Module INM433 – Visual Analytics

Lecture 04

Use of density-based clustering

given by
prof. Gennady Andrienko and
prof. Natalia Andrienko

Content and objectives

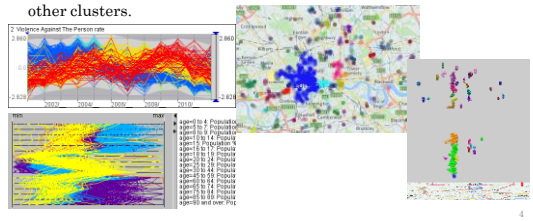
- The concept of density-based clustering will be introduced and the difference from the partition-based clustering explained.
- Possible uses of density-based clustering in analysis of spatial events and origin-destination movement data will be shown.
- The use of different distance functions depending on the data type and aims of analysis will be explained.
- We shall also consider how origin-destination movement data can be aggregated in space and time.

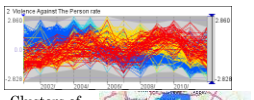
Two major types of clustering

Partition-based and density-based

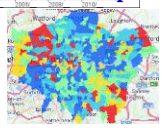
Generally, what is clustering?

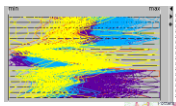
- Loose definition: clustering is the process of organising objects into groups whose members are close or similar in some way.
- A cluster is a group of objects which are "similar" or "close" between them and are "dissimilar" or "distant" to the objects belonging to other clusters.



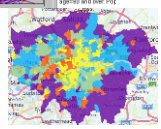


Clusters of similar time series

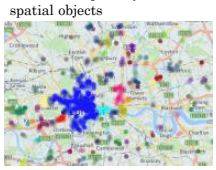




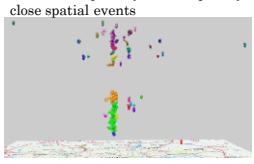
Clusters of similar multi-attribute value combinations



Clusters of spatially close spatial objects



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

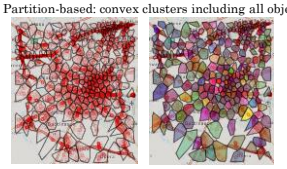
Two major types of clustering

- **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"

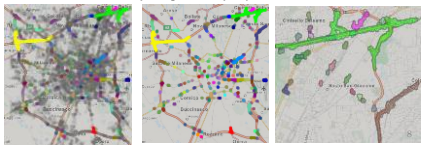
7

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

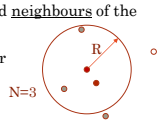
- **Partition-based:**
  - Typically applied to multiple thematic (non-spatial) attributes or to time series of thematic attributes
  - Objective: divide objects into groups such that objects within a group have similar attribute values and differ from the objects in the other groups
- **Density-based:**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

9

Density-based clustering (DBC)

Goal: find dense groups of close or similar objects

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



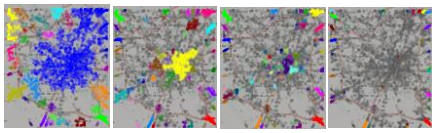
- To make a cluster:
  1) some core object with all its neighbours is taken;
  2) for each core object already included in the cluster, all its neighbours are also added to the cluster (if not added yet).
- Some objects may remain out of any cluster (when they have not enough neighbours and do not belong to the neighbourhood of any core object). These objects are treated as "noise".

10

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.



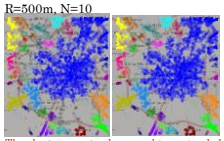
Grey: "noise"

11

Exploring the impact of the DBC parameters

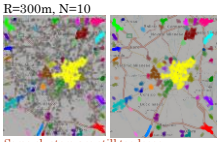
Example: DBC according to the spatial distances between point objects

R=500m, N=10



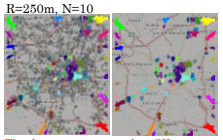
The clusters are too loose and too extended in space.

R=300m, N=10



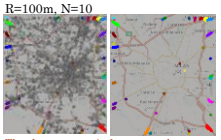
Some clusters are still too loose.

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.

12

Questions?

Density-based clustering; difference between two types of clustering

13

Density-based clustering of spatial events

by positions in space and time

14

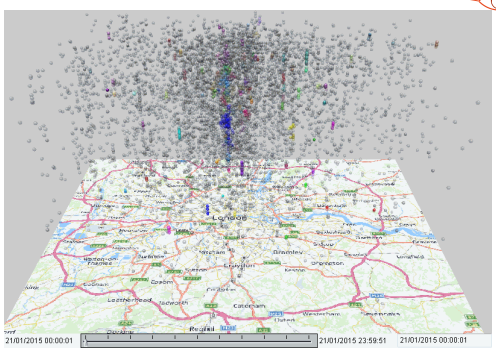
DBC by spatio-temporal distances

- 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_{space} = 300 m and R_{time} = 30 minutes.
- However, the clustering algorithm requires a single measure of distance and a single radius.

⇒ Spatial and temporal distances need to be combined together

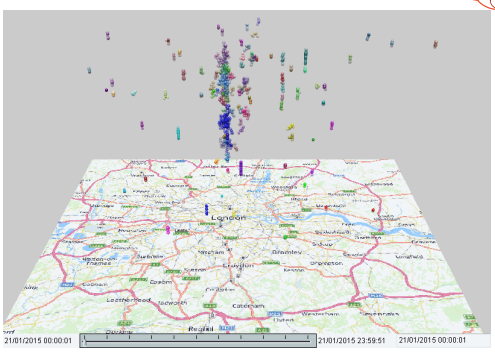
- e.g., $d = \max(d_{space}/R_{space}, d_{time}/R_{time}) * R_{space}$

15



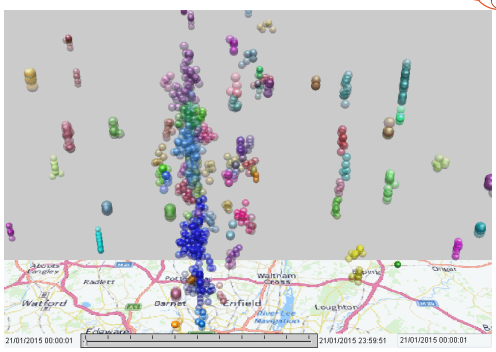
Example: clusters of London tweet events (21/01/2015; 15% sample) with R_{space}= 500 m, R_{time}= 20 minutes, and N = 3. Grey: noise.

16

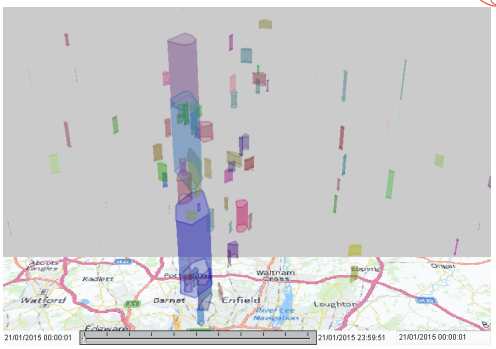


Example: clusters of London tweet events (21/01/2015; 15% sample) with R_{space}= 500 m, R_{time}= 20 minutes, and N = 3. The noise is filtered out.

17

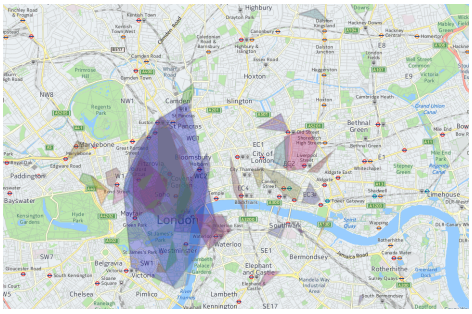


18



Spatio-temporal convex hulls of the event clusters.
Each cluster itself is a *spatial event* extended in space and time.

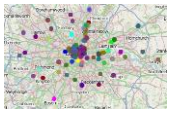
19



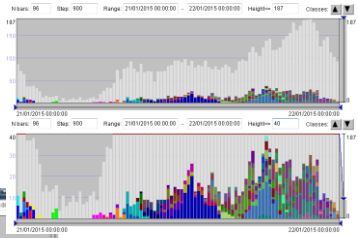
Spatio-temporal convex hulls of the event clusters projected on a map.

20

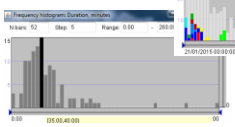
Investigation of clustering results



Spatial distribution of the dense clusters



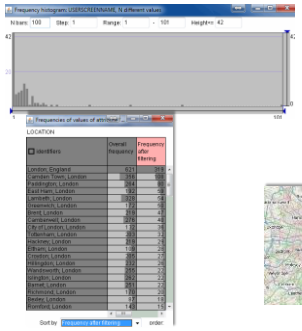
Temporal distribution of the dense clusters



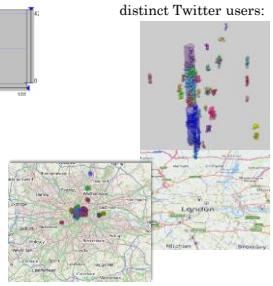
Cluster durations

21

Summarization of attribute values by the clusters



Clusters with at least 5 distinct Twitter users:



22

Extending the spatio-temporal distance function to additional thematic attributes

Distance function:

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

(a) – neighbourhood defined as a cube

(b) – neighbourhood defined as a sphere

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

Distance in time
(t_1, t_2 are intervals):

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

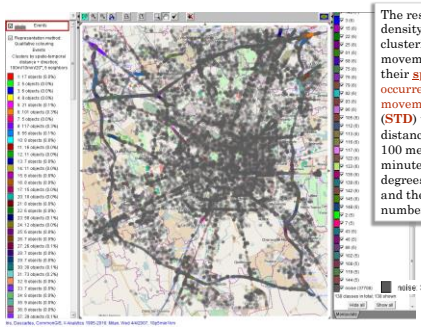
Distance for a cyclic attribute
(V is the cycle length):

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

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

23

Example: clusters of low speed car movement events

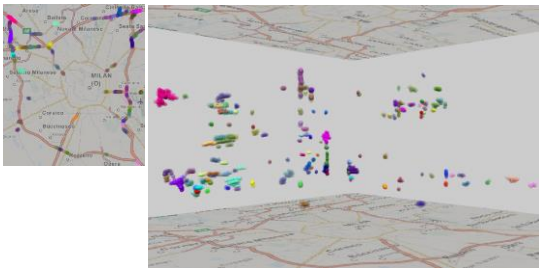


The result of the density-based clustering of the slow movement events by their spatial positions, occurrence times, and movement directions (STD) with the distance thresholds 100 meters, 10 minutes, and 20 degrees, respectively, and the minimum number of neighbors 5.

24

The STD-clusters, noise hidden

Each cluster represents a possible traffic congestion



25

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, …

26

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

27

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.

28

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

29

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

30

Questions?

Density-based clustering of spatial events

31

DBC of origin-destination movement data

32

Origin-destination movement data

Shortly: OD moves

- Movement data specify spatial positions of moving objects at different times
- Origin-destination data: specify only the positions and times of trip starts and ends; intermediate positions are not available.

33

A running data example

Use of shared bikes in London

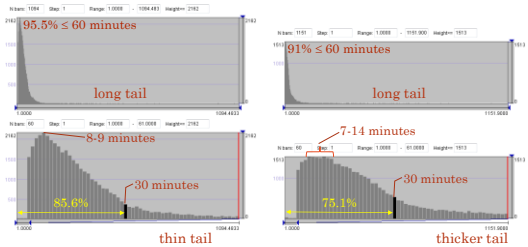
- Barclays Cycle Hire
- 569 docking stations to pick up and return bicycles
- The data are publicly available from **Transport for London** (http://www.tfl.gov.uk/info-for/open-data-users/our-feeds)
- Data describe the journeys made with the rented bikes:
  - Journey ID
  - Bike ID
  - Start date & time
  - End date & time
  - Start docking station ID
  - End docking station ID
    - The geographic coordinates of the docking stations are known and can be joined with the transaction data

34

OD moves as generic objects

Examples of analysing the attribute value distributions

Frequency distributions of the bike trip **durations** on 25/07/2012 (Wednesday; 41,380 trips) and 28/07/2012 (Saturday; 40,990 trips)

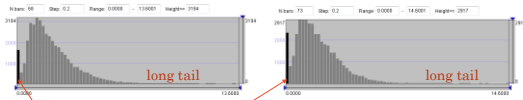


35

OD moves as generic objects

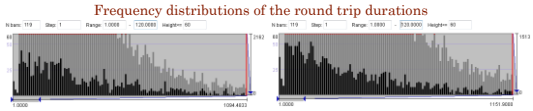
Examples of analysing the attribute value distributions

Frequency distributions of the bike **displacement distances** on 25/07/2012 (Wednesday; 41,380 trips) and 28/07/2012 (Saturday; 40,990 trips)



Round trips (possibly, for leisure): It makes sense to analyse them separately.

Frequency distributions of the round trip durations

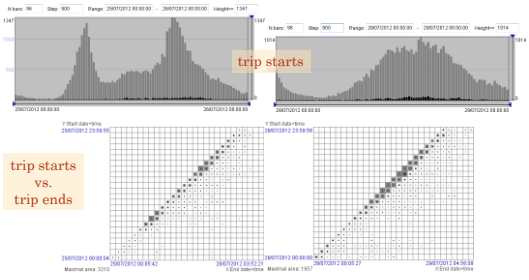


36

OD moves as temporal objects (events)

Analysis and comparison of temporal distributions

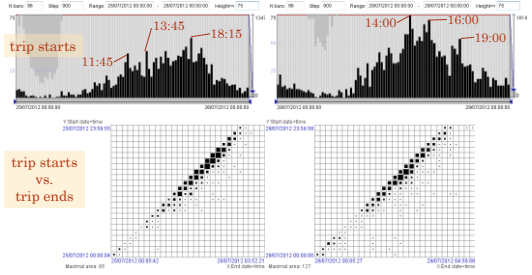
25/07/2012 (Wednesday; 41,380 trips) vs. 28/07/2012 (Saturday; 40,990 trips)



37

Temporal distributions of the round trips

25/07/2012 (Wednesday; 1,641 trips, 4%) vs. 28/07/2012 (Saturday; 2,322 trips, 5.7%)

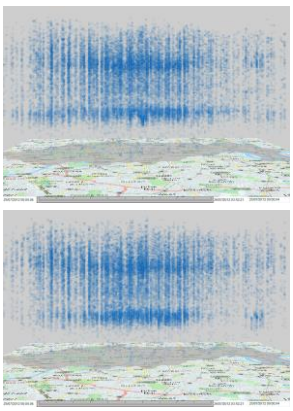


38

Relation of OD moves to spatial events

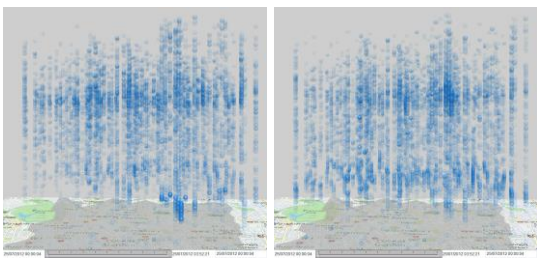
- Spatial events are objects having positions in space and times of occurrence or existence (= positions in time).
  - Space and time (and space × time) can be treated as containers of events.
  - Analysis is generally concerned with the spatio-temporal distribution of events or values of their thematic attributes.
- An OD move includes 2 instant spatial events: start and end.
  - The spatio-temporal distributions of these events may be of interest in analysis.
  - These events can be analysed with the same aims and the same methods as any other spatial events.

39



Example: instant point events extracted from the London bike trip data from Wednesday, 25/07/2012. The upper STC image shows the spatio-temporal distribution of the trip start events and the lower image represents the trip end events.
There are noticeable differences on the east; the other parts of the territory need to be explored with the help of spatial filtering, to decrease the display clutter and overplotting.

40



After applying spatial filtering and zooming in to the central part of the territory, the differences between the spatio-temporal distribution of the start (left) and end (right) events in this part become more vivid.

All visualisation and analysis methods applicable to point events can be used for analysis of the trip start and end events.

41

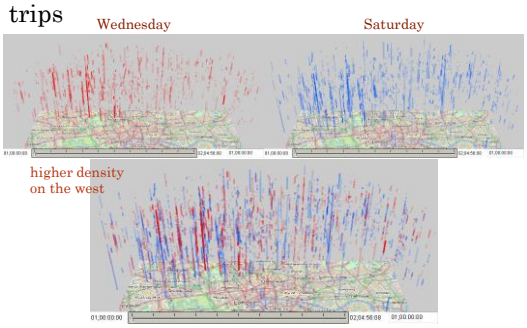
Relation of OD moves to spatial events

(continued)

- An OD move as a whole is a spatial event.
  - It has existence time = the time of the trip.
  - It has a position in space consisting of the start and end positions.
    - Such discontinuous spatial positions also occur for other spatial objects, e.g., a university.
    - The spatial position of a move can be represented by a directed line segment (vector), which is a spatial object.
- However, the visualisation and analysis methods designed for spatial events may not be applicable to this kind of events.
  - Most methods assume that the spatial positions of the events do not change during the time of event existence.
  - Moreover, most of them assume that the spatial positions of the events can be represented as points in space.
- Hence, OD moves require special approaches.

42

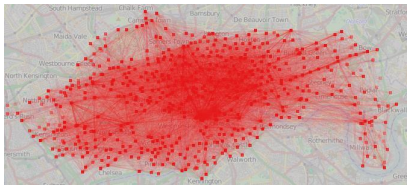
Spatio-temporal distributions of the round trips



Time transformation to the daily cycle

43

OD moves as spatio-temporal objects are hard to visualise



OD moves with differing origins and destinations can be represented by straight lines on a map or in a space-time cube connecting the positions of the starts and ends. However, such a display is very often extremely cluttered due to numerous line intersections.
Here: 39,737 bike trips from Wednesday are shown on a map with 99% transparency.

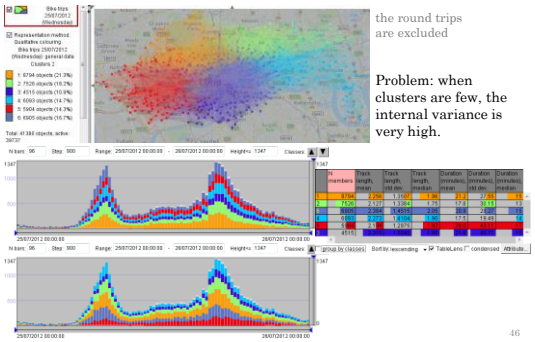
44

General approaches to dealing with large data and display clutter

- **Aggregation**: divide attribute domains into bins; obtain counts of objects and summaries of their thematic attributes for the bins; analyse the distributions of the aggregates.
  - In particular, aggregate spatio-temporal objects by areas and time steps.
- **Partition-based clustering**: divide objects into groups by similarity or closeness; analyse and compare group summaries (aggregates) and internal variations.
  - In particular, cluster spatio-temporal objects by their positions in space and time.
  - PBC is efficient when a relatively small number of clusters is sufficient.
- **Density-based clustering**: find dense groups (concentrations) of similar or close objects; analyse their relations to the remainder.
  - In particular, find spatio-temporal concentrations of spatio-temporal objects; analyse where and when they occurred.

45

PBC by spatial positions (starts + ends)



the round trips are excluded

Problem: when clusters are few, the internal variance is very high.

46

DBC: what distance function to use?

- DBC requires setting a distance threshold ⇒ there should be a meaningful notion of the distance between objects.
- The **spatial distance** between OD moves can be defined as the mean of the spatial distances between the origins and between the destinations:
  - $s\_distance(m_1,m_2) = (s\_distance(o_1,o_2) + s\_distance(d_1,d_2)) / 2$
- Analogously, the **temporal distance** may be defined as
  - $t\_distance(m_1,m_2) = (t\_distance(o_1,o_2) + t\_distance(d_1,d_2)) / 2$
- The **spatio-temporal distance** between OD moves can be defined taking the same approach as for instant spatial events:
  - Spatial distance threshold (radius) R_S + temporal distance threshold R_T
  - $s\_t\_distance = \max(s\_distance/R_S + t\_distance/R_T)*R_S$

47

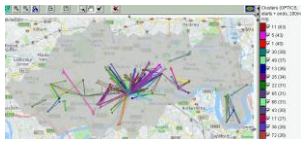
Distance functions and analysis tasks

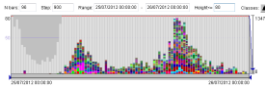
- DBC with different distance functions is used for answering different questions.
- **Spatial distance**: What moves (in terms of the origin and destination *regions*) were the most frequent throughout the whole time?
  - Further questions: What is the proportion of the frequent moves in the whole set? How are the frequent moves distributed over time? What are their thematic characteristics?
- **Spatio-temporal distance**: What similar trips occurred in close times?
  - A spatio-temporal cluster means collective movement of multiple objects between some origin and destination regions.
  - Further questions: How frequent are the occurrences of collective movement? Based on their positions in space and time, how can they be interpreted?

48

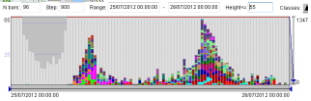
DBC by the spatial distance (examples)

25/07/2012 (Wednesday); round trips excluded.
Distance threshold R = 200 m; minimal number of neighbours N = 10.
Result: 93 spatial clusters with sizes from 10 to 63 include 1,618 trips (4%); the noise consists of 38,119 trips (96%).
Hence, not many very similar trips occurred throughout the day.





49

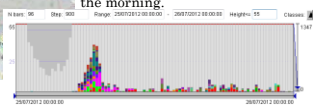


Further observations:

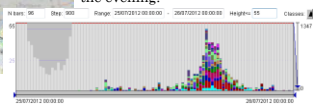
- Waterloo is the most popular area of trip origins and destinations: 940 out of 1,618 clustered trips (58%) originate from or end in this area.
- Some spatial clusters are related to particular time periods, i.e., either morning or evening.

50

Out-trips from Waterloo:
19 clusters containing 390 trips.
Most of the out trips occurred in the morning.



In-trips to Waterloo:
26 clusters containing 550 trips.
Most of the out trips occurred in the evening.



51

DBC by spatial distance

Testing other parameter settings

Distance threshold R = 250 m; minimal number of neighbours N = 10.
Result: 153 spatial clusters with sizes from 10 to 72 include 2,707 trips (6.8%); the noise consists of 37,030 trips (93.2%).
Relaxing the notion of similarity increased the proportion of clustered trips.

52

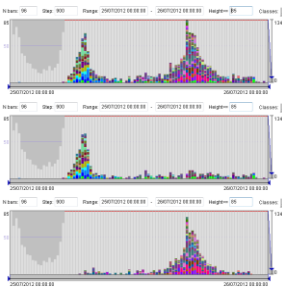
Clustered trips that were previously in the "noise":

Compare to the previously obtained clusters with R = 200 m:

Evidently, some of the trips from the former "noise" have been grouped together with previously clustered trips; i.e., the previously existing clusters have become bigger. Besides, a number of new clusters have appeared.

53

Consistency between the clustering results

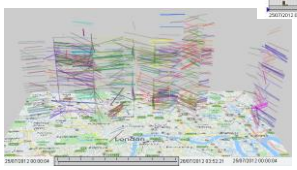


Popularity of the Waterloo area:
58 spatial clusters with 1,358 outgoing and incoming trips. The trip number increased by 44% from previous 940, but the proportion among the clustered trips decreased to 50% from the previous 58%.
Outgoing trips: 29 clusters with 627 trips; the majority occurred in the morning.
Incoming trips: 30 clusters with 742 trips; the majority occurred in the evening.
There were 11 trips that both started and ended in the Waterloo area.

54

The trip clusters outside of the Waterloo area:

94 clusters; 1,349 trips



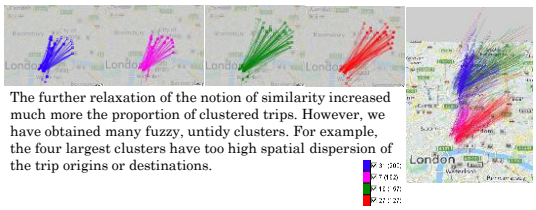
55

We can notice cluster pairs with opposite movements. In some of these pairs, one cluster mostly appears in the morning and the other in the evening.

However, there is no such vivid separation between the morning and evening movement directions as for the Waterloo area.

56

Distance threshold R = 300 m; minimal number of neighbours N = 10.
Result: 277 spatial clusters with sizes from 10 to 200 include 6,174 trips (15.5%); the noise consists of 33,563 trips (84.5%).



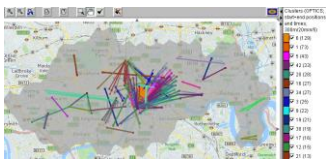
The further relaxation of the notion of similarity increased much more the proportion of clustered trips. However, we have obtained many fuzzy, untidy clusters. For example, the four largest clusters have too high spatial dispersion of the trip origins or destinations.

57

What similar trips occurred in close times?

Distance thresholds R_{space} = 300 m and R_{time} = 20 minutes; minimal number of neighbours N = 5.

Result: 75 spatio-temporal clusters with sizes from 5 to 129 include 916 trips (2.3%).



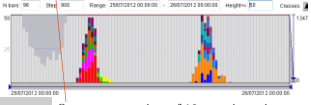
Groups of similar trips (collective or mass movements) mostly occurred in the morning and in the evening.

58

Groups with ≥ 10 members

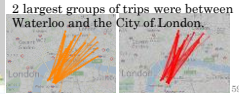
18 groups with 536 members in total.
All occurred in the morning or in the evening.
All but 3 groups have origins or destinations in the Waterloo area.

Two groups with 13 and 10 trips end near King's Cross in the evening.



One group consists of 10 morning trips from East India to Canary Wharf.

2 largest groups of trips were between Waterloo and the City of London.



59

Analysis of OD moves using spatial clustering

Iterative clustering + visualisation+ interactive operations

- In this example, we applied the principles of visual analytics:
  - Iterative application of computational methods (here clustering) with different parameter settings.
  - Visual investigation of the results, comparison between them.
  - Division of data into subsets and comparison between the subsets.
- In the investigation, we used following interactive operations:
  - Display linking by highlighting
  - Filtering by attribute values (e.g., displacement distance)
  - Propagation of cluster colours
  - Filtering by cluster selection
  - Spatial filtering by visited areas
  - Temporal filtering
- As a result, we gained useful information from quite a large dataset despite of initially extremely cluttered views.

60

What we have learned so far

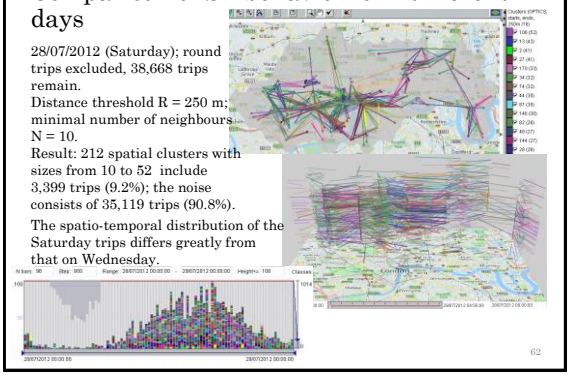
- High variety of trips; not many similar trips.
- A large number of similar trips from Waterloo in the morning and to Waterloo in the evening.
  - These may be trips of commuters who travel to London by railway and then get to the final destinations by bike.
- Existence of opposite groups of trips.
  - Some of these may be trips of people going to their work or study places in the morning and back in the evening.
- Existence of groups of similar trips occurring close in time.
  - Most such groups occur in the morning or in the evening and go from or to areas around railway stations.

61

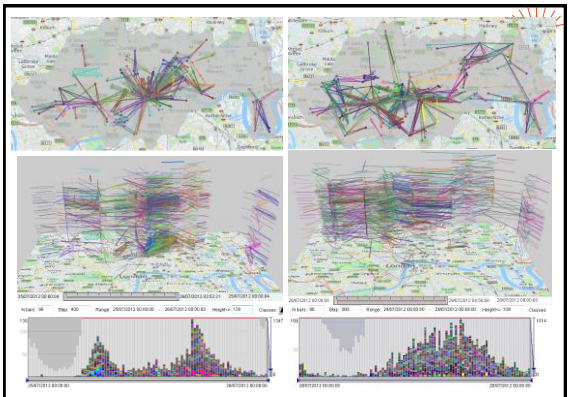
Comparison of ST behaviours in different days

28/07/2012 (Saturday); round trips excluded, 38,668 trips remain.
Distance threshold R = 250 m; minimal number of neighbours N = 10.
Result: 212 spatial clusters with sizes from 10 to 52 include 3,399 trips (9.2%); the noise consists of 35,119 trips (90.8%).

The spatio-temporal distribution of the Saturday trips differs greatly from that on Wednesday.



62



63

Questions?

Analysis of OD movement data by means of density-based clustering

64

Analysis of OD movement data by means of spatio-temporal aggregation

65

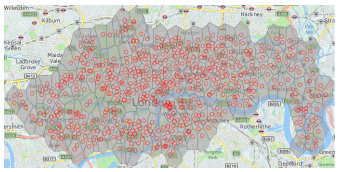
Spatio-temporal aggregation of spatio-temporal objects (events)

- The space (territory) containing the events is divided into compartments.
- The time span of the data is divided into intervals, usually of equal length.
- For each pair (compartment × time interval), the relevant events are summarized into event counts + summaries of thematic attributes
  - Relevant **OD moves** for a pair (compartment × time interval): moves that start in this compartment during this time interval + moves that end in this compartment during this time interval
    - The starts and ends are counted separately.
- Aggregation result: spatial time series A(S,T)
  - S: the set of space compartments
  - T: the set of time intervals
  - A: aggregate attributes (counts + statistical summaries of thematic attributes)

66

Space (territory) division possibilities

- Regular grid (rectangular or hexagonal) with a chosen cell size
- Meaningful pre-existing division (e.g., administrative units)
- Irregular grid respecting the spatial distribution of the objects



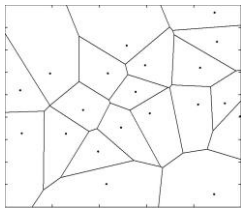
Example: an irregular grid for the London cycle hire data has been built on the basis of the spatial distribution of the docking stations, i.e., the origin and destination places of the trips.

67

Voronoi tessellation (a.k.a. Voronoi diagram)

Used for building irregular grids

- The partitioning of a plane with N points into convex polygons *(cells),* such that
  - each polygon contains exactly one generating point
  - every point in a given polygon is closer to its generating point than to any other.
- The generating points are also called *seeds*.
- A Voronoi diagram is also known as a Dirichlet tessellation.
- The cells are called Dirichlet regions, Thiessen polytopes, or Voronoi polygons.



68

How do we obtain a tessellation?

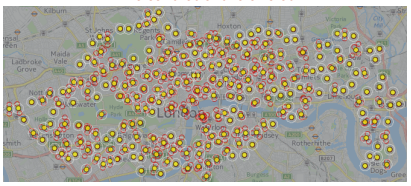
Step 1: we apply a special algorithm for clustering of points based on their spatial proximity.

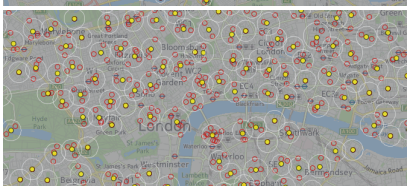
The main idea of the algorithm: put the points into circles with a given maximal radius R.

We chose R = 400 m.

69

The centres of the circles

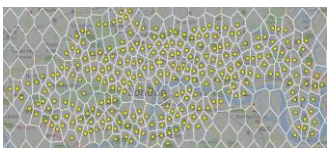




70

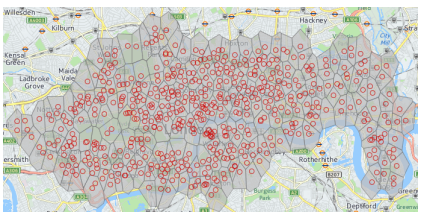
How do we obtain a tessellation?

Step 2: we use the centres of the point clusters (i.e., the centres of the circumcircles of the clusters) as the generating seeds for the Voronoi tessellation.

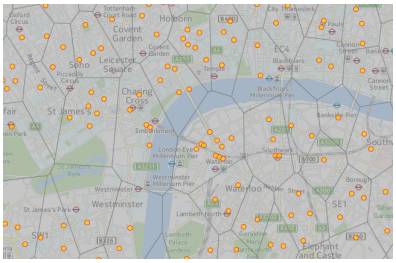


71

The empty cells (not containing any stations) can be removed.

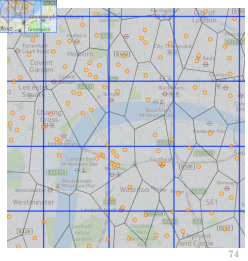


72



The resulting irregular grid is built so that near stations tend to be included in the same cell. This allows trips with close origins or destinations to be counted together, e.g., trips from or to the area around Waterloo.

73



Compared to a regular grid, an irregular grid can give us a better idea about the spatial distribution of objects (in our example, trip origins and destinations).

74

Time division possibilities

- Treat the time as a directed line.
  - Divide the range from the earliest to the latest time value in the data into consecutive intervals.
  - To aggregate: for each time value in the data, find the containing interval.
- Treat the time as a cycle.
  - Choose a relevant cycle for the data: daily, weekly, annual, domain-specific, e.g., production cycle
  - Divide the chosen cycle into intervals.
  - To aggregate: for each time value in the data, determine its position in the cycle and find the containing interval of this position.
    - I.e., the absolute time value is transformed to relative w.r.t. the time cycle.

75

Aggregation result: spatial time series

- Data with two referrers: a set of space compartments and a set of time intervals
- Can be viewed in two complementary ways:
  - as a set of time series of attribute values associated with the spatial compartments
  - as a set of *spatial situations* associated with the time steps
    - A *spatial situation* is a distribution of attribute values over the space in some time step

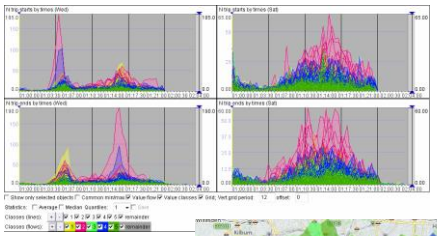
... ...

76

Analysis of spatial time series *(a reminder)*

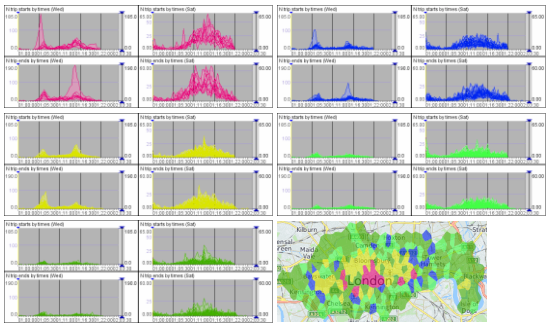
- Analysis tasks address two aspects of the overall behaviour:
  - Spatial distribution of the local temporal variations of the attribute values in different compartments
  - Temporal variation of the overall spatial distribution of the attribute values in different time moments
- Supporting visual analytics techniques *(considered in the previous lecture)* include two-way partition-based clustering
  - Grouping of places (compartments) by similarity of the local time series
    - followed by visual exploration of the distribution of the group members over the space
  - Grouping of time steps by similarity of the spatial situations
    - followed by visual exploration of the distribution of the group members over time

77

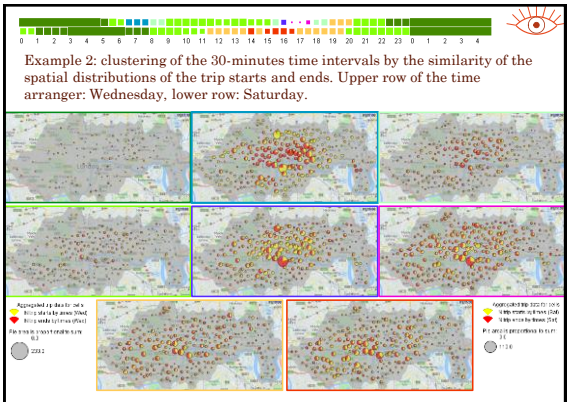


Example 1: the spatial compartments (grid cells) have been clustered by the similarity of the time series of the counts on the trip starts and ends on Wednesday and Saturday.

78



79



Example 2: clustering of the 30-minutes time intervals by the similarity of the spatial distributions of the trip starts and ends. Upper row of the time arranger: Wednesday, lower row: Saturday.

80

Questions?

Analysis of aggregated OD movement data referring to places (space compartments)

81

Trip aggregation by OD pairs

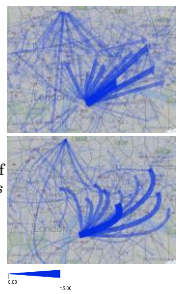
- The ST aggregation considered before loses information about the connections between the trip origins and destinations.
- To preserve this information, we need to aggregate data differently: by pairs of compartments and time intervals.
  - From the set of the spatial compartments S, we create the set of all possible pairs of compartments, i.e., the Cartesian product S × S
  - As previously, we divide the time (line or cycle) into a set of intervals T*
  - To aggregate the trip data (OD moves):
    - For each pair (s_o, s_d) and each time interval t, find and summarize all trips that started in s_o and ended in s_d within the interval t*.

* It would also be appropriate to create the set of all possible pairs of time intervals (t_i, t_j), where $t_i \le t_j$, and aggregate the data by the combinations (s_o, s_d, t_i, t_j); however, the result would be too complex for the analysis. For simplification, trips can be assigned to time intervals based on the times of their ends or starts.

82

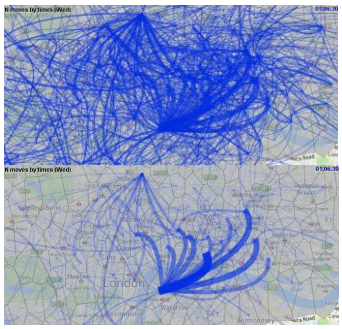
Result of trip aggregation by OD pairs

- Aggregated trip data have 3 referrers *(it would be 4 without the temporal simplification* : S × S × T
  - Attributes A(s_o, s_d, t): trip counts + statistical summaries of the trip characteristics (e.g., durations)
- It is very hard to visualise and analyse data with 3 referrers
- Approach: treat pairs (s_o, s_d) as a special type of geographic objects, called *spatial links*, or *flows*
  - Treat the aggregate attributes as usual spatial time series referring to these spatial objects
- Visual representation of flows on a map: flow symbols; widths and/or colours may represent values of thematic attributes.



83

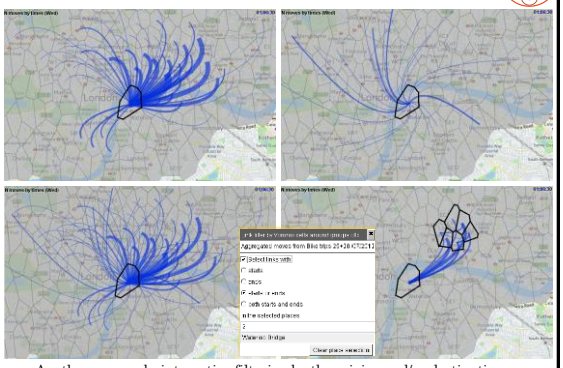
Problems of flow visualisation



Representing flows on a map typically results in extremely cluttered images due to numerous intersections of the flow symbols.

The problem may be reduced by varying the transparency of the flow symbols depending on the attribute values that are represented.

84



Another approach: interactive filtering by the origins and/or destinations.

85

Flows with coinciding origin and destination places may be represented by rings with the thickness proportional to values of a numeric attribute, such as the count of the moves.

For decreasing the display clutter, flows with small attribute values can be hidden by means of interactive focusing.

86

Spatial link-referenced time series

- We transformed the complex data structure $S \times S \times T \rightarrow A$ into a simpler data structure $L \times T \rightarrow A$
  - L is the set of *spatial links* defined as pairs (s_o, s_d), $s_o \in S$, $s_d \in S$.
- Hence, we obtained "normal" space-referenced time series, in which attribute values refer to spatial objects (specifically, the spatial links) and time steps.

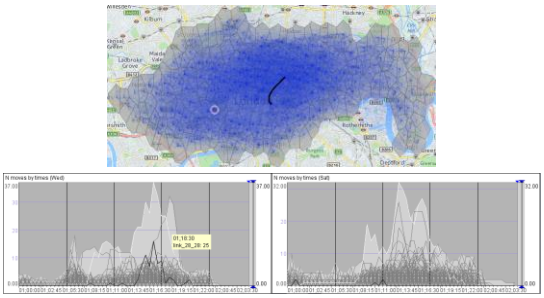
⇒ We can apply the same visual analytics techniques as for usual time series

- … but we need to deal with the difficulties in the visual representation of the spatial links in a map
  - ☹ This is a visualisation problem that is not completely solved yet.

87

Aspect 1

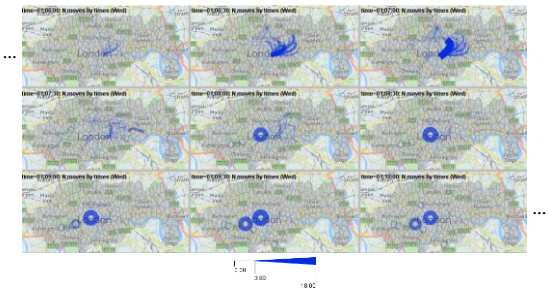
Set of time series associated with the links



88

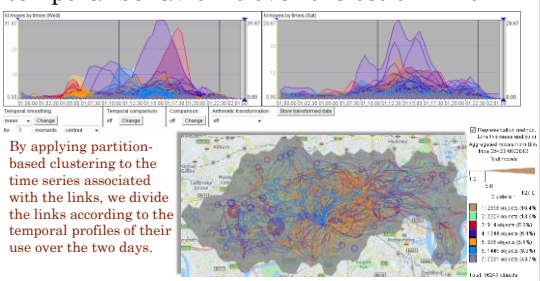
Aspect 2

Set of spatial situations associated with the time steps



89

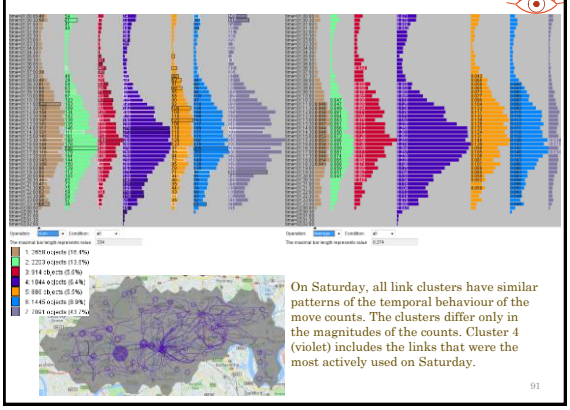
Investigating the distribution of the temporal behaviours over the set of links



By applying partition-based clustering to the time series associated with the links, we divide the links according to the temporal profiles of their use over the two days.

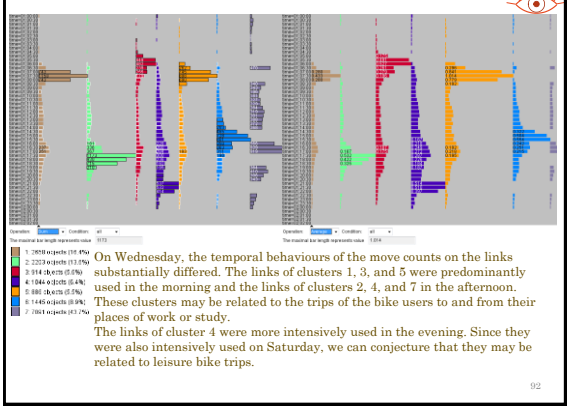
The map shows only the links that were used at least 5 times during the two days. The widths of the flow symbols are proportional to the total counts of the moves.

90



On Saturday, all link clusters have similar patterns of the temporal behaviour of the move counts. The clusters differ only in the magnitudes of the counts. Cluster 4 (violet) includes the links that were the most actively used on Saturday.

91



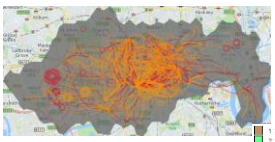
On Wednesday, the temporal behaviours of the move counts on the links substantially differed. The links of clusters 1, 3, and 5 were predominantly used in the morning and the links of clusters 2, 4, and 7 in the afternoon. These clusters may be related to the trips of the bike users to and from their places of work or study.

The links of cluster 4 were more intensively used in the evening. Since they were also intensively used on Saturday, we can conjecture that they may be related to leisure bike trips.

92

Wednesday morning link clusters

Leisure time link clusters



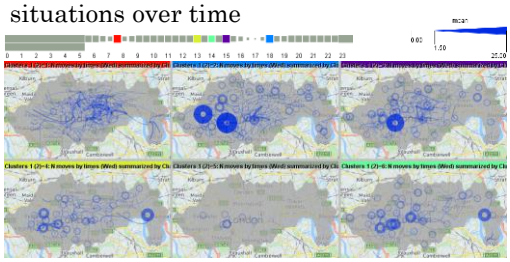
Wednesday afternoon link clusters

The interpretation of the clusters is not completely "clean". Thus, it appears that clusters 3 and 5 may include "leisure" links, which might also be used by some people on Wednesday morning.

Even when clustering does not give us clear separation, it still reveals the major patterns and facilitates our understanding.

93

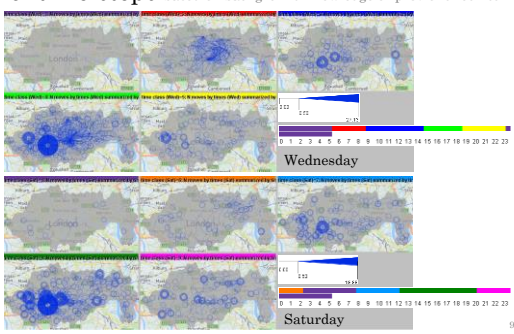
Investigating the variation of the spatial situations over time



Partition-based clustering has been applied to the combinations of the move count values for the links corresponding to the different time intervals of the Wednesday. Only the links with the total move counts ≥ 3 were used (4,897 of 16,241, or 30.2%). The clustering does not produce meaningful groups but just isolates certain intervals.

94

Interactive manual partitioning of the set of time steps based on background knowledge or previous results



95

Questions?

Aggregation of OD movement data by OD pairs and analysis of flows

96

Analysis of spatial events and OD moves

Complementarity of DBC and spatio-temporal aggregation

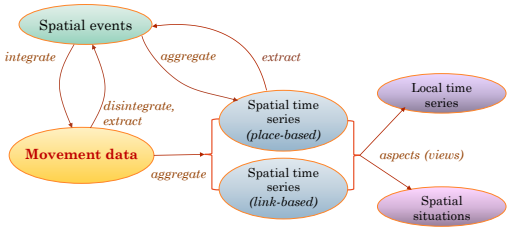
- Task: analyse the spatio-temporal distribution of events/moves
  - Space and time are treated as containers of the events/moves
- Spatio-temporal aggregation:
  - Creates a data structure (spatial time series) that appropriately represents the distribution:
    - Space and time become referrers; attributes express containment of objects by spatio-temporal bins
  - However, it may conceal important features in the ST distribution
    - particularly, spatio-temporal concentrations of events and spatio-temporal coincidences of moves with close origins and/or destinations
    - Possible reasons:
      - Large bins: features become concealed by averaging
      - Small bins: features become dispersed over multiple bins

⇒ST aggregation need to be complemented by ST density-based clustering, which reveals concentrations and coincidences.

97

Transformations of spatio-temporal data

a reminder



98

Purposes of data transformations

a reminder

- **Aggregation**
  - Supports abstraction, gaining an overall view of characteristics and behaviour
  - Reduces large data
  - Simplifies complex data
- **Extraction** *of events, etc.*
  - Selects a portion of data relevant to a task, enables focusing
  - Allows dealing with complex data portion-wise
- **Integration, disintegration, projection** (taking one of possible aspects)
  - Adapts data to analysis tasks

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.

100

Questions?

Analysis of spatial events and OD moves

101

Lecture summary

- Density-based clustering: distance, neighbourhood, core object, cluster, noise
- Difference between density-based and partition-based clustering
- Creation of distance measures (distance functions) to account for spatial positions, times, and thematic attributes
- Density-based clustering of spatial events
- Origin-destination movement data
  - Structure and properties
  - Density-based clustering with different distance functions
  - Spatio-temporal aggregation, complexity of link-based aggregates

102