INM433: Session 02

Using Partition-Based Clustering in Visual Analytics

INM433 Visual Analytics

Last week

- Visual Analytics
- Role of interactive visualisation in Visual Analytics
  - Visual variables and when to use
  - Types of visualisation display and when to use
  - Types of interaction
  - Coordinated linked views
- (Practical) how to use Mondrian and Tableau

This week

- Data types and structure
  - And how these affect analysis and interpretation
- How **partition-based clustering** combined with interactive visualisation can help deal with large complex datasets
  - Density-based is the other type of clustering that will be covered next week
- Practical: Using R with Mondrian and Tableau

Data structure

Semantic role of data components

- **Reference**: What is described?
- **Characteristic**: What is known about it?

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

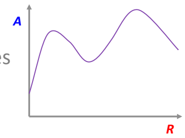
There may be multiple referrers

- 2 referrers
  - year, state, (id is stateId)
- Many characteristics (attributes)

year	id	State	Population	Index offenses	Violent crime	Murder	Forcible rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft
1960	1	Alabama	3266740	39920	6097	406	281	898	4512	33823	11626	19344	2853
1960	2	Alaska	226167	3730	236	23	47	64	102	3494	751	2195	548
1960	4	Arizona	1302161	39243	2704	78	209	706	1711	36539	8926	23207	4406
1960	5	Arkansas	1786272	18472	1924	152	159	443	1170	16548	5399	10250	899
1960	6	California	15717204	546069	37558	616	2859	15287	18796	508511	143102	311956	53453
1960	8	Colorado	1753947	38103	2408	73	229	1362	744	35695	9996	21949	3750
1960	9	Connecticut	2535234	29321	928	41	103	236	548	28393	8452	16653	3288
1960	10	Delaware	446292	9642	375	33	41	157	144	9267	2661	5867	739
1960	11	District of Co	763956	20725	4230	81	111	1072	2966	16495	4587	9905	2003
1960	12	Florida	4951560	133919	11061	527	403	4005	6126	122858	39966	73603	9289

Considering data as variables & functions

- Data components
  - Referrers: independent variables
  - Attributes: dependent variables
- Consider data as a function
  - *f(indepVar)=depVar*
  - E.g. crime rates vary over space and time



Study the behaviour of the function

- The general aim of analysis is to study the behaviour of the function:
  - **Describe**: how attributes vary
  - **Locate**: referrers and/or subsets for which particular bahaviours or attribute value apply
  - **Compare:** behaviours between different **attributes** or different **subsets**
  - **Relate:** find similar behaviours
- Analysis uses specific versions of these generic tasks

Behaviour: describe

- Describe the behaviour
  - What is the distribution of values
- Examples
  - Has crime been increasing over the past decade?
  - How normal is the distribution tweets per twitter user?

Behaviour: locate

- Locate the behaviour
  - Which referrers exhibit a particular behaviour?
  - Identify
- Examples
  - Which places have both high unemployment and a high proportion of under 30s?
  - Which days of the week have low burglary?
  - Which students are above average height?

Behaviour: compare

- Compare two or more behaviours (find similarities and differences)
  - Different attributes over the same set of referrers
  - Same attributes over different subsets of referrers
- Examples
  - Does spatial distribution pubs compare to that of craft beer bars?
  - How do the salaries of men and women compare?

Behaviour: relate

- Relate behaviours of two or more attributes
  - Is there a correlation between two or more attributes?
- Examples
  - Is there a relationship between density of CCTV cameras and amount of recorded crime?
  - Does Tweet frequency relate to hour of day?

Data formats and structure

Semantics independent of structure

- Data can be structured in different ways
  - Trees: XML, JSON
  - Fields: e.g. sea surface temperature
  - Networks: e.g. social networks
  - Geometry: geographical areas
  - Tabular: CSV, Excel, tab-delimited, ASCII
- Data can be represented in different ways
- But may have the same semantics

Tabular data is common

- Many software rely on table-structured data
  - Often use relational database theory to represent **1:n, n:1** or ***n:n*** relations.
  - Tableau uses a table-based format for geometry
    - http://kb.tableau.com/articles/knowledgebase/polygon-shaded-maps
  - Mondrian used a text-based representation:
    - http://www.theusrus.de/Mondrian/Mondrian.html#ASCII

Transforming tables

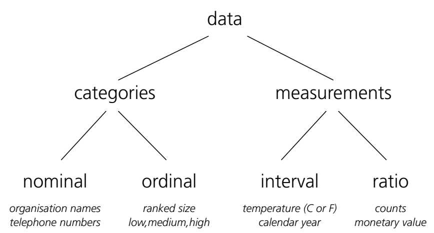
- Referrers and attributes are are not always in columns
  - Some software needs you to transform the data
  - In **r**, the `melt()` and `dcast()` functions in the `reshape2` package will do this – see practical.

```
#   ozone solar.r wind temp month day
# 1    41     190  7.4   67     5   1
# 2    36     118  8.0   72     5   2
# 3    12     149 12.6   74     5   3
# 4    18     313 11.5   62     5   4
# 5    NA      NA 14.3   56     5   5
# 6    28      NA 14.9   66     5   6
```

```
#   variable value
# 1    ozone    41
# 2    ozone    36
# 3    ozone    12
# 4    ozone    18
# 5    ozone    NA
# 6    ozone    28
```

Data types

Measurement level



Measurement value

- Affects the domain* (set of possible values)
  - Finite/infinate
  - Discrete/continuous
  - Ordered/not ordered
  - Has distance/no distance

Klir, G.J. (1985). *Architecture of Systems problem Solving*. Plenum, New York.

Types of referrers

- Object (sometimes referred to as "population")
  - No ordering, no distances, discrete
    - Temporal objects: 1D ordered, maybe 2D/3D ordered
    - Spatial objects: 2D ordered
- Time
  - 1D ordering, has distance, continuous
- Space (2D, 3D)
  - 2D ordering, has distances, continuous

Types of dataset (by referrer)

- 1 referrer
  - Object-referenced
  - Time-referenced (time-series)
  - Space-referenced (spatial data)
- May have multiple attributes
  - Multivariate/multi-demensional/high dimensional
  - Multi-dimensional time-series
  - Multi-dimensional spatial data

Types of dataset (by referrer)

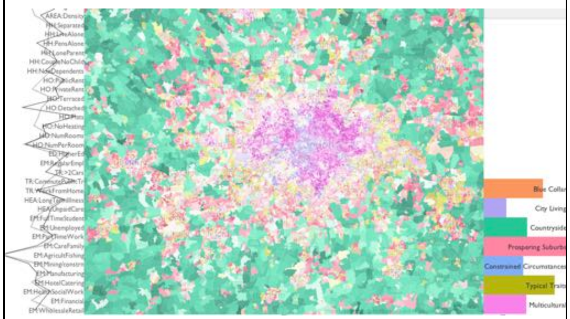
- 2 referrers
  - Object-referenced time-series
  - Spatial time-series
- May have multiple attributes
  - multidimensional spatial time series

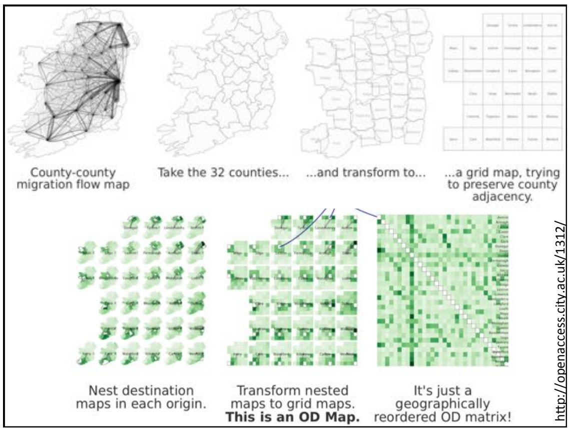
Types of Object (1)

- Generic
- Spatial objects
  - locations in space, sometimes with areal extent
- Temporal objects (events)
  - **temporal categories**: month, year, hour aggregate
  - **instant events**: e.g. tweet postings, bank transactions
  - **events with duration**: e.g. holidays, electoral campaigns, classes, breaks, TV shows

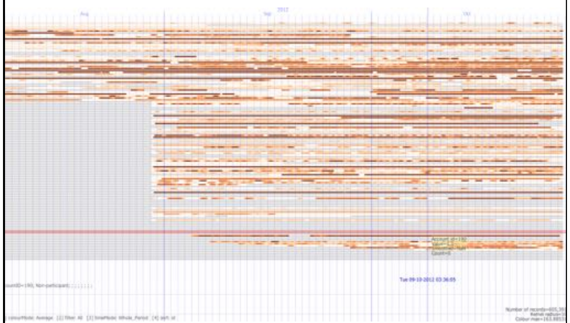
Types of (Spatiotemporal) Object (2)

- Spatial events
  - events with location: e.g. lightning strikes, geolocated tweet postings, earthquakes, traffic jams
- Moving objects
  - object which change their location over time
  - time-series of spatial locations (trajectories)
  - E.g. people, animals, vehicles, storms, oil spills,
- Trajectories
  - May have other attributes: shape, travelled distance, mean speed





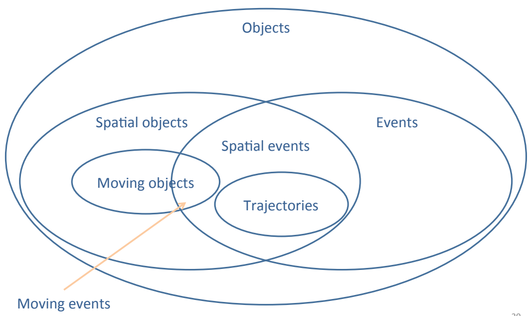
http://openaccess.city.ac.uk/1312/



http://openaccess.city.ac.uk/2618/

http://openaccess.city.ac.uk/2528/

Types of object



30

Data types

- Data types based on referrers and attributes
  - Object, space, time
  - Their combination
- Objects
  - Generic, space, time, spatio-temporal

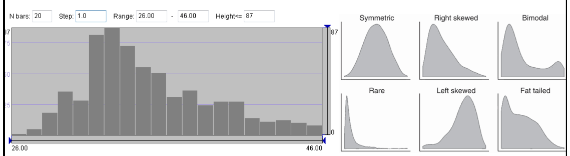
Analysing multidimensional object-referenced data

Aim

- Study the distribution of the attribute values over the set of objects
  - describe
  - locate
  - compare
  - relate

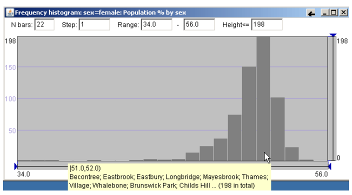
Task: Describe

- **Task:** Describe the value distribution of a single numeric attribute
  - Use a frequency histogram to look at the shape of the distribution and any outliers



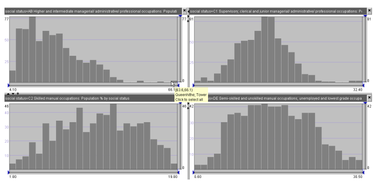
Task: Locate

- **Task:** Locate referrers with attributes of various values
  - Use interaction on a frequency histogram



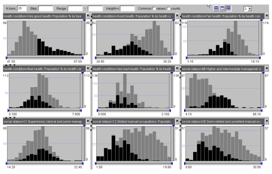
Task: Compare

- **Task:** Compare value distributions of several attributes
  - Juxtapose multiple distributions in histograms
    - make sure they scaled appropriately!



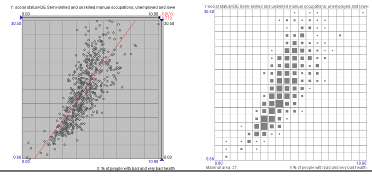
Task: Relate

- **Task:** Relate value distributions of several attributes
  - Relate a subset frequency to the whole dataset, juxtaposing those for different attributes



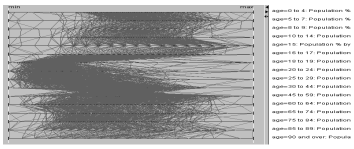
Task: Relate

- **Task:** Relate value distributions of two attributes
  - Use a scatterplot (left) for pairwise comparison looking for apparent correlations, clusters and outliers
  - Use a binned scatterplot (right) if lots of data



Task: Describe

- **Task:** Describe the joint value distribution of multiple attributes
  - Tricky where there are lots of variables
  - Try using summary statistics (deciles, interquartiles)
  - Try using transparency and/or binning



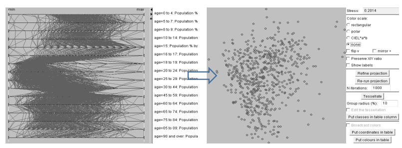
Simplify the data

- Where there two many data points, simplify.
- Two approaches:
- 1. Reduce the number of **attributes**
  - Dimension-reduction: create two synthetic variables that try to represent the variation
- 2. Reduce the number of objects (referrers)
  - **Group** the objects into a (much) smaller set of *representative* groups
    - low within variation and high between variation
  - **Summarise** the attributes by group

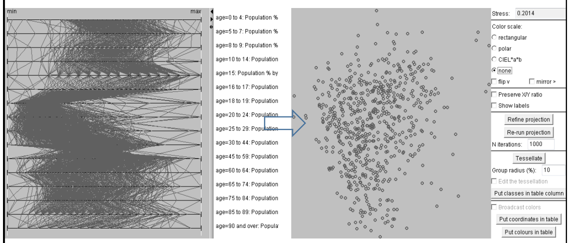
Approach 1: Dimension-reduction

Approach 1: Dimension-reduction

- (Reducing the number of attributes)
- There are many approaches to reduce many dimensions to two (so they can be plotted)
  - multidimensional scaling (MDS), principal component analysis (PCA), Sammon's mapping



Approach 1: Dimension-reduction

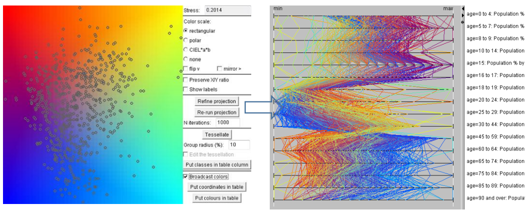


Approach 1: Dimension-reduction

- Is it useful?
  - Less clutter
  - Shows objects that (likely) have **similar** characteristics (close to each other)
    - similar = similar combinations of attribute values
  - Shows (likely) **outlier** objects
  - Shows **groups** of objects that are (likely) similar
- But…
  - These two dimensions are not interpretable
  - Need to link them back to the original data

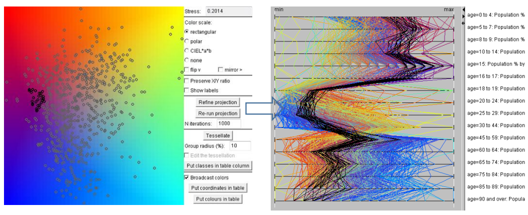
Approach 1: Dimension-reduction

- We can use hue!
  - "Sammon's mapping" provides a 2D-varying hue
  - We can relate object in two projections with hue



Approach 1: Dimension-reduction

- We have use interactive brushing to visually link them



Approach 1: Dimension-reduction

- Problems
  - Inherent distortions – cannot whether close=similar
  - The projection itself only gives n approximate understanding of the data distribution.
  - Try different methods of simplification
    - consistent results can be trusted
    - discrepancies require detailed investigation

Approach 2: Partition-based clustering

Approach 2: Partition-based clustering

- (Reducing the number of objects)
- Two major types of clustering
  - **Partition-based clustering**: all object allocated
  - **Density-based clustering**: not all object allocated
- We'll use **partition-based clustering** to group our objects with similar attribute "signatures". Aim:
  - objects **within** a group should be similar.
  - objects should be dissimilar **between** groups.
- How's it work?
  - http://shabal.in/visuals/kmeans/2.html

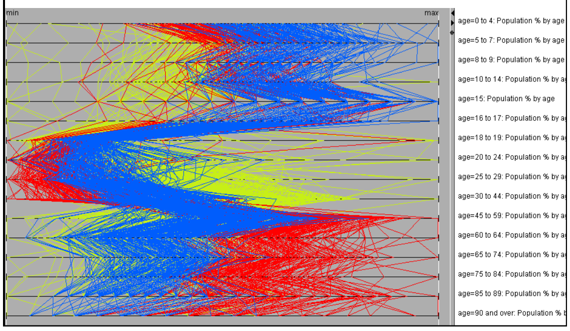
Partition-based clustering

- Most methods ask for:
  - the desired number of clusters
  - the features with which to cluster (and weighting)
  - sometime other parameters
- It's hard to choose good ones, so try a few:
  - you need to discriminate objects based on characteristics that are of interest to your analysis
  - don't forget the purpose is to help your analysis!
  - use interactive graphics to compare

Partition-based clustering

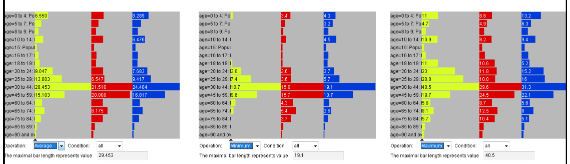
- Output
  - A clusterID assigned to each object
- Interpret
  - How do attributes vary within & between clusters
- Note that:
  - There's often a stochastic element, so slightly different solutions each time
  - Hue is a good way to show cluster (why?)
  - Often hard to relate alternative cluster solutions – often not a 1:1 mapping.

A 3-cluster solution



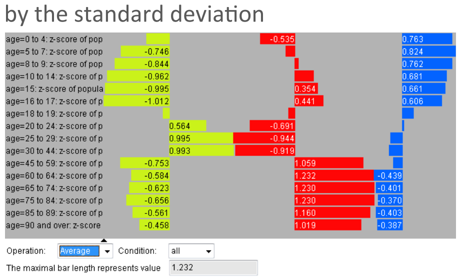
Attribute values by cluster

- Mean (yellow), min (red), max (blue)
- But large magnitude differences make differences between clusters hard to see

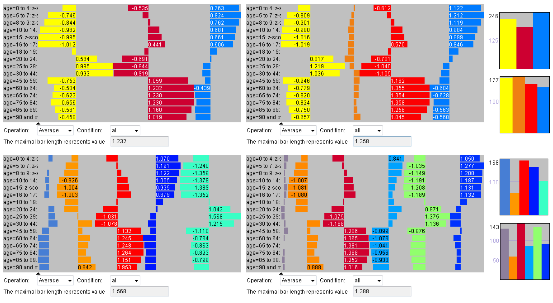


Transformation to z-scores

- z-score is the deviation from the mean divided by the standard deviation



Impact of number of clusters (k)



No “right” and “wrong” groupings!

- Choose a solution that supports your analysis
- All groupings are a (huge) simplification
  - and information loss
- Try different solutions and explore these with interactive graphics

Analyse space-referenced data

Aim

- Study the **spatial** distribution of the attribute values
  - describe
  - locate
  - compare
  - relate

“Spatial is special”

- Concepts usually that are usually important
  - location
  - distance between items
  - neighbourhoods
  - spatial distribution

Tobler's first law of geography

- “Everything is related to everything else, but near things are more related than distant things”
  - Spatial dependence
  - neighbouring objects or locations expected to have similar attribute values
  - outliers are values that deviate from that
  - But these may not hold once we spatially aggregate data (e.g., by district)

Tobler, W. (1970). "A computer movie simulating urban growth in the Detroit region". **Economic Geography**, 46(2): 234-240.

Direction

- Space has inherent 2D ordering
  - spatial objects can be arranged using 2D position
- However, we may choose to analyse spatial data using **distance** or **direction**
  - Frees up a visual variable

Geographical space

- Geographical space contains physical features
  - rivers, motorways, coastlines, land use
  - this interferes with the First Law of Geography
- So, taking geographical context into account is often important
  - distance to coastline
  - Altitude
  - sources of noise/pollution

Map

- Both dimensions of **position** (visual variable)
- Can show
  - spatial distribution of spatial objects
  - spatial distribution of space-referenced attributes
  - distance/neighbourhood relations
  - geographical context (same coordinate system)

Maps: some problems

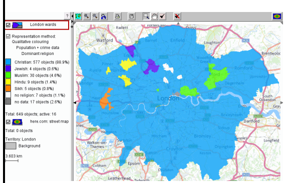
- Geographical distortions
  - use an appropriate cartographic projection
  - **all** of these introduce some kinds of scale and/or angular distortion
    - Extremely small for area of small geographical extent
    - Particularly high for areas with large latitudinal extent at high latitudes

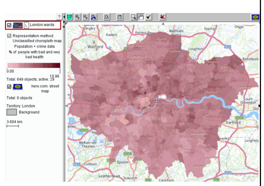
Maps: some problems

- Cartographic space
  - We're often more interested in area where there's a high density of objects
  - This doesn't give us much cartographic space
  - Lots of possible solutions
    - Cartograms (see last week) and other alternative projections
    - Interactions that limit the amount of data shown at any one time
    - Don't use maps

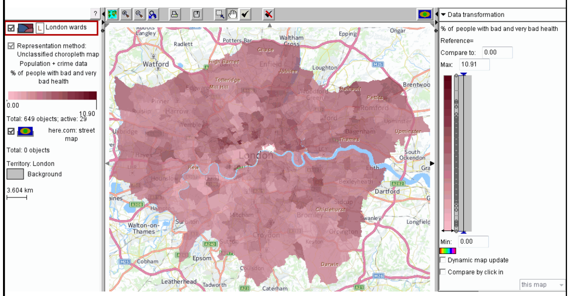
Choropleth maps

- Good for maps of spatial distributions of attribute values

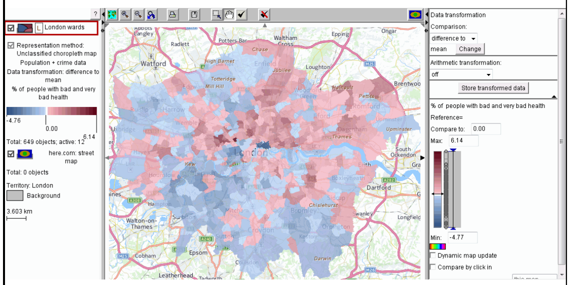




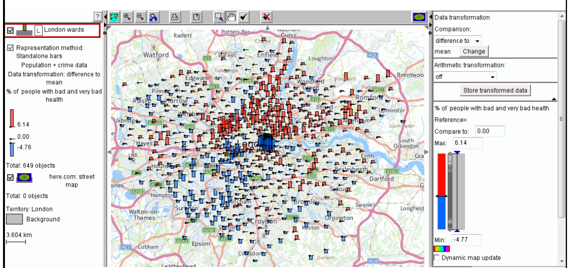
Choropleth maps: sequential



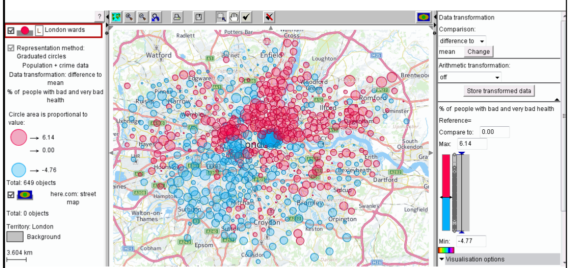
Choropleth maps: sequential



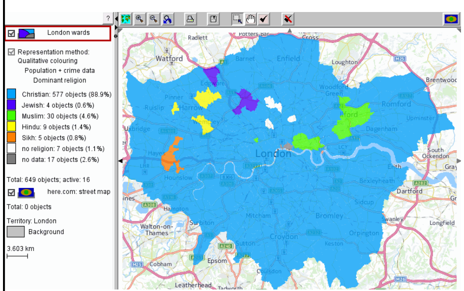
Proportional symbol map



Proportional symbol map: diverging



Choropleth map: qualitative



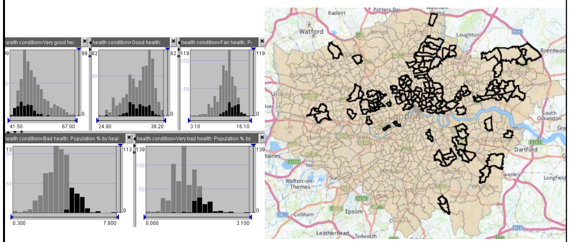
Space-referenced vs object-referenced

- Object-referenced techniques are suitable for space-referenced data
  - They just don't take space into account
- But they can be combined with spatial views using interactive linked views

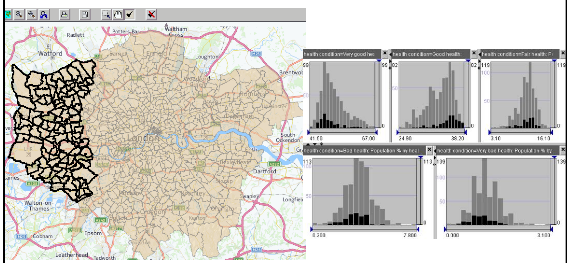
Linking spatial to attribute

- Select objects on a plot, diagram, or histogram
  - see the spatial distribution of the selected objects on a map
  - are there any spatial patterns?
- Select spatial objects on a map
  - See those objects in the attribute displays
  - which values and value combinations occur in the selected part of space?

Attribute displays to map

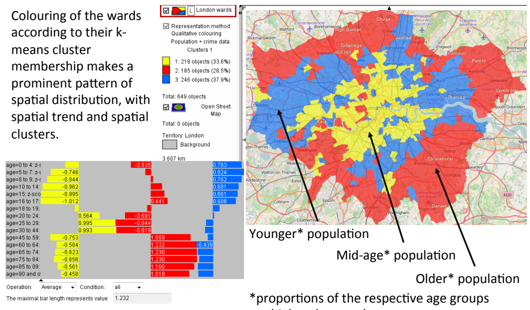


Map to attribute displays



Partition-based clustering of space-referenced data

Colouring of the wards according to their k-means cluster membership makes a prominent pattern of spatial distribution, with spatial trend and spatial clusters.



Younger* population

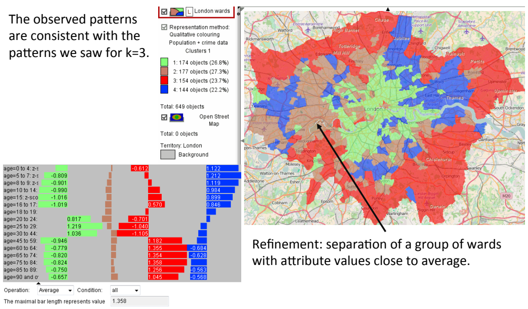
Mid-age* population

Older* population

*proportions of the respective age groups are higher than on the average

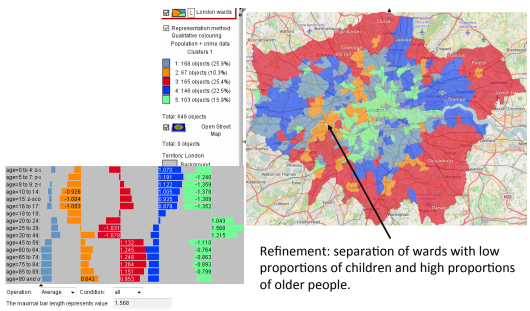
Running clustering with different settings

The observed patterns are consistent with the patterns we saw for k=3.



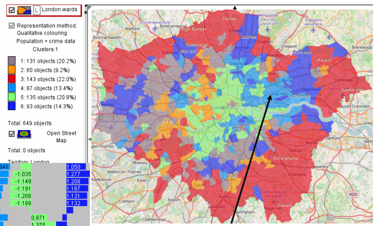
Refinement: separation of a group of wards with attribute values close to average.

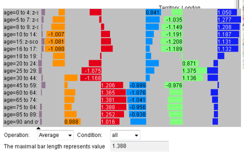
Running clustering with different settings



Refinement: separation of wards with low proportions of children and high proportions of older people.

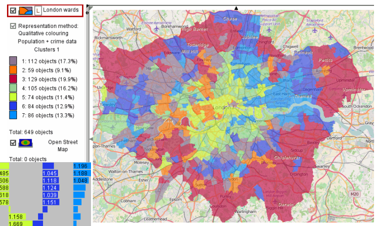
Running clustering with different settings

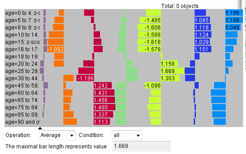




Further refinement: separation of wards with low proportions of older people.
Still, the observed spatial pattern remains consistent with the previous results.

When to stop further refinement?





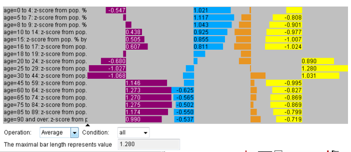
When it reveals only subtle differences between groups and does not bring significant new information.
We use clustering to simplify data; too many clusters mean high complexity.

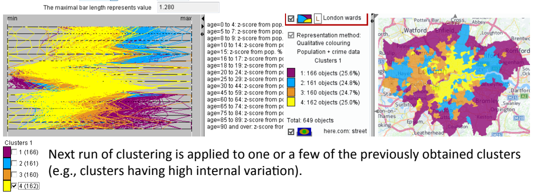
Use of partition-based clustering in visual data analysis

How to choose suitable parameter settings?

- Run clustering with different settings and investigate how the results change
- Select the settings bringing the "best" results:
  - makes sense? (e.g., understandable spatial patterns)
  - internal variance within the clusters is sufficiently low
  - fit to the purpose (e.g., the intended analysis scale may require coarser or finer division)
- Use **progressive clustering** for targeted refinement of clusters with high internal variance.

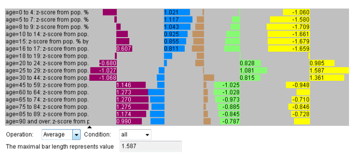
Interactive progressive clustering

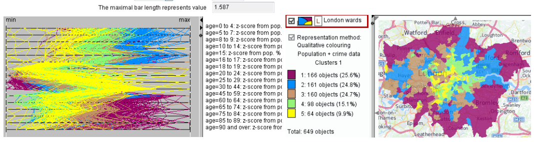




Next run of clustering is applied to one or a few of the previously obtained clusters (e.g., clusters having high internal variation).

Interactive progressive clustering



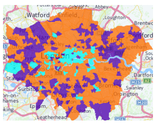


Former cluster 4 has been subdivided into clusters 4 and 5, which differ in the proportions of children. The internal variation was thereby decreased.

Can apply more generally

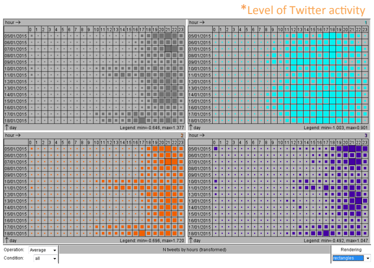
- Applied to combinations of values of multiple attributes associated with any objects
- Partition-based clustering is also applicable to multiple time series
  - Object-referenced time series
  - Space-referenced time series

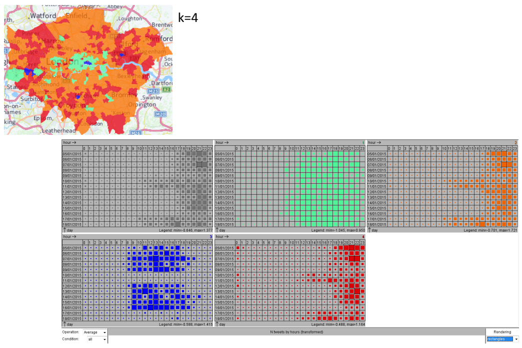
An example of applying clustering to space-referenced time series

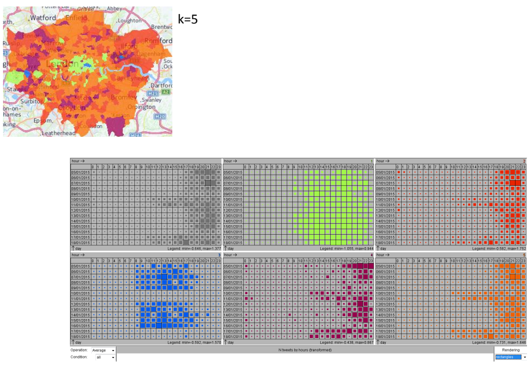


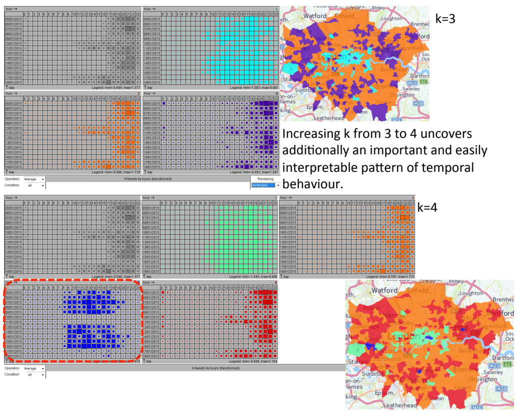
Attribute* values refer to hourly time intervals over a period of two weeks.

*Level of Twitter activity

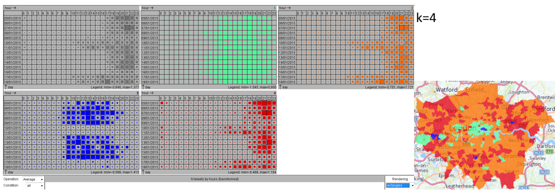








Increasing k from 3 to 4 uncovers additionally an important and easily interpretable pattern of temporal behaviour.



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

Wrap up

Partition-based clustering

- Groups objects into clusters by similarity of attribute values
  - ✓reduces and simplifies the data to analyse
  - ✓facilitates abstraction
  - ☹but involves large information losses
- To decrease the information loss, interact:
  - Vary parameter settings & compare different groups
  - Examine internal variance and refine clusters by progressive clustering

Visualisation of clustering results

- Colour objects by cluster colour across multiple views
- Summarise and visualise attribute values by cluster
- Link back to original data

More clustering to come!

- How clustering algorithms work
  - machine learning module.
- Two-way application of partition-based clustering to multiple time series
  - this module
- Density-based clustering
  - in this module.
- Progressive clustering with different distance functions
  - in this module.

Wider context

- Not only about clustering
- A good example of the general principle of visual analytics
- Principles
  - Iteratively vary parameters and refine your results
  - Visualise **all** your results!

Intended learning outcomes

- Data types and structure
  - And how these affect analysis and interpretation
- How **partition-based clustering** combined with interactive visualisation can help deal with large complex datasets
  - Density-based is the other type of clustering that will be covered next week
- Practical: Using R with Mondrian and Tableau