Practical 03

Clustering as an instrument for interactive visual analysis

given by

prof. Gennady Andrienko and

prof. Natalia Andrienko

General description

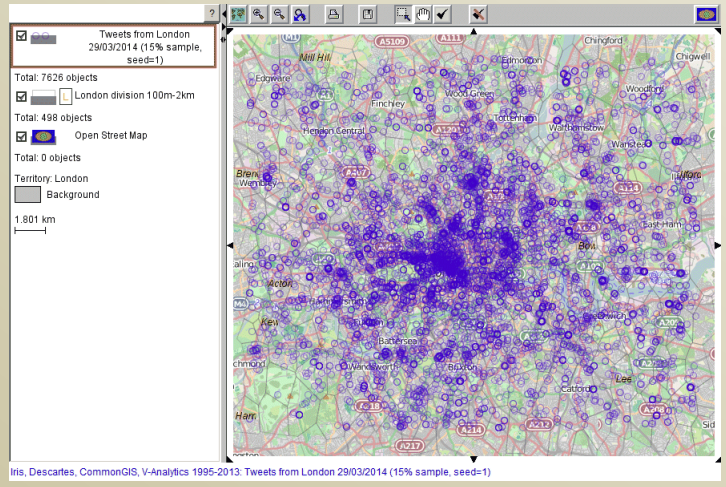
- Two major topics:
  - Exploration of **spatial events**
  - Exploration of **spatial time series**
- Data: geo-located tweets from London
  - Spatial events: a sample of tweets from one day
  - Spatial time series: tweets from one week aggregated into counts by territory compartments and hourly time intervals
- 2-3 student groups work with different datasets (same territory, different days and weeks)
- Software: V-Analytics (simplified UI, reduced functionality)

Note: the following illustrations do not show the specific results the students are supposed to obtain but show how the results may look like.

Topic 1: Exploration of spatial events

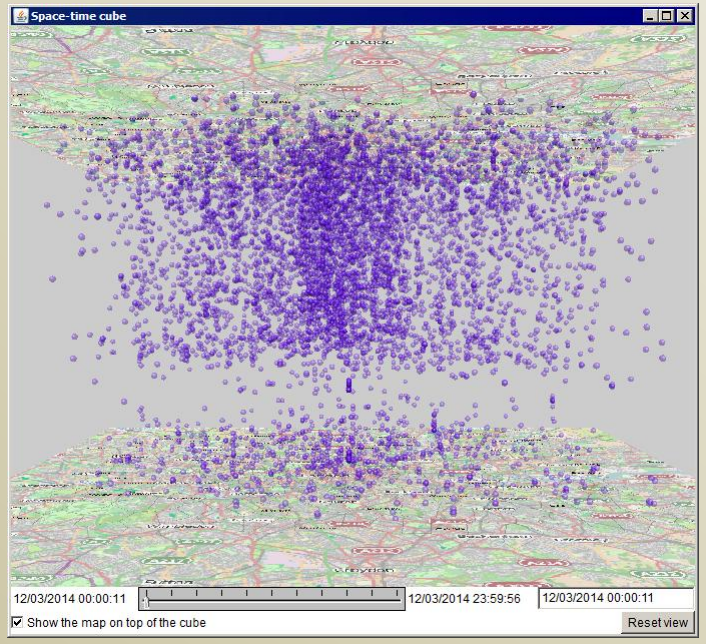
Plan and preparation

- Plan
  - Visualization of spatial events in a space-time cube
  - Spatio-temporal density-based clustering
  - Exploration of clustering results using visual displays
  - Exploration of clustering results using basic text analysis functions (extraction of frequent terms, text cloud display)
  - Spatio-temporal aggregation by territory compartments and time intervals
- Preparation to the exercise
  - Start V-Analytics
  - Load project “events.app”
    - Menu “File” > “Load project” > button “Browse” > open folder named by your group number (“1”, “2”, or “3”) > load file “events.app”
  - The events are loaded and shown on a map



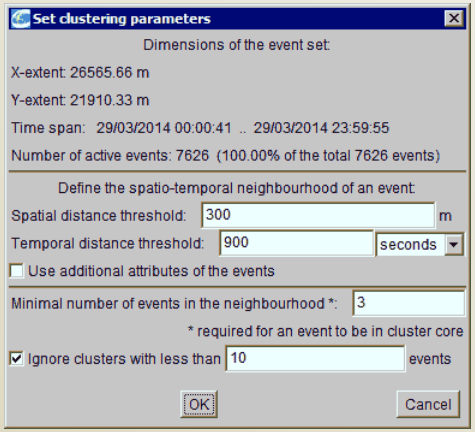
1.1. Space-time cube

- Visualize the events in a space-time cube
  - Menu “Display” > “Space-time cube” > the layer with the events is selected in the list > press button “OK” > a new window with a space-time cube appears
- Observe the spatio-temporal distribution of the events. How did the event density vary over time? Can you see spatio-temporal clusters of events?
- Interactive operations:
  - Switch on and off the upper map: checkbox “Show the map on top of the cube”
  - Rotate the view left or right: press RMB (right mouse button) and move the mouse left or right
  - Move closer to or farther from the viewpoint: press RMB and move the mouse down or up
  - Shift left, right, up, down, etc.: press LMB (left mouse button) and move the mouse
  - Rotate the view forward or backward: press RMB while pressing Control key and move the mouse down or up
  - Reset the view: double-click or button “Reset view”



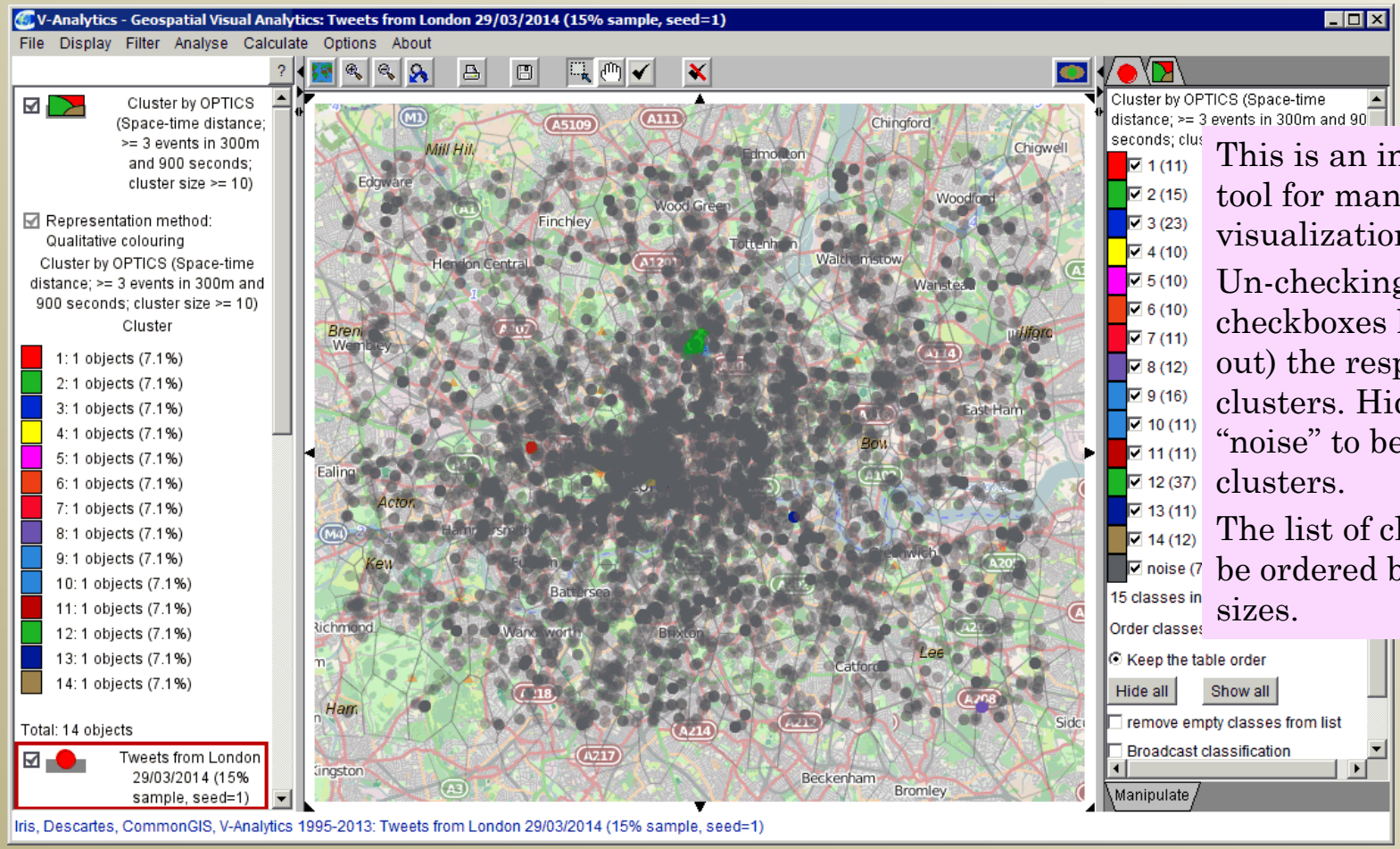
1.2. Density-based clustering of events - 1

- Activate the clustering tool:
  - Menu “Analyse” > “Events: density-based clustering” > a dialog appears; the layer with the events is pre-selected > press OK
- Set the clustering parameters
  - The suggested default parameters can be used. If you wish to obtain more clusters and/or bigger clusters, try to change the temporal distance threshold to 20 minutes. Pressing OK starts the clustering.
- After the clustering finishes, the system shows the results in two ways:
  - The dots representing the events in the map and space-time cube are coloured according to their cluster membership. Grey colour is used for “noise”.
  - For each cluster, excluding the “noise”, the system builds its convex hull. A new map layer with the hulls of all clusters is added to the map. The interiors of the hulls are painted in the same colours as the dots from the respective clusters.



1.2. Density-based clustering of events - 2

Representation of results by colouring of dots on the map and in the STC



This is an interactive tool for manipulating the visualization.

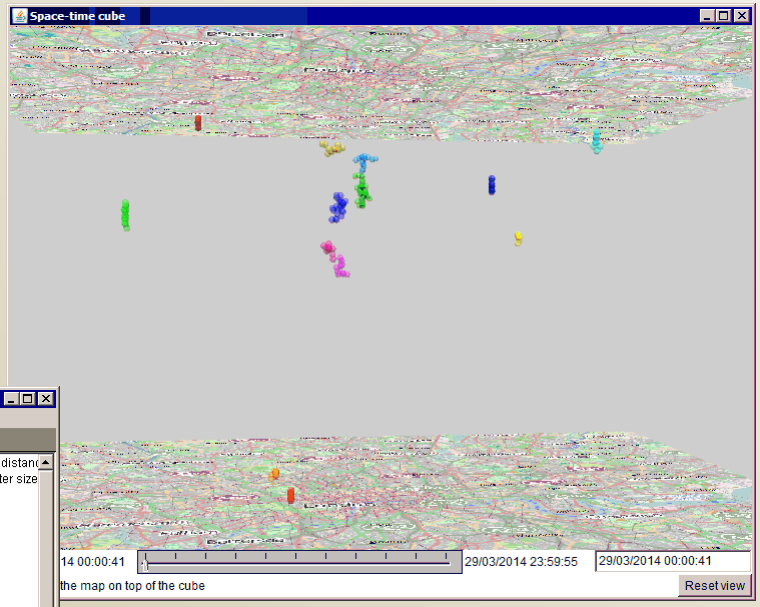
Un-checking the checkboxes hides (filters out) the respective clusters. Hide the "noise" to better see the clusters.

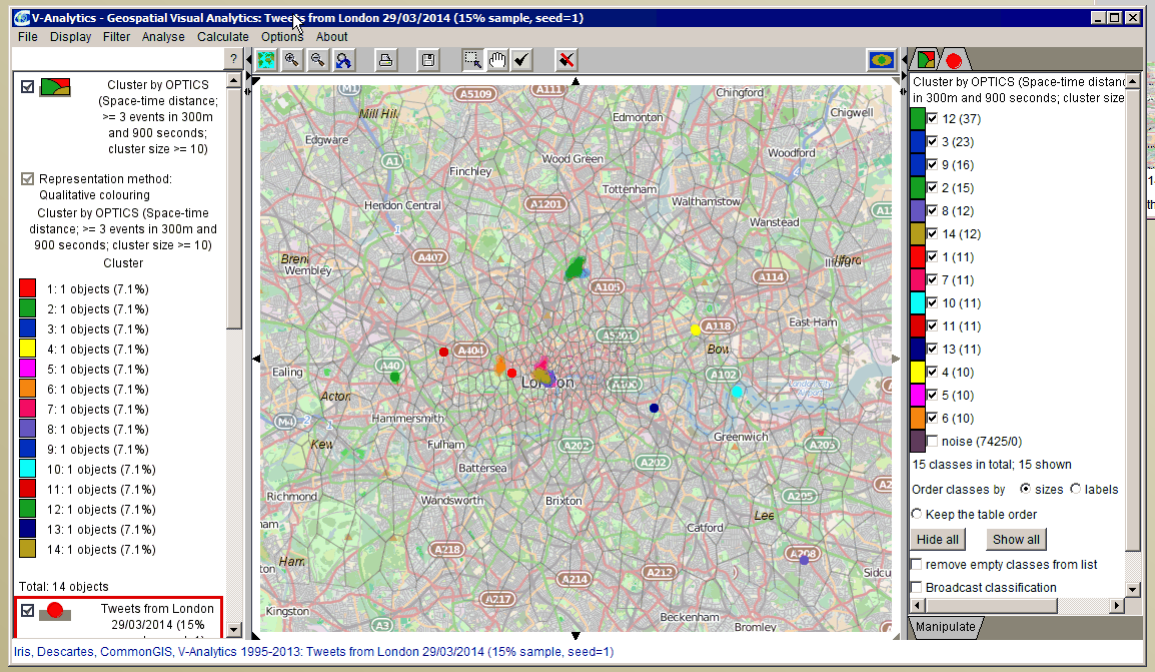
The list of clusters can be ordered by cluster sizes.

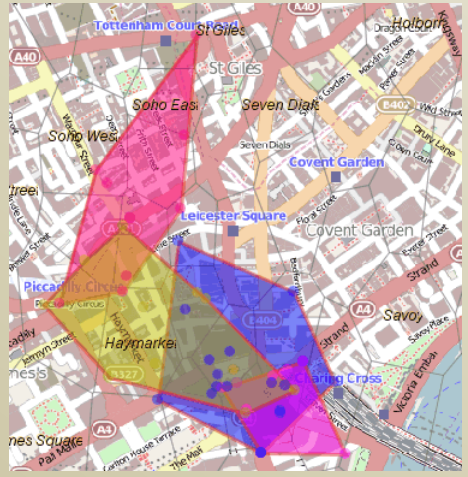
1.3. Exploration of clusters using visual displays

Hiding "noise", selecting clusters to view

Observe the spatial and temporal positions of the event clusters using the map and space-time cube. Where are the biggest clusters, areas with multiple clusters, clusters with longest durations (most extended vertically in the STC)?







1.3. Exploration of clusters using visual displays

Viewing cluster summaries

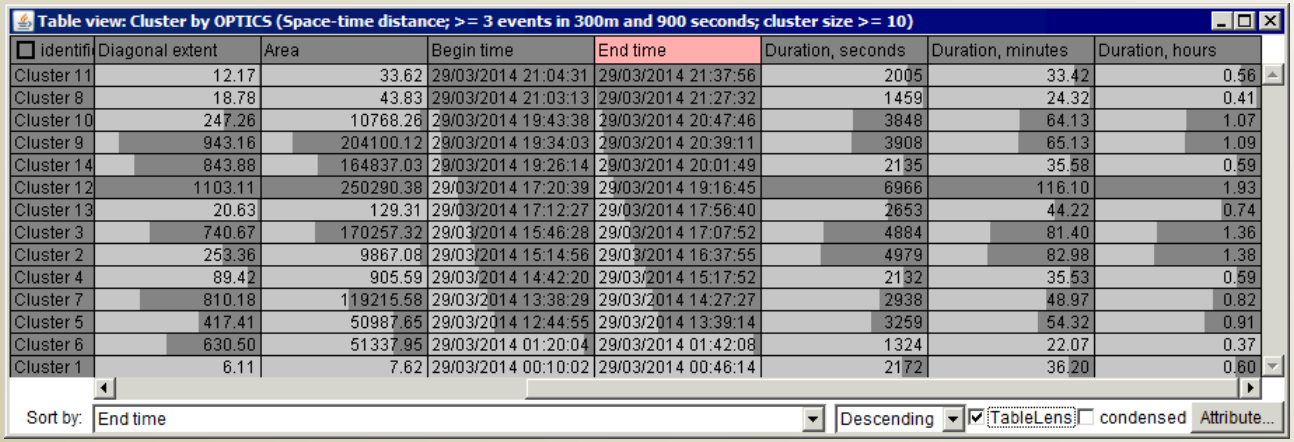


Table view: Cluster by OPTICS (Space-time distance; >= 3 events in 300m and 900 seconds; cluster size >= 10)

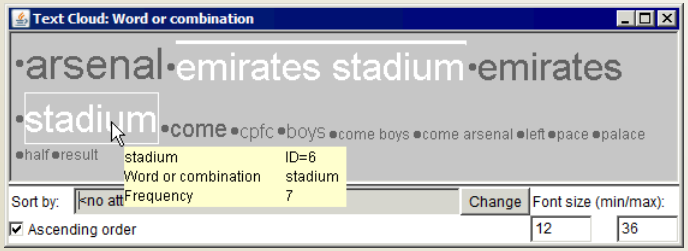
identifi	Diagonal extent	Area	Begin time	End time	Duration, seconds	Duration, minutes	Duration, hours
Cluster 11	12.17	33.62	29/03/2014 21:04:31	29/03/2014 21:37:56	2005	33.42	0.56
Cluster 8	18.78	43.83	29/03/2014 21:03:13	29/03/2014 21:27:32	1459	24.32	0.41
Cluster 10	247.26	10768.26	29/03/2014 19:43:38	29/03/2014 20:47:46	3848	64.13	1.07
Cluster 9	943.16	204100.12	29/03/2014 19:34:03	29/03/2014 20:39:11	3908	65.13	1.09
Cluster 14	843.88	164837.03	29/03/2014 19:26:14	29/03/2014 20:01:49	2135	35.58	0.59
Cluster 12	1103.11	250290.38	29/03/2014 17:20:39	29/03/2014 19:16:45	6966	116.10	1.93
Cluster 13	20.63	129.31	29/03/2014 17:12:27	29/03/2014 17:56:40	2653	44.22	0.74
Cluster 3	740.67	170257.32	29/03/2014 15:46:28	29/03/2014 17:07:52	4884	81.40	1.36
Cluster 2	253.36	9867.08	29/03/2014 15:14:56	29/03/2014 16:37:55	4979	82.98	1.38
Cluster 4	89.42	905.59	29/03/2014 14:42:20	29/03/2014 15:17:52	2132	35.53	0.59
Cluster 7	810.18	119215.58	29/03/2014 13:38:29	29/03/2014 14:27:27	2938	48.97	0.82
Cluster 5	417.41	50987.65	29/03/2014 12:44:55	29/03/2014 13:39:14	3259	54.32	0.91
Cluster 6	630.50	51337.95	29/03/2014 01:20:04	29/03/2014 01:42:08	1324	22.07	0.37
Cluster 1	6.11	7.62	29/03/2014 00:10:02	29/03/2014 00:46:14	2172	36.20	0.60

Sort by: End time | Descending | TableLens | condensed | Attribute...

- Open the table view for the table describing the clusters (the system has automatically created the table and attached it to the map layer with the cluster hulls).
  - Menu “Display” > “Table view” > a dialog for table selection appears; select the table “Cluster by OPTICS (...)” > press OK > a dialog for attribute selection appears > select attributes (you may use the “Select all” button) > press OK > the table view appears
- Determine the life times (intervals and durations) of the clusters.
- Select the checkbox “Table lens” > the attribute values are represented by darker grey bars in the table cells.
- By clicking on column titles, you can sort the table rows by the values contained in the columns. Repeated clicks toggle between ascending and descending ordering.
- Find which clusters were the earliest, latest, longest by duration, largest by the spatial extent (area).

1.4. Exploration of clusters using basic text analytics

Extraction of frequent terms

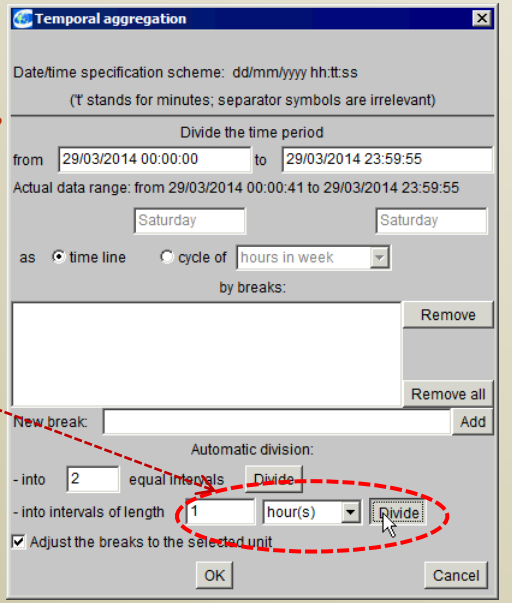


- Start the text summarization tool:
  - Menu “Analyse” > “Texts: extract frequent terms” > a dialog for table selection appears; select the table “Tweets from London ...” and press OK > a dialog for table column selection appears; select MESSAGETEXT and press OK > a dialog for setting tool parameters appears. You do not need to change the default settings.
    - Optionally: you may load a list of stop words from a file. Press the button “Take words from text file”, then browse and select the file “stop_words.txt” from the folder with the data.
  - Press “OK” in the dialog.
- The tool runs and creates a text cloud display with the terms extracted from the currently active tweets (i.e., those that are not filtered out).
- Select the clusters of tweets one by one (using the checkboxes on the right of the map). The tool re-runs, extracts frequent terms from the active tweets, and updates the text cloud display.
- Try to explain some of the clusters based on the terms and cluster locations (e.g., what public events might cause people gathering and active twitting).

1.5. Spatio-temporal aggregation of events

By territory compartments and time intervals

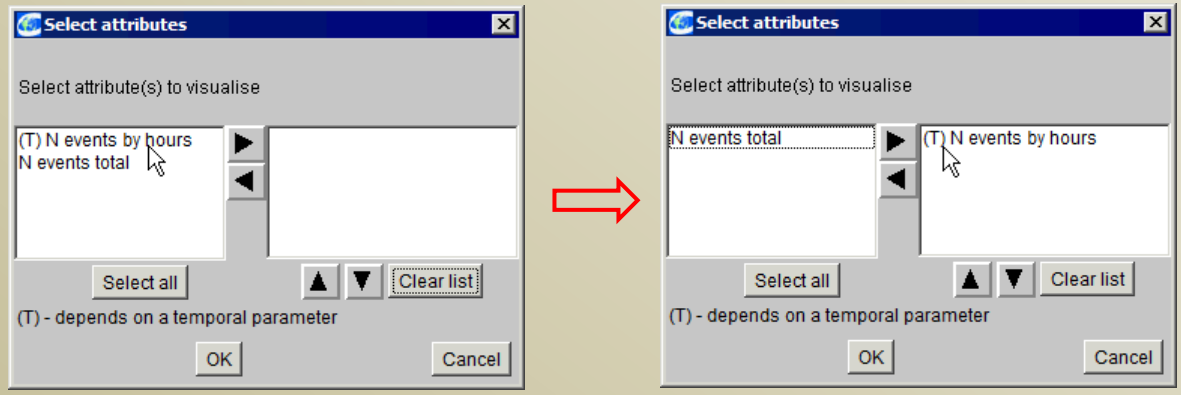
- Preparation:
  - Close all additional windows, except the main window with the map and the menu.
  - In the cluster selection panel on the right of the map, press the button “Show all”. All events must be active (visible on the map).
- Activate the event aggregation tool:
  - Select menu “Analyse” > “Events: spatio-temporal aggregation” > a dialog with 2 list boxes appears > select “Tweets from London ...” in the upper list > select “London division ...” in the lower list > press OK > a dialog “How to aggregate?” appears; “by time intervals” is selected > press OK
  - A dialog for setting time breaks appears; divide the time span of the data into hourly intervals > after the breaks have appeared in the list, press OK > confirm removal of useless breaks > press OK
  - In the following sequence of dialogs, press OK (default settings will be used), except the dialog “Compute the number of events in the neighbourhood of each area (...)?” (here press “No”)
- The tool computes time series of event counts for the compartments and puts them in the table associated with the map layer containing the territory division.



1.5. Spatio-temporal aggregation of events

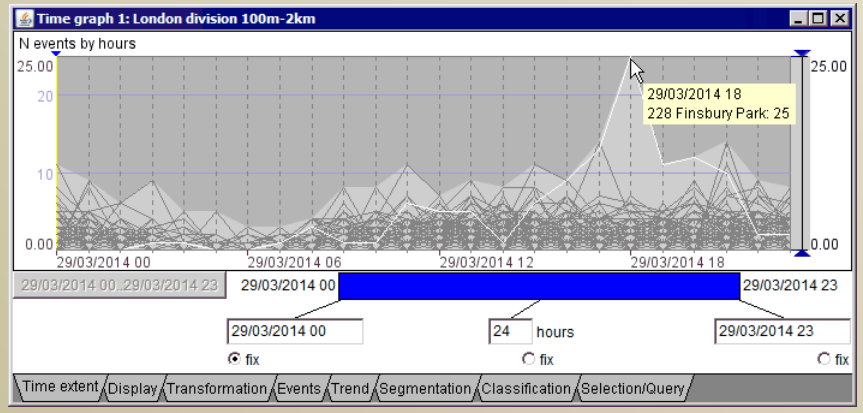
Visualization of the aggregation results (spatial time series)

- All possible visualizations for spatial time series are obtained as follows:
  - Select menu “Display” > “Display wizard” > a dialog for table selection appears > select table “London division ...” > press OK > a dialog for attribute selection appears; time-variant attributes (time series) are marked by (T) > select a time-variant attribute, i.e., move it from the left list to the right list (double-click in the list or click and press the right-directed arrow) > press OK



  - The system shows the options: Animated map, Map with value flow diagrams, Time graph.

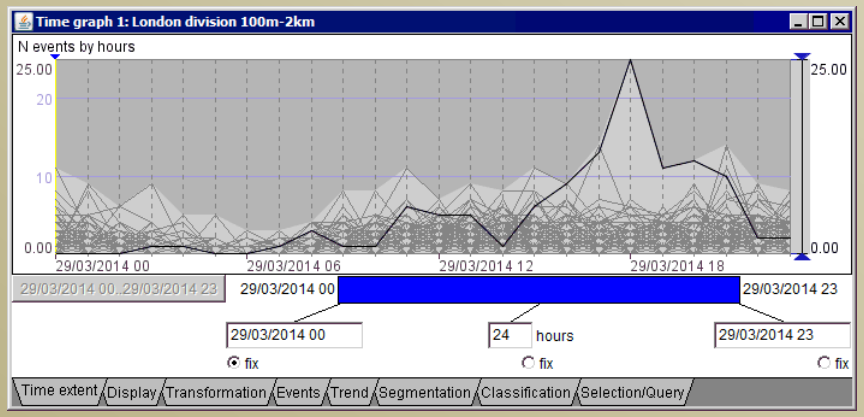
- Open a time graph display with the time series of the event counts.

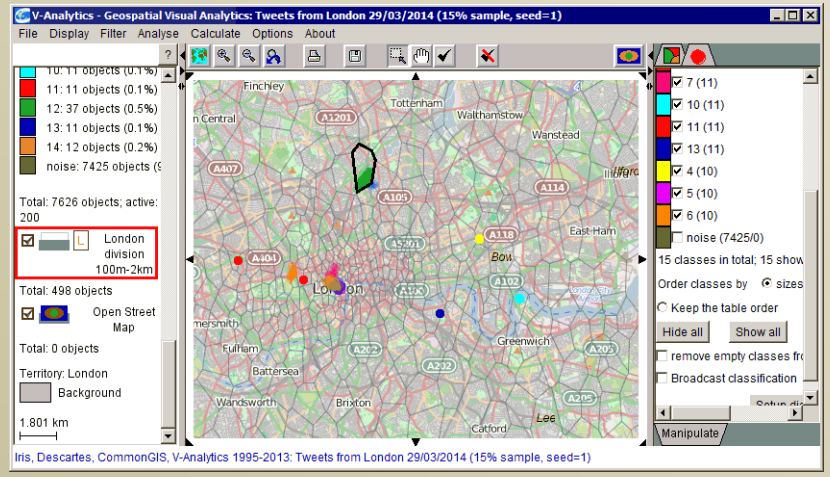


1.6. Visual exploration of spatial time series

Using the links between the map and the time graph

- Links between the map and the time graph:
  - Activate the layer “London division ...” in the map by clicking on the layer name in the map legend; a red frame in the legend marks the currently active layer.
  - Areas of the active layer in the map and curves in the time graph can be selected and deselected by clicking. Selected areas and curves are highlighted in black.
- Using these operations, find the area with the highest peak. Does it contain one or more event clusters? (Hide the “noise” in the layer with the events and make the division layer active again)
- For some other area containing an event cluster, find the corresponding curve in the time graph. Is there a peak in the time interval when the cluster existed?





Questions for discussion

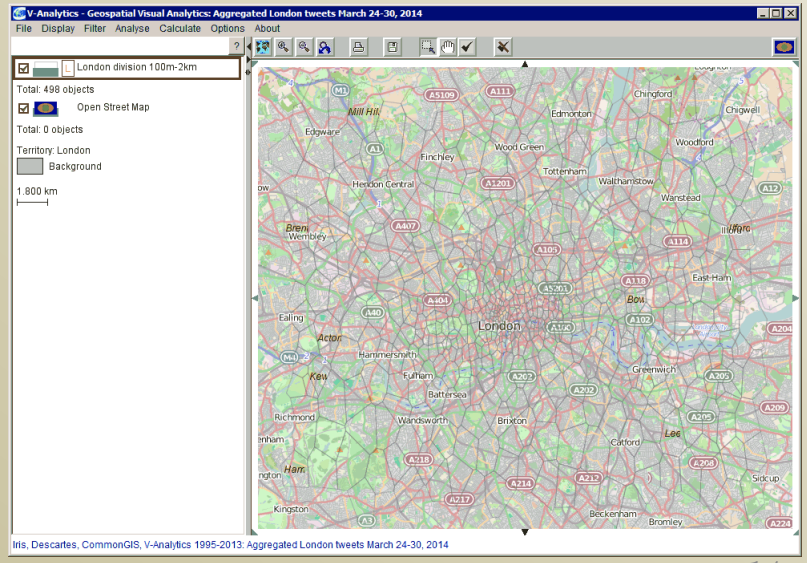
Topic: exploration of spatial events

- What does a space-time cube represent? What patterns can be seen in it?
- What is a spatio-temporal cluster of events?
- What can a dense spatio-temporal cluster of tweets mean?
- What does density-based clustering do?
- Did the density-based clustering of the tweets uncover some interpretable clusters? When and where did they occur? What were their reasons?
- What kind of data results from spatio-temporal aggregation of events? What do these data tell us about different places (territory compartments)?

Topic 2: Exploration of spatial time series

Plan and preparation

- Plan
  - Visualization of spatial time series: time graph, 2d time histogram
  - Transformation of time series
  - **Partition-based clustering** of places by similarity of the time series; exploration of clustering results
  - **Partition-based clustering** of time intervals by similarity of the spatial situations; exploration of clustering results
- Preparation to the exercise
  - Start V-Analytics
  - Load project “time_series.app”
    - Menu “File” > “Load project” > button “Browse” > open folder named by your group number (“1”, “2”, or “3”) > load file “time_series.app”
  - A map with territory division appears



Explanation of the data

- The layer with the territory division has an associated table containing time series of event counts by hourly intervals. The duration of the time series is one week, i.e., 168 (= 24 x 7) time steps.
  - Note: The table is not immediately visible. It can be visualized through menu “Display” > “Table view” (not required for the exercise).
- Table structure:

places

times →

Table view: Hourly event counts

identifiers	hour=24/03/2014,00: N events by hours	hour=24/03/2014,01: N events by hours	hour=24/03/2014,02: N events by hours
Jubilee Gardens	1	4	1
British Museum	1	0	0
Charing Cross	6	2	2
Russel Square	0	2	0
Tate Modern	9	0	1
Piccadilly Circus	0	4	0
North Kensington	7	2	7
Royal Festival Hall	0	0	0
Covent Garden Market	0	3	0
London Aquarium	1	0	1
Imperial College	0	0	1
Golden Jubilee Bridges	0	0	0
Victoria and Albert Museum	0	1	0
Leicester Square	2	1	1
Buckingham Palace	0	3	0
Tottenham Court Road	0	2	2
Parliament Square	0	2	1
Tower Bridge	2	2	1

,19: s	hour=30/03/2014,20: N events by hours	hour=30/03/2014,21: N events by hours	hour=30/03/2014,22: N events by hours	hour=30/03/2014,23: N events by hours
2	3	8	1	1
0	2	2	1	1
9	3	5	5	4
2	3	11	5	21
2	0	0	0	0
7	4	7	4	5
0	0	2	1	1
13	13	8	3	0
7	2	2	4	2
1	1	1	0	0
1	0	0	0	0
1	1	2	0	3
		5	0	1
		5	3	3
		1	5	1
		2	1	1
		1	2	3
		3	1	5

...

The times in your data example may be different

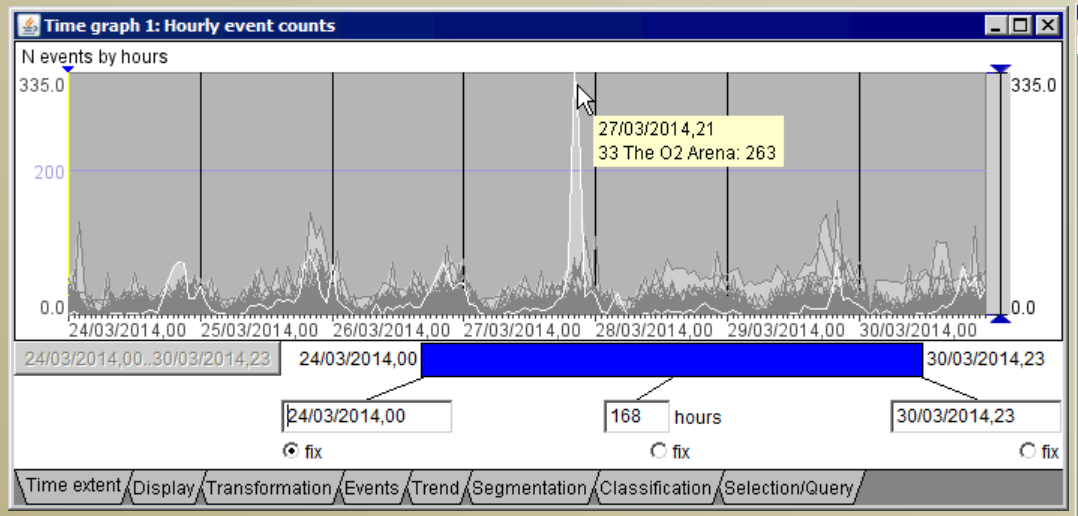
hour ∈ [24/03/2014, 00 .. 30/03/2014, 23] : temporal *parameter*

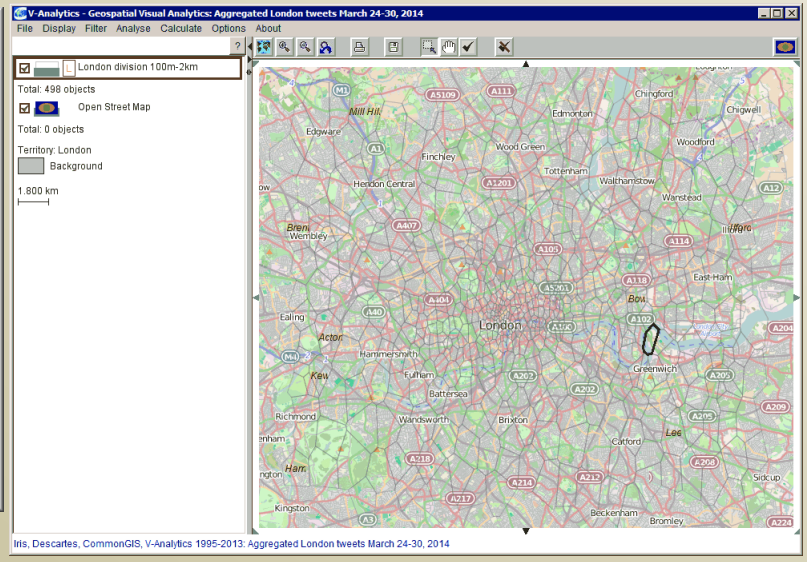
N events by hours : parameter-dependent attribute

2.1. Visualization of spatial time series - 1

Time graph

- Open a time graph display
  - "Display" > "Display wizard" > attribute selection dialog appears; select the time series "(T) N events by hours" > dialog with possible visualization options appears; select "Time graph".
- Find the times and places (territory compartments) of the highest peaks; locate the places on the map using the link between the map and the graph.

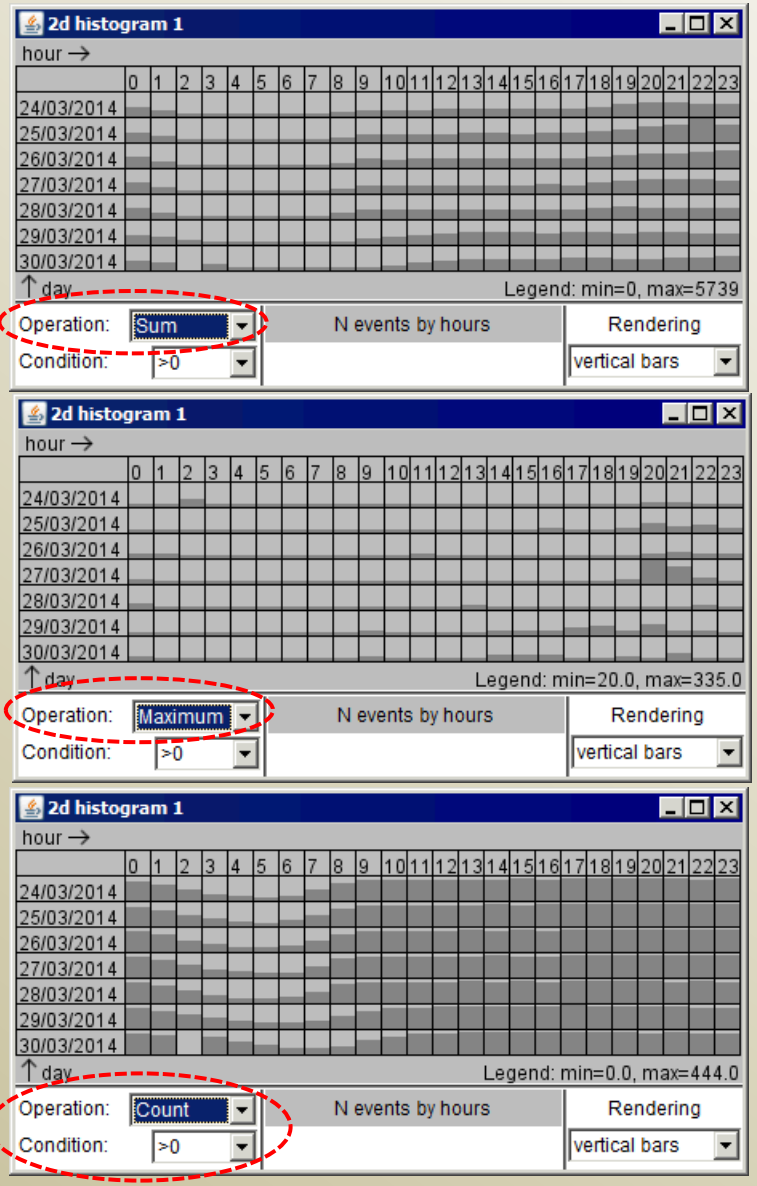




2.1. Visualization of spatial time series - 2

2-dimensional time histogram

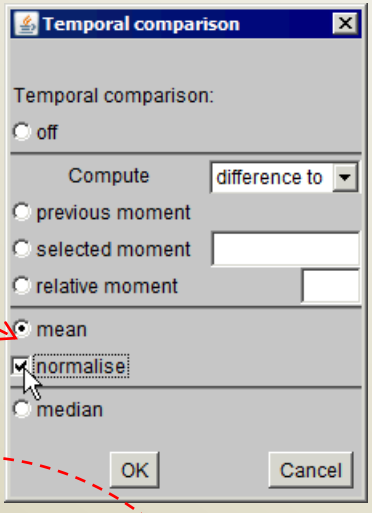
- Open a 2d time histogram display analogously to the time graph.
  - "Display" > "Display wizard" > attribute selection dialog appears; select the time series "(T) N events by hours" > dialog with possible visualization options appears; select "2-dimensional histogram".
- The rows in the histogram correspond to the days and the columns to the day hours. The cells contain summary statistics computed from the time series. The default view shows sums of event counts from all places. Other options include maximum, minimum, average, and count of places with positive values.

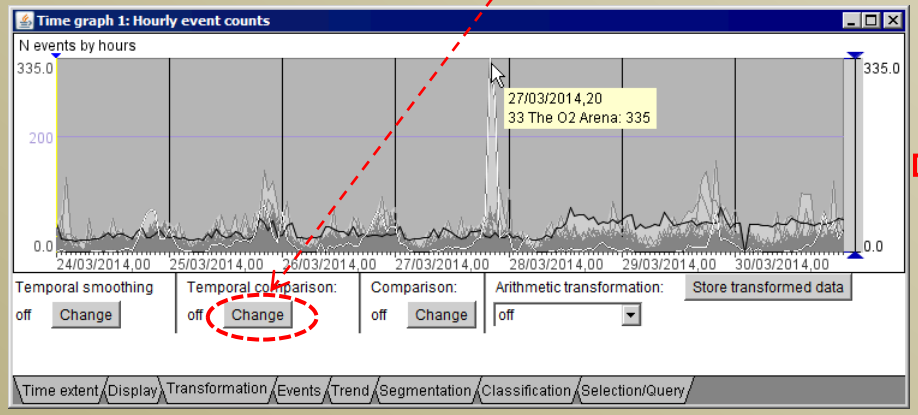


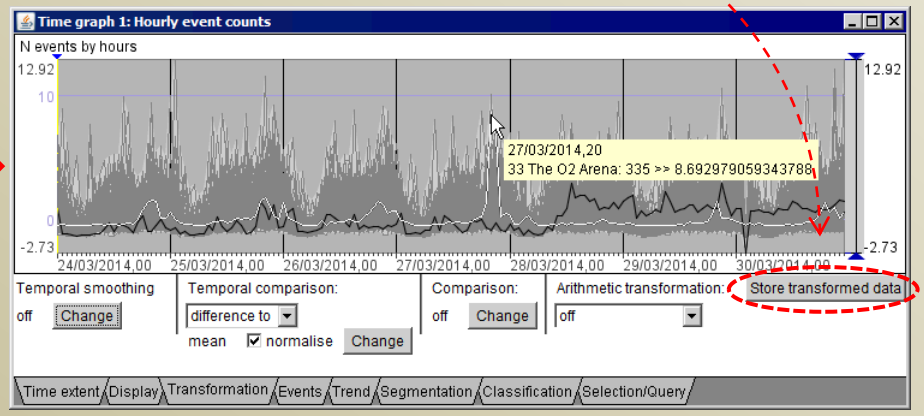
2.2. Transformation of time series - 1

From absolute values to relative with respect to the time series means

- In the time graph display, open the tab “Transformation”, find section “Temporal comparison”, press “Change”.
- Dialog “Temporal comparison” appears; select “mean” and “normalise”, press OK.
  - For each time series, the system computes the mean M and standard deviation S. Then each value V is substituted by V′ = (V – M) / S.
- Observe the change of the time graph.
- To put the transformation results in the table, use the button “Store transformed data” (extend the time graph window if not fully visible).



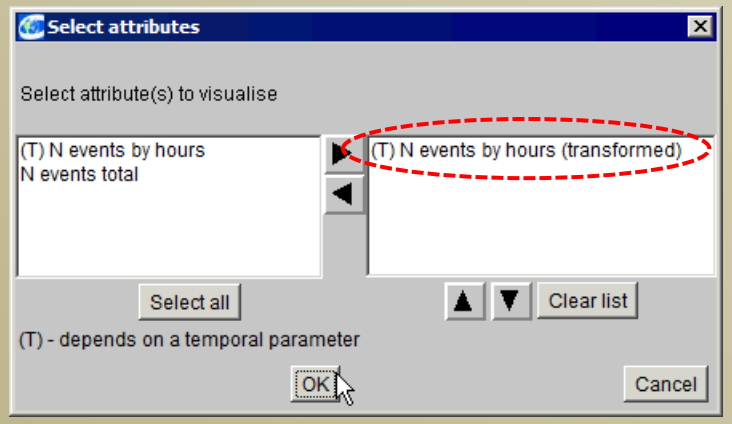


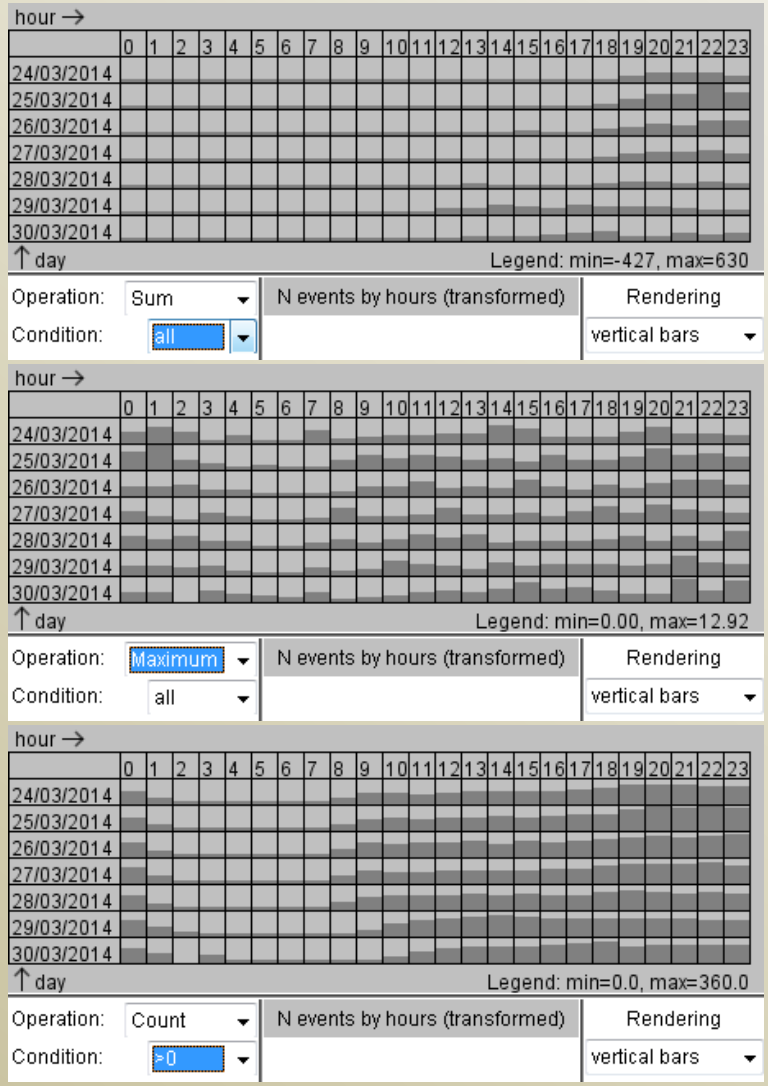


2.2. Transformation of time series - 2

Visualisation of transformed data in a 2d time histogram

- Open a 2d time histogram for the transformed time series (note that the table now contains two time series, original and transformed). Select different aggregation options (sum, maximum, count) and compare the histograms for the original and transformed time series.
  - Note: for “Sum”, select condition “All”; for “Count”, select condition “>0”

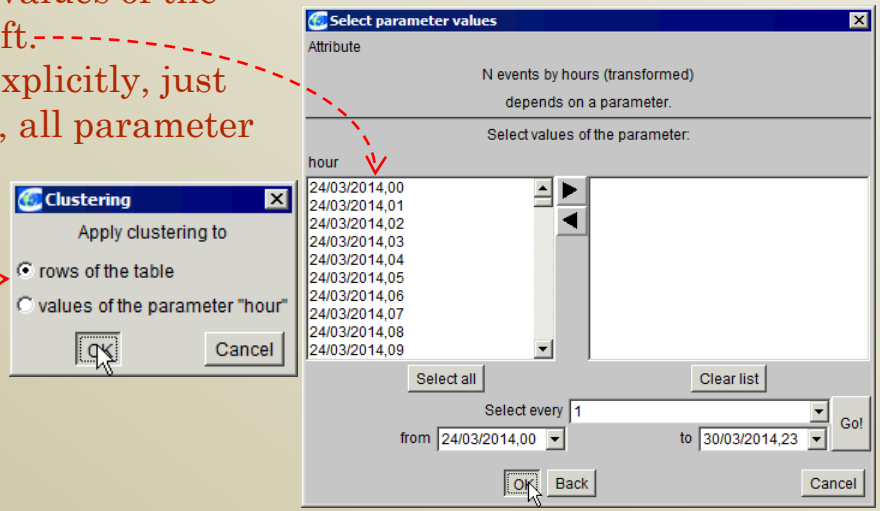


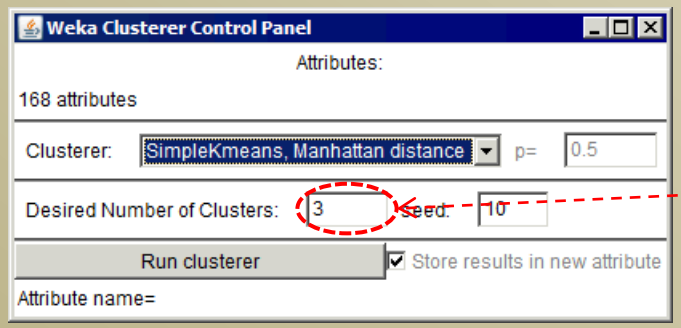


2.3. Partition-based clustering of places - 1

By similarity of their time series of attribute values

- Start the clustering tool *(it uses the partition-based clustering algorithm k-means)*
  - Menu “Analyse” > select “K-means clustering” > attribute selection dialog appears; select the time series “(T) N events by hours (transformed)”, i.e., the result of the transformation
  - Dialog “Select parameter values” appears. All values of the temporal parameter “hour” are listed on the left. You **do not need** to select parameter values explicitly, just **press OK**. When nothing is explicitly selected, all parameter values will be taken for the analysis.
  - Dialog “Apply clustering to ...” appears: make sure that checkbox “rows of the table” is selected; press OK.
  - You receive the following window with controls for setting clustering parameters:



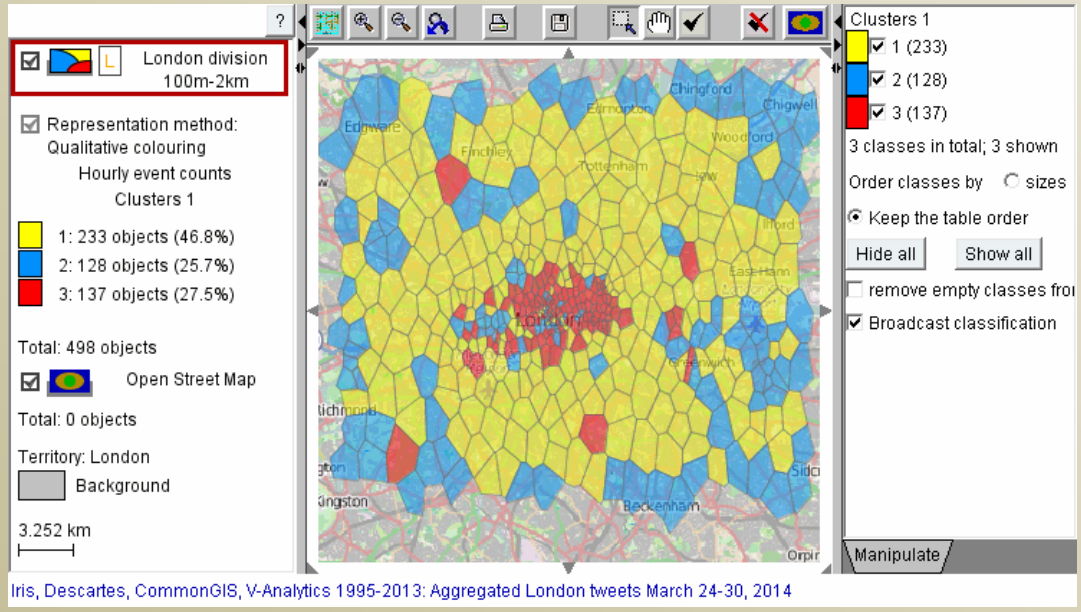


desired number of clusters

2.3. Partition-based clustering of places - 2

By similarity of their time series of attribute values

- Run the clusterer for the default number of clusters k = 3 (press button "Run clusterer")
- When the k-means algorithm finishes, the results are put in the table and visualised on the map.
  - A new column is created in the table. For each place, the number of the cluster it belongs to (1, 2, or 3) is written in this column.
  - The system automatically visualizes the clustering results by assigning each cluster some colour and painting the places in the map in these colours.
- Observe the spatial patterns of the colour distribution on the map.
  - Note: the cluster colours you will obtain may differ from the colours in this example.

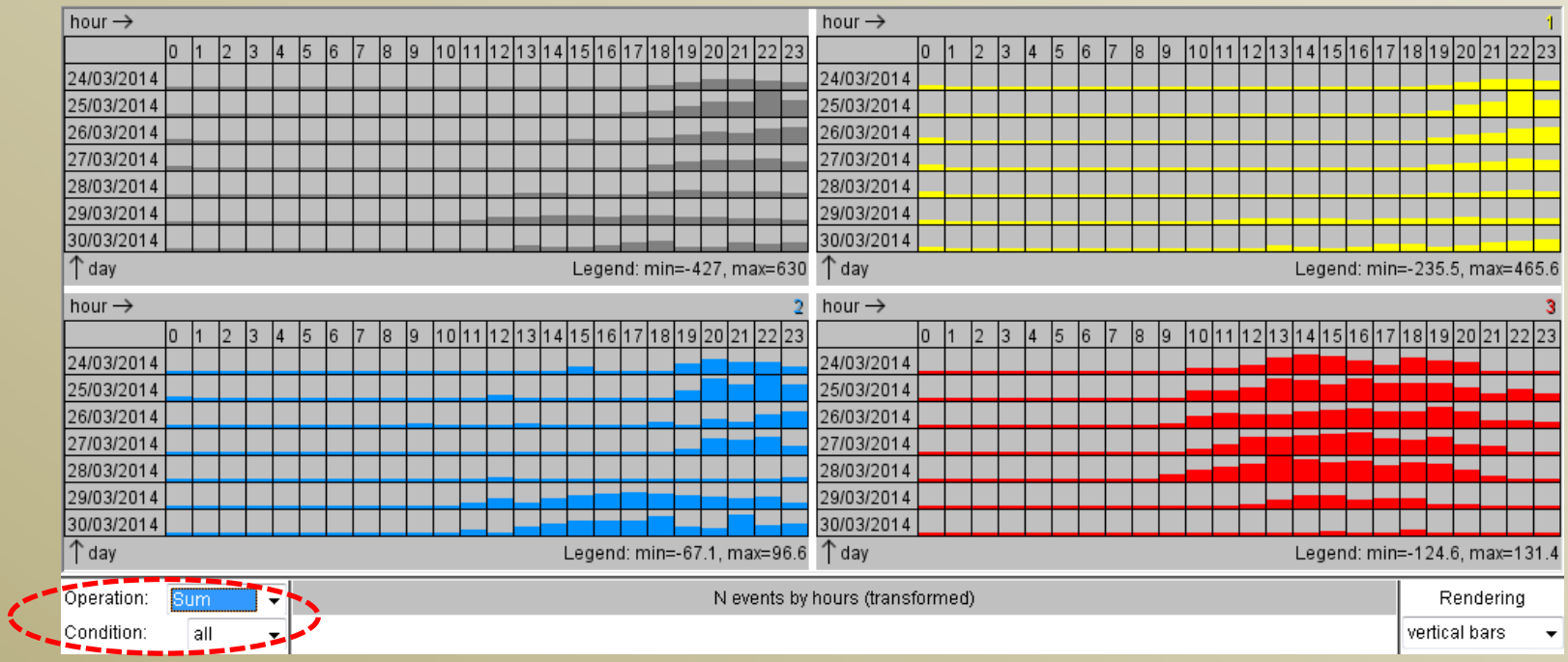


2.3. Partition-based clustering of places - 3

Interpretation of the clusters using 2d time histograms - 1

- Find the checkbox “Broadcast classification” to the right of the map and check it. Information about the cluster membership and colours will be transmitted to all displays, i.e., to the time graph and 2d time histograms.
- The 2d time histogram displays will multiply the histograms. To see all histograms, extend the windows of the displays.

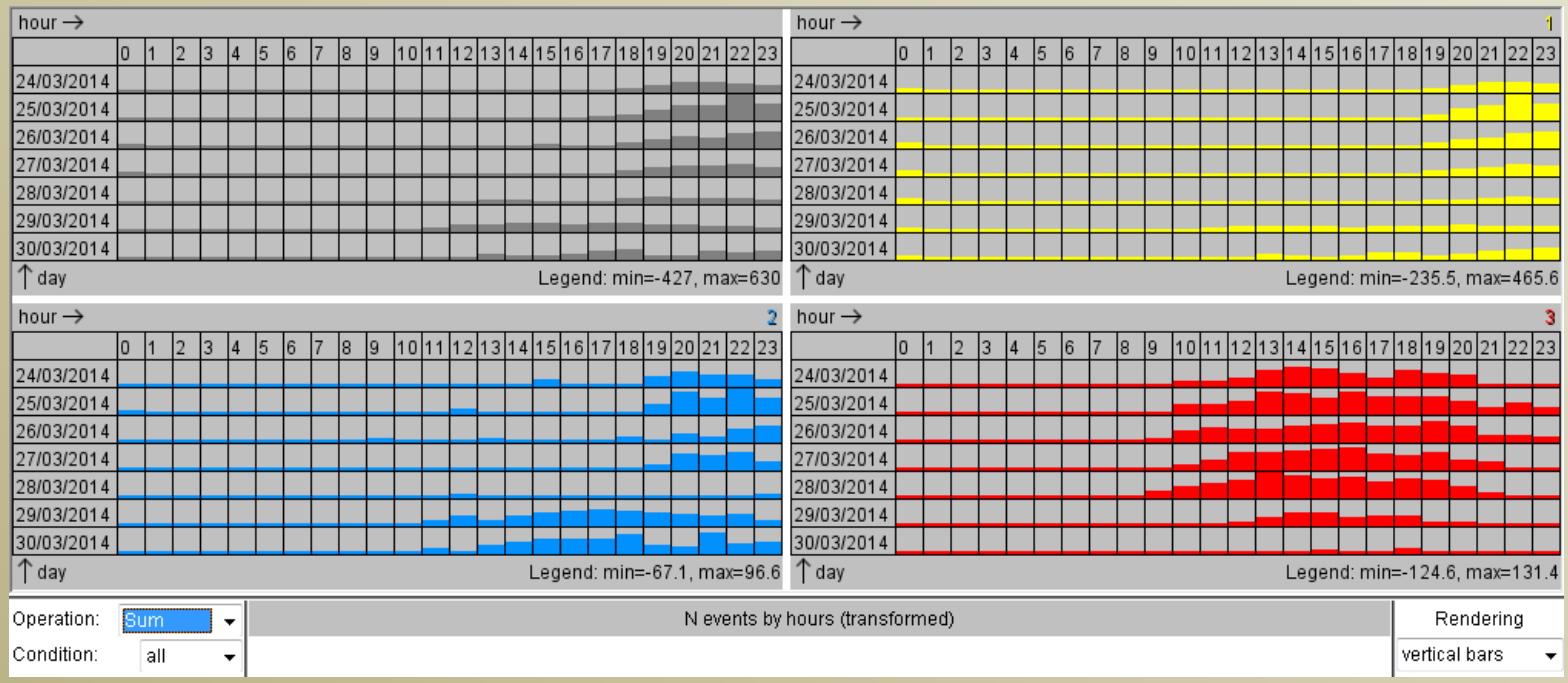
The first histogram (top left) corresponds to the whole set of places and the others to the clusters.



2.3. Partition-based clustering of places - 4

Interpretation of the clusters using 2d time histograms - 2

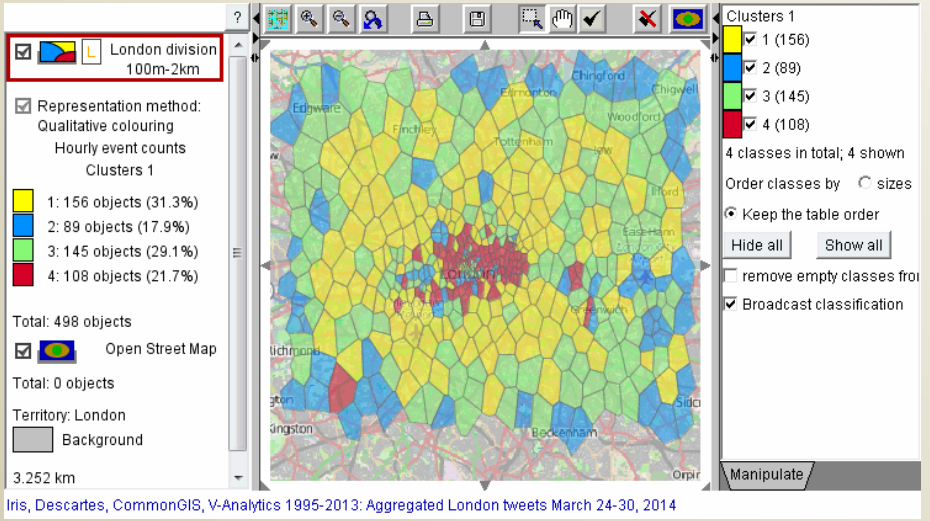
- Compare the weekly temporal patterns of the Twitter activities for the different clusters of places. Which temporal patterns might correspond to mostly residential areas, to business/work areas, to leisure areas? *(Note: the patterns you will get for your data may differ from the ones shown here)*.
- Find some places known to you (e.g., City University, Hyde park, ...), determine their cluster membership and the characteristic features of the respective temporal patterns.

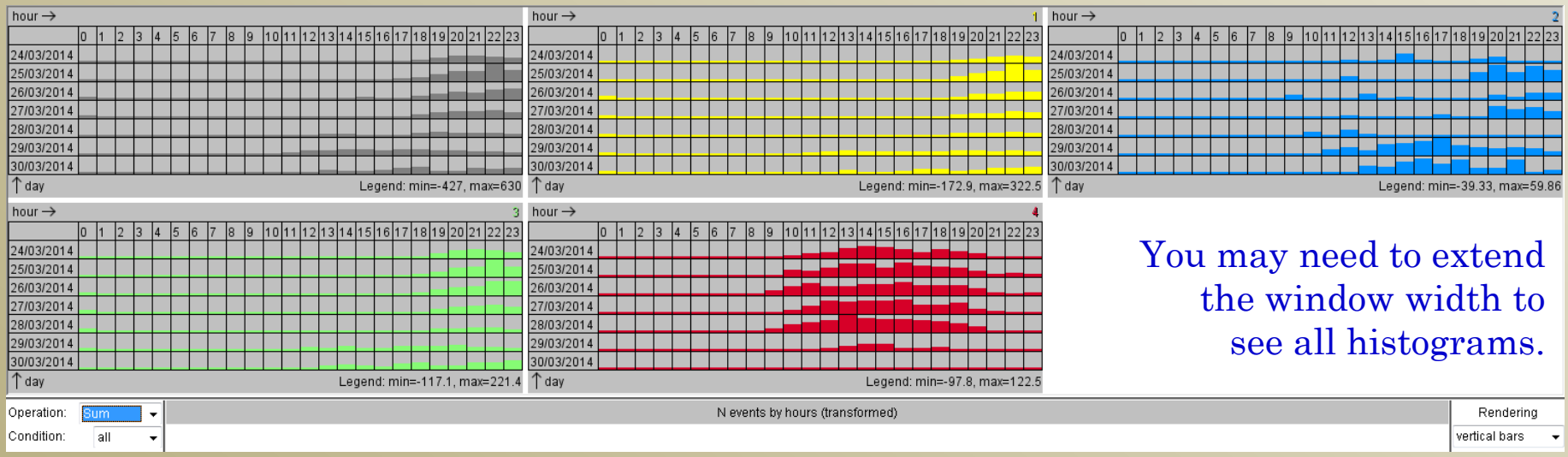


2.3. Partition-based clustering of places - 5

Testing the impact of parameter k (number of clusters)

- Set the desired number of clusters (k=4) in the clustering control window and run the clusterer again. The map, the time graph, and the time histograms are updated to show the new results.
  - Did the increase of the number of clusters result in uncovering new temporal patterns or in clearer distinctions between the patterns of the different clusters?
- Repeat the experiment for k=5.





You may need to extend the window width to see all histograms.

Questions for discussion

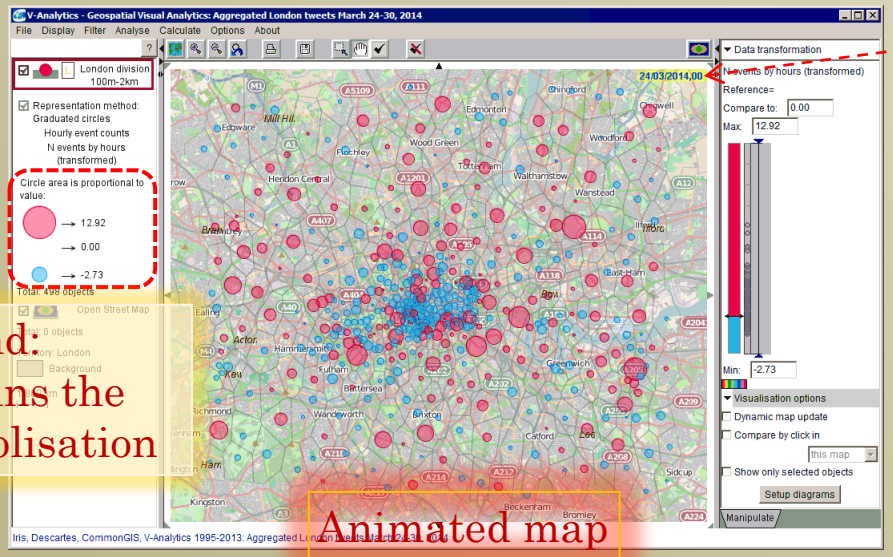
Topic: exploration of spatial time series

- Why do we need to combine a map with temporal displays (time graph, 2d time histogram)?
- When are 2d time histograms useful?
- What does partition-based clustering do?
- How does the clustering by the Twitter time series divide London? Does the central part differ from the remaining territory? If so, describe the differences.
- What value of parameter *k* (number of clusters) was sufficient for getting interpretable clusters of time series?
  - Re-cluster the time series with this parameter value and describe the result.

2.4. Partition-based clustering of times

Preparation

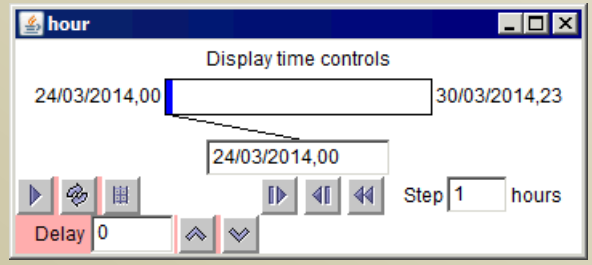
- Close all windows except the main window with the map.
- "Clean" the map, i.e., remove the place colouring
  - Menu "Display" > "Clean the map"
- Visualize the time series with an animated map.
  - Menu "Display" > "Display wizard" > attribute selection dialog appears; select the time series "(T) N events by hours (transformed)"; press OK > a dialog for selecting visualisation type appears; select "Animated map"; press OK > a dialog for selecting cartographic visualization method appears; select "Graduated circles"; press OK



current time step

Legend: explains the symbolisation

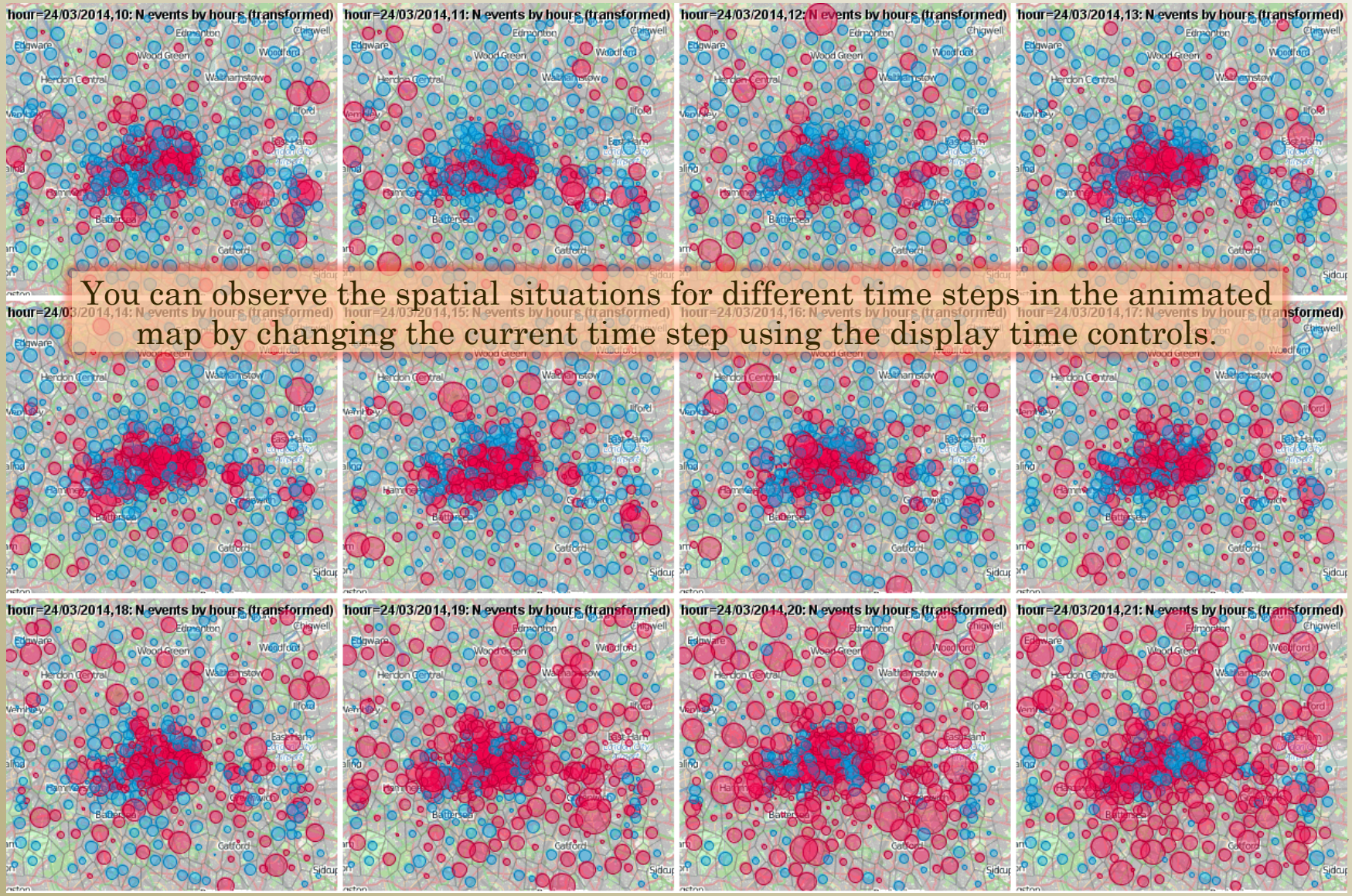
Animated map



Controls for changing current time step

Spatial situations

Spatial distributions of attribute values (Twitter activities) in different time steps

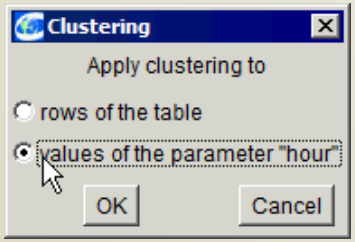


You can observe the spatial situations for different time steps in the animated map by changing the current time step using the display time controls.

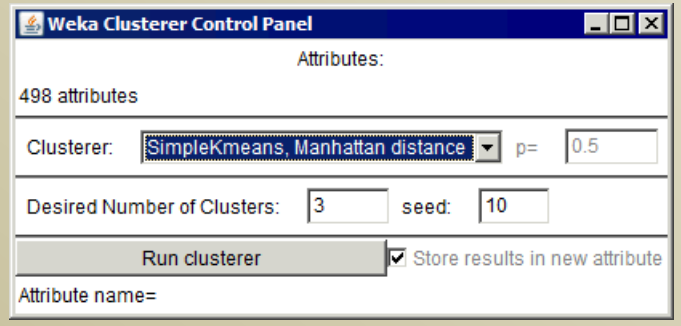
2.4. Partition-based clustering of times - 1

by the similarity of the spatial situations

- Start the k-means clustering tool for the values of the parameter "hour"
  - Menu "Analyse" > K-means clustering > table selection dialog appears; select table "Hourly event counts"; press OK > attribute selection dialog appears; select the time series "(T) N events by hours (transformed)"; press OK > dialog "Select parameter values" appears; press OK (no need to select values explicitly – all will be selected automatically)
  - Dialog "Apply clustering to ..." appears. Important: select the checkbox "**values of the parameter "hour"**" *(not "rows of the table"!)*; press OK
  - The window for setting clustering parameters appears, as previous time.



- This time, the clustering will be applied to table columns, which correspond to different values of the parameter "hour" (recall the table structure). More precisely, the system will create a new table, in which the rows and columns are transposed (times > rows; places > columns), and send this new table to the clusterer. The clusterer will divide the set of time steps into groups according to the similarity of the distributions of the attribute values across the set of places.

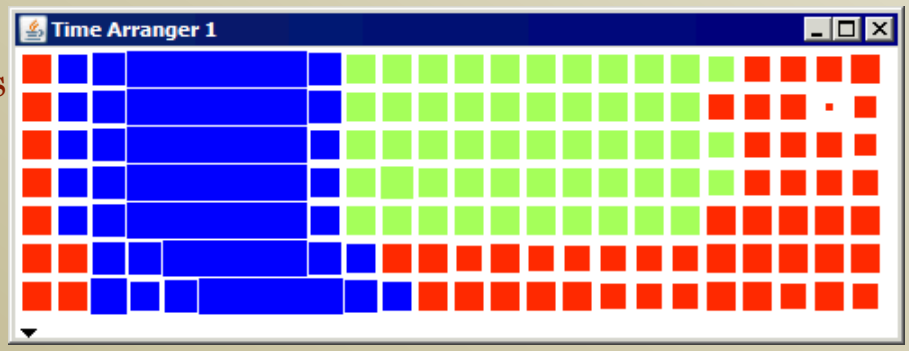


2.4. Partition-based clustering of times - 2

by the similarity of the spatial situations

- Run the clusterer with $k = 3$ (*k* is the desired number of clusters)
- After obtaining the results, the system automatically visualises them in a Time Arranger display *(you may get different colours than shown in this example)*

rows: days



columns: day hours

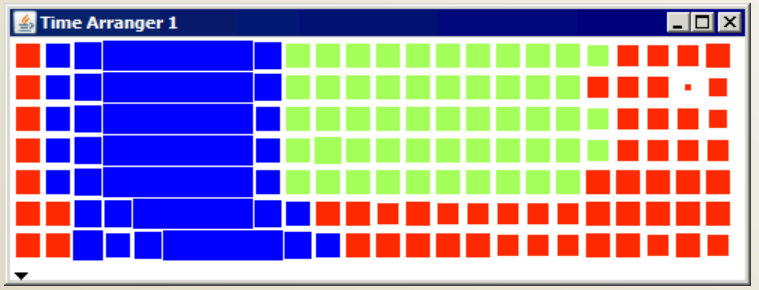
colours: cluster membership of the hourly time intervals

square sizes: how typical the spatial situation is for this cluster (i.e., how close it is to the cluster's average)

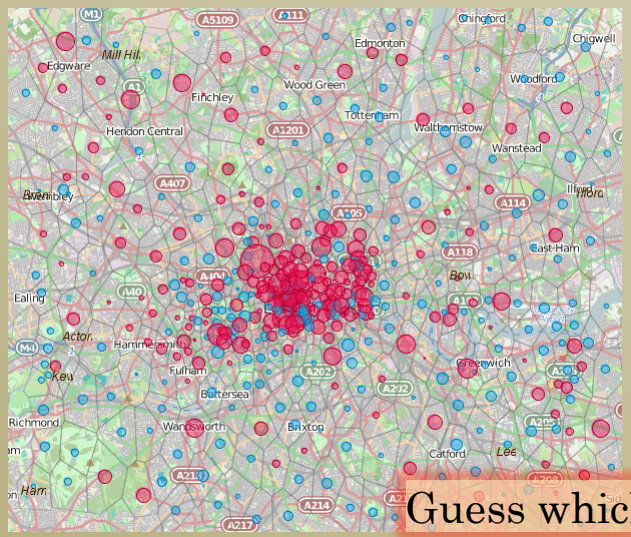
- To see examples of situations for some cluster, click on the squares coloured in this cluster's colour. The animated map will show the situations for the time intervals represented by the squares.

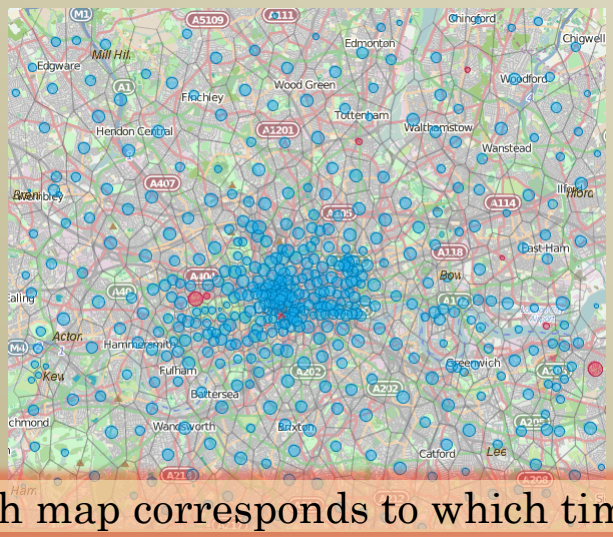
2.4. Partition-based clustering of times - 3

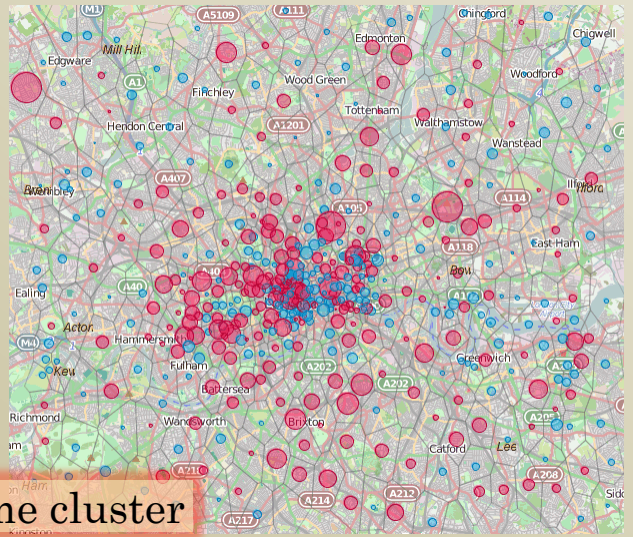
Interpretation of the time clusters



- The pattern you will see in the time arranger display for your data may slightly differ from this example, but, very probably, you will also see a division of the time steps into morning, working day hours, and evenings + weekend, and high periodicity of the overall pattern.
- Describe and compare the typical spatial situations (distributions of the Twitter activities) for these groups of time steps.
  - Note that the **red**-coloured circles on the map represent values that are **higher** than the *place averages*; the **blue**-coloured circles represent **lower** values.







Guess which map corresponds to which time cluster

2.4. Partition-based clustering of times - 4

Testing the impact of parameter k (number of clusters)

- Run the clusterer with $k = 4, 5, \ldots$ Observe in the time arranger how the weekly pattern is refined with increasing k.

Questions for discussion

Topic: spatial time series

- What are two different possible perspectives for looking at spatial time series?
  - The different perspectives can be seen in different visualizations: time graph (containing a curve for each place) and animated map (showing a spatial distribution for each time interval).
- What is the difference between the two experiments on applying partition-based clustering?
- What aspects of the overall behaviour of the Twitter activities in London could be studied in these two experiments?
- What did you learn about the Twitter activities over London at different times?