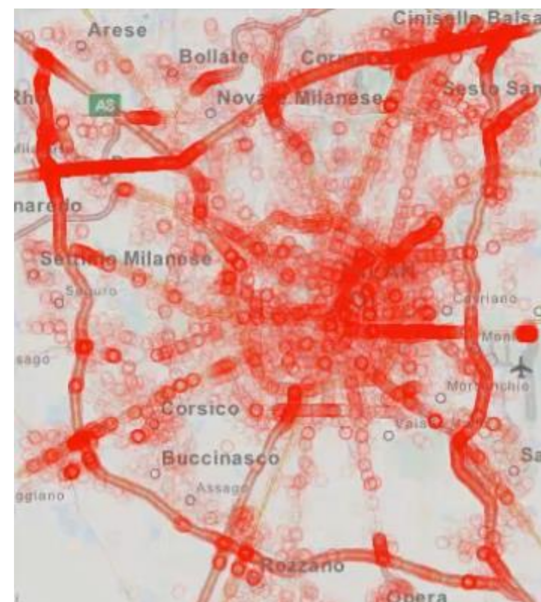
  Trajectories → Events → Places → Spatial time series

Step 1: extract low speed events from the trajectories

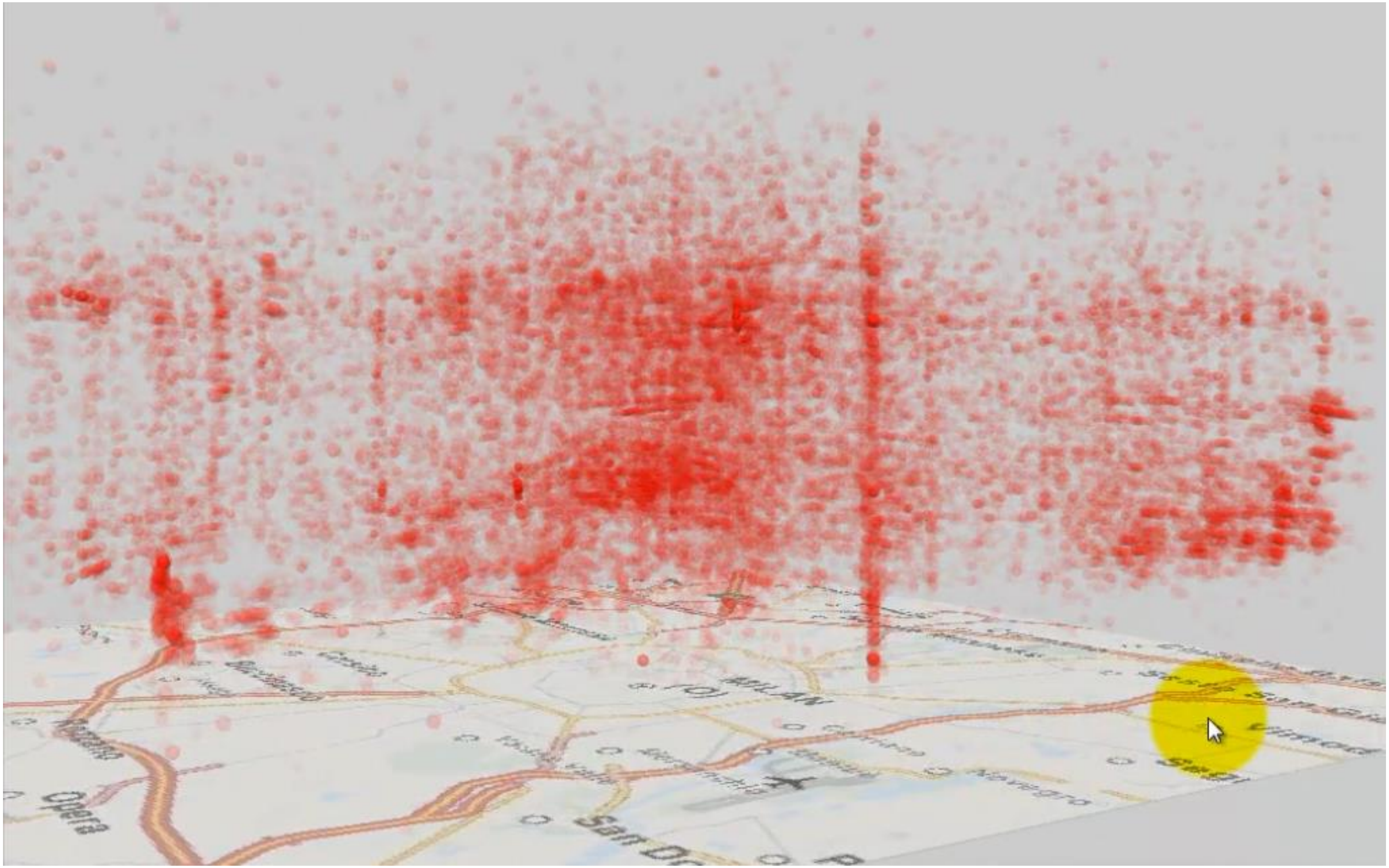








Low speed := speed ≤ 10 km/h



Vertical dimension ← time

Step 2: density-based clustering of events

by spatio-temporal positions and directions

Distance function:

$$d = \begin{cases} \infty, & if\ (d_s > D_s)\ or\ \exists i \mid (d_i > D_i), \quad i = 0..n \\ \begin{cases} D_s * max\left(\frac{d_s}{D_s}, \frac{d_0}{D_0}, \ldots, \frac{d_n}{D_n}\right), & if\ (a) \\ D_s * \sqrt{\left(\frac{d_s}{D_s}\right)^2 + \sum_{i=0}^{n}\left(\frac{d_i}{D_i}\right)^2}, & if\ (b) \end{cases} \end{cases}$$

(a) – neighbourhood defined as a cube

(b) – neighbourhood defined as a sphere

D_s – spatial distance threshold; $D_0,D_1,\ldots,D_N$ - distance thresholds for other attributes

$d_s,d_0,d_1,\ldots,d_N$ – distances; d_s – distance in space

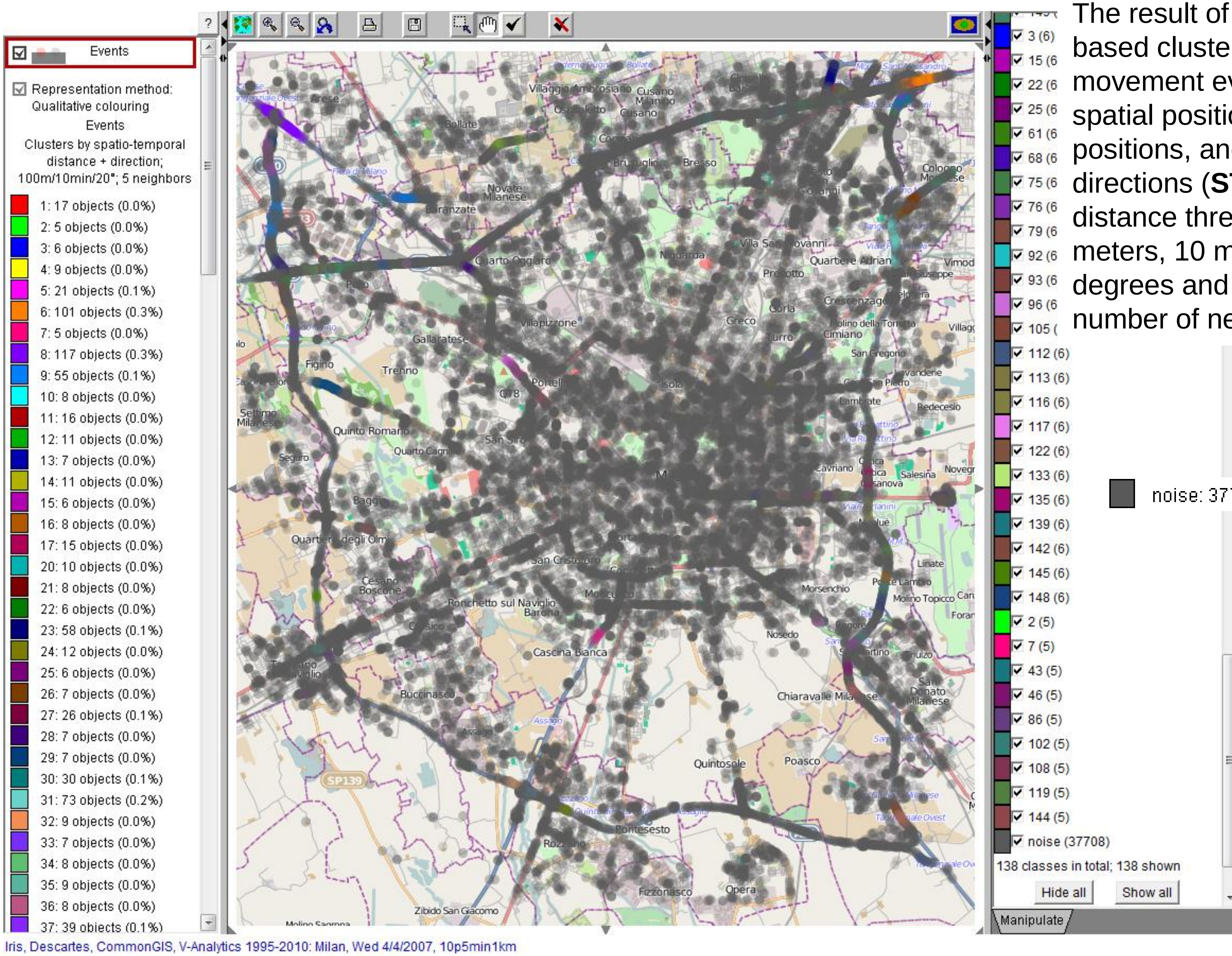
Distance in time (t_1, t_2 are intervals):

$$d_t(t_1,t_2) = \begin{cases} t_2^{start} - t_1^{end} & if\ t_1^{end} < t_2^{start} \\ t_1^{start} - t_2^{end} & if\ t_1^{start} > t_2^{end} \\ 0 & otherwise \end{cases}$$

Distance for a cyclic attribute (V is the cycle length):

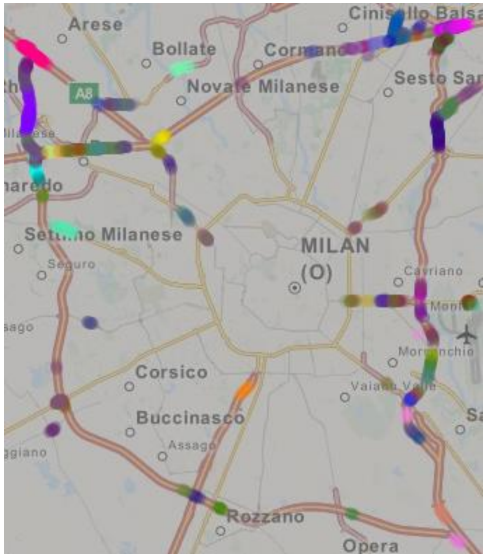
$$d(v_1, v_2, V) = \begin{cases} |v_1 - v_2|, & |v_1 - v_2| < V/2 \\ V - |v_1 - v_2|, & otherwise \end{cases}$$

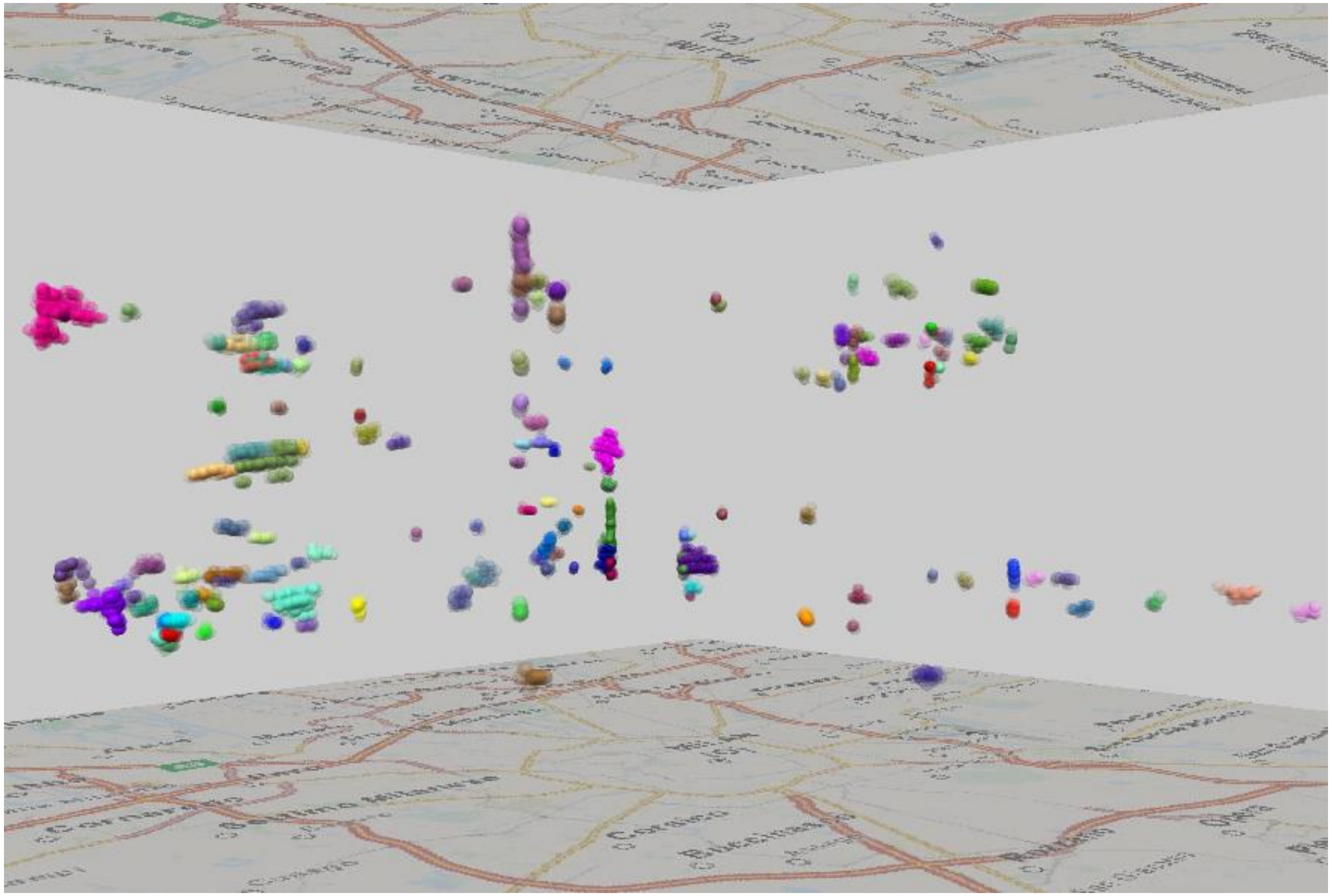
E.g., direction: V = 360°; d(5°,355°, 360°) = 10°



The result of the density-based clustering of the slow movement events by their spatial positions, temporal positions, and movement directions (**STD**) with the distance thresholds 100 meters, 10 minutes, and 20 degrees and the minimum number of neighbors 5.

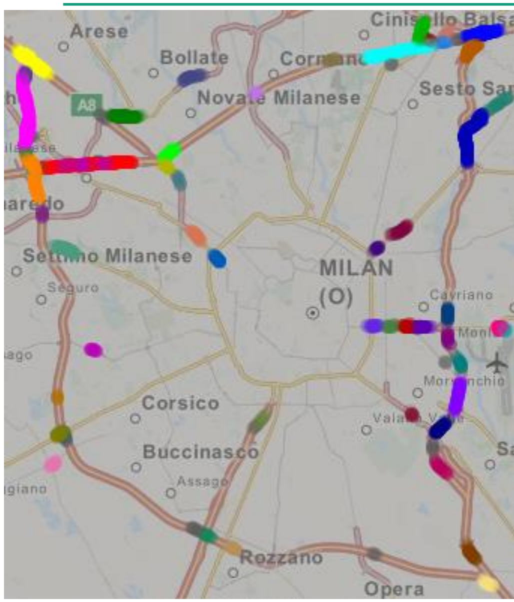
The STD-clusters, noise hidden



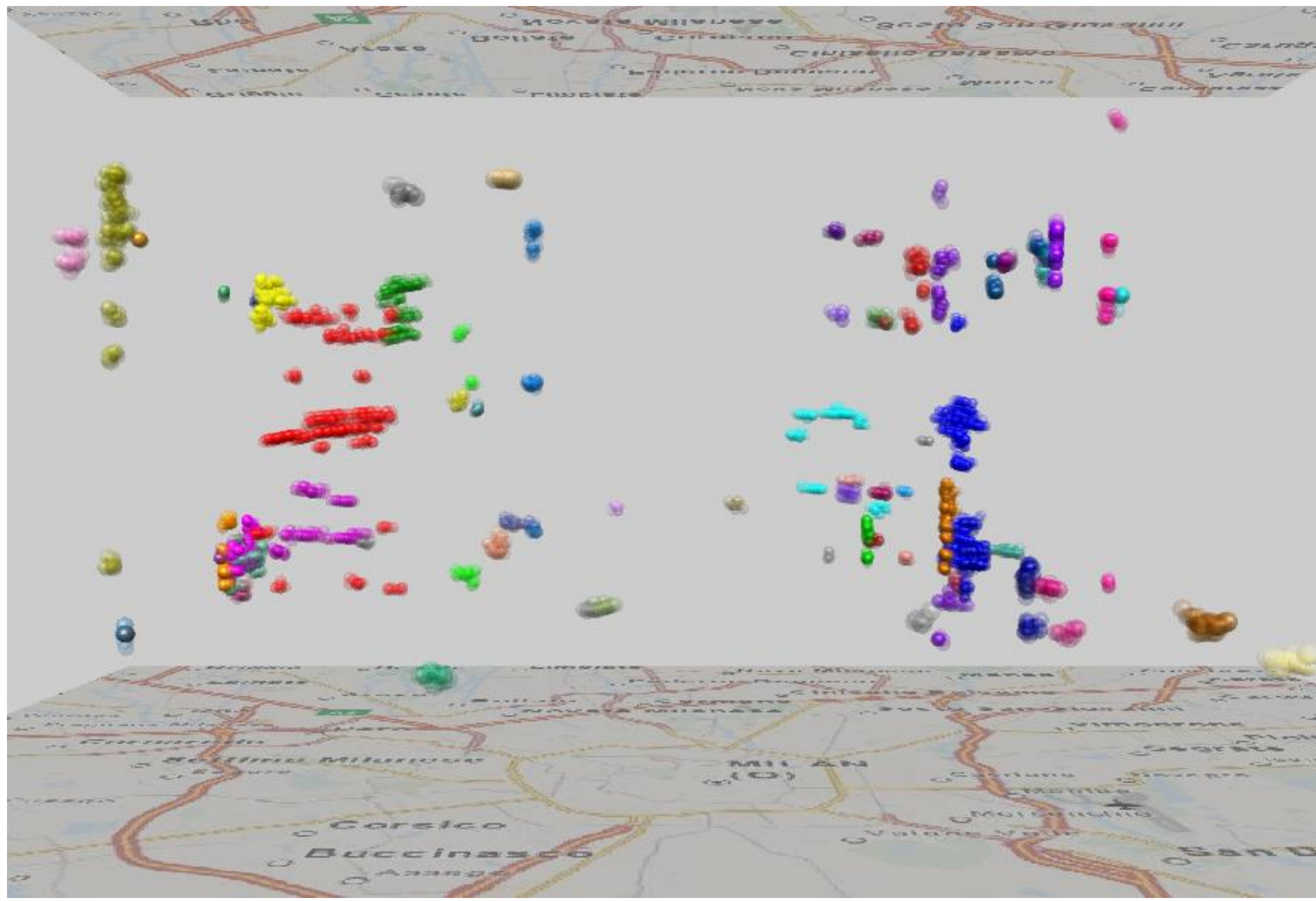


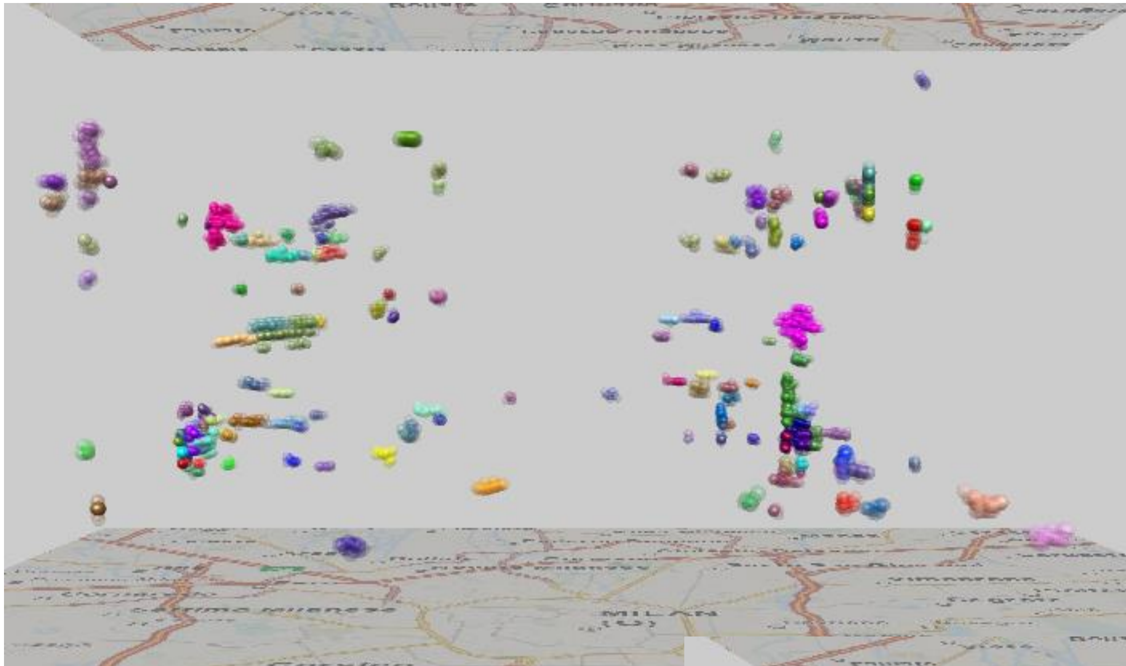
Step 3: unite STD-clusters in SD-clusters

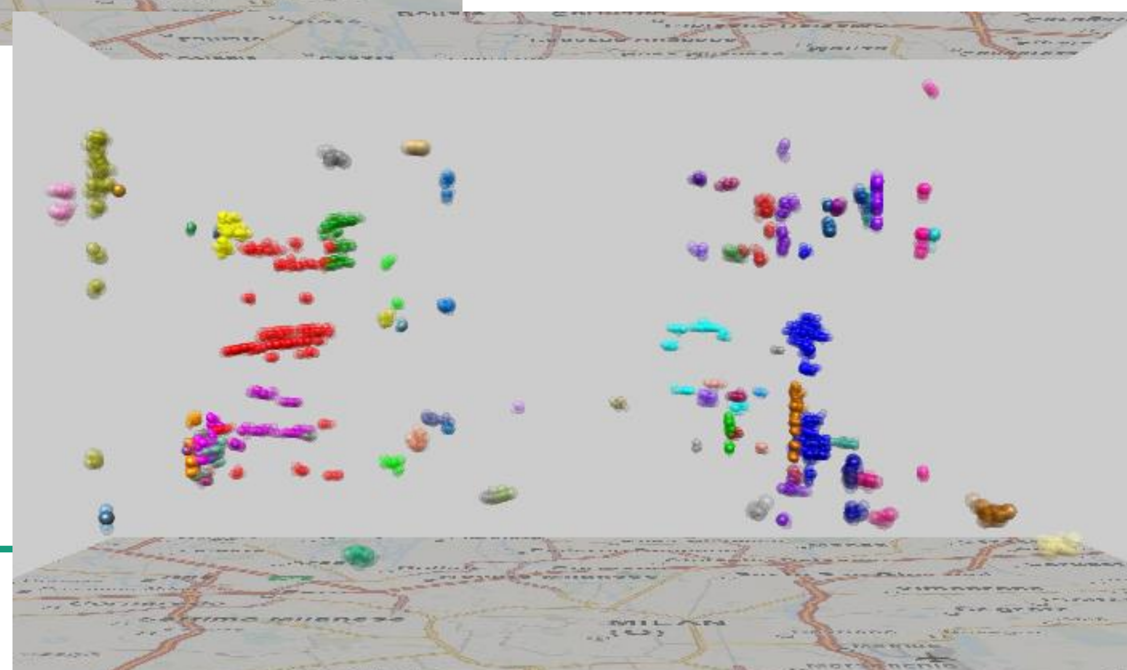
Cluster the events from the STD-clusters by the spatial positions and directions



The result of the density-based clustering with the spatial distance threshold of 100 m and direction distance threshold of 20°





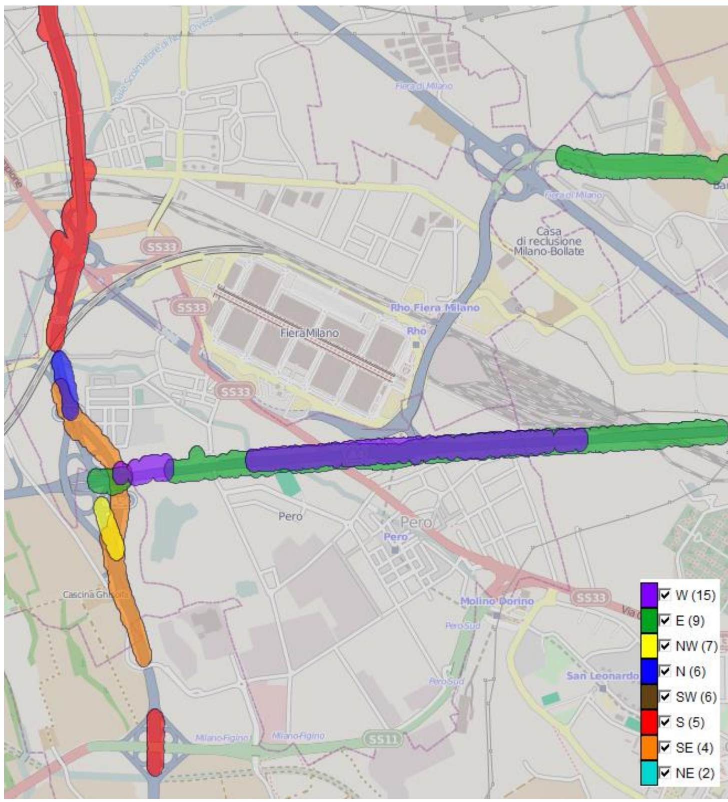


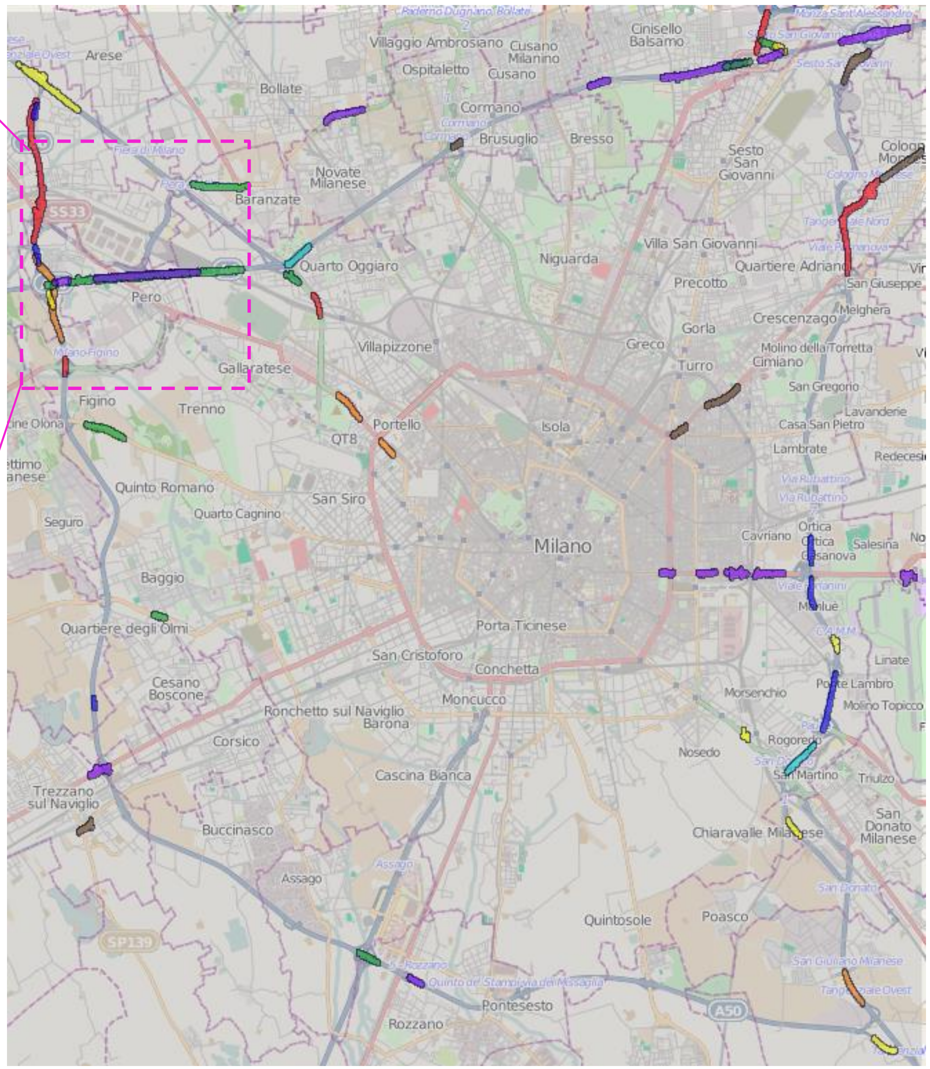
Events that occurred in same or close places but in different times were formerly in different clusters, but now they are in the same clusters.

One SD-cluster includes one or several STD-clusters.

Step 4: outline the places of interest

Build spatial buffers around the SD-clusters of events



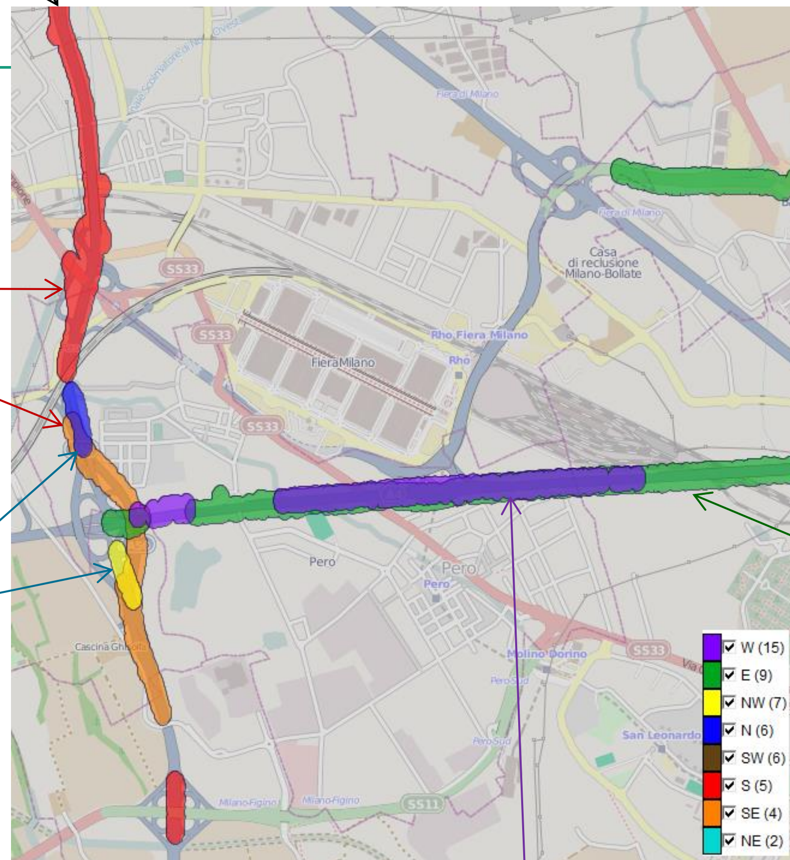


The places are painted according to the prevailing movement directions of the respective events.

Belt road north-south on the east of the city (A50)

Extended areas of congested traffic directed to the south and southeast

Smaller areas of obstructed movement directed to the north and northwest



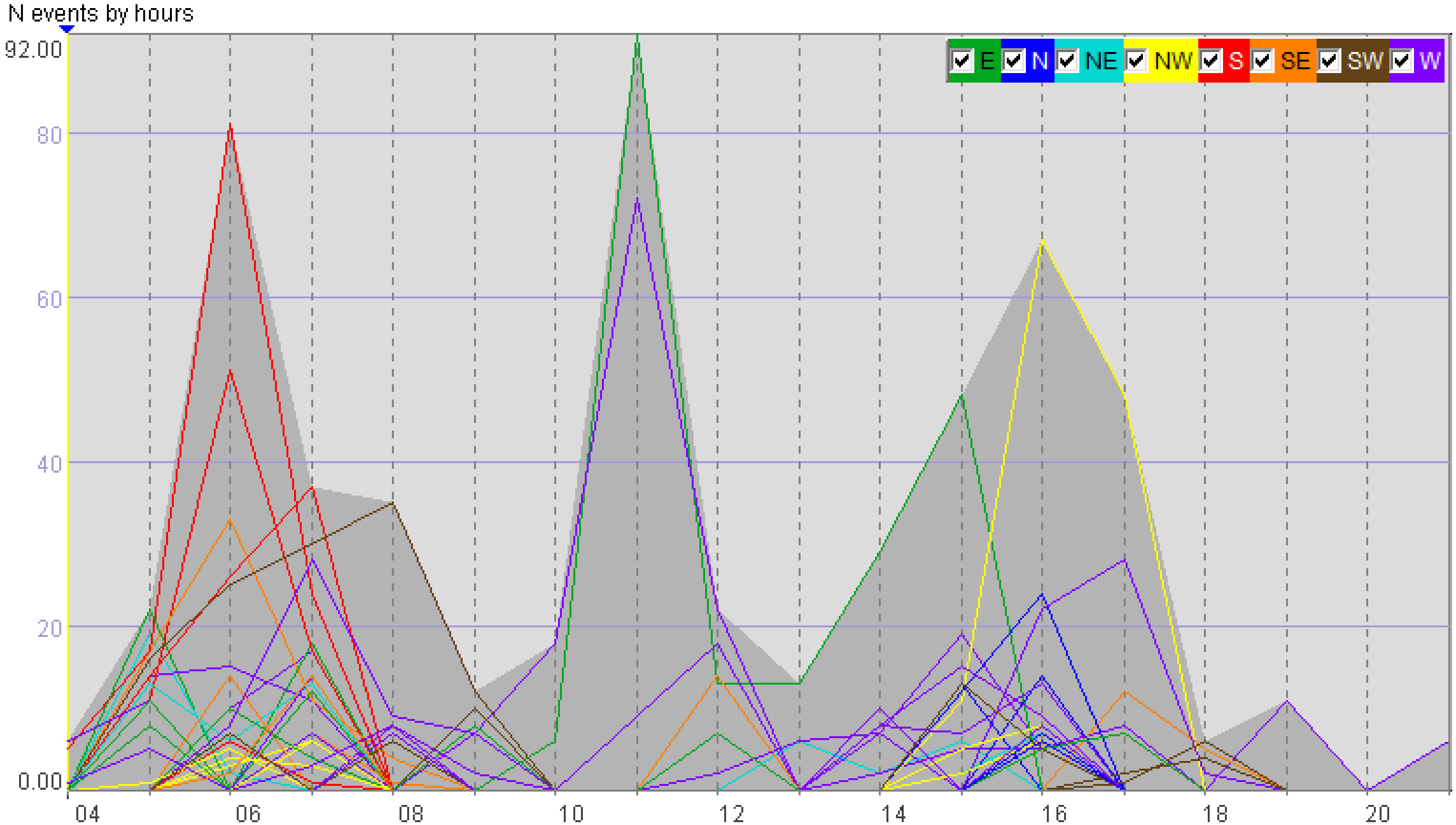
Belt road west-east on the north of the city (A4)

Very long area of congested traffic directed to the east

Long area of congested movements directed to the west

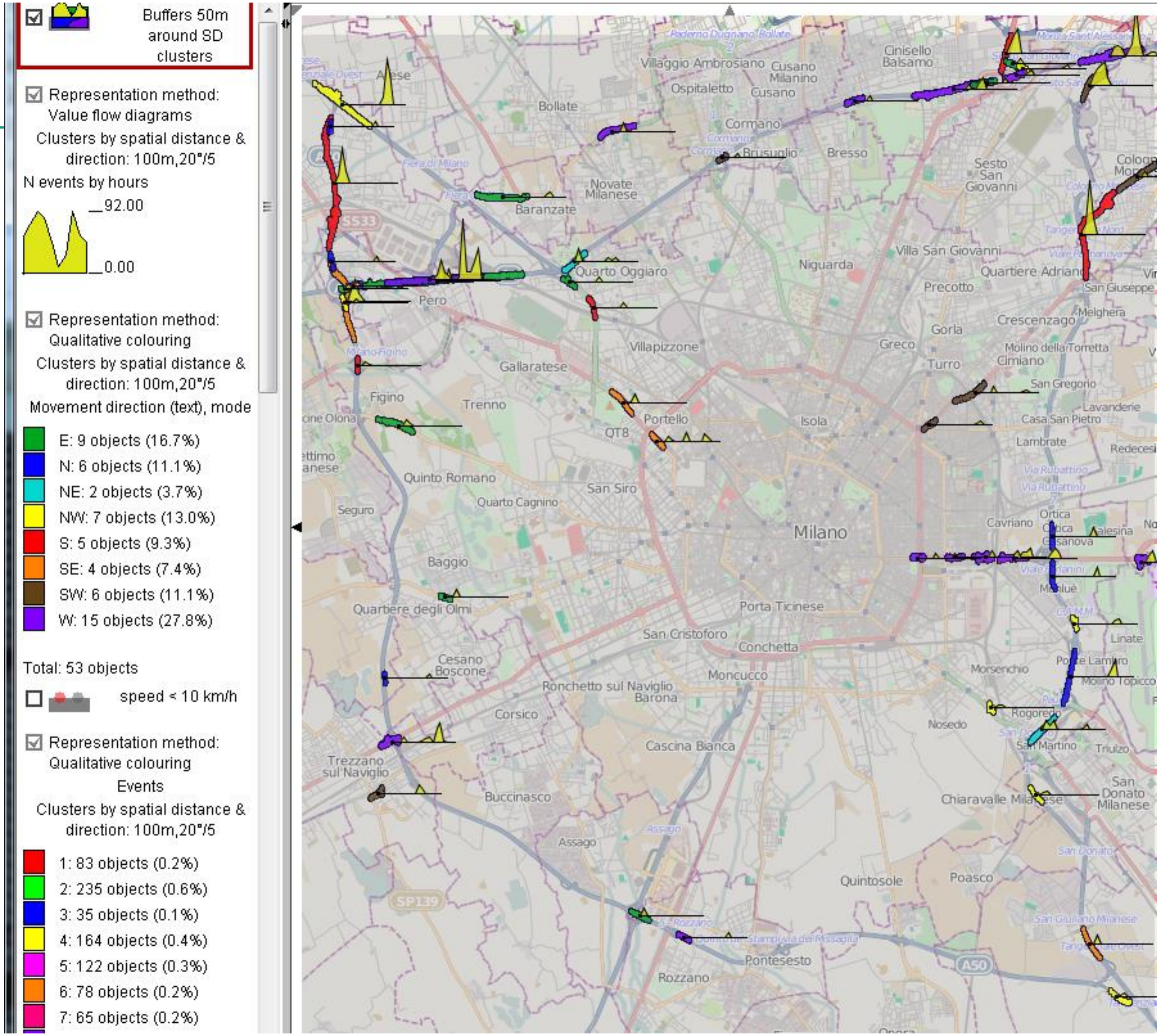
Step 5: aggregate data by the places

and by suitable time intervals, e.g., hourly

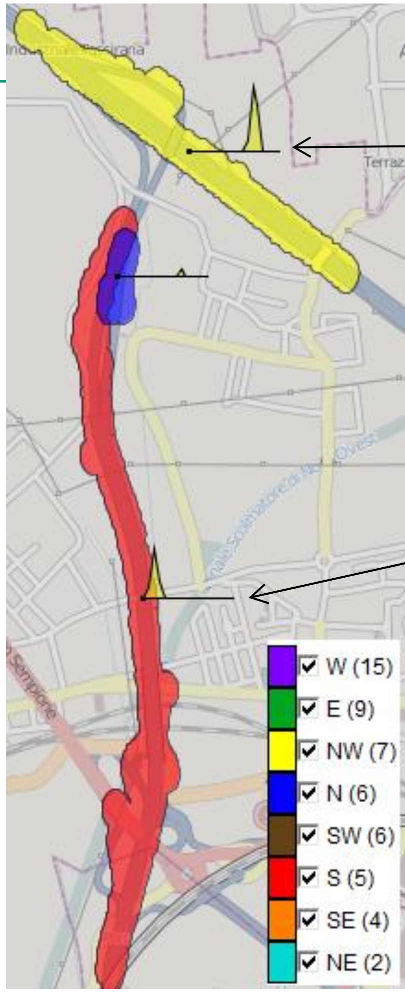


Place-referenced time series of the counts of slow movement events

The temporal diagrams show the variation of the attribute value (vertical dimension) over time (horizontal dimension).



Map fragment (northwest) enlarged

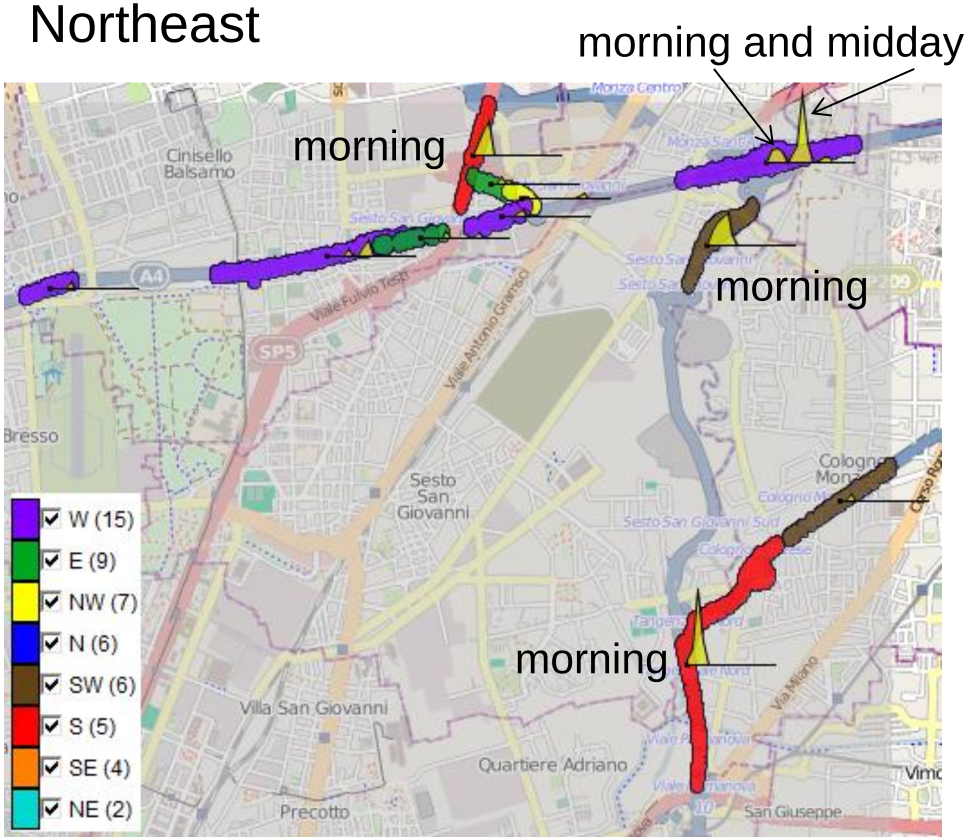


Congested traffic in the afternoon in the direction out of the city (northwest)

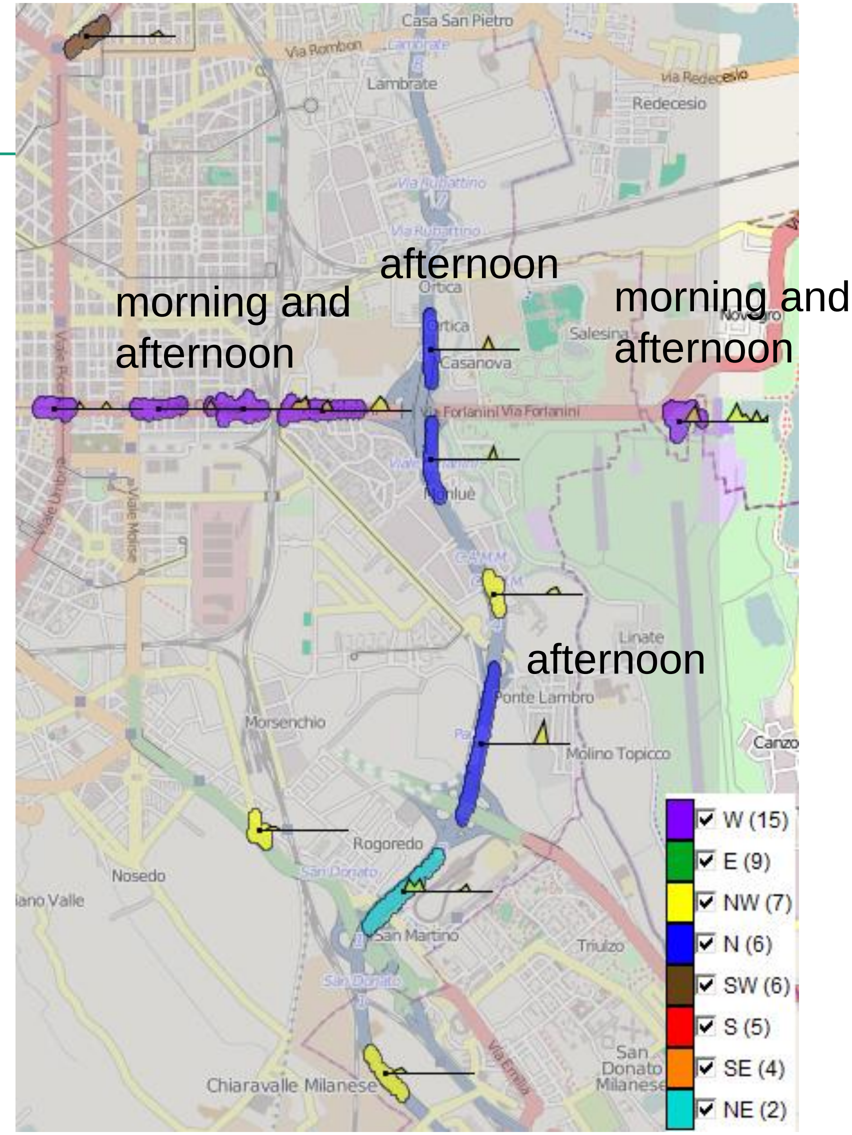
Congested traffic in the morning in the direction to the south

Other map fragments enlarged

Northeast



East



Where to read more

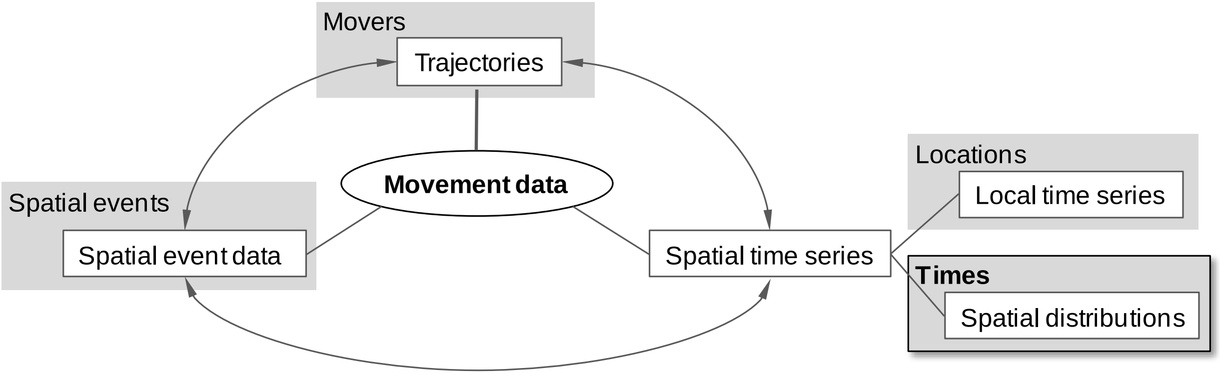
- IEEE VAST 2011 paper (**best paper** award)

  G.Andrienko, N.Andrienko, C.Hurter, S.Rinzivillo, S.Wrobel
  From Movement Tracks through Events to Places:
  Extracting and Characterizing Significant Places from Mobility Data
  IEEE Visual Analytics Science and Technology (VAST 2011),
  Proceedings, IEEE Computer Society Press, 183-192

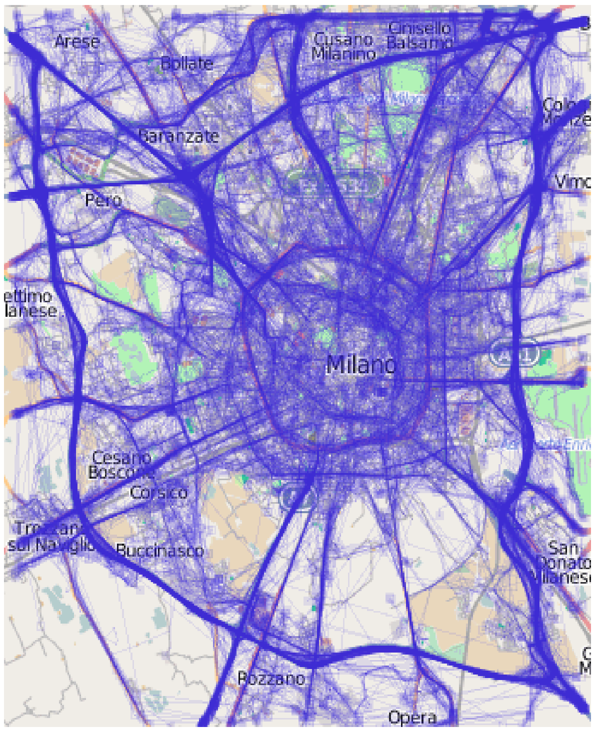
- Extended version, covering also scalable clustering of events

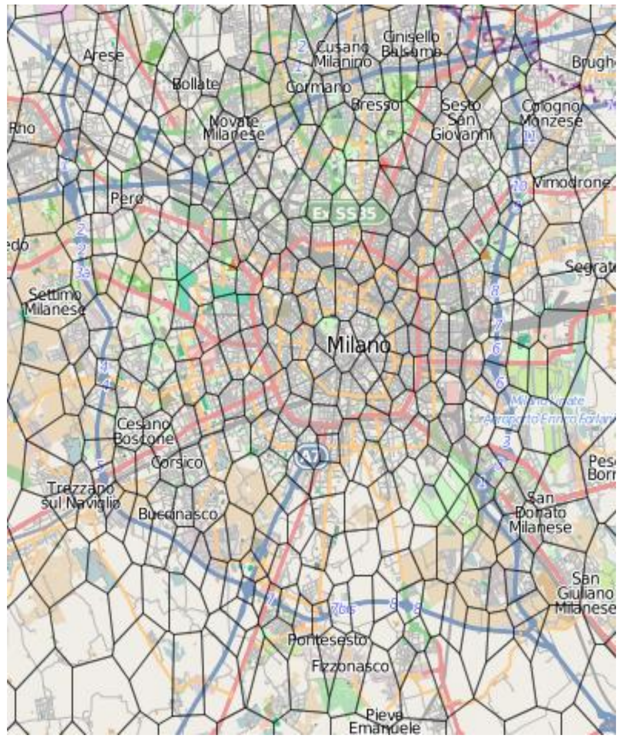
  G.Andrienko, N.Andrienko, C.Hurter, S.Rinzivillo, S.Wrobel
  Scalable Analysis of Movement Data for Extracting and Exploring Significant Places
  IEEE Transactions on Visualization and Computer Graphics,
  2013, 19(7), 1078-1094

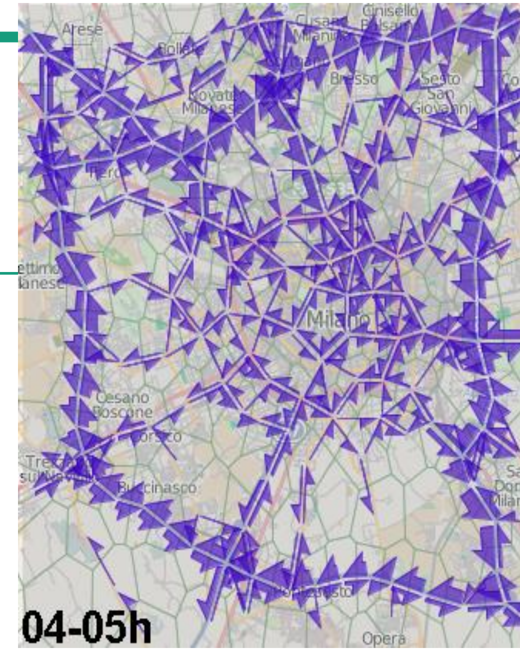
Perspective 3:
Movement data in the form of spatial situations

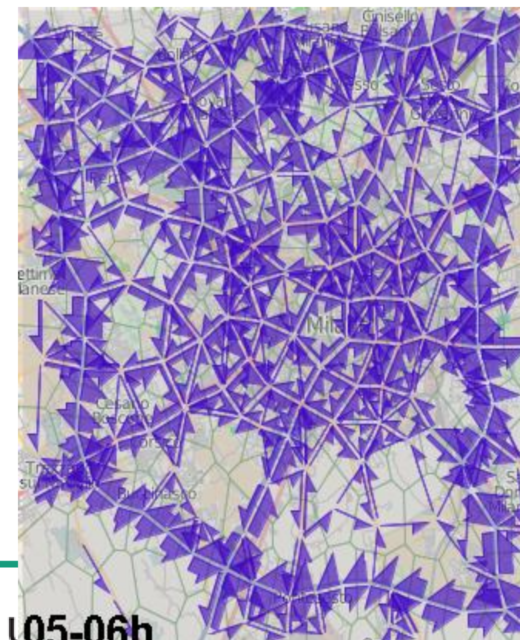


Spatio-temporal aggregation of trajectories



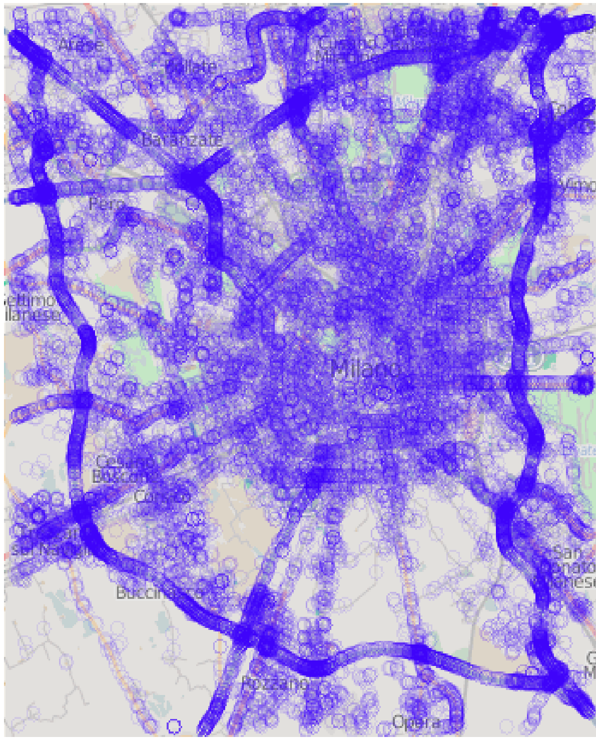




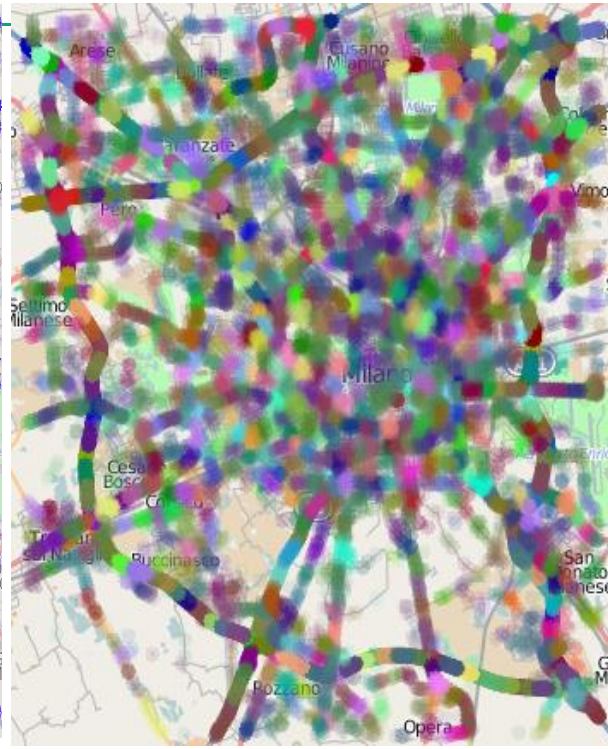


Division of the territory

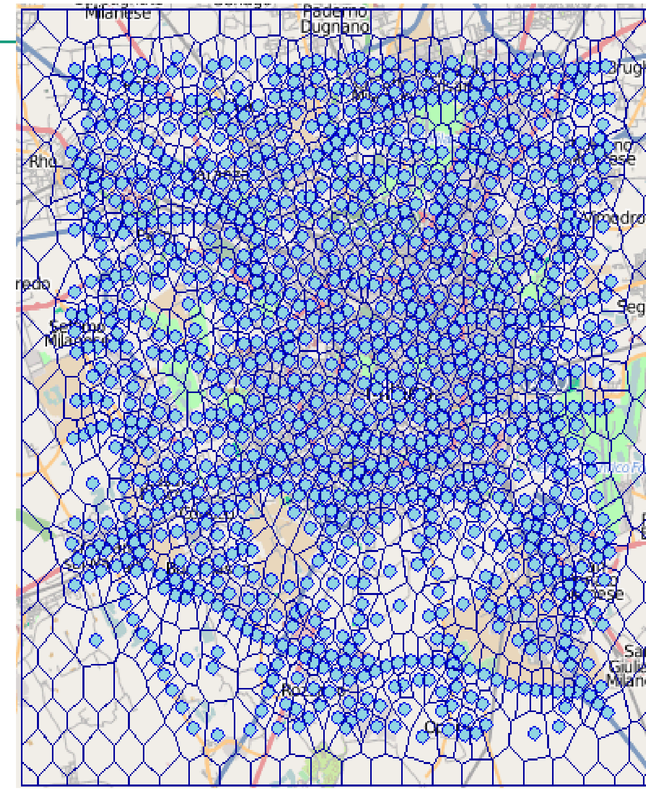
Characteristic points from the trajectories



Spatial clusters of characteristic points



Cluster centres → seeds for Voronoi tessellation



Details:

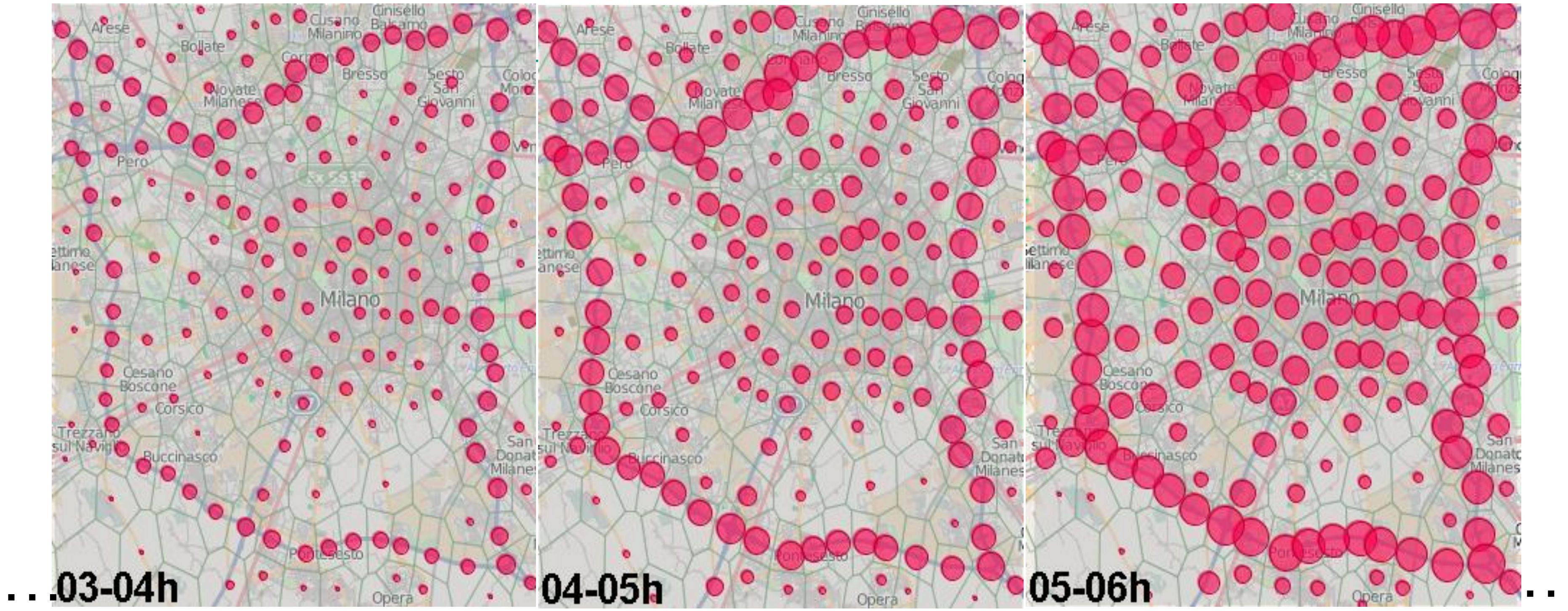
Natalia Andrienko, Gennady Andrienko

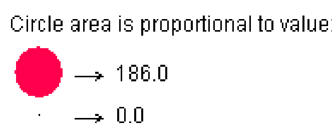
Spatial Generalization and Aggregation of Massive Movement Data

IEEE Transactions on Visualization and Computer Graphics (TVCG), 2011, v.17 (2), pp.205-219

http://doi.ieeecomputersociety.org/10.1109/TVCG.2010.44

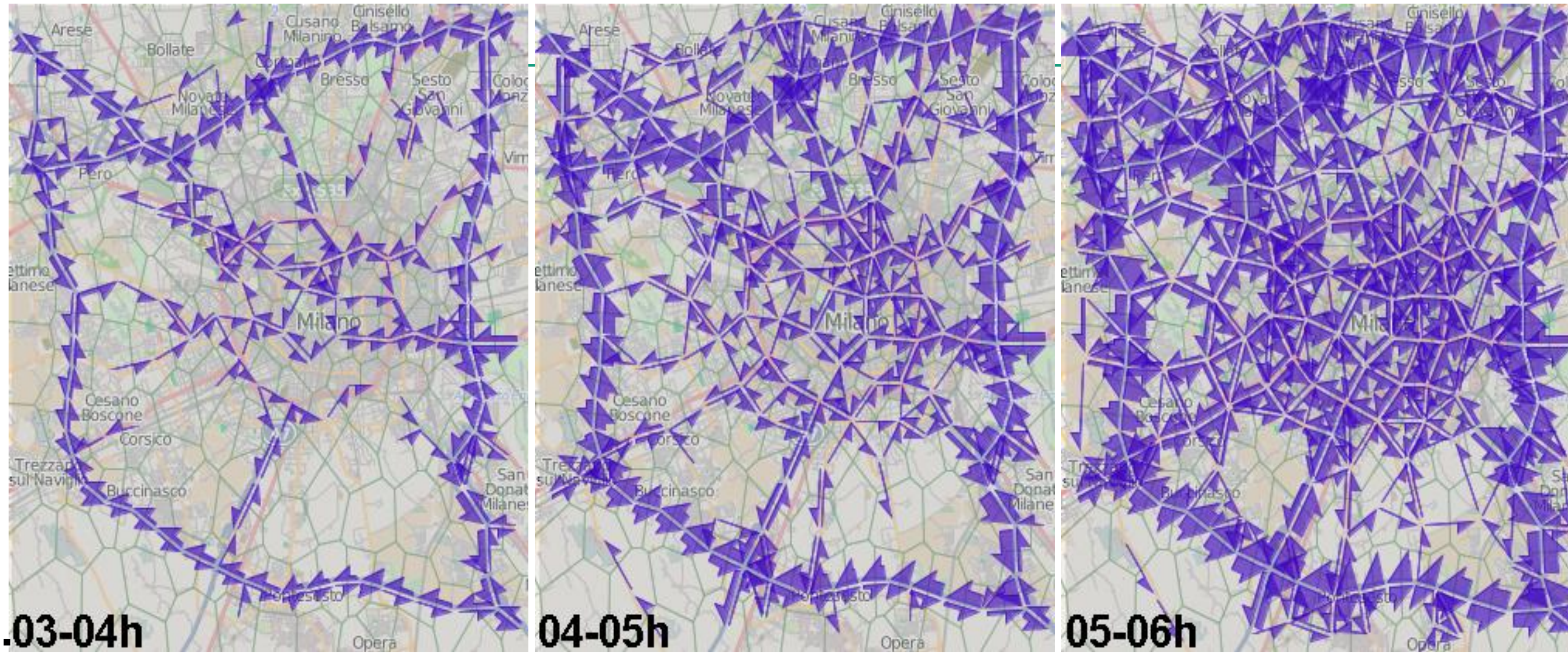
Spatial situations: presence



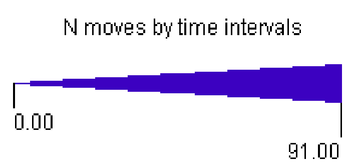


Situation is described by tuples like
<Place_id, time, attribute(s)>

Spatial situations: flows

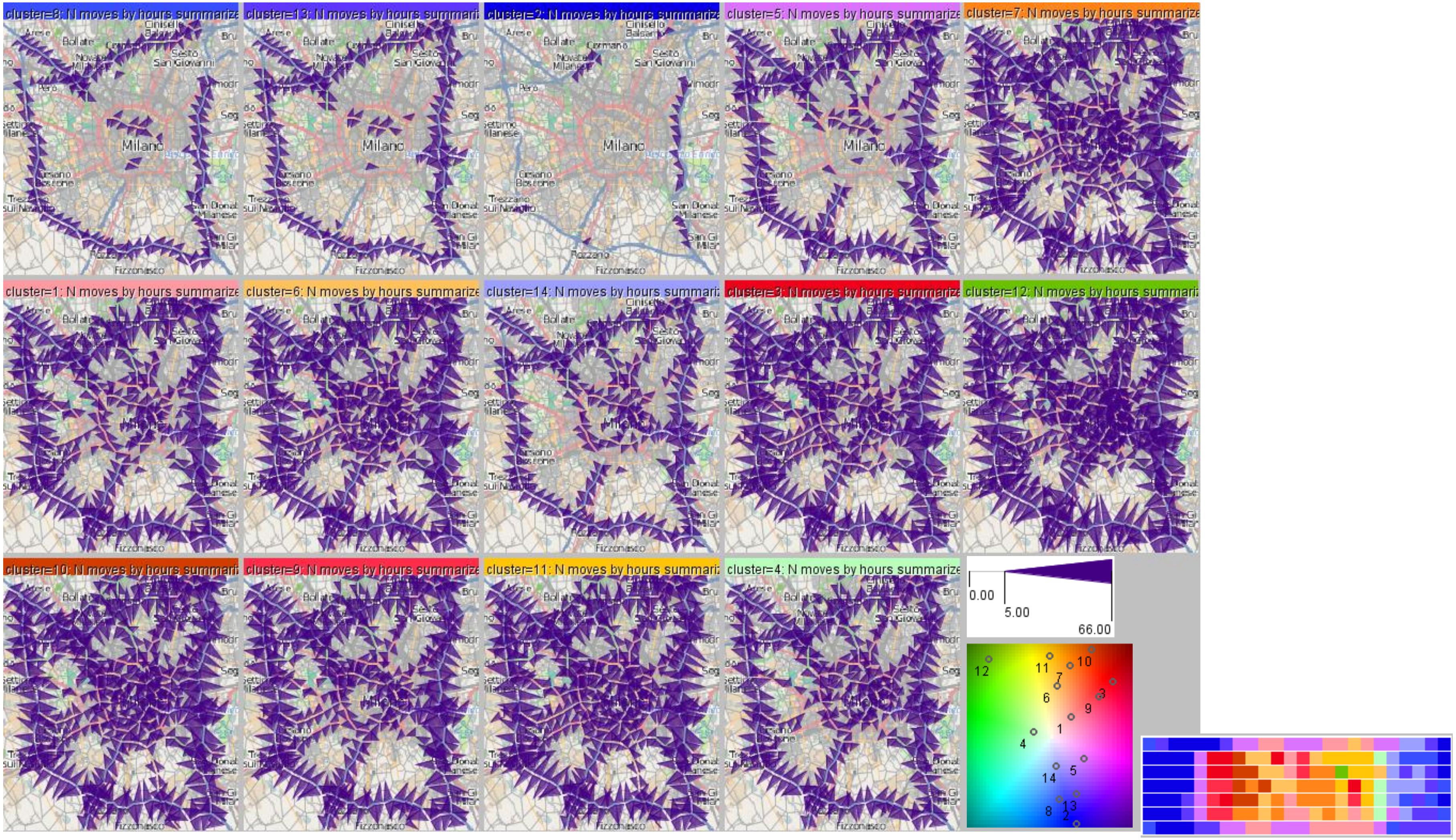
. . .


. . .

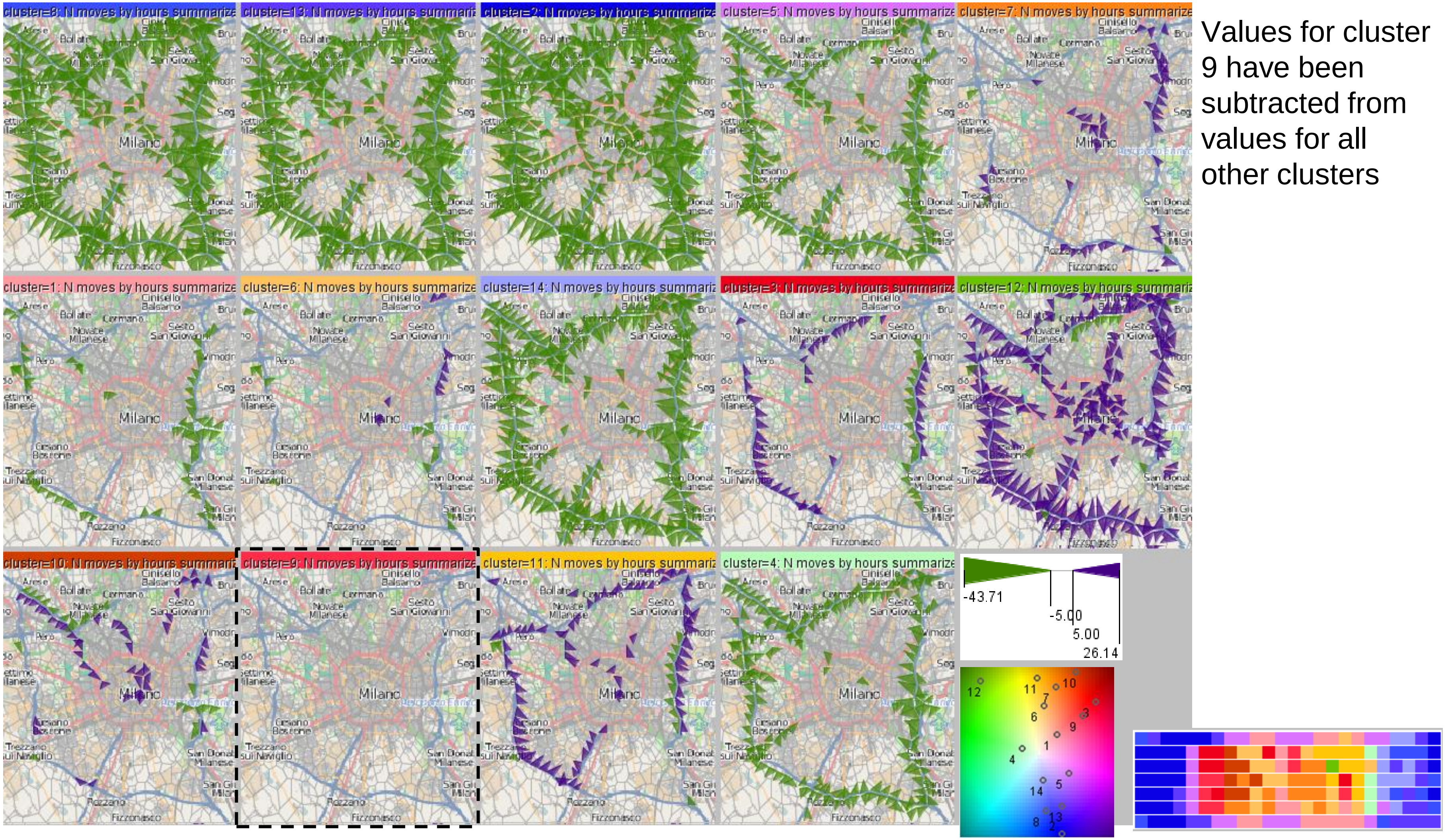


Situation is described by tuples like <Link_id, time, attribute(s)>

Clustering of spatial (flow) situations by similarity



Comparison of clusters of spatial situations



Values for cluster 9 have been subtracted from values for all other clusters

Where to read more

N.Andrienko, G.Andrienko, H.Stange, T.Liebig, D.Hecker

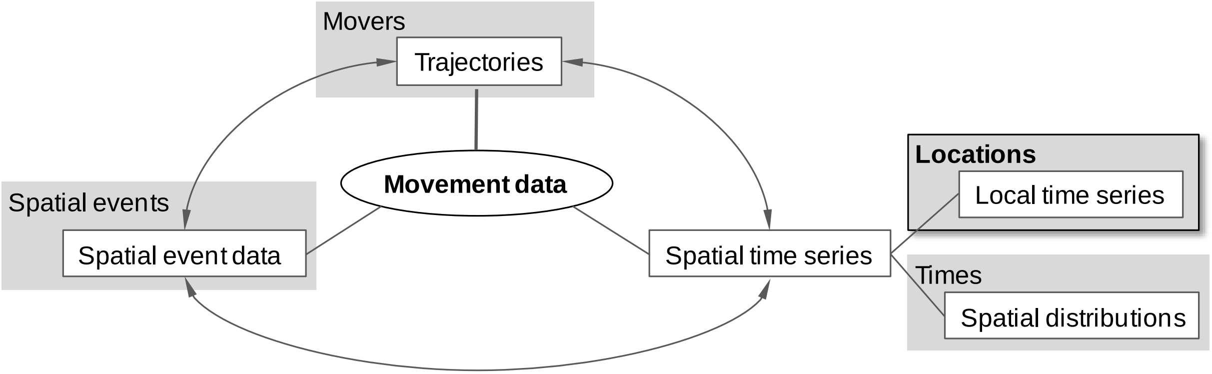
Visual Analytics for Understanding Spatial Situations from

Episodic Movement Data

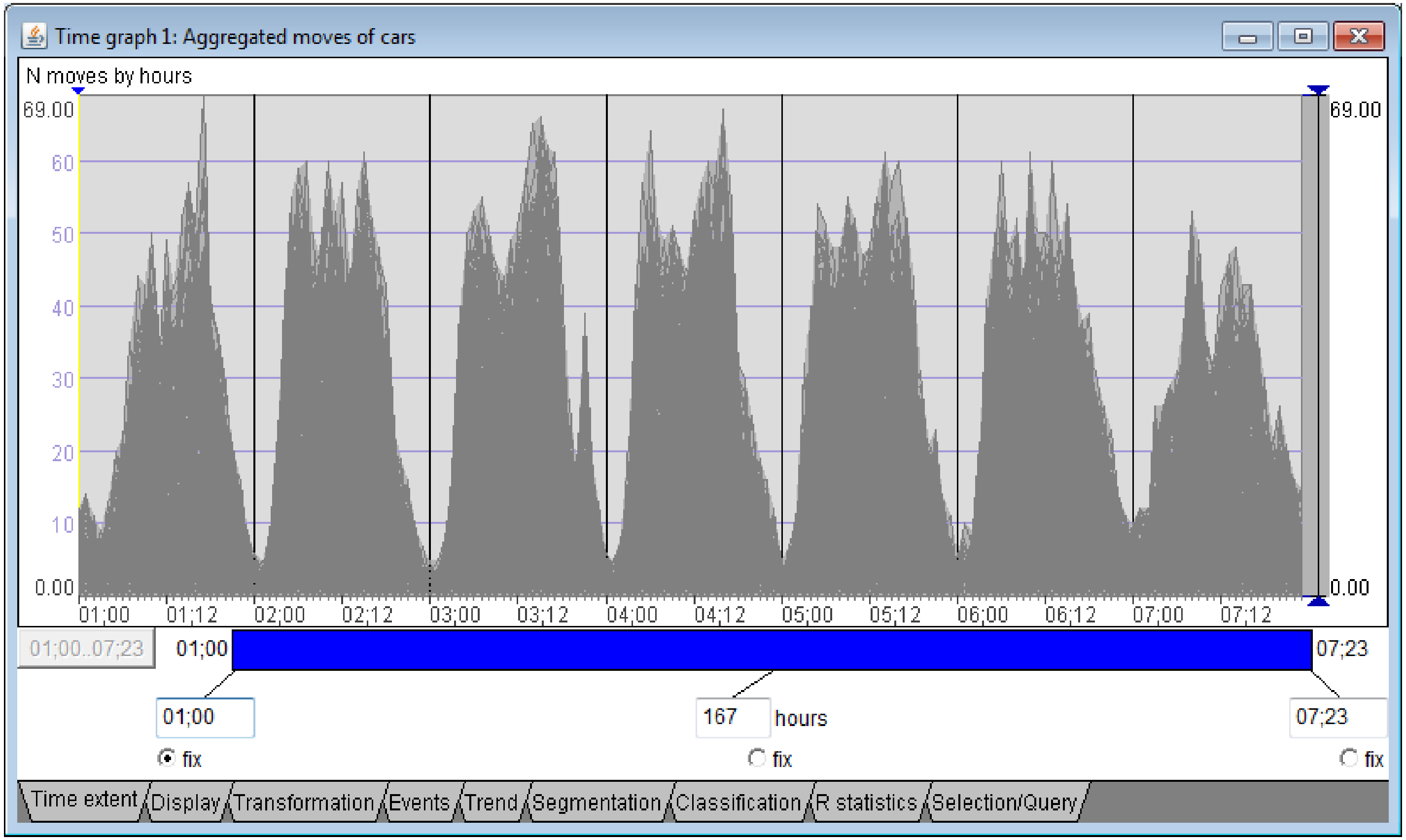
Künstliche Intelligenz, 2012, v.**26** (3), pp.241-251

http://dx.doi.org/10.1007/s13218-012-0177-4

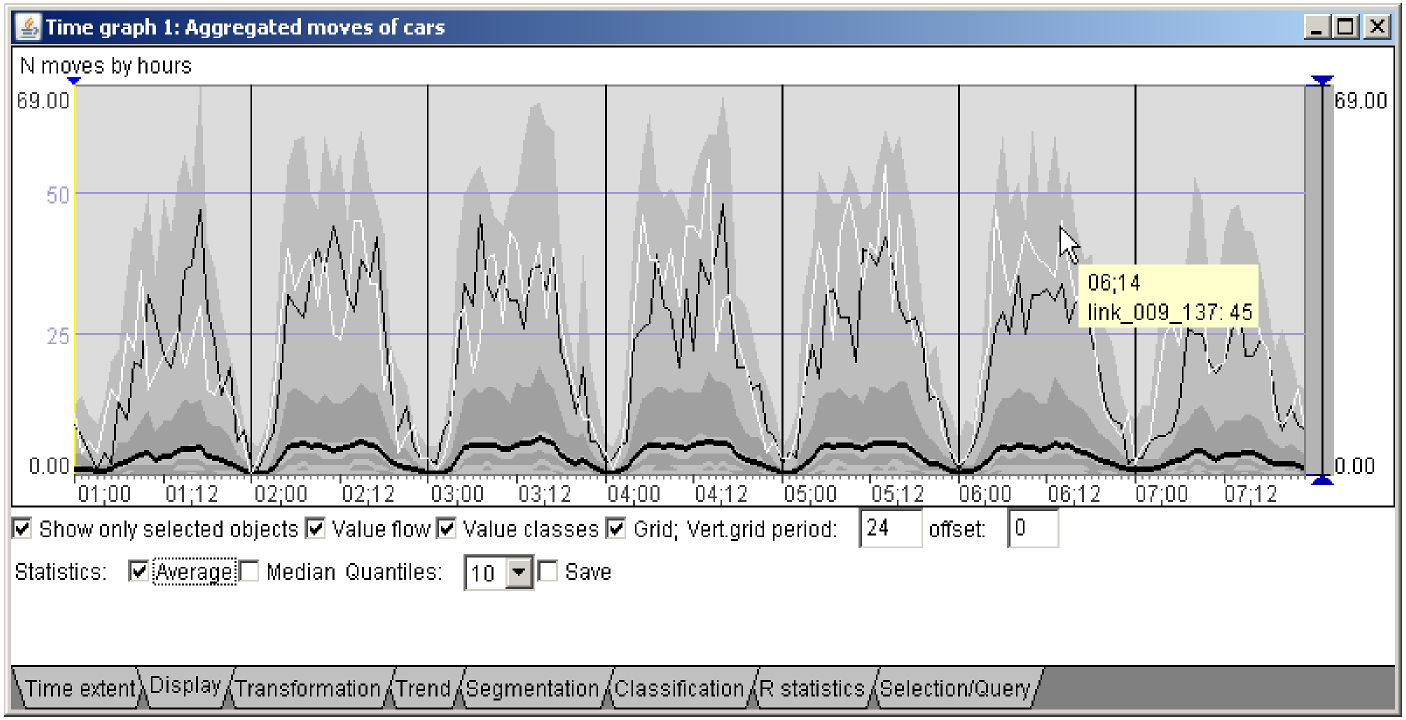
Perspective 4: Movement data in the form of local time series



An alternative view of spatial time series: a set of local time series



An alternative view of spatial time series : a set of local time series



We wish to represent the essential characteristics of the ST-variation explicitly by a formal model or a set of models.

Methods for spatio-temporal modelling (e.g. STARIMA)

- Account for spatial and temporal dependencies
- Require prior specification of multiple weight matrices expressing impacts among locations for different temporal lags
  - may be difficult (the impacts are not easy to quantify)
- Build a single global model of the entire spatio-temporal variation
  - It does not necessarily perform better than a set of local temporal models
- Assume spatial smoothness of the modelled phenomenon, i.e., closer places are more similar than more distant ones
  - May be not very suitable for spatially abrupt phenomena

Existing techniques for time series modelling

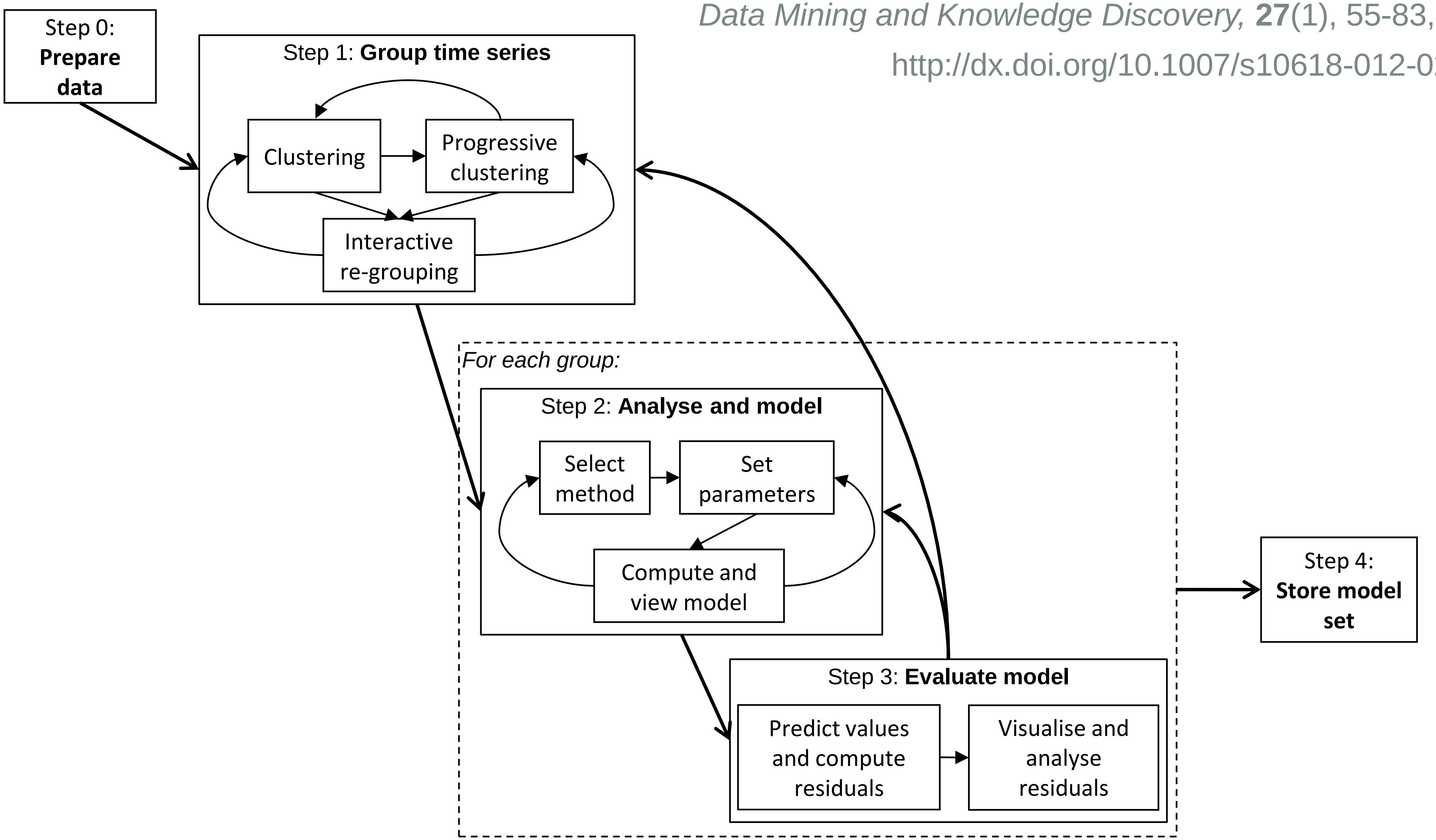
- \+ Widely available in numerous statistical packages and libraries → can be applied to spatially referenced time series
- \- The modelling methods are designed to deal with singular time series → hard to use for a large number of time series
- \- Separate consideration of each time series ignores the phenomenon of spatial dependence (relatedness and similarities among spatial locations or objects)
- \- Separate consideration of each time series does not allow data abstraction and generalisation over space

Combination of spatial and temporal modelling

- Approach 1:
  1. Model the temporal variation independently for each location
  2. Model the spatial variation of the parameters of the temporal models, e.g., as a random field
  - Assumes that the character of the temporal variation is the same everywhere and only the parameters differ
- Approach 2:
  - Model the spatial variation independently for each time step, e.g., as a random field
  - Model the temporal variation of the parameters of the spatial models at each location
- Both approaches assume spatial smoothness of the phenomenon

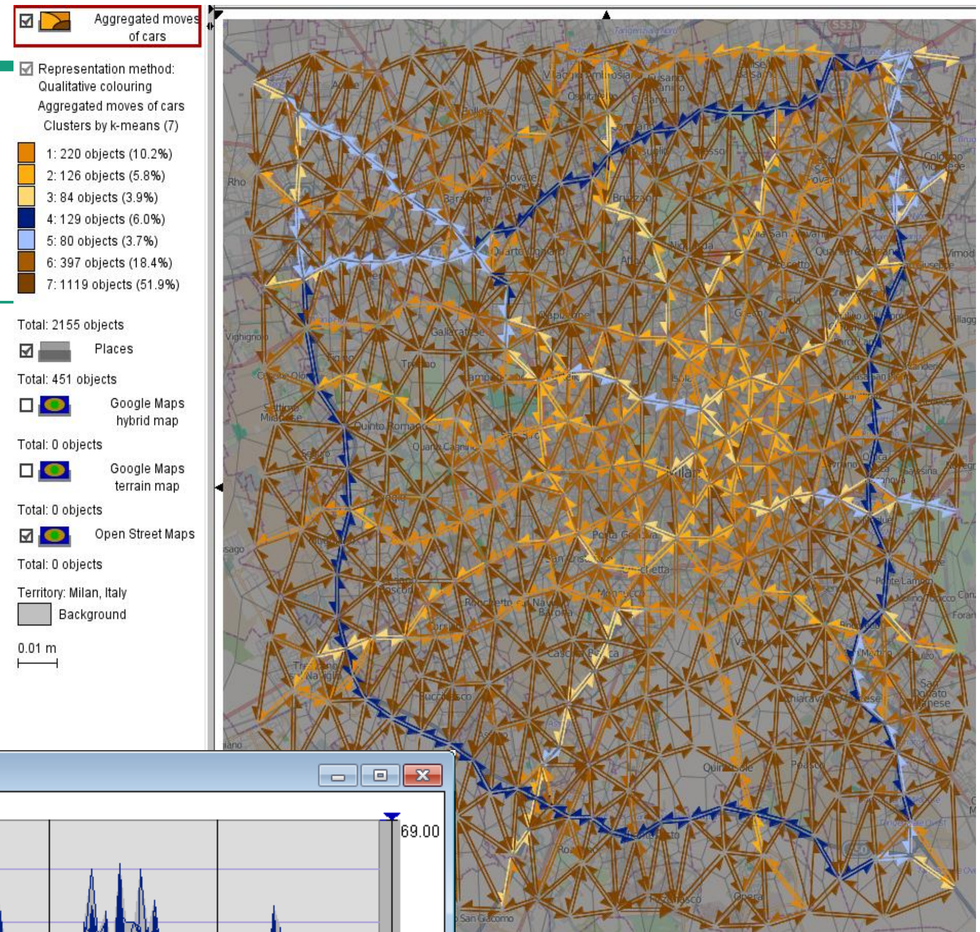
Our approach

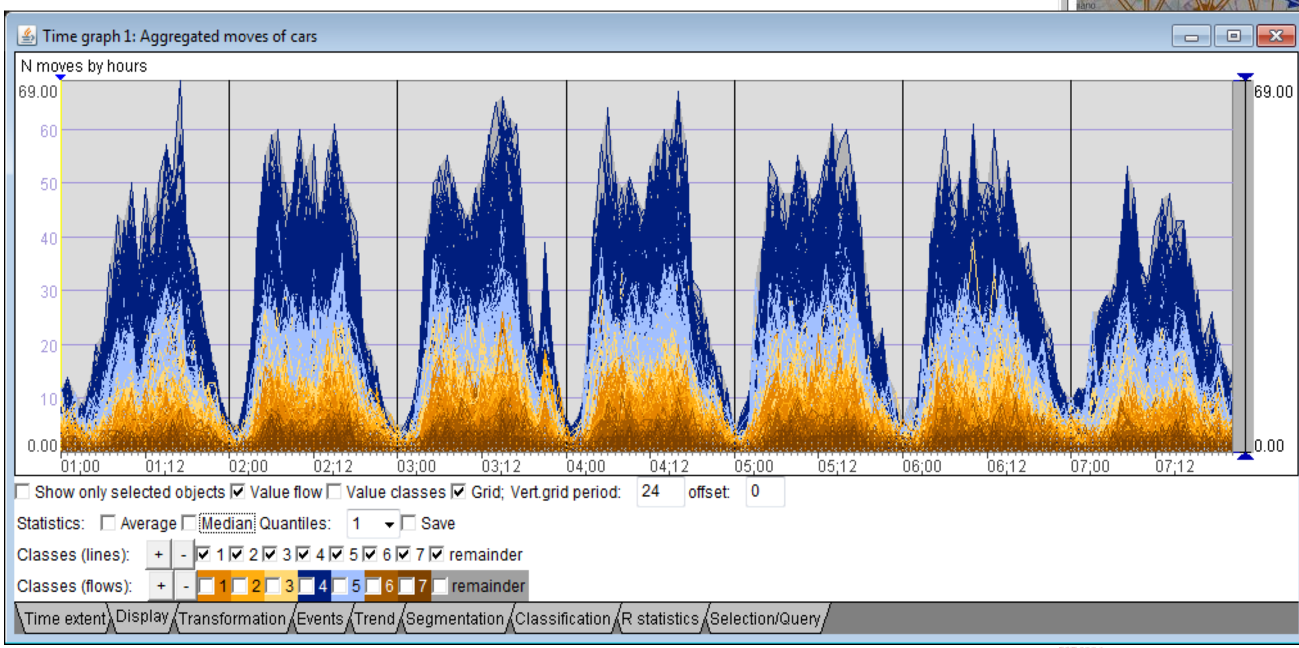
Details:
Natalia Andrienko, Gennady Andrienko
A Visual Analytics Framework for Spatio-temporal Analysis and Modelling
Data Mining and Knowledge Discovery, **27**(1), 55-83, 2013
http://dx.doi.org/10.1007/s10618-012-0285-7



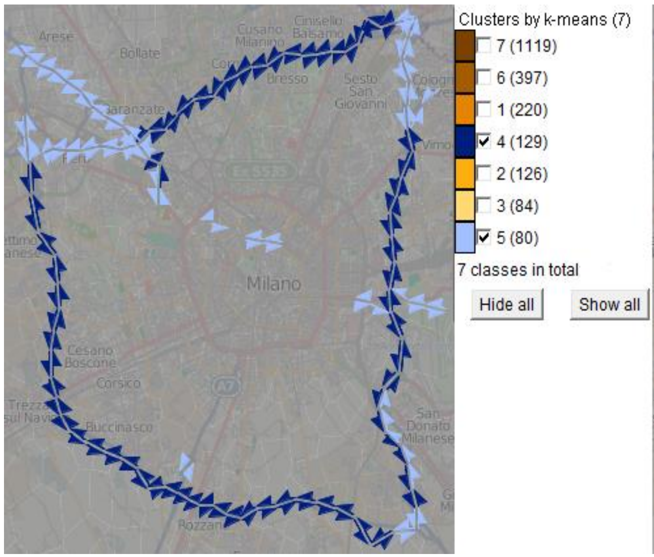
Step 1: Clustering of local TS

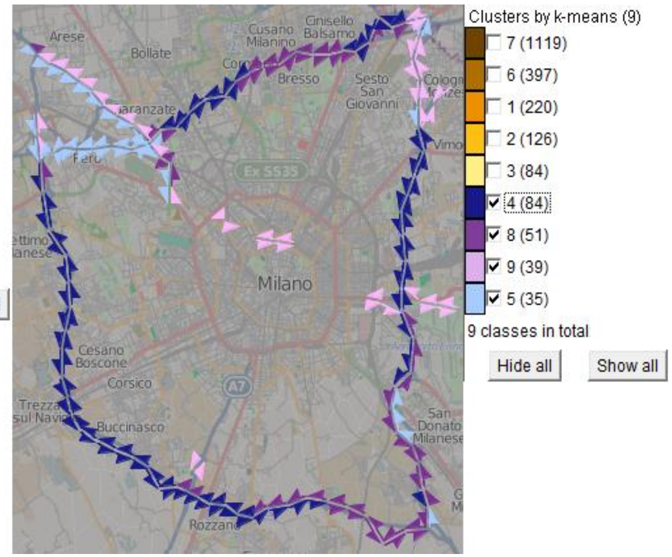
- Here: k-means (Weka) but may be another partition-based method
- Tried different k from 5 to 15
- Immediate visual response facilitates choosing the most suitable k

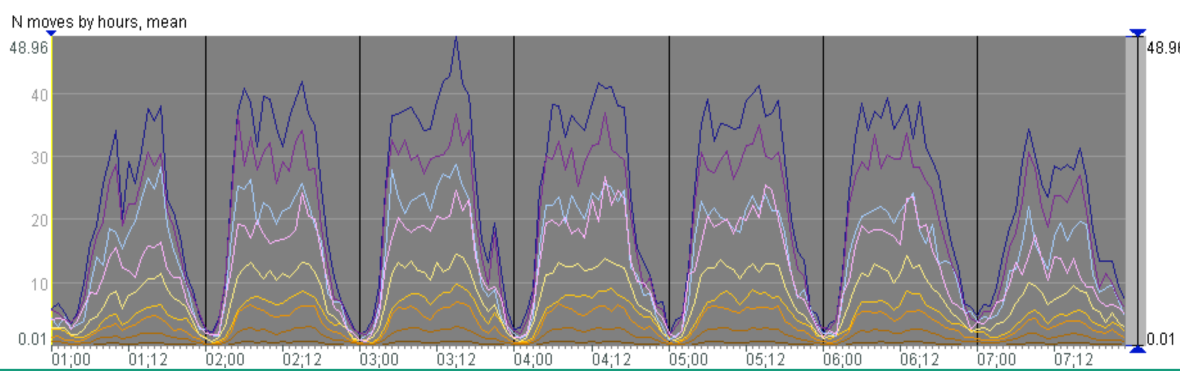




Step 1: Re-grouping by progressive clustering

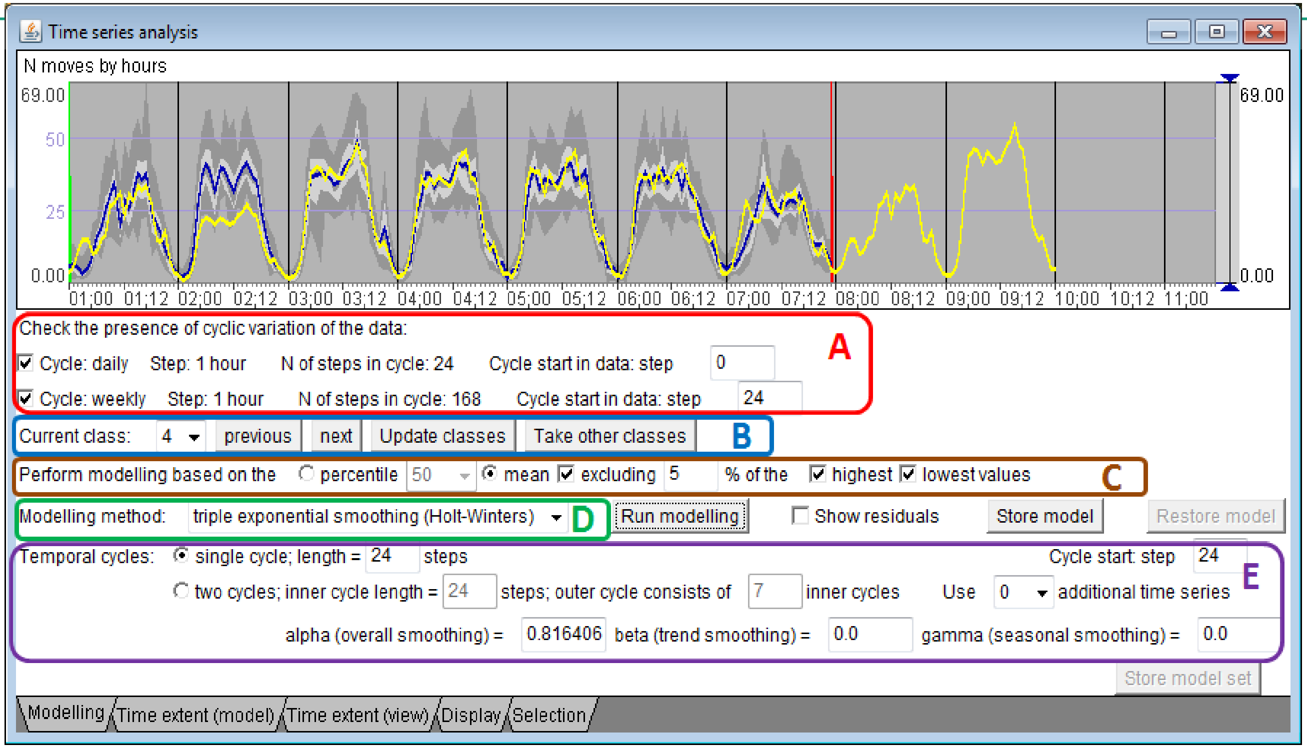






CITY UNIVERSITY LONDON
EST 1894
Fraunhofer
IAIS

Step 2: Analysis and modelling



A) Check automatically detected time cycles in the data.

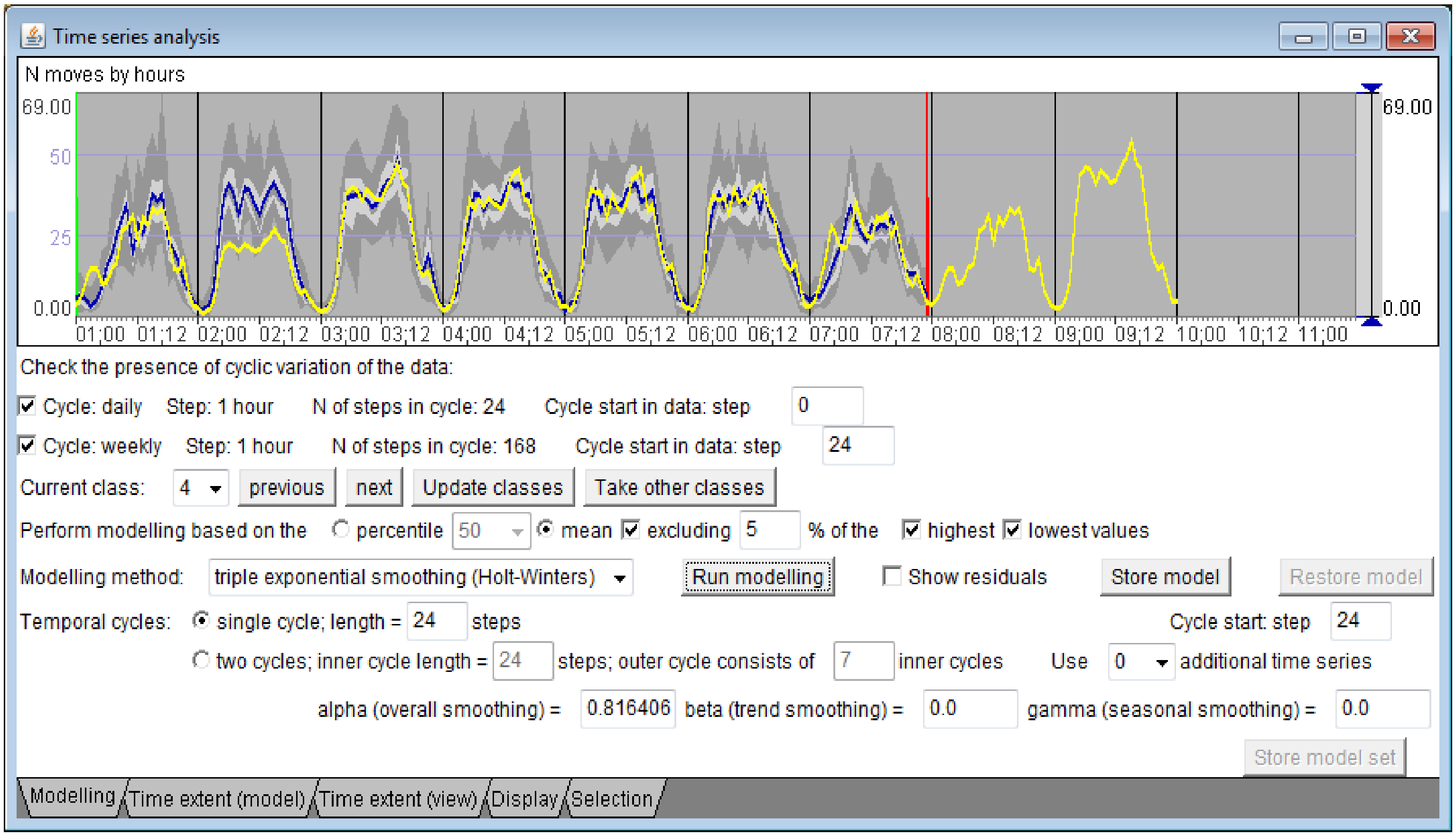
B) Select the current class (cluster) for the analysis and modelling.

C) Build the representative TS.

D) Select the modelling method.

E) View and modify model parameters (this section changes depending on the selected modelling method).

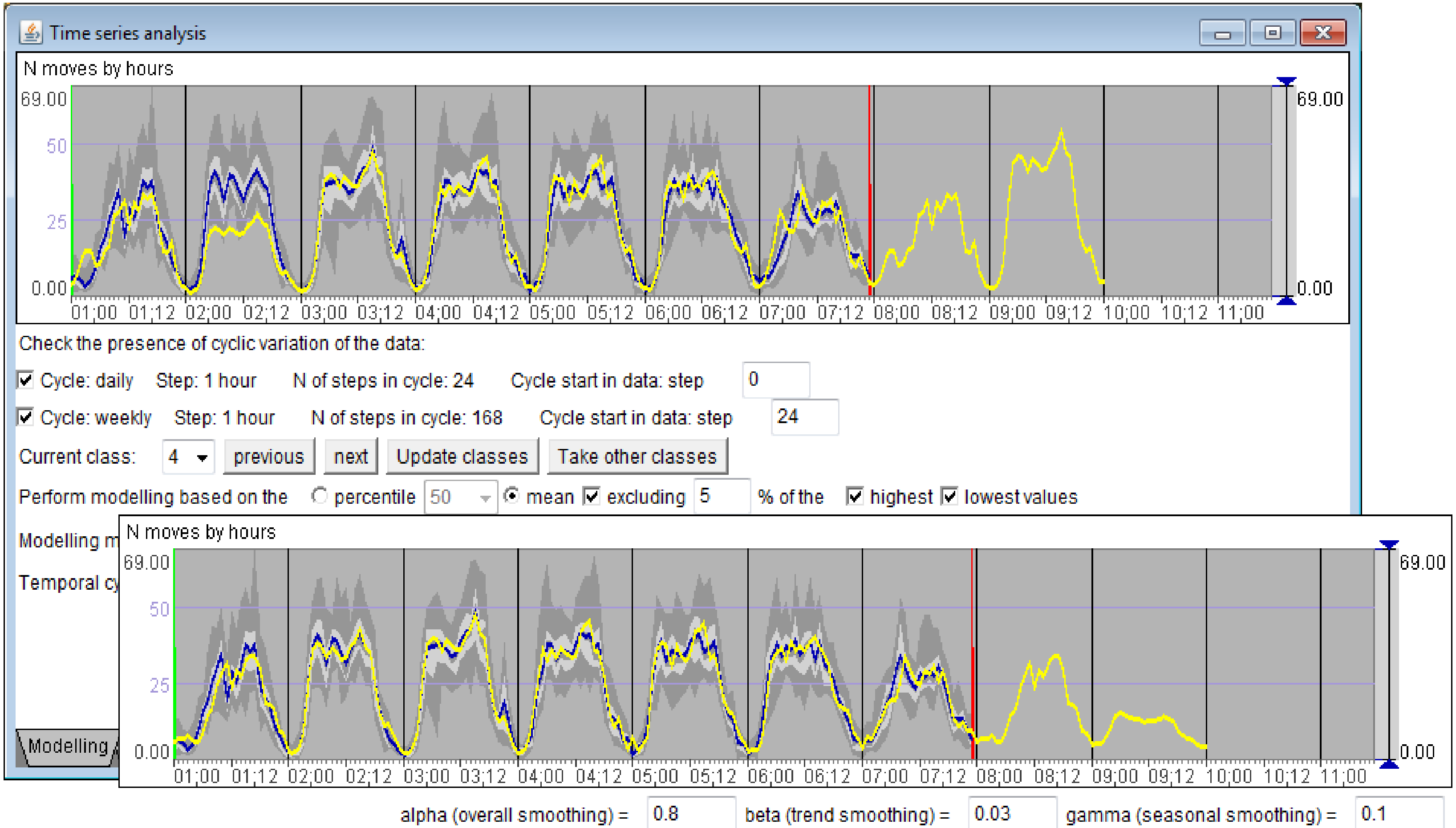
Step 2: Analysis and modelling



CITY UNIVERSITY LONDON

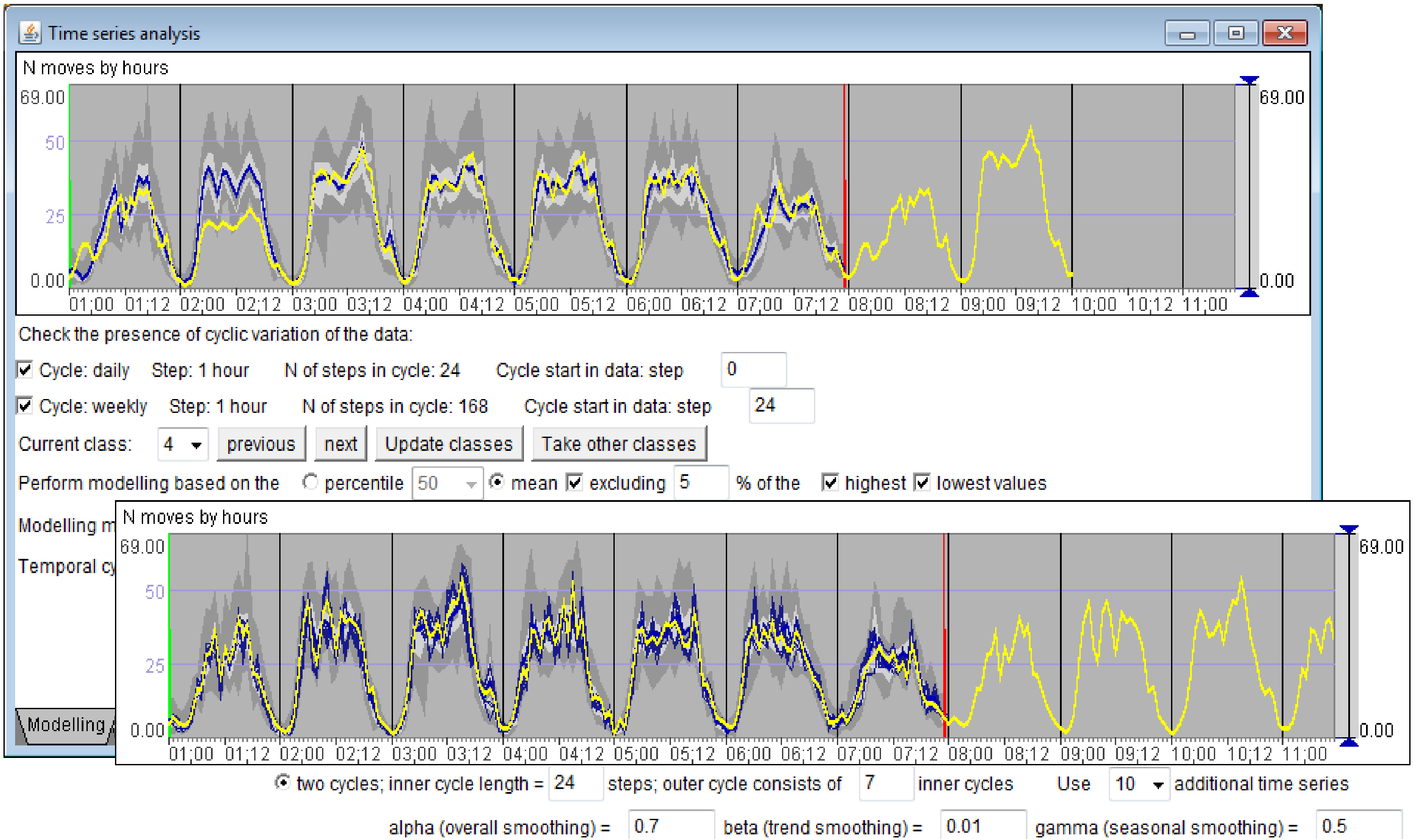
Fraunhofer IAIS

Step 2: Analysis and modelling



CITY UNIVERSITY LONDON Fraunhofer IAIS

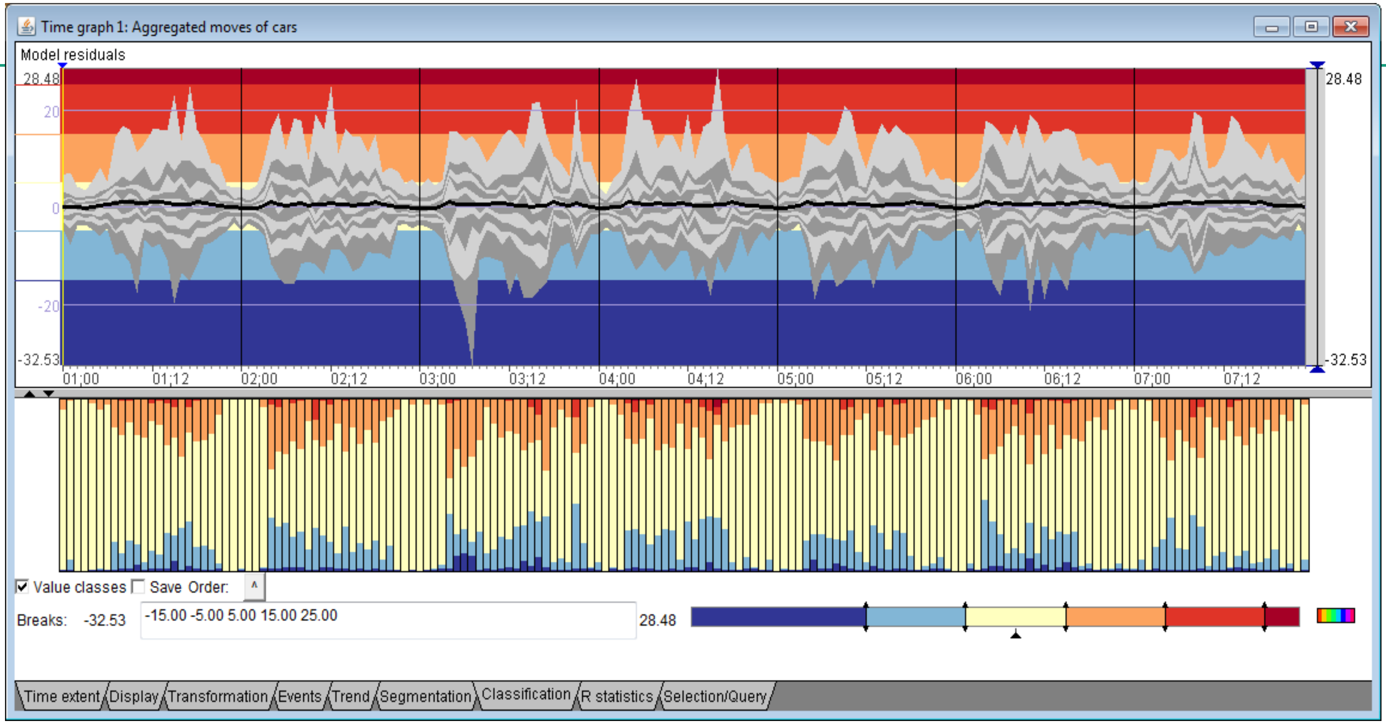
Step 2: Analysis and modelling



Step 3: Model evaluation (analysis of residuals)

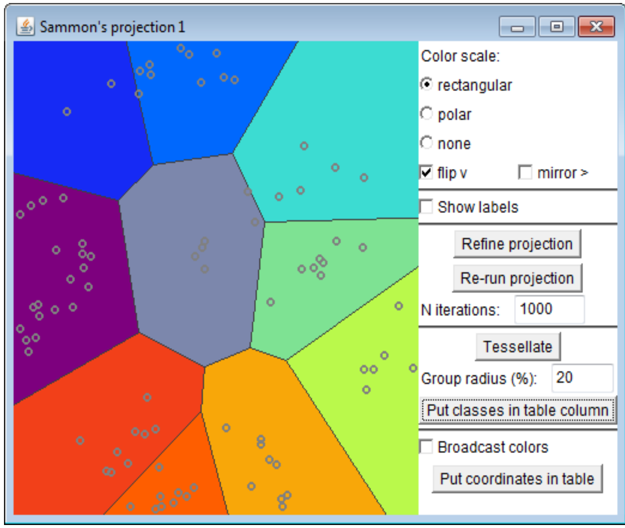
- The goal is not to minimise the residuals
  - The model should not reproduce all fluctuations and outliers present in the data
  - This should be an abstraction capturing the characteristic features of the temporal variation
  - High values of the residuals do not mean low model quality
- The goal is to have the residuals randomly distributed in space and time (no detectable patterns)
  - This means that the model correctly captures the characteristic, non-random features of the temporal variation

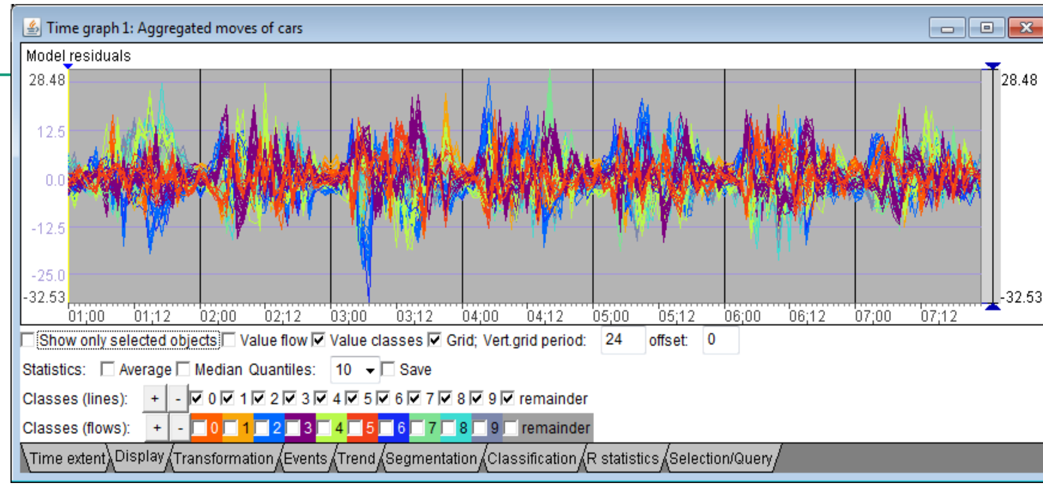
Analysis of residuals (example)

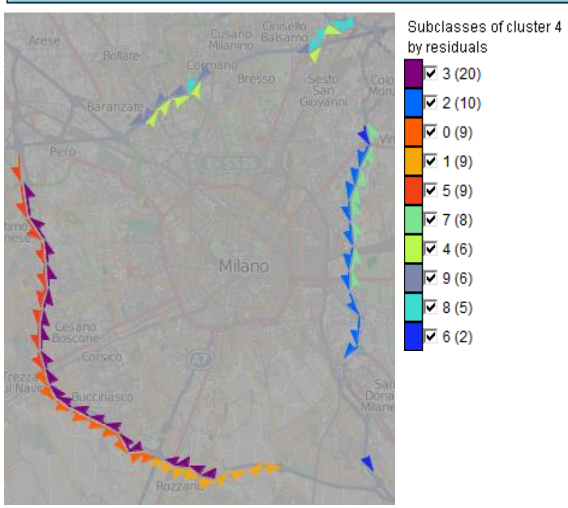


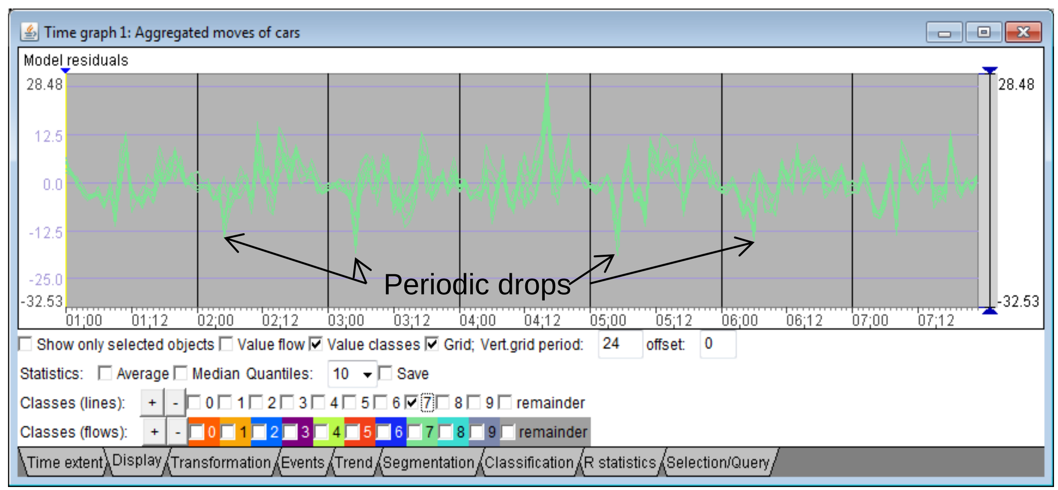
- No systematic bias: approximately equal numbers of positive and negative errors in each time step
- No periodic increases and decreases at the level of the whole group
- However, we are not sure about individual objects

More detailed analysis by subgroups









It may be reasonable to consider this subgroup separately -> back to re-grouping

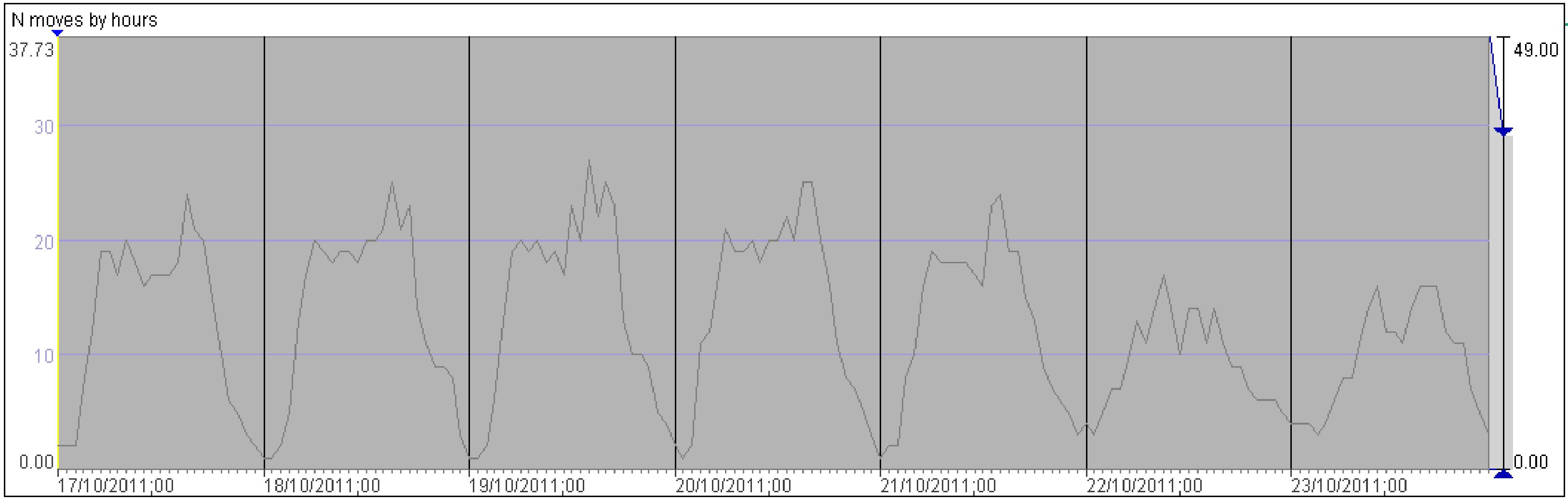
Use of a model for prediction

- We obtain a common model for a group (cluster) of time series
  - Predicts the same values for all objects/places of the group
  - The statistical properties of the distribution of the predicted values in each place differ from the distribution of the original values
- Adjustment of the prediction for individual objects/places:
  - Compute and store the basic statistics (quartiles) of the original values for each object/place i: $Q1_i$, M_i, $Q3_i$
  - Compute the statistics of the model-predicted values for the same time steps as the original values: Q1, M, Q3 (common for the cluster)
  - Shift *(level adjustment)*: $S = M_i - M$
  - Scale factors *(amplitude adjustment)*: $F_{low} = \dfrac{M_i - Q1_i}{M - Q1}$ $\quad F_{high} = \dfrac{Q3_i - M_i}{Q3 - M}$
  - Let v^t be the model-predicted value for an arbitrary time step t and v^t_i the individually adjusted value for the place/object i

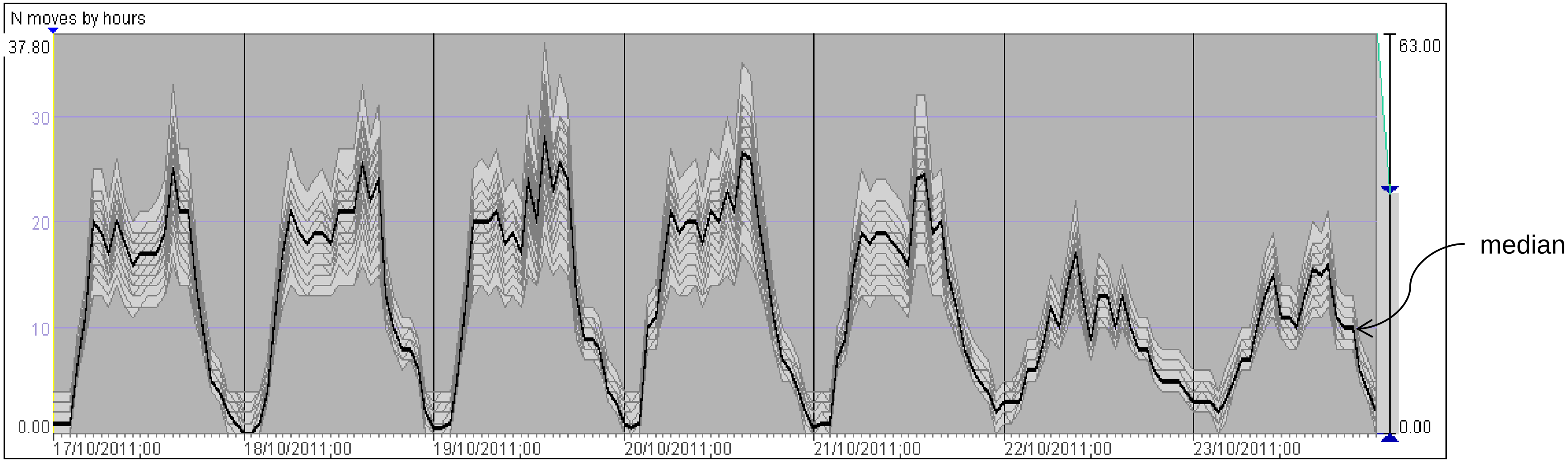
$$v^t_i = \begin{cases} M + F_{low} \cdot (v^t - M) + S, \text{ if } v^t < M \\ M + F_{high} \cdot (v^t - M) + S, \text{ otherwise} \end{cases}$$

Use of a model for prediction: example

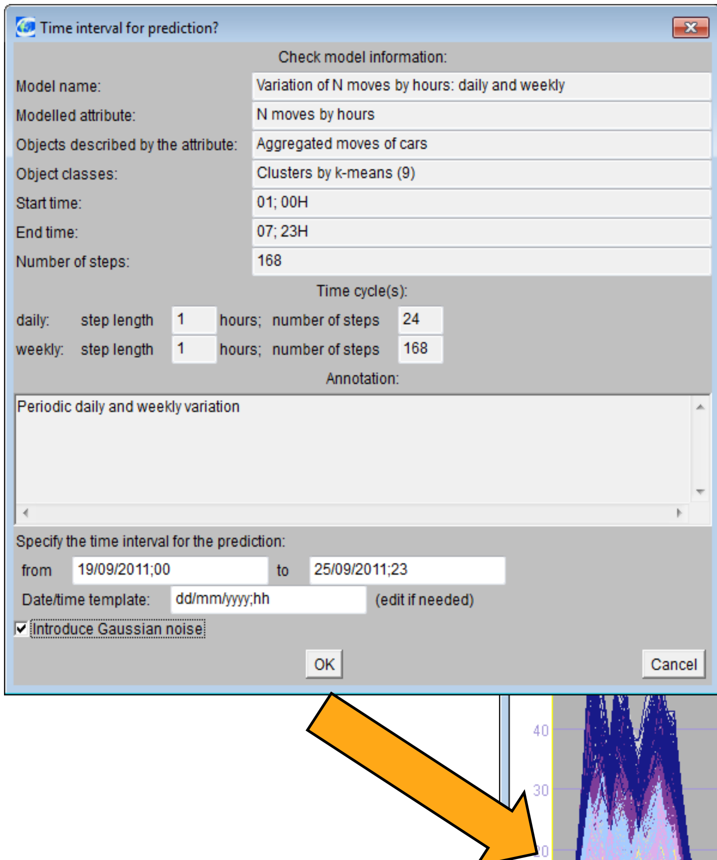
Common prediction for a cluster:

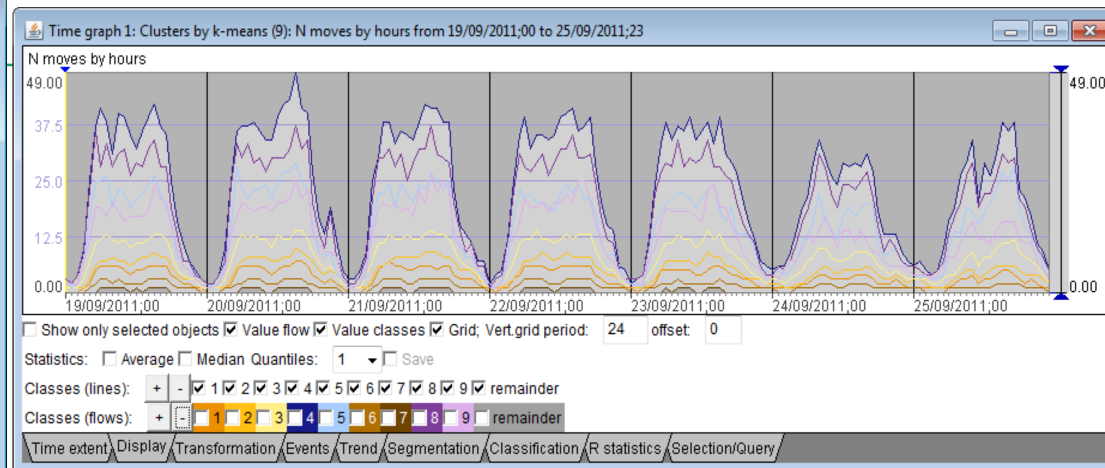


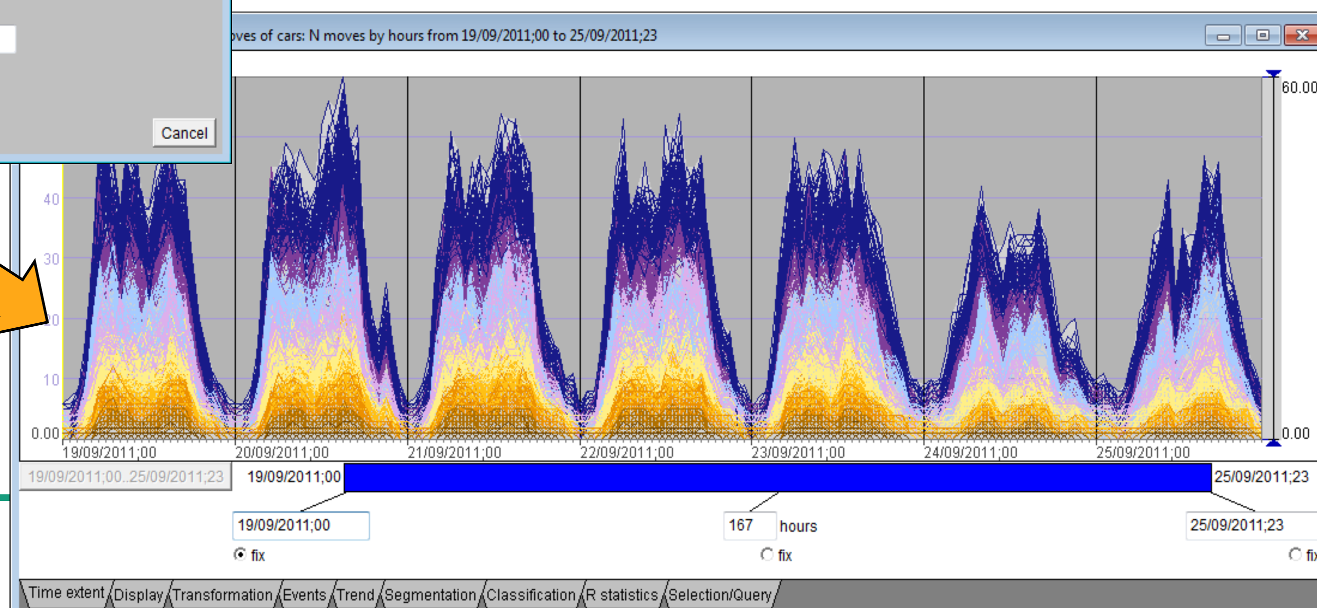
Set of individually adjusted predictions for this cluster:



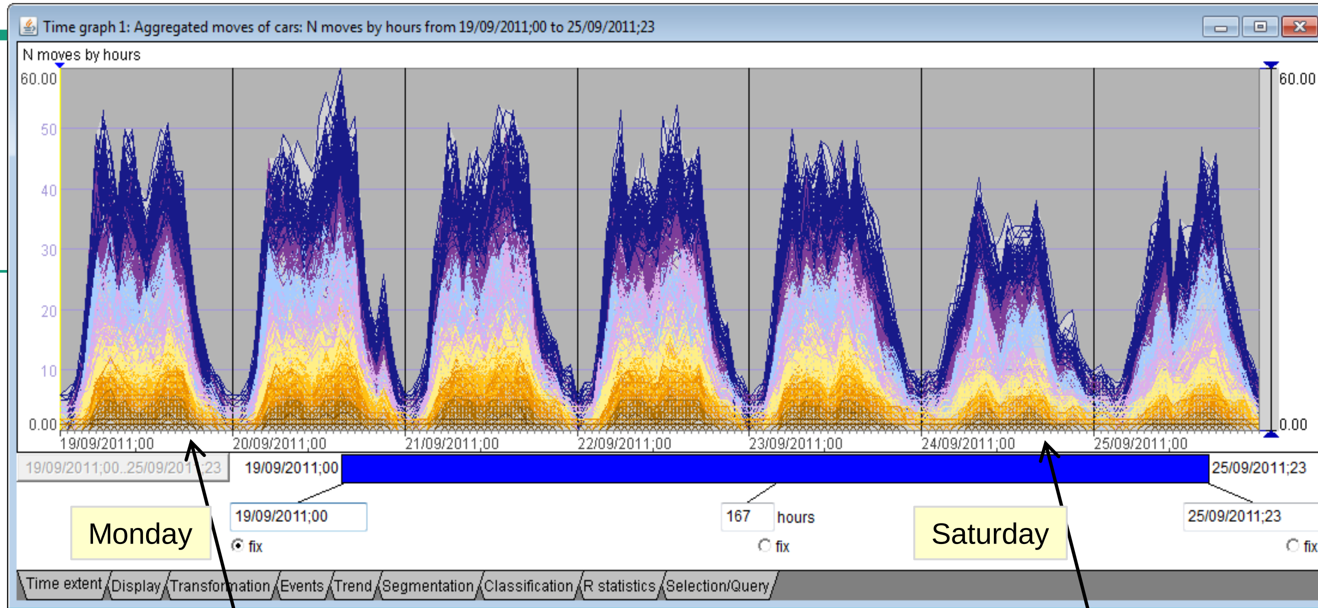
Prediction based on the models



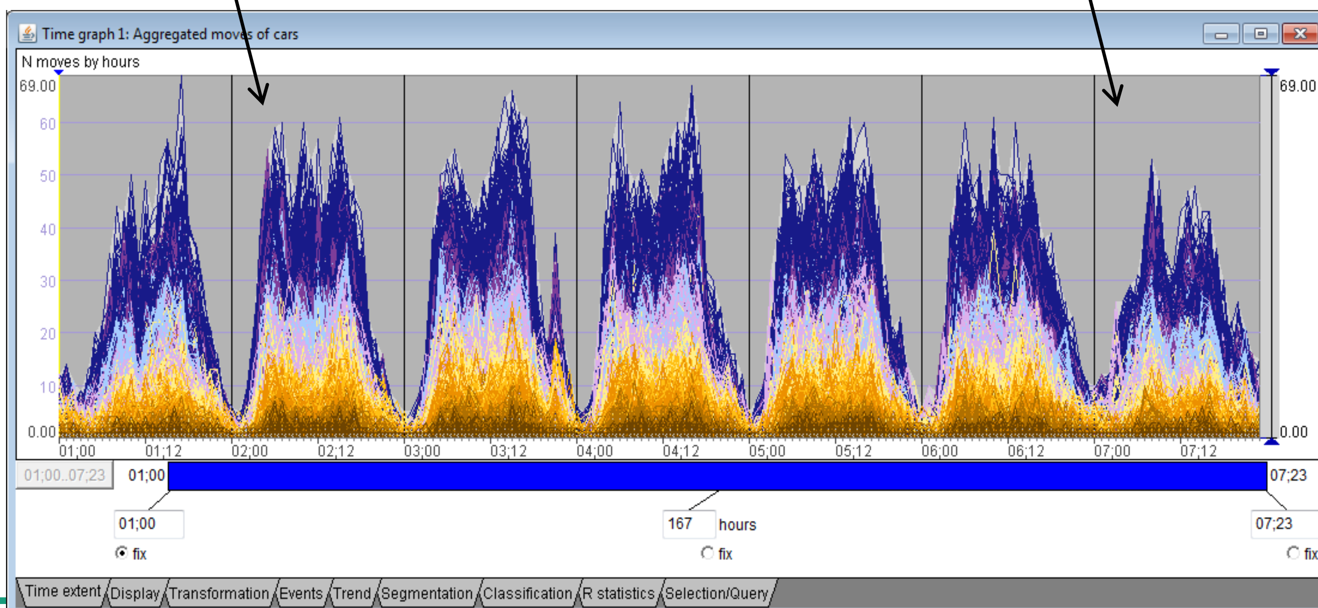




Predicted:

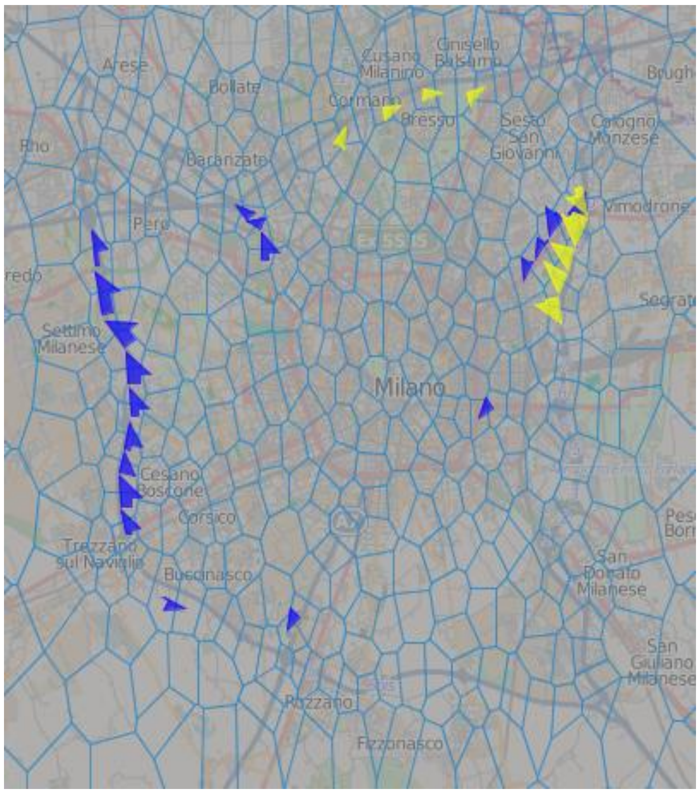


Original:

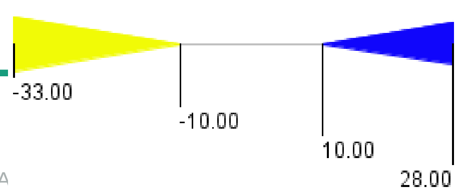


Comparison of actual values with predicted (e.g., in monitoring)

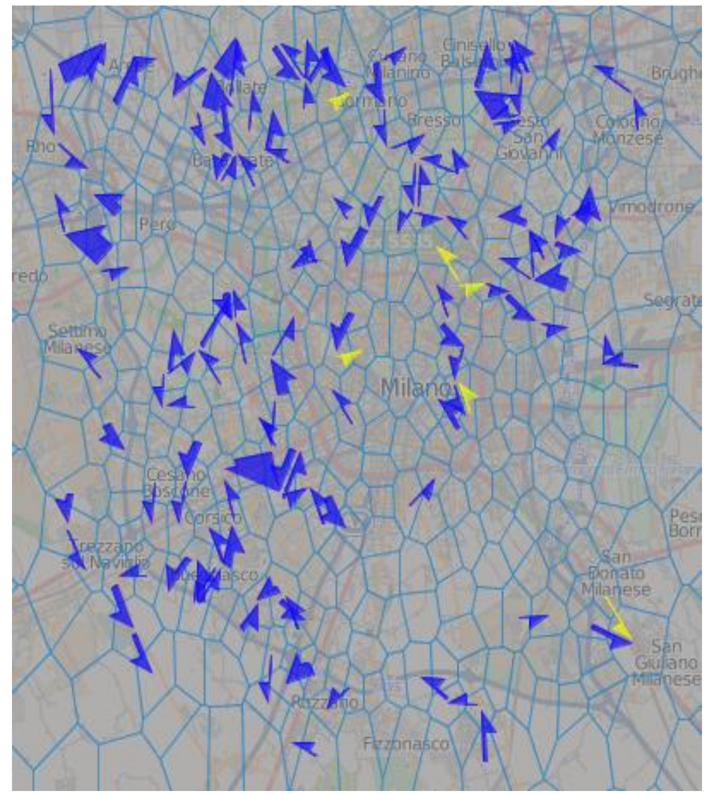
Absolute differences



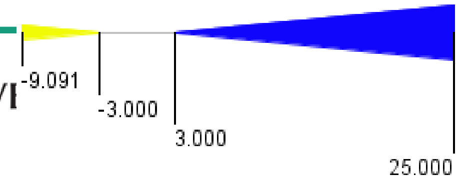
Actual N moves by hours - predicted



Normalized differences

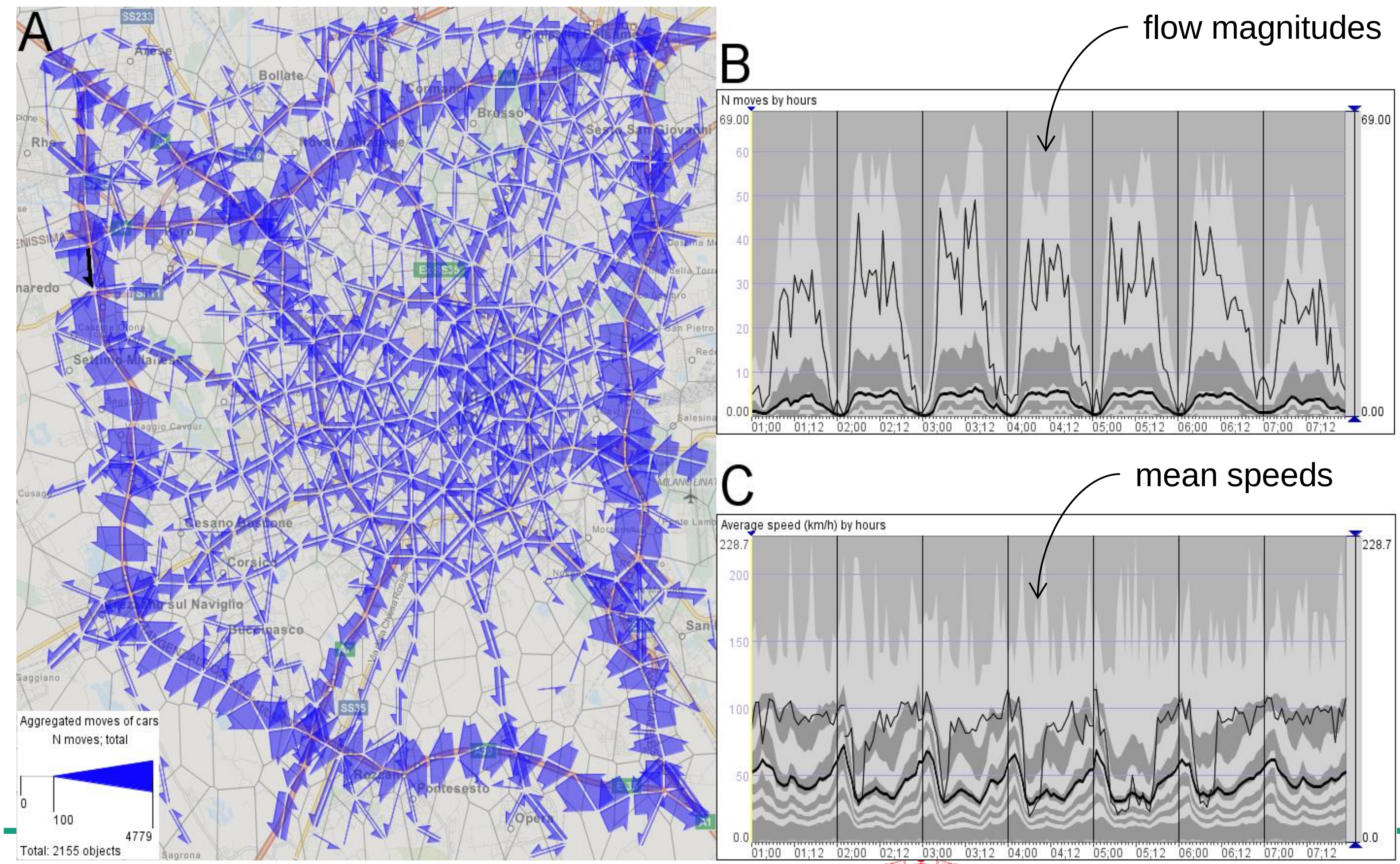


Actual N moves by hours - predicted divided by variance of predicted



CITY UNIVI LONDON

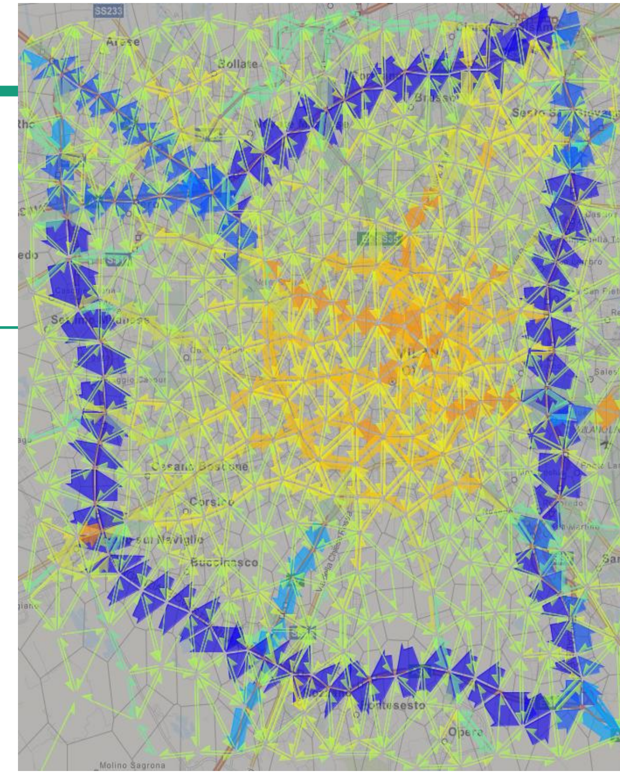
Analysis and modelling of relationships between two time-variant attributes

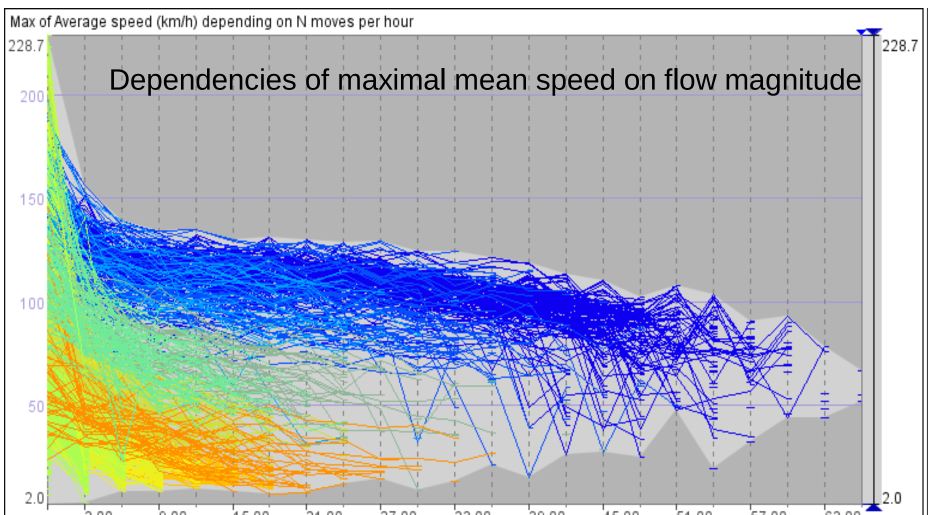


CITY UNIVERSITY LONDON Fraunhofer IAIS

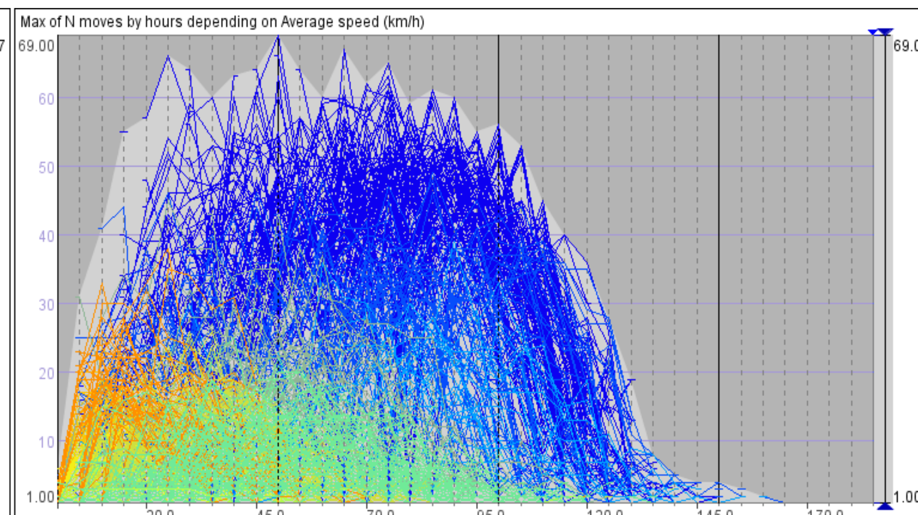
Data transformation and clustering

- Dependency of attribute A(t) on attribute B(t):
  - Divide the value range of B into intervals
  - For each interval, collect all values of A that co-occur with the values of B from this interval
  - Compute statistics of the values of A: minimum, maximum, median, mean, percentiles …
  - For each of these, there is a series B → A, or A(B)



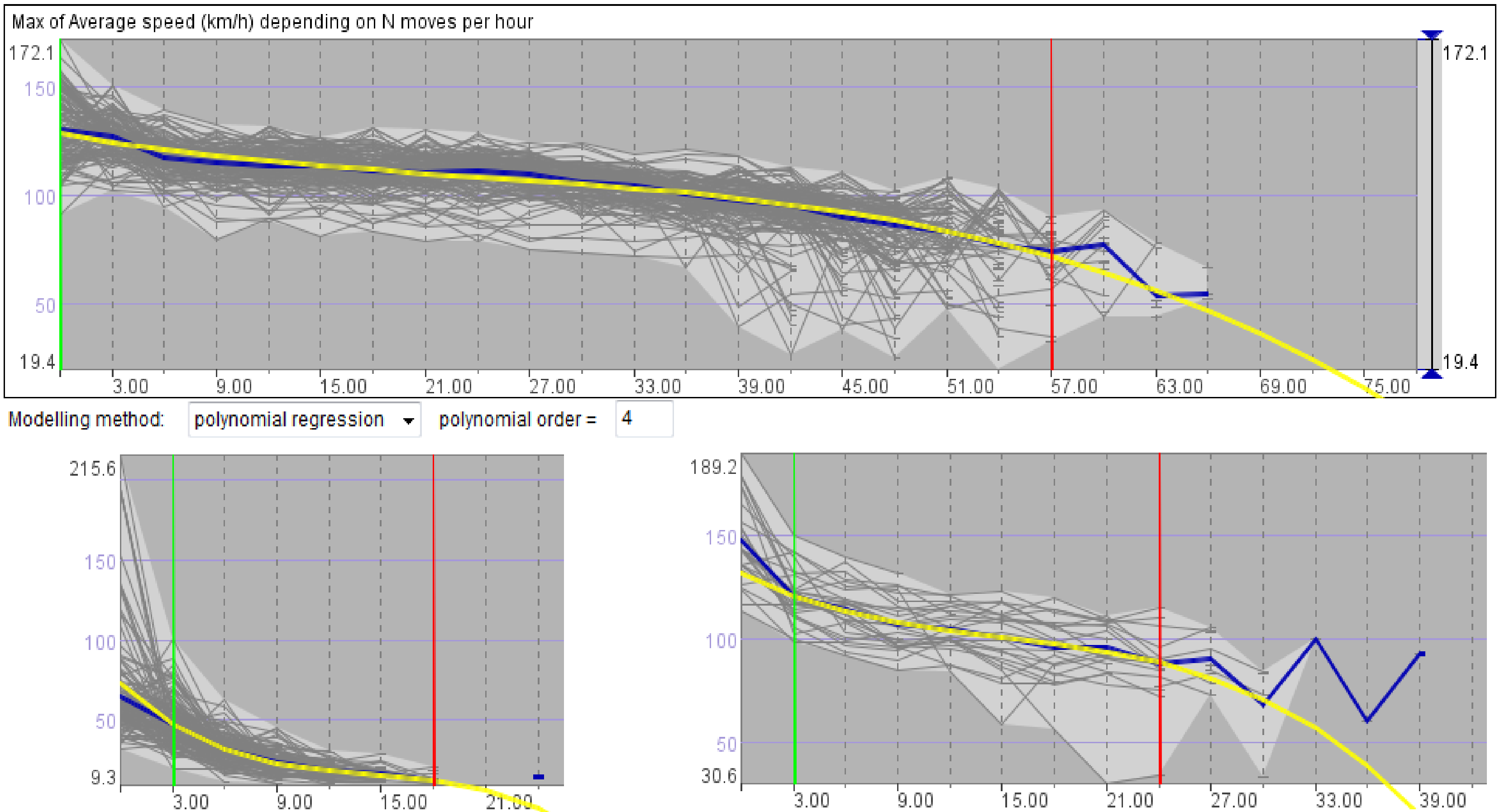


Dependencies of maximal mean speed on flow magnitude



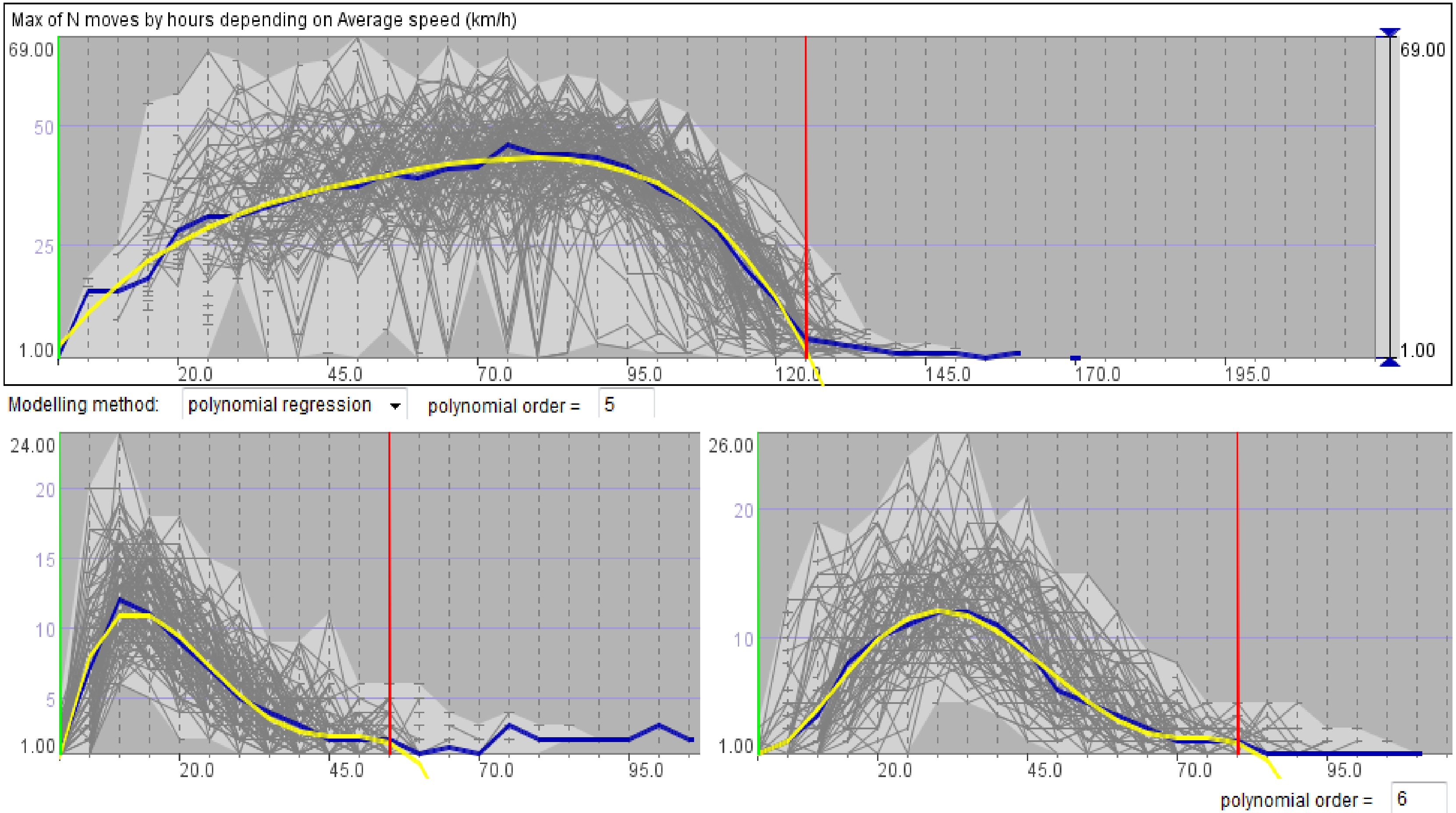
Dependencies of maximal flow magnitude on mean speed

Dependency modelling: flow → maximal mean speed



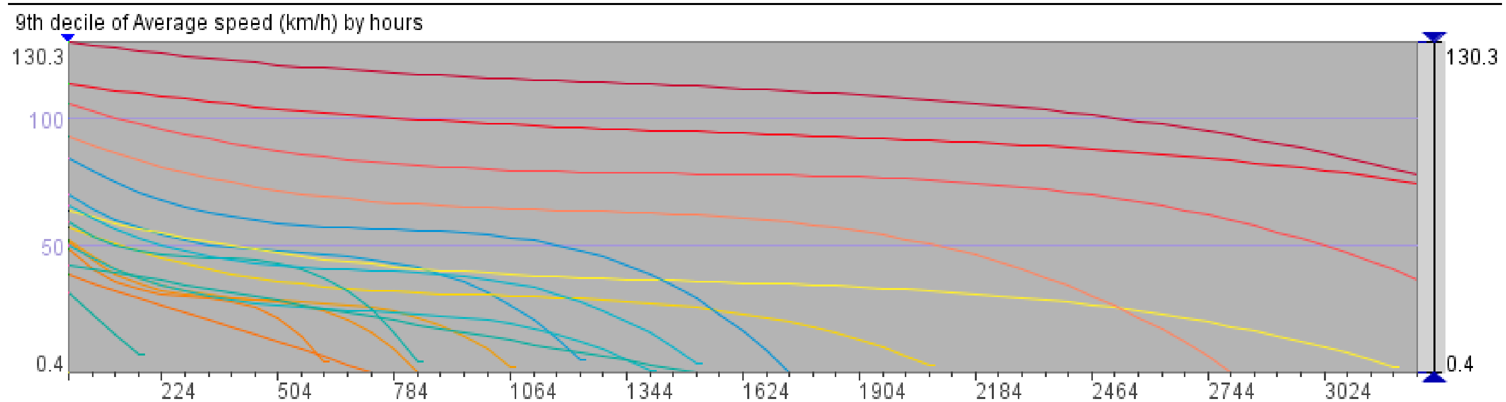
CITY UNIVERSITY LONDON
Fraunhofer IAIS

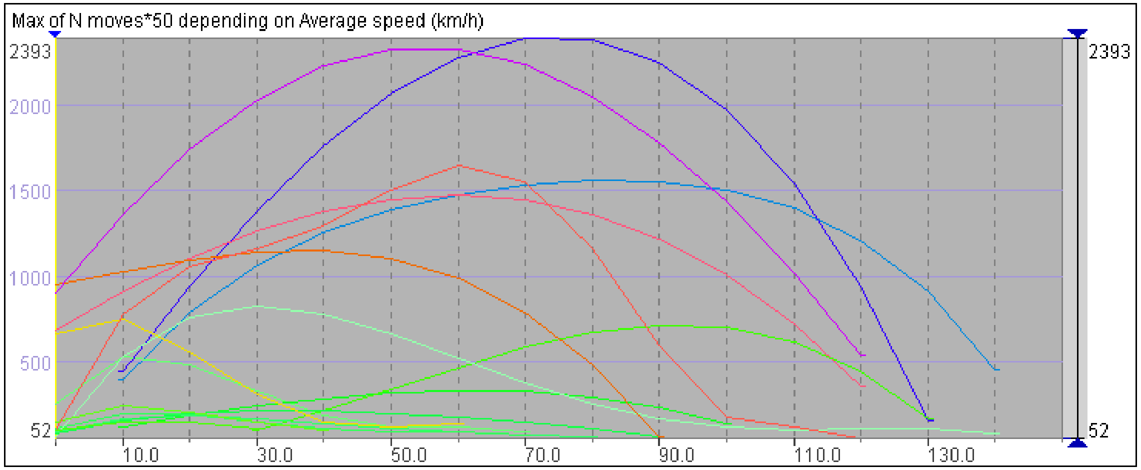
Dependency modelling: mean speed → maximal flow



CITY UNIVERSITY LONDON
Fraunhofer IAIS

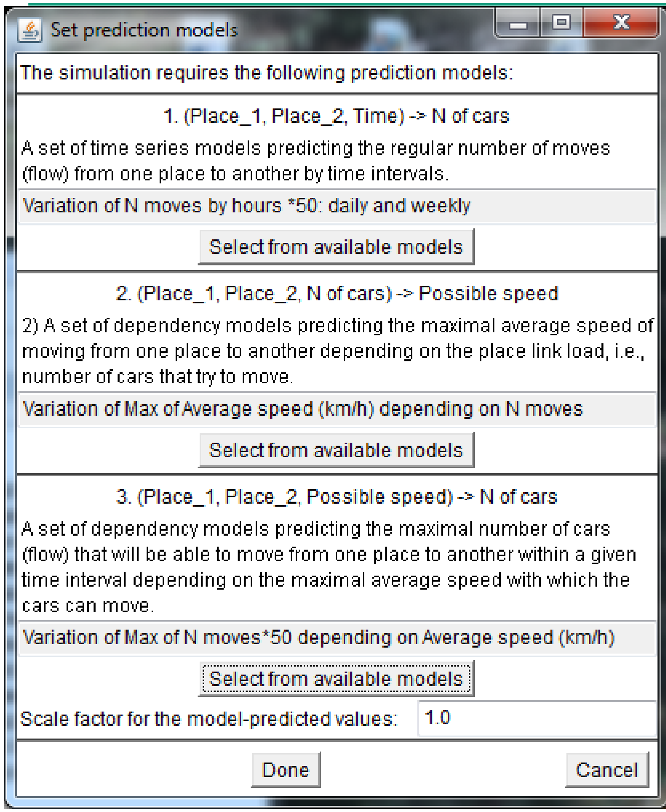
Graphical representation of the models built

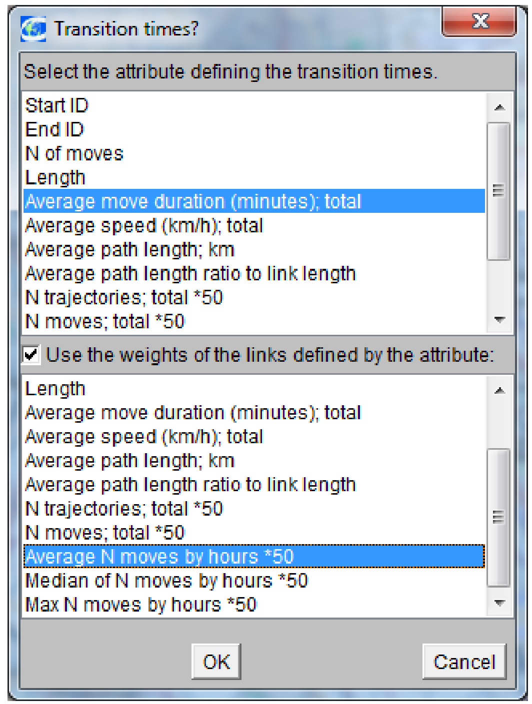


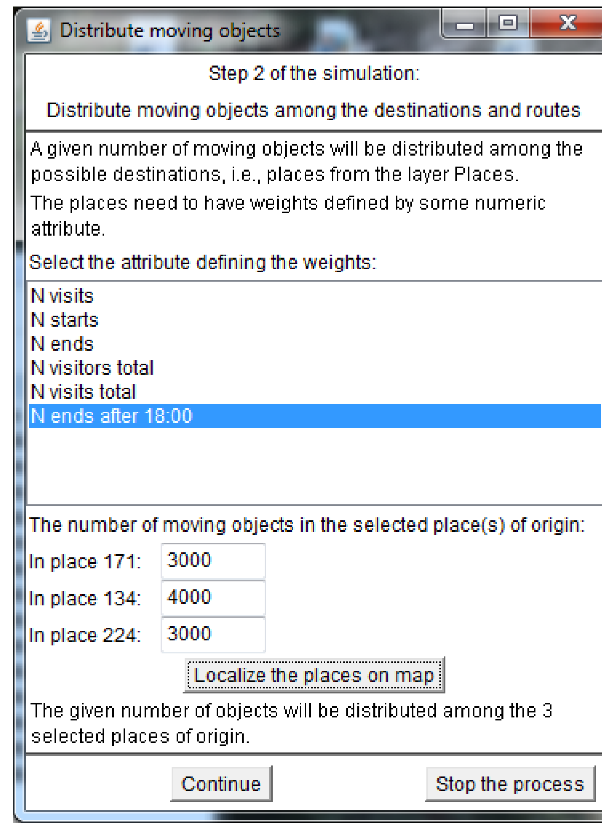


Use of the models:

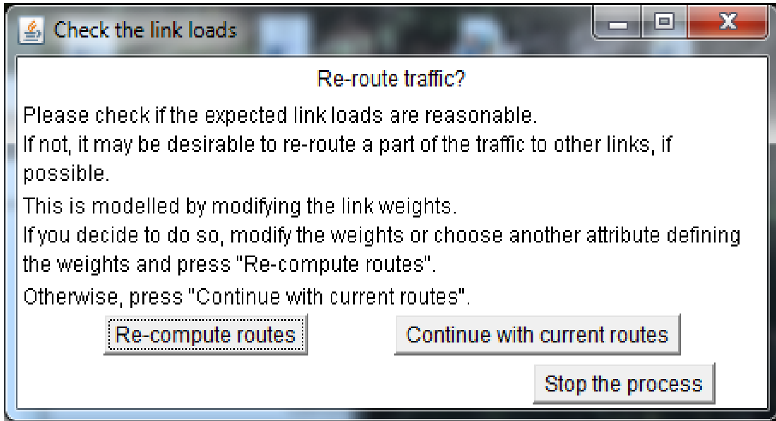
simulation of extraordinary traffic from given places

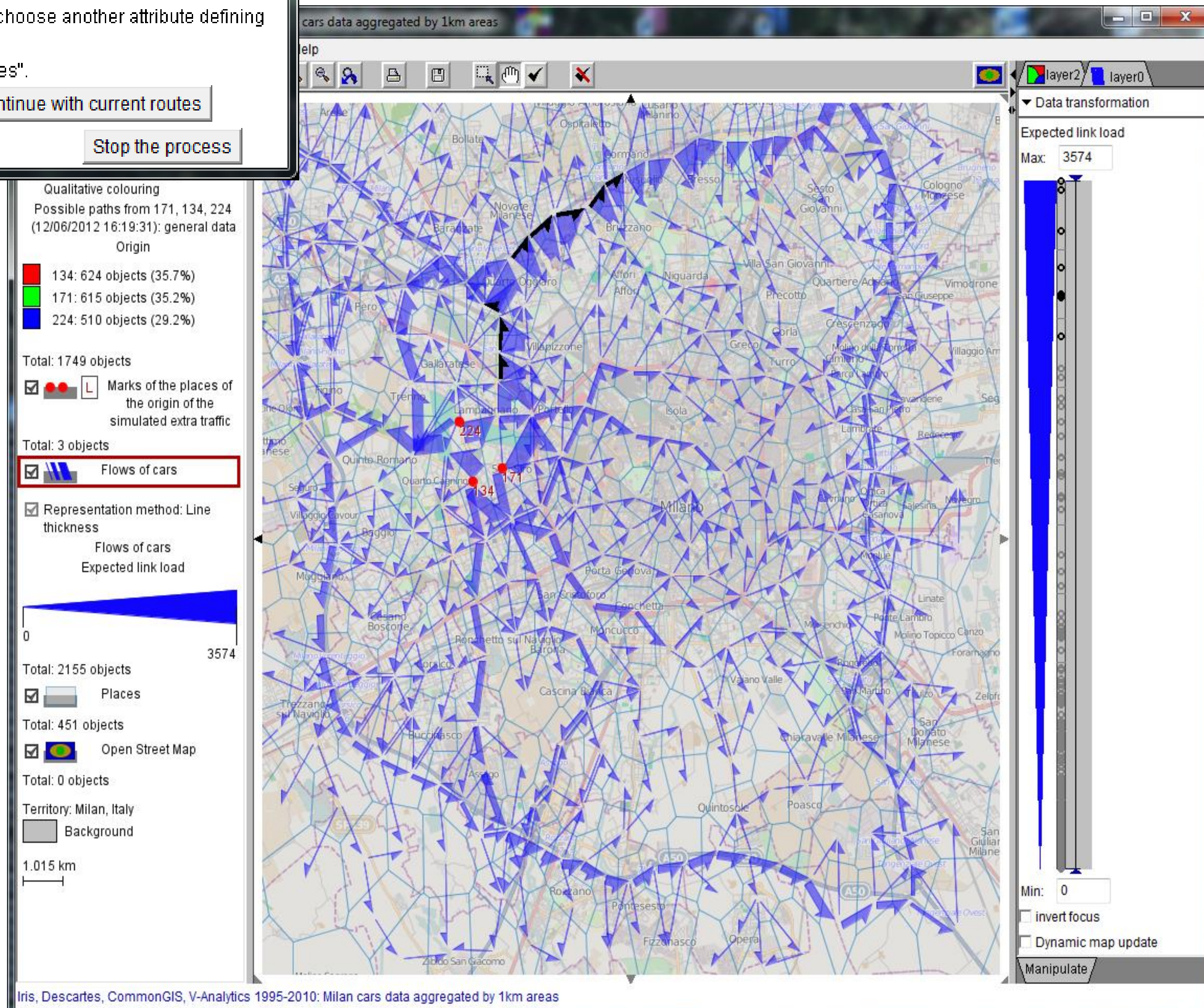




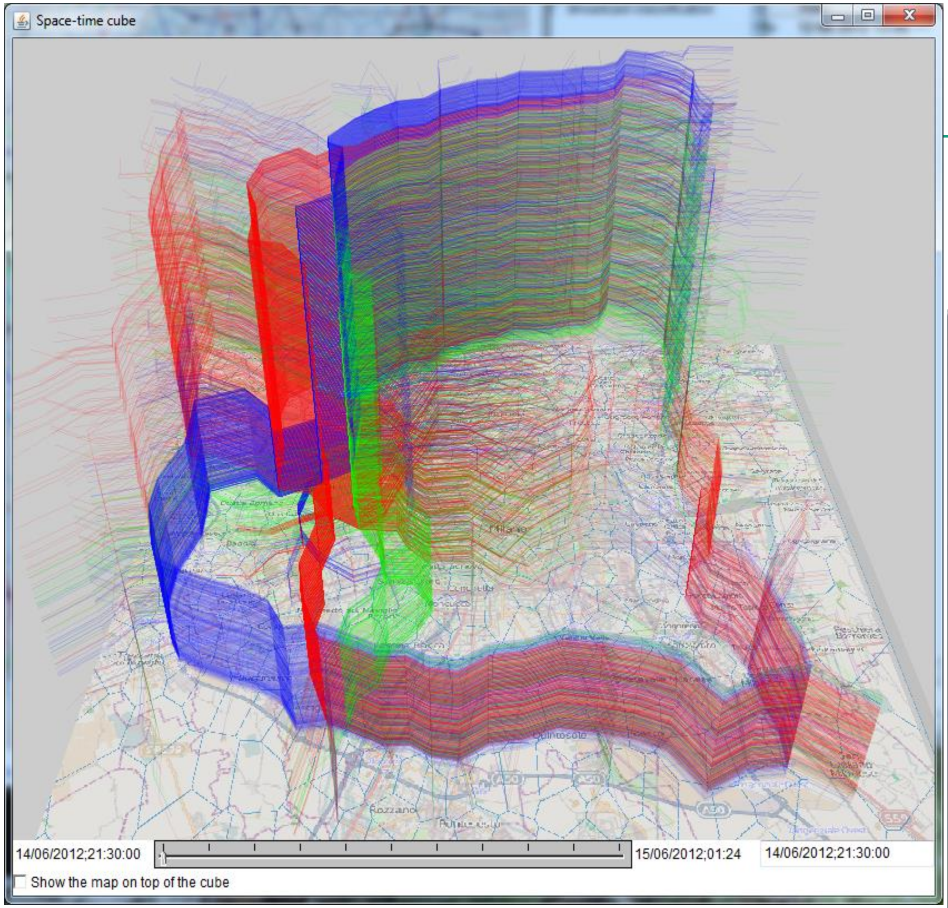


- The bottlenecks can be revealed even before the simulation

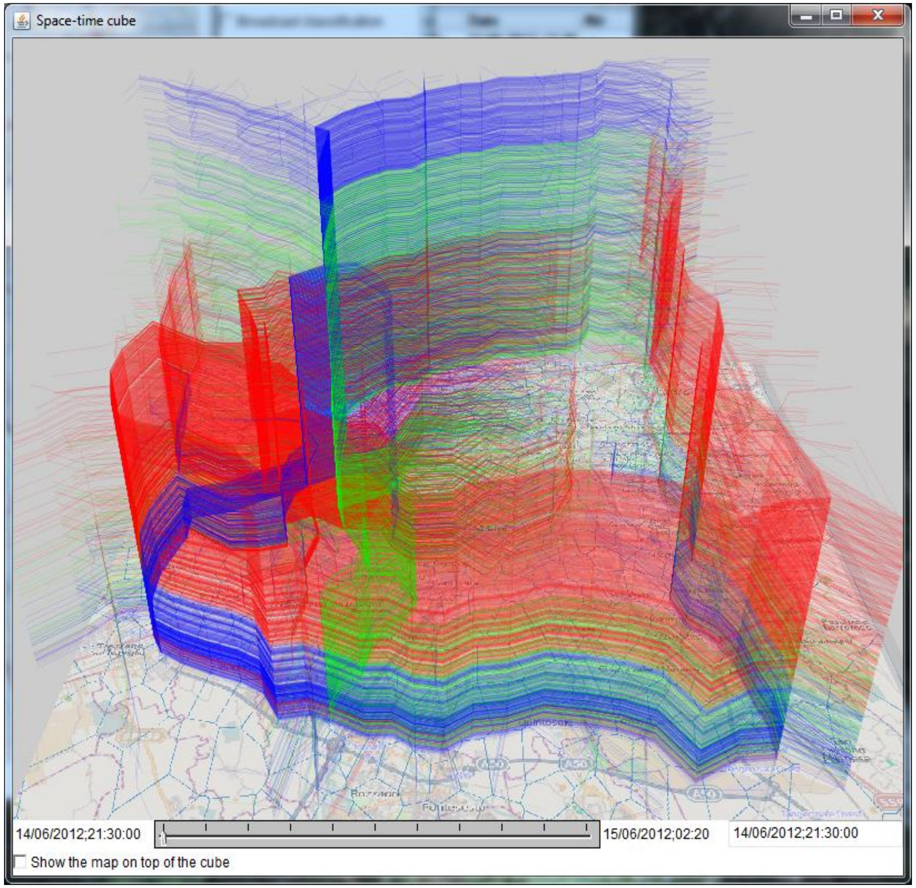




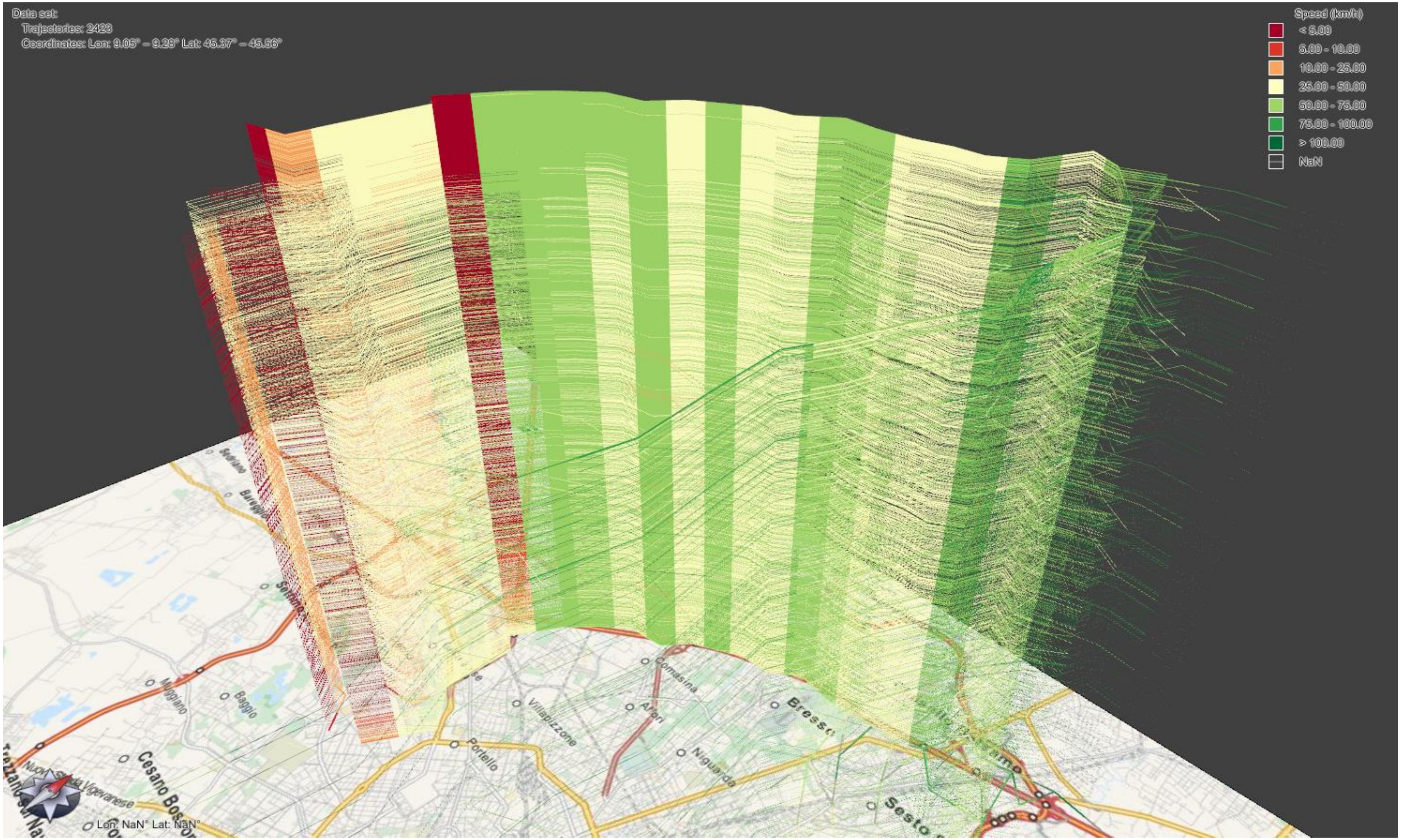
Simulated trajectories



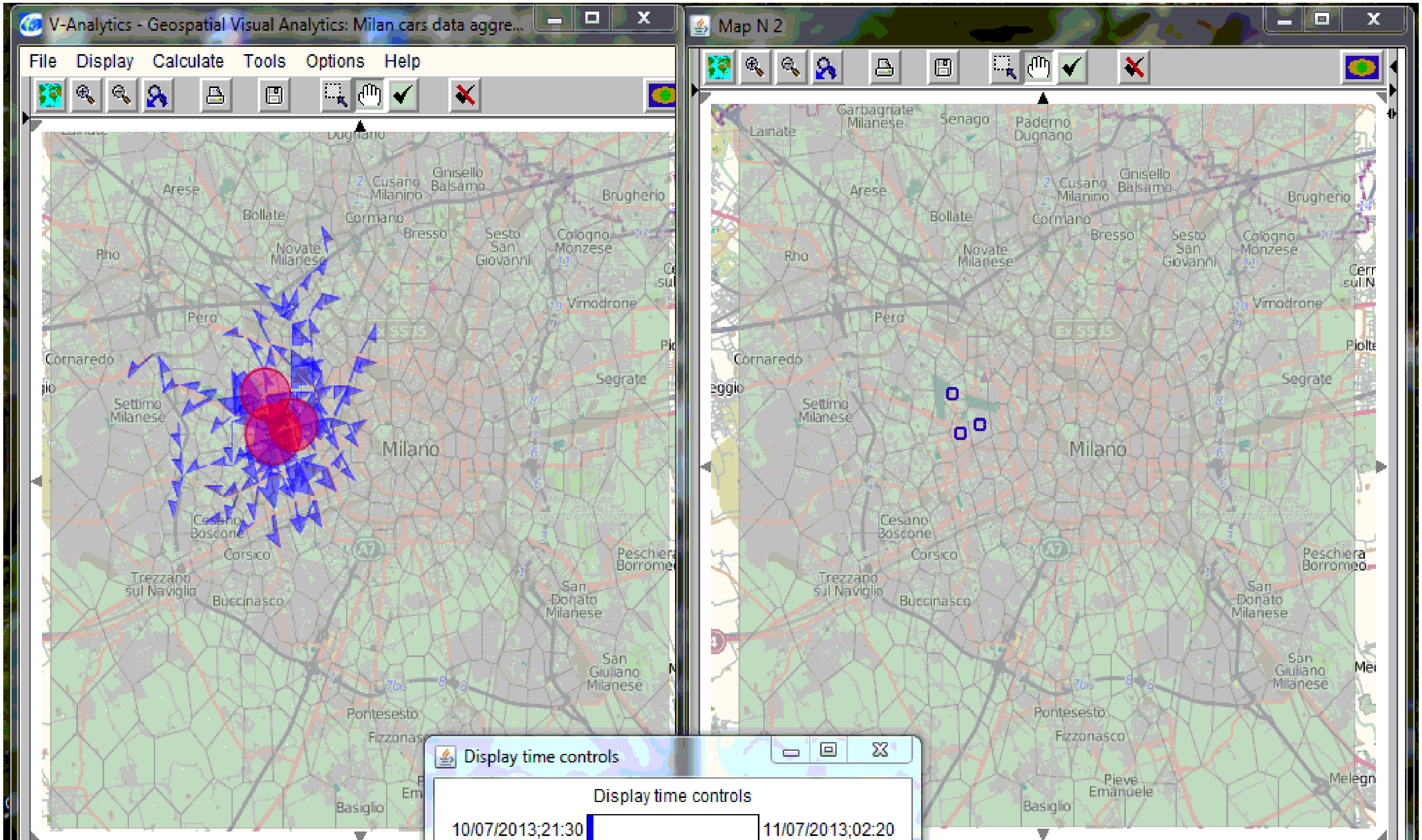
Some traffic re-routed to the south:



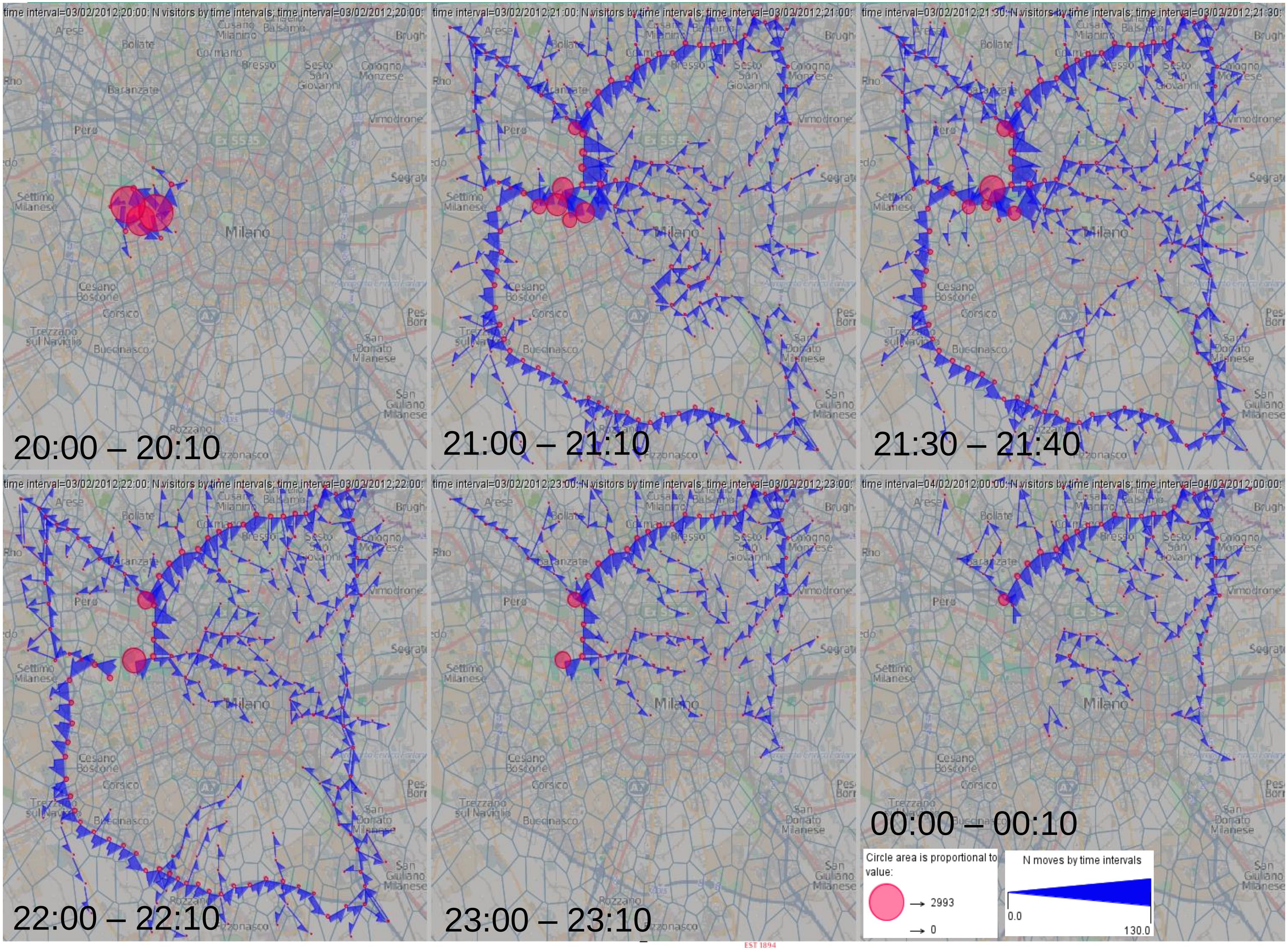
The speeds on the northern motorway



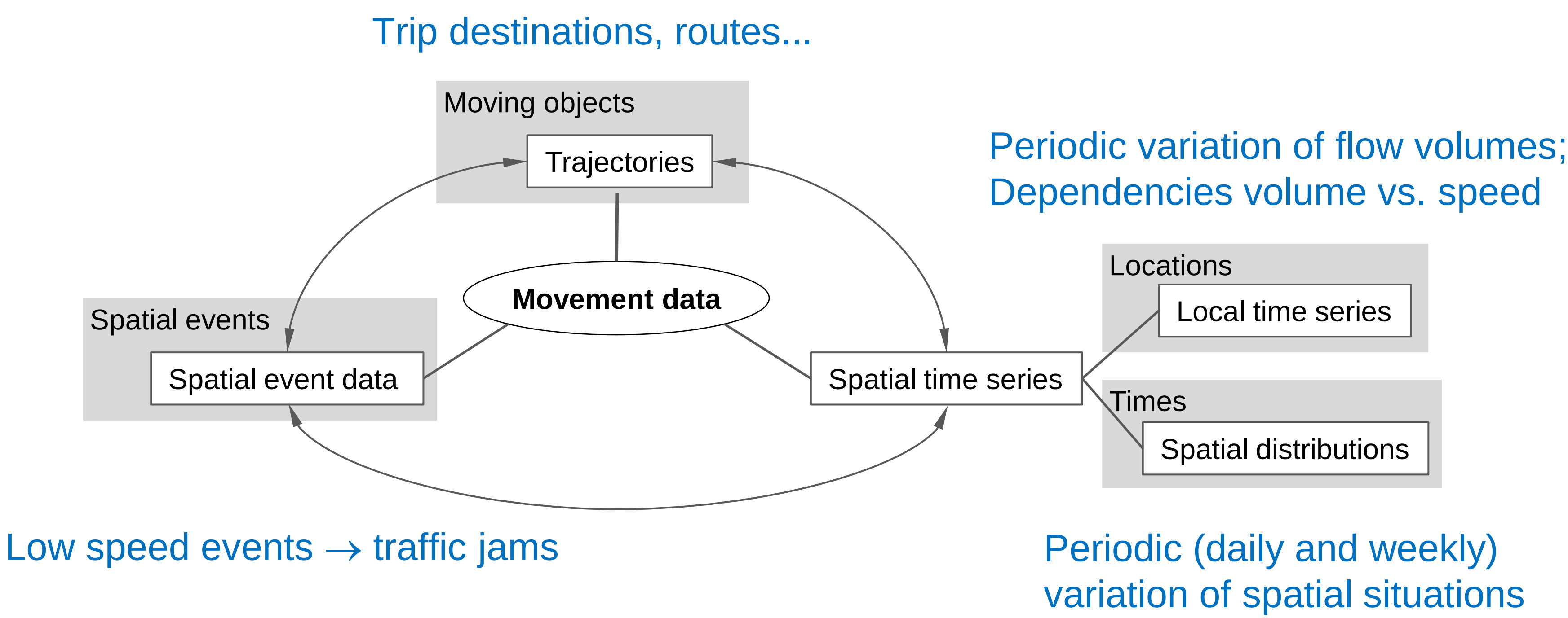
Animation of simulation results



Presence and flows for selected time intervals



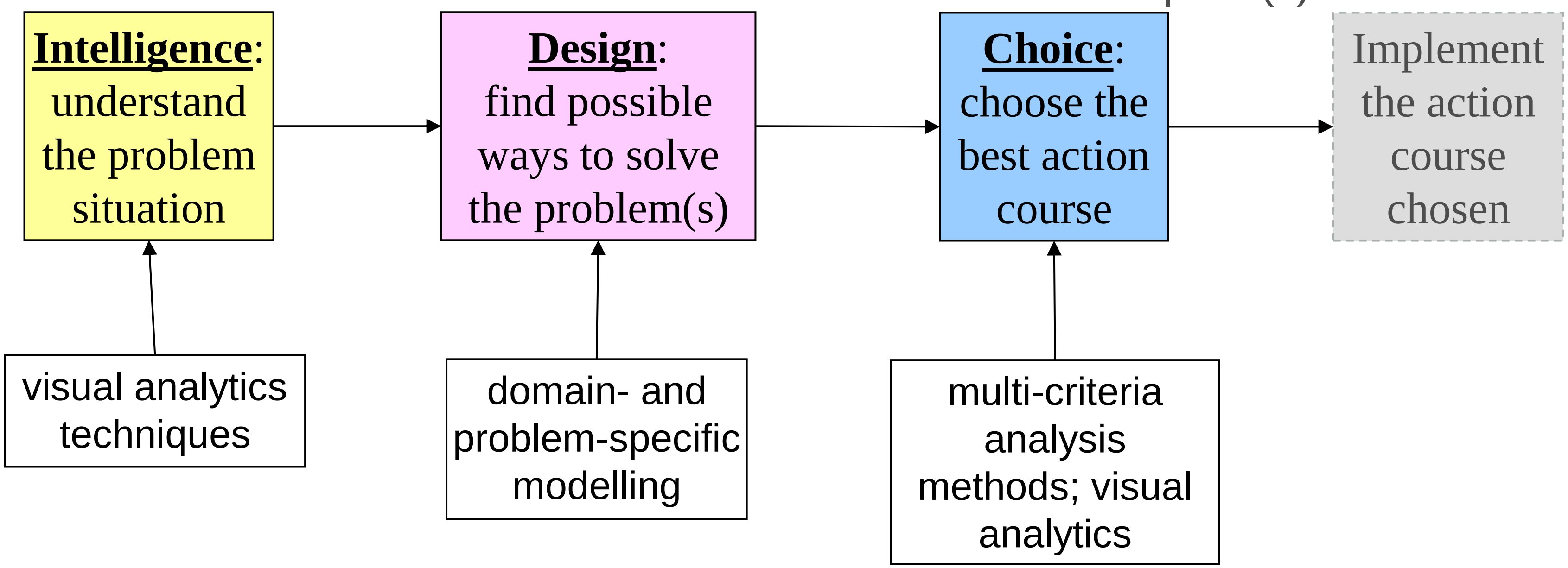
Multi-perspective analysis of movement



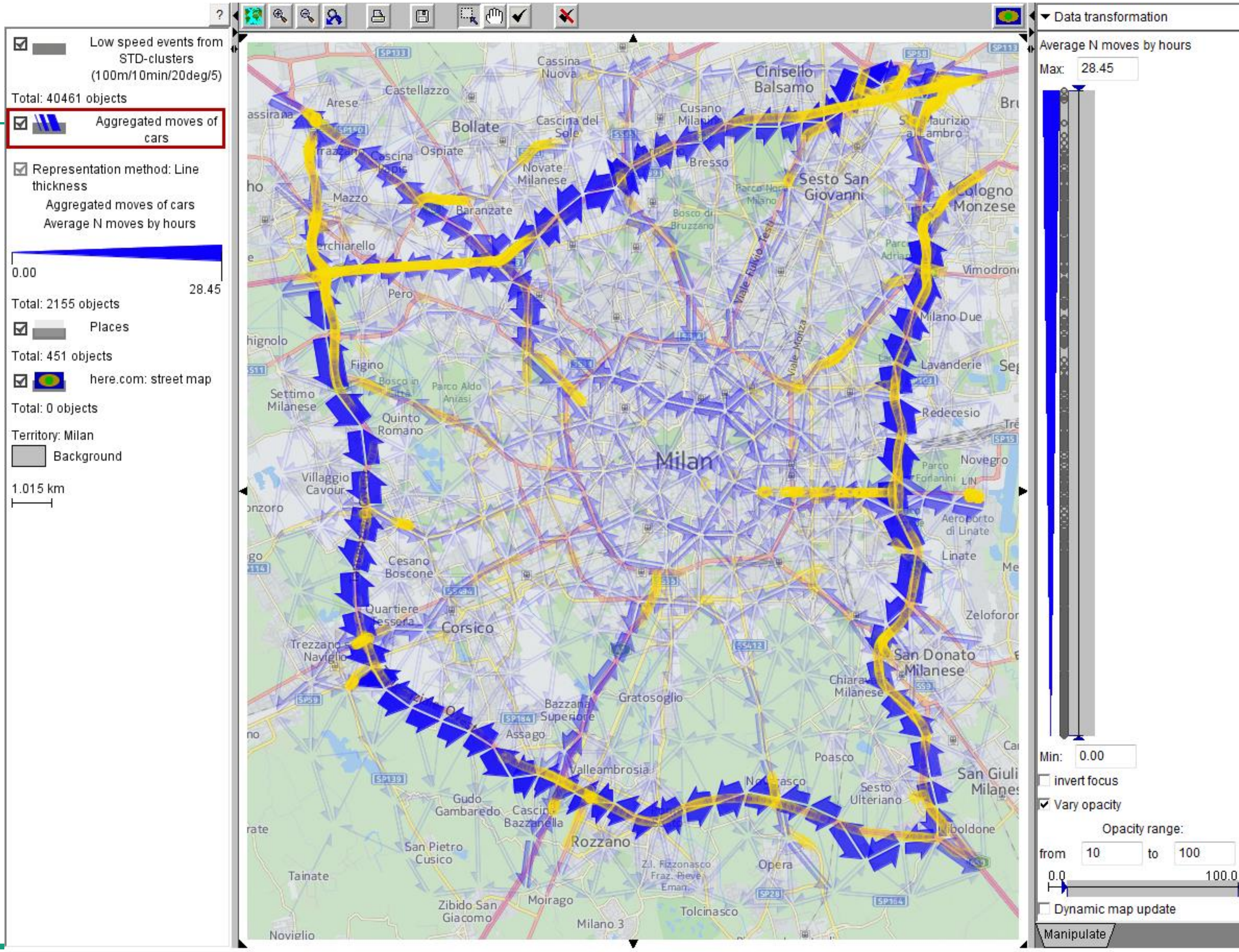
Bonus track: spatial decision support

collect and integrate data; explore the data, identify problems and opportunities

analyse and evaluate the options; select the most suitable option(s)

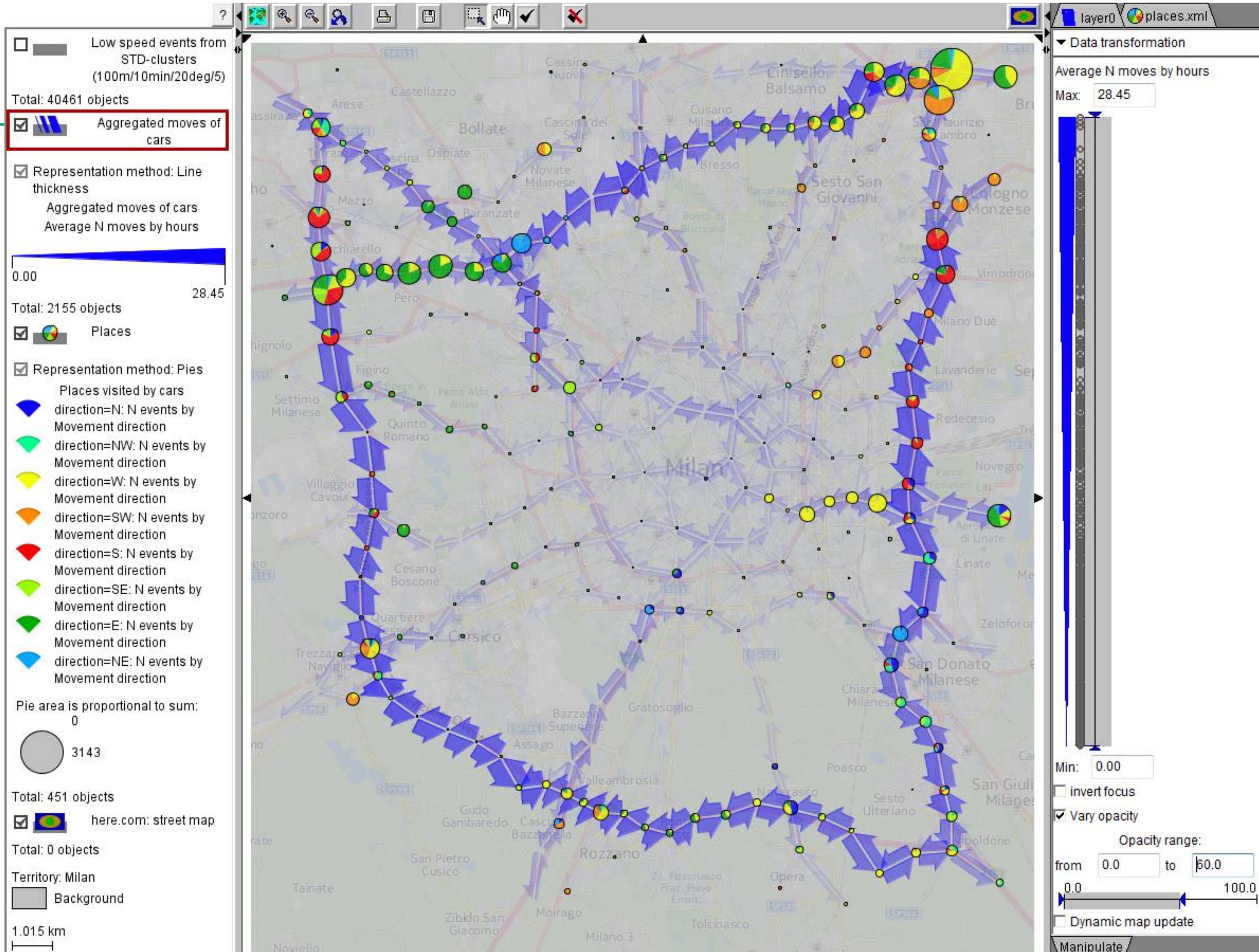


Average hourly traffic flows and traffic jam events



Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

Traffic jam events summarized by areas and directions

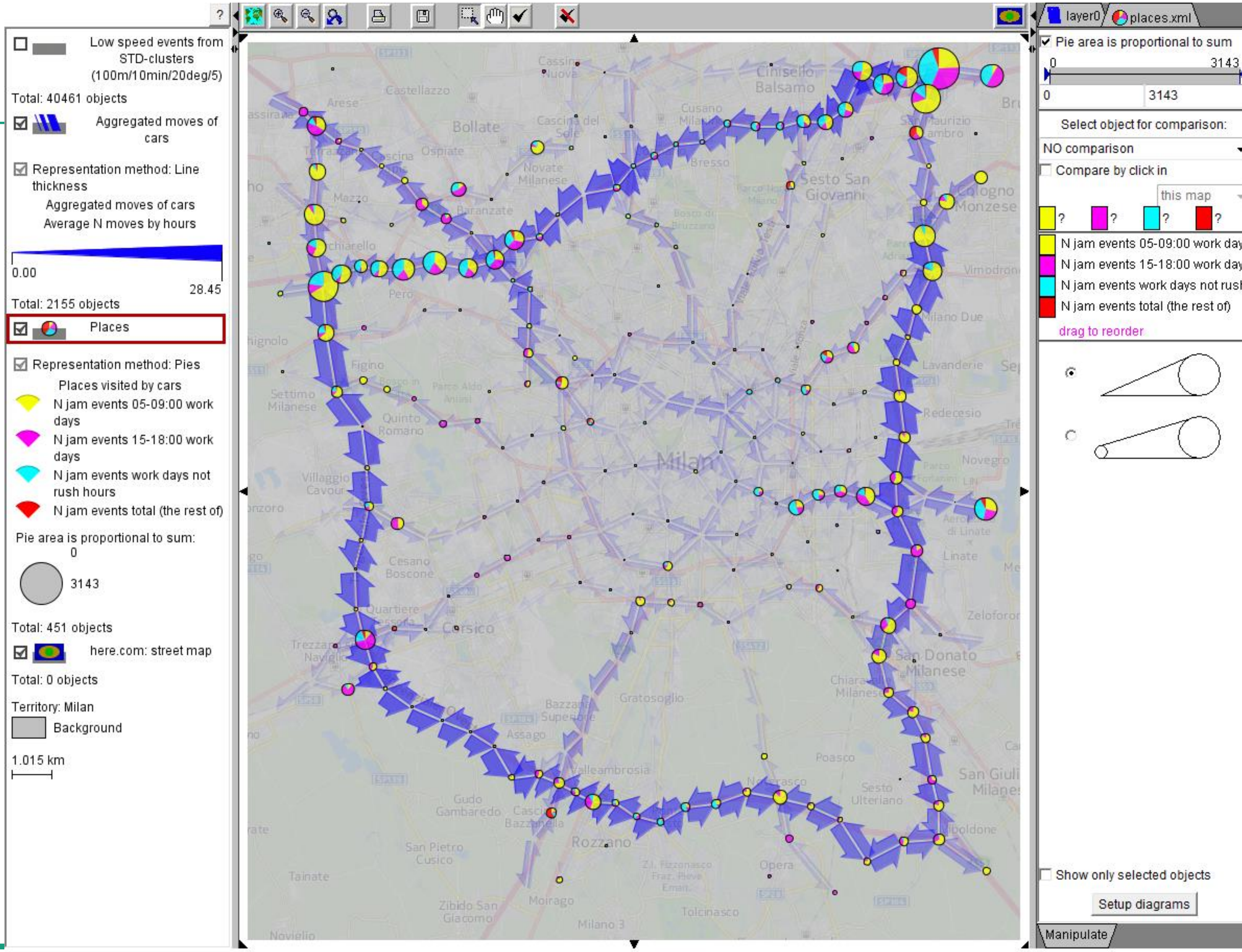


Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

CITY UNIVERSITY LONDON

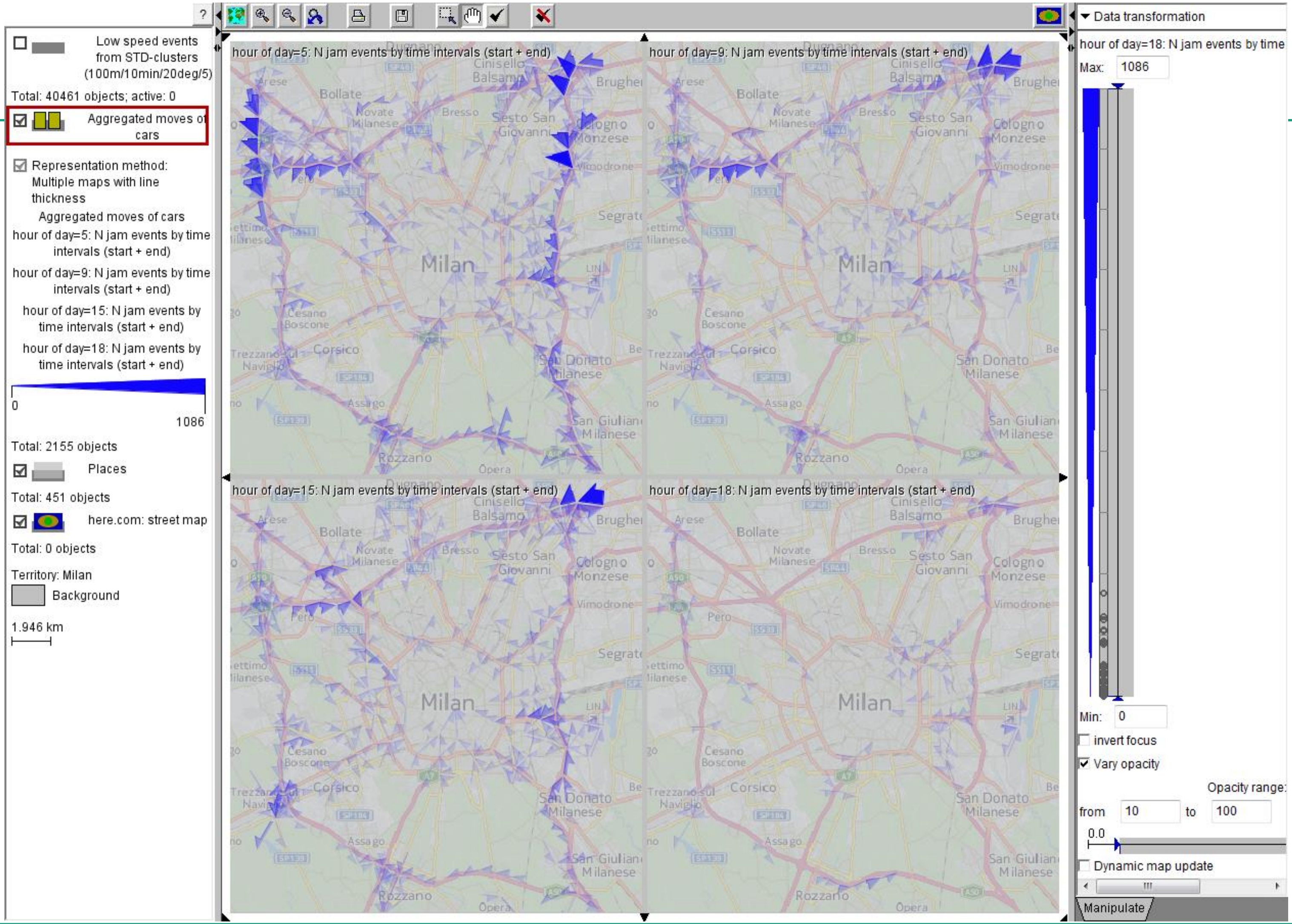
Fraunhofer IAIS

Traffic jam events summarized by areas and day times

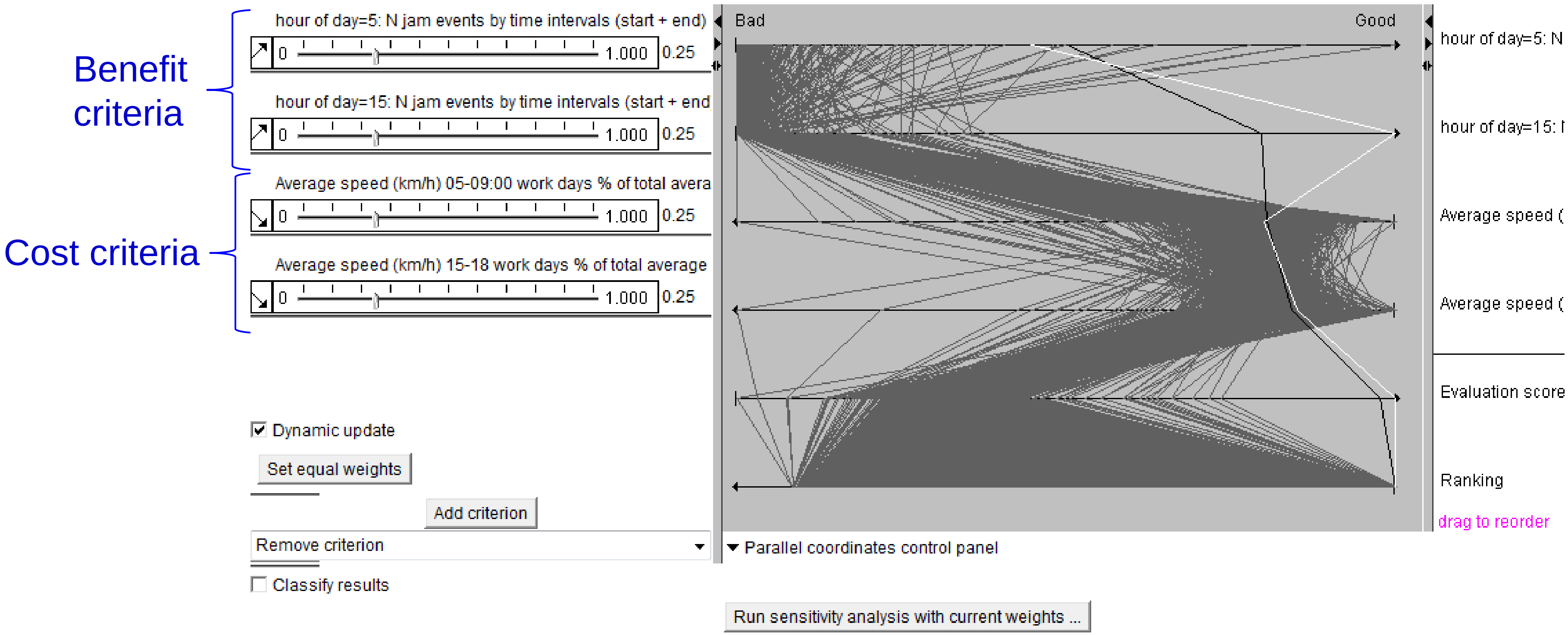


Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

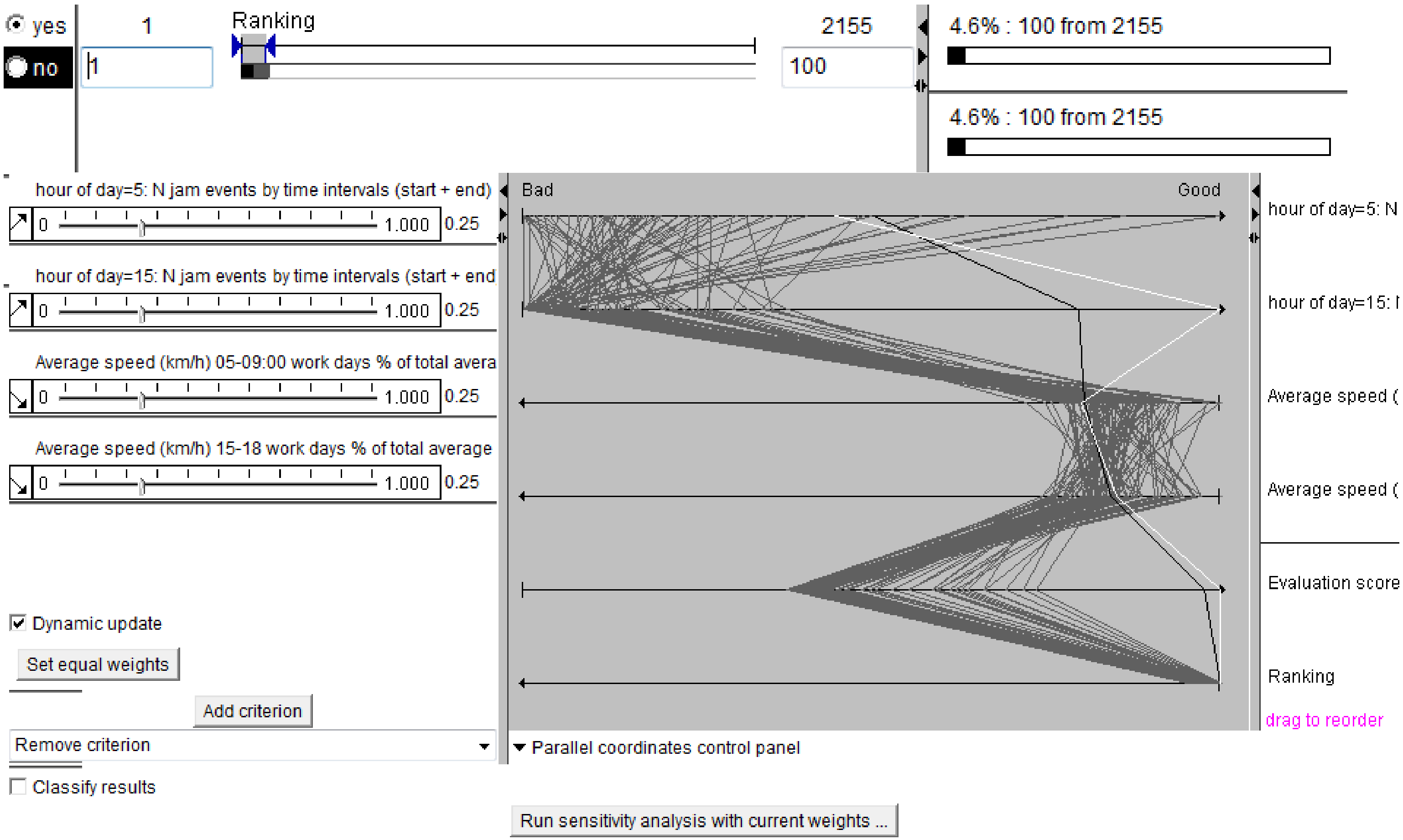
Traffic jam events summarized by links and day times



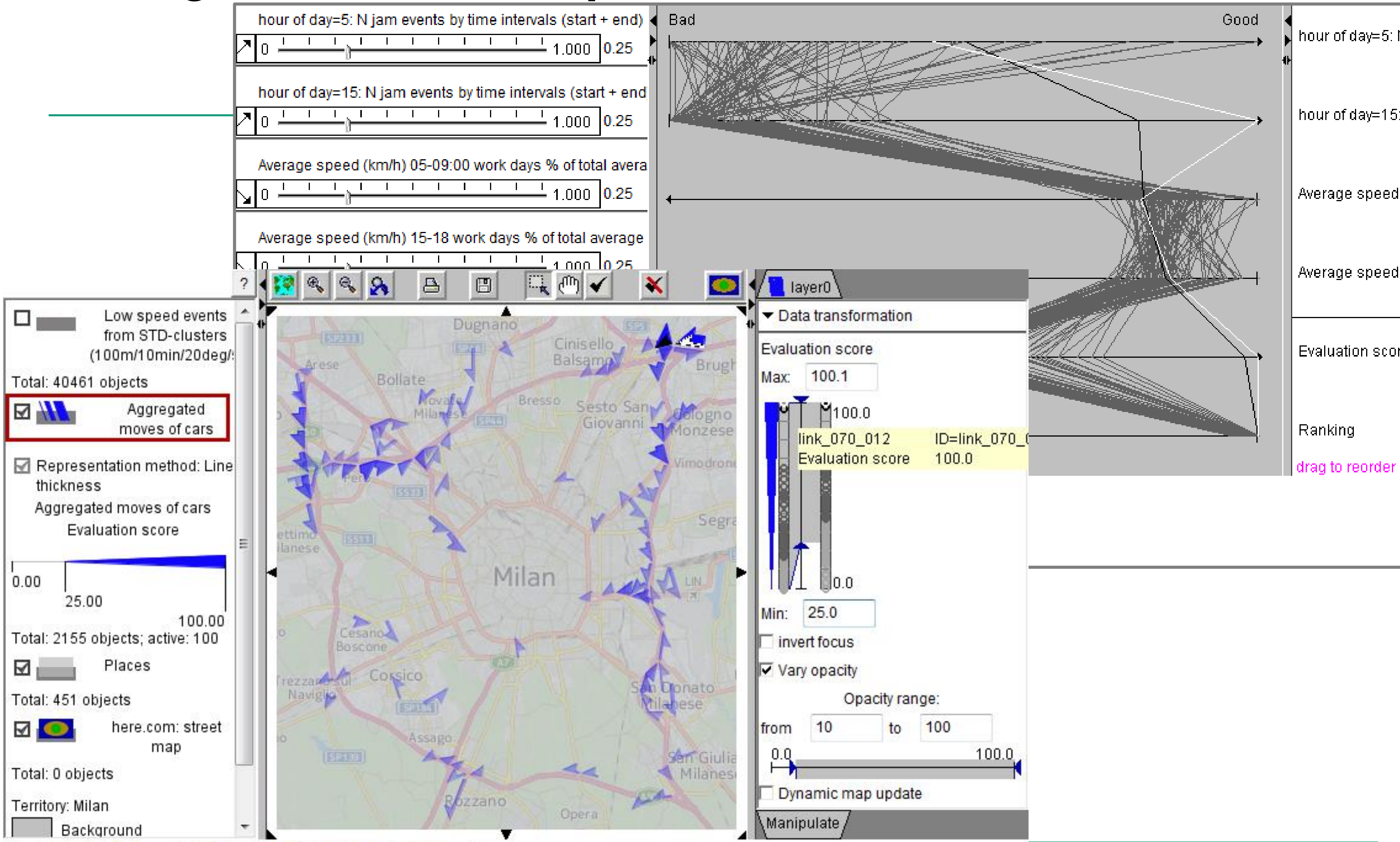
Multi-criteria choice support tool



Filtering by rank



Seeing results on a map

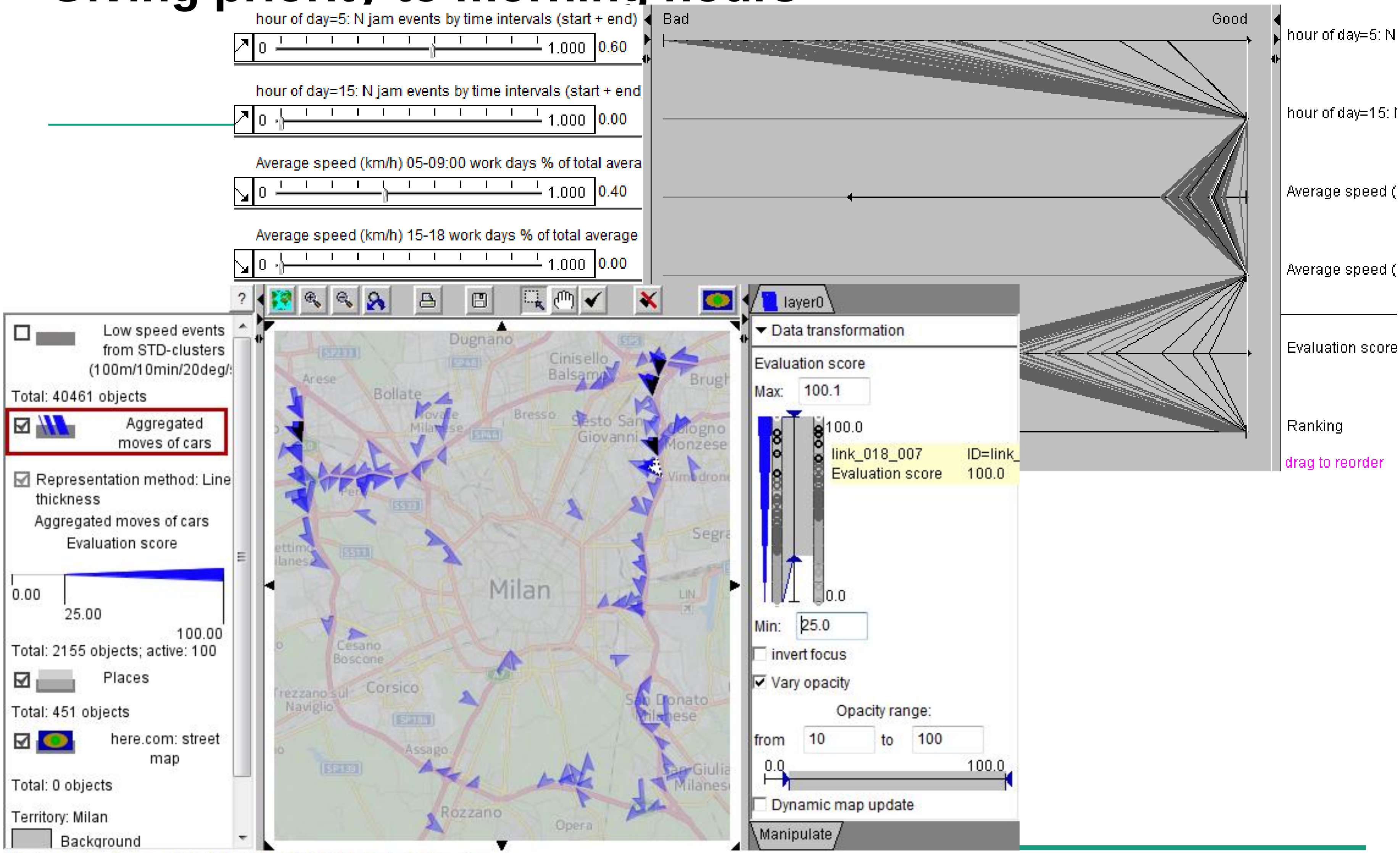


Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

CITY UNIVERSITY LONDON

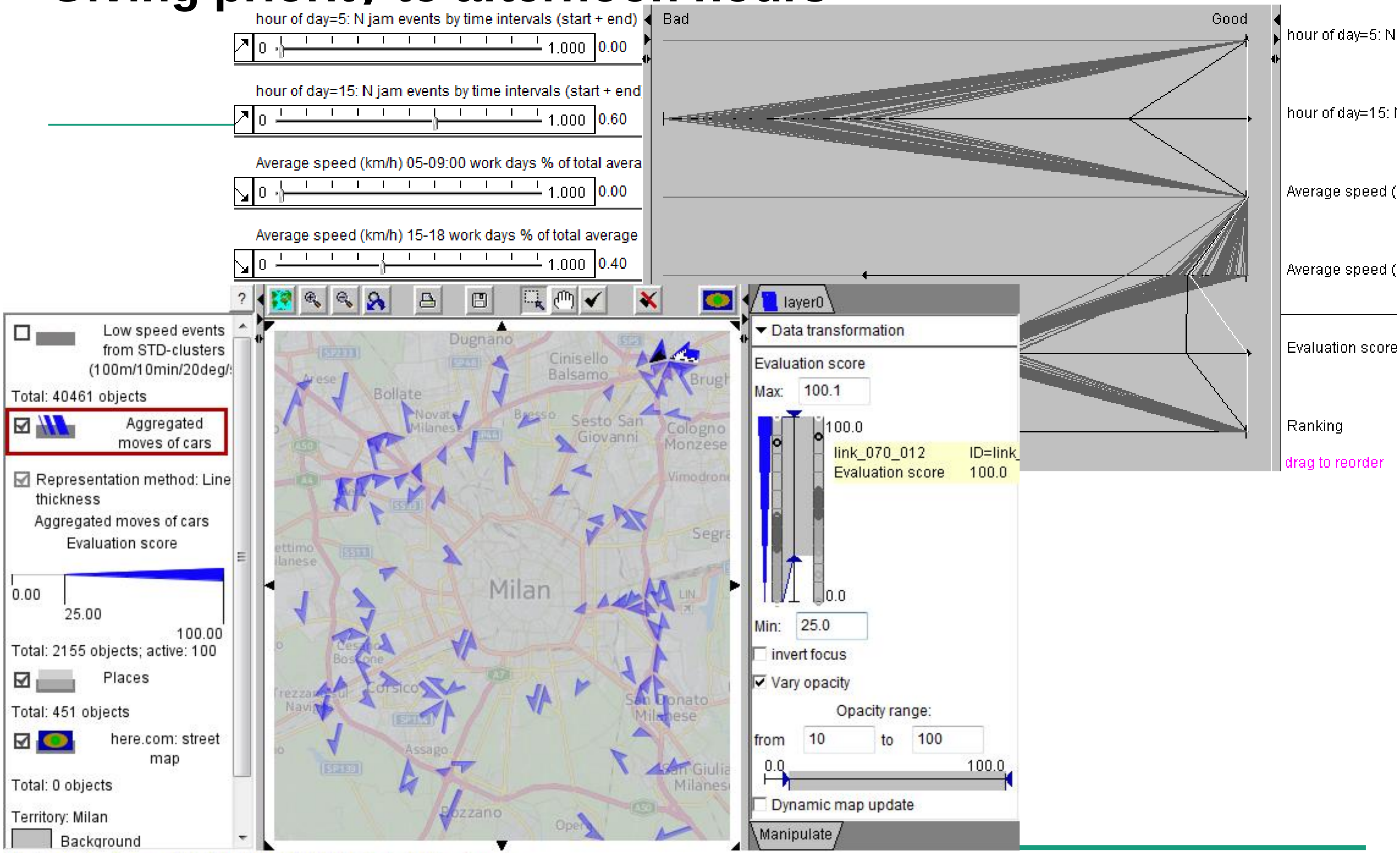
Fraunhofer IAIS

Giving priority to morning hours



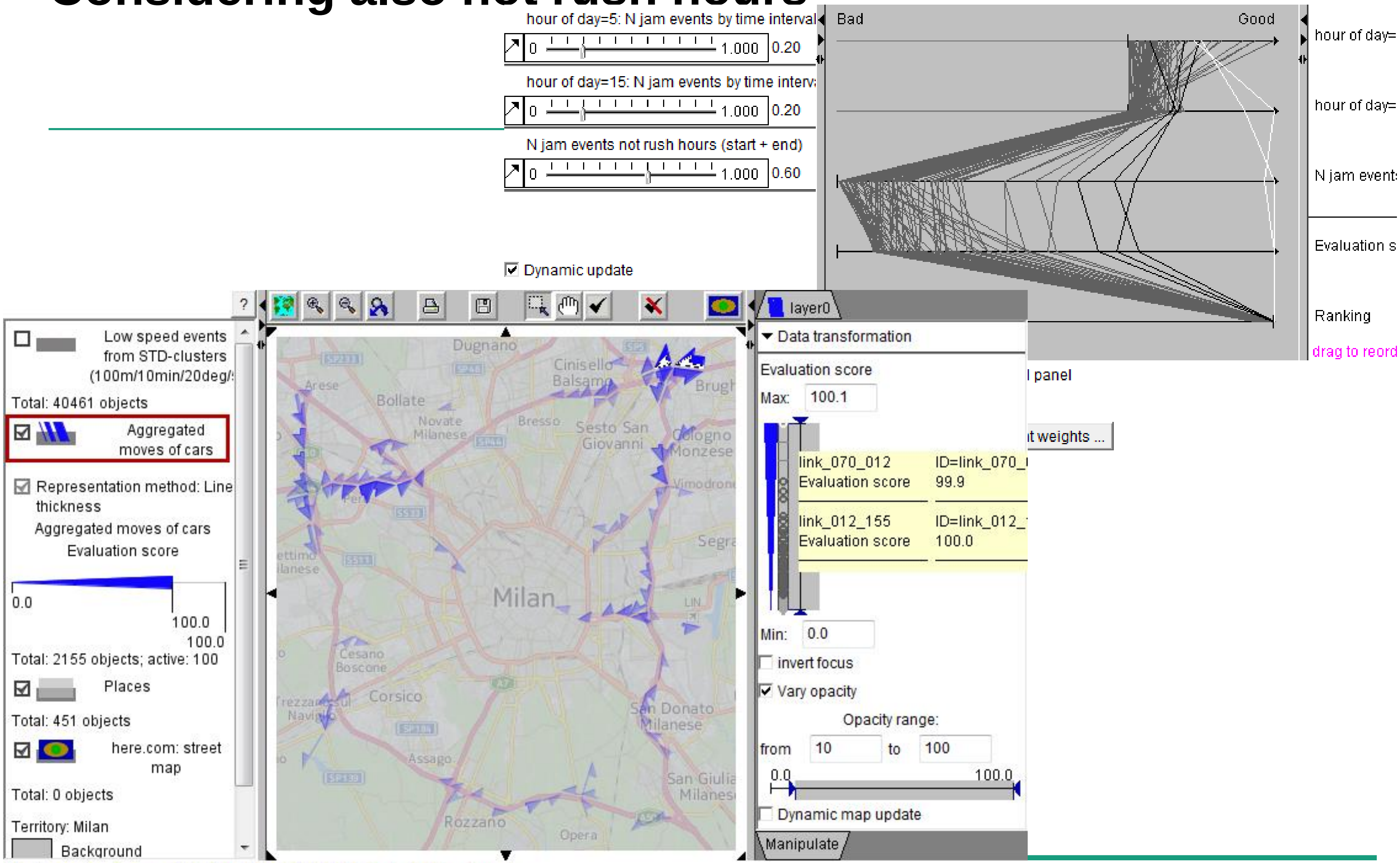
Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

Giving priority to afternoon hours



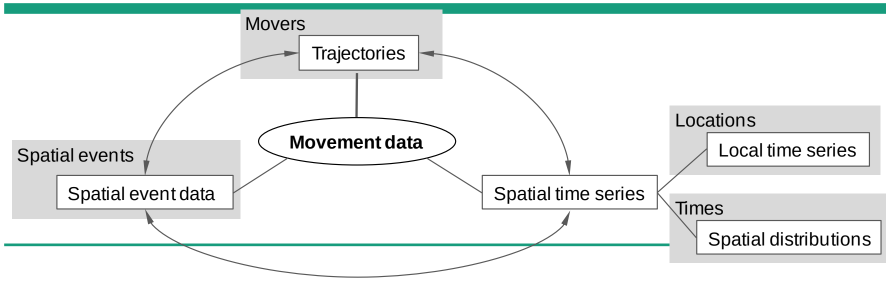
Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

Considering also not rush hours



Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Traffic in Milan, 1 week

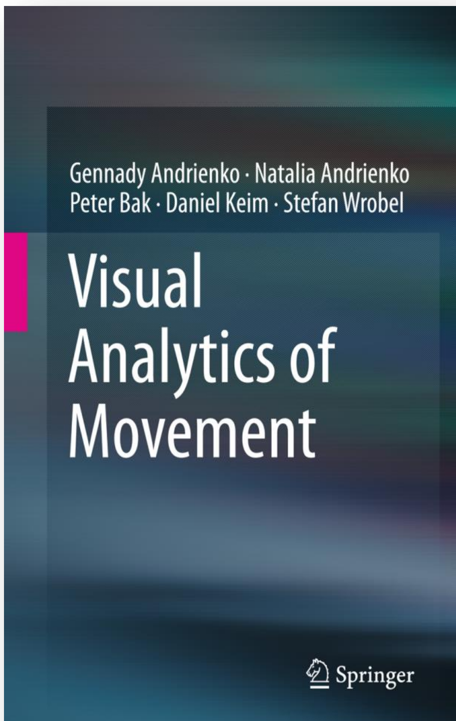
CITY UNIVERSITY LONDON
Fraunhofer IAIS



Springer, June 2013

ISBN 978-3-642-37582-8

397 p. 200 illus., 178 in colour



CITY UNIVERSITY LONDON

Fraunhofer IAIS