



Module INM433 – Visual Analytics

Lecture 03

Complex data structures Use of clustering

given by

prof. Gennady Andrienko and

prof. Natalia Andrienko



Content and objectives

- We shall consider two classes of data structure: *object-referenced time series* and *space-referenced time series*. These are data with 2 referrers: objects \times time or places \times time.
- Such complex data may be hard to analyse by purely interactive visual techniques.
- Clustering is one of the computational tools that can help us to analyse complex data.
- The lecture will introduce two major types of clustering: *partition-based* and *density-based*. It will explain how each type of clustering is used in data analysis and how clustering results are investigated using interactive visualisations.
- In the practical, you will try to apply the two types of clustering.



Complex data structures and complex behaviours

Data with 2 or more referrers



Complex behaviour

- In case of 2 (or more) referrers, we need to analyse the behaviour of the attributes over a complex reference set consisting of all available combinations of values of the referrers.
 - I.e., the behaviour of A over the Cartesian product $R_1 \times R_2 \times \dots$
- Such a complex behaviour cannot be represented by a single image and observed as a whole.
- To study and describe a complex behaviour, we need to decompose it into slices and aspects



Decomposing a complex behaviour

Data with 2 referential components: $X \times Y \rightarrow A$ (e.g., $S \times T \rightarrow A$)

- The overall behaviour of A over the set $X \times Y$:

$$B_{X \times Y}(A(x, y))$$

- **Slices** of the overall behaviour:

$$B_Y(A(x, y) \mid x = \text{const}); B_X(A(x, y) \mid y = \text{const})$$

- For a selected element from X , what is the behaviour of A over Y ?
- For a selected element from Y , what is the behaviour of A over X ?

- **Aspects** of the overall behaviour:

$$B_X(B_Y(A(x, y))); B_Y(B_X(A(x, y)))$$

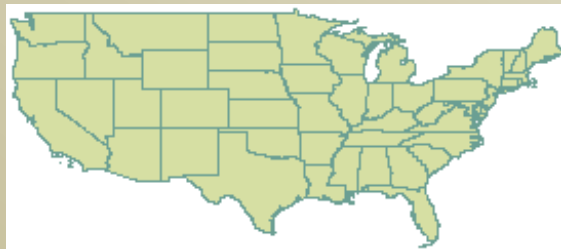
- How are the behaviours of A over Y distributed over X ?
- How are the behaviours of A over X distributed over Y ?



Decomposing a complex behaviour: slices (*illustrated for $S \times T \rightarrow A$, or $A(s,t)$*)

Overall behaviour: $B_{S \times T}(A(s,t))$

Space as a whole



Selected time

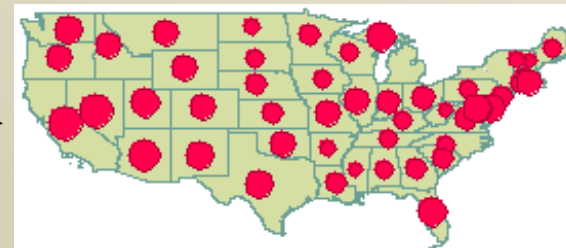
t_k

Selected place



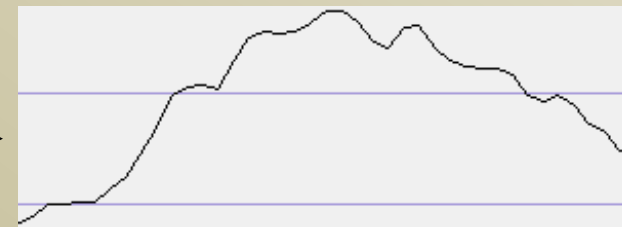
Time as a whole

Spatial **behaviour** at this time



$B_S(A(s,t) | t=t_k)$

Temporal **behaviour** in this place

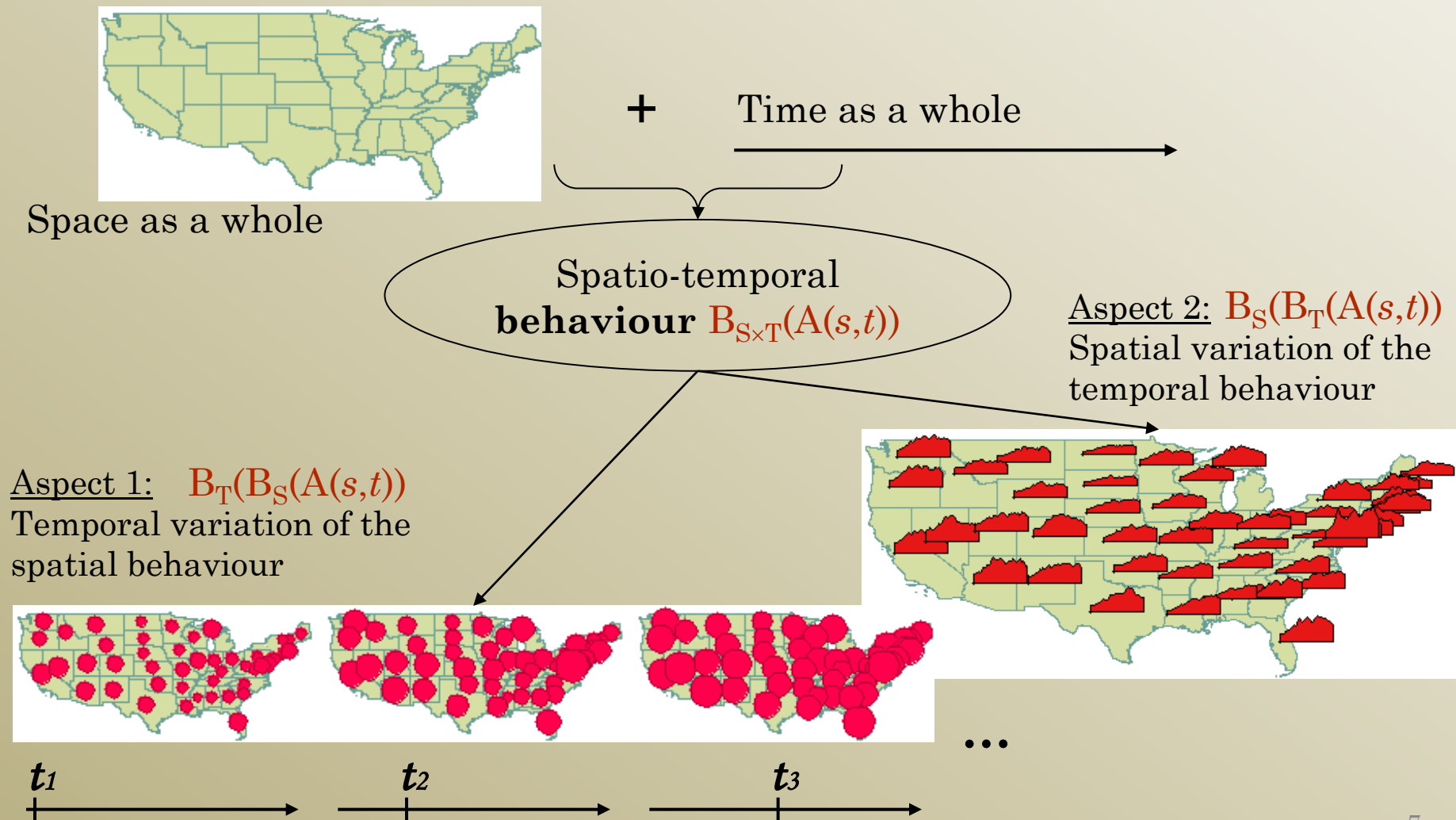


$B_T(A(s,t) | s=s_j)$



Decomposing a complex behaviour: aspects

(illustrated for $S \times T \rightarrow A$, or $A(s,t)$)





Relationship between a complex behaviour and its aspects

- A complex behaviour includes its aspects, but is it true that a complex behaviour is a sum (union) of its aspects?

☹ No! A toy counter-example:

- $A(x,y)$: $x \in \{1,2,3\}=X$; $y \in \{a,b,c\}=Y$; $A(x,y) \in \{\text{black}, \text{white}\}$
- $B_{X \times Y}(A(x,y)) \neq B_X(B_Y(A(x,y))) \cup B_Y(B_X(A(x,y)))$

| | | | | |
|---|---|---|---|--------------------------|
| a | ■ | ■ | ■ | $B_{X \times Y}(A(x,y))$ |
| b | □ | ■ | □ | |
| c | □ | ■ | □ | |
| | 1 | 2 | 3 | |

- Aspects \approx projections of the overall behaviour
 - Similar to 2D projections of the shape of a 3D object:
 - ☹ The object shape is not a sum or union of its projections.
 - ☺ But we can imagine (i.e., reconstruct) the shape in our mind by looking at the projections.

\Rightarrow We can reconstruct the overall complex behaviour in our mind by studying the aspectual behaviours

- ☹ It may be quite complex
- ☺ It is not always necessary – depends on the analysis goals.



Questions?

Complex data structures and complex
behaviours



Object- or space-referenced time series

Data with 2 referrers:

- time
- set of objects or set of spatial locations



Object-referenced time series

- Referrers:
 - **O**: set of objects
 - **T**: time, i.e., sequence of time moments or intervals, called *time steps*
- + One or more thematic* attributes **A** \Rightarrow function **A(o,t)**
 - * \neg spatial & \neg temporal
- Overall behaviour **B_{O×T}(A(o,t))** : distribution of the attribute values over the objects and time
- Aspect **B_O(B_T(A(o,t)))**: distribution of the temporal variations of **A** over the set of objects
- Aspect **B_T(B_O(A(o,t)))**: temporal variation of the distribution of **A** over the set of objects

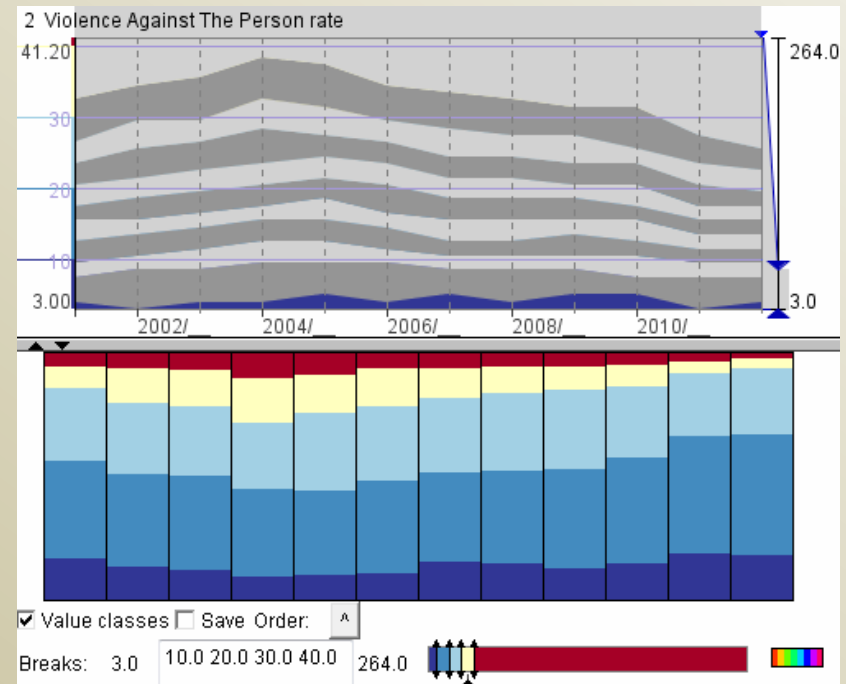
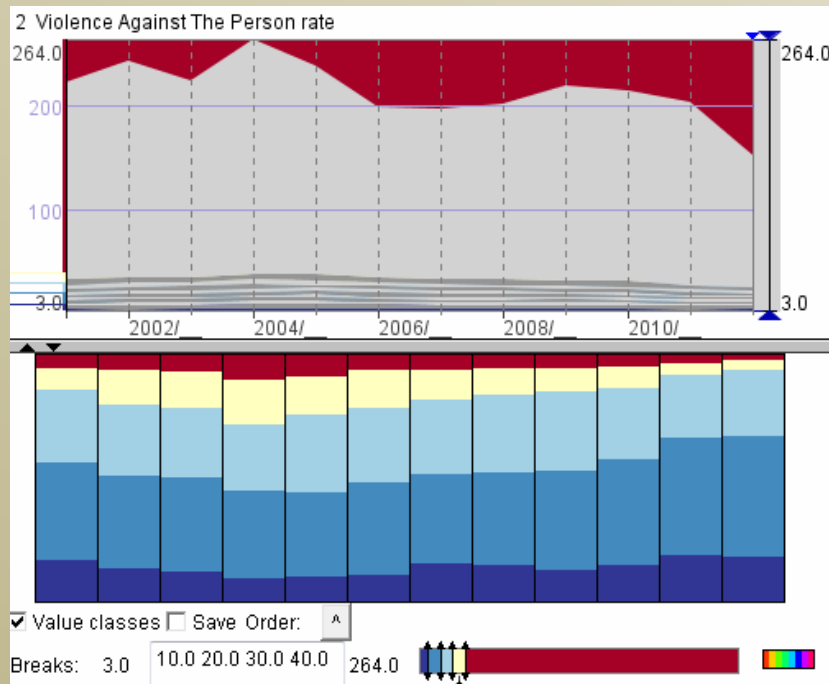


Space-referenced time series (a.k.a. spatial time series)

- Referrers:
 - **S**: set of spatial locations or spatial objects
 - **T**: time, i.e., sequence of time moments or intervals, called *time steps*
- + One or more thematic* attributes **A** \Rightarrow function **A**(s,t)
 - * \neg spatial & \neg temporal
- Overall behaviour **B_{S×T}**(**A**(s,t)) : distribution of the attribute values over the space and time
- Aspect **B_S**(**B_T**(**A**(s,t))): spatial distribution of the temporal variations of **A**
- Aspect **B_T**(**B_S**(**A**(s,t))): temporal variation of the spatial distribution of **A**



Temporal variation of the distribution of A over the set of objects $B_T(B_O(A(o,t)))$

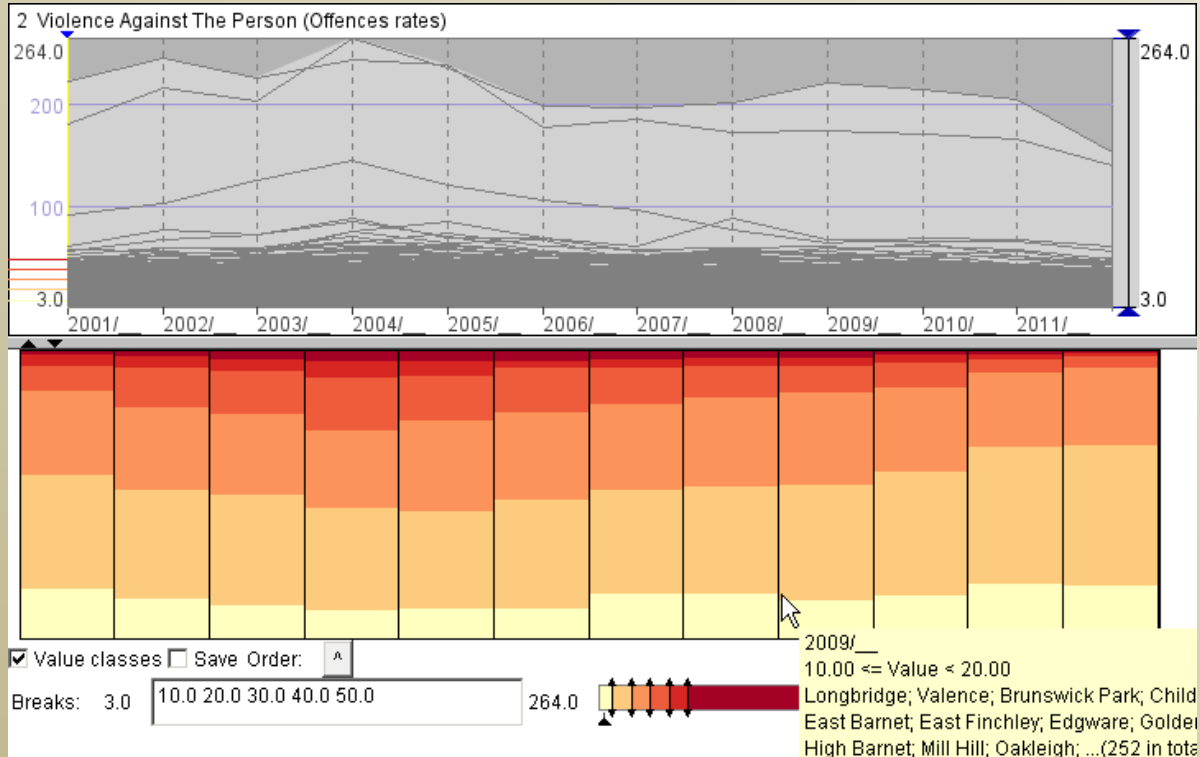


Please recall these techniques introduced in lecture 1



Segmented bars (time histogram)

A reminder from lecture 1



This technique is applicable to data with **two referrers**: one linearly ordered (such as time) and one arbitrary.

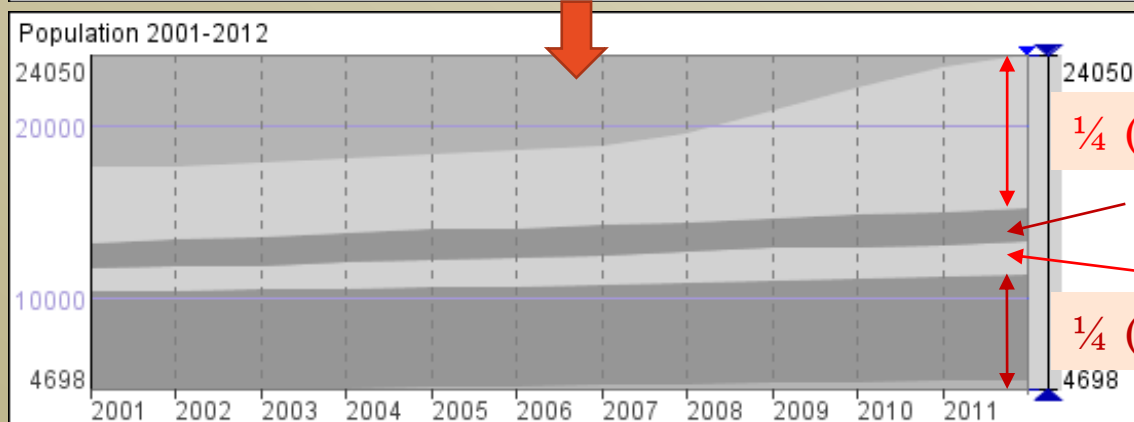
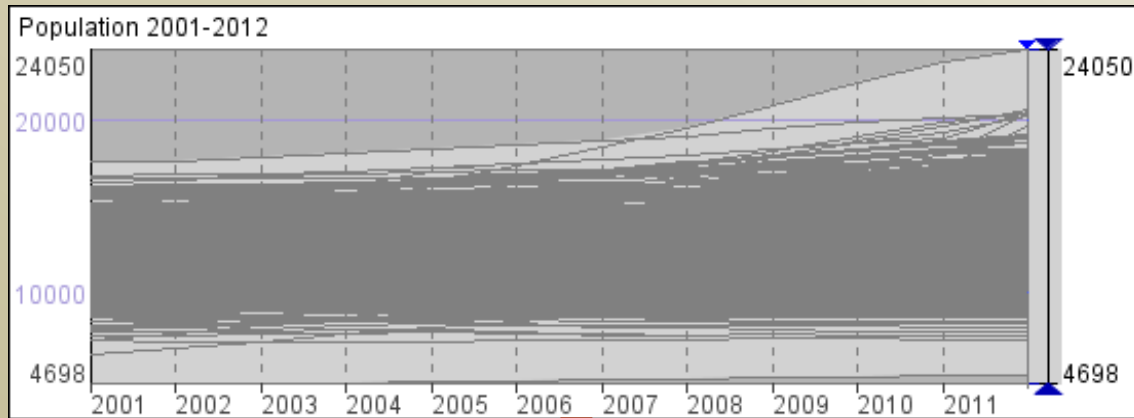
Note that a line graph with multiple curves may be unreadable due to line overlapping.

Each bar stands for one time step (generally, one value of referrer 1). The value range of the attribute is divided into intervals. The height of each segment is proportional to the number of values of referrer 2 for which the values in this time step lie within the respective interval. The intervals are represented by the segment colours.



Quantile graph

A reminder from lecture 1



$\frac{1}{4}$ (25%) of the values

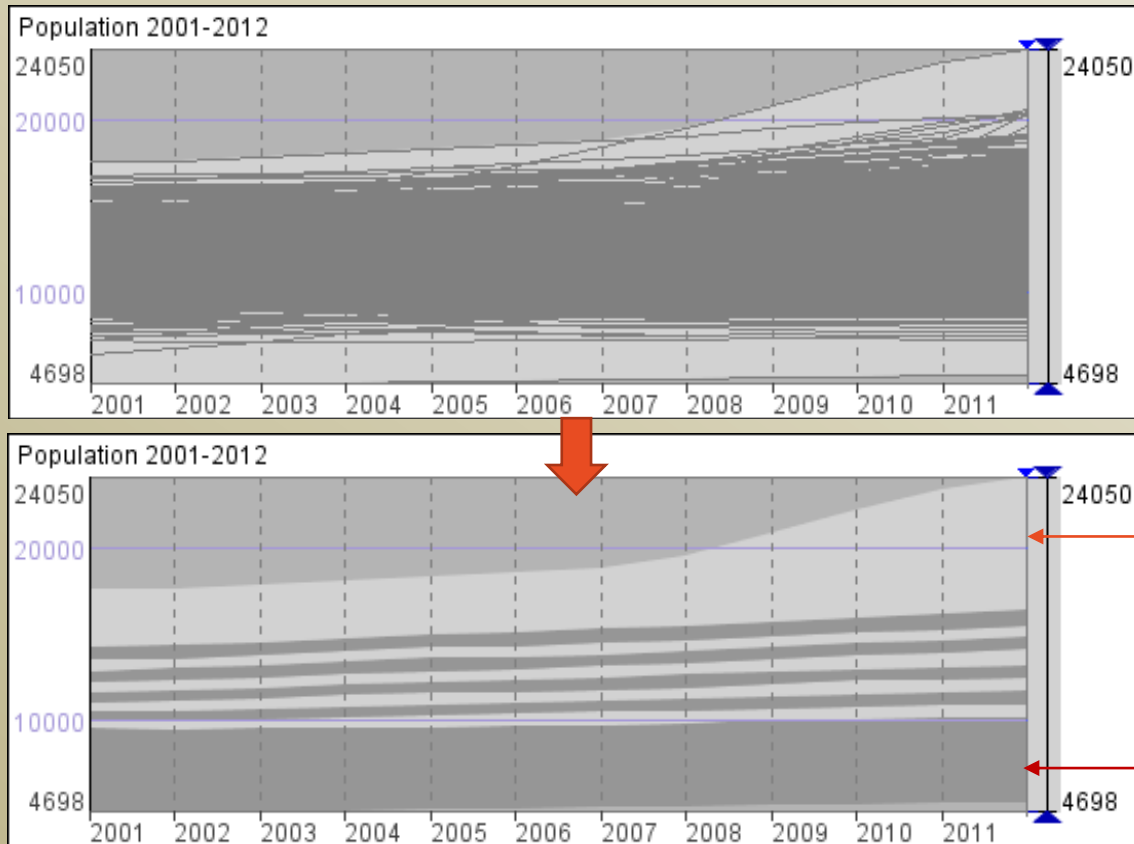
$\frac{1}{4}$ (25%) of the values

$\frac{1}{4}$ (25%) of the values

$\frac{1}{4}$ (25%) of the values



Quantile graph (*continued*)



$1/_{10}$ (10%) of the values

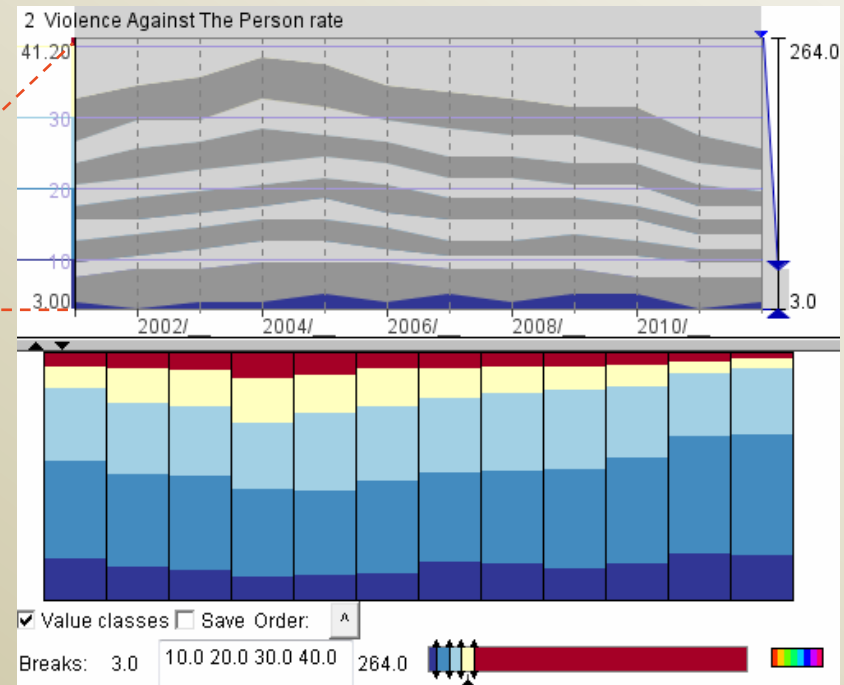
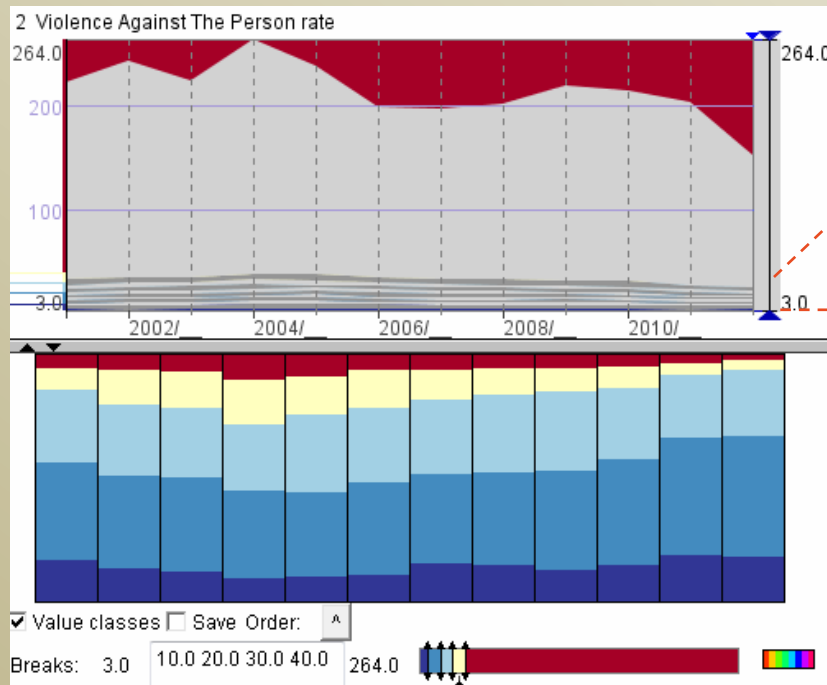
...

$1/_{10}$ (10%) of the values

Segmented bars and quantile graphs show how the value distribution w.r.t. one referrer (set of objects, places, etc.) varies over the range of linearly ordered values of another referrer (such as time).



Temporal variation of the distribution of A over the set of objects $B_T(B_O(A(o,t)))$



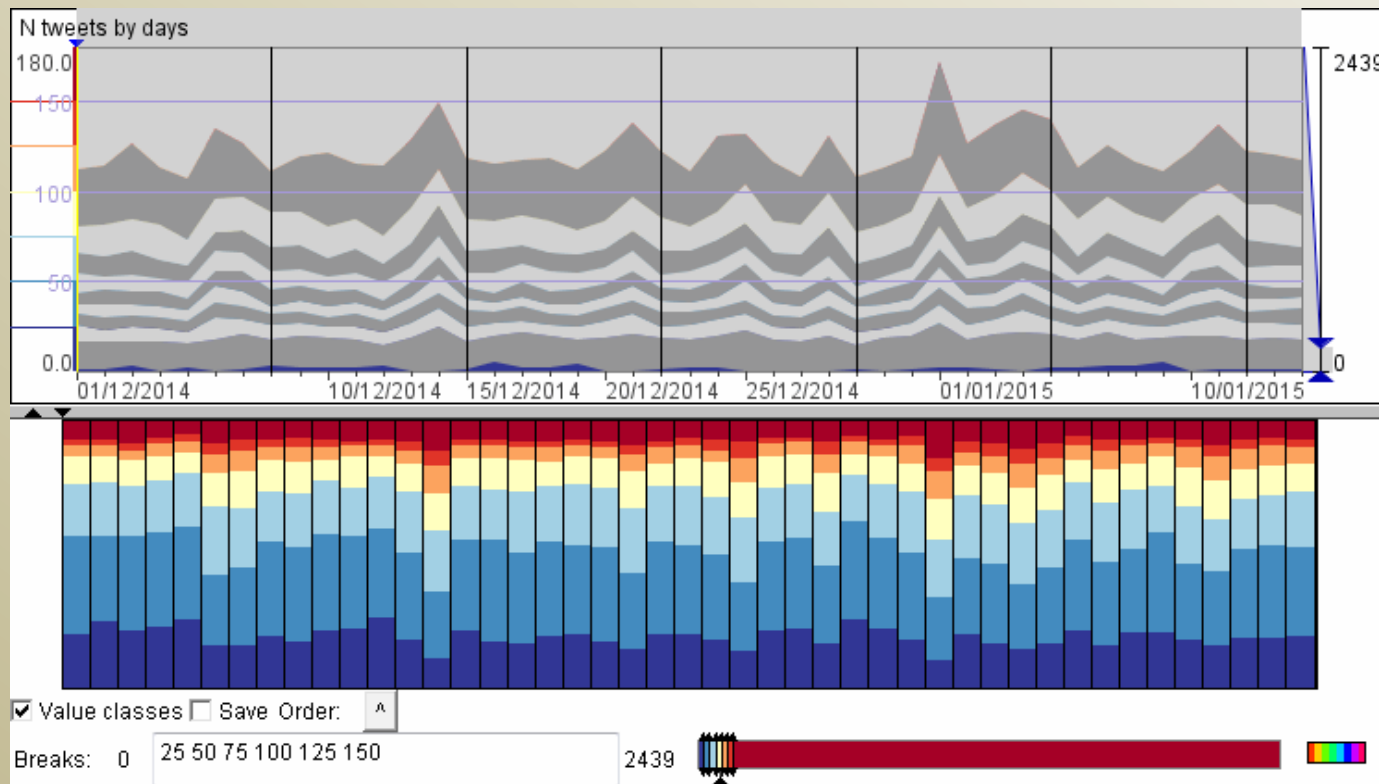
Quantile time graph (top) and segmented bars (bottom) show the temporal variation of the statistical value distribution of a numeric attribute.

Generally, the distribution did not change much. The number of higher values increased by year 2004/2005 and then gradually decreased from year to year.

Temporal variation patterns (behaviour types): constancy, temporal trend (increase, decrease).



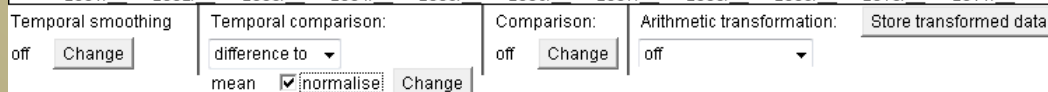
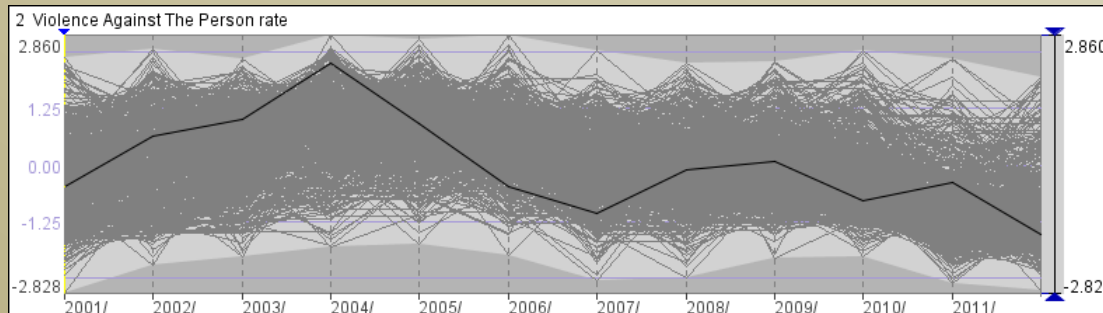
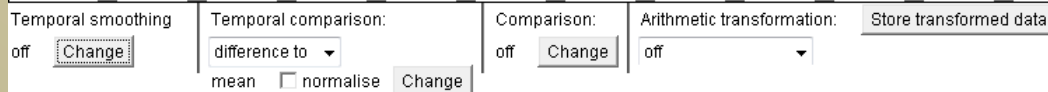
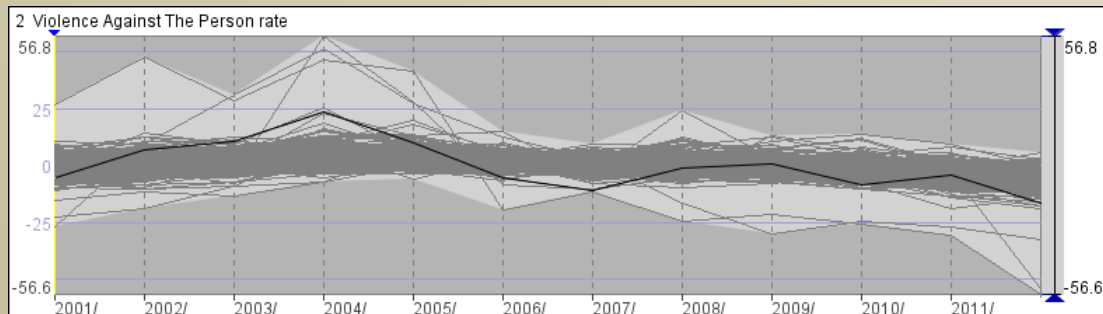
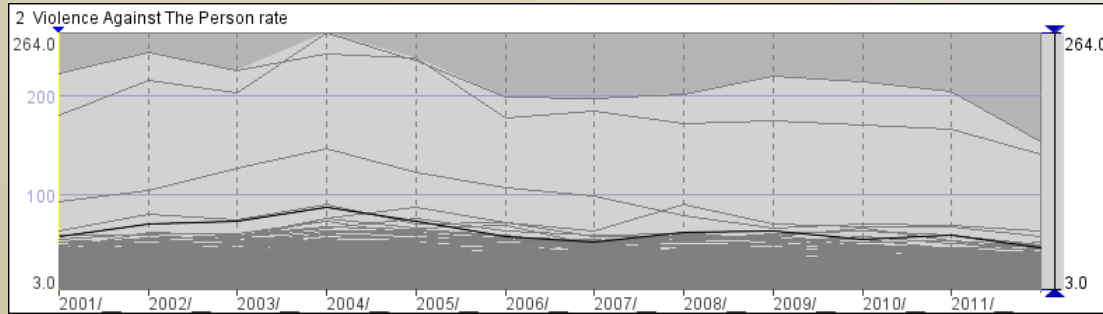
Temporal variation of the distribution of A over the set of objects $B_T(B_O(A(o,t)))$



For another attribute and longer time series, we see a periodic temporal pattern. The data refer to daily intervals; vertical lines mark Mondays. We observe a stable (nearly constant) distribution over the week days and increase of the number of high values in the weekends. This periodic behaviour is interrupted by special days: Christmas and New Year.



Transformations of time series



Let

$$T = \{t_1, t_2, \dots, t_i, \dots, t_N\}$$

$$O = \{o_1, o_2, \dots, o_k, \dots, o_M\}$$

v_i^k : attribute value for object o_k in time step t_i .

The attribute values have been transformed to the differences w.r.t. the mean of each time series:

$$v_i^k \rightarrow v_i^k - \mu^k$$

$$\mu^k = \sum (v_i^k \mid 1 \leq i \leq N) / N$$

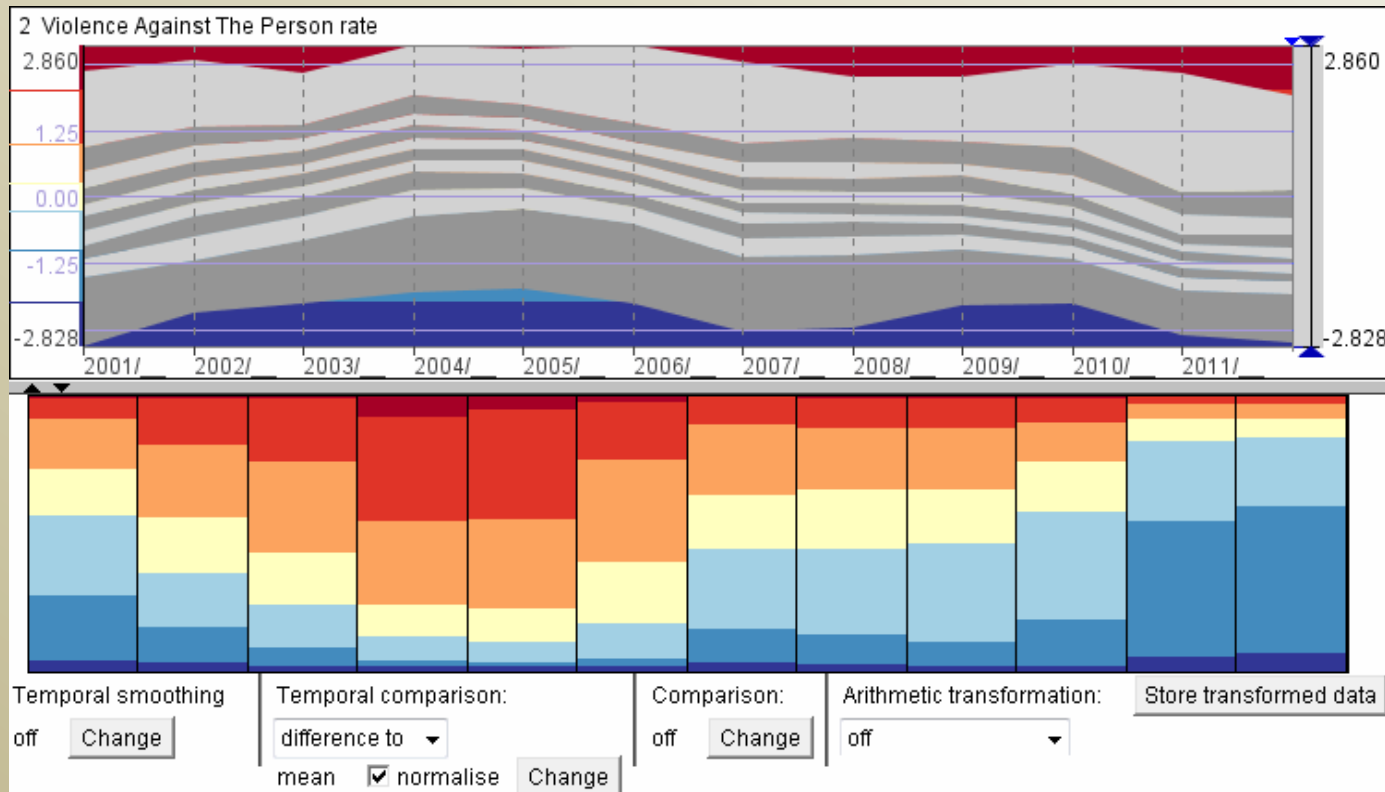
The attribute values have been transformed to the differences w.r.t. the time series means normalised by the standard deviations:

$$v_i^k \rightarrow (v_i^k - \mu^k) / \sigma^k$$

$$\sigma^k = \sqrt{\sum (v_i^k - \mu^k \mid 1 \leq i \leq N) / (N-1)}$$



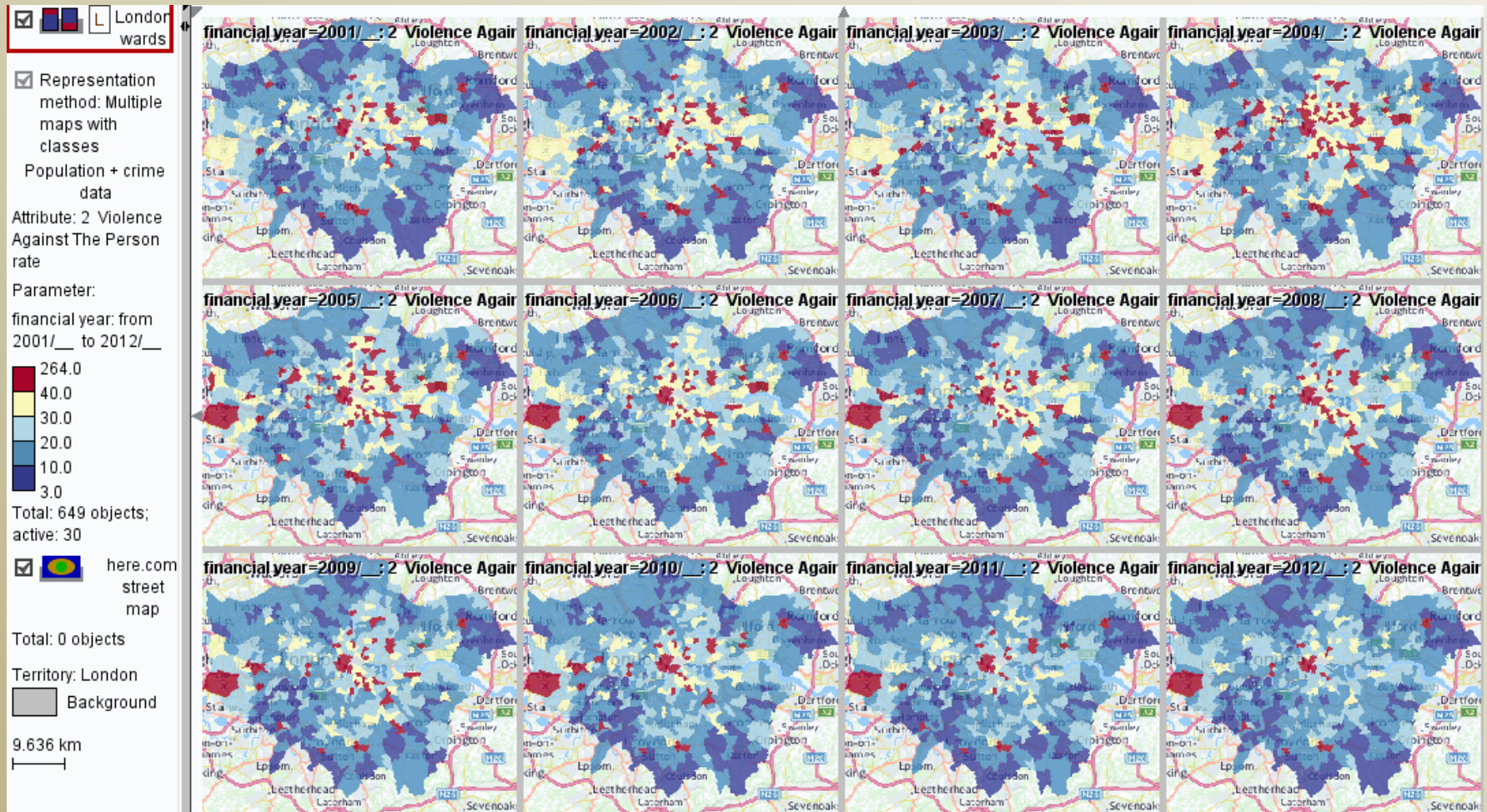
Transformation of time series + aggregation



The increasing and decreasing temporal trends have become more prominent. We can detect time intervals with particular distributions of values, e.g., >80% of objects with values above the mean or about 90% of objects with values below the mean.



Temporal variation of the distribution of A over space $B_T(B_S(A(s,t)))$



When the time series are short, the technique “small multiples” is applicable.



Temporal variation of the distribution of A over space $B_T(B_S(A(s,t)))$



Multiple classified or unclassified choropleth maps can show the temporal variation of the spatial distribution for a small number of time steps.



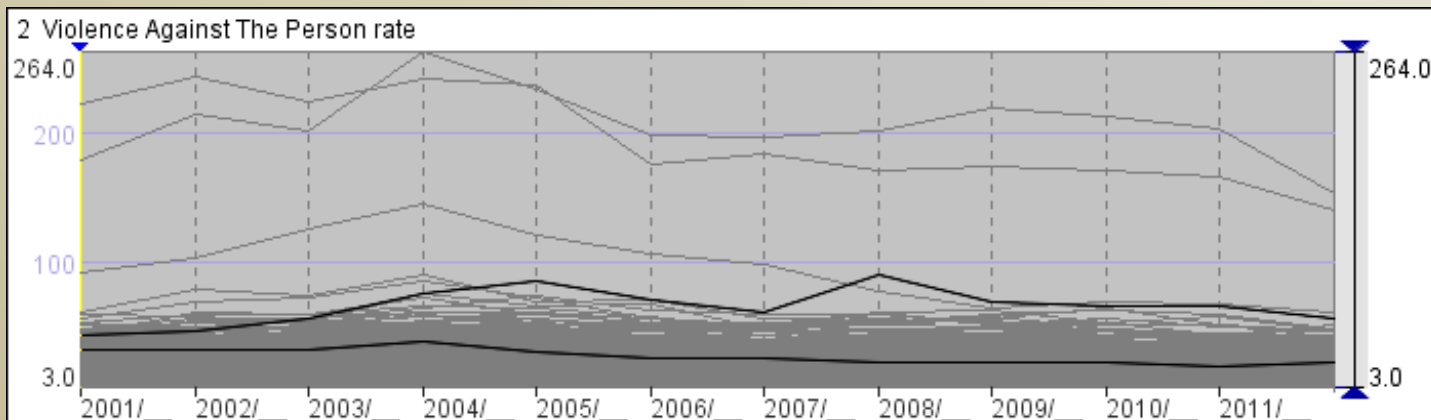
Limitations of multiple maps

- Applicable only to short time series
 - Temporal patterns are not easy to detect
 - E.g., the increasing and decreasing temporal trends are not so clear in the previous examples
 - It may be even more complex to detect periodic patterns
- ⇒ It is useful to combine multi-map displays with aggregated temporal displays (quantile time graph and segmented bars).
- Display linking through interactive selection, e.g., by clicking on bar segments.
 - For longer time series, only visual and interactive techniques are insufficient.
 - Clustering can help (*to be shown later*)



Distribution of the temporal variations of A over the set of objects $B_O(B_T(A(o,t)))$

- When the objects are few, the corresponding temporal variations (time series) can be directly compared, e.g., using a time graph.
- However, this may hardly be possible in case of many objects.

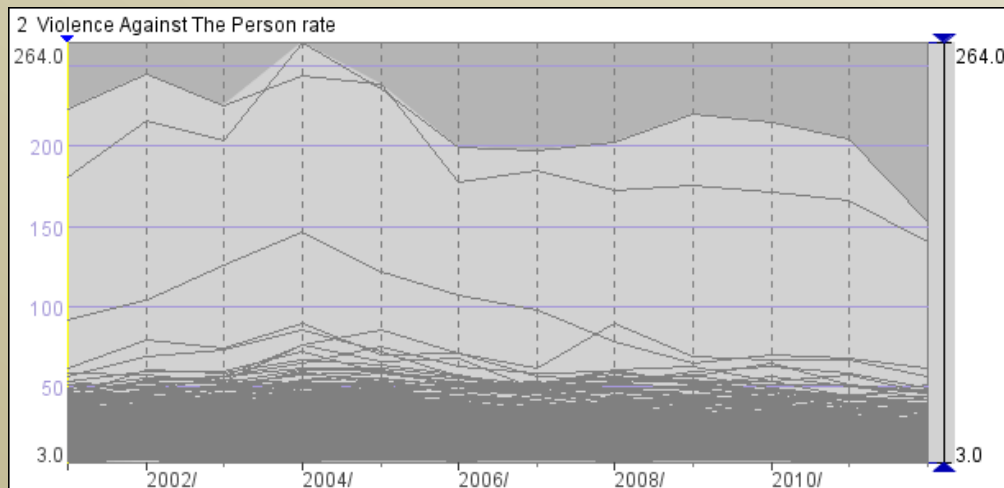


- A more feasible solution: group the time series by similarity and compare the groups.

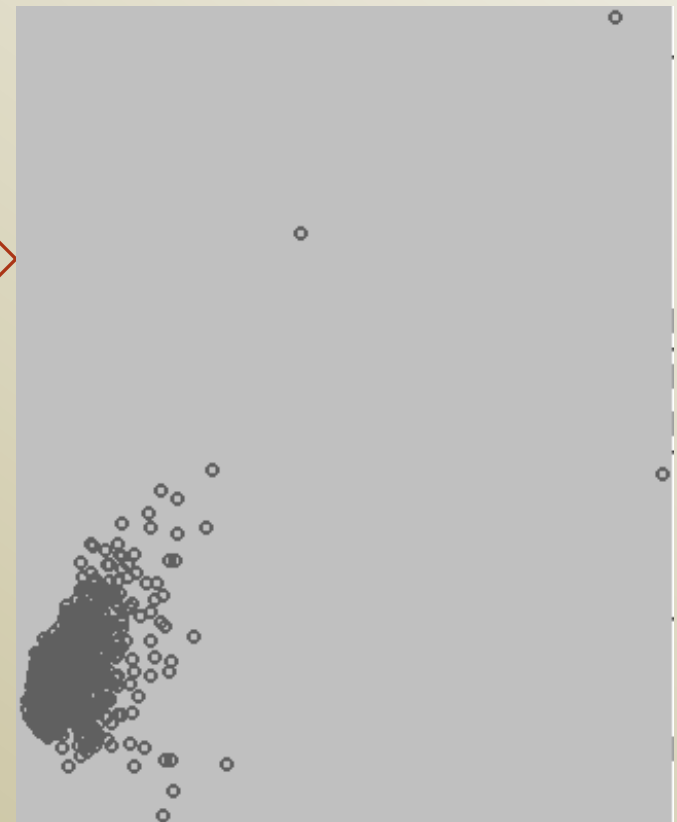


Grouping time series by means of projection (dimensionality reduction)

Step 1



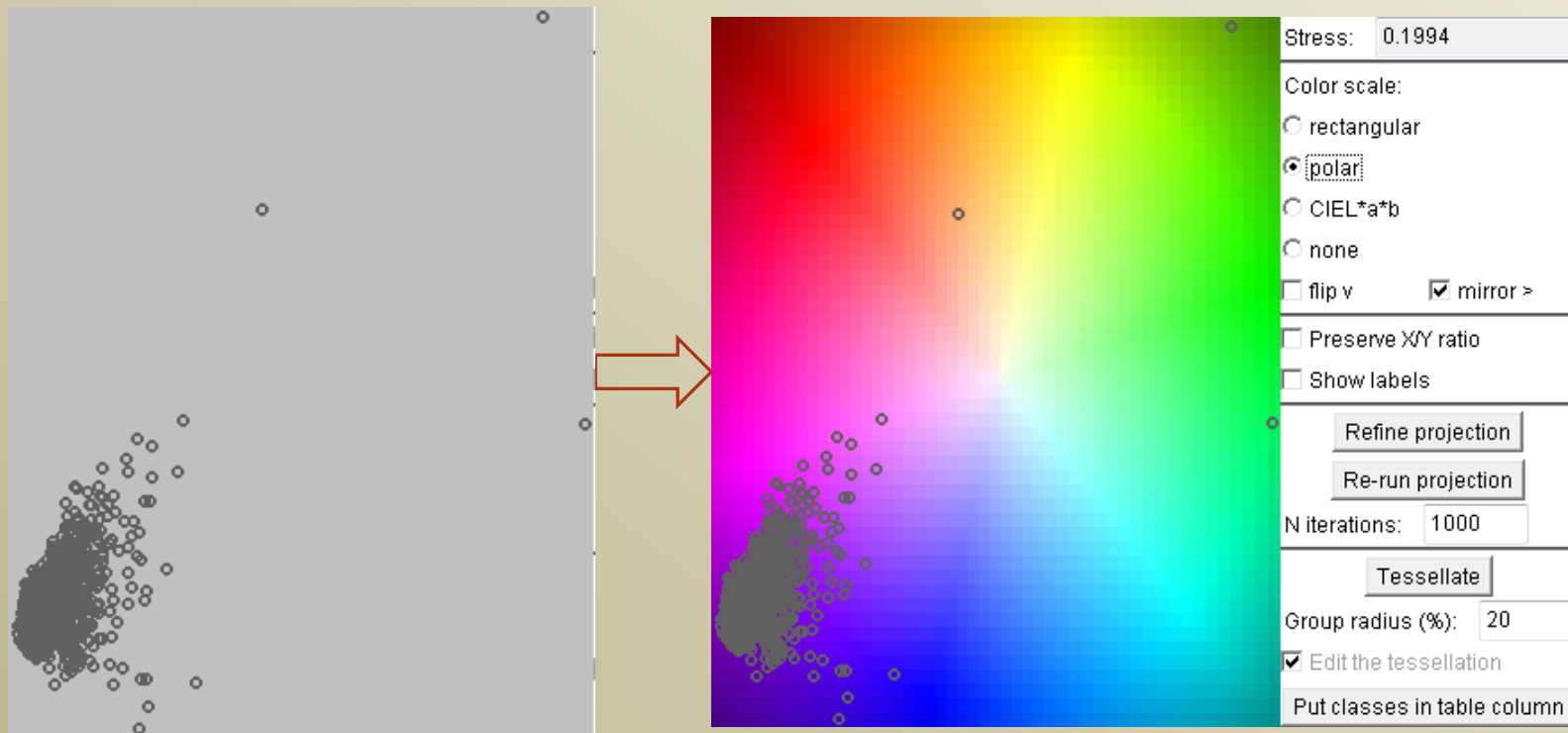
As we did earlier for multiple attributes, we apply projection (dimensionality reduction) to time series. Each time step is treated as a distinct attribute.





Grouping time series by means of projection (dimensionality reduction)

Step 2

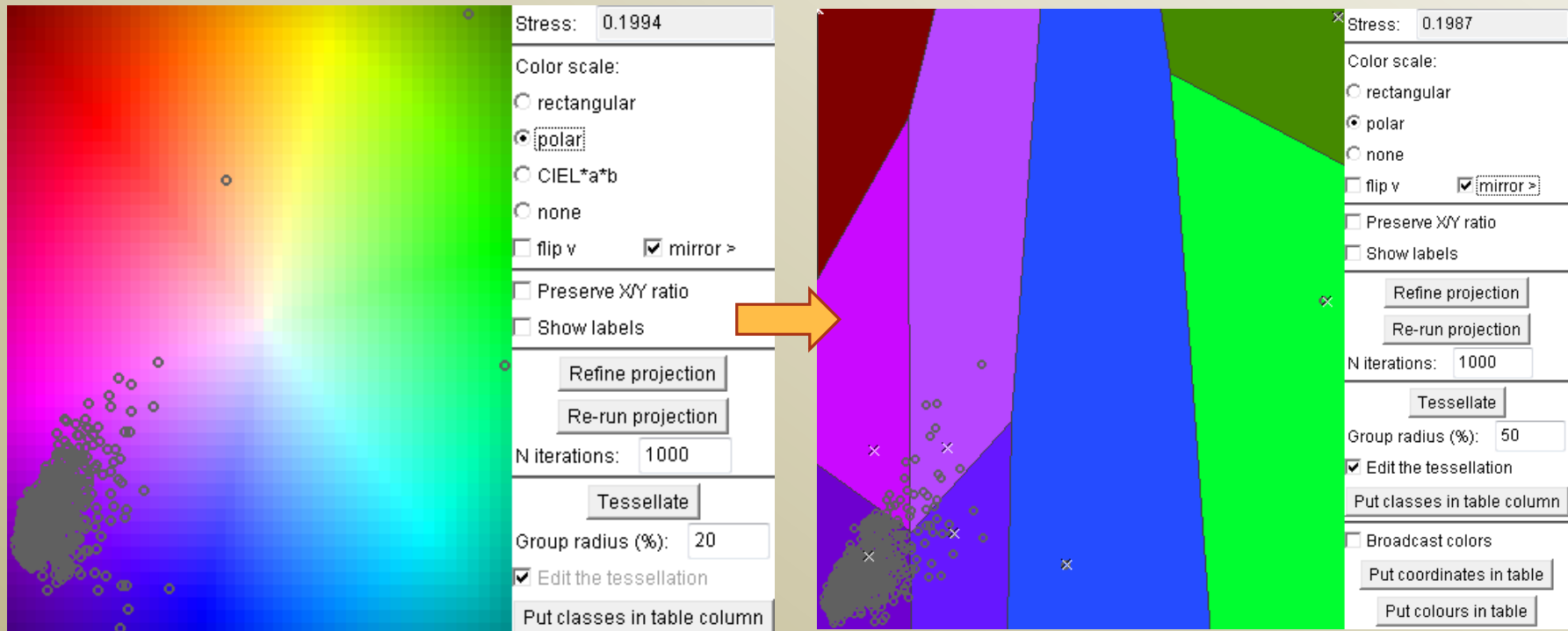


The projection is superimposed on a continuous 2D colour scale.



Grouping time series by means of projection (dimensionality reduction)

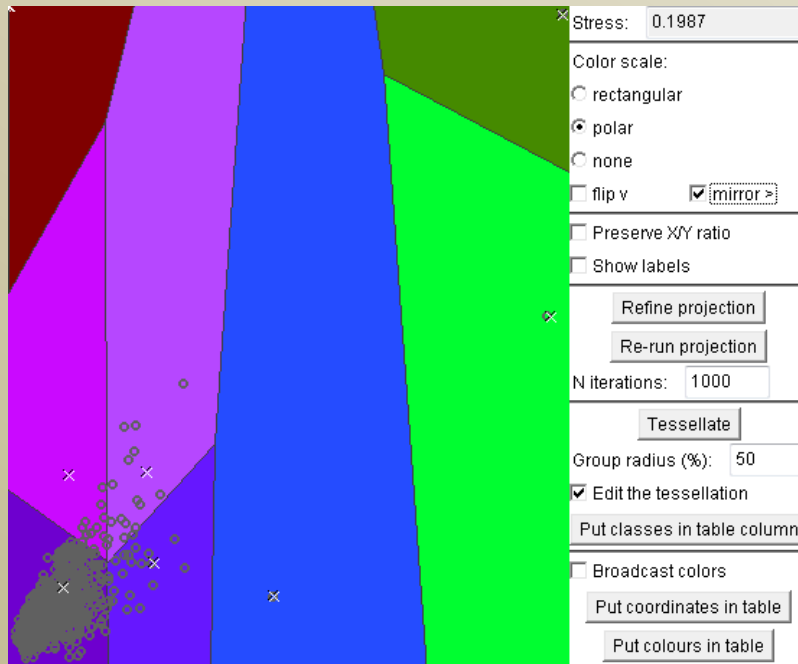
Step 3



By dividing the projection space into parts, we divide the points (and, hence, the time series represented by them) into groups. Small distance between points in the projection space means that the corresponding time series are similar.

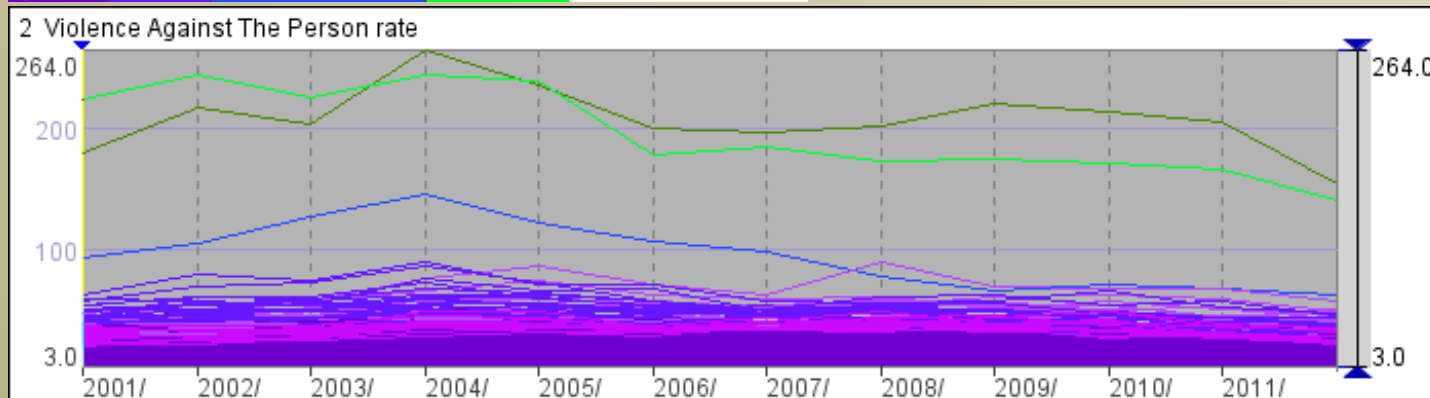


Using projection for grouping time series



As we did earlier for multiple attributes, we can apply projection (dimensionality reduction) to time series. Each time step is thereby treated as a distinct attribute.

After projecting the time series on a 2D coloured plane, we group them by tessellating the plane. The classes receive colours according to the positions of their centres. The colours are propagated to the time graph.



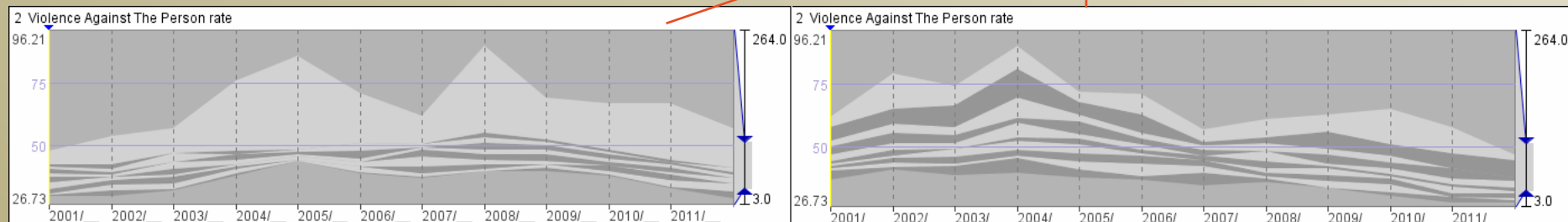
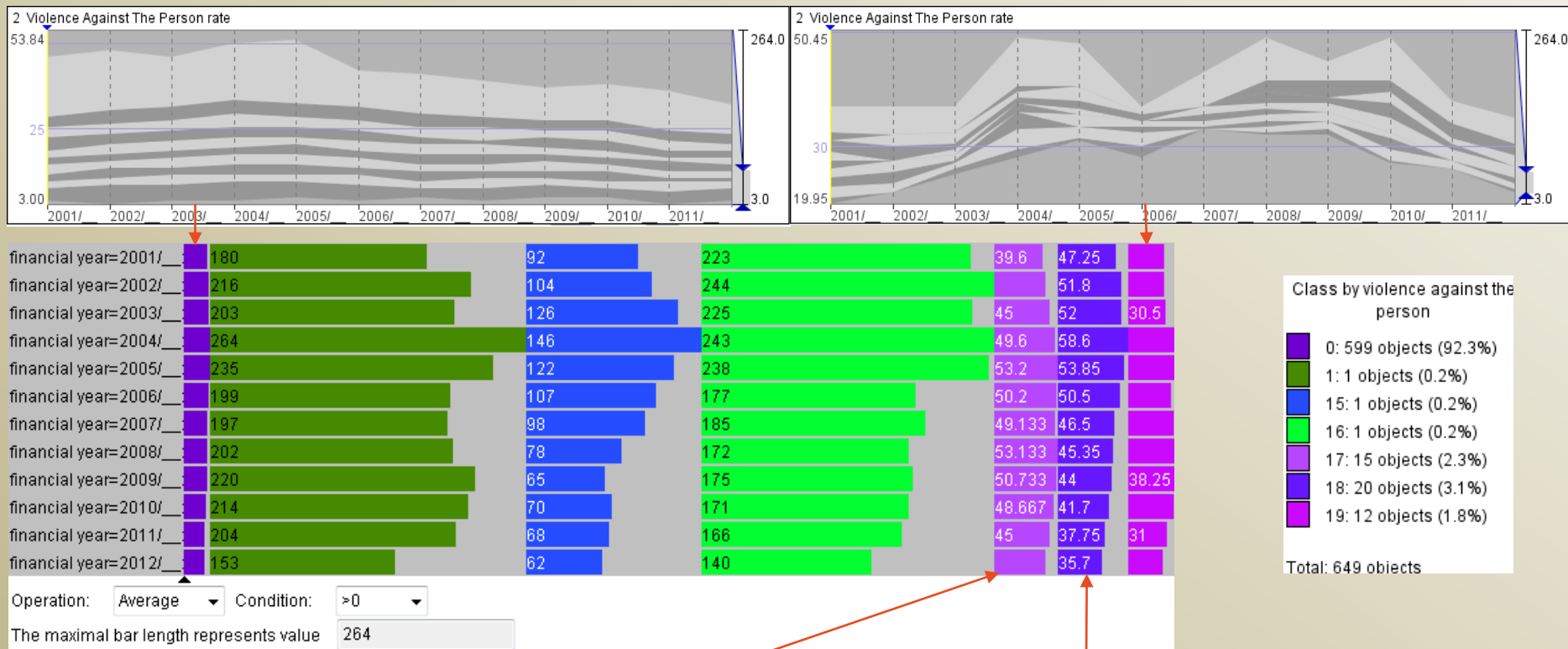
Class by violence against the person

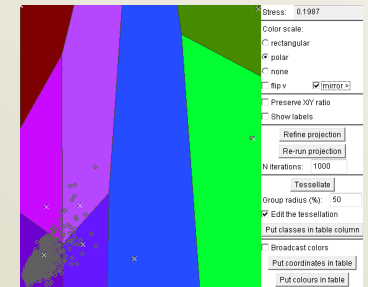
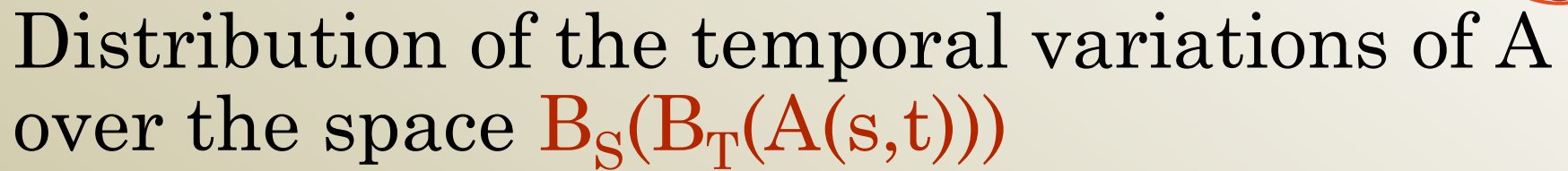
- 0: 599 objects (92.3%)
- 1: 1 objects (0.2%)
- 15: 1 objects (0.2%)
- 16: 1 objects (0.2%)
- 17: 15 objects (2.3%)
- 18: 20 objects (3.1%)
- 19: 12 objects (1.8%)

Total: 649 objects

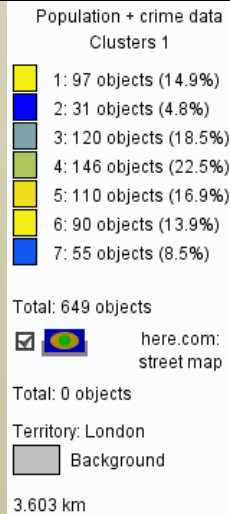
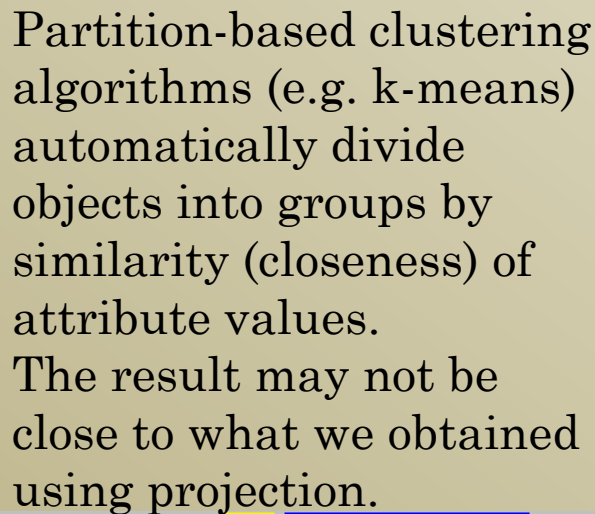


Comparison of groups (classes) of time series



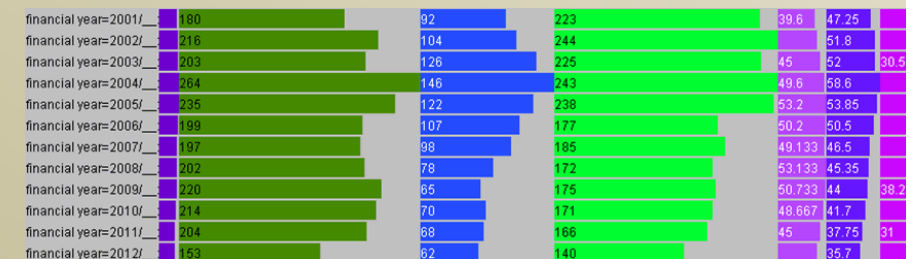
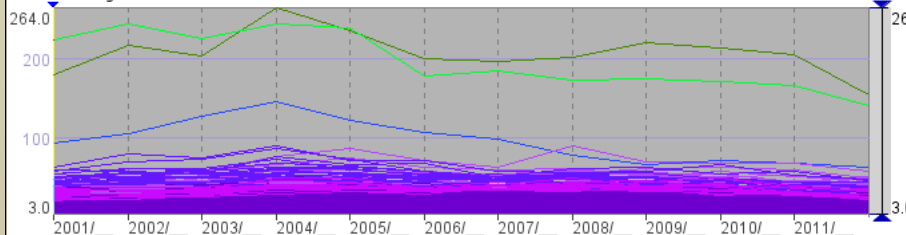


The map displays the extensive network of the London Underground, with lines radiating from central London to the surrounding suburbs. Key areas labeled include Watford, Harrow, Epsom, Dartford, and various districts within Greater London. The map also shows major roads and the River Thames.

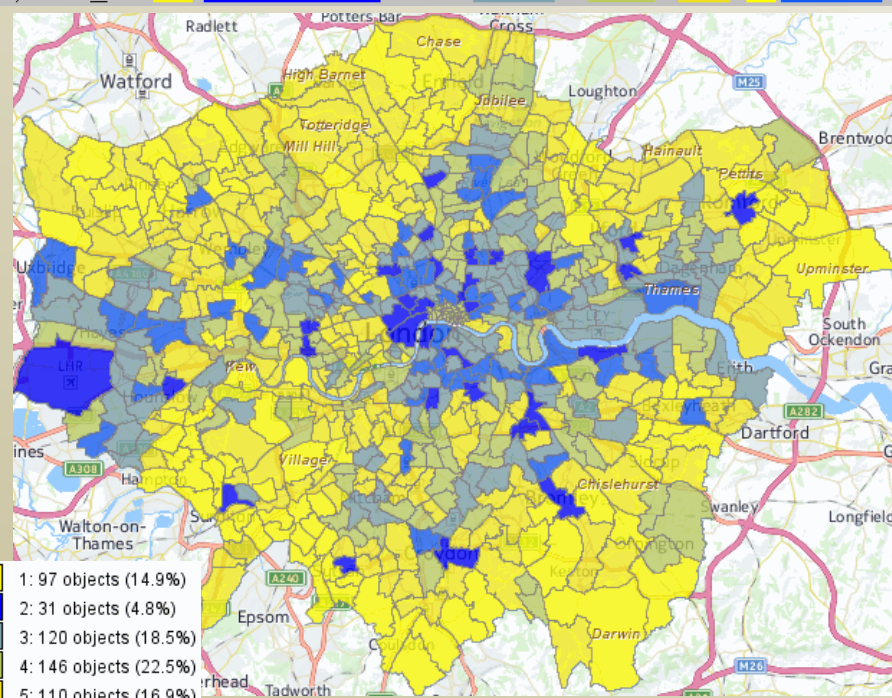
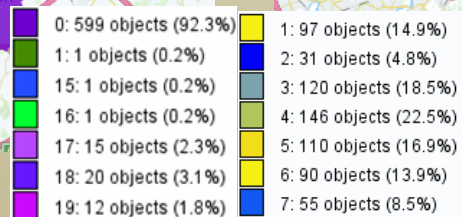
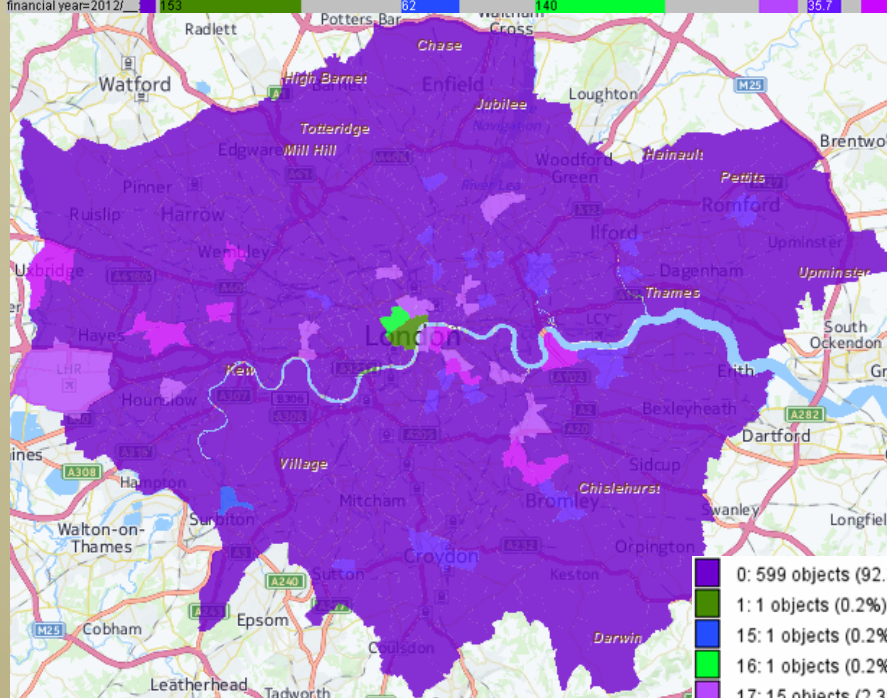
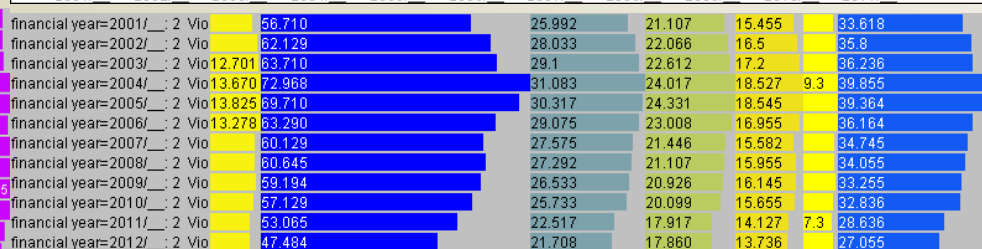
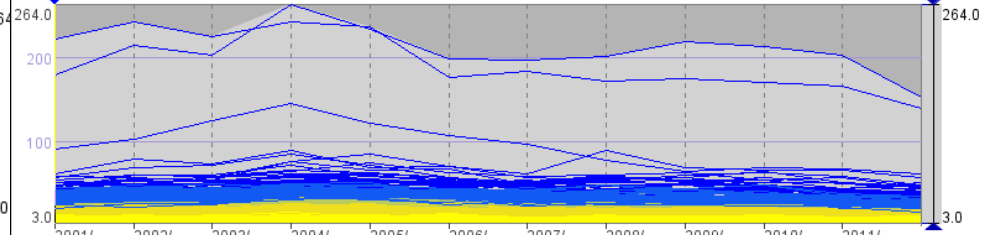




2 Violence Against The Person rate



2 Violence Against The Person rate





Different methods of grouping

- We have observed that the clustering algorithm applies different grouping principles than closeness of points in a projection space.
 - The grouping principles may also differ between different clustering algorithms.
 - You will learn how clustering algorithms work in the machine learning module.
- **There are no “right” and “wrong” groupings!**
 - Any grouping is a simplification (even an oversimplification!) of data \Rightarrow information loss.
 - We need simplification when data are large and complex.
 - To decrease the information loss:
 - Create and compare several groupings
 - Normally, results of different groupings should be consistent
 - Examine the intra-group variation



Object- and space-referenced time series

Preliminary summary (more to be said later)

- The complex overall behaviour $B_{O \times T}(A(o,t))$ or $B_{S \times T}(A(s,t))$ needs to be decomposed into aspects:
 - $B_T(B_O(A(o,t)))$ or $B_T(B_S(A(s,t)))$: how the distribution of the attribute values over the set of objects (B_O) or over the space (B_S) varies over time (B_T).
 - $B_O(B_T(A(o,t)))$ or $B_S(B_T(A(s,t)))$: how the temporal variations B_T of the attribute values are distributed over the set of objects (B_O) or over the space (B_S).
 - Each aspect is a kind of projection of the overall behaviour.
- General approach to analysis: grouping and aggregation
 - Attribute values are grouped by value intervals (in histograms).
 - Objects and places are grouped by similarity of their time series.
 - Time steps are grouped by similarity of the value distributions over the object sets or over space *(to be demonstrated later)*.



Questions?

Object- and space-referenced time series



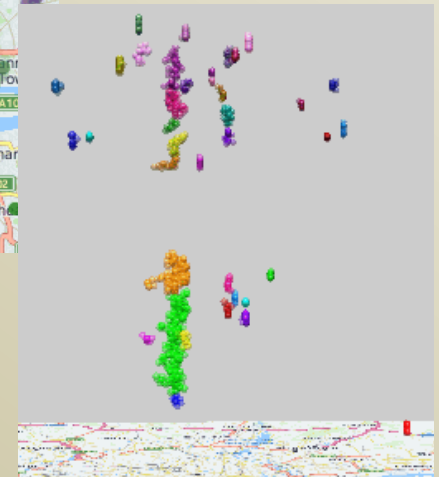
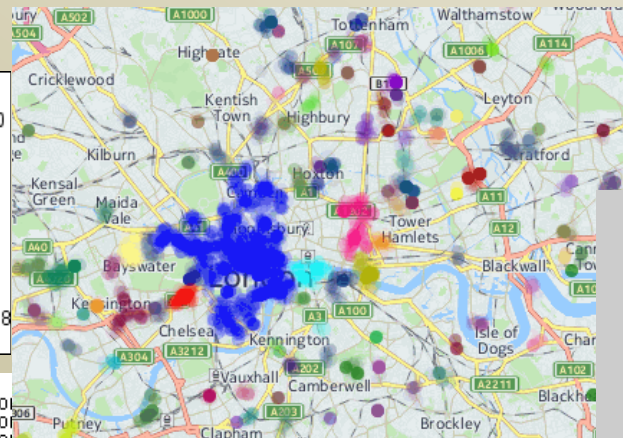
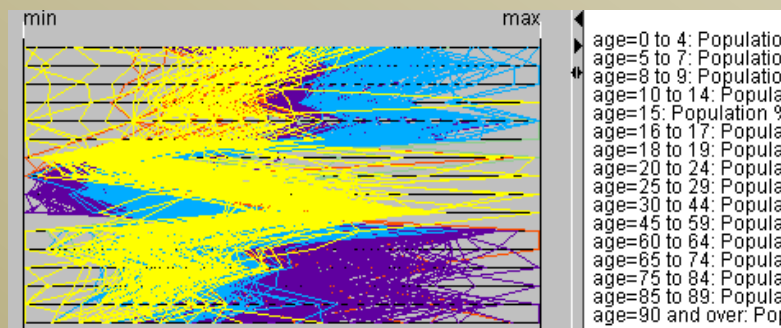
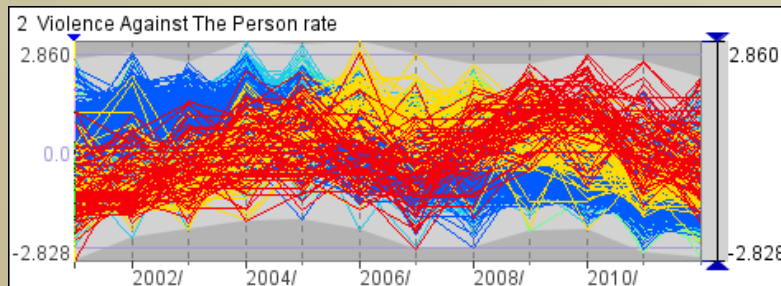
Clustering

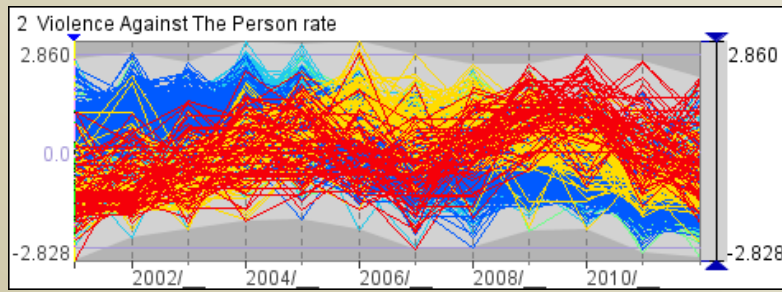
as an instrument for interactive visual analysis



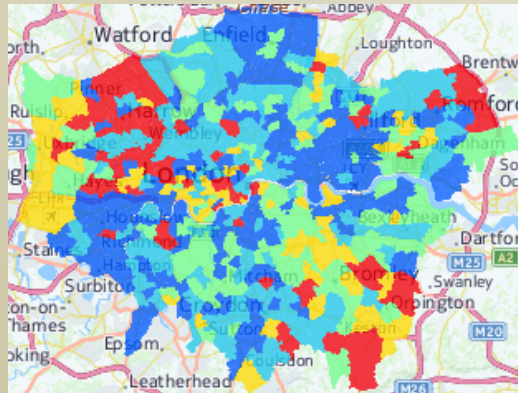
What is clustering?

- 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.

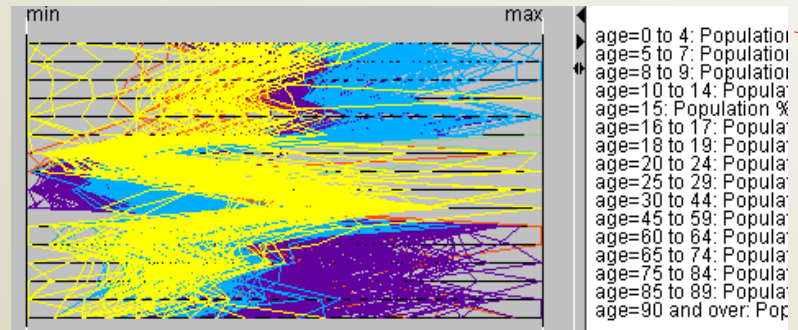
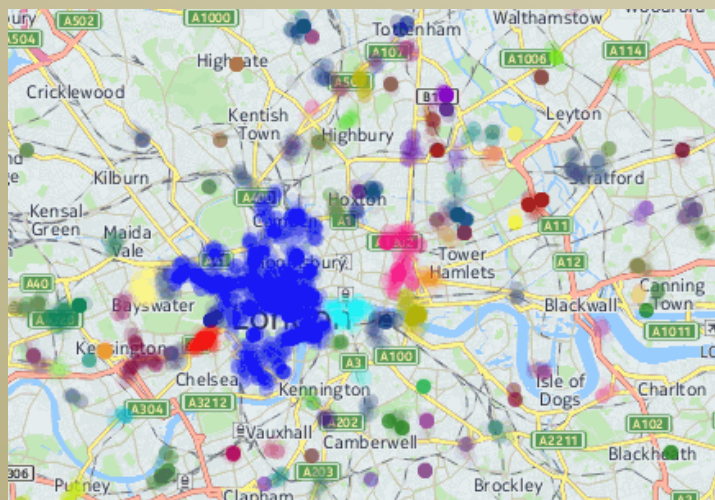




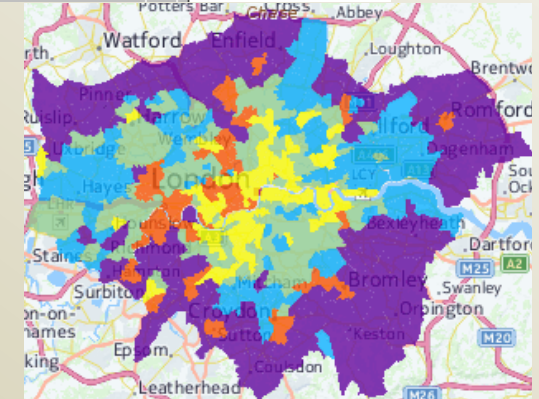
Clusters of similar time series



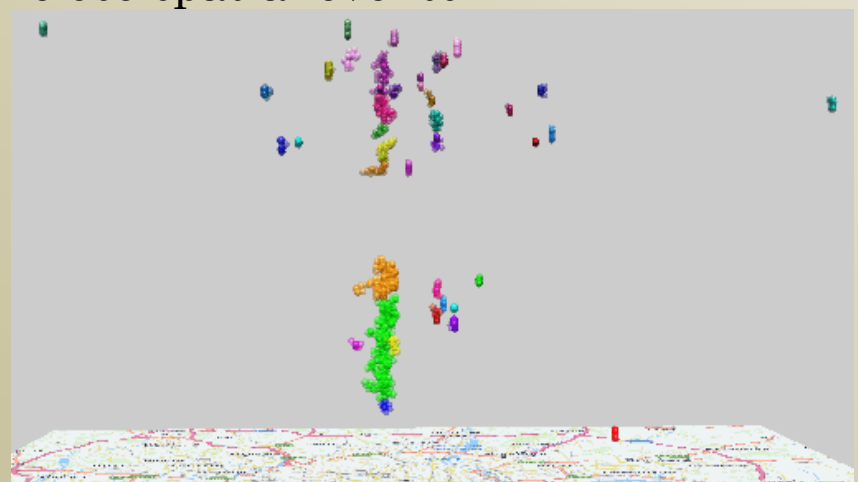
Clusters of spatially close spatial objects



Clusters of similar multi-attribute value combinations



Clusters of spatially and temporally close spatial events





Role of clustering in visual analytics

- Grouping of similar or close items plays an essential role in VA
 - as a tool supporting abstraction: elements → subsets; the subsets may be considered as wholes
 - as a tool to manage large data volumes
 - as a tool to find specific features of interest, e.g., event concentrations
 - as a tool to deal with multiple attributes and multiple time series, which are hard to visualise



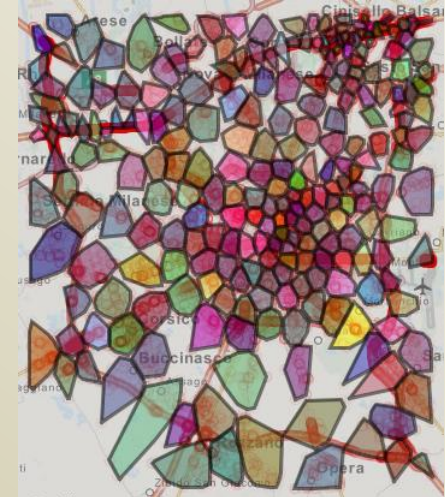
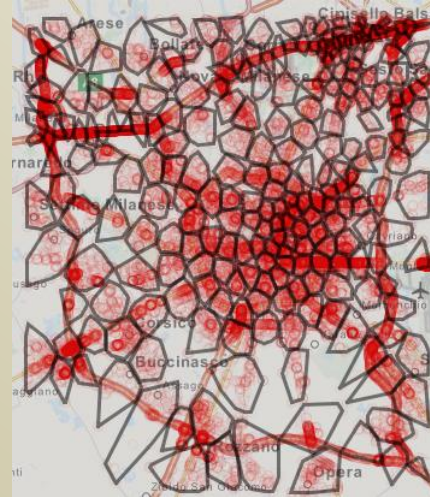
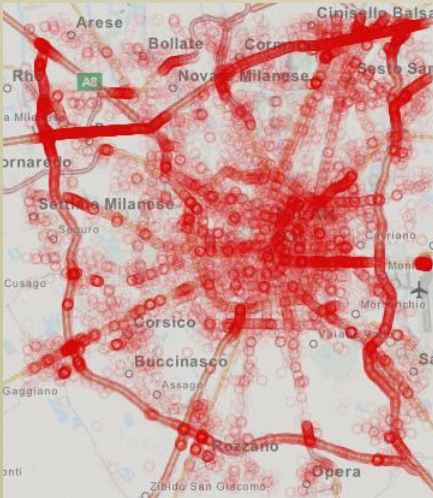
Two major types of clustering

- **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, hierarchical
 - 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”

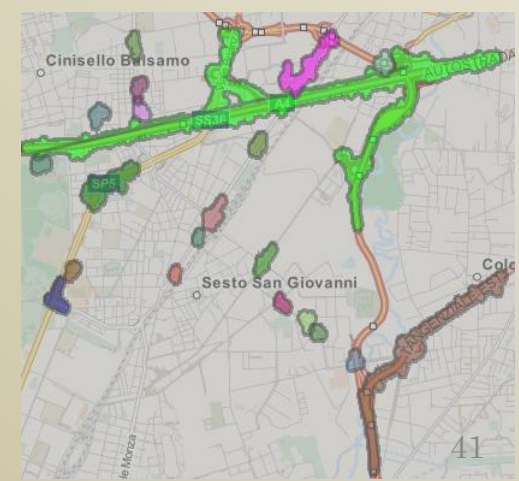
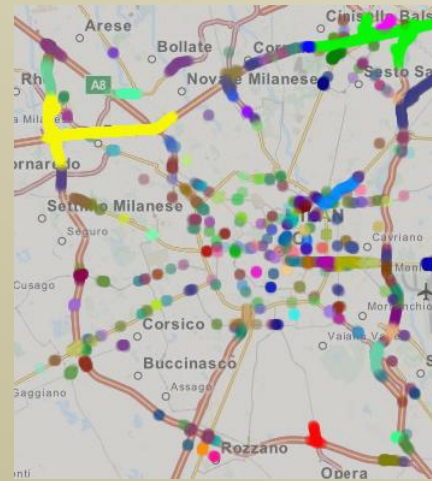
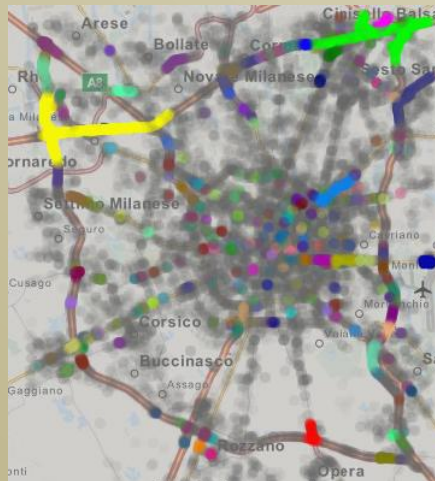


Two major types of clustering: an example

Partition-based: convex clusters including all objects



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





Use of the two types of clustering

- **Partition-based:**

- Typically applied to multiple thematic (non-spatial) attributes or to time series of thematic attributes
- Objective: divide objects into groups such that objects within a group have similar attribute values and differ from the objects in the other groups

- **Density-based:**

- Typically applied to spatial and temporal attributes of spatial or spatio-temporal objects
- Objective: find concentrations of objects in space or in space and time (i.e., groups of objects with close spatial locations and existence times)
 - concentrations of objects may have special meanings;
e.g., spatio-temporal cluster of low speed events \Rightarrow traffic jam



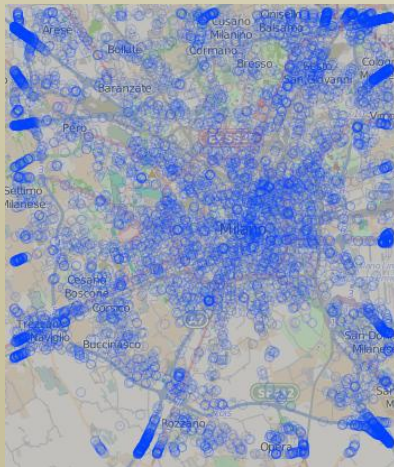
Partition-based clustering



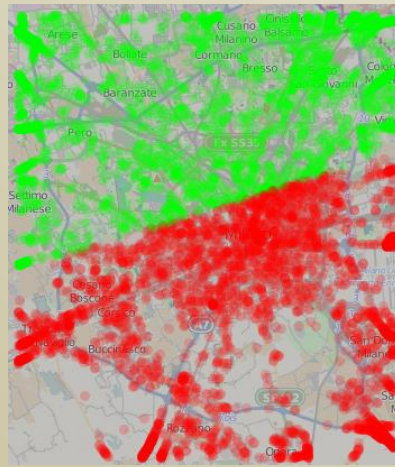
Partition-based clustering (PBC)

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

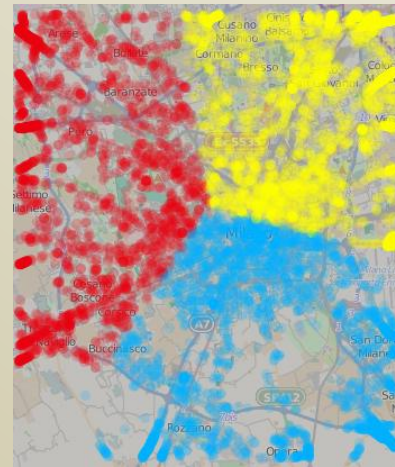
Data: 2D points (X,Y)



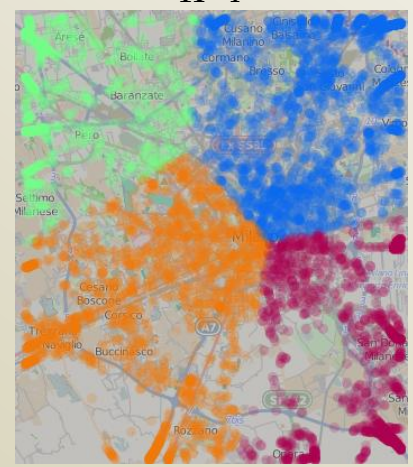
K=2



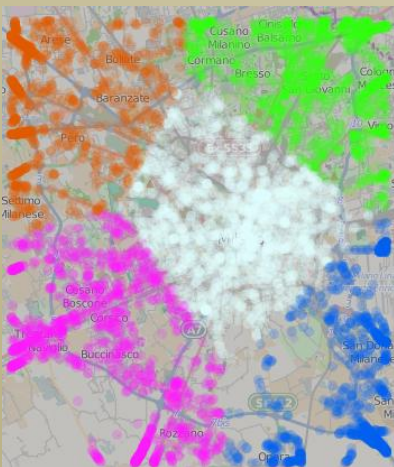
K=3



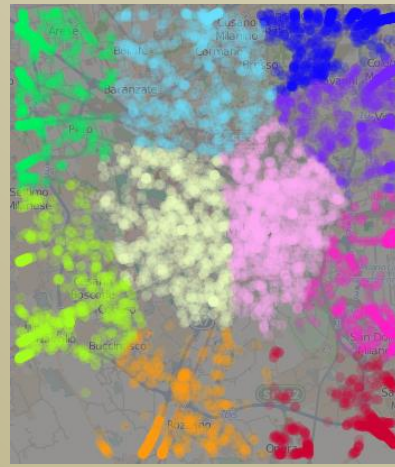
K=4



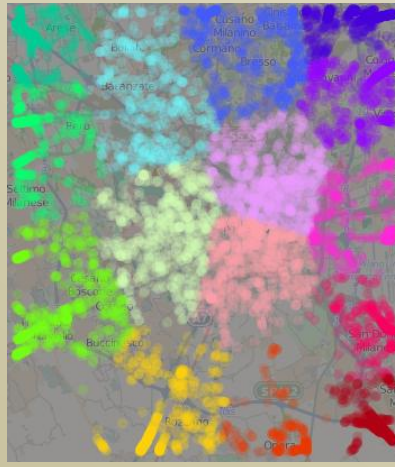
K=5



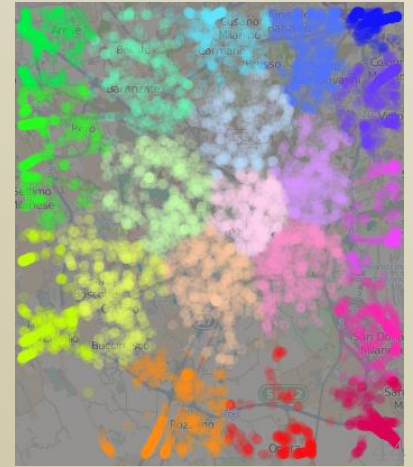
K=10



K=15



K=20

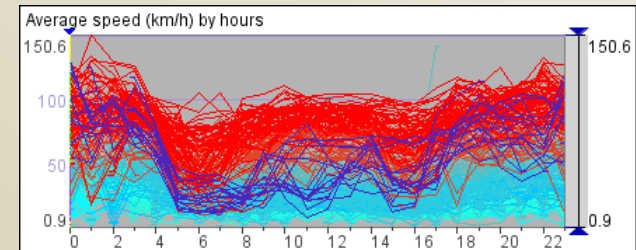
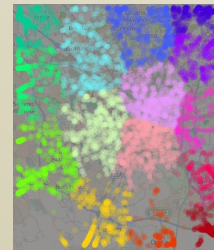




Problem: what value of k to choose?

Generally: for any computational tool, what parameter settings to choose?

- Typically not known in advance
- Computation results (such as clusters) need to be properly visualised and examined
 - Clustering results are often represented by colour-coding, which is applied to different visual objects, depending on the structure of the input data
- The analyst needs to run the tool with different settings and see how the results change
- The analyst 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)



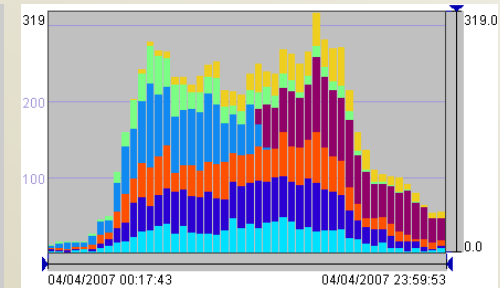
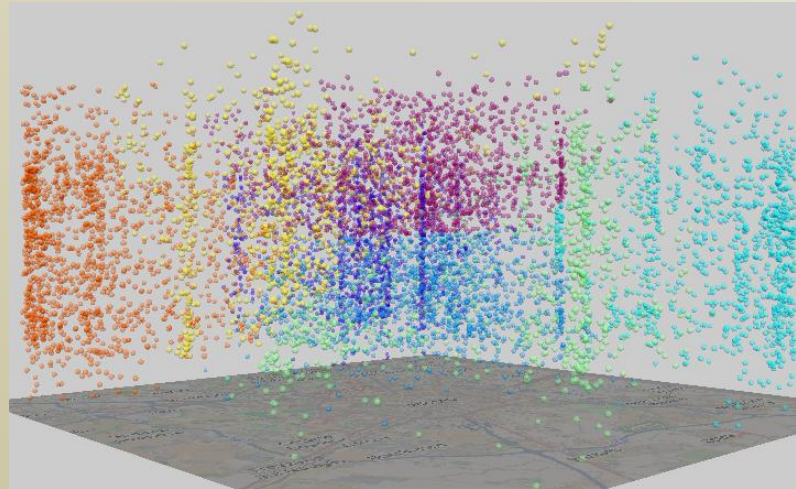
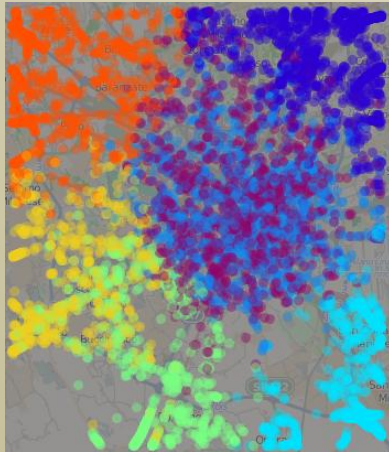


Visualization of clustering results

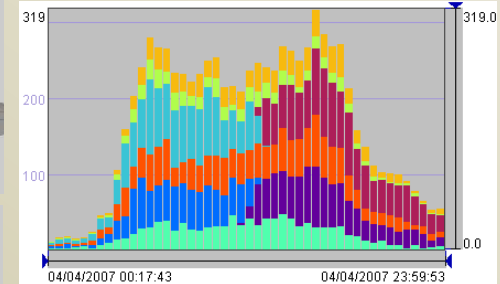
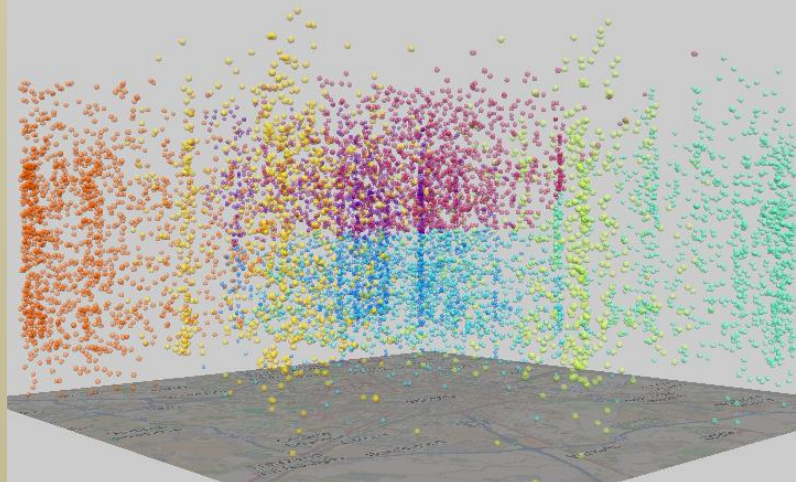
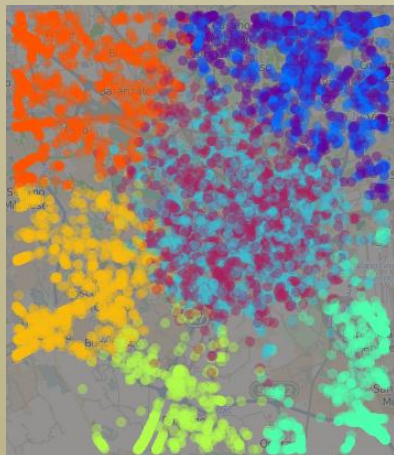
A single display may be not enough

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

K=7



K=8

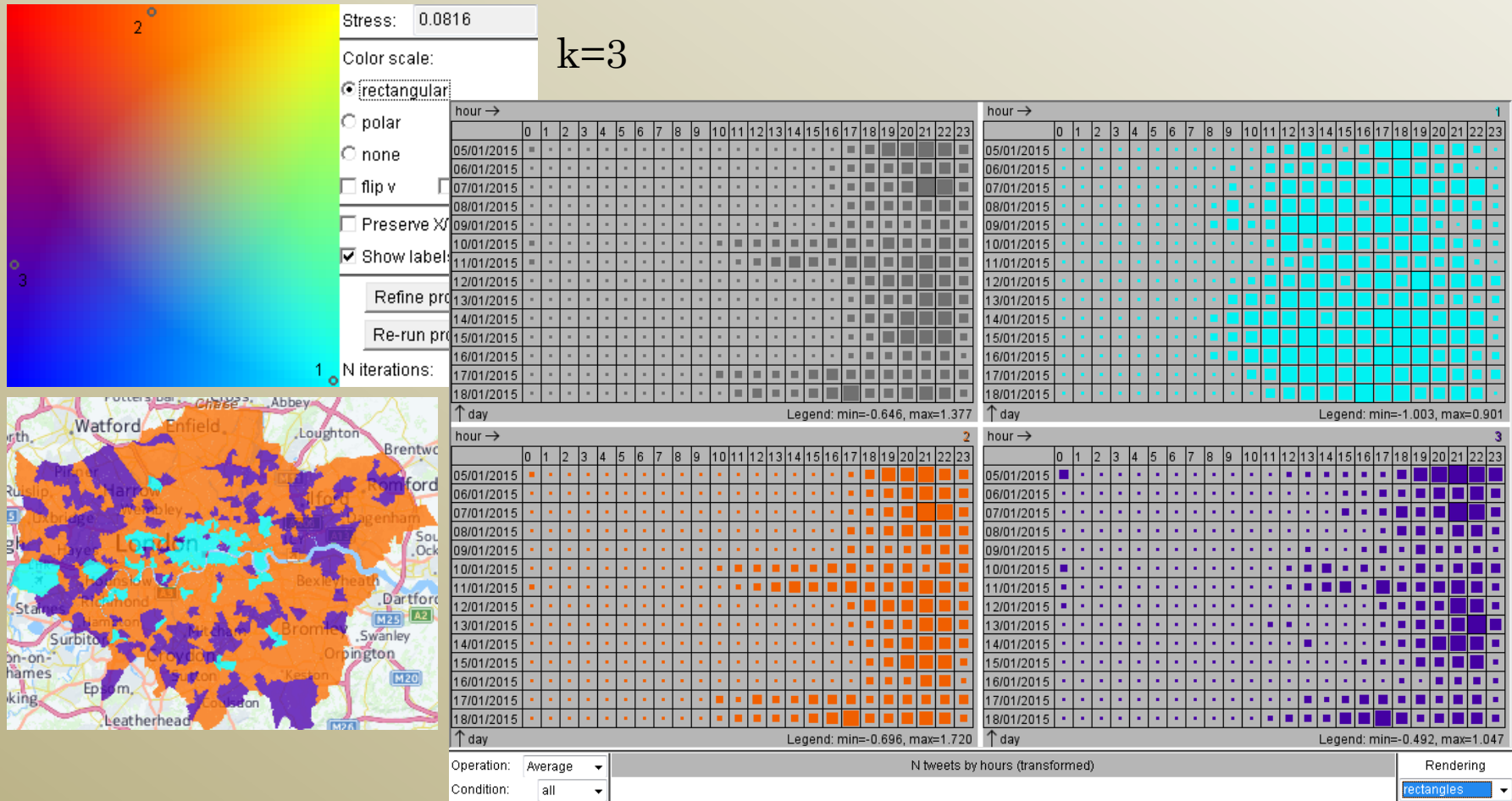


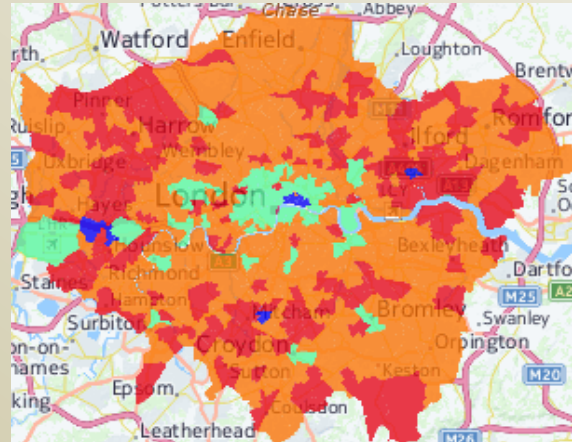
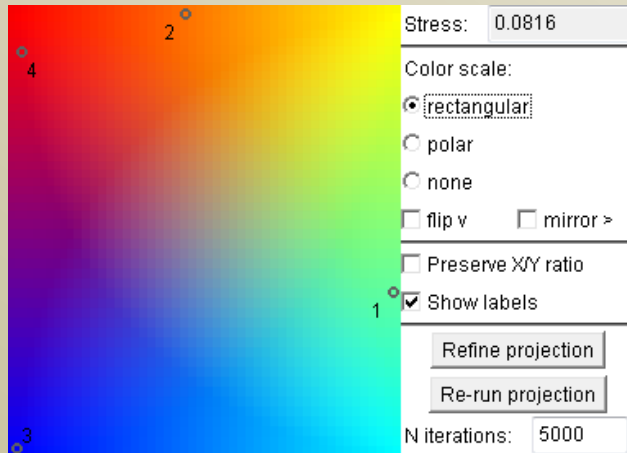
Time-based frequency histograms may show the temporal relations between the clusters in a clearer way than the space-time cube.



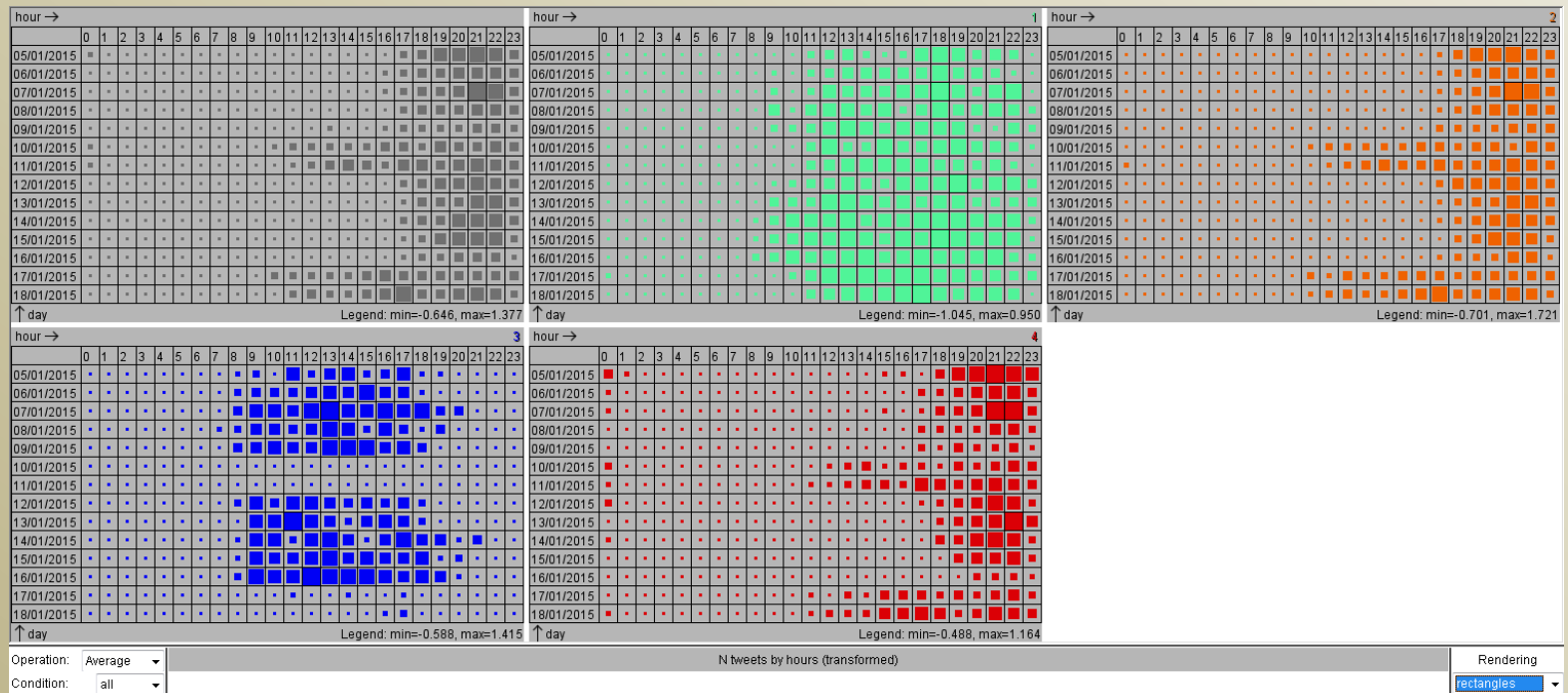
Interactive visual analysis by PBC

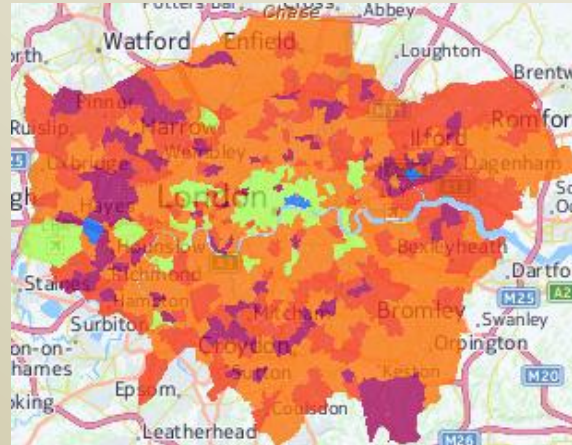
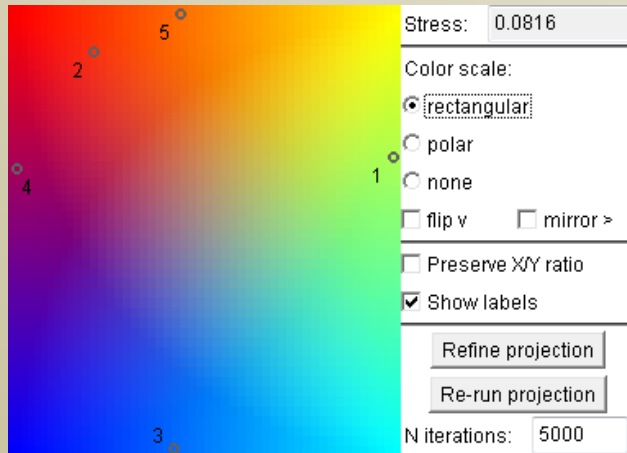
Trying different parameter settings



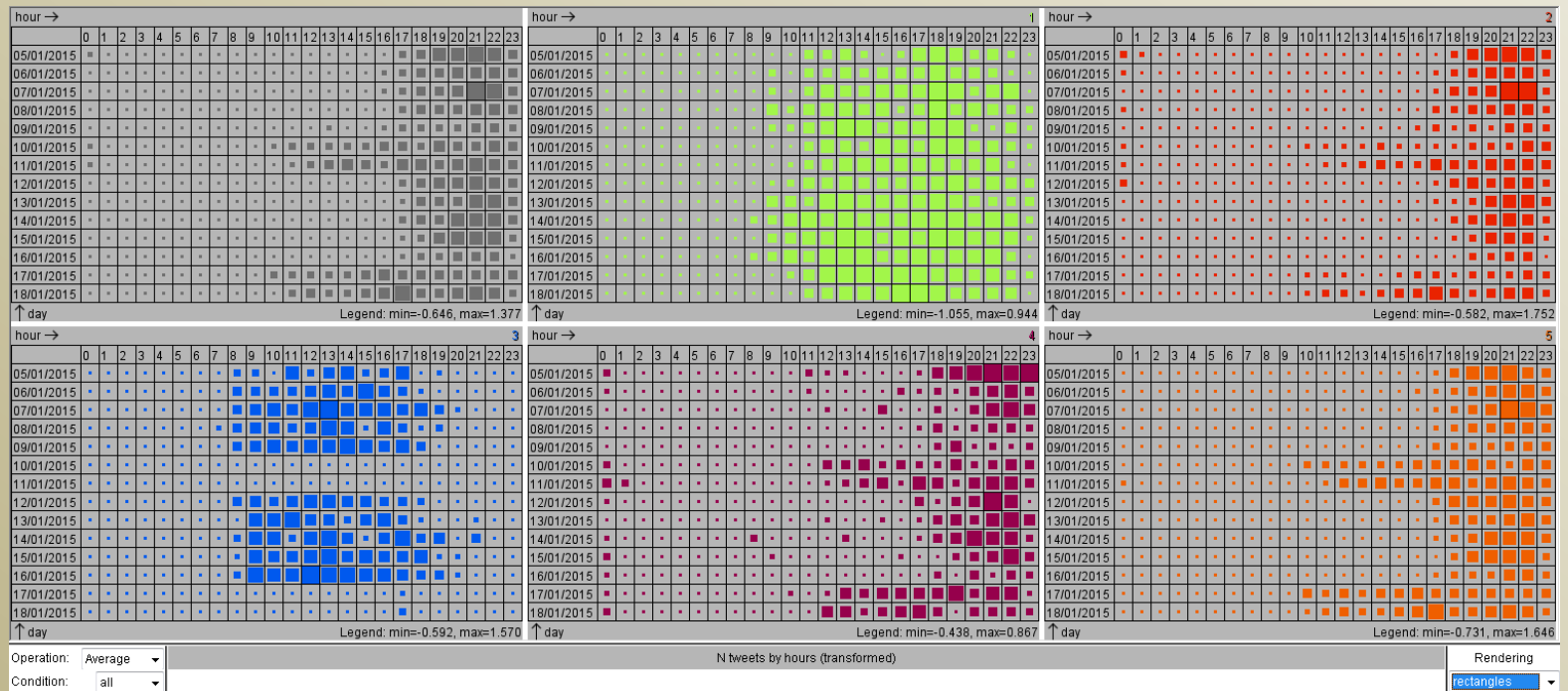


k=4



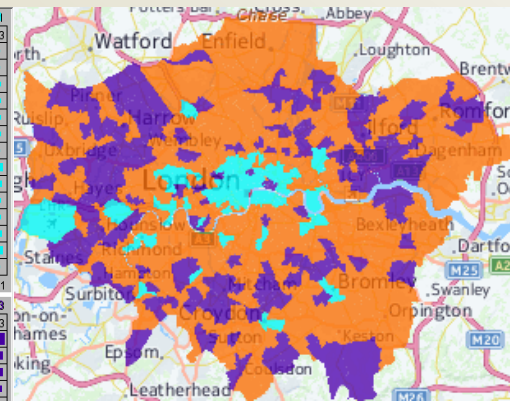
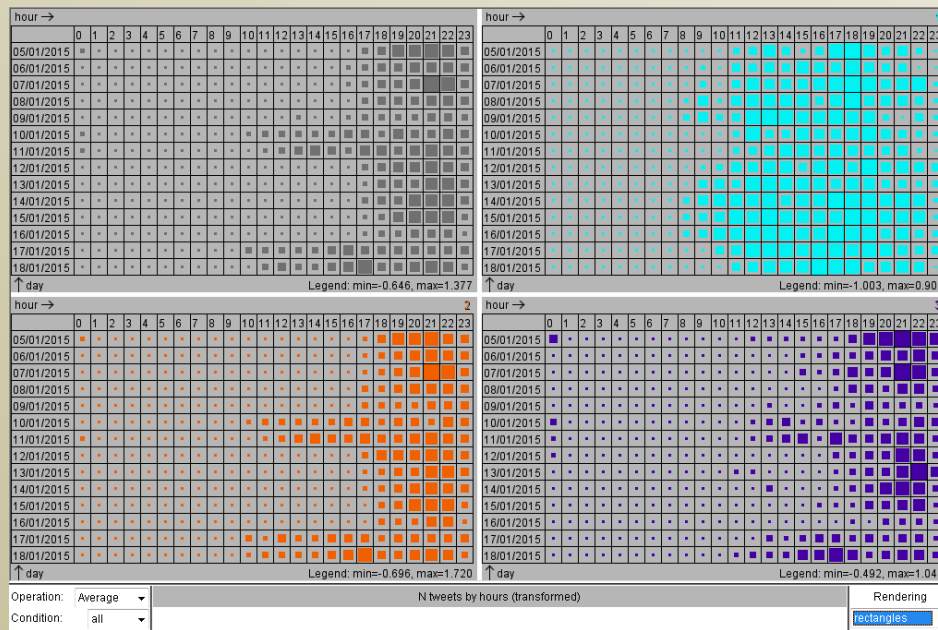


k=5

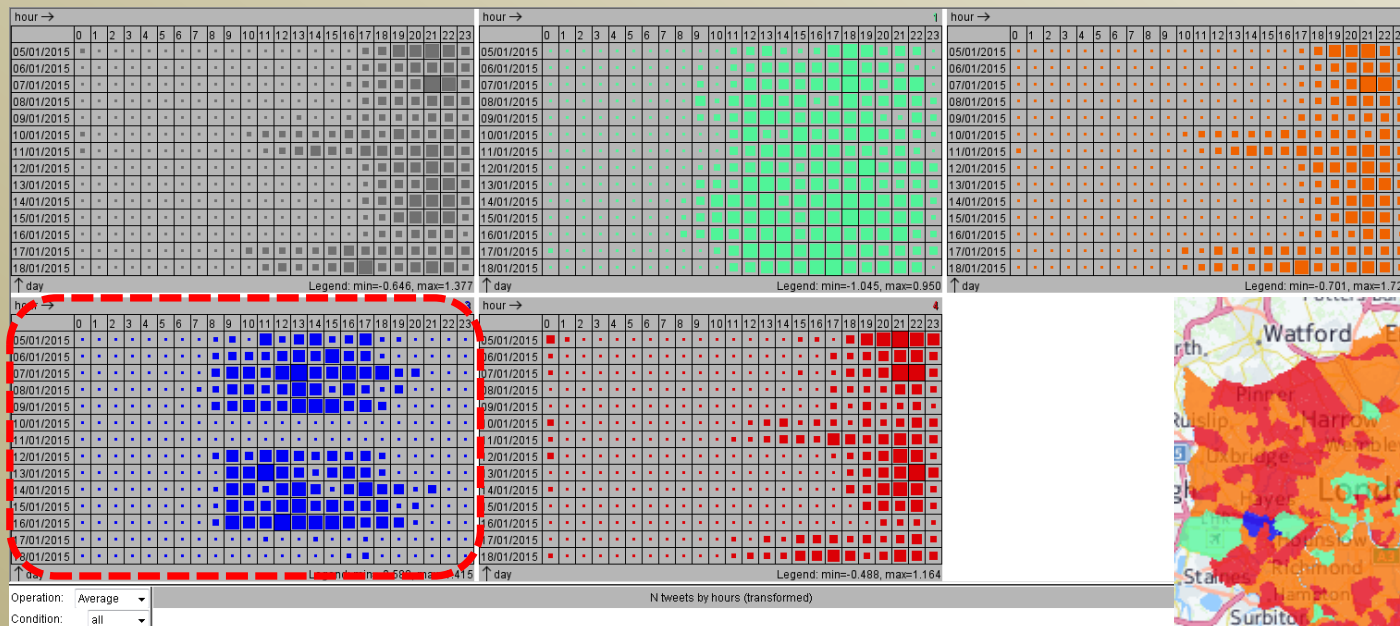




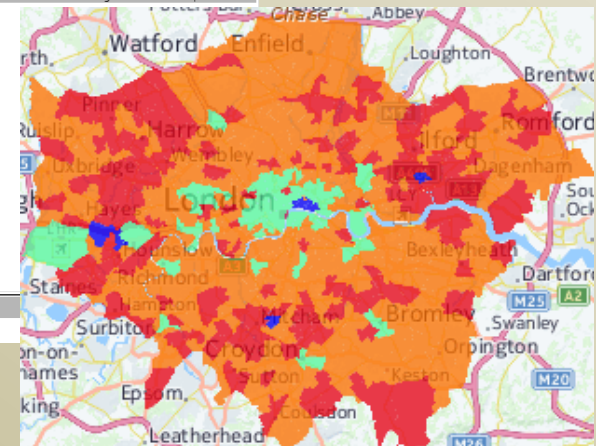
k=3



Increasing k from 3 to 4 uncovers additionally an important and easily interpretable pattern of temporal behaviour.

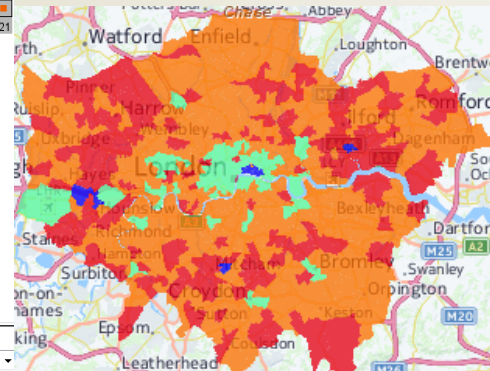
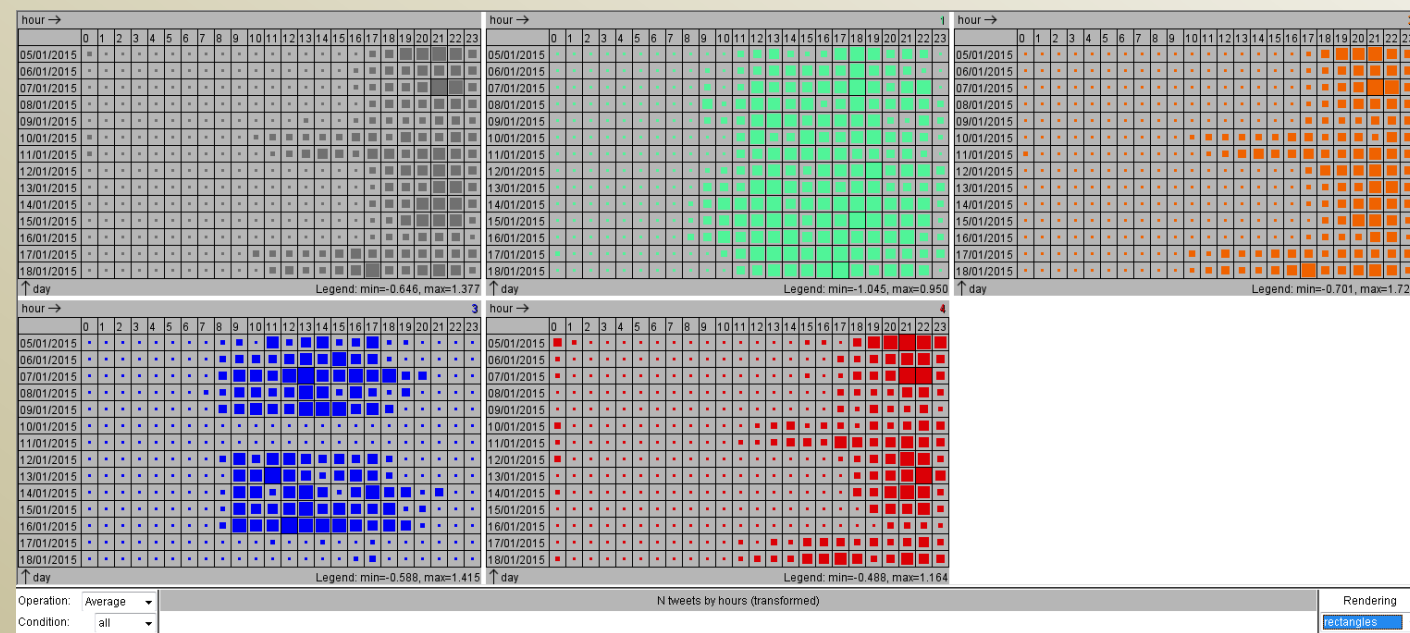


k=4

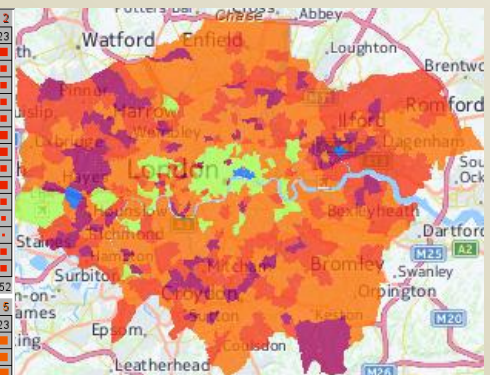




k=4



Increasing k from 4 to 5 does not add any significantly distinct temporal pattern.

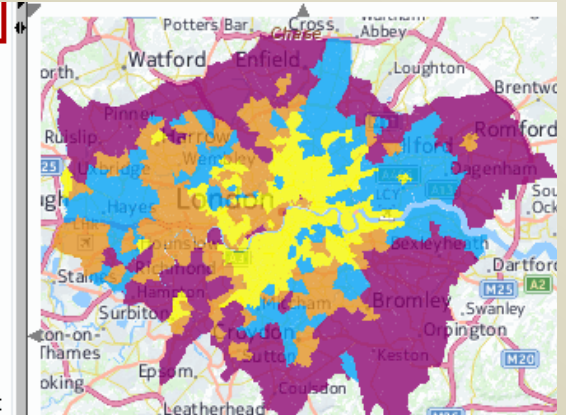
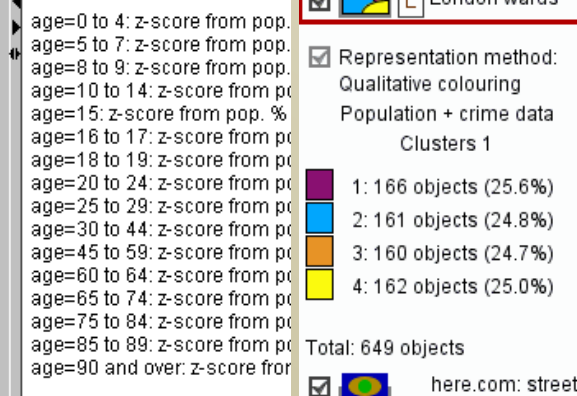
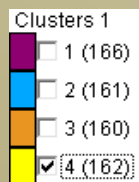
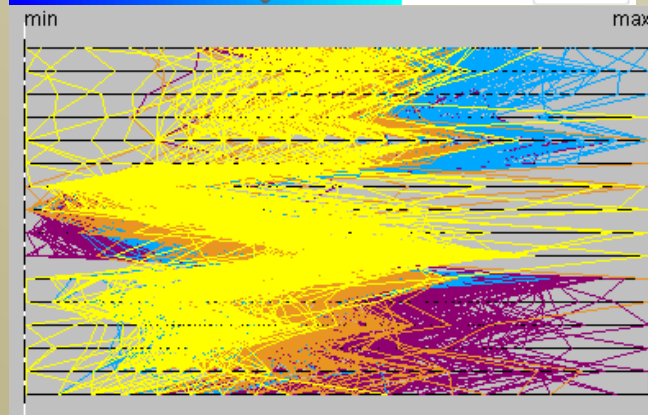
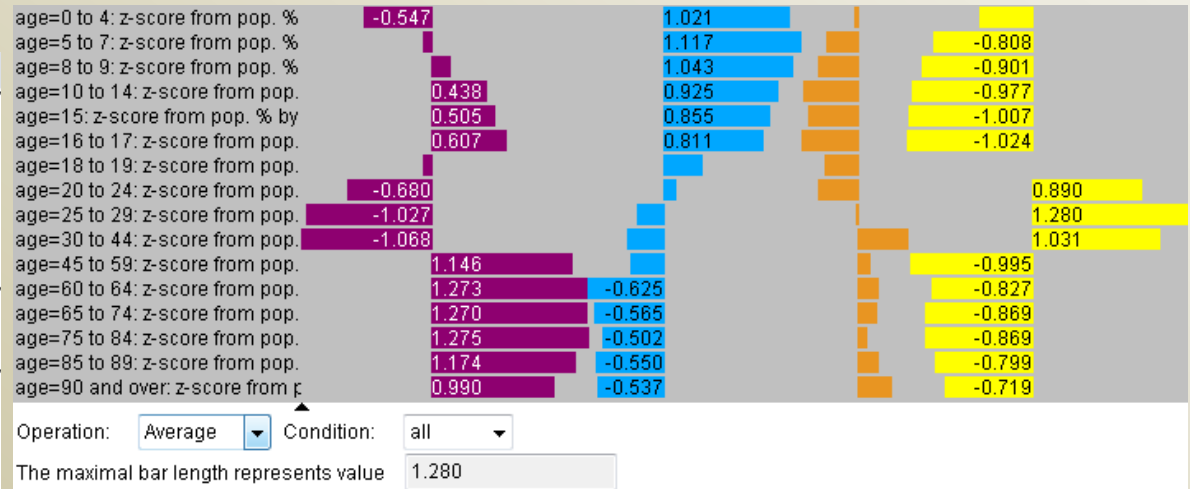
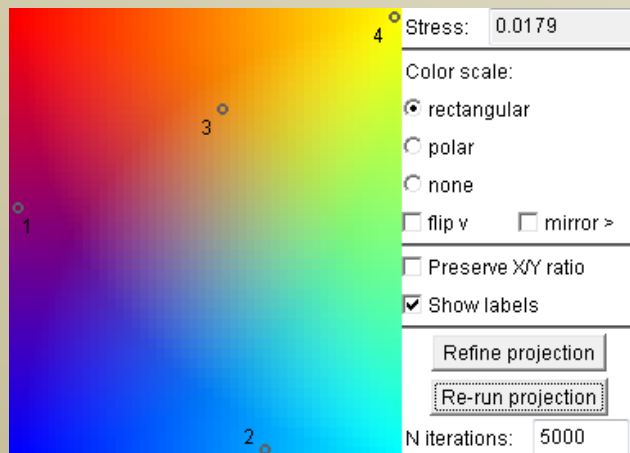


k=5



Interactive progressive clustering

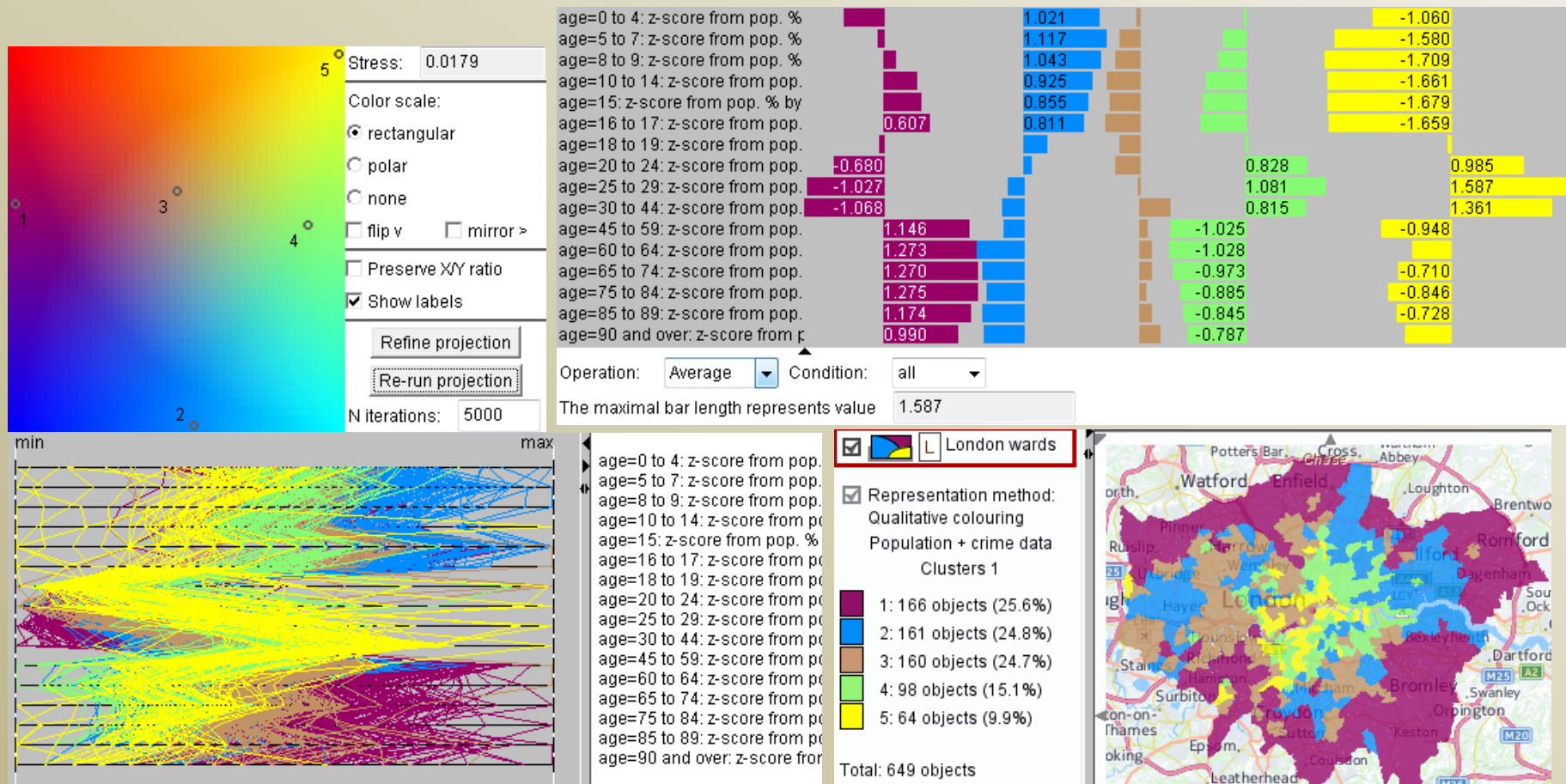
k=4



Next clustering is applied to one or a few of the previously obtained clusters. In this way, selected clusters may be refined.



Interactive progressive clustering



Former cluster 4 has been subdivided into clusters 4 and 5, which differ in the proportions of children.



Partition-based clustering of time steps

by similarity of attribute value distributions over objects or space

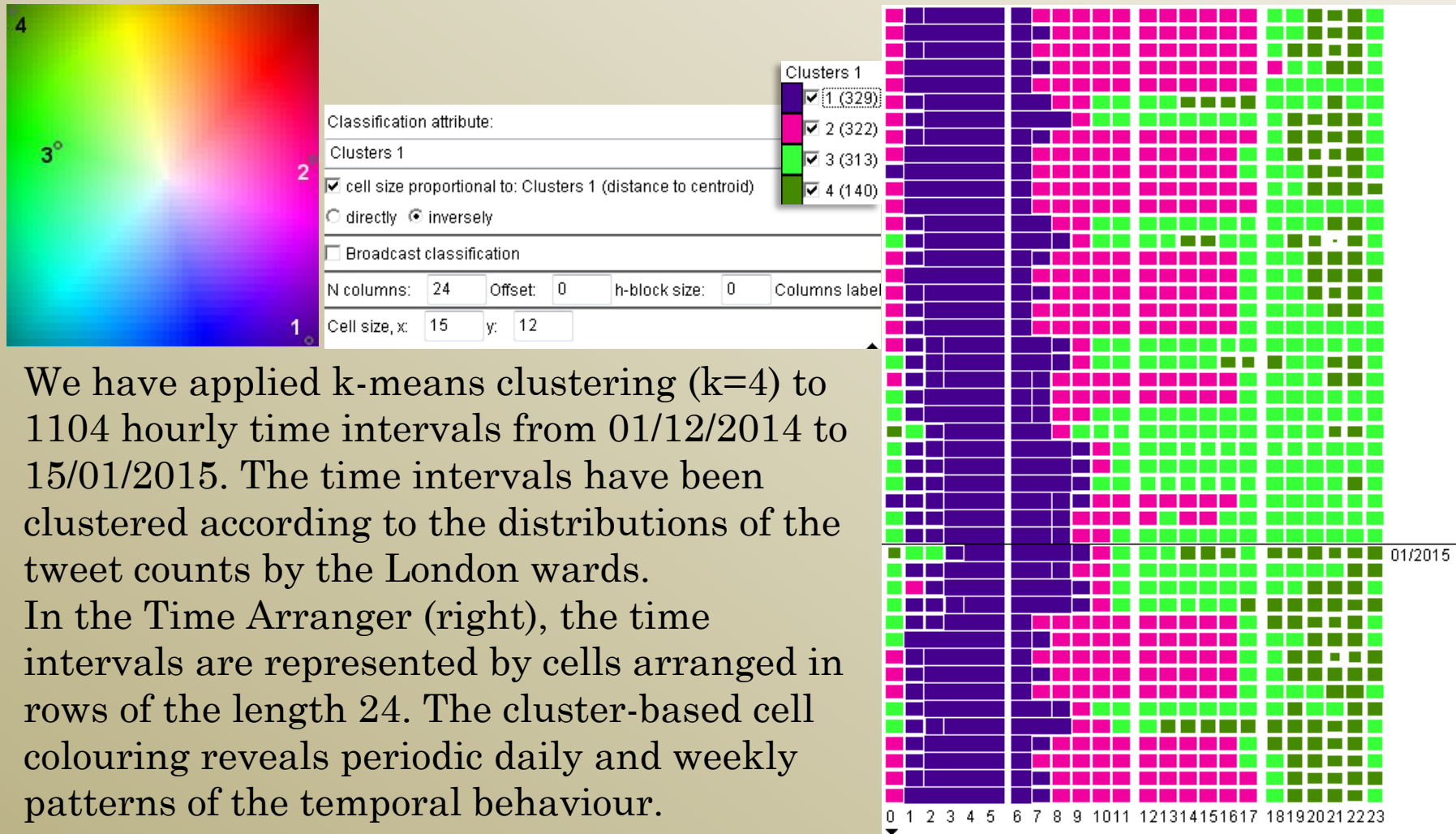
| <input checked="" type="checkbox"/> identifiers | hour=01/12/2014;00: N tweets by hours | hour=01/12/2014;01: N tweets by hours | hour=01/12/2014;02: N tweets by hours | hour=01/12/2014;03: N tweets by hours | hour=01/12/2014;04: N tweets by hours | hour=01/12/2014;05: N tweets by hours | hour=01/12/2014;06: N tweets by hours | hour=01/12/2014;07: N tweets by hours | hour=01/12/2014;08: N tweets by hours |
|---|--|--|--|--|--|--|--|--|--|
| E05000128 Belsize | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| E05000129 Bloomsbury | 10 | 4 | 5 | 4 | 0 | 2 | 2 | 16 | 10 |
| E05000130 Camden Town with Primrose | 3 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 6 |
| E05000131 Canteloves | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 4 |
| E05000132 Fortune Green | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| E05000133 Frogna1 and Fitzjohns | 4 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| E05000134 Gospel Oak | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 |
| E05000135 Hampstead Town | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| E05000136 Haverstock | 3 | 6 | 9 | 0 | 6 | 0 | 0 | 0 | 1 |
| E05000137 Highgate | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| E05000138 Holborn and Covent Garden | 5 | 4 | 1 | 1 | 0 | 0 | 1 | 9 | 22 |
| E05000139 Kentish Town | 3 | 1 | 1 | 0 | 0 | 1 | 0 | 4 | 2 |
| E05000140 Kilburn | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 6 |
| E05000141 King's Cross | 6 | 4 | 2 | 2 | 0 | 0 | 0 | 3 | 9 |
| E05000142 Regent's Park | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 2 | 9 |
| E05000143 St Pancras and Somers Town | 7 | 2 | 2 | 0 | 1 | 0 | 4 | 7 | 15 |
| E05000144 Swiss Cottage | 2 | 0 | 3 | 1 | 0 | 0 | 0 | 2 | 2 |
| E05000145 West Hampstead | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 12 |
| E05000001 Aldersgate | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| E05000005 Bishopsgate | 3 | 5 | 1 | 0 | 0 | 2 | 1 | 3 | 5 |
| E05000015 Cripplegate | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| E05000017 Farringdon Within | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 |
| E05000018 Farringdon Without | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| E05000021 Portsoken | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| E05000022 Queenhithe | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E05000023 Tower | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| E05000231 Brownswood | 7 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 2 |
| E05000232 Cazenove | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 0 |
| E05000233 Chatham | 1 | 4 | 0 | 8 | 2 | 0 | 0 | 0 | 0 |
| E05000234 Clissold | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| E05000235 Dalston | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 3 |
| E05000236 De Beauvoir | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| E05000237 Hackney Central | 40 | 41 | 32 | 0 | 0 | 0 | 1 | 0 | 8 |
| E05000238 Hackney Downs | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| E05000239 Haggerston | 7 | 3 | 2 | 0 | 1 | 0 | 0 | 4 | 18 |
| E05000240 Hoxton | 2 | 9 | 0 | 4 | 9 | 5 | 0 | 1 | 3 |

Object- or space-referenced time series can be organized in a table so that rows correspond to one referrer and columns to the other. Clustering can be applied to rows or to columns.



Partition-based clustering of time steps

by similarity of attribute value distributions over objects or space

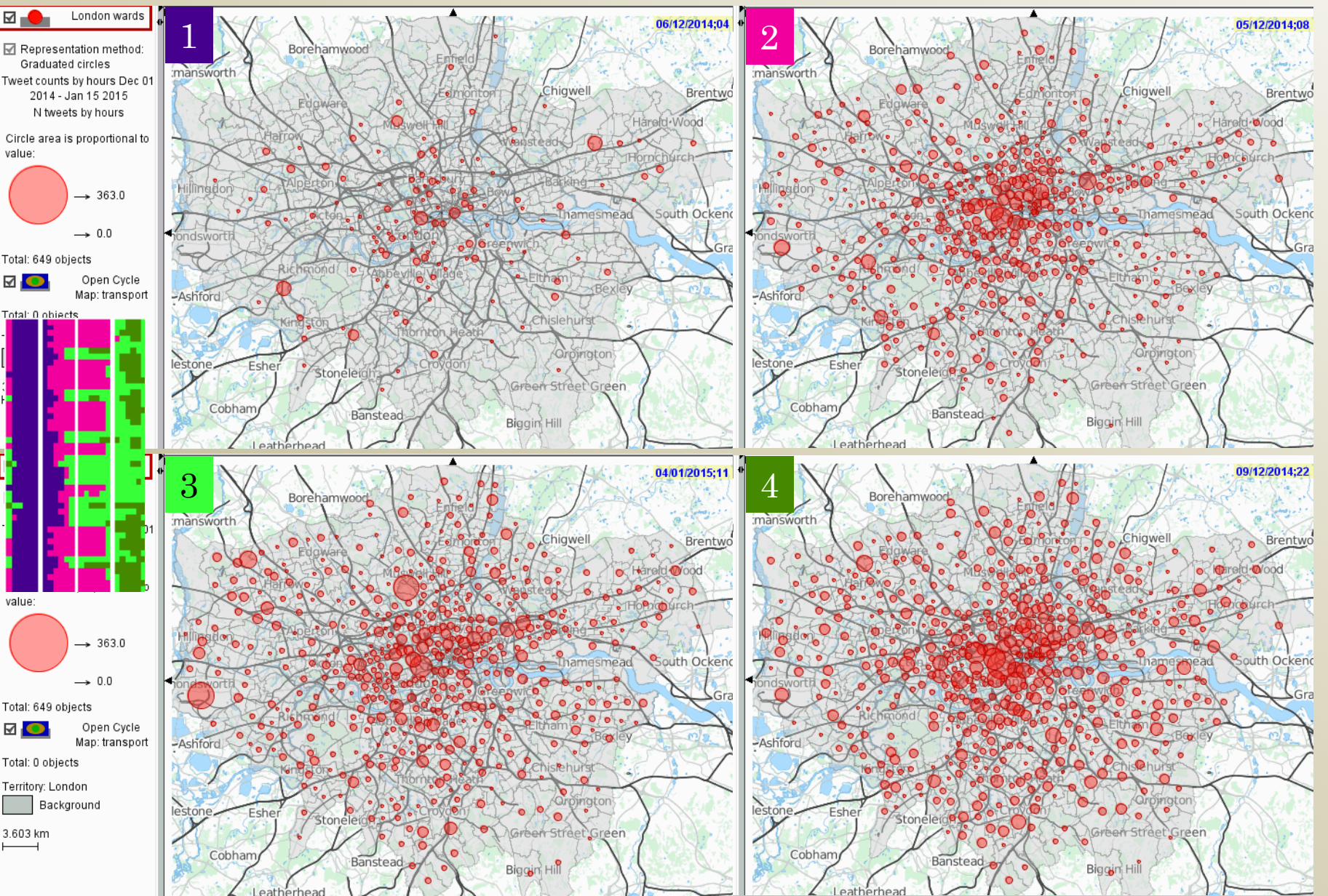


We have applied k-means clustering ($k=4$) to 1104 hourly time intervals from 01/12/2014 to 15/01/2015. The time intervals have been clustered according to the distributions of the tweet counts by the London wards.

In the Time Arranger (right), the time intervals are represented by cells arranged in rows of the length 24. The cluster-based cell colouring reveals periodic daily and weekly patterns of the temporal behaviour.

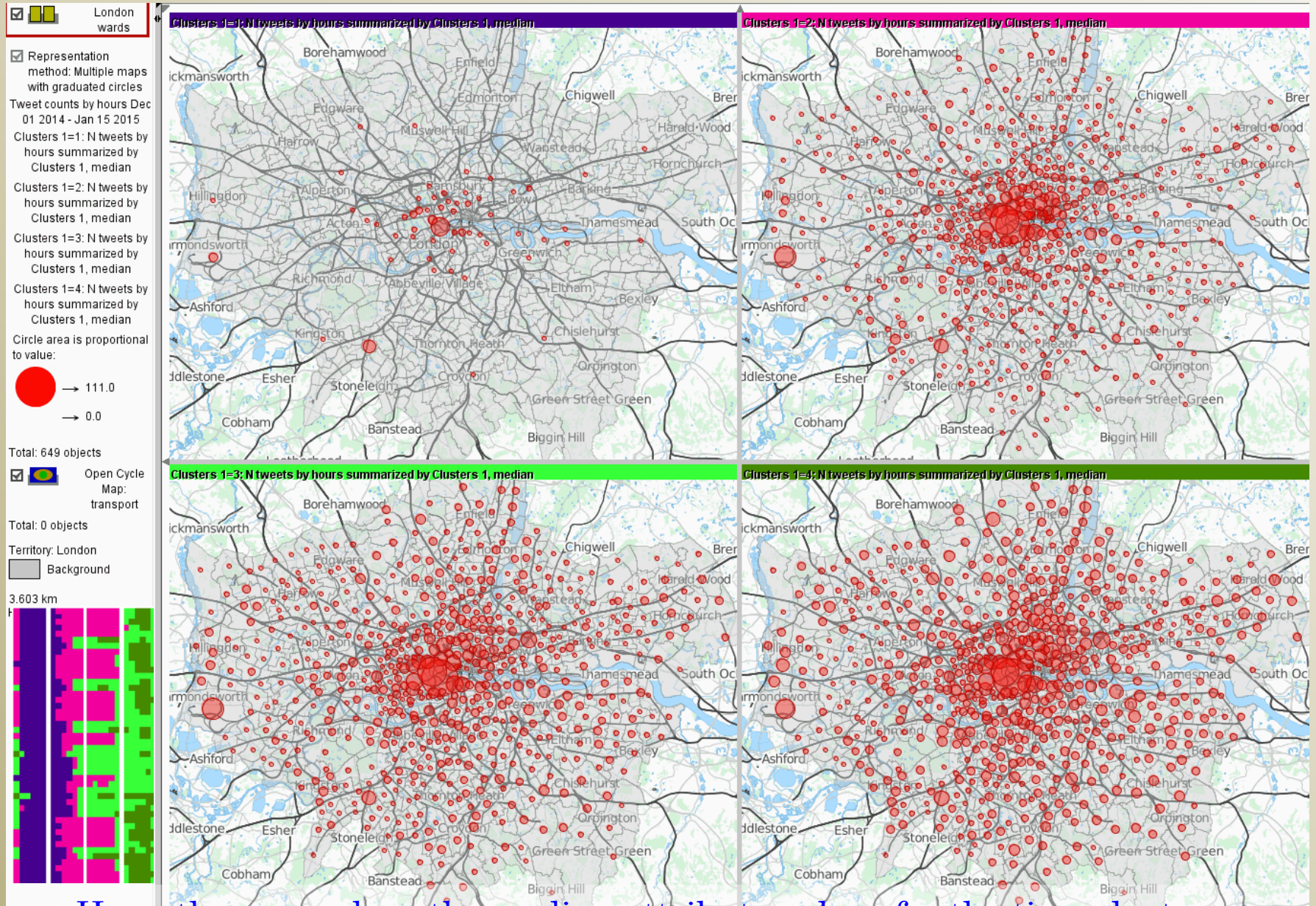



Cluster representatives may be interactively selected for viewing.





The value distributions can be summarized by the clusters of time steps.



☒  London wards

☒ Representation method: Multiple maps with graduated circles

Tweet counts by hours Dec 01 2014 - Jan 15 2015

Data transformation: difference to attribute

Clusters 1=2: N tweets by hours summarized by Clusters 1, median


Clusters 1=1: N tweets by hours summarized by Clusters 1, median

Clusters 1=2: N tweets by hours summarized by Clusters 1, median

Clusters 1=3: N tweets by hours summarized by Clusters 1, median

Clusters 1=4: N tweets by hours summarized by Clusters 1, median

Circle area is proportional to value:

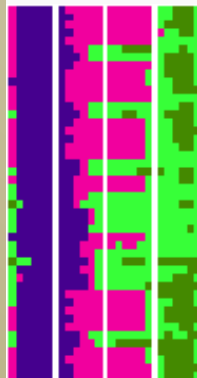
 → 26.00

→ 0.00

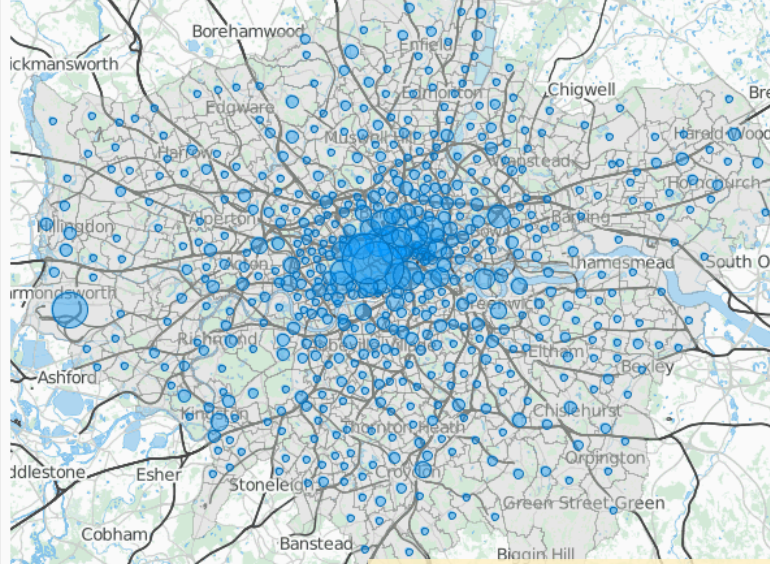
 → -60.00

Total: 649 objects

☒  Open Cycle Map: transport



Clusters 1=1: N tweets by hours summarized by Clusters 1, median



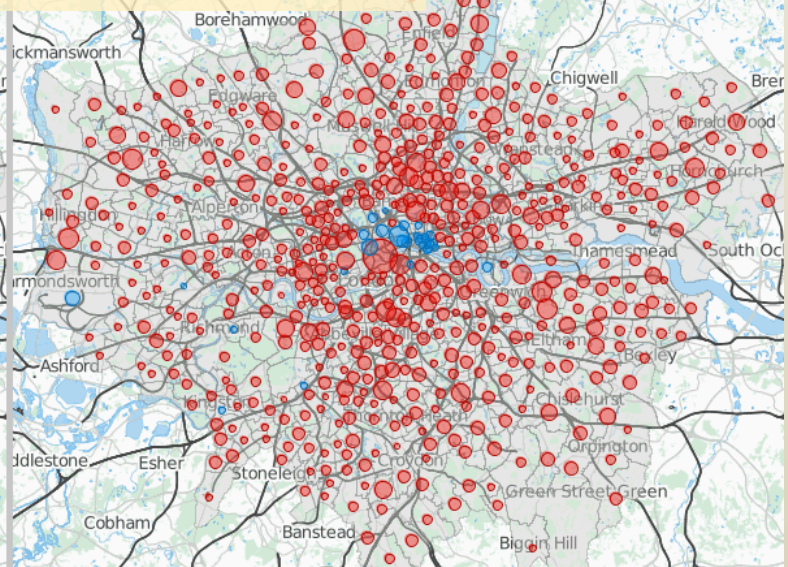
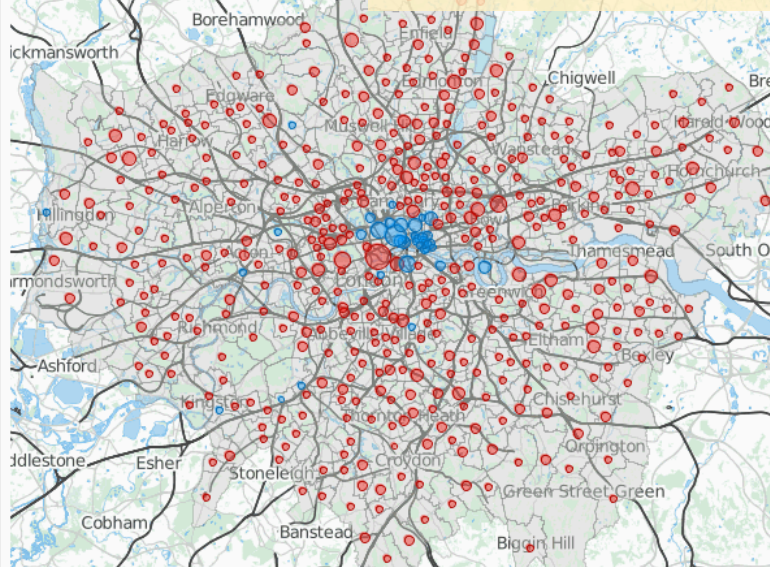
Clusters 1=2: N tweets by hours summarized by Clusters 1, median



Clusters 1=3: N tweets by hours summarized by

Visual comparison to time cluster 2

Clusters 1, median



For better seeing the differences, the summary (here: median) values of one time cluster (here: 2) are subtracted from the values of the other clusters.



London
wards

Representation
method: Multiple maps
with graduated circles

Tweet counts by hours Dec
01 2014 - Jan 15 2015

Data transformation:

difference to attribute

Clusters 1=1: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=2: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=3: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=4: N tweets by
hours summarized by

Clusters 1, median

Circle area is proportional
to value:

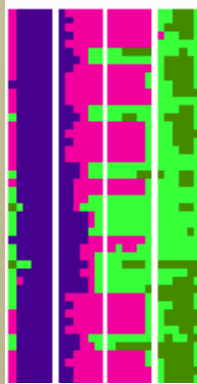
→ 12.00

→ 0.00

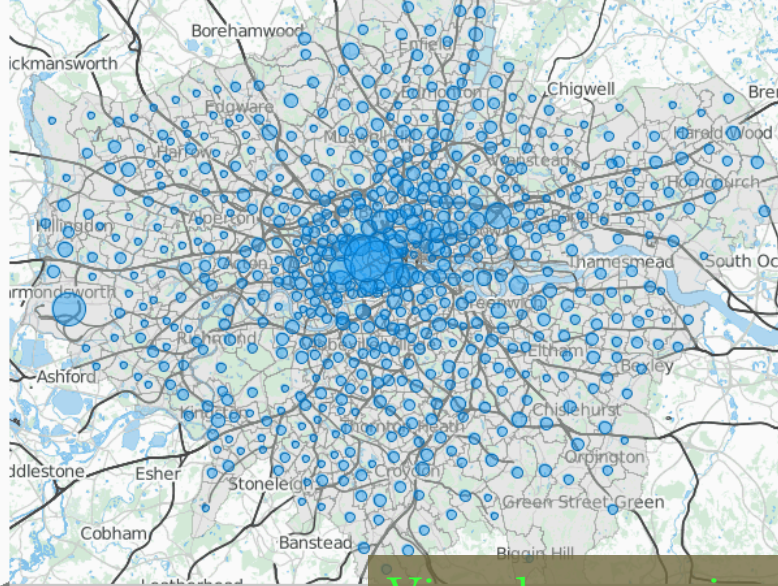
→ -74.00

Total: 649 objects

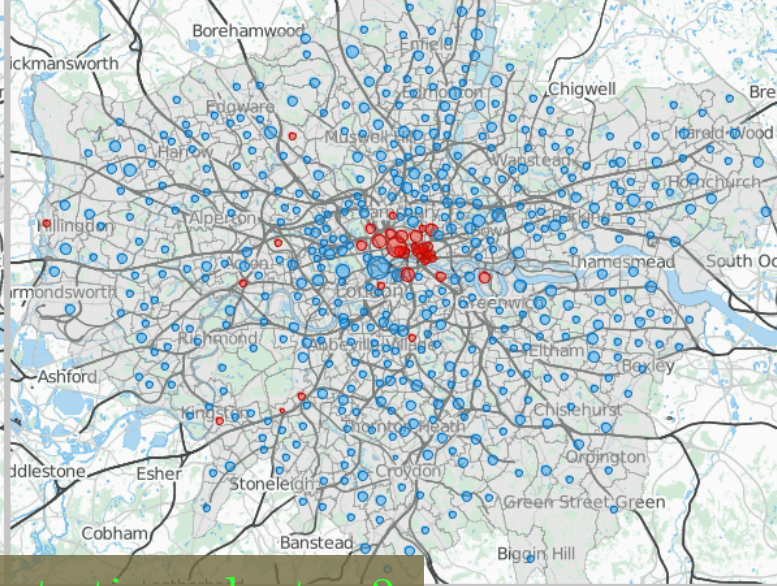
Open Cycle
Map:
transport



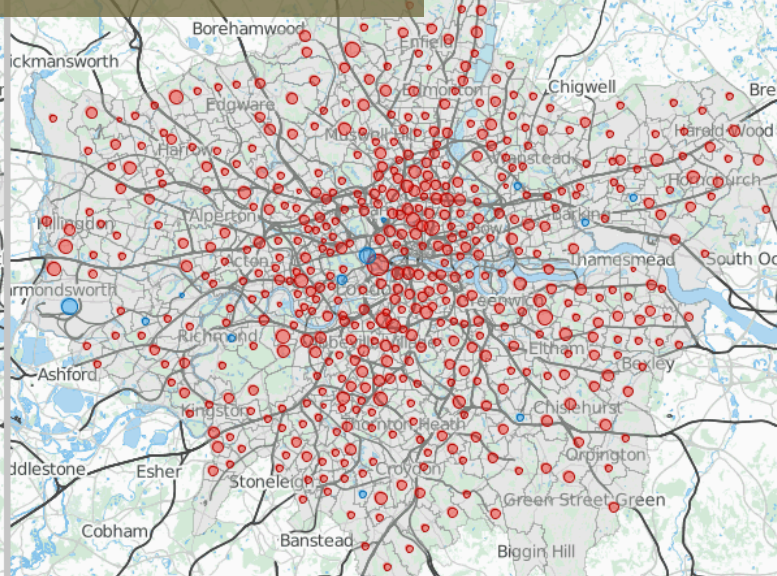
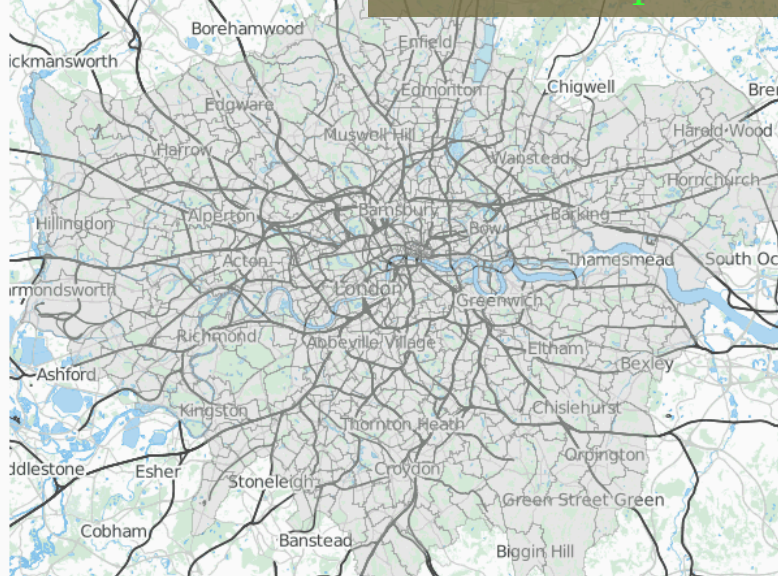
Clusters 1=1: N tweets by hours summarized by Clusters 1, median



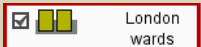
Clusters 1=2: N tweets by hours summarized by Clusters 1, median



Clusters 1=3: N tweets by hours summarized by Clusters 1, median



Visual comparison to time cluster 3



London
wards

Representation
method: Multiple maps
with graduated circles

Tweet counts by hours Dec
01 2014 - Jan 15 2015

Data transformation:
difference to attribute

Clusters 1=1: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=1: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=2: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=3: N tweets by
hours summarized by

Clusters 1, median

Clusters 1=4: N tweets by
hours summarized by

Clusters 1, median

Circle area is proportional
to value:

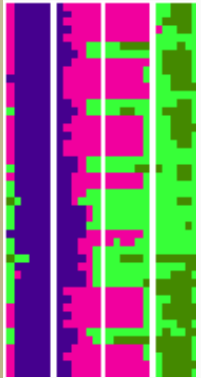
→ 9.00

→ 0.00

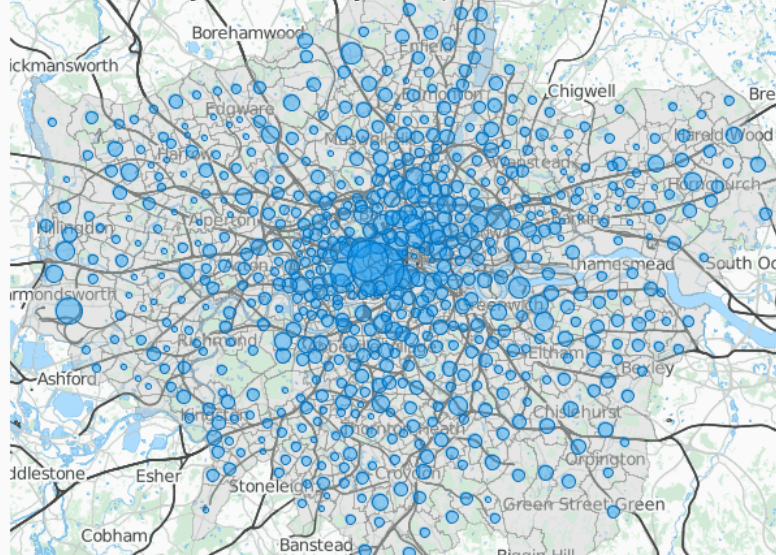
→ -86.00

Total: 649 objects

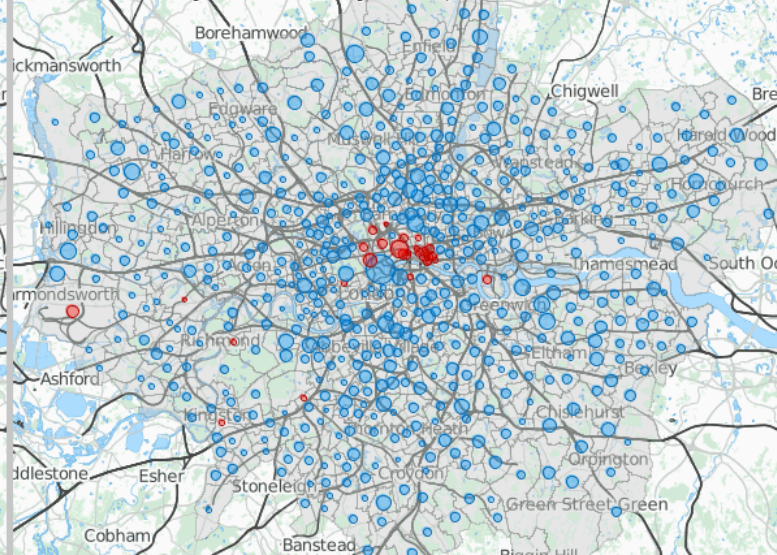
Open Cycle
Map:
transport



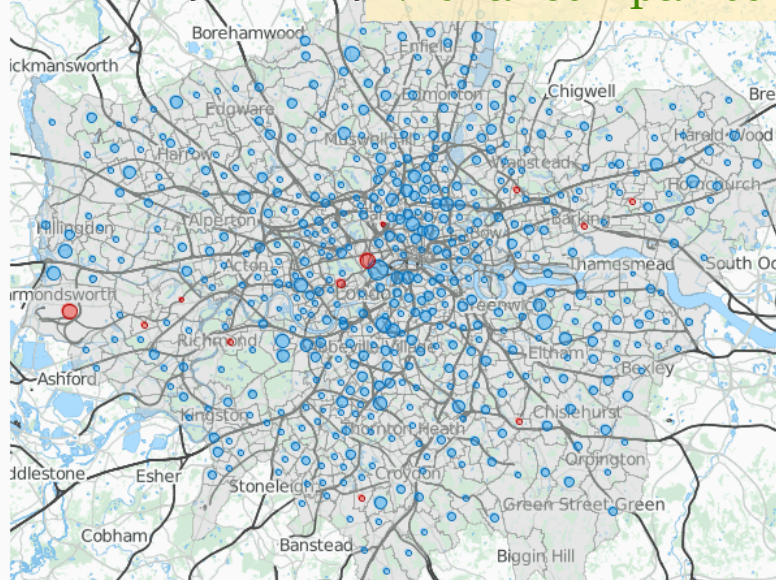
Clusters 1=1: N tweets by hours summarized by Clusters 1, median



Clusters 1=2: N tweets by hours summarized by Clusters 1, median



Clusters 1=3: N tweets by hours summarized by



Visual comparison to time cluster 3

Clusters 1, median





Partition-based clustering: a summary

- Divides references (objects, places, time steps, ...) into groups by similarity of the corresponding attribute values:
 - Multiple attributes (multidimensional data)
 - Time series of values of one or more attributes
 - Distributions of attribute values over a set of objects or places
- Grouping
 - ✓ reduces and simplifies the data to analyse
 - ✓ facilitates abstraction (data may be summarized by groups and represented in a compact way)
 - ☹ but involves large information losses
- To decrease the information loss, interact with the clustering tool
 - Vary the parameter settings and compare different groupings
 - Examine internal variance and refine clusters by progressive clustering



Visualisation of clustering results

- Distinct colours are assigned to the clusters and used for colouring display elements representing cluster members in diverse displays:
 - time graph, parallel coordinates plot, map, time arranger (*more generally, matrix layout of pixels*), scatter plot, etc.
- Aggregation of the attribute values by the clusters
 - allows more compact and/or less cluttered visual representation of the clustering results
 - facilitates comparison and interpretation of the clusters
 - facilitates abstraction
- Visualisation of cluster summaries (aggregated attribute values):
 - bar diagram (bars represent clusters), segmented histogram, multiple histograms (also 2D), multiple maps.
- Individual data displays are nevertheless needed for assessing the internal variation in the clusters.



Partition-based clustering:

use for object- or space-referenced time series

- PBC can help in studying both aspects of the complex overall behaviours $B_{O \times T}(A(o, t))$ and $B_{S \times T}(A(s, t))$.
- To study $B_O(B_T(A(o, t)))$ or $B_S(B_T(A(s, t)))$ (distribution of the temporal variations over the set of objects or places):
 - Cluster the objects or places by the similarity of the corresponding time series, i.e., apply clustering to time series.
- To study $B_T(B_O(A(o, t)))$ or $B_T(B_S(A(s, t)))$:
 - Cluster the time steps by the similarity of the corresponding attribute values associated with the objects or places, i.e., apply clustering to the combinations of attribute values in different time steps.



When data with two referrers ($X \times Y$, such as objects \times time, space \times time, etc.) are organized in a table (matrix) with the rows and columns corresponding to these X and Y , the two aspects of the overall behaviour can be studied by applying partition-based clustering to the table rows and to the table columns.

- Application to rows: $B_X(B_Y(A(x,y)))$
- Application to columns: $B_Y(B_X(A(x,y)))$

Referrer 1 (X) \downarrow Referrer 2 (Y) \rightarrow

| <input checked="" type="checkbox"/> identifiers | hour=01/12/2014;00: N tweets by hours | hour=01/12/2014;01: N tweets by hours | hour=01/12/2014;02: N tweets by hours | hour=01/12/2014;03: N tweets by hours | hour=01/12/2014;04: N tweets by hours | hour=01/12/2014;05: N tweets by hours | hour=01/12/2014;06: N tweets by hours | hour=01/12/2014;07: N tweets by hours | hour=01/12/2014;08: N tweets by hours |
|---|--|--|--|--|--|--|--|--|--|
| E05000128 Belsize | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| E05000129 Bloomsbury | 10 | 4 | 5 | 4 | 0 | 2 | 2 | 16 | 10 |
| E05000130 Camden Town with Primrose | 3 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 6 |
| E05000131 Cantelowes | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 4 |
| E05000132 Fortune Green | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| E05000133 Froggnal and Fitzjohns | 4 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| E05000134 Gospel Oak | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 |
| E05000135 Hampstead Town | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| E05000136 Haverstock | 3 | 6 | 9 | 0 | 0 | 6 | 0 | 0 | 1 |
| E05000137 Highgate | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| E05000138 Holborn and Covent Garden | 5 | 4 | 1 | 1 | 0 | 0 | 1 | 9 | 22 |
| E05000139 Kentish Town | 3 | 1 | 1 | 0 | 0 | 1 | 0 | 4 | 2 |
| E05000140 Kilburn | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 6 |
| E05000141 King's Cross | 6 | 4 | 2 | 2 | 0 | 0 | 0 | 3 | 9 |
| E05000142 Regent's Park | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 2 | 9 |
| E05000143 St Pancras and Somers Town | 7 | 2 | 2 | 0 | 1 | 0 | 4 | 7 | 15 |
| E05000144 Swiss Cottage | 2 | 0 | 3 | 1 | 0 | 0 | 0 | 2 | 2 |
| E05000145 West Hampstead | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 12 |
| E05000001 Aldersgate | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| E05000005 Bishopsgate | 3 | 5 | 1 | 0 | 0 | 2 | 1 | 3 | 5 |
| E05000015 Cripplegate | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| E05000017 Farringdon Within | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 |
| E05000018 Farringdon Without | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| E05000021 Portsoken | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| E05000022 Queenhithe | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E05000023 Tower | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| E05000231 Brownswood | 7 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 2 |
| E05000232 Cazenove | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 0 |
| E05000233 Chatham | 1 | 4 | 0 | 8 | 2 | 0 | 0 | 0 | 0 |
| E05000234 Clissold | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| E05000235 Dalston | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 3 |
| E05000236 De Beauvoir | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| E05000237 Hackney Central | 40 | 41 | 32 | 0 | 0 | 0 | 1 | 0 | 8 |
| E05000238 Hackney Downs | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| E05000239 Haggerston | 7 | 3 | 2 | 0 | 1 | 0 | 0 | 4 | 18 |
| E05000240 Hoxton | 2 | 9 | 0 | 4 | 9 | 5 | 0 | 1 | 3 |



Reading:

<http://dx.doi.org/10.1111/j.1467-8659.2009.01664.x>

Gennady Andrienko, Natalia Andrienko, Sebastian Bremm, Tobias Schreck,
Tatiana von Landesberger, Peter Bak, Daniel Keim

Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns

Computer Graphics Forum, 2010, v.**29** (3), pp. 913-922



Questions?

Partition-based clustering



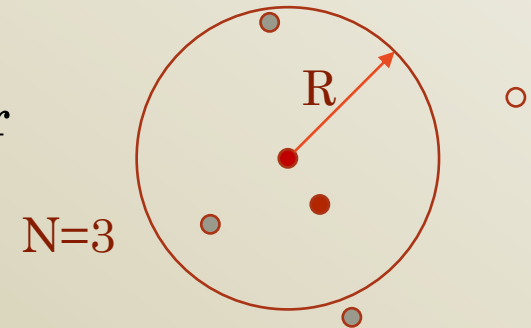
Density-based clustering



Density-based clustering (DBC)

Goal: find dense groups of close or similar objects

- For a given object \mathbf{o} , the objects whose distances from \mathbf{o} are within a chosen distance threshold (radius) \mathbf{R} are called neighbours of the object \mathbf{o} .
- An object is treated as a core object of a cluster if it has at least \mathbf{N} 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”.

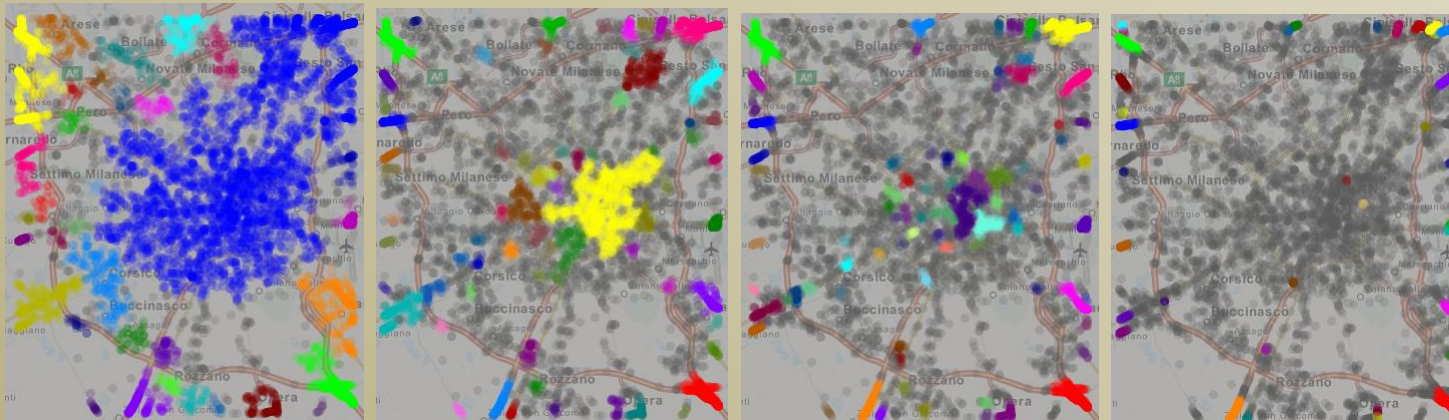




Density-based clustering

Parameters

- For DBC, the user needs to specify the neighbourhood radius (distance threshold) R and the minimum number of neighbours N .
⇒ 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 the analyst to find suitable values for R and N that lead to good results.



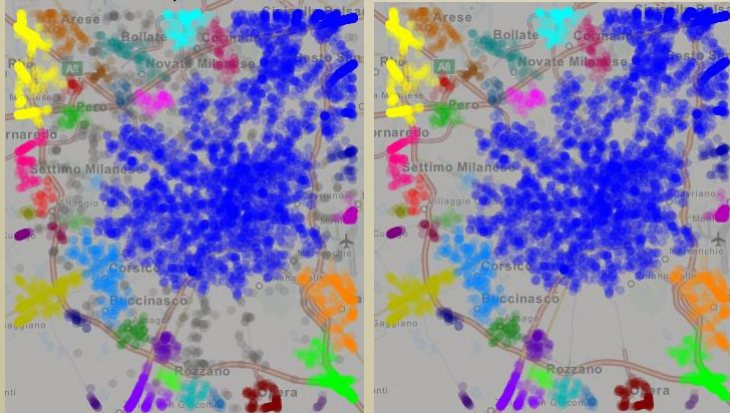
Grey: “noise”



Exploring the impact of the DBC parameters

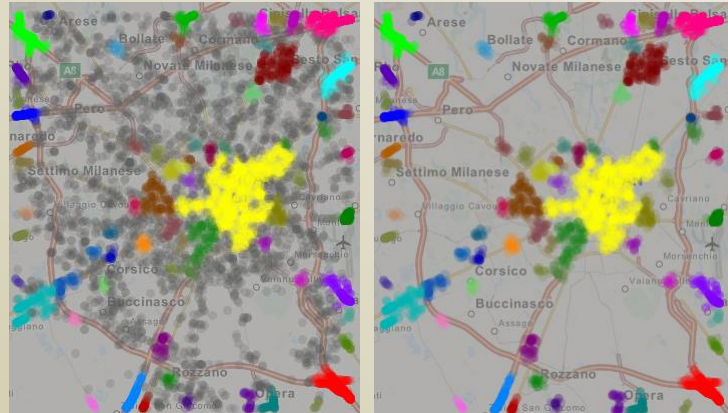
Example: DBC according to the spatial distances between point objects

$R=500m$, $N=10$



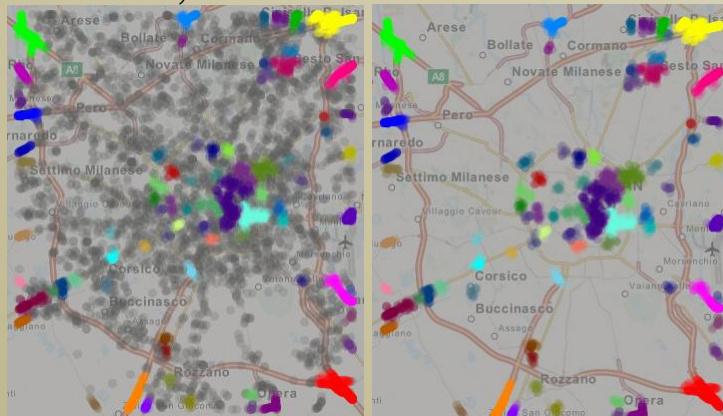
The clusters are too loose and too extended in space.

$R=300m$, $N=10$



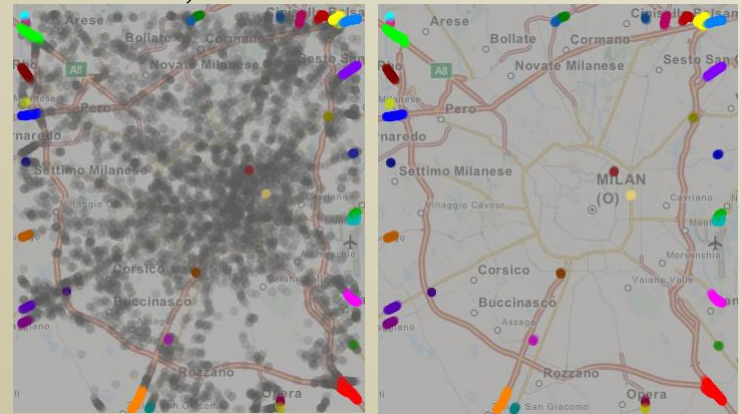
Some clusters are still too loose.

$R=250m$, $N=10$



The clusters are more or less OK.

$R=100m$, $N=10$



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



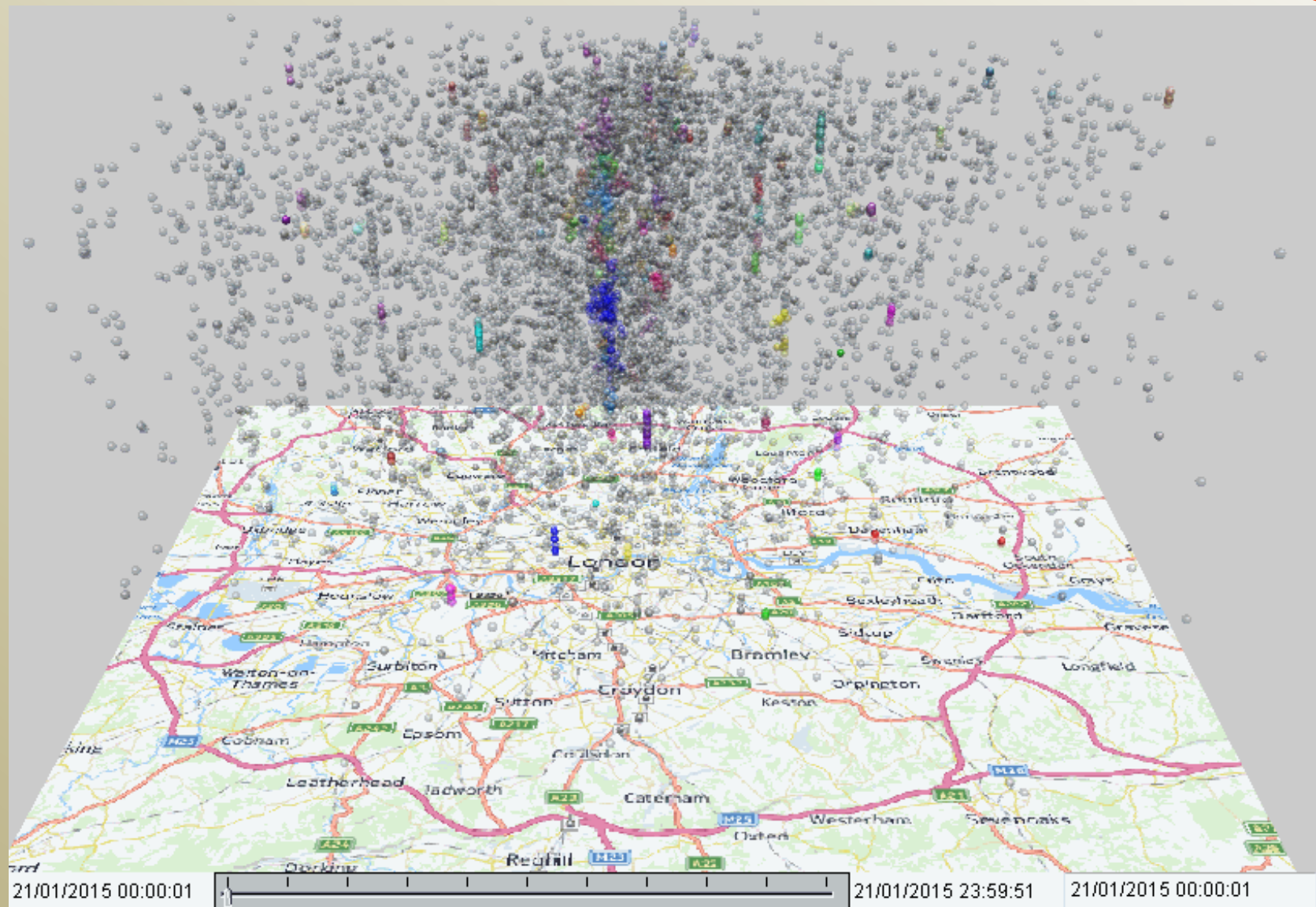
DBC by spatio-temporal distances

A possible application: clustering of spatial events

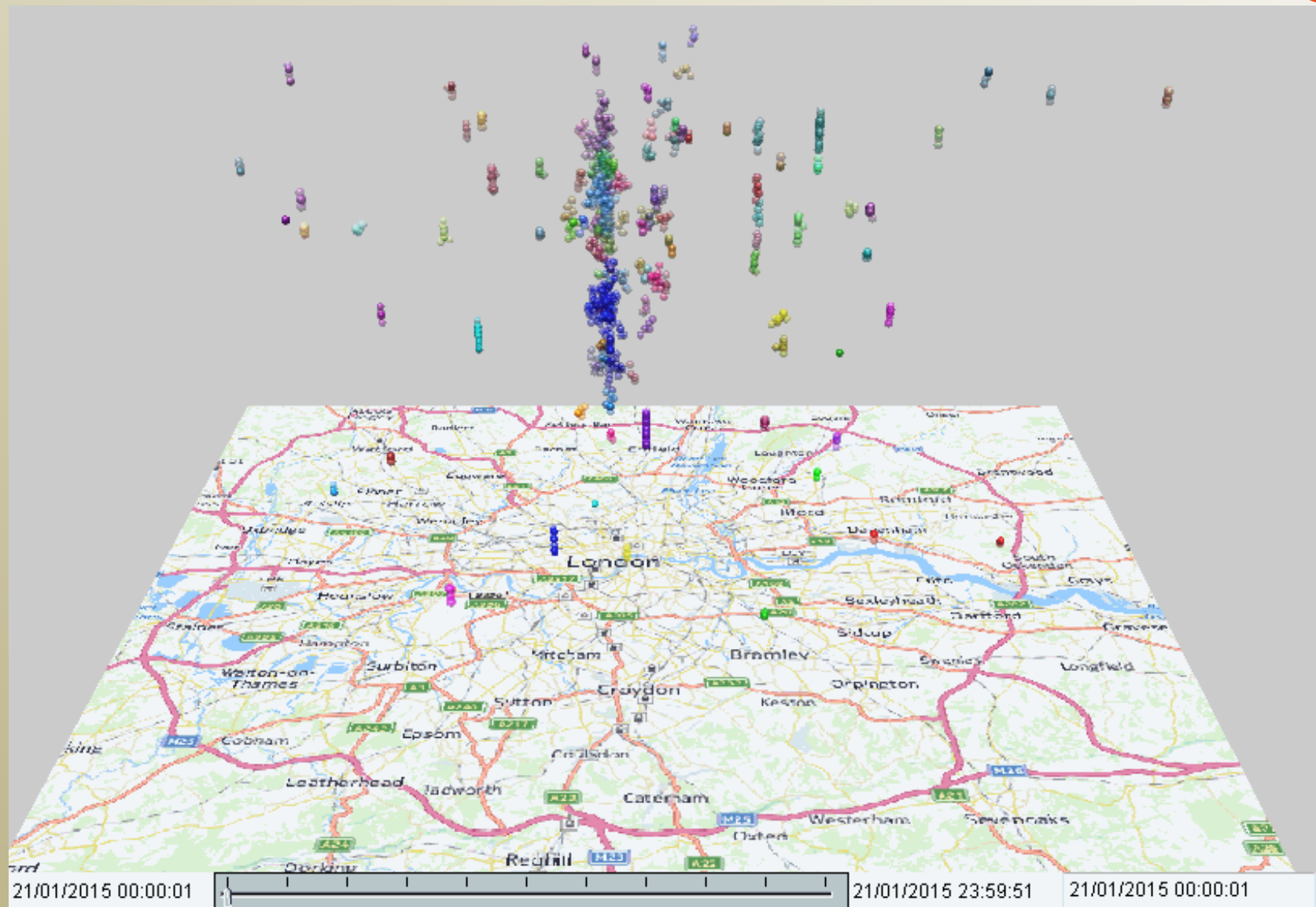
- For any two objects, there is a distance in space d_{space} and a distance in time d_{time} .
- To cluster the objects by their spatio-temporal proximity, the analyst may choose two neighbourhood radii R_{space} and R_{time}
 - e.g., $R_{\text{space}} = 300$ m and $R_{\text{time}} = 30$ minutes.
- However, the clustering algorithm requires a single distance and a single radius.

⇒ Spatial and temporal distances need to be combined together

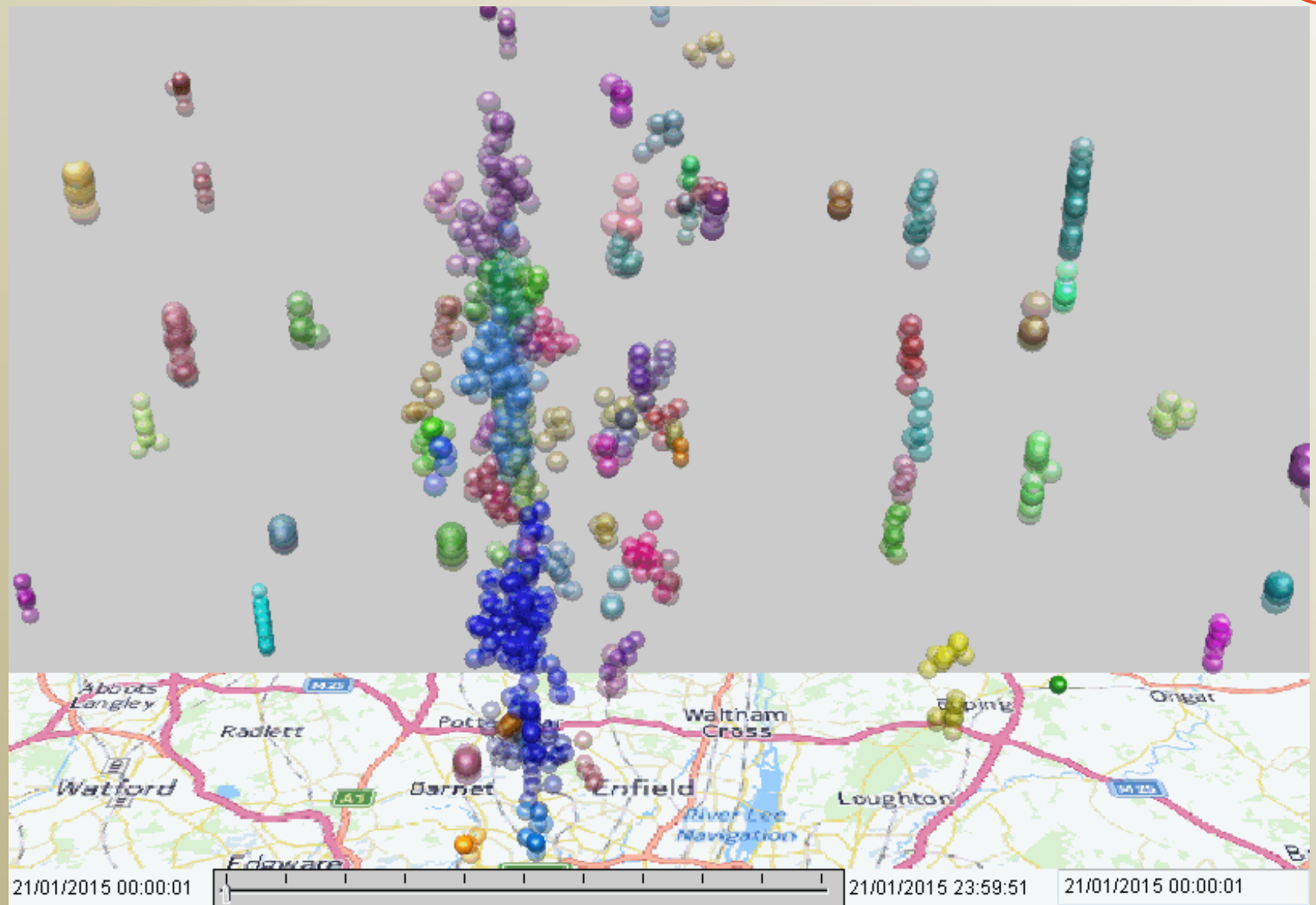
- e.g., $d = \max(d_{\text{space}}/R_{\text{space}}, d_{\text{time}}/R_{\text{time}}) * R_{\text{space}}$

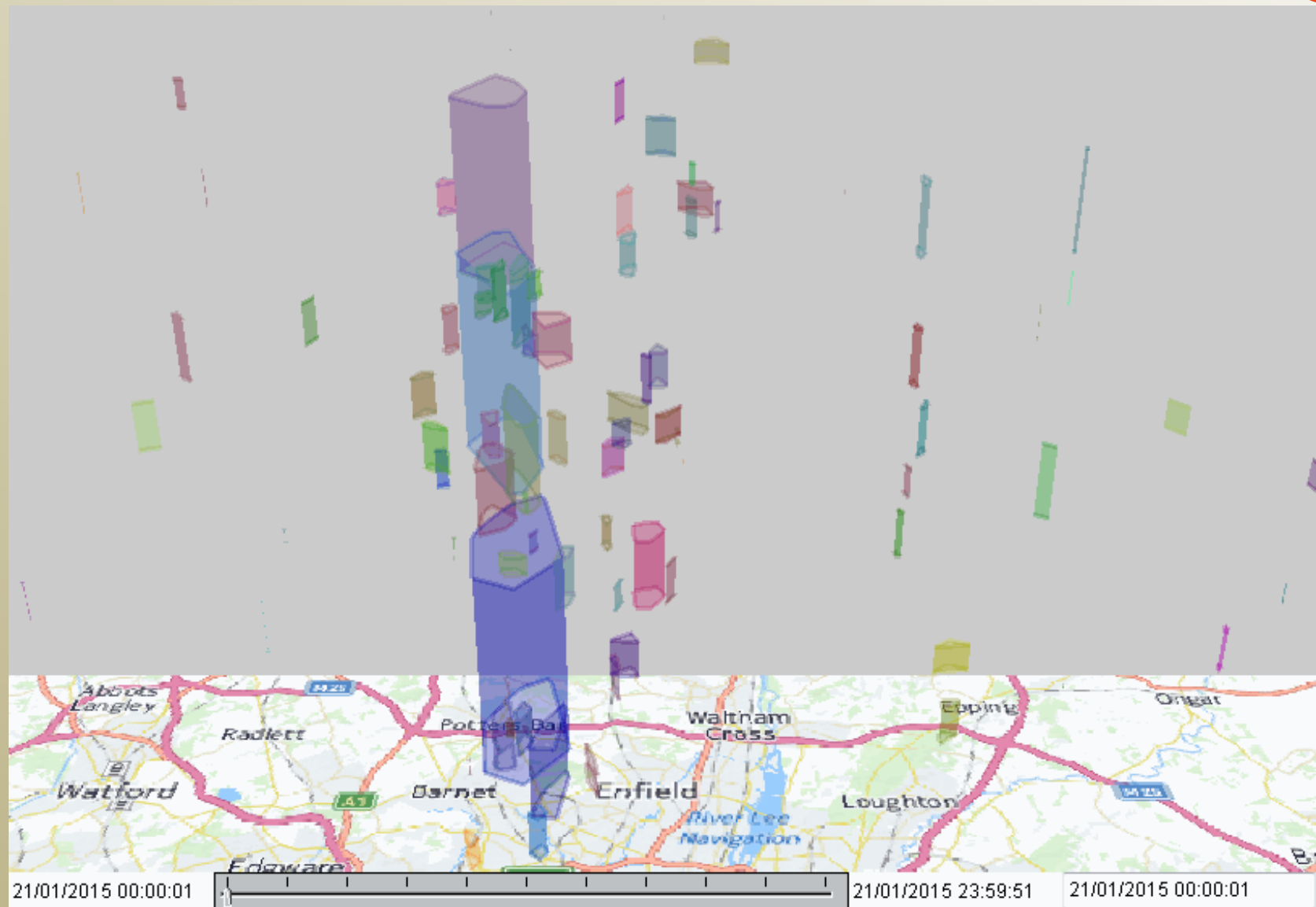


Example: clusters of London tweet events (21/01/2015; 15% sample) with $R_{\text{space}} = 500$ m, $R_{\text{time}} = 20$ minutes, and $N = 3$. Grey: noise.

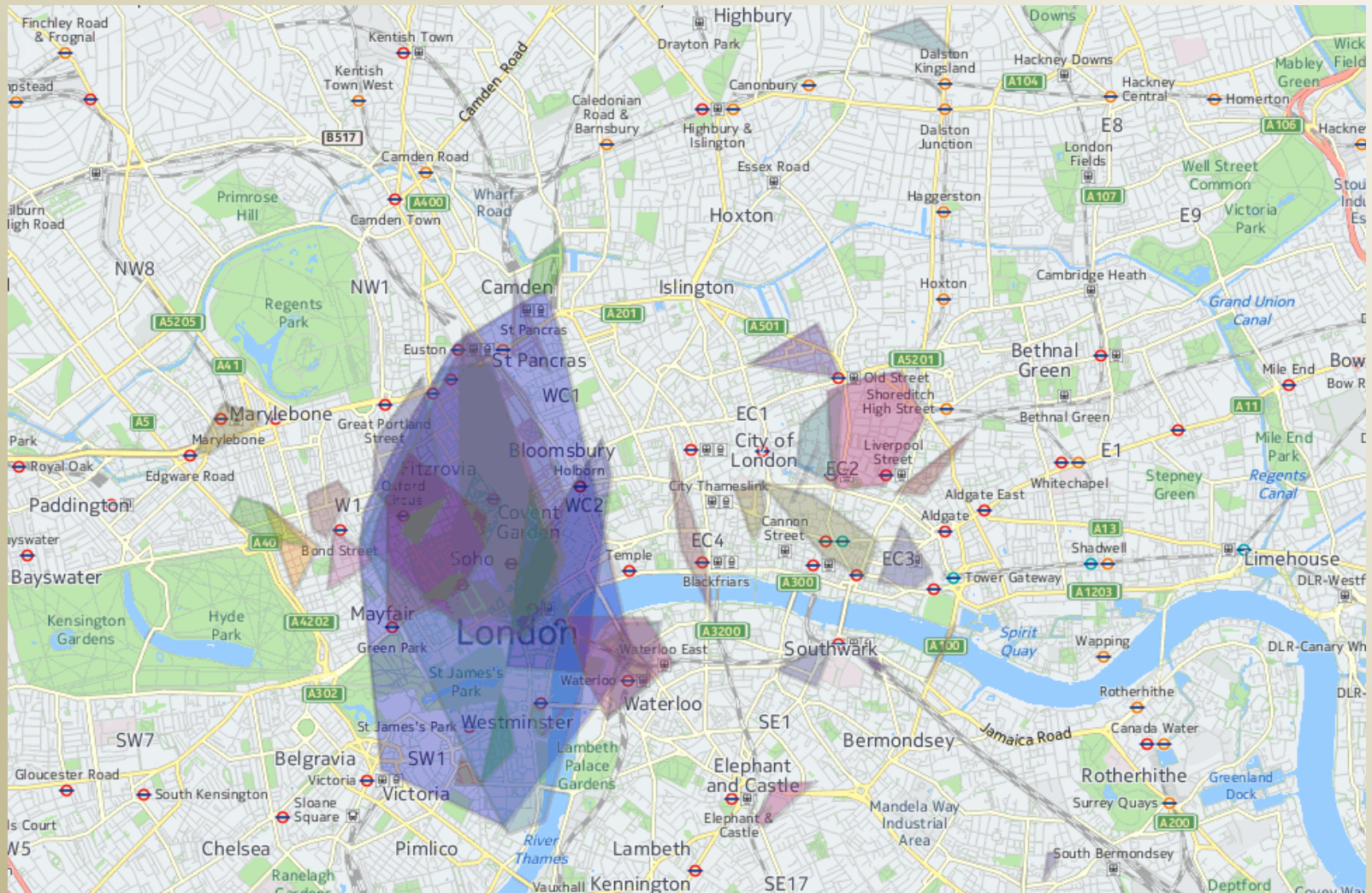


Example: clusters of London tweet events (21/01/2015; 15% sample) with $R_{\text{space}} = 500 \text{ m}$, $R_{\text{time}} = 20 \text{ minutes}$, and $N = 3$. The noise is filtered out.





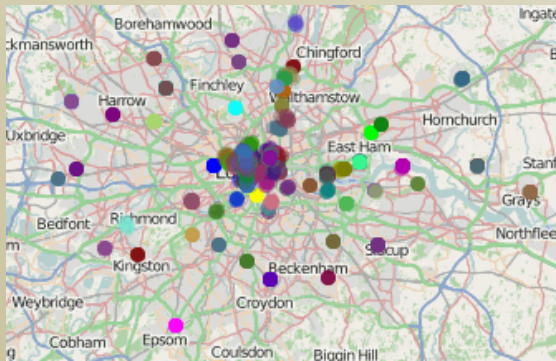
Spatio-temporal convex hulls of the event clusters.



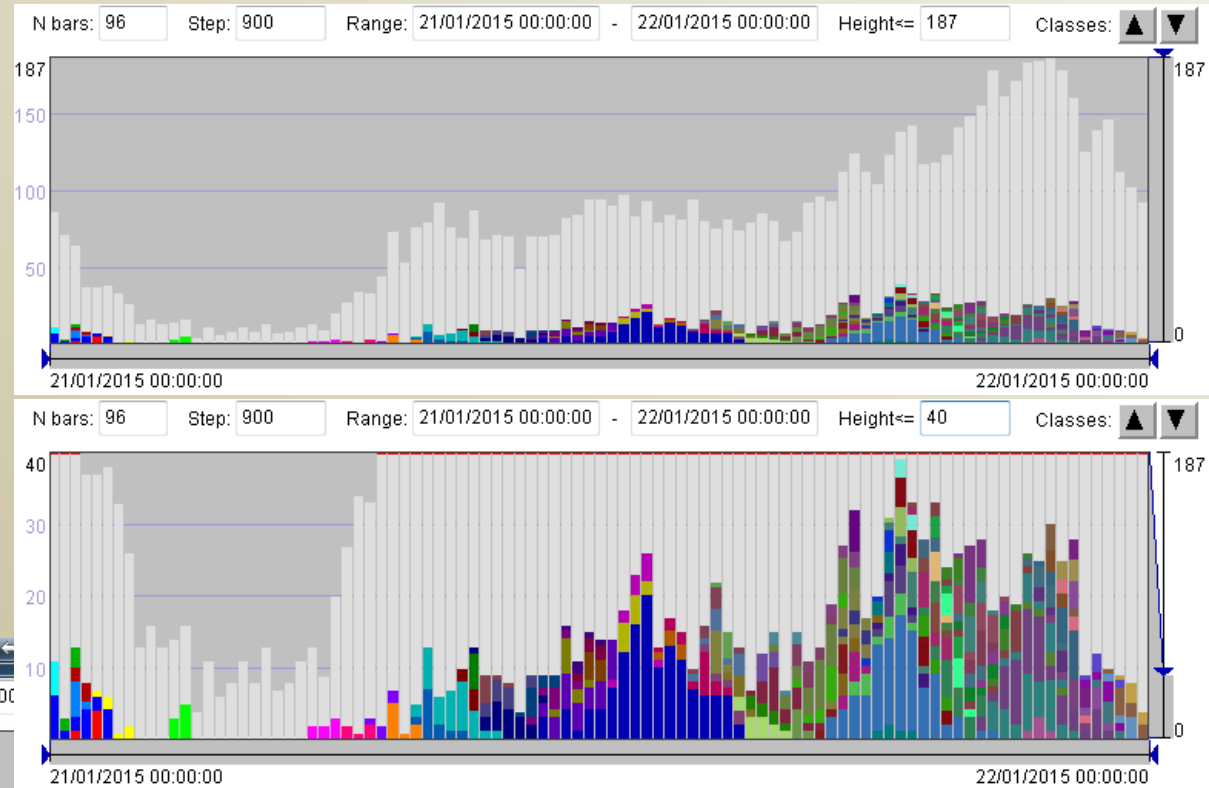
Spatio-temporal convex hulls of the event clusters
projected on a map.



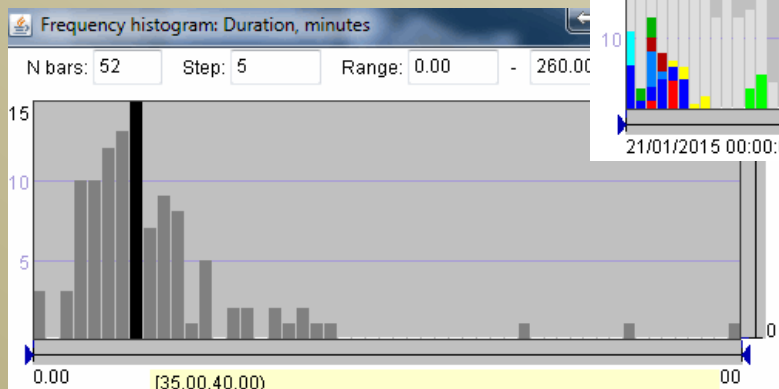
Investigation of clustering results



Spatial distribution of the dense clusters



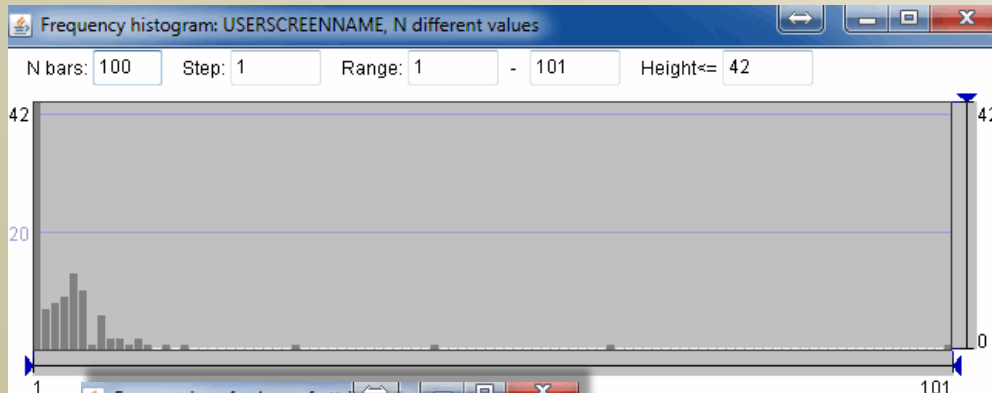
Temporal distribution of the dense clusters



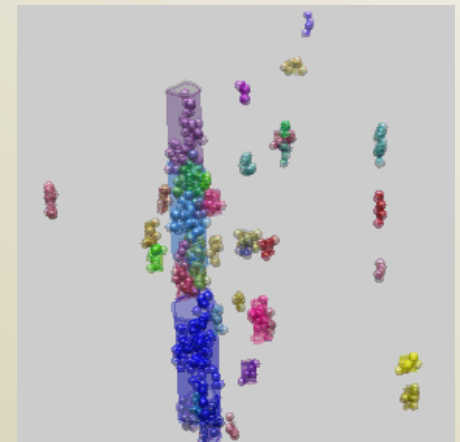
Cluster durations



Summarization of attribute values by the clusters



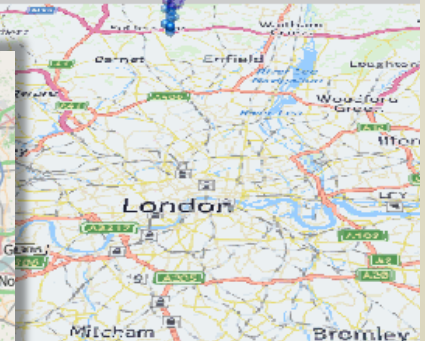
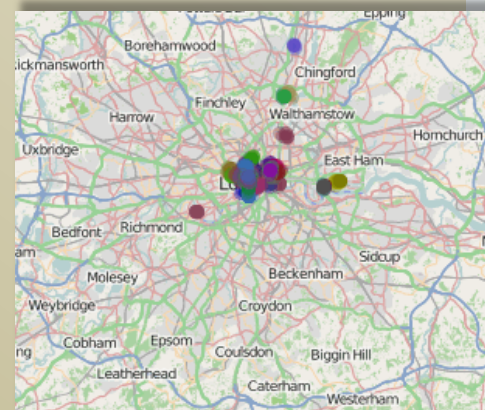
Clusters with at least 5 distinct Twitter users:



LOCATIONS

| <input type="checkbox"/> identifiers | Overall frequency | Frequency after filtering |
|--------------------------------------|-------------------|---------------------------|
| London; England | 621 | 319 |
| Camden Town; London | 356 | 108 |
| Paddington; London | 264 | 80 |
| East Ham; London | 192 | 59 |
| Lambeth; London | 328 | 54 |
| Greenwich; London | 172 | 50 |
| Brent; London | 219 | 47 |
| Camberwell; London | 276 | 40 |
| City of London; London | 132 | 36 |
| Tottenham; London | 203 | 32 |
| Hackney; London | 219 | 29 |
| Eltham; London | 109 | 28 |
| Croydon; London | 205 | 27 |
| Hillingdon; London | 232 | 26 |
| Wandsworth; London | 255 | 22 |
| Islington; London | 262 | 22 |
| Barnet; London | 251 | 22 |
| Richmond; London | 170 | 20 |
| Bexley; London | 87 | 18 |
| Romford; London | 143 | 15 |

Sort by: Frequency after filtering order:





Extending the spatio-temporal distance function to additional thematic attributes

Distance function:

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

- D_s – spatial distance threshold
- D_0, D_1, \dots, D_N - distance thresholds for other attributes
- $d_s, d_0, d_1, \dots, d_N$ – distances; d_s – distance in space

Distance in time

(t_1, t_2 are intervals):

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

Distance for a cyclic attribute

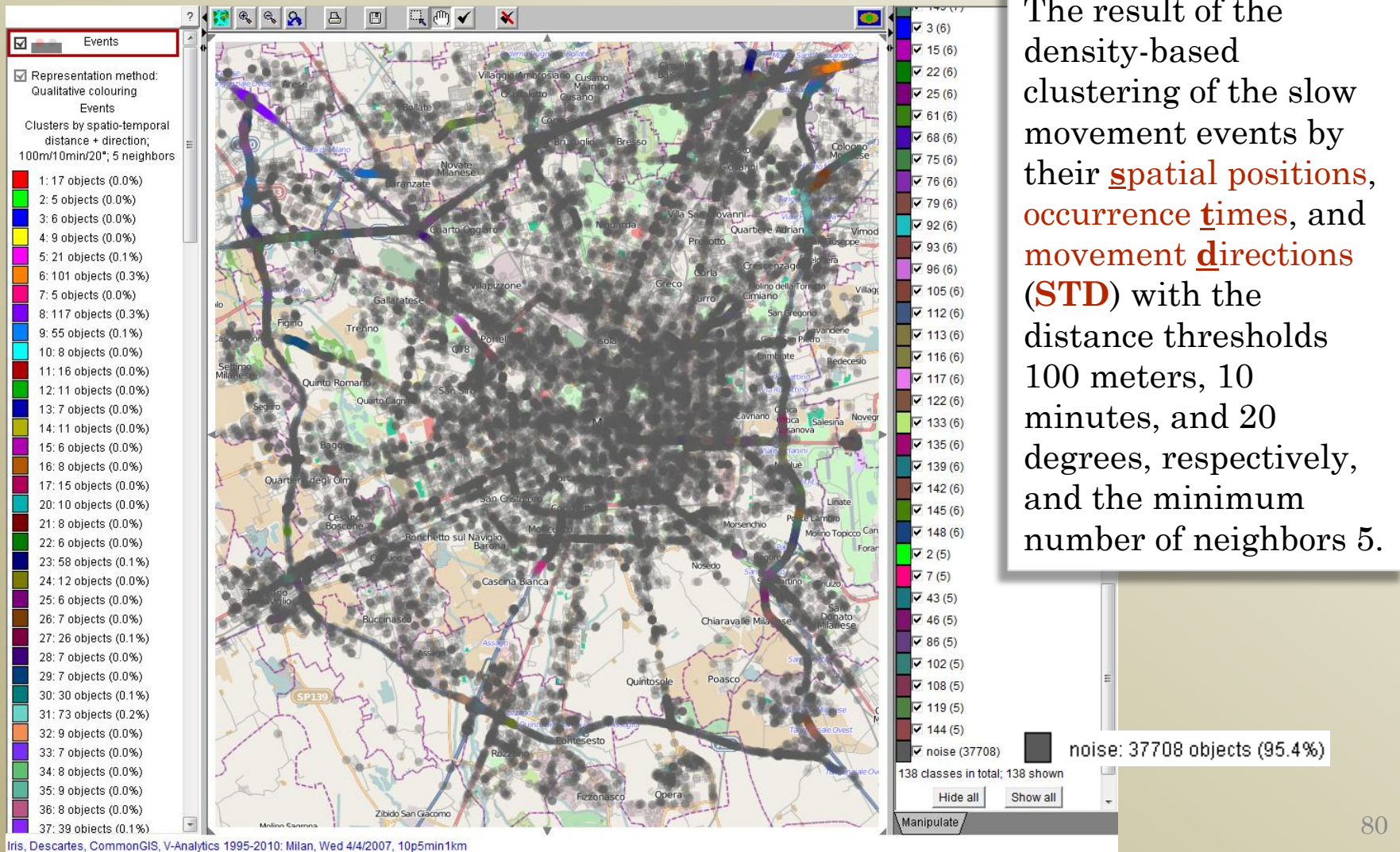
(V is the cycle length):

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

E.g., direction: $V = 360^\circ$; $d(5^\circ, 355^\circ, 360^\circ) = 10^\circ$



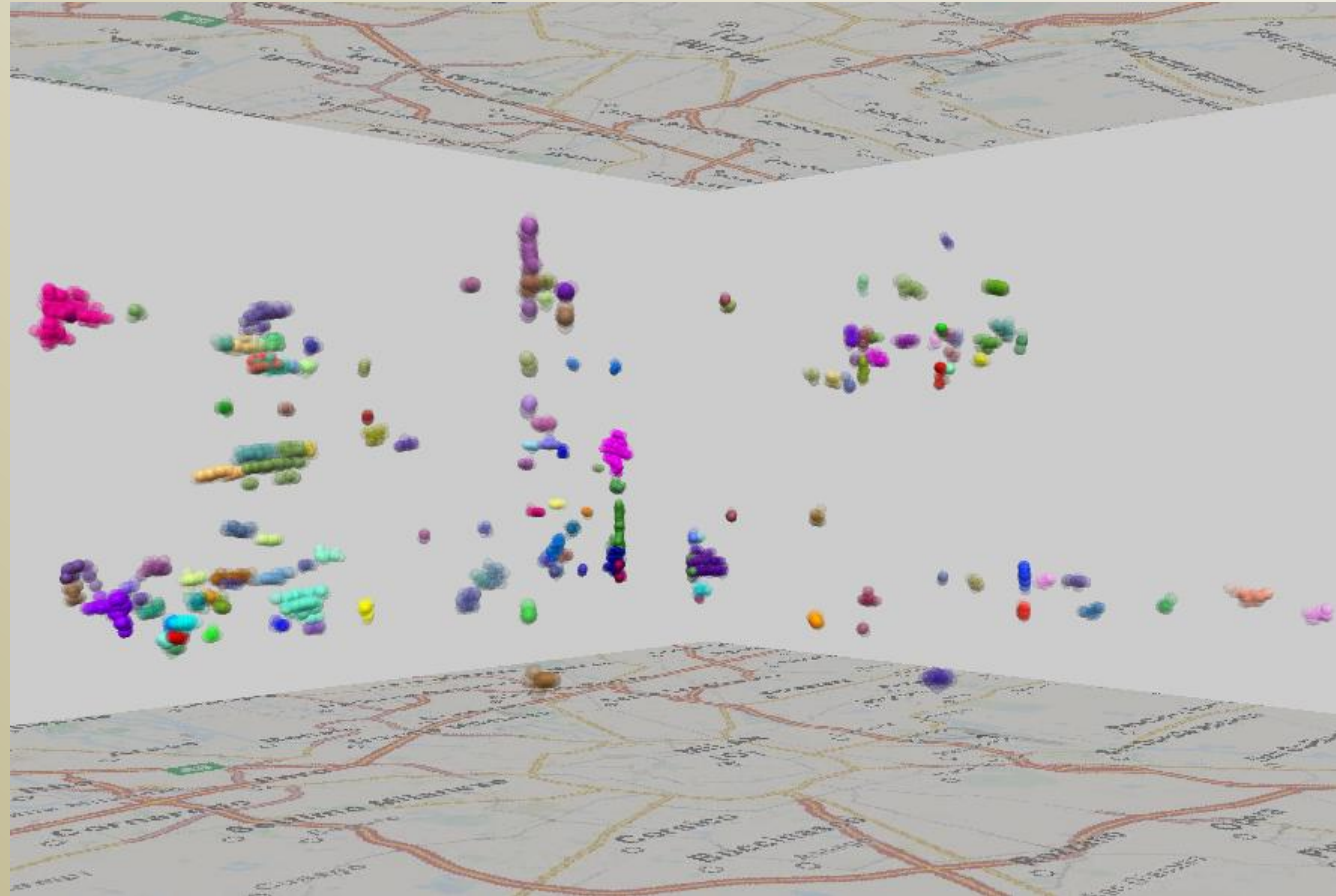
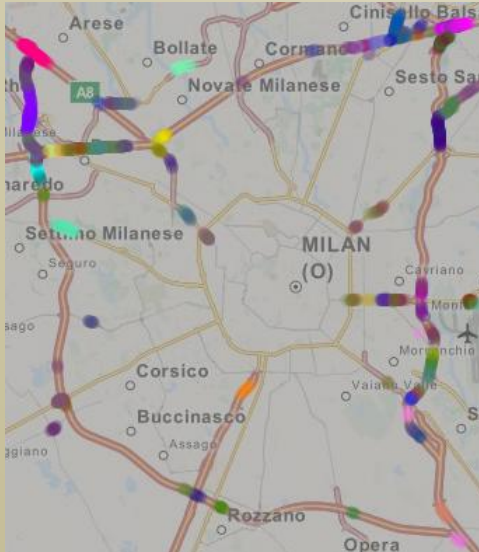
Example: clusters of low speed car movement events





The STD-clusters, noise hidden

Each cluster represents a possible traffic congestion





Density-based clustering: a summary

- Goal: find groups of highly similar (close) items and separate from them items that are less similar (more distant) to others.
- DBC is often applied to spatial and spatio-temporal objects
 - to find spatial and spatio-temporal concentrations of objects;
 - to find groups of objects with similar spatial or spatio-temporal properties
- Parameters:
 - distance threshold (neighbourhood radius) **R**
 - minimal number of neighbours of a cluster core object **N**
- The analyst needs to set a meaningful distance threshold
 - ⇒ Well understandable distances between items must exist
 - ✓ Spatial distance, temporal distance, difference between directions, ...



Distance functions in DBC

- Elementary distances: spatial, temporal, difference of values of a single thematic attribute
- It may be necessary to group objects on the basis of two or more elementary distances, e.g., spatial and temporal
⇒ A distance function integrating the elementary distances is needed
- General approach:
 - 1) Set a separate threshold for each elementary distance
 - 2) Transform the absolute elementary distances to relative w.r.t. the respective thresholds
 - 3) Combine the relative distances:
 - take their maximum or compute the Euclidean or Manhattan distance
- Defining more complex distance functions is also possible
 - May be needed for complex objects, such as trajectories (*to be considered later*)



Investigation of parameter impact

- The results of DBC greatly depend on the parameter settings (values of R and N)
- ⇒ It is necessary to run the clustering tool multiple times with different parameter settings
- Choose clear, easily interpretable results
 - Results from different runs may complement each other and contribute to better understanding
- Interactive visual interfaces are used for investigating the results of different runs.



Visual investigation of DBC results

- Analogously to PBC, clusters are given distinct colours, which are used for colouring marks in a map, space-time cube, various graphs and plots, segments in histograms, ...
 - Noise is usually shown in grey
 - The analyst should be able to interactively hide and unhide the noise
- Problems in visualising density-based clusters:
 - ☹ DBC may produce more clusters than there are distinguishable colours
 - ⇒ The analyst should not rely too much on cluster colours but use them mainly for distinguishing clusters from noise.
 - ☹ Visual displays showing individual cluster members may be too cluttered
 - ⇒ The clusters need to be represented in a summarized form
 - E.g., by spatial or spatio-temporal convex hulls



Reading

- 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
<http://dx.doi.org/10.1109/TVCG.2012.311>



Questions?

Density-based clustering



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, hierarchical
 - 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”
- DBC and PBC can be combined: first use DBC to find and filter out the noise (i.e., outliers), then apply PBC to the remaining data.
 - This may result in cleaner and clearer clusters.



That's not yet all about clustering!

- How clustering algorithms work: to come in the machine learning module.
- Specific distance functions for trajectories: in this module.
- Progressive clustering using different distance functions: in this module.



Not only clustering ...

- By example of clustering, we presented the general principles of using computational methods in data analysis.
 - We cannot consider all classes of computational methods, but the principles apply to any of them.
- The principles:
 - Apply any method iteratively, don't rely on a single run
 - Test diverse parameter settings, compare results, understand the parameter impact
 - Subdivide the data, apply the method to data subsets, compare results, understand differences, refine your (mental) model
 - Visualise all results (don't trust numbers!) and apply analytical reasoning
 - Try to see the whole
 - Note and interpret visual patterns
 - Attend to particulars: global and local outliers, gaps, intrusions or interruptions in patterns, ...



Visual Analytics technology:

- combine visual and computational analysis methods

Goal: divide the labour between humans and computers so as to enable their synergistic work.

