Module INM433 – Visual Analytics

Practical 04

Density-based clustering of spatial events and OD moves

given by
prof. Gennady Andrienko and
prof. Natalia Andrienko

1

General description

- Two major topics:
  - Application of density-based clustering to **spatial events**
  - Application of density-based clustering to OD moves
- Data:
  - Spatial events: a sample of geo-located tweets from one day
    - 2-3 student groups work with different datasets (same territory, different days and weeks)
  - OD moves: a sample of the London bike trips from 25/07/2012 (Wednesday)
- Software: V-Analytics

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

2

Topic 1: density-based clustering of spatial events

- 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"
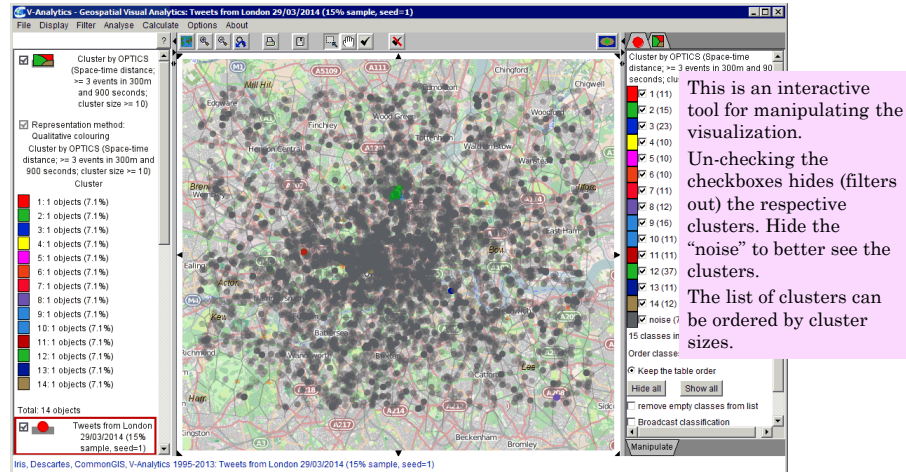 - The events are loaded and shown on a map
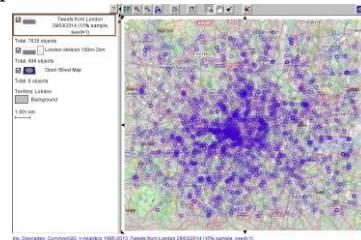
3

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.

4

Density-based clustering of events - 2

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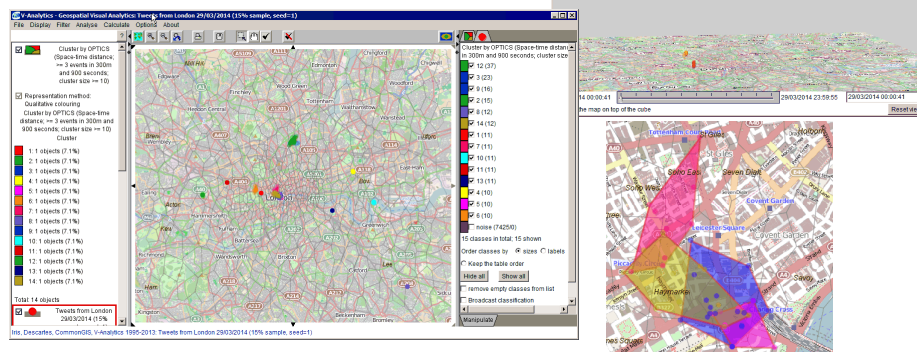

This is an interactive tool for manipulating the visualization.

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.

5

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)?

6

3

Exploration of clusters using visual displays

Viewing cluster summaries

- 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).

7

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).

8

Questions for discussion

- 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?

9

# Topic 2: density-based clustering of OD movement data

- Preparation to the exercise
  - Load project "trips_Wednesday.app" from folder practicals/04_DB_clustering/OD_moves
    - Menu "File" > "Load project" > button "Browse" > open folder practicals/04_DB_clustering/OD_moves > load file "trips_Wednesday.app"
  - The data are loaded and shown on a map
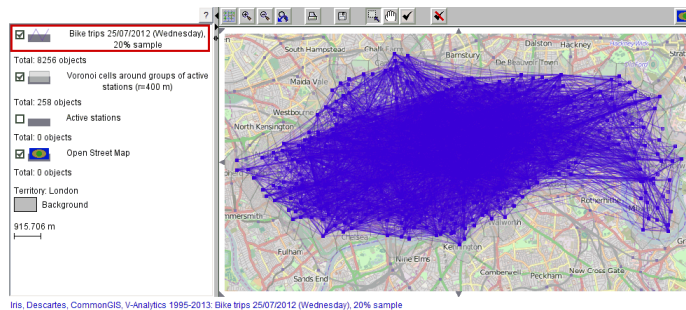
10

# Exercise 2.1: Visual representation of OD moves as spatio-temporal objects
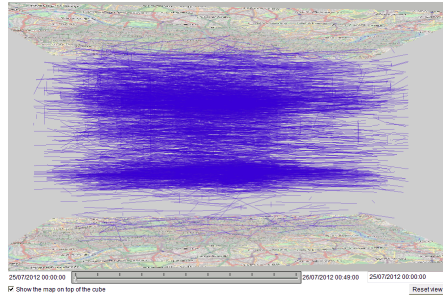
1. Open a space-time cube view with the bike trip data *(Display > Space-time cube)*.
2. For temporal zooming in the STC, use the time filter *(Filter > Time filter)*.
   • Manipulate the length and position of the time slider.
3. Pay attention to the differences in the line slopes. What do they mean?
4. Observe the variation of the line density throughout the day.

11
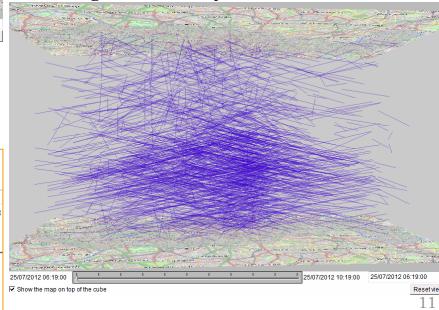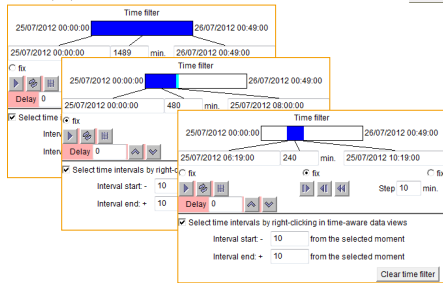
Find vectors representing round trips

- Use attribute-based filter
  - Filter > Attribute-based filter > Select table "Bike trips 25/07/2012…" >Select attribute "Track length" > set the upper bound to 0.2 km
- Clear the time filter
- To better see the vectors in the STC, increase the line thickness
  - For changing the drawing settings, click on the layer's icon in the map legend.
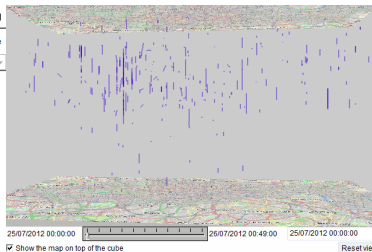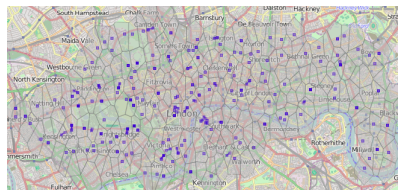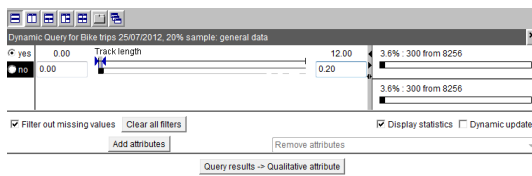
12

Question to exercise 2.1

- What characteristics of the trips are represented by the following properties of the lines (vectors)?
  - length in the spatial dimension
  - length (height) in the temporal dimension
  - inclination

13

Exercise 2.2: DB clustering of OD moves by the spatial positions of their starts and ends

- Invert the attribute-based filter (i.e., filter out the round trips)
- Activate the clustering tool:
  - Menu "Analyse" > "OD moves: density-based clustering" > a dialog appears; the layer with the moves is pre-selected > press OK
- Set the clustering parameters
  - Uncheck the check box "start and/or end times" (the times will not be taken into account in this exercise).
  - Set the spatial distance threshold to 300m.
  - Set the minimal size of the clusters you are interested in to 10 (in the text field following the check box "Ignore clusters with less than" at the bottom of the dialog).
  - Pressing OK starts the clustering.
- After the clustering finishes, the system shows the results by colouring the lines on the map and in the STC according to the cluster membership.

Dimensions of the trip set:
X-extent: 15359.13 m
Y-extent: 6559.44 m
Time span: 25/07/2012 00:00:40 .. 26/07/2012 00:48:03
Number of active trips: 7956 (96.37% of the total 8256 trips)
The trips to be clustered by
☑ start positions
☑ end positions
☐ start and/or end times ⊙ absolute ○ time of the day
Define the spatial and temporal neighbourhood of a trip:
Spatial distance threshold: 300   m
Temporal distance threshold: 900   seconds ▾
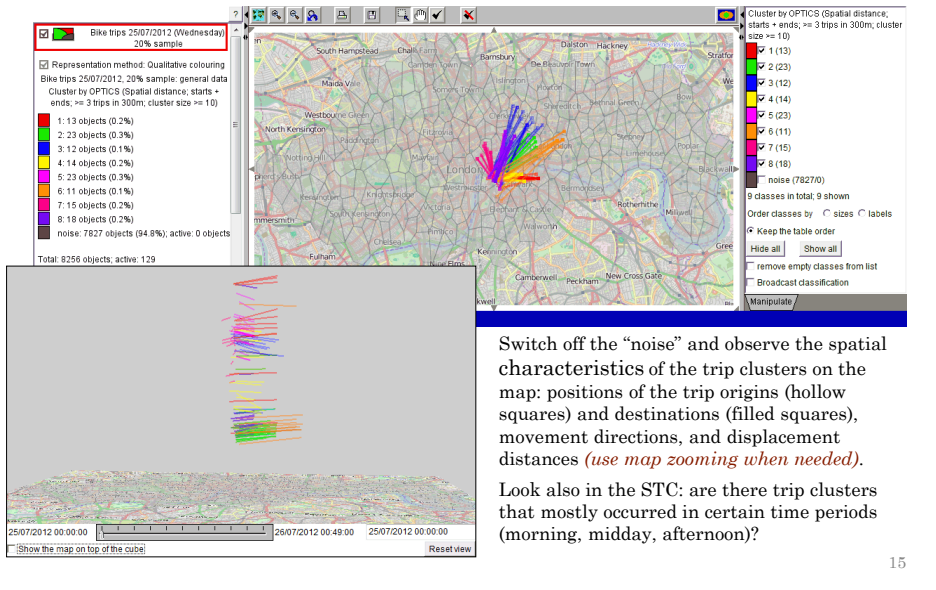☐ Use additional attributes of the trips
Minimal number of trips in the neighbourhood *: 3
                        * required for a trip to be in cluster core
☑ Ignore clusters with less than 10   trips
                    OK   Cancel

14

Exploration of the clustering results

Switch off the "noise" and observe the spatial characteristics of the trip clusters on the map: positions of the trip origins (hollow squares) and destinations (filled squares), movement directions, and displacement distances *(use map zooming when needed)*.

Look also in the STC: are there trip clusters that mostly occurred in certain time periods (morning, midday, afternoon)?

15

Testing the impact of the distance threshold

- Switch on the "noise", to have all moves visible.

- Move the current visualisation to another window:
  - "Display" > "Move the map to another window"

- Start again the clustering tool as previously

- In the dialog for setting the parameters, uncheck again the check box "start and/or end times" and change the spatial distance threshold to 350 m instead of the previous 300 m.

- Run the clustering algorithm and visually explore the results. Compare with the results of the previous clustering.

16

Questions to Exercise 2.2

- Describe the differences between the results of the clustering with the different distance thresholds. Consider the following aspects:
  - Number of clusters
  - Sizes of the clusters
  - Amount of "noise": how many moves that were earlier in the "noise" have been included in clusters?
    - Hint: switch off the "noise" in the new map and look in the legend of the older map.
  - Spatial properties of the additional clusters.
  - Internal variance within the clusters.
    - Hint: select several largest clusters one by one after ordering the clusters by the sizes.

17

# Exercise 2.3: Density-based spatio-temporal clustering of OD moves

- Switch on the visibility of all clusters and "noise".

- Move the current visualisation to another window:
  - "Display" > "Move the map to another window"

- Start again the clustering tool as previously

- In the dialog for setting the parameters:
  - The check box "start and/or end times" must be checked – now the times will be taken into account.
  - Set the temporal distance threshold to 30 minutes.
  - Set the minimal cluster size to 5.

- Run the clustering algorithm and visually explore the results using the map and the space-time cube.

Dimensions of the trip set:
X-extent: 15359.13 m
Y-extent: 6559.44 m
Time span: 25/07/2012 00:00:40 .. 26/07/2012 00:48:03
Number of active trips: 7956 (96.37% of the total 8256 trips)
The trips to be clustered by
☑ start positions
☑ end positions
☑ start and/or end times ⦿ absolute ○ time of the day
Define the spatial and temporal neighbourhood of a trip:
Spatial distance threshold: 350 m
Temporal distance threshold: 30 minutes ▾
☐ Use additional attributes of the trips
Minimal number of trips in the neighbourhood *: 3
                * required for a trip to be in cluster core
☑ Ignore clusters with less than 5 trips
        OK   Cancel

18

Questions to exercise 2.3

- Describe the spatial and temporal characteristics of the spatio-temporal clusters of OD moves.

- How are the largest spatio-temporal clusters (size>=10) related to the earlier obtained spatial clusters (results of the previous run)?
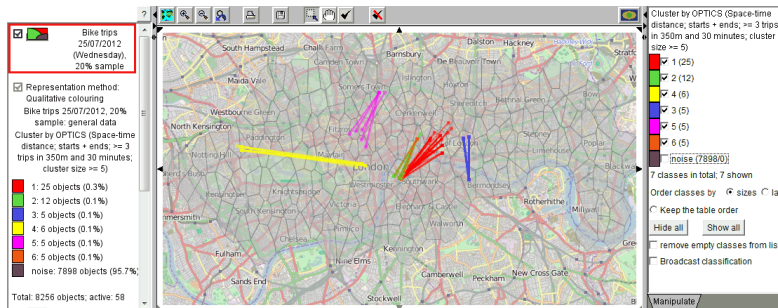

19

Questions on density-based clustering

- For what purposes did we apply DBC?

- What distance functions did we use?

- How do the distance functions differ in terms of
  - types of data they can be applied to?
  - clustering outcomes (meaning of the clusters)?

- Compare the complexity of the different distance functions and explain the differences

20

## Exercise 2.4: Spatio-temporal aggregation of OD moves

- Cancel all filters (cluster selection and attribute-based filter).

- Close all additional windows (leave only the main window) and clean the main map (remove the cluster visualisation).

- Start the aggregation tool:
  - "Analyse" > "OD moves: spatio-temporal aggregation" > a dialog appears with two pre-selected map layers (with the moves and with the space compartments) > press "OK"
  - A time division dialog appears; the temporal resolution "hour(s)" and the interval length "1" are proposed at the bottom > Press "Divide" > A dialog showing the number of breaks appears > press "Yes" > The list of breaks appears in the list box > Press "OK" > A dialog asking about finding useless breaks appears > Press "Yes" > The tool informs about removing a useless break > Press OK
  - Press "OK" in each of the following dialogs (i.e., agree to the use of the default settings) .
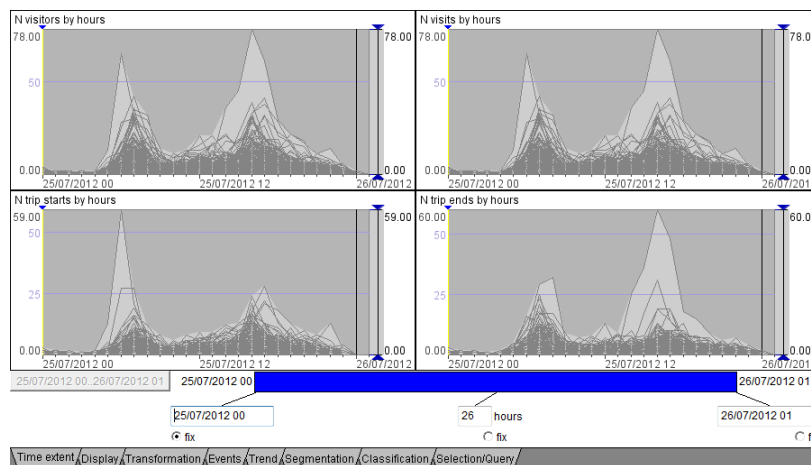
21
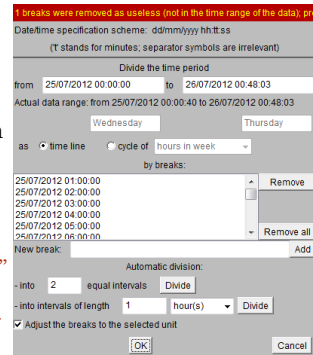
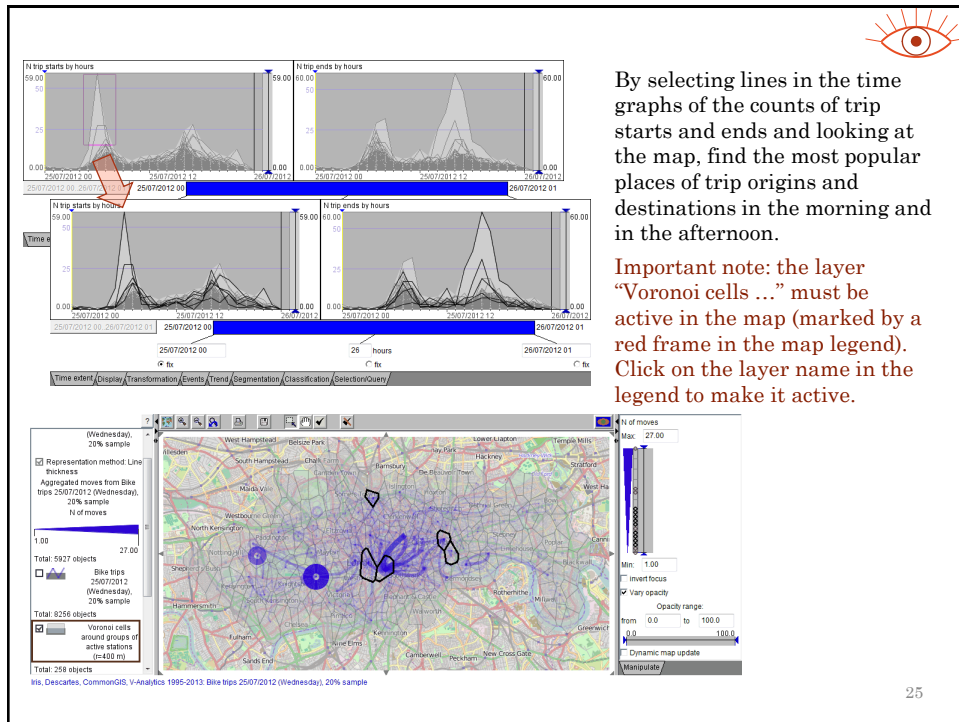## Visual representation of flows resulting from the aggregation
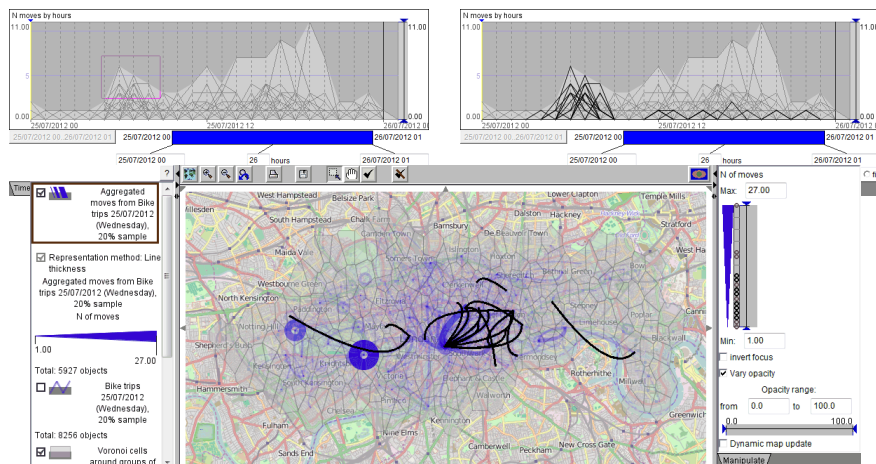

22
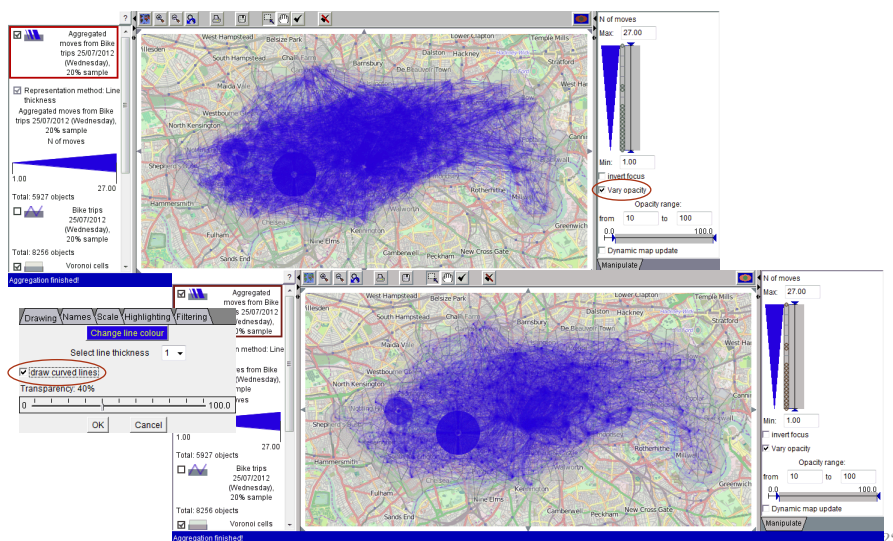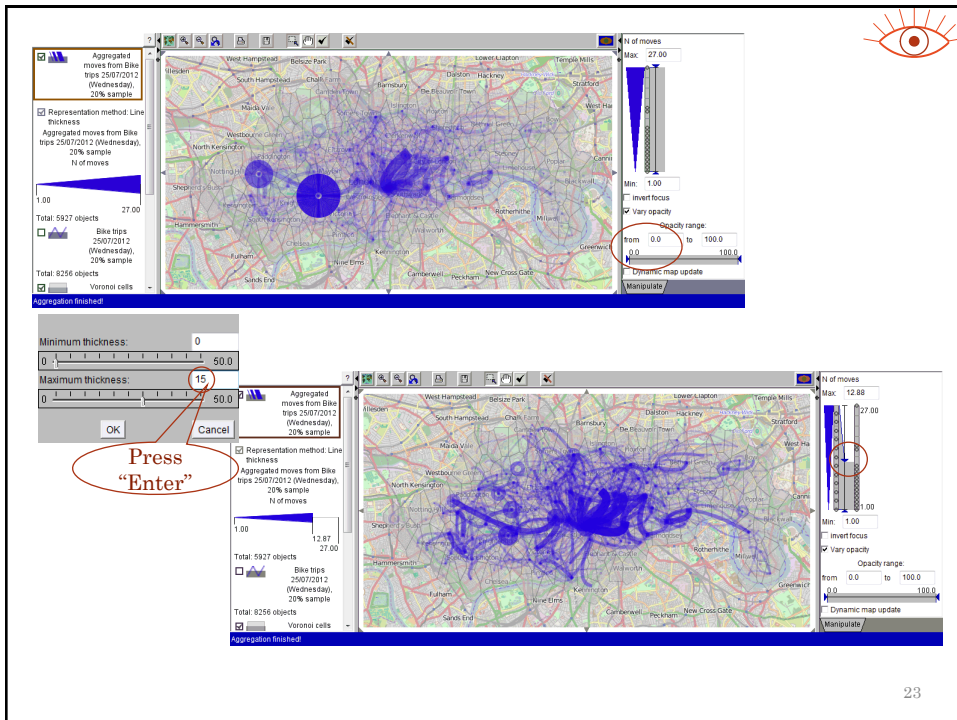
Place-referenced time series

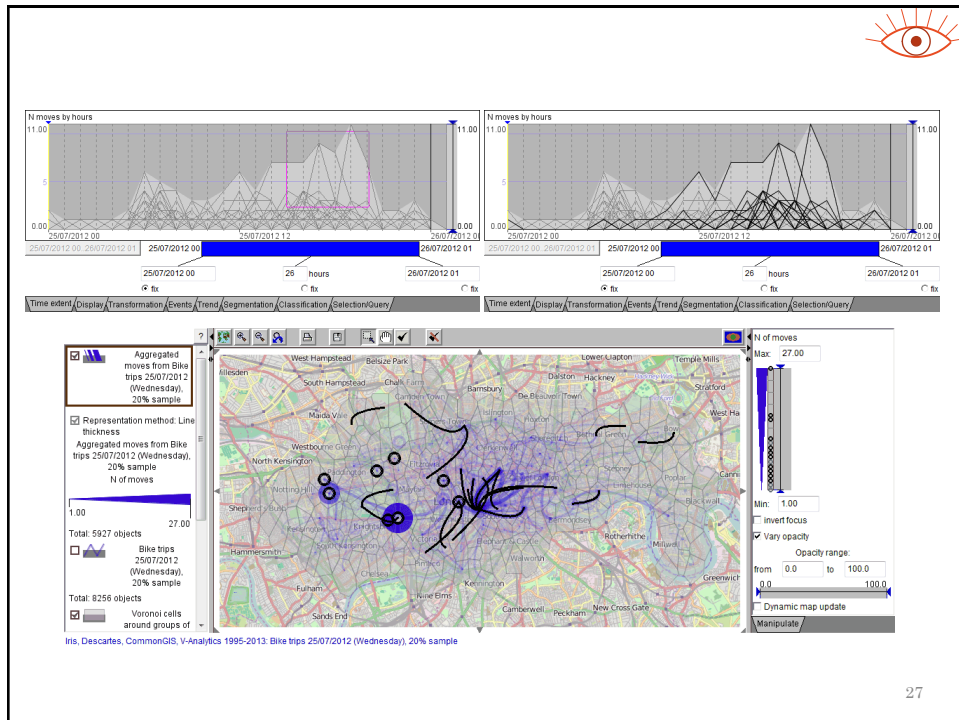Display > Display wizard > Select table "Voronoi cells …" > Select time-variant attributes > Select "Time graph"

By selecting lines in the time graphs of the counts of trip starts and ends and looking at the map, find the most popular places of trip origins and destinations in the morning and in the afternoon.

Important note: the layer "Voronoi cells …" must be active in the map (marked by a red frame in the map legend). Click on the layer name in the legend to make it active.

Iris, Descartes, CommonGIS, V-Analytics 1995-2013: Bike trips 25/07/2012 (Wednesday), 20% sample

25

Link-referenced time series

Create a time graph for table "Aggregated moves from bike trips …" and attribute "N moves by hours". Select the highest morning flows, then the highest afternoon flows. Before that, make sure that the layer "Aggregated moves …" is active in the map.

26

27

Questions to Exercise 4

- Describe the main features of the collective movement behaviour that could be learned by exploring the aggregated data:
  - Popular trip origins and destinations in different times of the day
  - Frequencies of trips within the same areas (represented by rings)
  - Major flows in different times of the day
    - origins, destinations, directions

28

Try on your own

- Bike trip data from one week (03-09/09/2012) aggregated by hourly intervals
  - Project file aggr_hours_week_2012_09_03-09.app
- Try partition-based clustering for place-based and link-based time series
  - The same operations as for the aggregated events in practical 3 can also be used for these data.
  - Generally, the partition-based clustering tools are used in standard ways for spatial time series of any origin.

29