Lecture 02

Data structures and types

given by
prof. Gennady Andrienko and
prof. Natalia Andrienko

Content and objectives

- Depending on the data structure, properties of the data components, and relationships between them, different analytical tasks may be meaningful and different methods and tools for the data analysis may be useful. In this lecture, you will be acquainted with the major classes of data structures.
- Three classes of data structures, the corresponding analysis tasks, and the visual and interactive tools supporting them will be introduced in more detail.
- In the practical, you will acquire your own experience in visually analysing examples of data of these three classes.

Semantic roles of data components

a reminder

Semantic roles of data components

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

referents

references

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

characteristics (attribute values)

Referential component (referrer)

Characteristic components (attributes)

Data may have >1 referrers

Referrer 1: time

Referrer 2: place

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
•••													
1972	54	West Virgini	1781000	25584	2299	109	146	562	1482	23285	7356	13976	1953
1972	55	Wisconsin	4520000	133382	4358	126	376	1661	2195	129024	28862	89642	10520
1972	56	Wyoming	345000	10461	511	14	48	117	332	9950	2057	7190	703
1973	1	Alabama	3539000	91389	12390	468	751	2809	8362	78999	31754	39206	8039
1973	2	Alaska	330000	16313	1269	33	147	221	868	15044	3852	9456	1736
1973	4	Arizona	2058000	137966	9877	167	637	3031	6042	128089	40301	76560	11228
1973	5	Arkansas	2037000	56149	5905	180	398	1456	3871	50244	18088	29204	2952
1973	6	California	20601000	1298872	116563	1862	8357	49531	56813	1182309	407824	643488	130997
1973	8	Colorado	2437000	133933	10088	193	944	3970	4981	123845	38963	70931	13951
1973	9	Connecticut	3076000	112717	6421	102	342	2589	3388	106296	31661	58742	15893
•••													
2000	44	Rhode Islanc	1048319	36444	3121	45	412	922	1742	33323	6620	22038	4665
2000	45	South Carolir	4012012	209482	32293	233	1511	5883	24666	177189	38888	123094	15207
2000	46	South Dakot	754844	17511	1259	7	305	131	816	16252	2896	12558	798
2000	47	Tennessee	5689283	278218	40233	410	2186	9465	28172	237985	56344	154111	27530
2000	48	Texas	20851820	1033311	113653	1238	7856	30257	74302	919658	188975	637522	93161
2000	49	Utah	2233169	99958	5711	43	863	1242	3563	94247	14348	73438	6461
2000	50	Vermont	608827	18185	691	9	140	117	425	17494	3501	13184	809
2000	51	Virginia	7078515	214348	19943	401	1616	6295	11631	194405	30434	146158	17813
2000	53	Washington	5894121	300932	21788	196	2737	5812	13043	279144	53476	190650	35018
2000	54	West Virgini	1808344	47067	5723	46	331	749	4597	41344	9890	28139	3315
2000	55	Wisconsin	5363675	172124	12700	169	1165	4537	6829	159424	25183	119605	14636
2000	56	Wyoming	493782	16285	1316	12	160	70	1074	14969	2078	12318	573

References are not always in columns

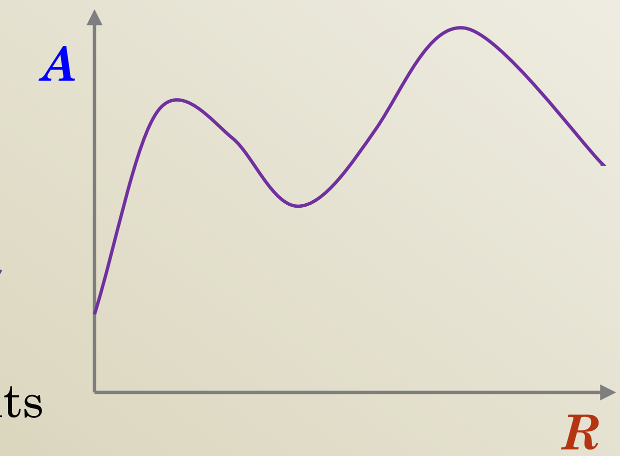
Referrer 1: place

Referrer 2: time →

identifiers	financial year=2001/__: 1 Total offences (Offences rates)	financial year=2002/__: 1 Total offences (Offences rates)	financial year=2003/__: 1 Total offences (Offences rates)	financial year=2004/__: 1 Total offences (Offences rates)	financial year=2005/__: 1 Total offences (Offences rates)	financial year=2006/__: 1 Total offences (Offences rates)	financial year=2007/__: 1 Total offences (Offences rates)	financial year=2008/__: 1 Total offences (Offences rates)	financial year=2009/__: 1 Total offences (Offences rates)	financial year=2010/__: 1 Total offences (Offences rates)	financial year=2011/__: 1 Total offences (Offences rates)
E05000026 Abbey	278	303	279	300	248	255	209	215	221	206	183
E05000027 Alibon	79	84	86	81	90	97	89	95	97	96	91
E05000028 Becontree	86	89	105	106	112	110	101	102	109	98	99
E05000029 Chadwell Heath	118	133	157	157	153	138	127	129	112	119	97
E05000030 Eastbrook	75	86	77	76	91	92	87	85	84	67	74
E05000031 Eastbury	100	96	109	119	105	134	121	105	97	106	81
E05000032 Gascoigne	277	239	220	190	170	162	133	123	107	129	122
E05000033 Goresbrook	78	75	99	103	99	86	89	89	93	94	84
E05000034 Heath	104	107	104	117	115	128	119	107	99	97	103
E05000035 Longbridge	77	67	82	89	71	89	77	73	75	81	79
E05000036 Mayesbrook	90	74	72	92	95	102	96	100	100	94	80
E05000037 Parsloes	77	79	70	92	85	97	92	89	84	75	76
E05000038 River	113	103	114	102	121	115	109	99	96	86	93
E05000039 Thames	263	226	254	227	215	209	180	178	176	163	130
E05000040 Valence	81	75	83	87	86	97	89	76	80	74	82
E05000041 Village	111	124	143	120	115	134	130	129	130	89	100
E05000042 Whalebone	109	93	96	97	114	119	99	102	107	93	86
E05000043 Brunswick Park	70	68	84	78	66	61	55	54	56	52	56
E05000044 Burnt Oak	88	108	102	114	95	95	79	70	76	67	64
E05000045 Childs Hill	119	125	129	142	136	127	118	106	106	105	100
E05000046 Colindale	93	114	114	107	95	93	76	76	70	77	71
E05000047 Coppetts	99	104	115	113	100	89	76	75	81	82	77
E05000048 East Barnet	61	62	80	88	82	61	66	65	61	57	59
E05000049 East Finchley	85	87	95	92	81	84	59	67	61	57	65
E05000050 Edgware	107	108	107	121	114	106	89	84	86	97	91
E05000051 Finchley Church	63	59	58	64	62	56	56	51	58	59	51
E05000052 Garden Suburb	87	83	88	105	70	72	74	65	66	71	79
E05000053 Golders Green	94	98	101	112	96	96	79	70	68	70	76
E05000054 Hale	58	63	65	70	72	70	54	54	55	53	54
E05000055 Hendon	88	83	101	94	106	86	84	93	82	73	72
E05000056 High Barnet	78	92	105	109	101	75	80	80	72	68	81
E05000057 Mill Hill	79	90	103	93	85	90	84	74	71	73	77
E05000058 Oakleigh	57	63	72	79	72	61	58	62	51	49	59

Data components viewed as variables

- Referrers ~ independent variables
- Attributes ~ dependent variables
- Data represent (are generated by) a *function*
  Referrers → Attributes
  - References (values of referrers): function's inputs
  - Characteristics (values of attributes): function's outputs
- Function's *behaviour*:
  how the outputs vary over the set of inputs →
  how the characteristics vary over the set of references
  - E.g., how the crime rates vary over space and time



Definitions of data types and structures

Types and properties of data components

Types of values:

- Numeric
- Textual
  - Predefined values (e.g., codes)
  - Free text
- Spatial
  - Coordinates
  - Place names
  - Addresses
- Temporal
- Other (image, video, audio, …)

Scales of measurement*:

- Nominal *(¬ order, ¬ distances)*
  - gender, nationality, …
- Ordinal *(✓order, ¬ distances)*
  - evaluations: bad, fair, good, excellent
- Interval *(✓order, ✓distances, ¬ ratios, ¬ meaningful zero)*
  - temperature, time, …
- Ratio *(✓order, ✓distances, ✓ratios, ✓meaningful zero)*
  - quantities, distances, durations, …

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

* Stevens, S.S. (1946). On the Theory of Scales of Measurement. *Science* **103** (2684): 677–680.

Essential properties of sets (value domains of data components)*

- *Ordering* between the elements
  - No ordering, partial ordering, full (linear) ordering
- Existence of *distances* between the elements
  - No distances, has distances
- *Continuity*
  - Discrete, continuous

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

- For any data component, there is a certain set (domain) of possible values

  ⇒ Data components can be characterised in terms of the properties of their domains

Three main types of referrers according to their value domains*

- Time
  - Has linear ordering, has distances, continuous
- Space (2D, 3D)
  - No ordering, has distances, continuous
- Population *(in a general sense, like statistical population, i.e., any set of material or abstract objects)*
  - No ordering, no distances, discrete
    - Further will be referred to as “discrete objects”, or simply “objects”, to avoid confusion with “population” in demographic sense

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

Main classes of data structures

according to the types of the referrers

- *Object-referenced data*:
  - attribute values refer to discrete objects
- *Time-referenced data*, a.k.a. *time series*:
  - attribute values refer to different times (moments or intervals)
- *Space-referenced data*, a.k.a. *spatial data*:
  - attribute values refer to different spatial locations (points, lines, areas, volumes in 3D space)
- *Object-referenced time series*:
  - attribute values refer to discrete objects and to different times (i.e., the data have 2 referrers)
- *Spatial time series*:
  - attribute values refer to different spatial locations and to different times (i.e., the data have 2 referrers)

Multi-dimensional data (a.k.a. multivariate data)

- Data including multiple attributes
- May have various types of references: discrete objects, spatial locations, times, or combinations of these
  - Multiple attributes referring to times: *multidimensional (multivariate) time series*
  - Multiple attributes referring to places: *multidimensional spatial data*
  - Multiple attributes referring to places + times: *multidimensional spatial time series*

An example: London crime data

- Referrers:
  - London wards – spatial
  - Financial years – temporal
- Attributes:
  - 10 crime rates (total + different types of crimes)
  - 10 crime counts (total + different types of crimes)

⇒ Multidimensional spatial time series

(T) 1 Total offences (Offences rates)
(T) 2 Violence Against The Person (Offences rates)
(T) 3 Sexual Offences (Offences rates)
(T) 4 Robbery (Offences rates)
(T) 5 Burglary (Offences rates)
(T) 6 Theft And Handling (Offences rates)
(T) 7 Fraud Or Forgery (Offences rates)
(T) 8 Criminal Damage (Offences rates)
(T) 9 Drugs (Offences rates)
(T) 10 Other Notifiable Offences (Offences rates)
(T) 1 Total (Offences numbers)
(T) 2 Violence Against The Person (Offences numbers)
(T) 3 Sexual Offences (Offences numbers)
(T) 4 Robbery (Offences numbers)
(T) 5 Burglary (Offences numbers)
(T) 6 Theft And Handling (Offences numbers)
(T) 7 Fraud Or Forgery (Offences numbers)
(T) 8 Criminal Damage (Offences numbers)
(T) 9 Drugs (Offences numbers)
(T) 10 Other Notifiable Offences (Offences numbers)

Types of objects

based on their properties and attributes***

- **Generic objects**
    - Wine varieties, car models, plant specimens, …
- **Spatial objects**
  - Have locations in space ⇒ the attributes include the location
    - Districts, buildings, streets, rivers, …
- **Temporal objects**, a.k.a. ***events***
  - Have limited existence time ⇒ the attributes include the time of existence
    - Instant objects: have no duration; only the appearance time needs to be specified
      - Tweet postings, bank transactions, lightning strokes …
    - Durable objects: have duration; the attributes need to include the time of appearance + the time of disappearance or the duration
      - Holidays, electoral campaigns, classes, breaks, TV shows, …

---

* What is observed in the reality   ** Representations of properties in the data

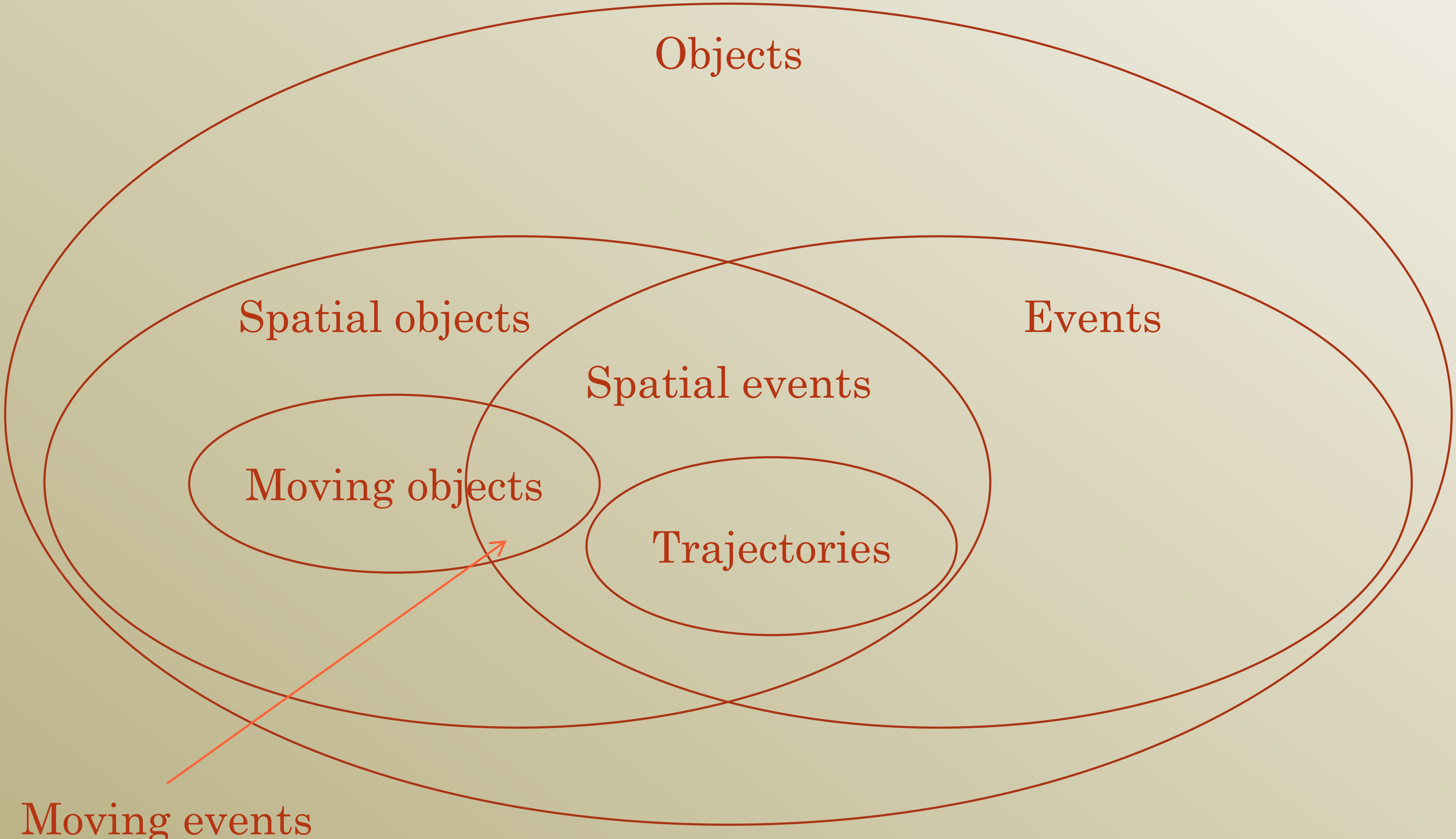
Types of objects *(continued)*

spatio-temporal objects

- **Spatial events**
  - Events that have location in space ⇒ the attributes include the spatial location and the existence time (instance or interval)
    - Lightnings, geolocated tweet postings, earthquakes, traffic jams, …
- **Moving objects**
  - Spatial objects whose locations change over time ⇒ the attributes include spatial locations at different times, i.e., time series of spatial locations, called ***trajectories***
    - People, animals, vehicles, storms, oil spills, …
- **Trajectories** can be treated as spatio-temporal objects
  - Spatial location = spatial footprint
  - Existence time = time from the beginning till the end of the movement
  - Other properties → attributes: shape, travelled distance, mean speed

Relationships between types of objects

Venn diagram



Classes of data to be considered in this lecture

- Object-referenced attributes
  - References: generic objects (have no temporal or spatial properties, or these properties are ignored)
- Space-referenced attributes
  - References: spatial locations or stationary spatial objects
- Spatial event data
  - References: spatial events

➢ All these are data with single referrers

➢ Next lecture: data with 2 referrers

Questions?

Definitions of data types and structures

Following material

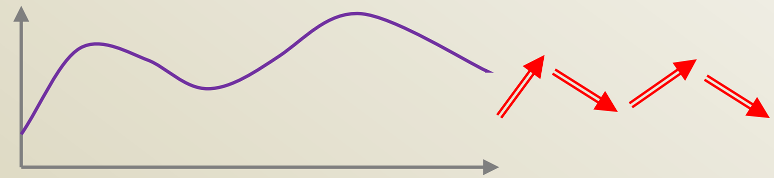
- Classes of data
  - Object-referenced attributes
  - Space-referenced attributes
  - Spatial event data
- For each class of data, we shall consider the typical *synoptic tasks* and the methods to accomplish them
- Synoptic tasks (a reminder):
  - Distinguishing property: a set or subset of data items is considered in its entirety; individual items are not of interest.

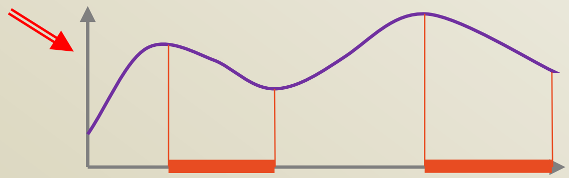
  $\Rightarrow$ Synoptic tasks entail *abstraction* from individual items to sets

  $\Rightarrow$ Synoptic tasks require analysis methods and visualisations supporting abstraction

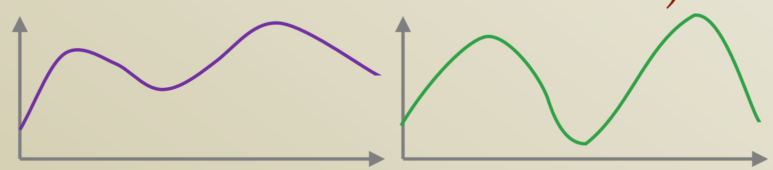
Major classes of synoptic tasks : a reminder

- Describe the behaviour of one or more attributes

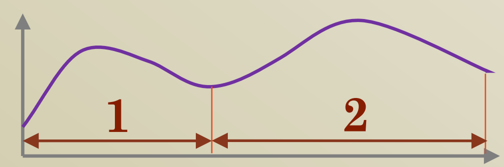




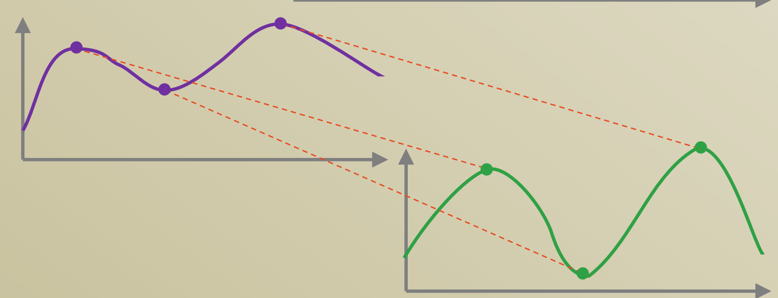
- Find subsets of references where attributes have particular behaviours
- Compare two or more behaviours (find similarities and differences)
  - Different attributes over the same set of references



  - Same attributes over different subsets of references



- Relate behaviours of two or more attributes



Classes of synoptic tasks with examples

- Describe the behaviour of one or more attributes
    - Describe the variation of the crime rates over time and space
- Find subsets of references where attributes have particular behaviours
    - Find time periods of decreasing burglary rates
- Compare two or more behaviours (find similarities and differences)
    - Different attributes over the same set of references
        - Compare the variations of the burglary and robbery rates
    - Same attributes over different subsets of references
        - Compare the temporal variations of the burglary rates in the eastern and western parts
- Relate behaviours of two or more attributes
    - Relate the crime rates to socio-demographic characteristics

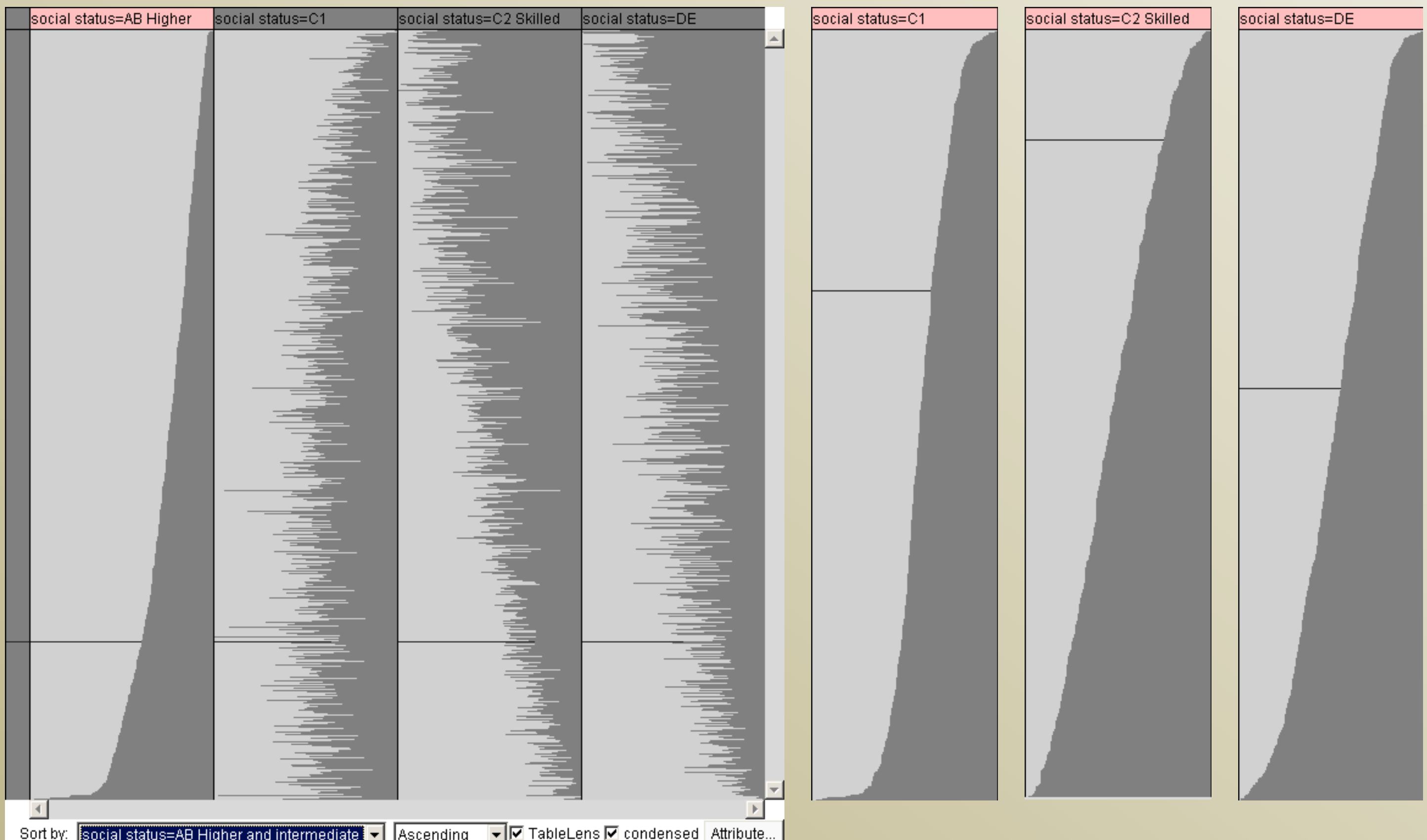
Analysis of object-referenced data

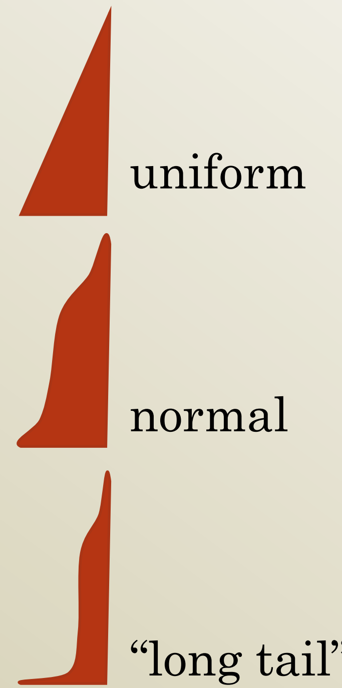
Attributes referring to discrete objects

Major classes of synoptic tasks

- Describe the behaviour of one or more attributes →
  Describe the distribution of the attribute values and value combinations over the set of objects
- Find subsets of references where attributes have particular behaviours →
  Find subsets (groups) of objects with particular attribute values or combinations, or with value combinations substantially differing from the bulk
- Compare two or more behaviours →
  Compare two or more value distributions
  - Different attributes over the same set of objects
  - Same attributes over different subsets of objects
- Relate behaviours of two or more attributes →
  Relate value distributions of two or more attributes

Task: Describe the distribution of a single numeric attribute

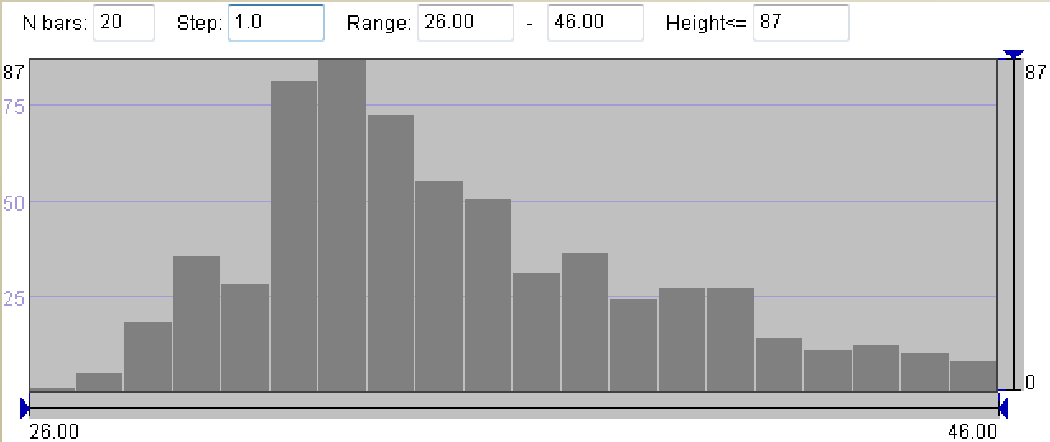




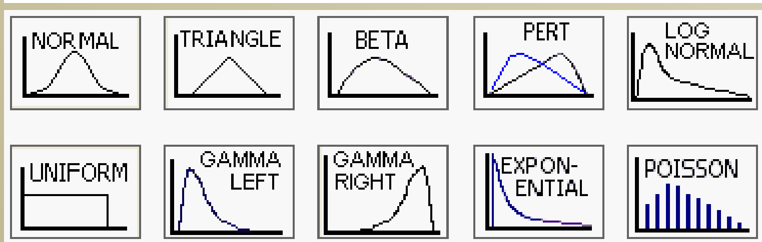
A possible approach: bar diagram (sorted). However, for supporting synoptic tasks, bars for all objects must be seen simultaneously ⇒ it is limited to quite a small number of objects.

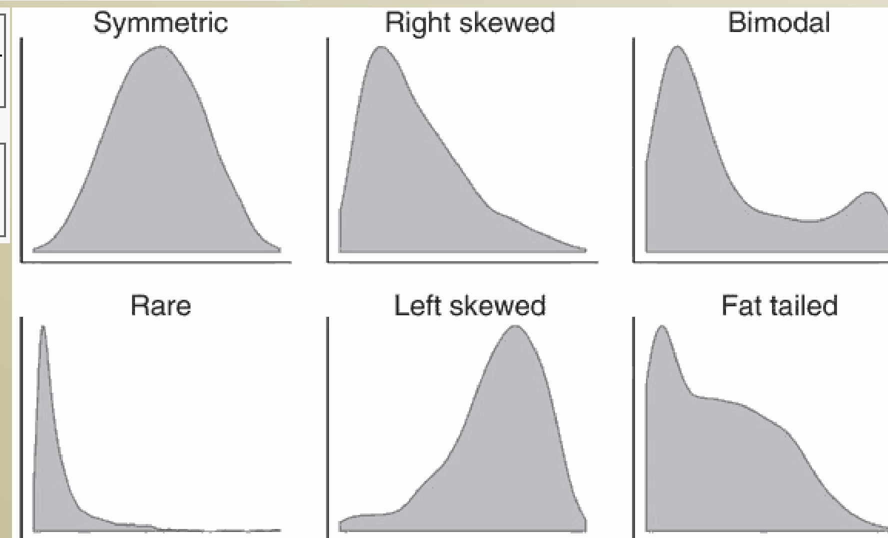
Task: Describe the distribution of a single numeric attribute

A better option: frequency histogram

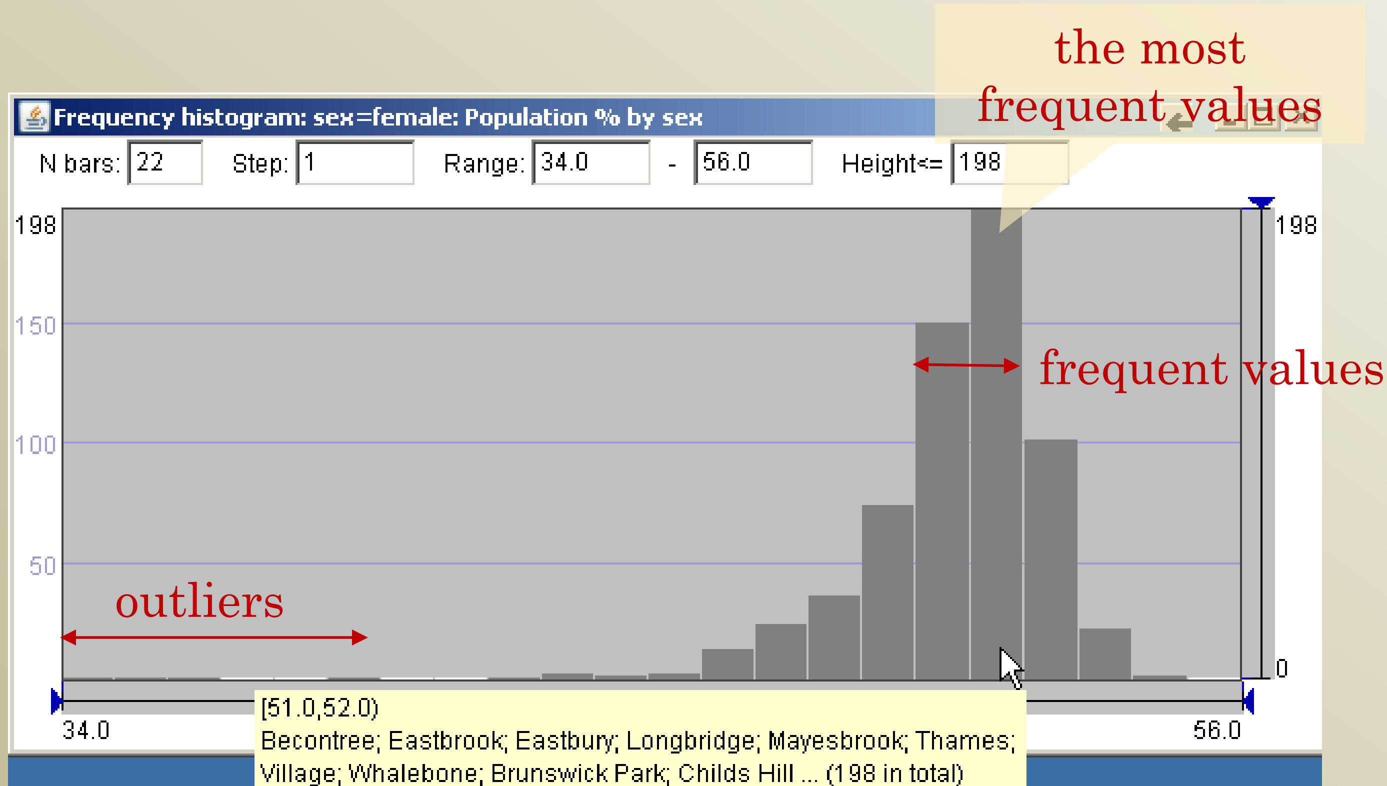


Shows the most popular value intervals and exhibits outliers. The shape can be compared with the “model” distribution shapes defined in statistics.

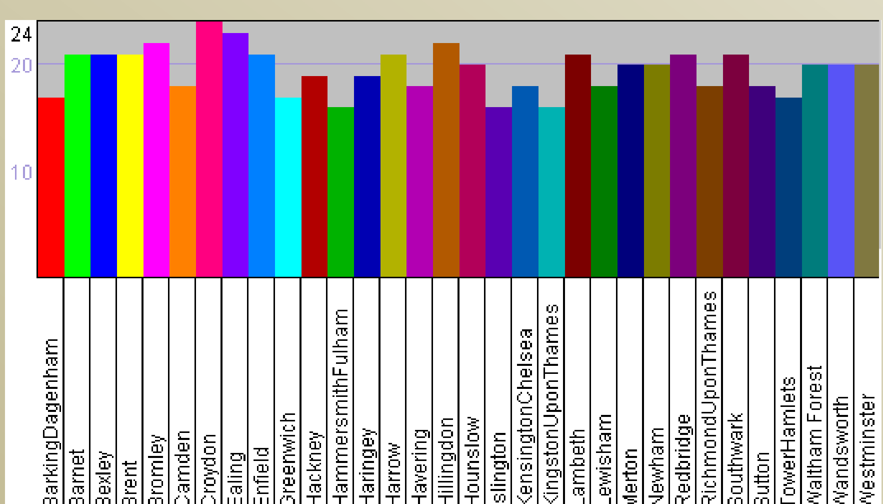


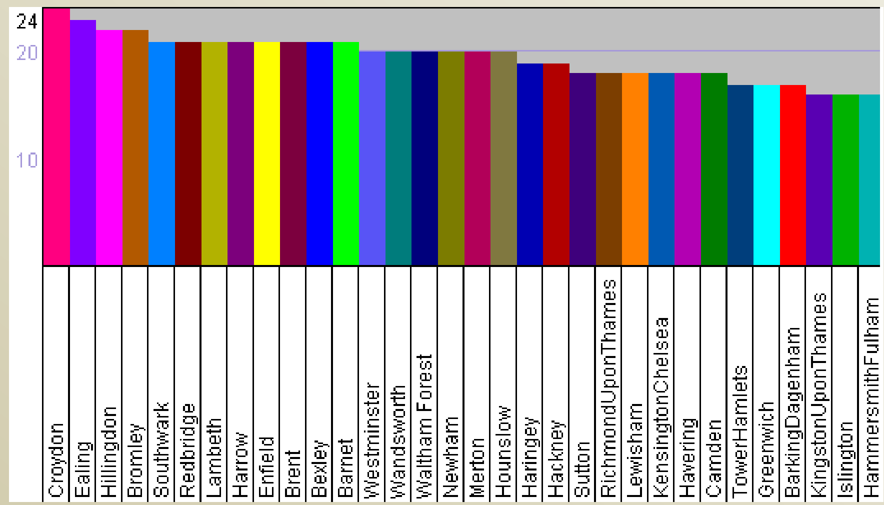


A frequency histogram example



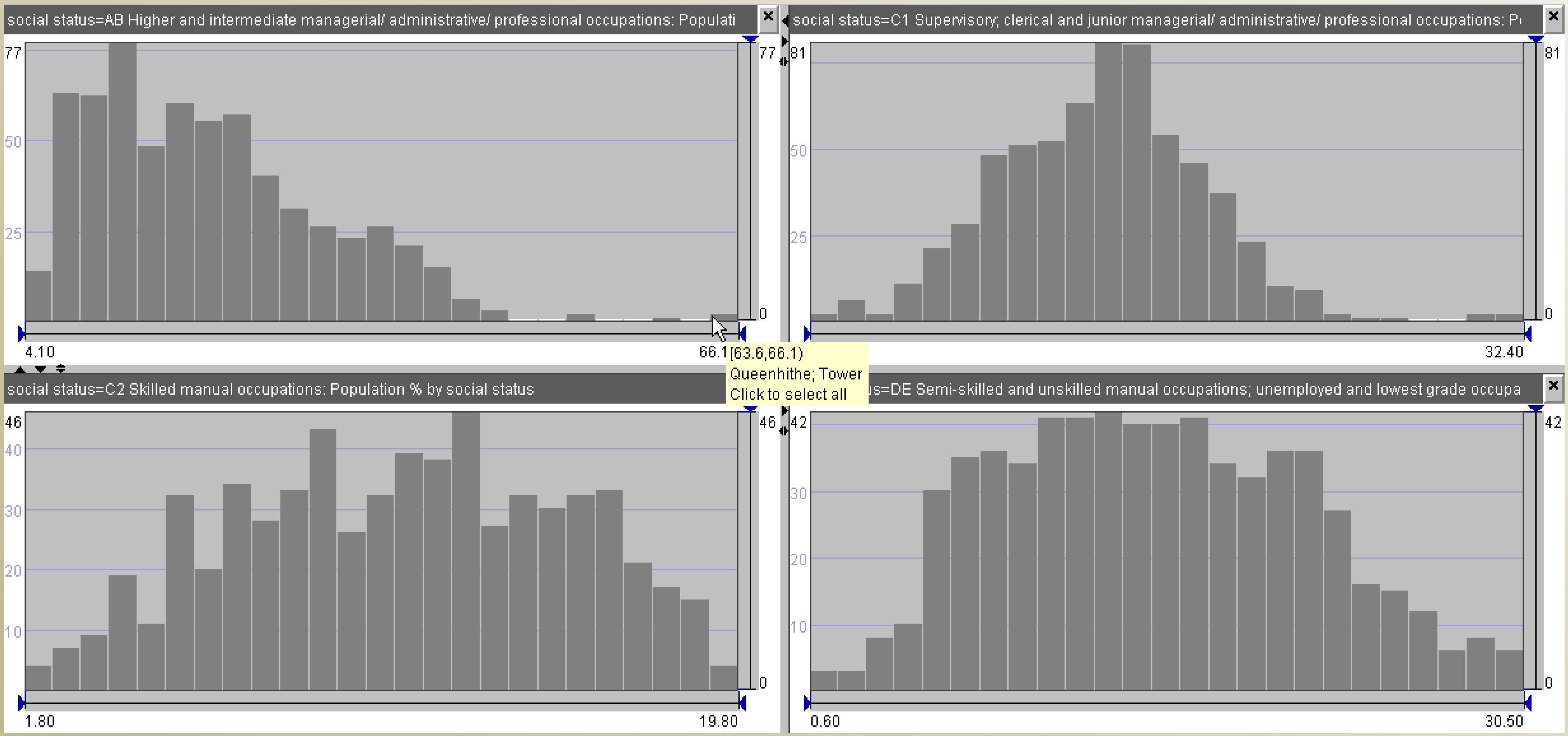
Frequency histogram for a qualitative (nominal) attribute





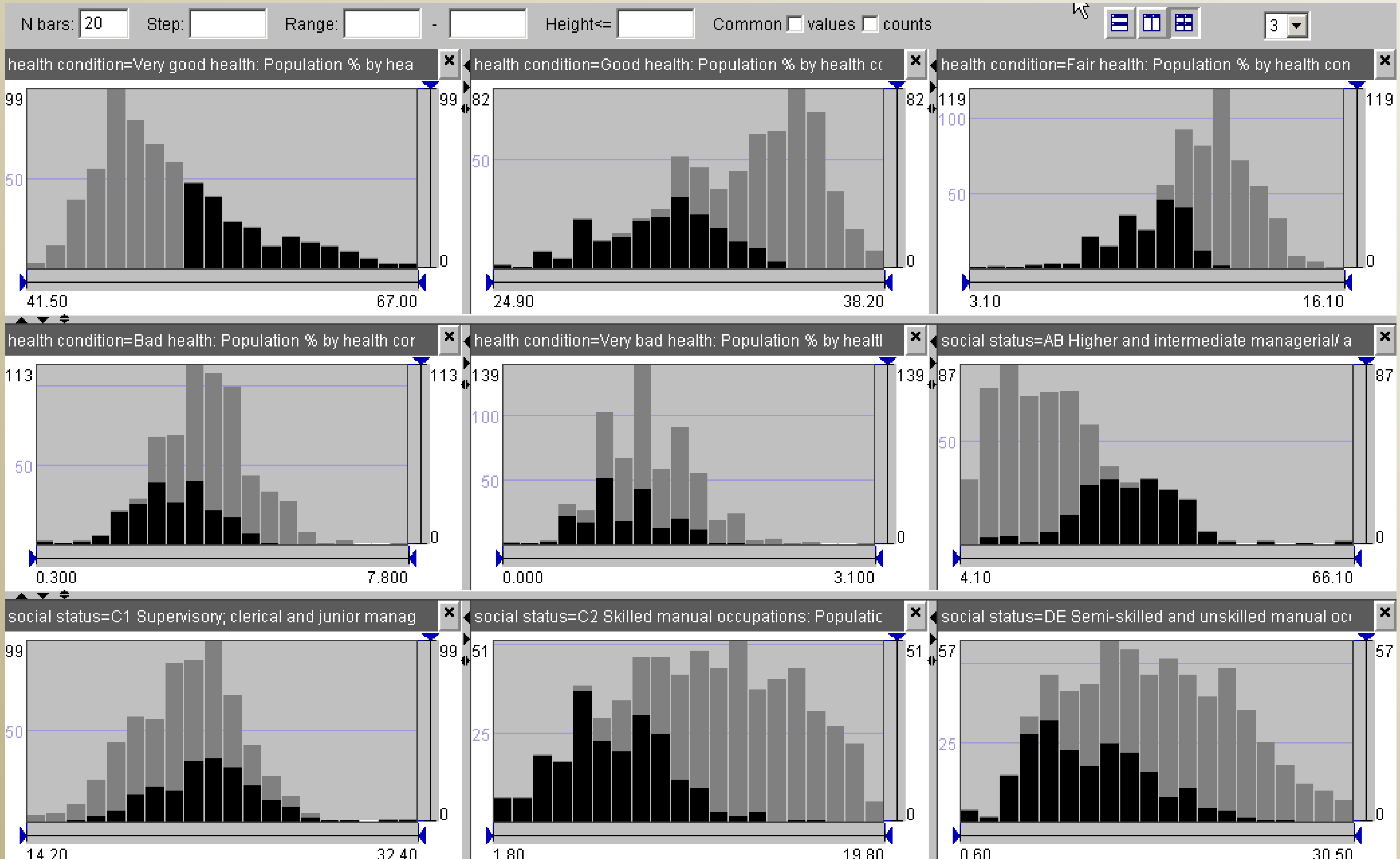
May be better when sorted

Other synoptic tasks supported by interactive frequency histograms



Compare value distributions of several attributes over the same object set.

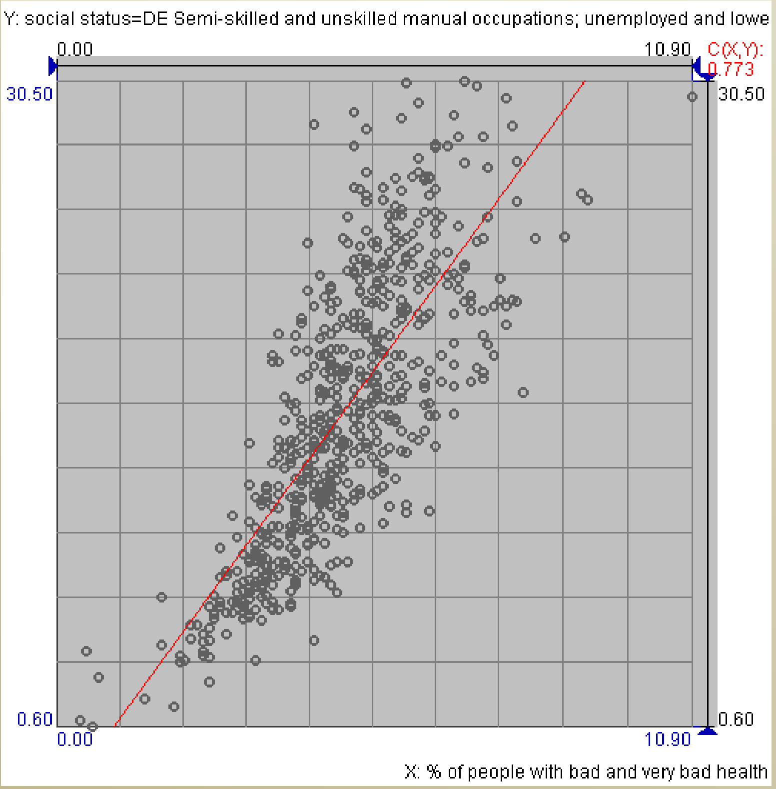
Other synoptic tasks supported by interactive frequency histograms



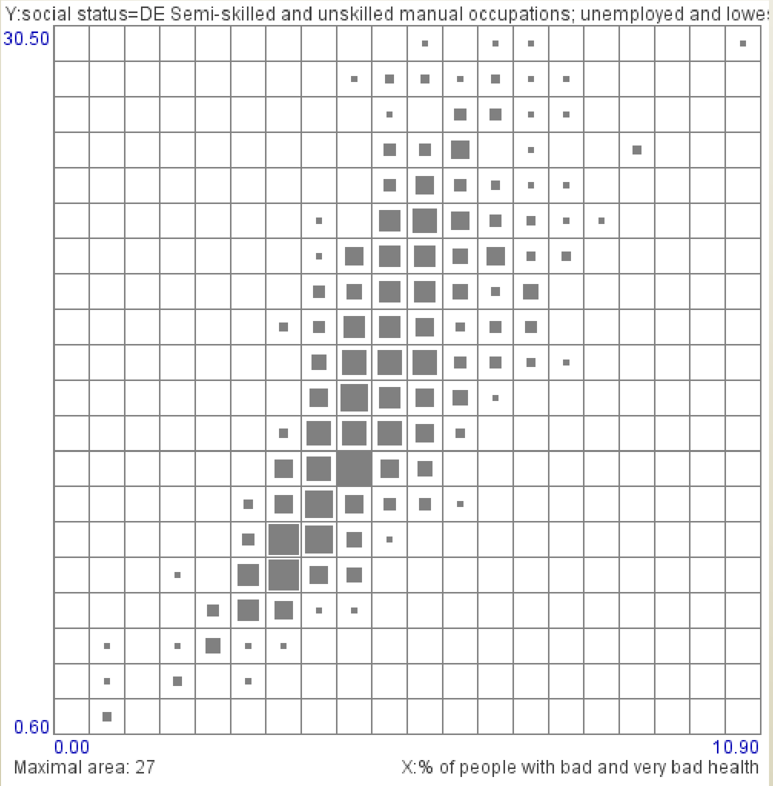
Relate value distributions of several attributes.

Task: Relate value distributions of two attributes

Scatter plot



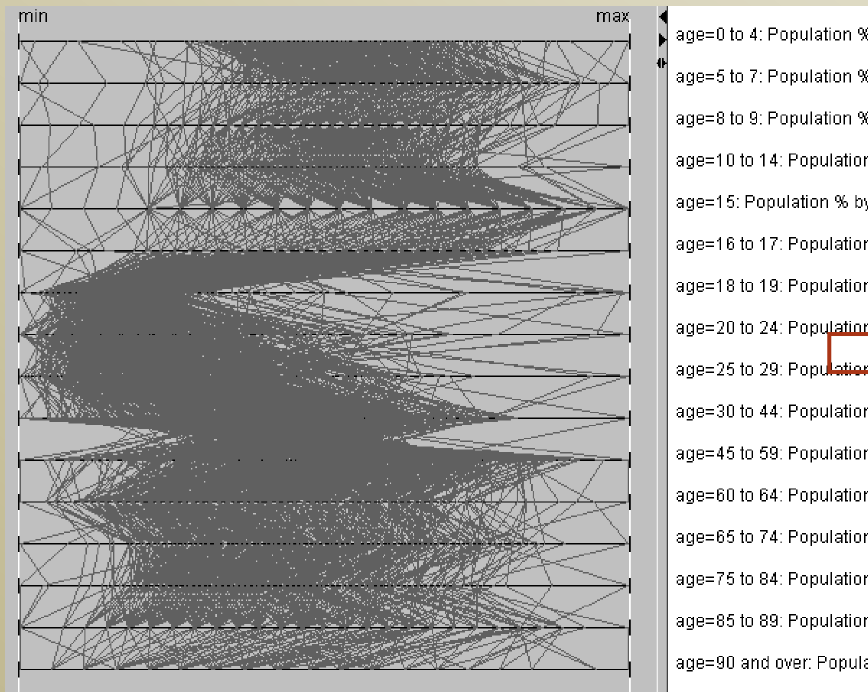
2D histogram (binned scatter plot)

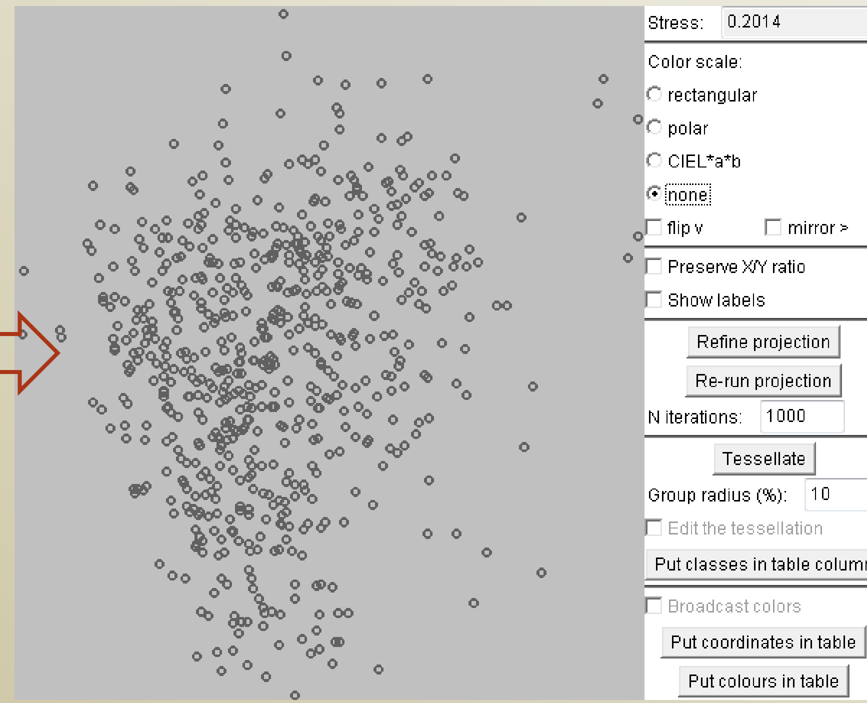


Possible relationships: positive correlation, negative correlation, independence; clusters (groups of similar value combinations).

Task: Describe joint value distribution of multiple attributes

Step 1: projection (dimensionality reduction)





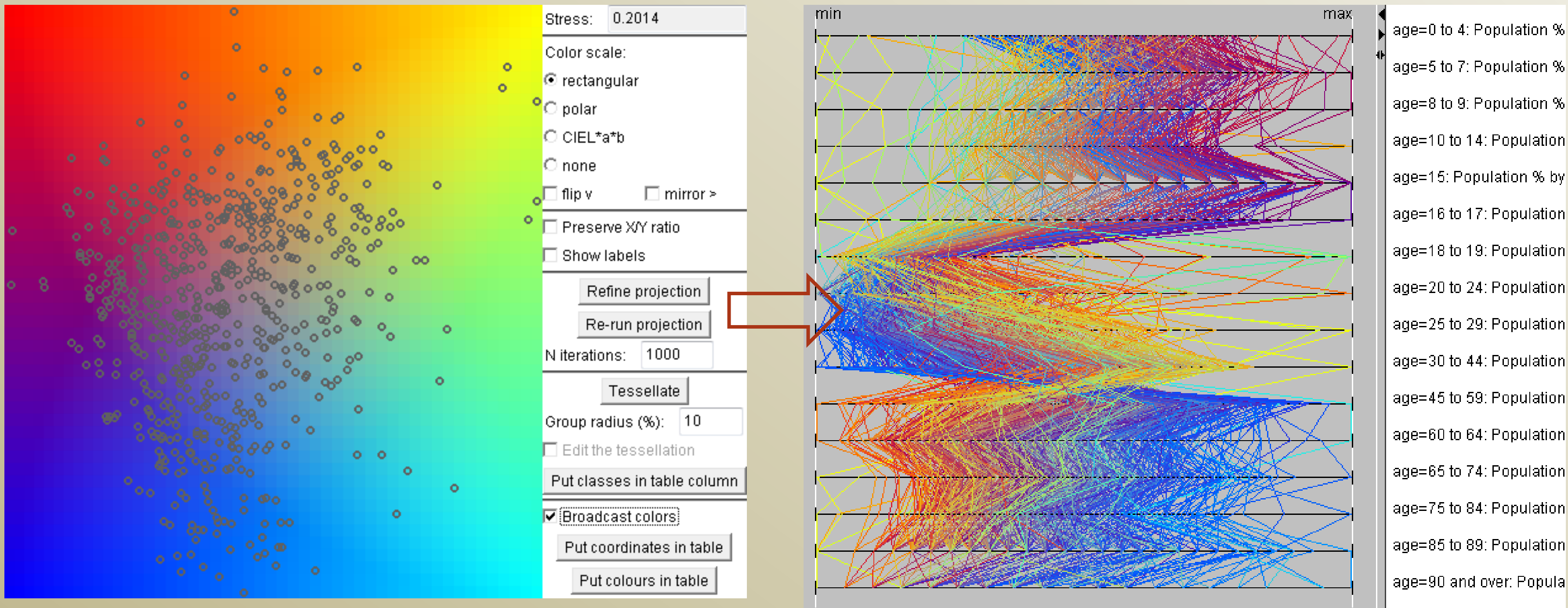
Combinations of values of multiple attributes

Positions in 2D space

Projection method here: Sammon's mapping; may be also MDS, PCA, …

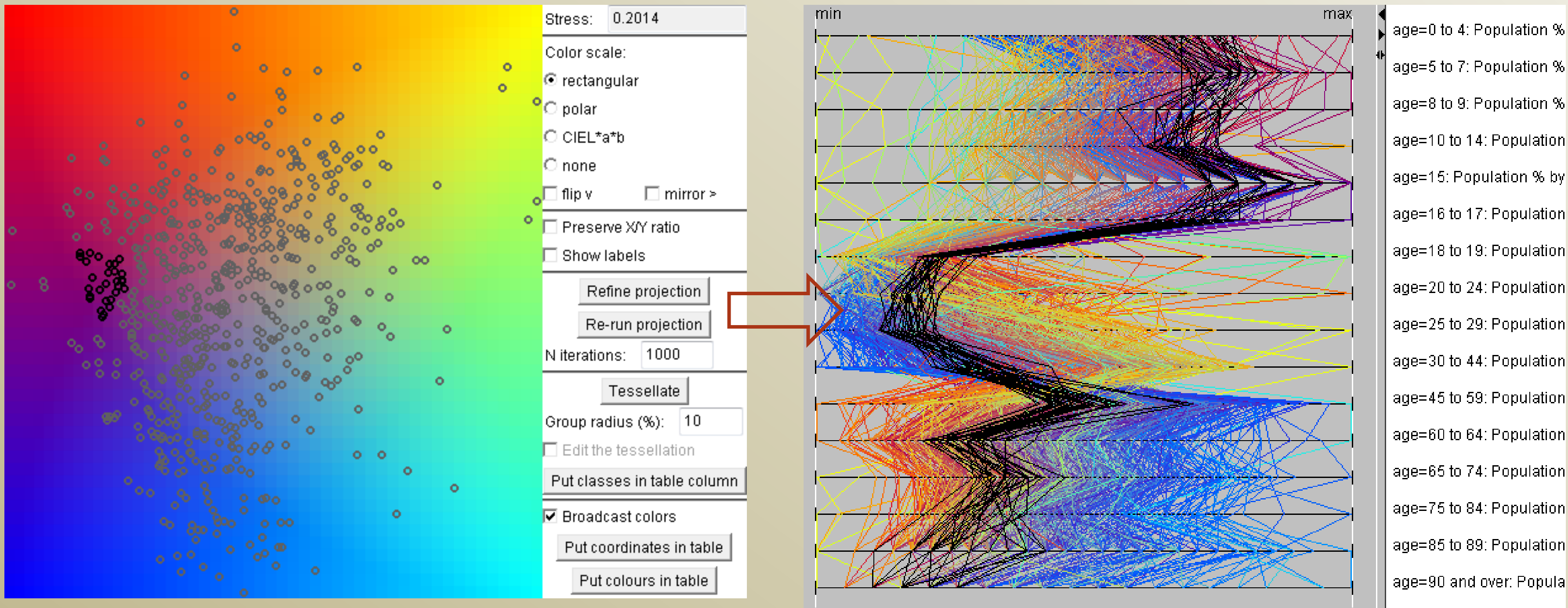
Task: Describe joint value distribution of multiple attributes

Step 2: superimpose the projection on a continuous 2D colour scale



A parallel coordinates plot helps to understand what value combinations correspond to different areas of the projection plot.

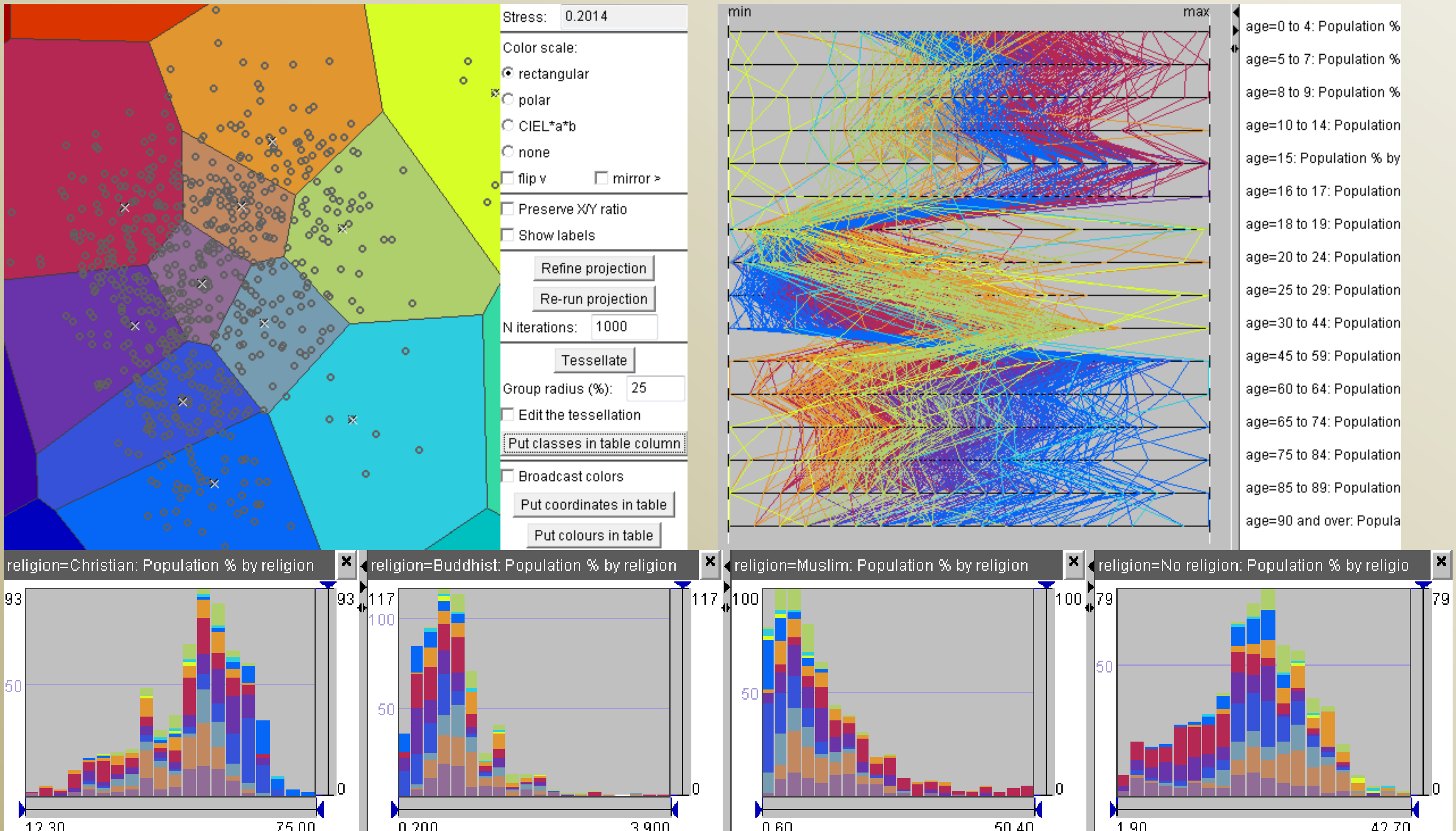
Tasks: describe joint value distribution of multiple attributes; find groups of objects with similar value combinations



Select (by clicking or dragging) a group of close points in the projection plot; then look at the corresponding marks in the PCP and in other displays.

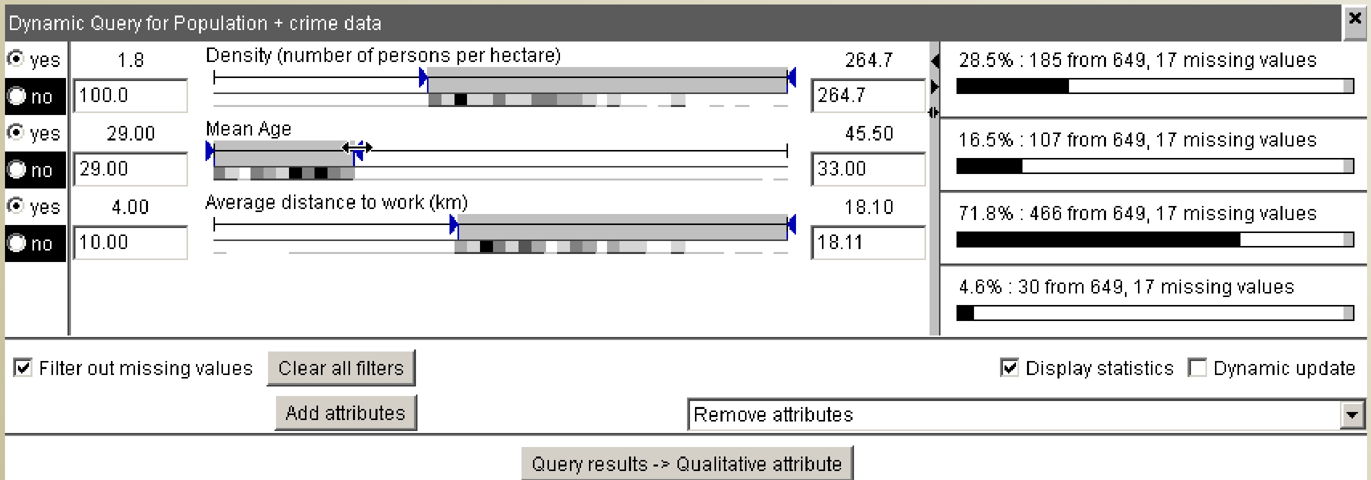
Further use of projection

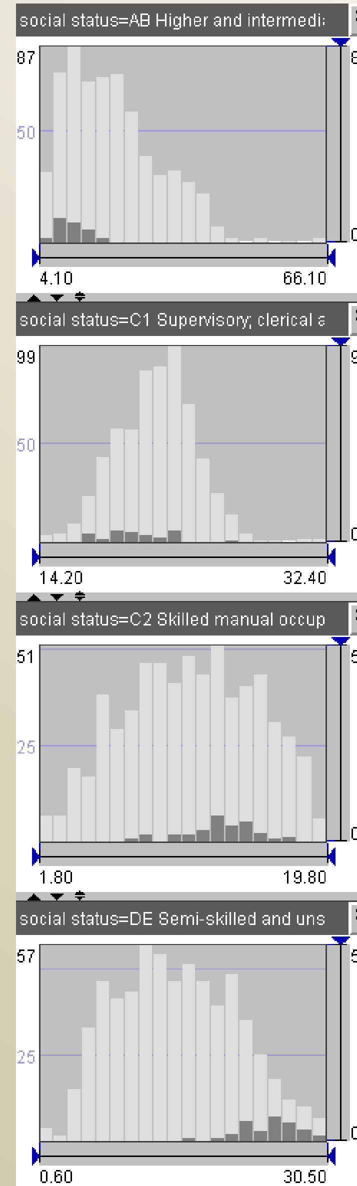
Grouping of objects by projection space tessellation



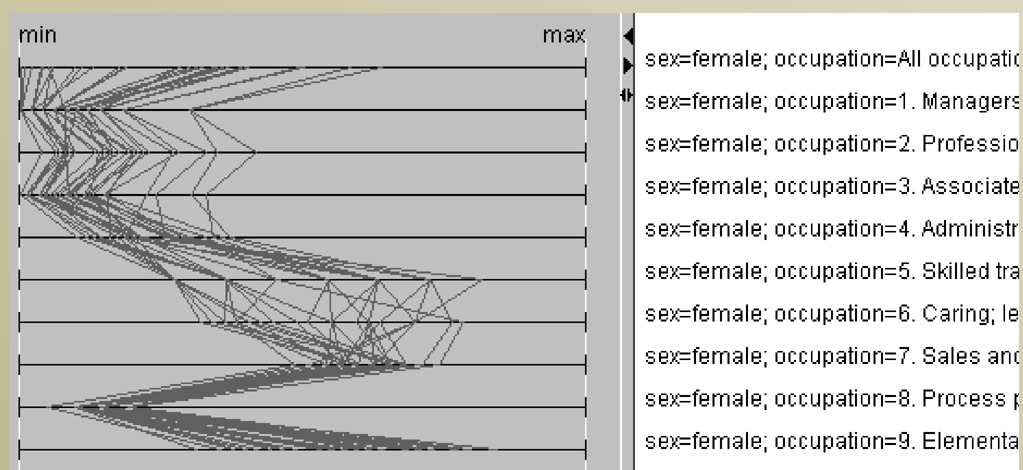
Creates object classes, which can be propagated to other displays.

Task: find groups of objects with particular value combinations





Supporting interactive tool: filtering by attribute values.



Techniques for analysing object-referenced data

- Visualisations: bar diagram, parallel coordinates plot (both suitable for small number of objects), frequency histogram, scatter plot, 2D histogram, projection plot.
- Transformations: aggregation (used in histograms), projection (dimensionality reduction).
  - Aggregation: helps to deal with numerous objects; supports abstraction.
  - Projection: helps to deal with multiple attributes.
- Interactive operations: selection, grouping (classification), filtering.
  - Support relating multiple attributes and finding groups of objects with particular/ close/ outstanding value combinations.

Questions?

Analysis of object-referenced data

Analysis of space-referenced data

Attributes referring to spatial locations or spatial objects

Space as a special reference domain

- Essential property: distances between elements
  ⇒ Spatial locations and spatial objects have neighbourhoods
- Behaviour of attributes over space = spatial distribution of attribute values and combinations
- In analysing and describing spatial distributions, spatial neighbourhood is taken into account
- Types (patterns) of spatial distribution based on neighbourhood
  - Spatial clusters: groups of neighbouring locations or objects having similar attribute values
  - Spatial smoothness: attribute values of neighbouring locations or objects do not differ much

Tobler's first law of geography

- Everything is related to everything else, but near things are more related than distant things*.
- A.k.a. *spatial dependence*.
- Implication: neighbouring objects or locations are expected to have similar attribute values.

⇒ Spatially smooth or spatially clustered attribute values are a kind of “normal” spatial distribution

  - *Spatial outliers*: locations or objects whose attribute values differ much from those of the neighbours.

- However, the first law of geography may not keep for spatially aggregated data (e.g., by districts).

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

Spatial directions

- Space has no natural ordering between locations.
- However, introducing a coordinate system also defines directions, such as compass directions in the geographic space.
- Type (pattern) of spatial distribution based on directions:
  - Spatial trend: increase or decrease of attribute values
    - in some direction w.r.t. the coordinate system (e.g., from north to south) or
    - in all/many directions from some location (e.g., from the city centre to the periphery)

Geographic space

- Spatially referenced data typically refer to geographic rather than abstract space.
- The geographic space is *heterogeneous*: consists of land and sea, plains, mountains, and valleys, rivers, lakes, and deserts, forests and built areas, …
- The spatial distribution of attribute values may be affected by the variation of the properties of the underlying geographic space.
    - This interferes with the first law of geography: near things may be less related or less similar due to the differences of the underlying space.

Distribution of attribute values over geographic space

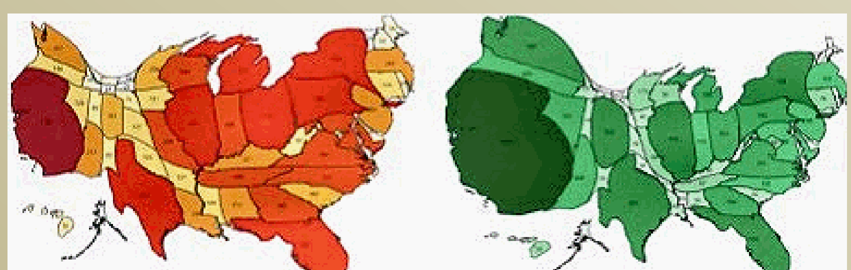
- In analysing spatial distributions, the geographic background should be taken into account.
- Types (patterns) of spatial distribution related to heterogeneous geographic background:
  - Differentiation based on altitude, land cover, land use, etc.
  - Alignment of similar values (elongated spatial cluster) or trend (increase or decrease) along a linear natural or man-made feature (river, road, …)
  - Spatial trend towards/outwards a boundary line (e.g., a coastline)
  - Concentration of high/low/specific values around some feature (e.g., a factory polluting air)

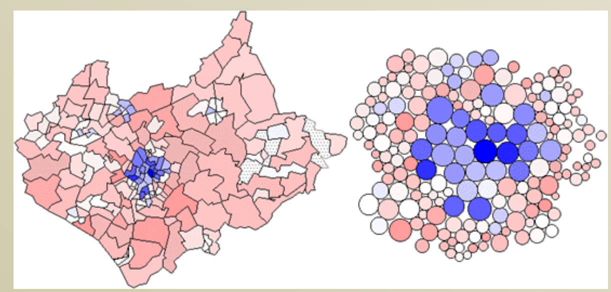
Major classes of synoptic tasks

- Describe the behaviour of one or more attributes →
  Describe the *spatial distribution* of the attribute values and value combinations
- Find subsets of references where attributes have particular behaviours →
  Find parts of space with particular attribute values or combinations, or with value combinations substantially differing from the bulk
- Compare two or more behaviours →
  Compare two or more spatial distributions of attribute values
  - Different attributes over the same part of space (territory)
  - Same attributes over different parts of space
- Relate behaviours of two or more attributes →
  Relate value distributions of two or more attributes

Map: the main instrument in analysing spatial data

- A map can show spatial distribution of spatial objects and space-referenced attribute values.
  - It allows the analyst to observe the spatial distribution patterns based on distance/neighbourhood relationships and on spatial directions.
- For geographic space, a map can also show the heterogeneous geographic background.
  - It thereby allows the analyst to observe the spatial distribution patterns based on relationships with the background properties and features.

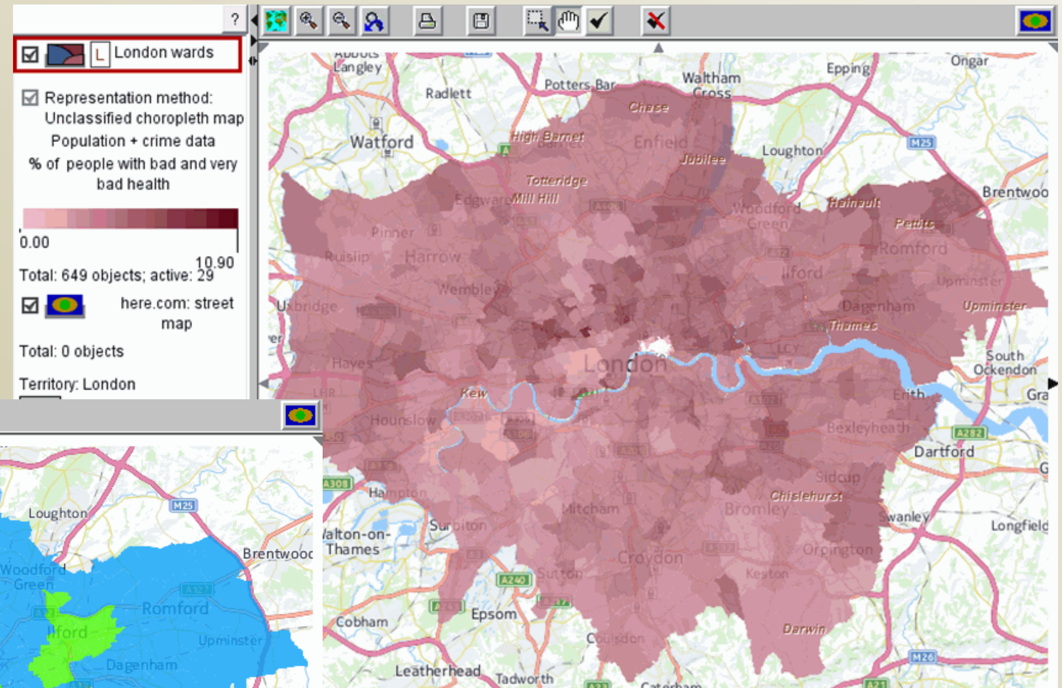


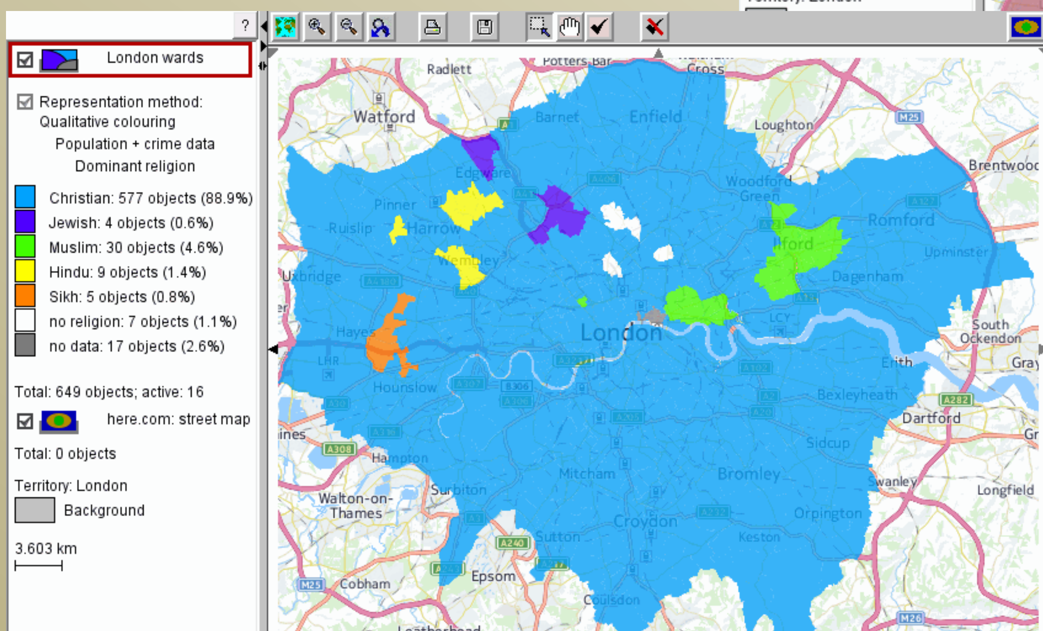


Cartograms:
- distort the spatial directions
- distort the distance and (often) neighbourhood relationships
- the geographic background can hardly be shown together with data

Efficient maps of spatial distributions of attribute values

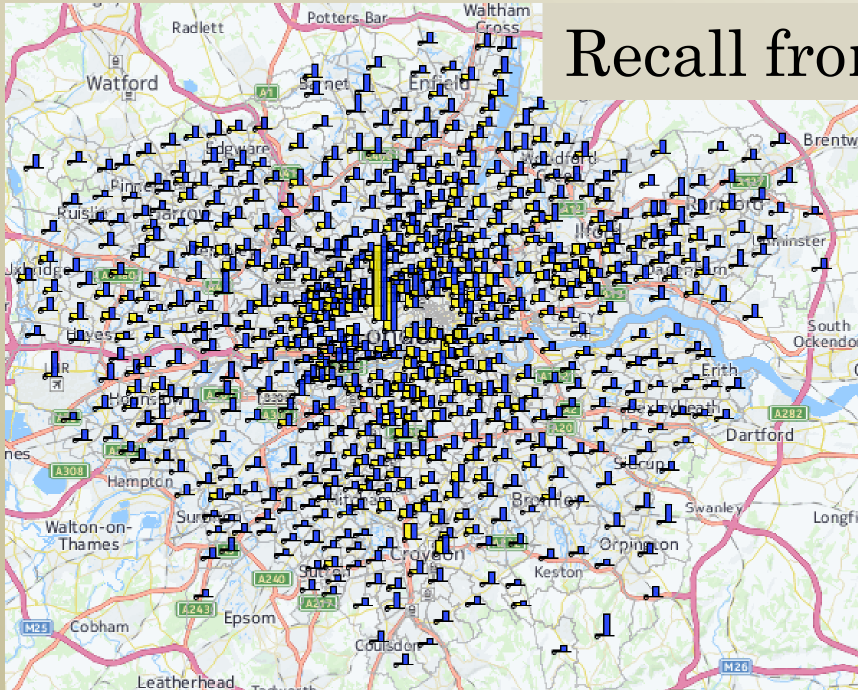
The most efficient map of a spatial distribution is a map from which the distribution is perceived as a single image, i.e., “in the minimum instant of vision” (J.Bertin).



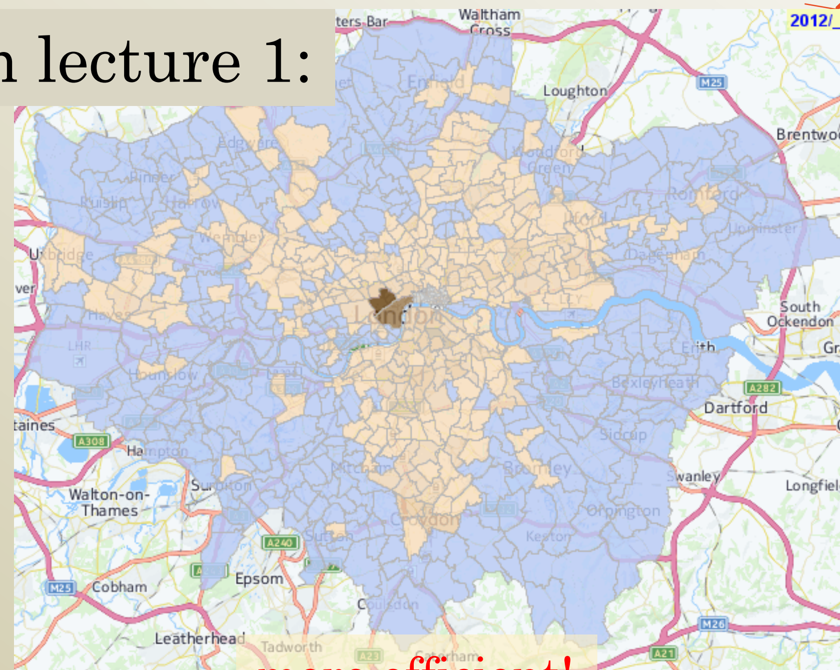


The use of colour darkness (above) or colour hue (left) creates an efficient map of the spatial distribution of a single numeric (above) or qualitative (left) attribute.

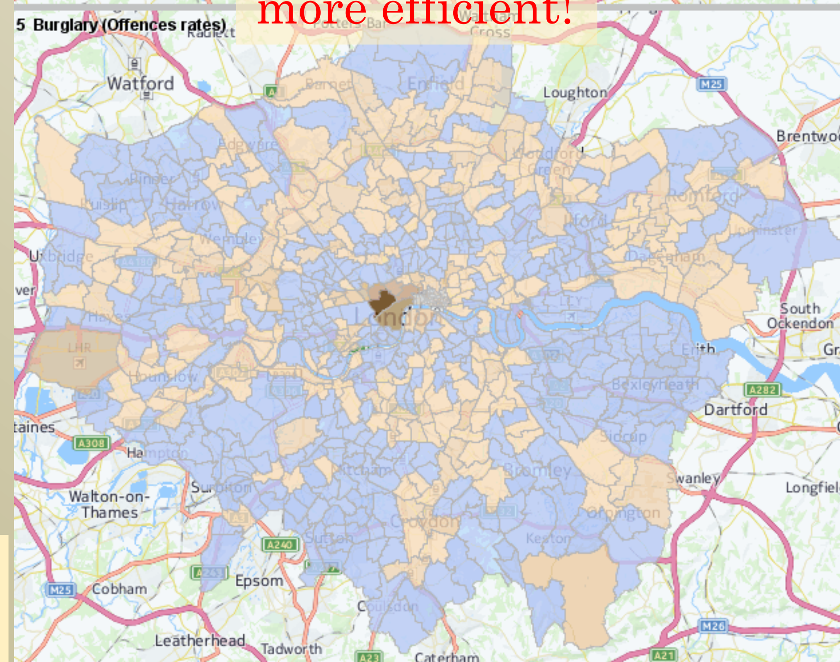
Recall from lecture 1:



Each diagram on the map requires an individual instance of perception, i.e., N of images = N of diagrams = N of districts.



more efficient!

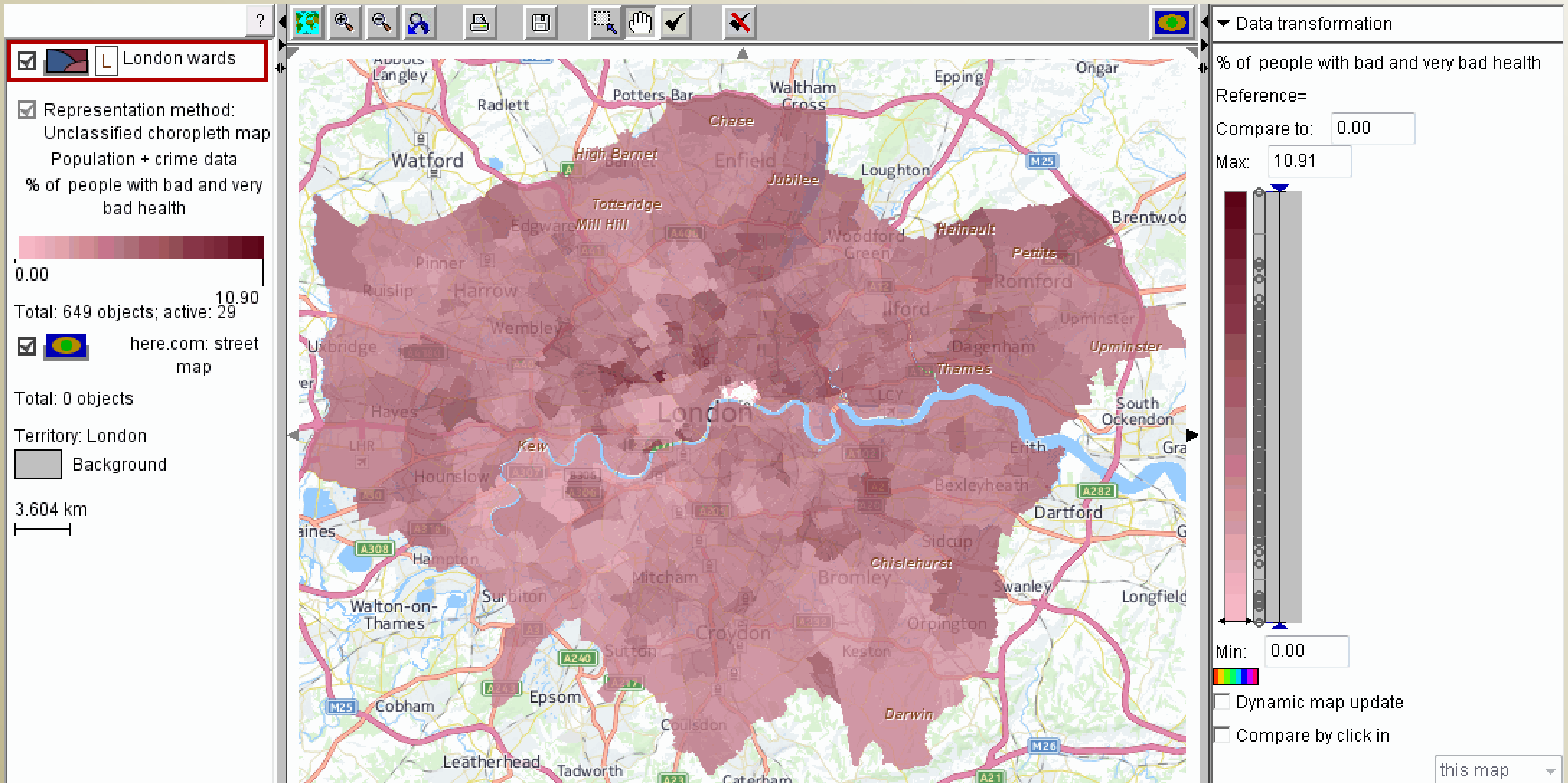


Each map requires one instance of perception, i.e., N of images = 2.

Limitation regarding the number of attributes

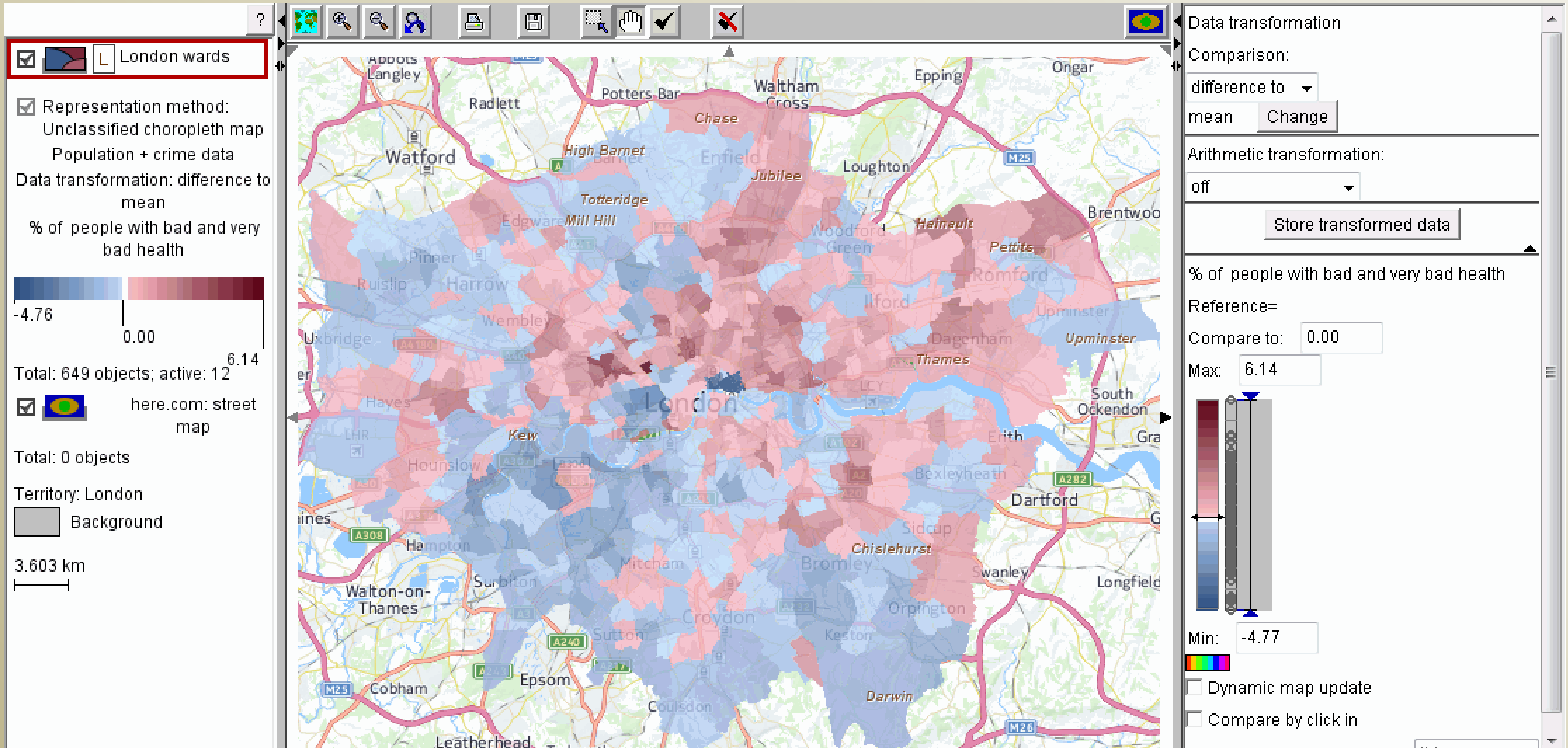
- Bertin: an image can be composed using at most two display dimensions and one retinal visual variable.
- In a map, the two display dimensions are used for representing spatial positions.
  - ⇒ Attribute values can only be represented by retinal variables: size, value (colour darkness), colour (hue), orientation, shape, or texture.
  - ⇒ Creating a single image is only possible for a single space-referenced attribute …
    - ➢ … or several attributes need to be somehow integrated into a single attribute.

Choropleth map *(a reminder)*



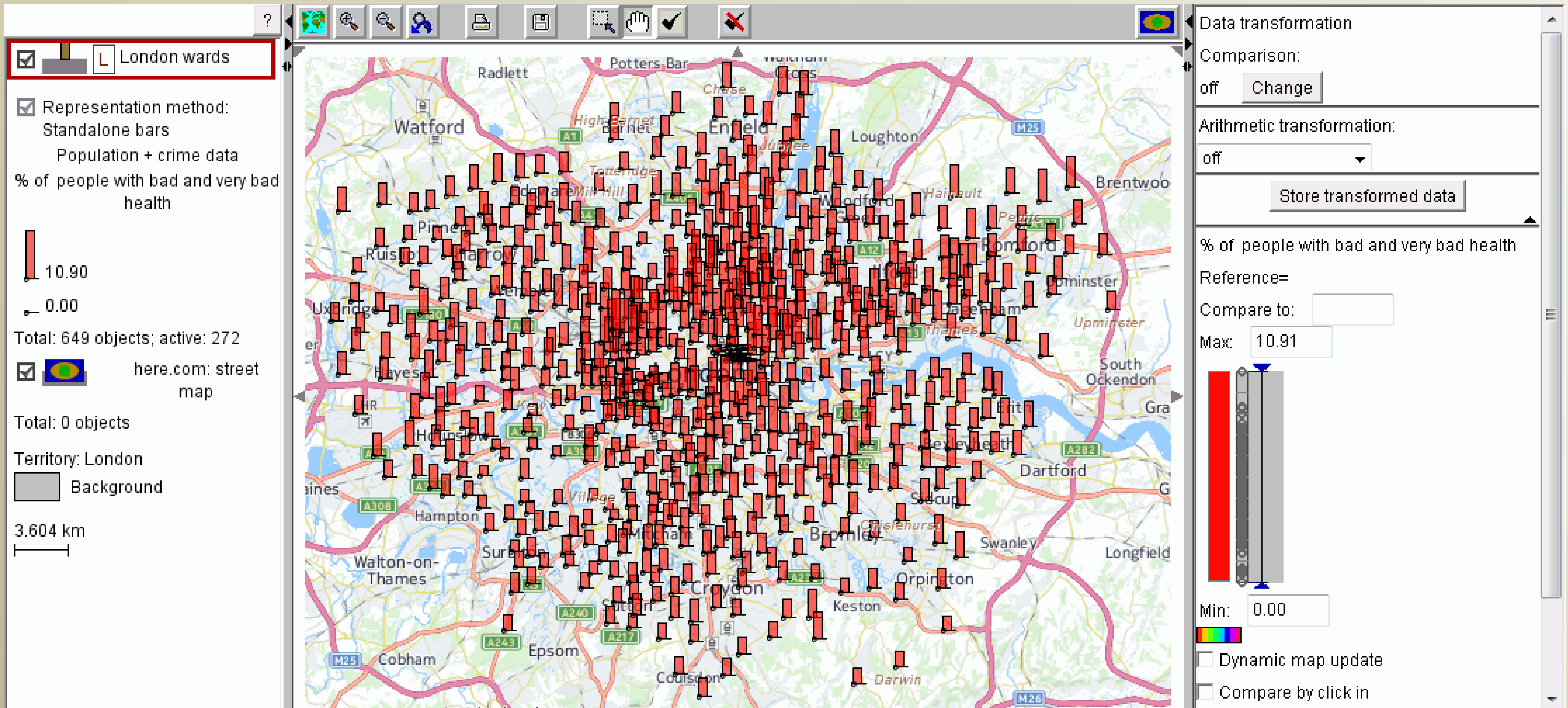
Values of a numeric attribute are represented by the visual variable ‘value’ (darkness). Darker shades correspond to higher attribute values. The shades are used for painting areas or objects on the map. The spatial distribution is perceived as a single image.

Choropleth map: diverging colour scale

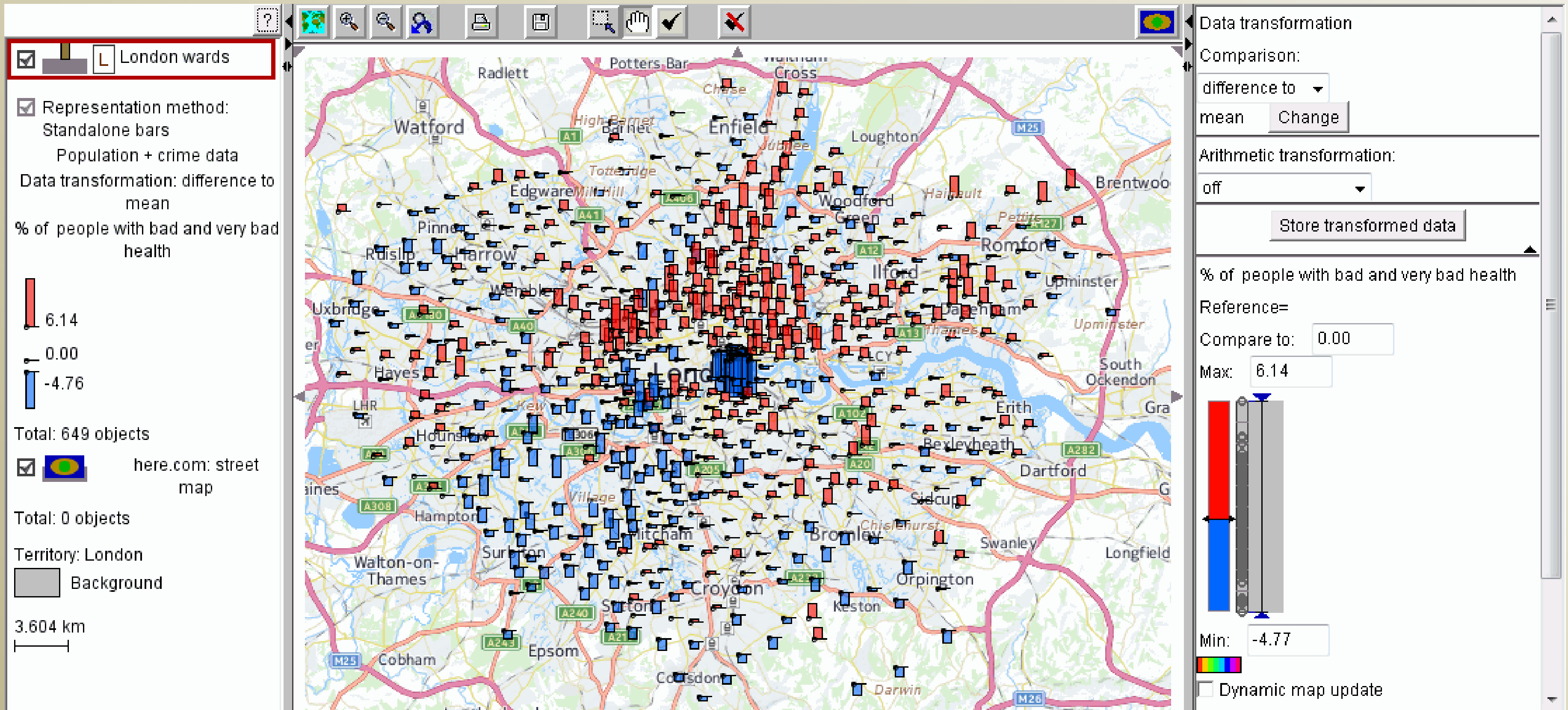


A diverging scale of colour shades uses two distinct colour hues for representing positive and negative values or positive and negative differences from a chosen central value, such as the overall mean. Darker shades correspond to larger differences. The map is perceived as a single image and exposes spatial clusters of high and low values.

Map with proportional symbols (bars)

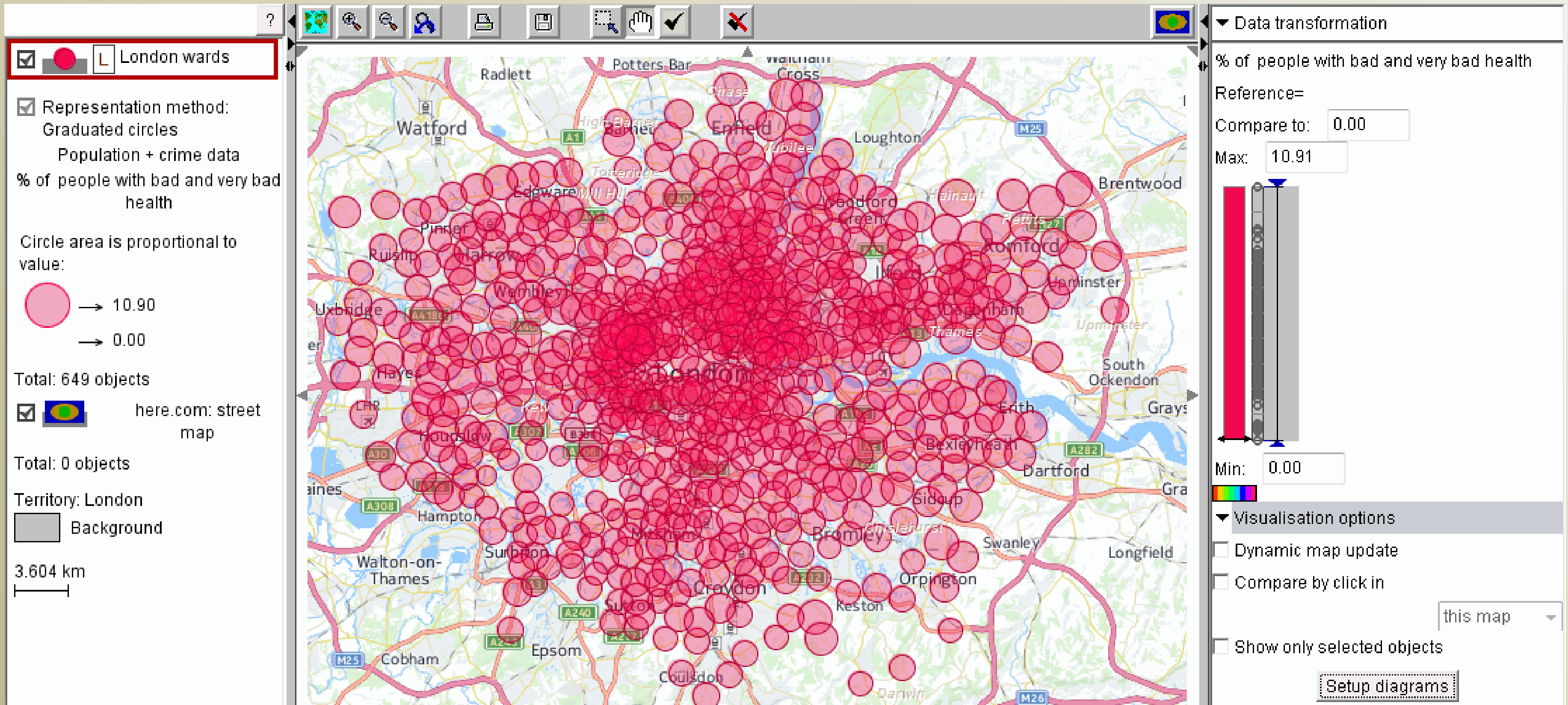


Map with proportional symbols (bars)



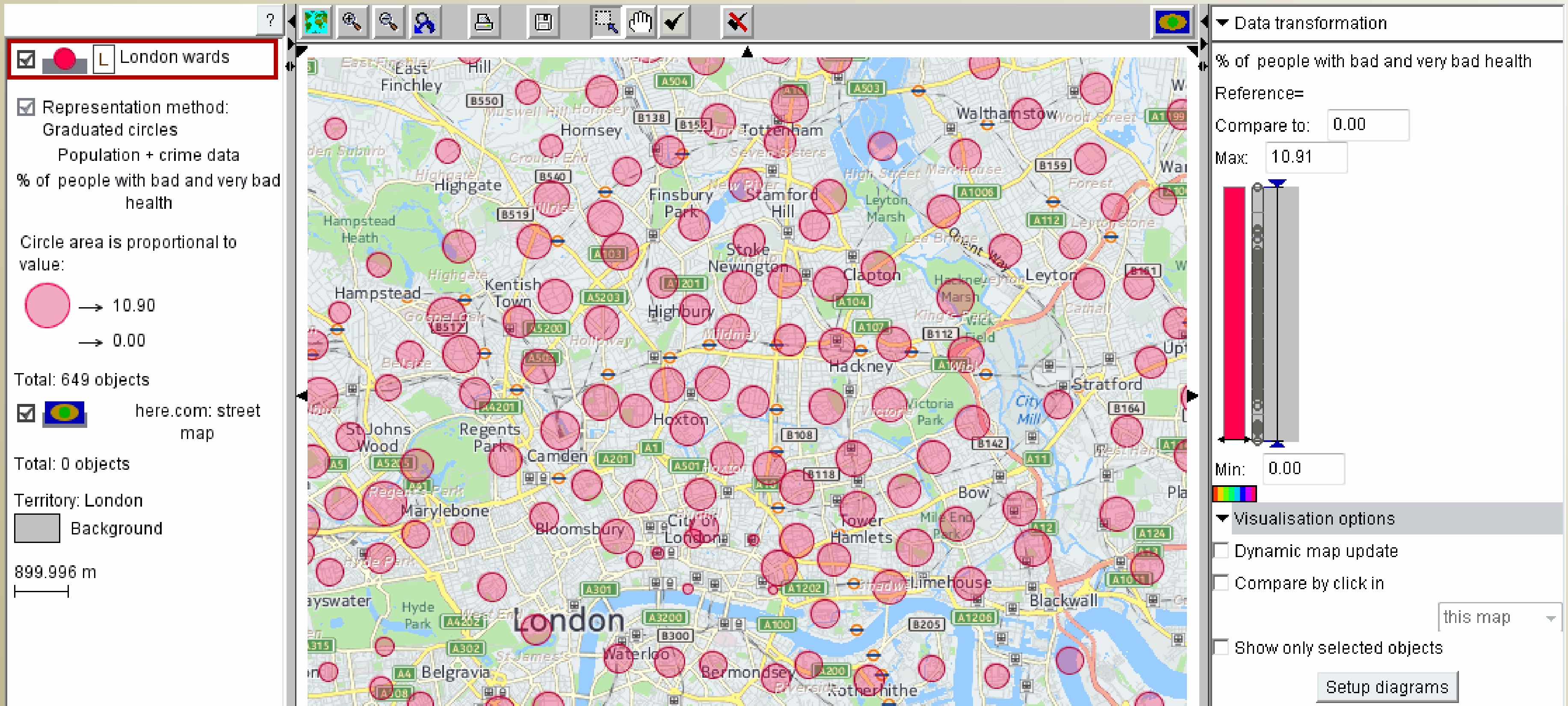
Diverging bars use orientation (upward and downward) and, complementarily, two distinct colour hues for showing positive and negative attribute values or differences from a selected central value, such as the overall mean.

Map with proportional symbols (circles)



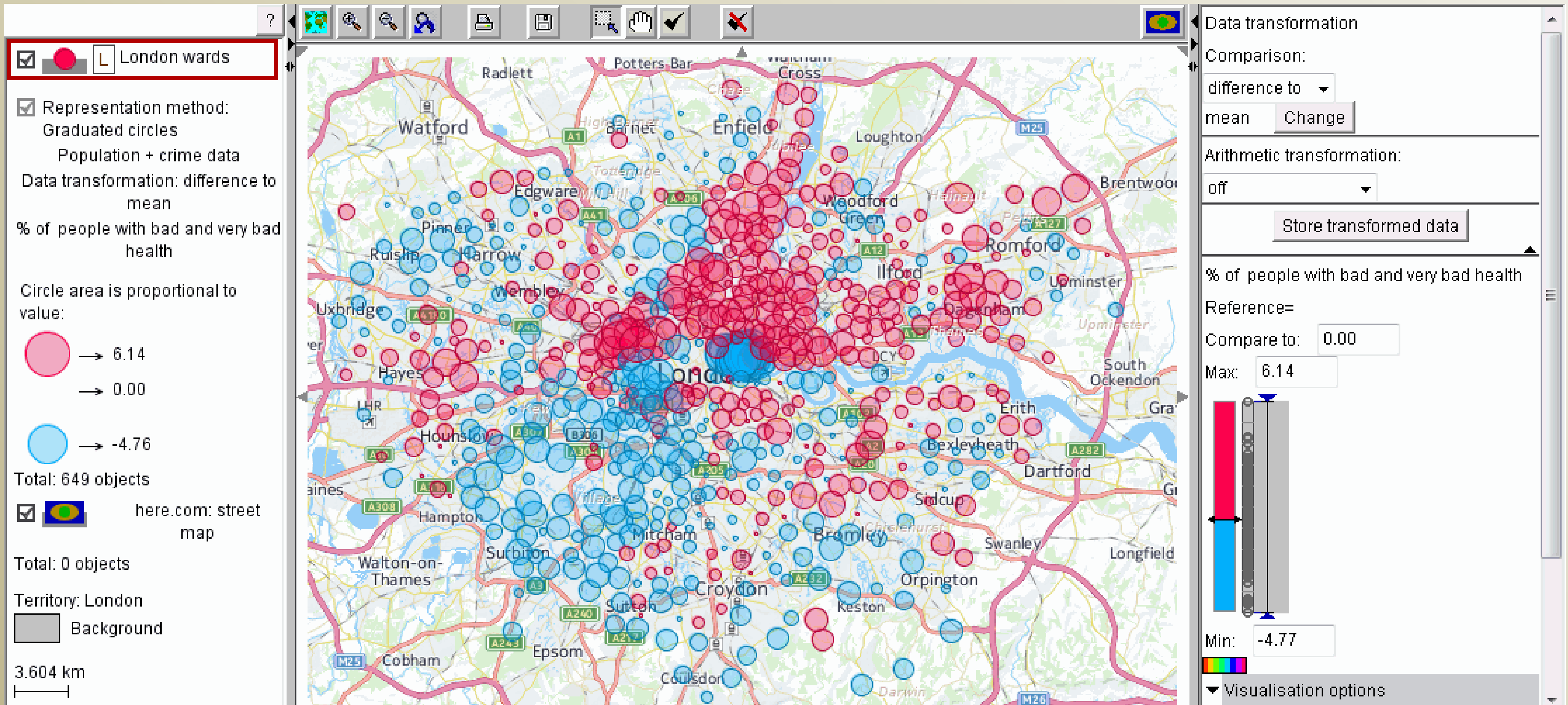
Problem: too much visual clutter; requires zooming for discerning the symbols.

Map with proportional circles



The symbols are now discernible, but the overall view is lost.

Map with proportional circles, diverging scale

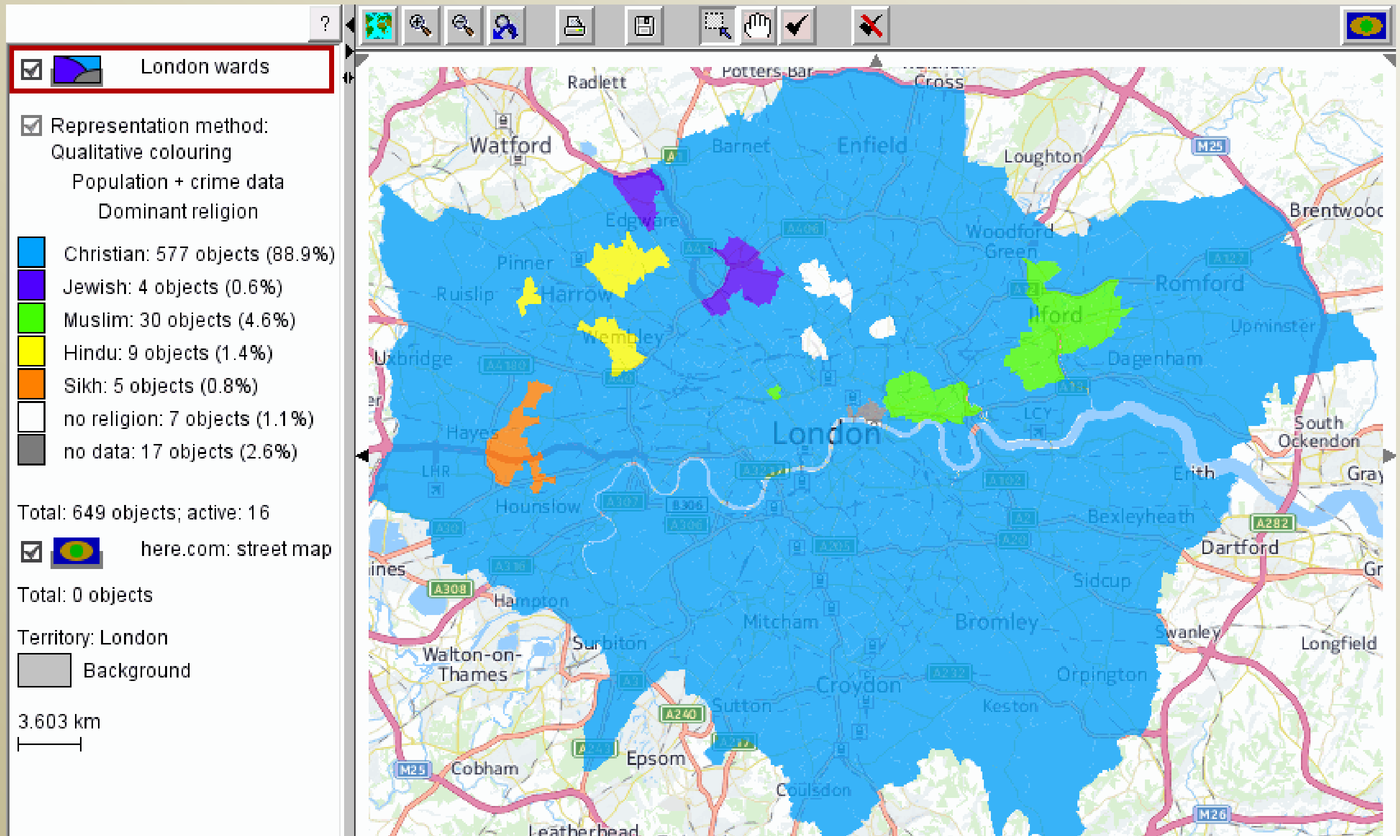


Similarly to bars, circles of two distinct colour hues can show positive and negative values or differences (the orientation cannot be used in this case).

Choropleth maps vs. proportional symbol maps

- Choropleth maps and maps with proportional symbols are used for representing the spatial distribution of a single numeric attribute.
- Advantages of a choropleth map:
  - provides a single image and thereby supports synoptic tasks;
  - is free from overlapping and clutter;
  - small differences in shades may be easier detectable than small differences in sizes.
- Maps with proportional symbols can also be efficient when they are not visually cluttered.
- Both choropleth maps and proportional symbol maps become more expressive and effective when diverging scales are applied.

Mapping of qualitative attribute values



Qualitative attribute values can best be represented by the retinal visual variable ‘colour’ (hue).

Synoptic tasks supported by maps alone

- Describe the spatial distribution of values *of a single attribute*.
- Compare several spatial distributions:
  - Different attributes over the same part of space (territory)
  - Same *single attribute* over different parts of space
  - Comparisons are supported by several juxtaposed maps (“small multiples” technique).
- For other tasks, maps need to be combined with other display types and interaction techniques.

Combining maps with other techniques

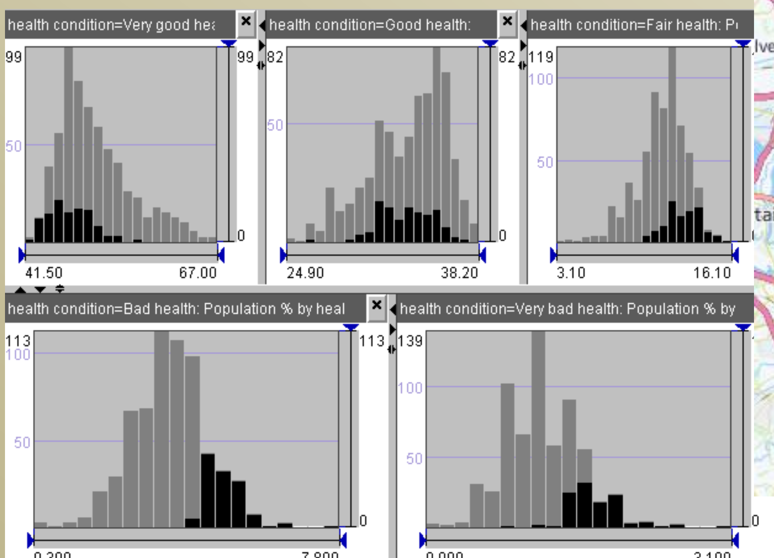
- All visualisation and interaction techniques supporting analysis of object-referenced data can be applied to space-referenced data.
- The specific tasks of spatial data analysis require these techniques to be used in combination with maps.
- Combining information from different views can be supported by interactive techniques for display linking:
  - Selection
  - Classification

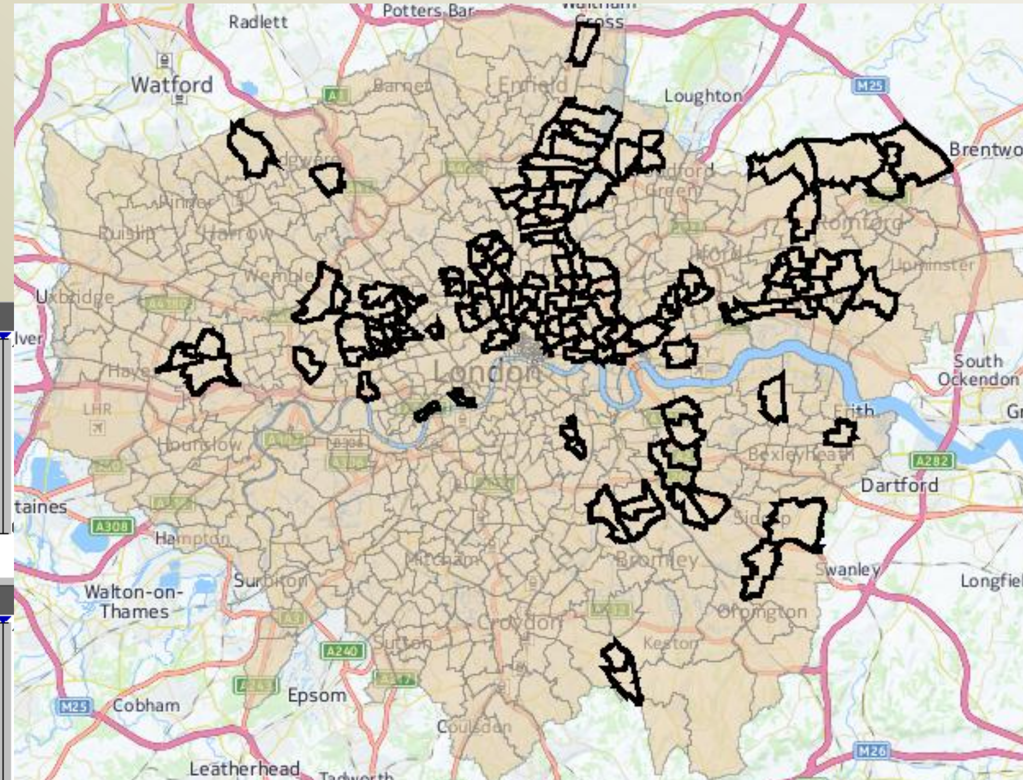
Two possible approaches to linking maps with attribute displays

- Attribute displays → map
  - Select or classify objects on a plot, diagram, or histogram and observe the spatial distribution of the selected objects or object classes on a map
    - Are there any spatial patterns?
- Map → attribute displays
  - Select or classify spatial objects on a map based on their spatial positions and observe on the attribute displays what values and value combinations occur in the selected part of space or are characteristic for the different parts of space (i.e., spatial classes of objects).
    - Are there any distinguishing value subsets or combinations?

Selection: attribute displays → map

Attribute displays (frequency histograms): we have selected districts with relatively high percentages of people having bad or very bad health.

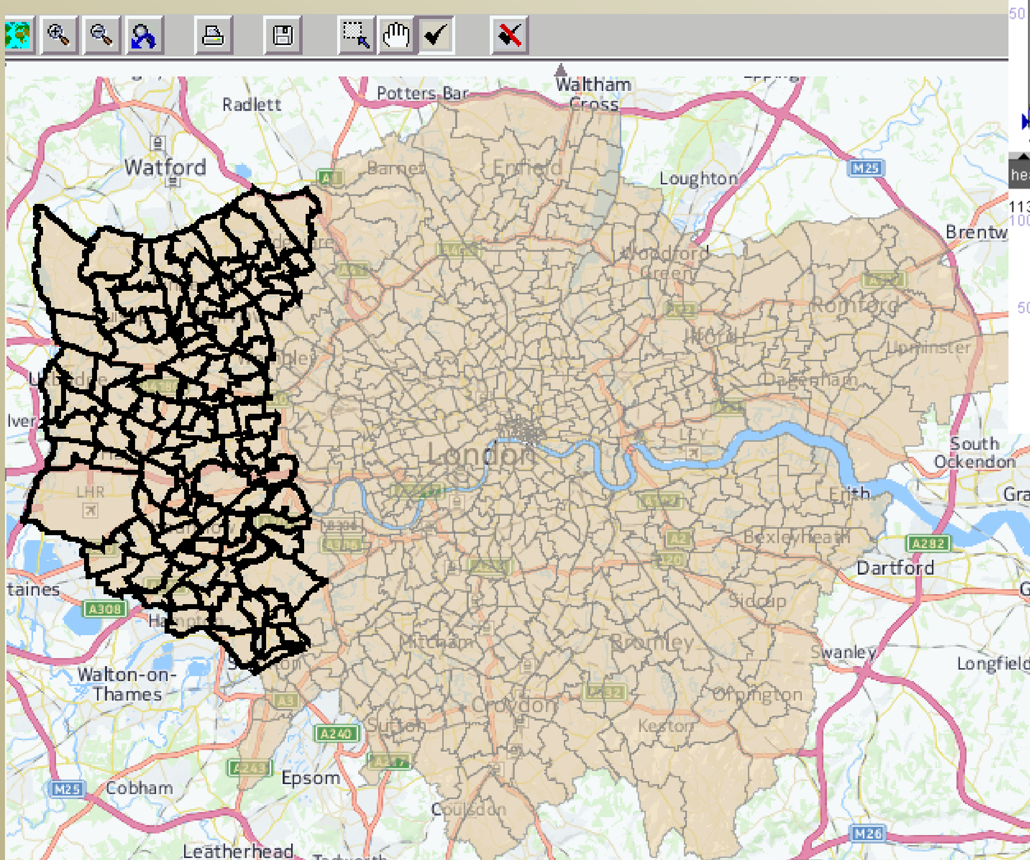


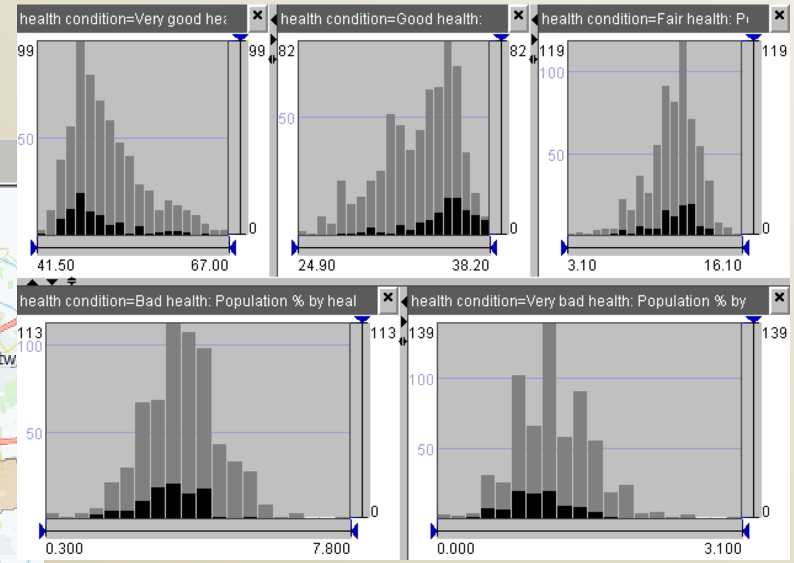


Map: shows us that the districts where many people have bad or very bad health are grouped in spatial clusters.

Selection: map → attribute displays

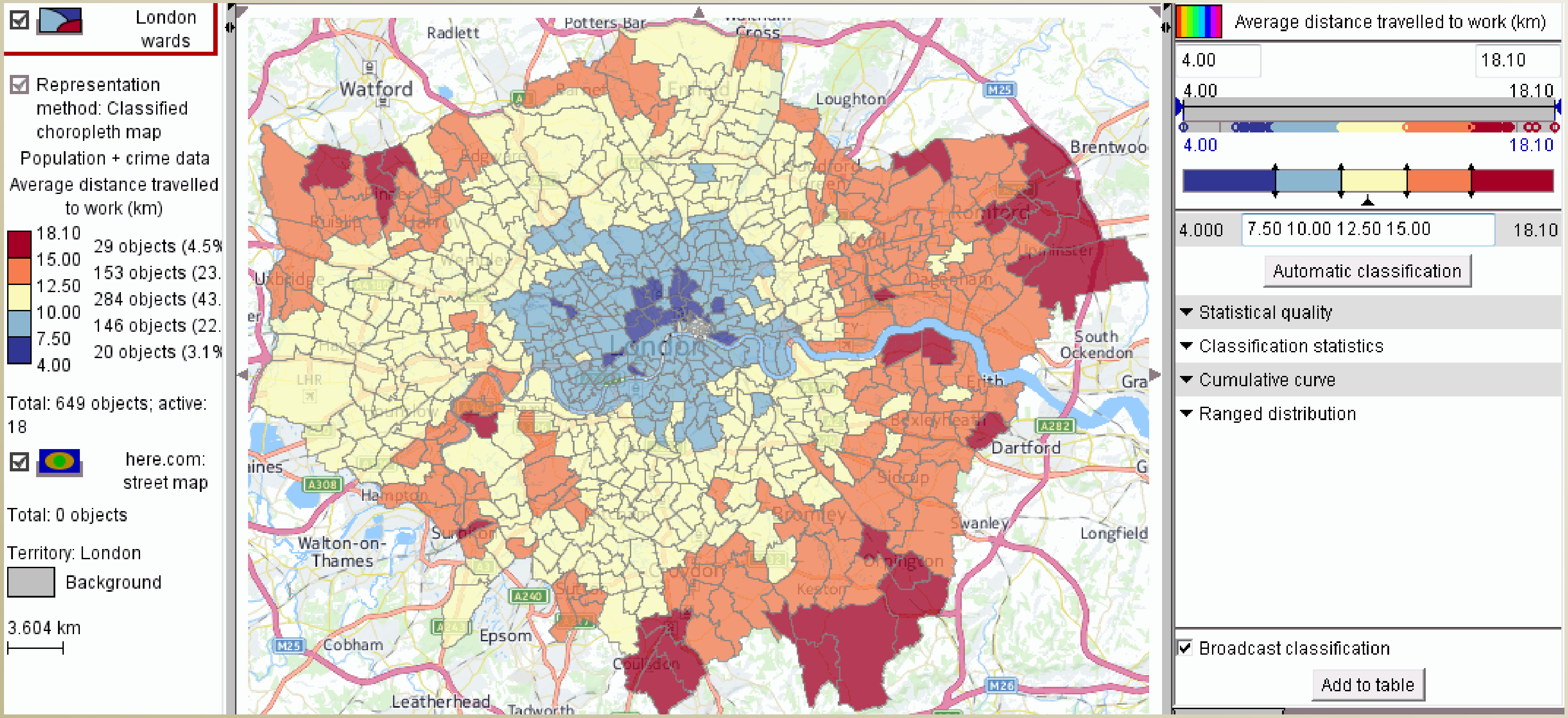
Map: we have selected the districts on the west of London.



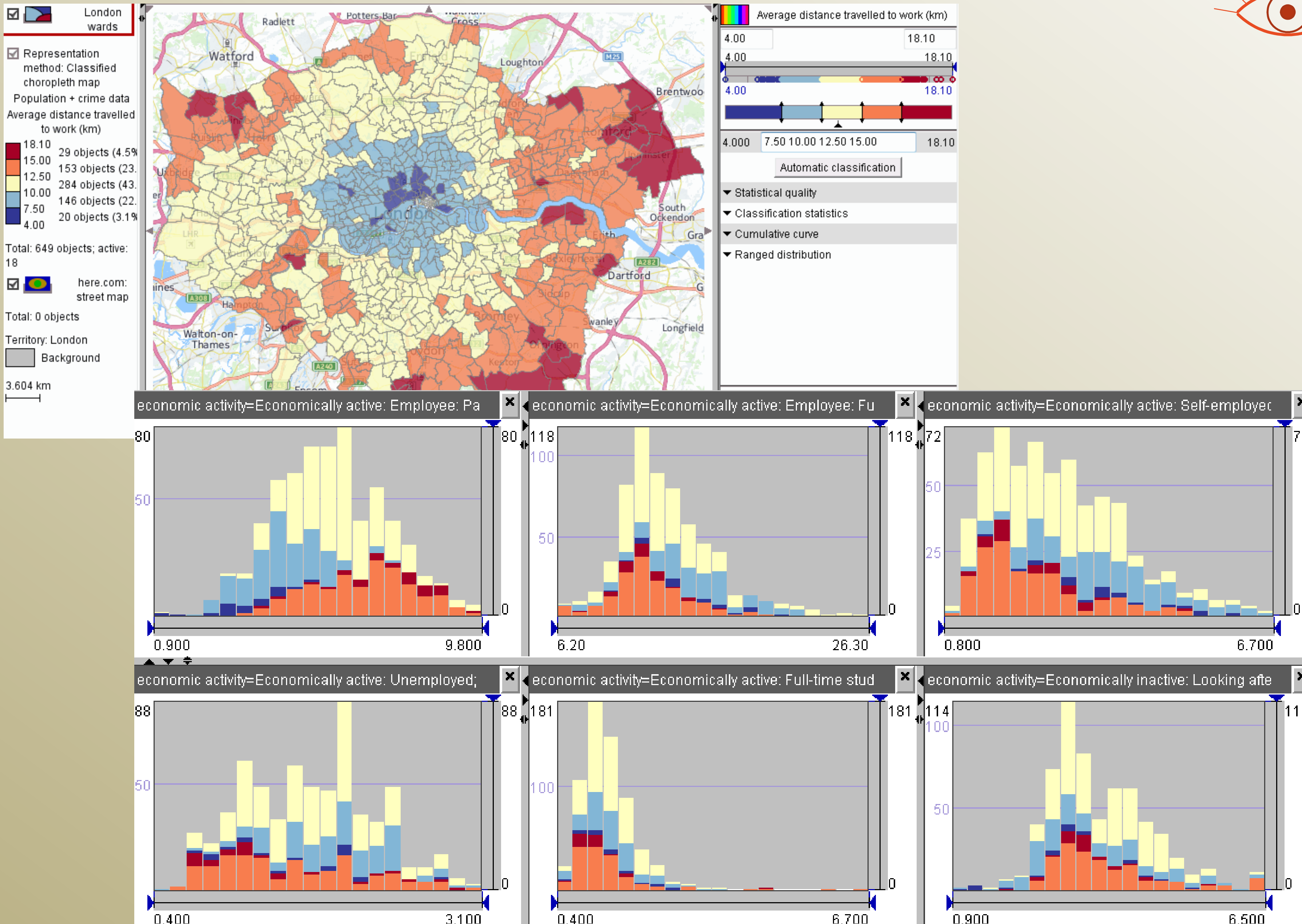


Attribute displays show us that the western districts tend to relatively high proportions of people having good health, average proportions of those having fair health, and lower proportions of people with bad or very bad health.

Attribute-based classification → map



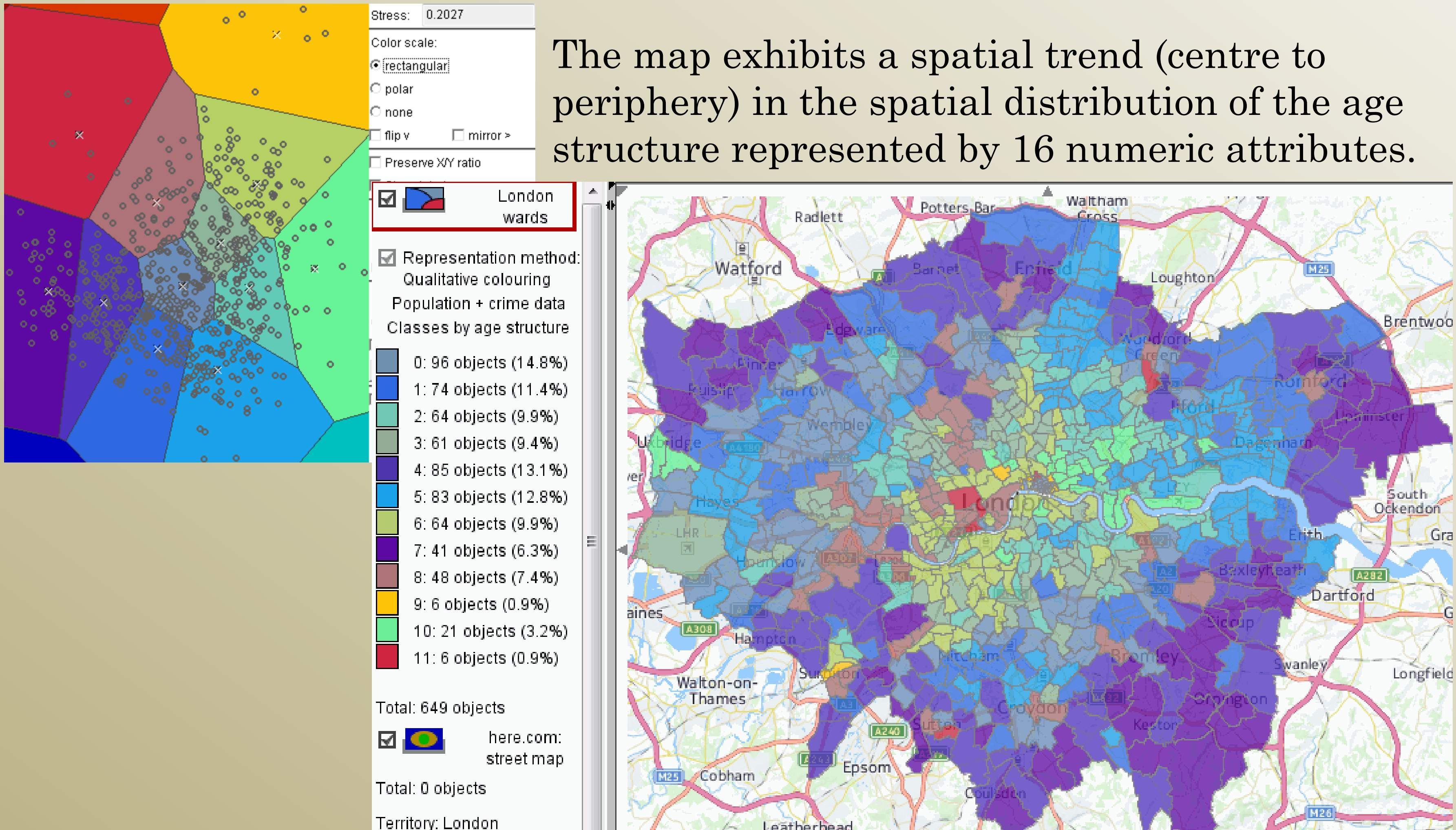
We see that the spatial distribution of the attribute "average distance travelled to work" has a spatial trend of increasing values from the centre to the periphery. This could also be seen on an unclassified choropleth map; however, classes can be propagated to other displays.



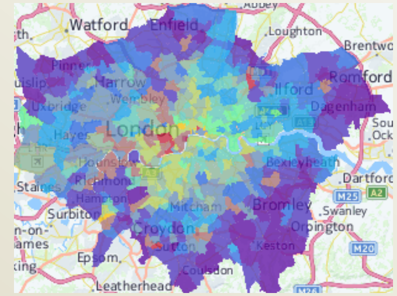
% of female population by economic activity

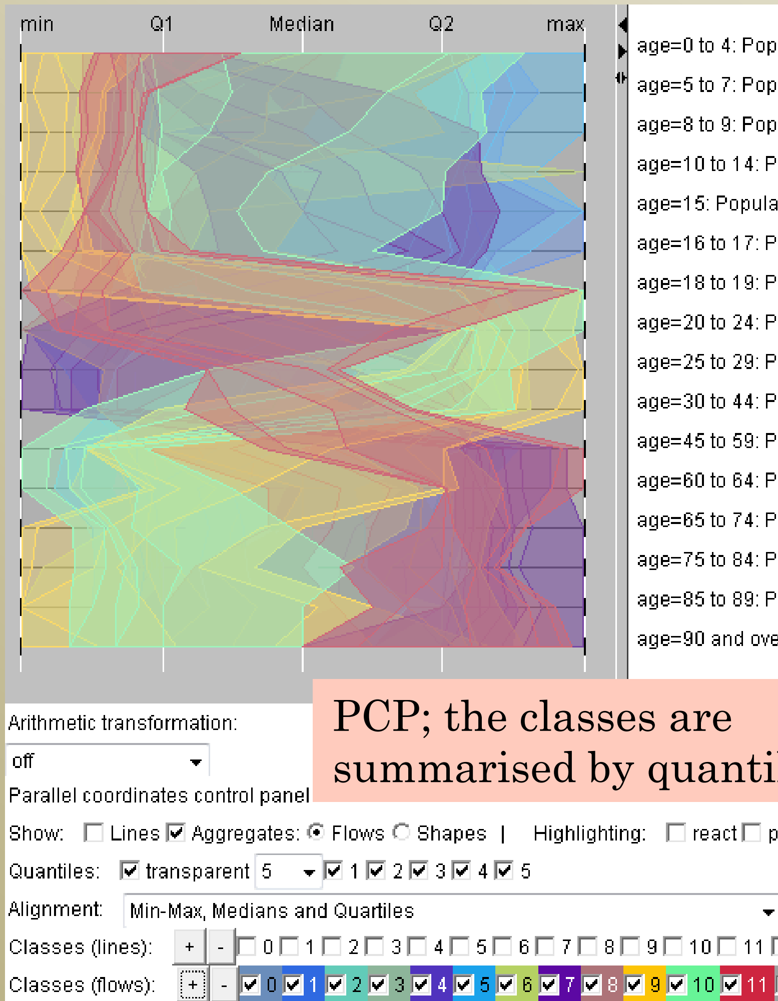
Classification based on multiple attributes (through projection) → Map

The map exhibits a spatial trend (centre to periphery) in the spatial distribution of the age structure represented by 16 numeric attributes.

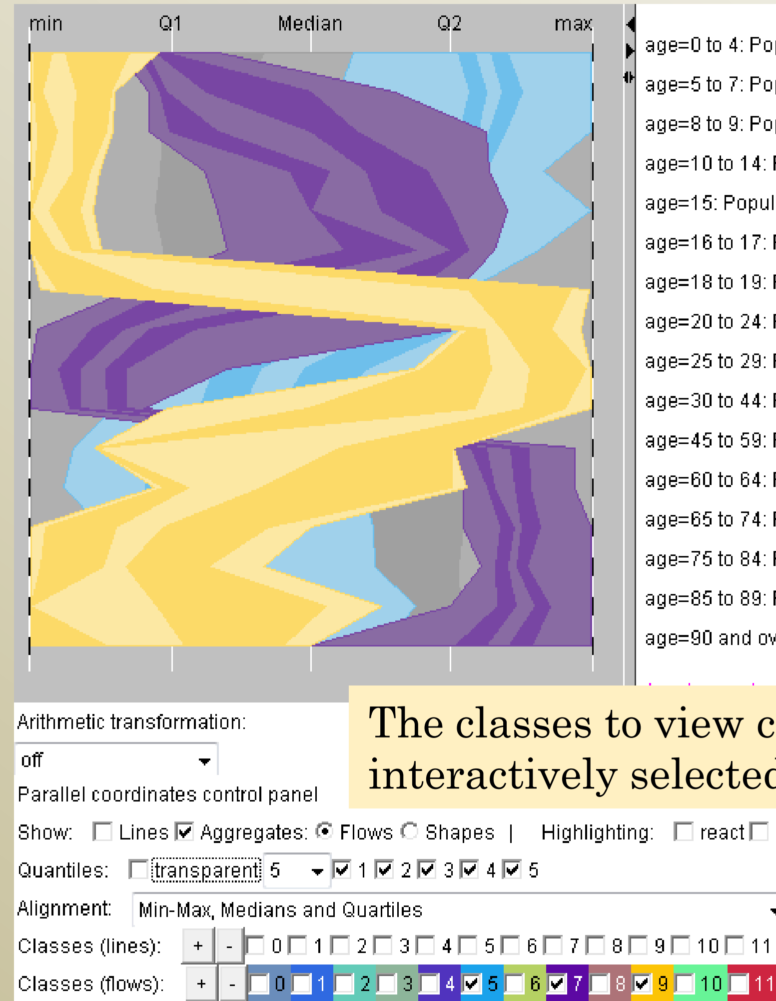


Interpreting the colours



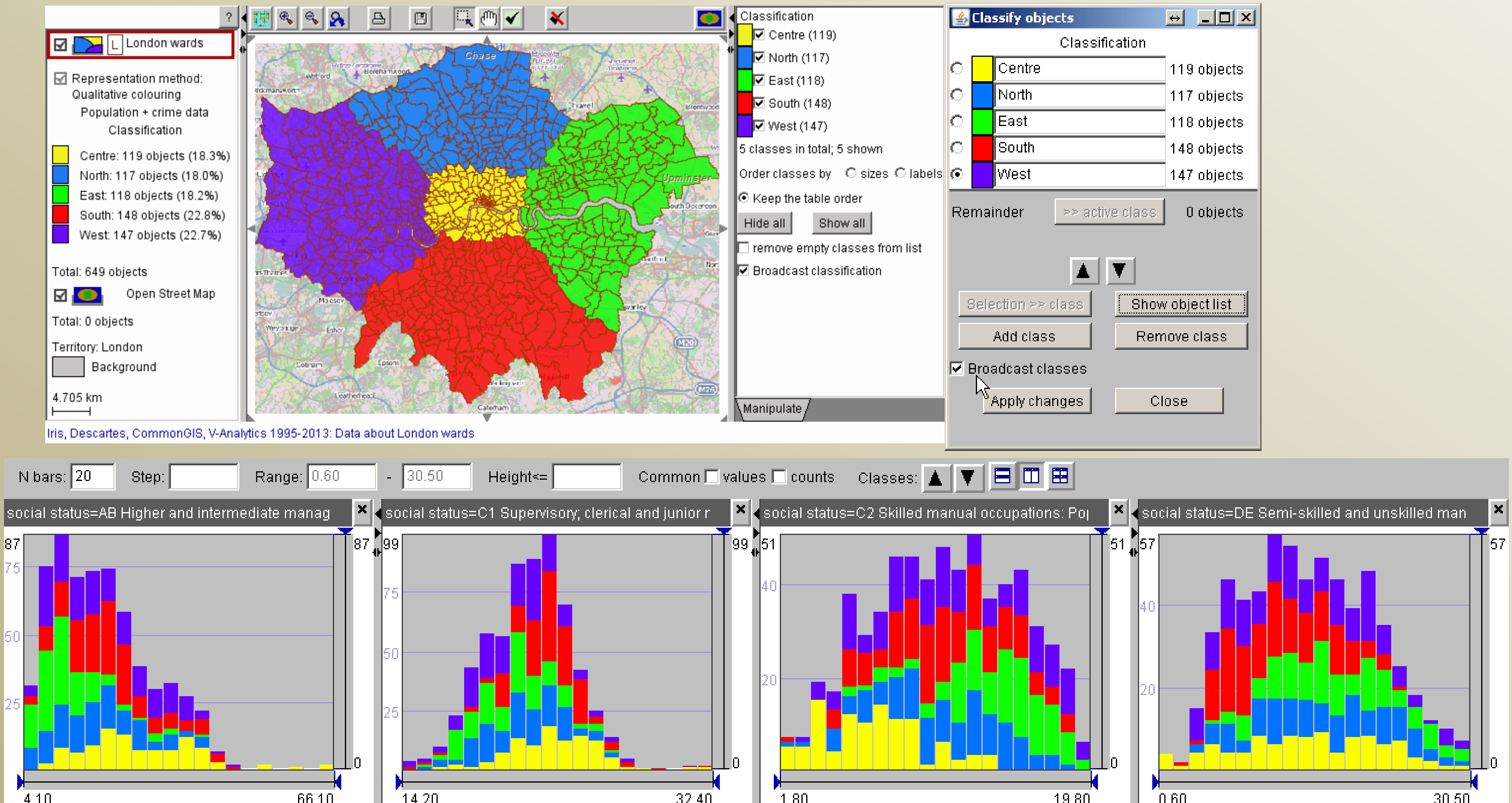


PCP; the classes are summarised by quantiles



The classes to view can be interactively selected.

Space-based interactive classification on a map → attribute displays



Techniques for analysing space-referenced data

- Describe the spatial distribution of a single attribute; compare spatial distributions of singular attributes
  - ➢ Maps with colouring
    - Retinal variable ‘value’ (darkness): numeric or ordinal attributes
    - Retinal variable ‘colour’ (hue): qualitative attributes
  - Several juxtaposed maps (“small multiples” technique) support comparisons of spatial distributions
- Describe the spatial distribution of value combinations of multiple attributes; find parts of space with particular or outstanding value combinations; relate attribute behaviours
  - ➢ Maps + attribute displays linked by interactive techniques
    - Selection
    - Classification

Questions?

Analysis of space-referenced data

Analysis of spatial events

Objects having spatial locations and existence times

Spatial event data

- Spatial event data structure:
  - 1 referrer: set of objects
  - 2 mandatory attributes: *spatial location* + *time of existence*
  - any other attributes, further called *thematic attributes*
- Spatial location and existence time are *attributes* of the objects

  ⇒The main synoptic tasks address the behaviour (distribution) of the spatial locations and existence times over the set of objects.
- However, space and time can also be considered as independently existing *containers* of the objects.

  ⇒The main synoptic tasks may be re-formulated to address the distribution of the objects over the space and time, i.e., the *spatio-temporal distribution* of objects

Space and time as object containers

- Considering space and time as containers of objects is quite intuitive.
  - We can easily imagine space and time without objects. It is difficult to do the same for other attributes (e.g., size).
- In visualisation, it is typical to represent space and time by display dimensions and objects by marks located within the display space.
  - I.e., the display conveys the idea of the objects being contained in space and/or time.
  - This representation is usual for people and therefore easily understandable.

➢ We take this *absolute view* of time and space as object containers and formulate the synoptic tasks accordingly.

Major classes of synoptic tasks

1) Attributes: space and time

- Describe the behaviour of one or more attributes →
  Describe the *spatio-temporal distribution* of the events
- Find subsets of references where attributes have particular behaviours →
  Find subset of objects forming particular spatio-temporal patterns, such as spatio-temporal concentration (cluster), alignment, …
- Compare two or more behaviours →
  Compare spatio-temporal distributions of event (sub)sets
  - Different events (by thematic attributes) in the same space and time
  - Events from different time periods in the same space
  - Events in different parts of space
- Relate behaviours of two or more attributes →
  Relate the spatial distribution to time and vice versa

Major classes of synoptic tasks

2) Attributes: thematic attributes

- Spatial events are *objects*.

  ⇒Spatial event data are a special case of *object-referenced data*.

  ⇒All synoptic tasks defined for object-referenced data apply, in particular, to spatial event data.

  - I.e., tasks addressing the distribution of the attribute values over the set of objects.

  ⇒All techniques supporting the analysis of object-referenced data can be used for spatial event data.

- Spatial events are *spatial objects*.

  ⇒Spatial event data are a special case of *space-referenced data*.

  ⇒All synoptic tasks defined for space-referenced data apply, in particular, to spatial event data.

  - I.e., tasks addressing the spatial distribution of the attribute values.

  ⇒All techniques supporting the analysis of space-referenced data can be used for spatial event data.

Major classes of synoptic tasks

2) Attributes: thematic attributes *(continued)*

- Additionally, spatial events are *spatio-temporal objects*.

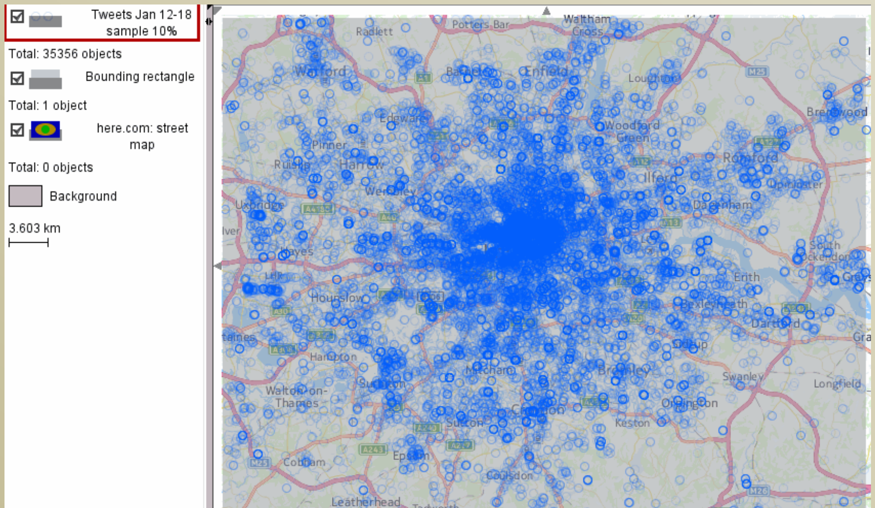
  ⇒Values of the thematic attributes refer to spatial locations and time moments and intervals.

  ⇒It is appropriate to consider the behaviour of the attribute values over space and time, i.e., the *spatio-temporal distribution of the attribute values*.

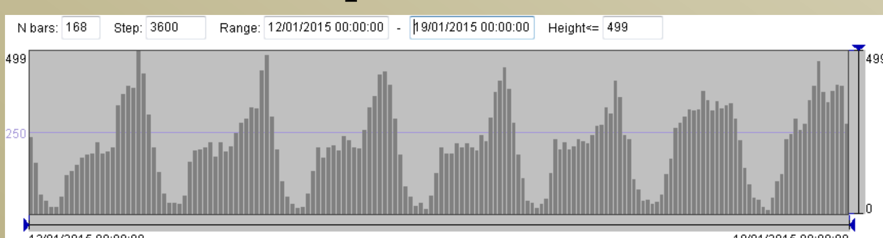
  - The tasks are formulated similarly to the tasks addressing the spatio-temporal distribution of the events.

Visual displays of spatial events

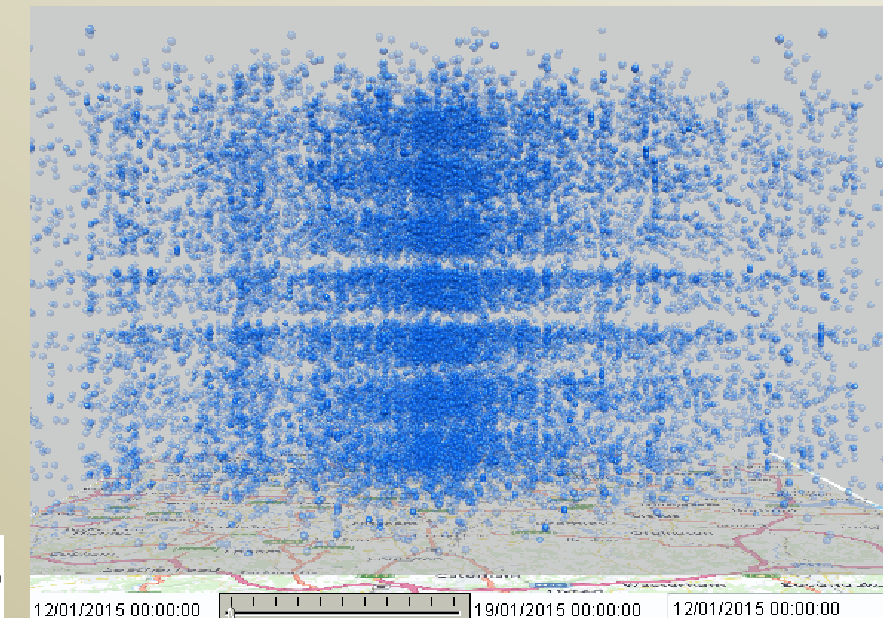
Map: shows the spatial distribution



Temporal display (e.g., frequency histogram of the occurrence times): shows the temporal distribution

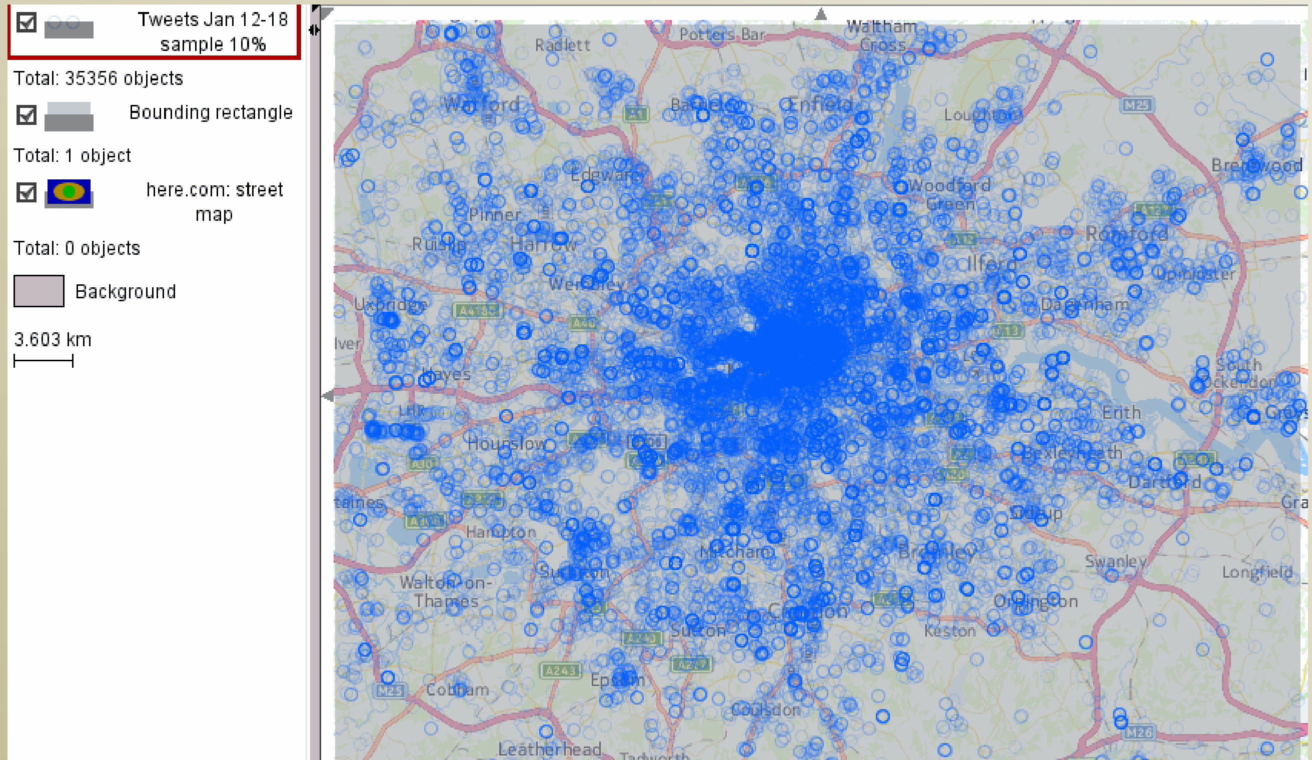


Space-time cube*: represents 2D space + time as a single 3D continuum, in which the objects (spatial events) are positioned.



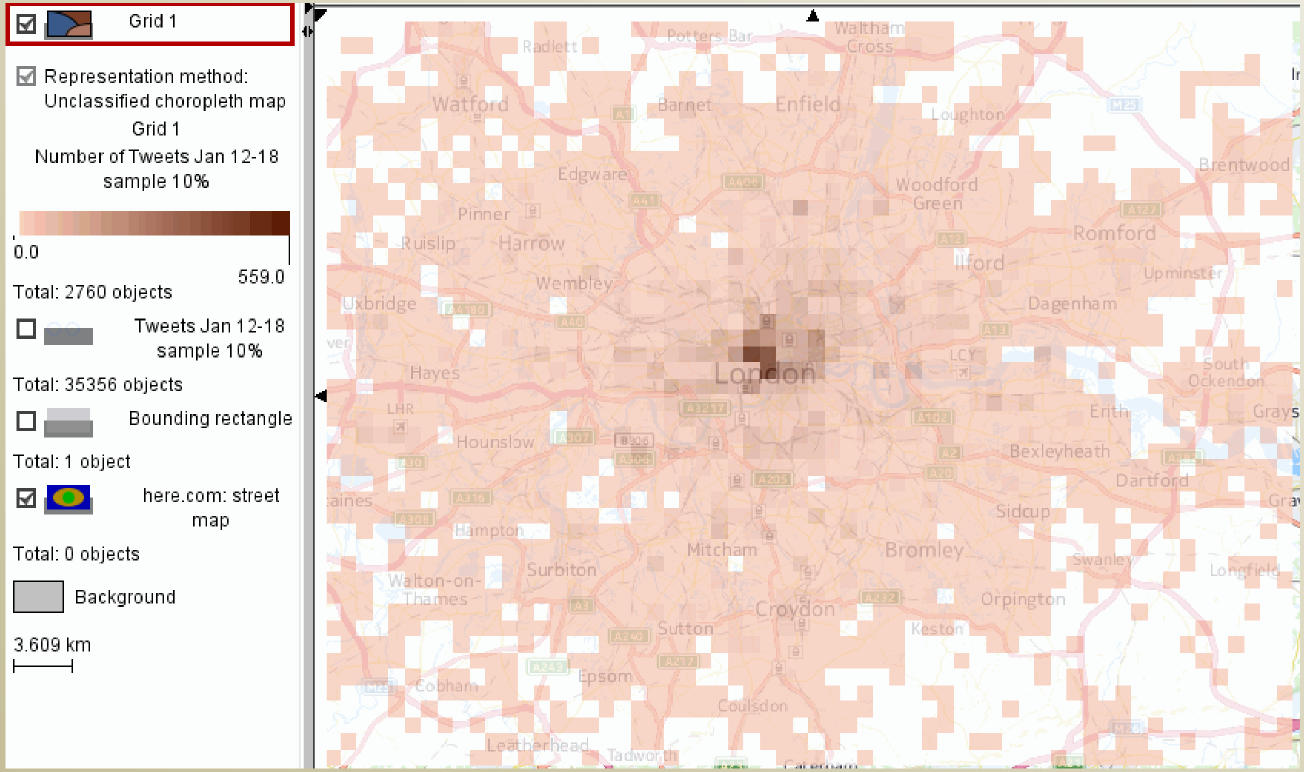
Hägerstrand. T. (1970).
"What about people in regional science?"
Papers of the Regional Science Association; 24:7-21.

Spatial distribution of events



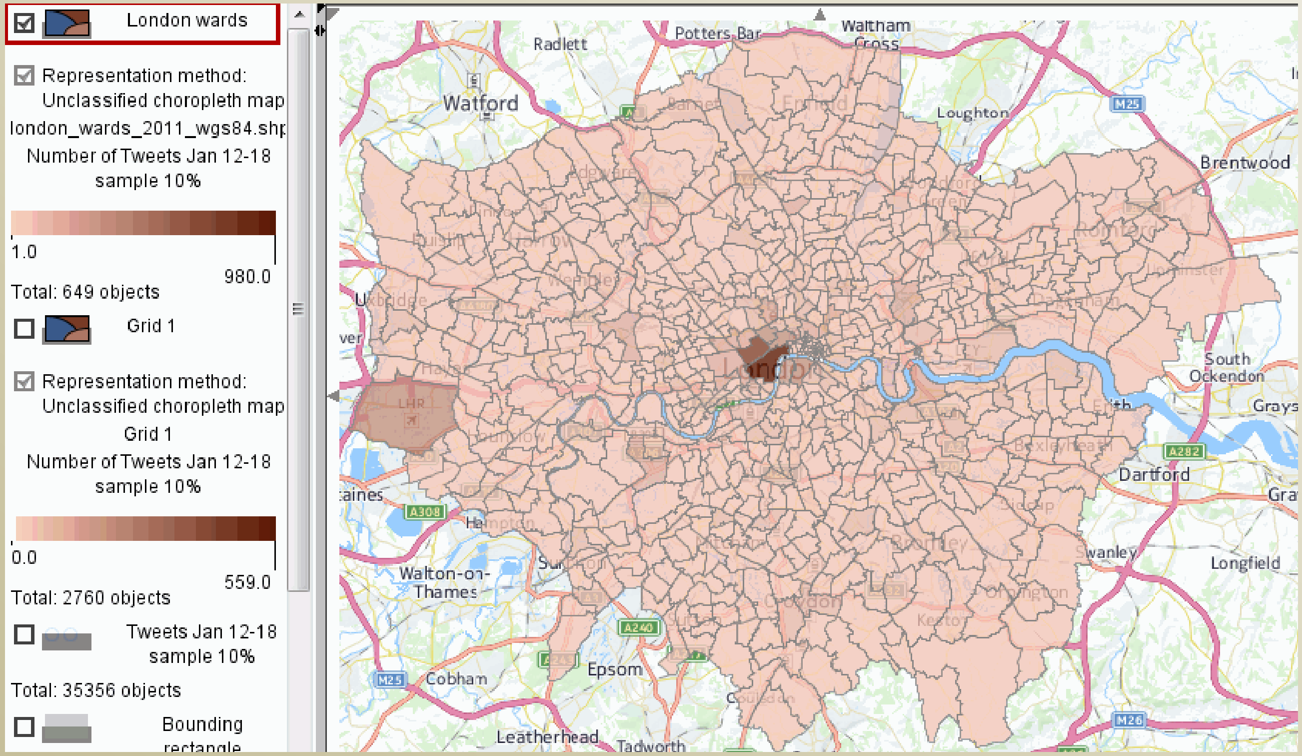
Point events (i.e., having no spatial extents) are typically represented on a map by dot symbols (small circles). Very often the dots overlap ⇒ semi-transparent drawing is recommended. The degree of transparency is adjusted to the dot density. Such a way of rendering helps us to observe the variation of the event density over the territory and detect spatial clusters of events.

Spatial aggregation by a regular grid



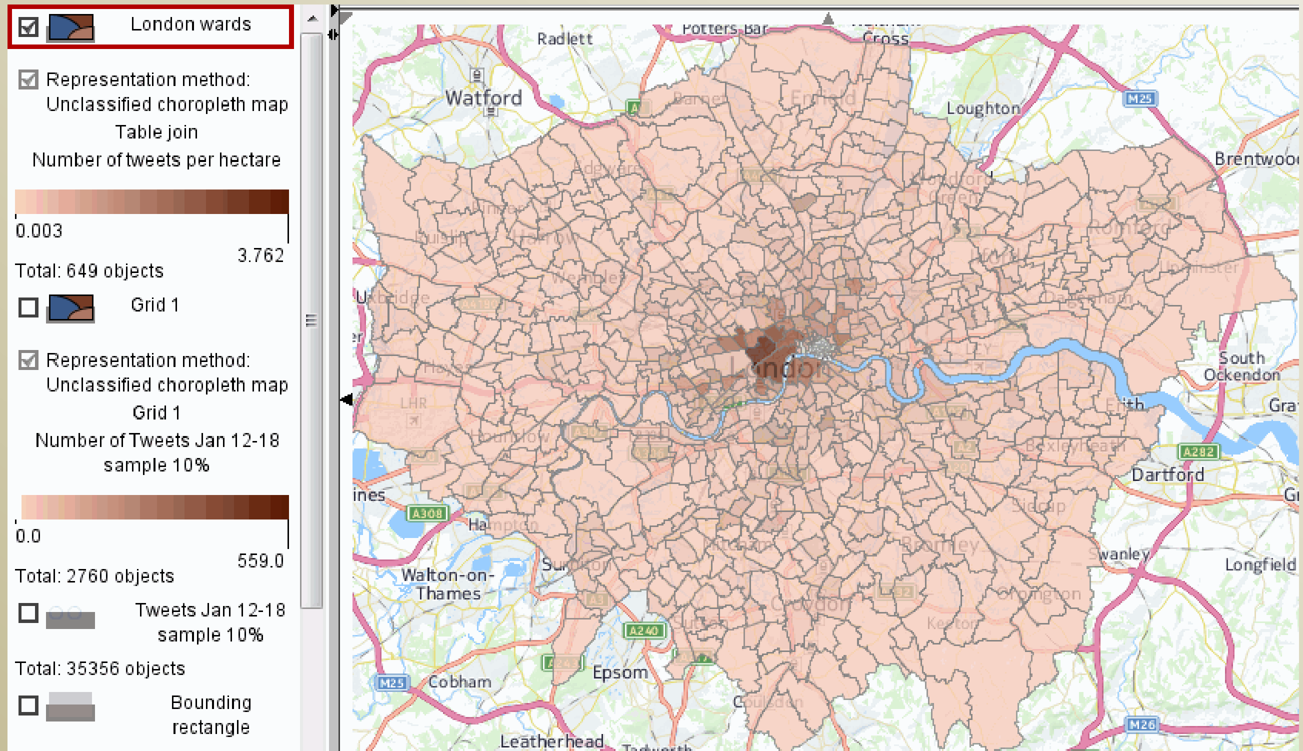
Investigation of the spatial variation of the event density can be supported by spatial aggregation of the events, e.g., by cells of a regular grid (here: 1×1 km). For each cell, the number of events is counted; additionally, thematic attributes can be summarized by computing the mean, mode, median, etc. Resulting data type: space-referenced attributes (considered before).

Spatial aggregation by arbitrary areas



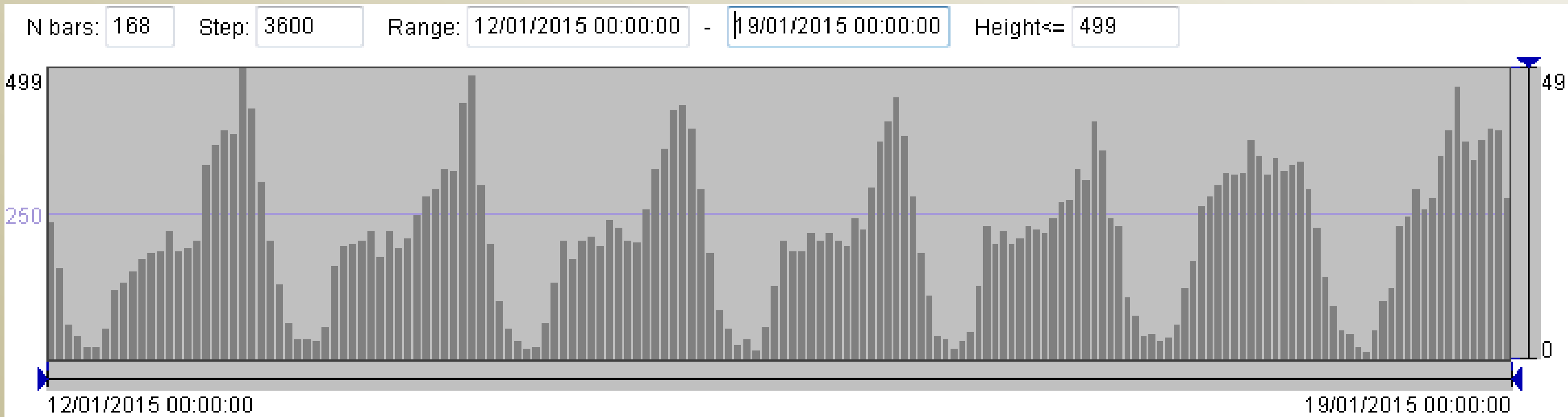
Events can also be aggregated by arbitrary areas, such as administrative districts (e.g., to relate event numbers and/or attributes to other characteristics of the areas). However, arbitrary areas may significantly differ in sizes, which distorts the perception of the event density. ⇒ Choropleth maps, like this one, should **not** be used for representing counts by areas.

Aggregation: absolute counts transformed to relative

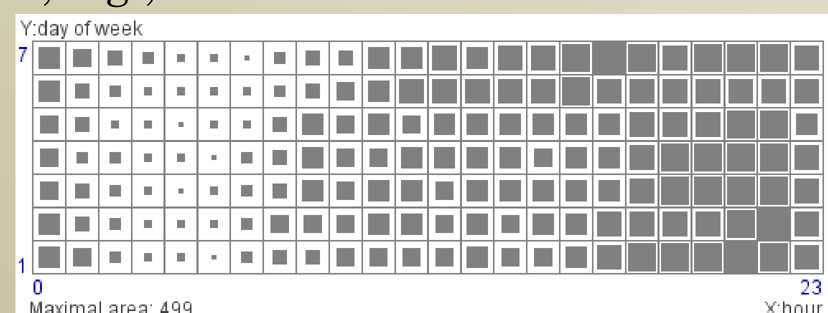


A correct image of the variation of the event density can be obtained by transforming the absolute event counts to relative w.r.t. the areas of the districts. When the events are related to lives and activities of people (e.g., tweets), it can also make sense to compute the relative counts w.r.t. the population of the districts.

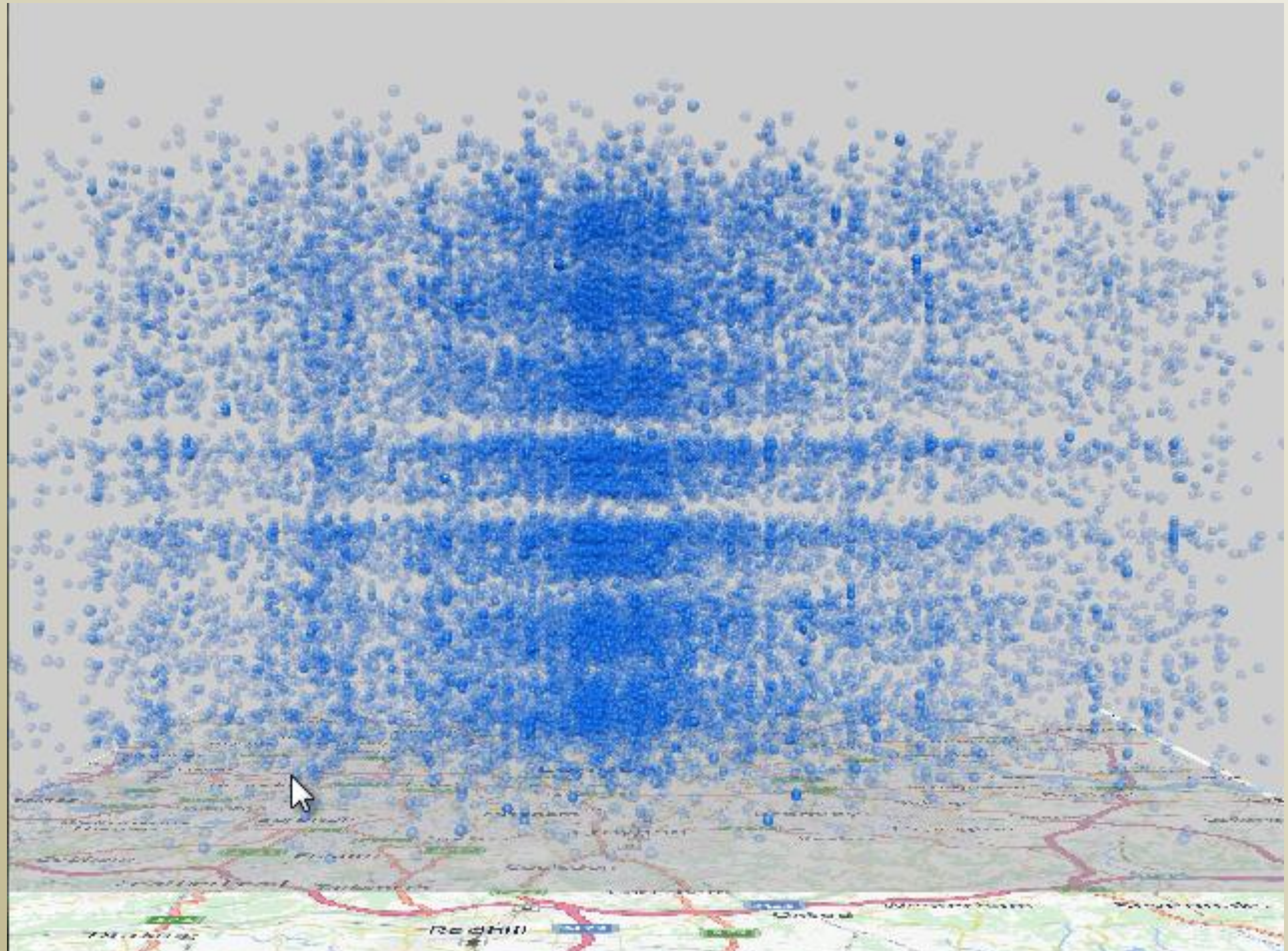
Temporal distribution of events



A frequency histogram of the event occurrence times shows the overall temporal distribution of the events and exhibits temporal patterns such as temporal trends and periodicity. A periodic temporal distribution can be additionally explored using a 2D histogram, e.g., with the dimensions corresponding to the days of the week and hours of the day.

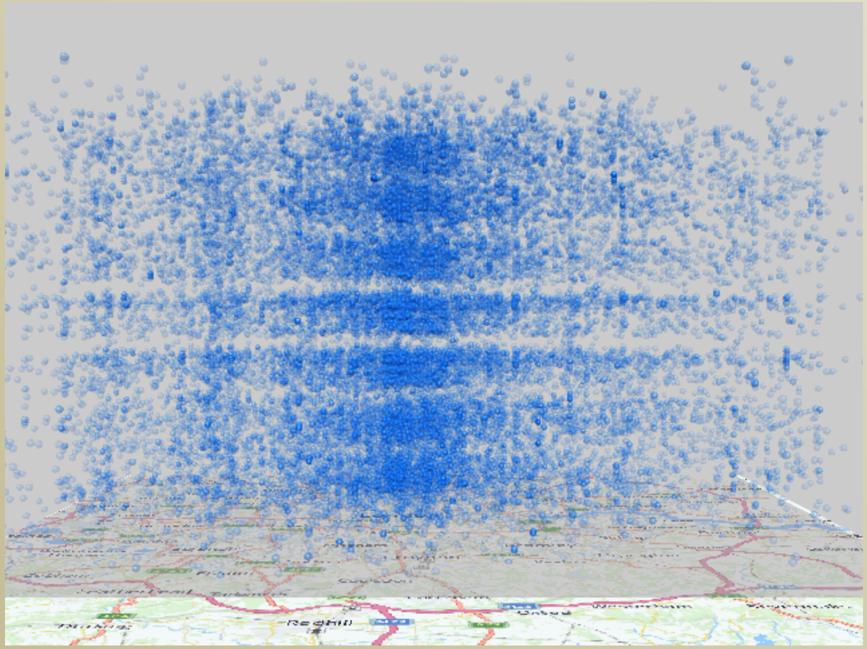


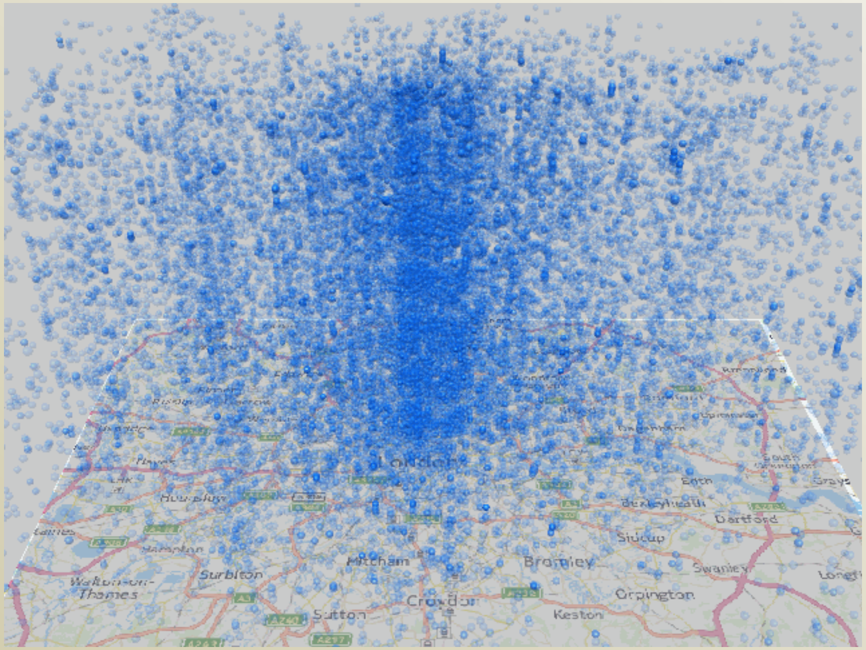
Spatio-temporal distribution of events



A static perspective view of a 3D representation is insufficient for observing and exploring the content. Interactive operations for changing the perspective (rotation, shifting, and tilting) are necessary.

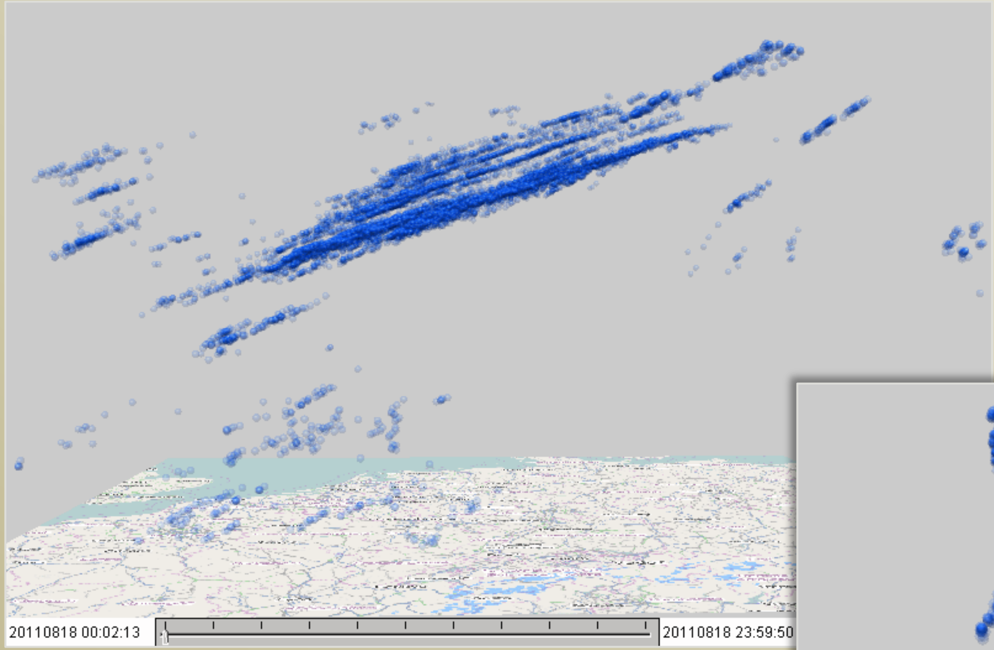
Spatio-temporal distribution of events



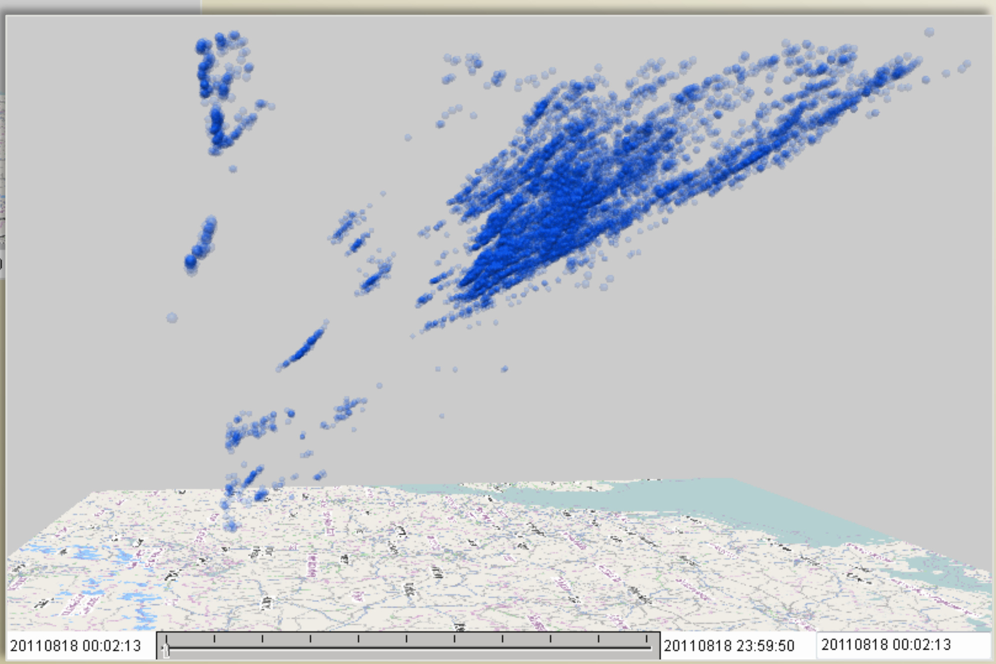


A space-time cube (STC) can exhibit several patterns of spatio-temporal distribution (types of spatio-temporal behaviour) appearing as horizontal “layers” and/or “gaps” (= periods of high and low event density), vertical “columns” (= high event density in some area for a long time), and “lumps” (= spatio-temporal clusters, i.e., groups of events that occurred closely in space and time).

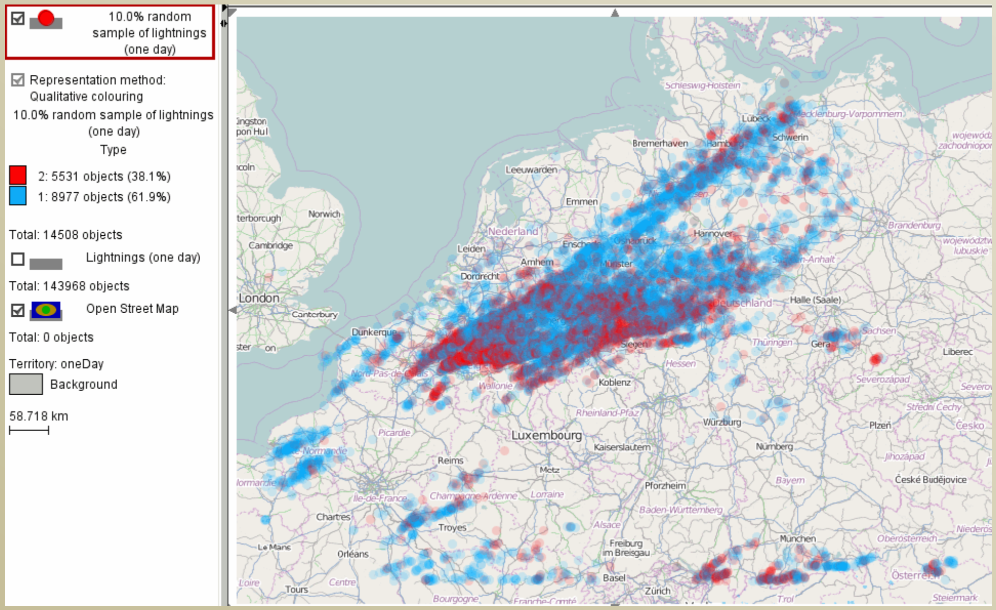
Spatio-temporal distribution of events

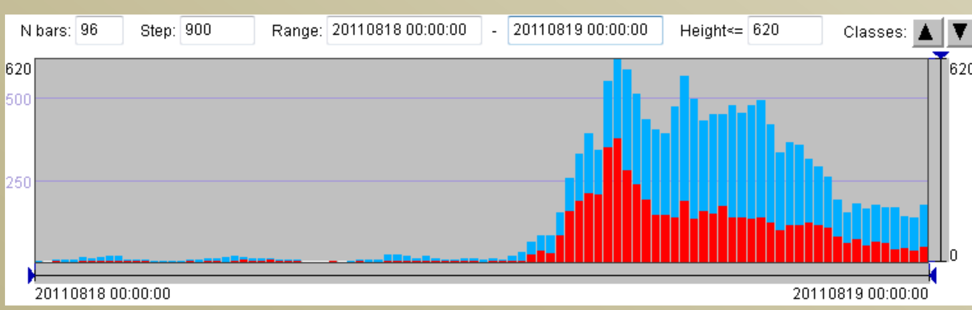


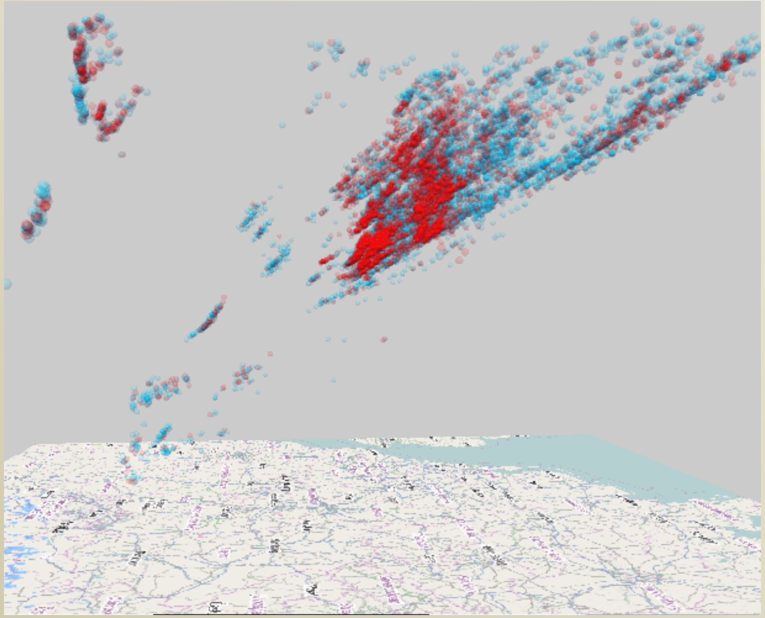
Diagonally elongated clusters may mean that the events were caused by a moving phenomenon, such as a storm or parade.
This example: lightning events.



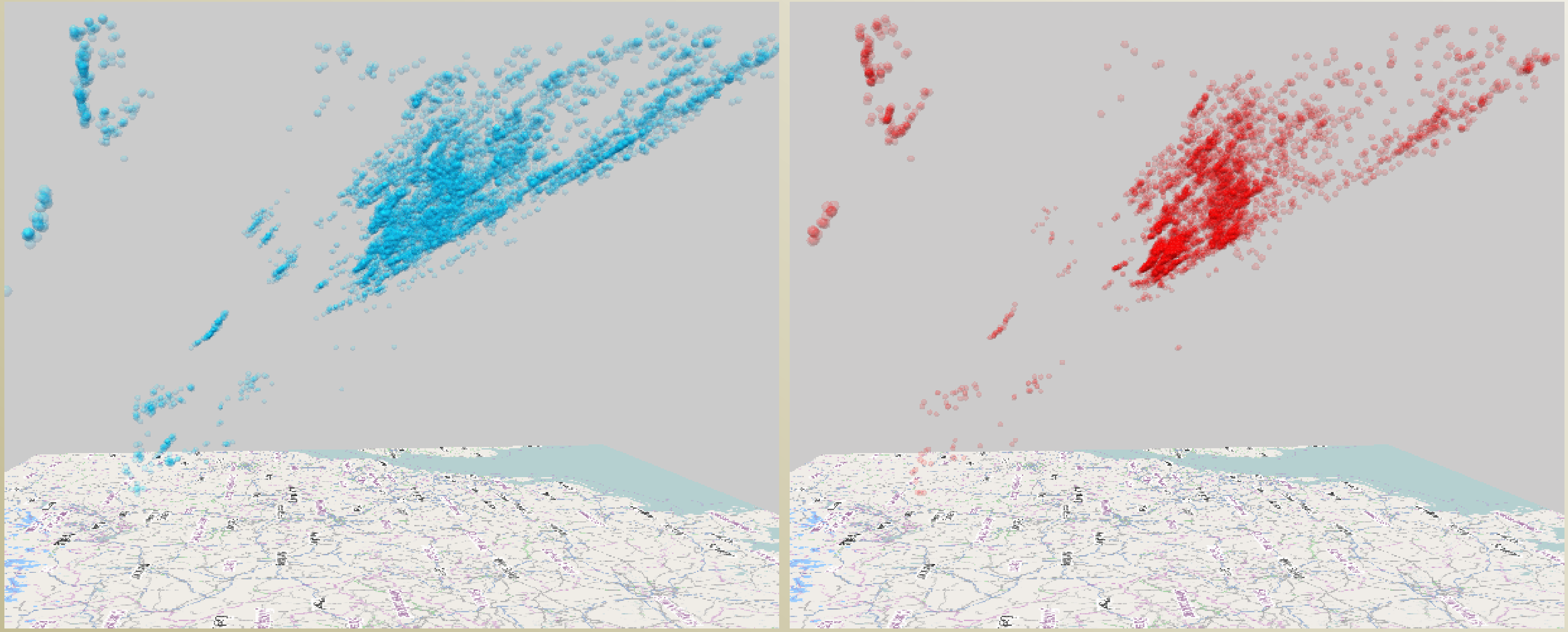
Spatial, temporal, and spatio-temporal distribution of thematic attribute values





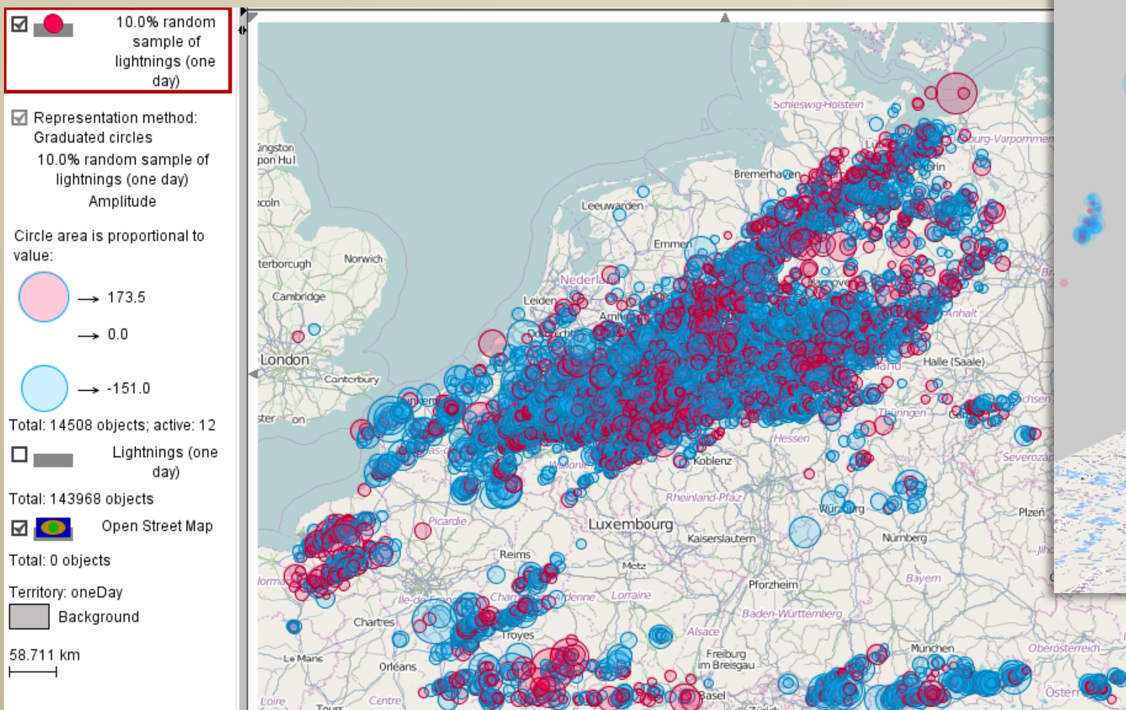


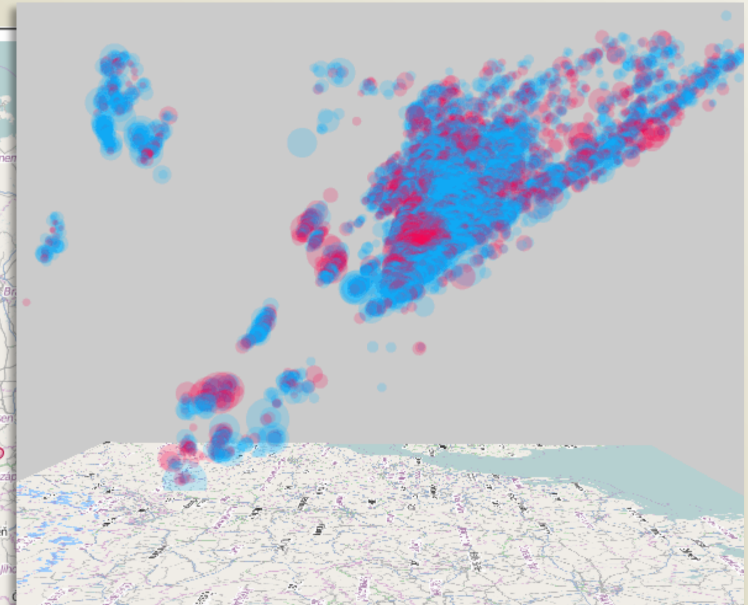
Spatio-temporal distributions of different attribute values



Different attribute values are selected using interactive attribute-based filtering.

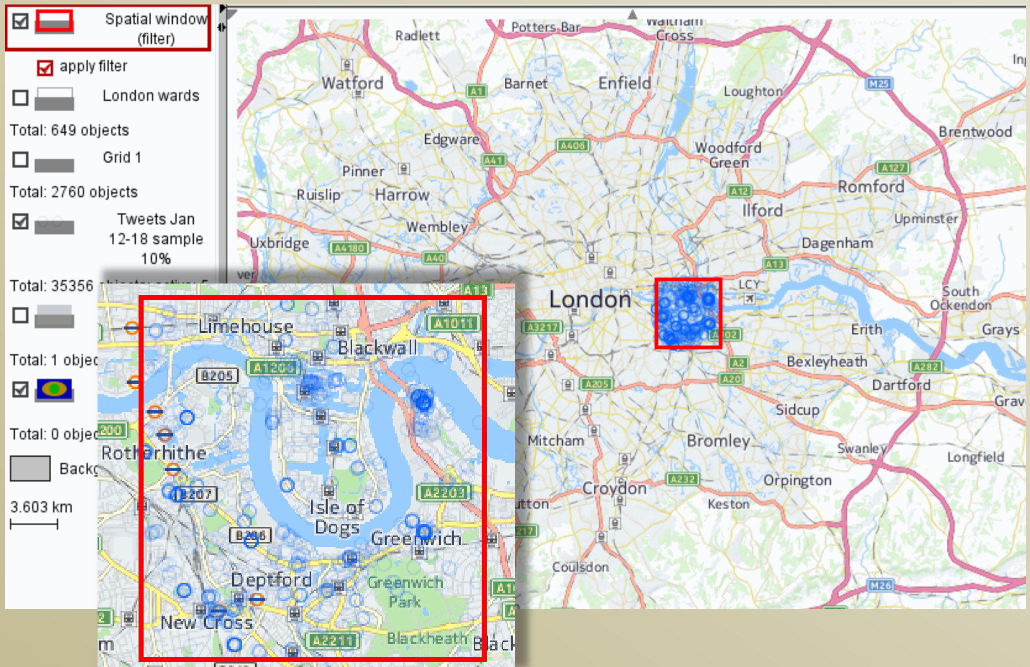
Visualisation of thematic attributes of events

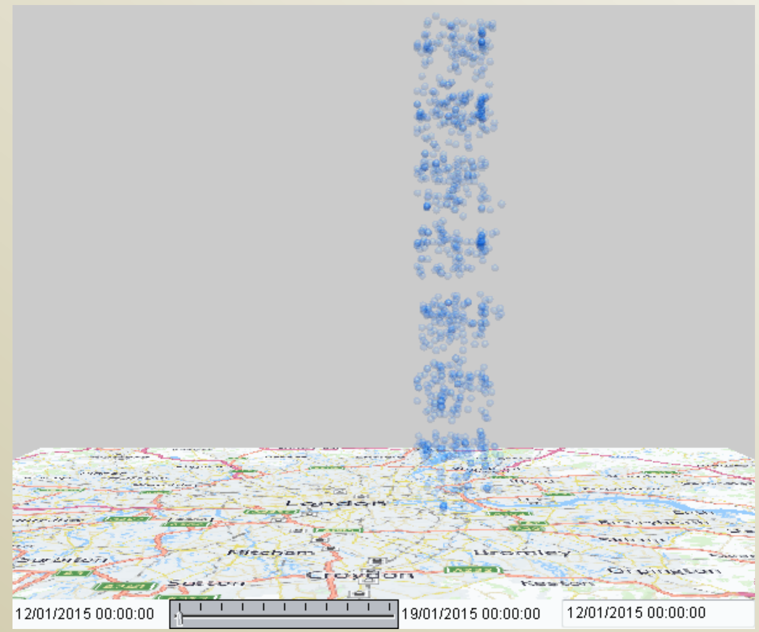


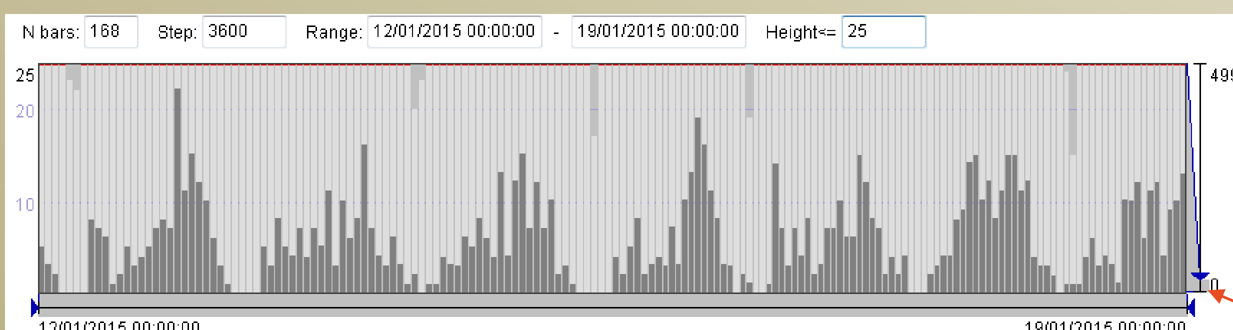


Although retinal visual variables can be used for representing the values of thematic attributes of spatial events on a map and in a space-time cube, the visual clutter and overlapping of marks make the displays illegible and practically useless for the analysis.

Spatial filtering of events

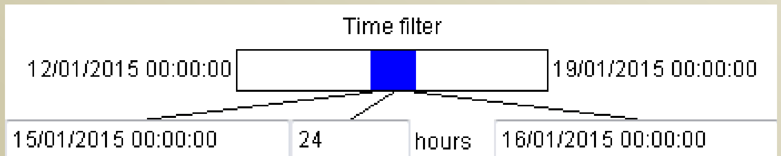


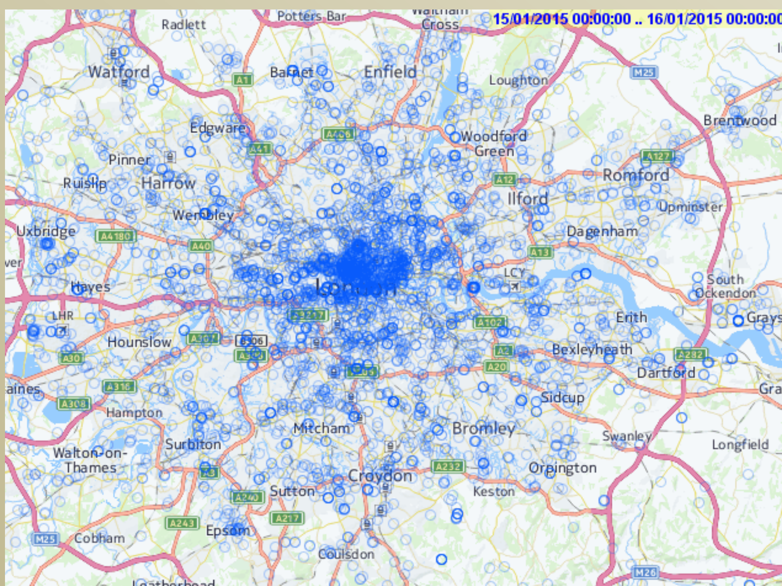


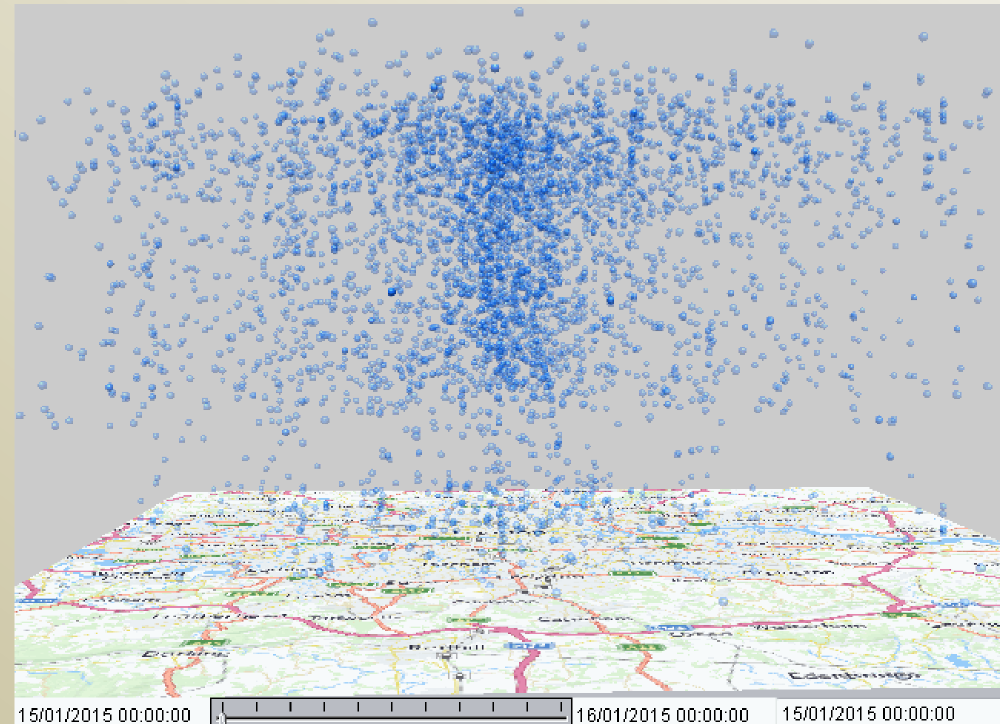


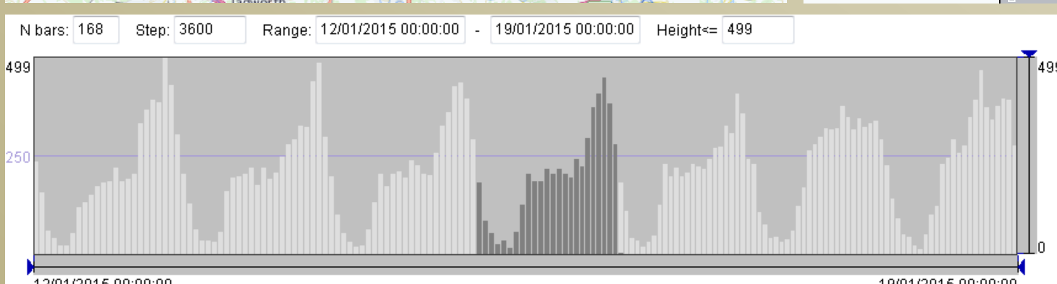
Note the focuser position

Temporal filtering of events









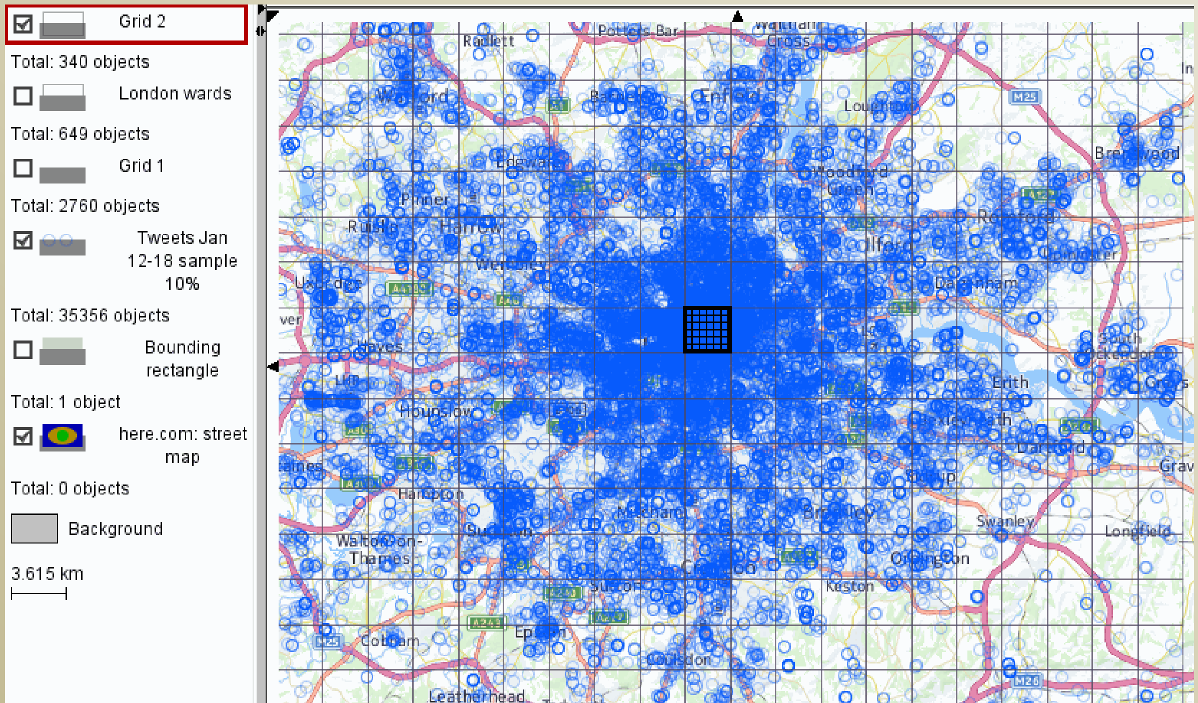
Insufficiency of visualisation and interaction techniques

- Map and STC do not work well when the events are numerous and dense.
  - The spatial and spatio-temporal distribution of the events themselves and their thematic attributes cannot be effectively analysed due to display clutter and greatly overlapping marks.
- Besides, STC may not work well when the time span of the data is long.
  - Spatio-temporal distribution patterns are not well visible.
- Interactive filtering can only partly reduce the display clutter while eliminating the overall view.
- Data transformations and computational techniques are strongly needed.
  - E.g., computational detection of clusters – to be considered later

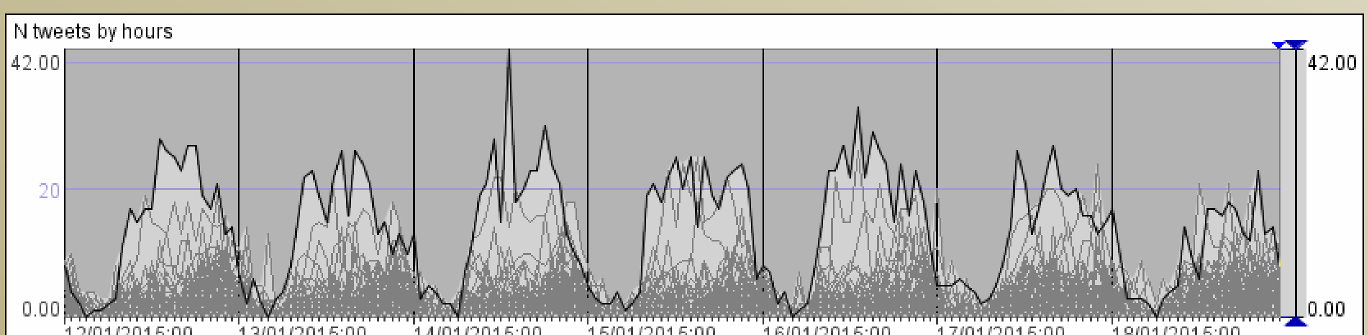
Spatio-temporal aggregation of events

- Spatial events and their thematic attributes can be aggregated spatially by areas (as considered before) and, simultaneously, by time intervals.
- Resulting data type: space-referenced time series of
  - event counts, densities, counts per capita, …
  - statistical summaries of thematic attributes: mean, median, mode, minimum, maximum, quantiles, …
- Structure of the resulting data: $S \times T \rightarrow A$
  - S: the areas by which the events have been aggregated
  - T: the time intervals by which the events have been aggregated
  - A: the aggregate attribute values

Spatio-temporal aggregation of events: an example



Geolocated tweets have been aggregated by cells of a regular grid and hourly time intervals. For each grid cell, there is a time series of hourly tweet counts.



Techniques for analysis of spatio-temporal event data: a summary

- Visual displays: map, space-time cube, time-based frequency histograms, various display types for thematic attributes.
- Interactive techniques: selection, classification, filtering.
- Data transformations: spatial aggregation, spatio-temporal aggregation.

Questions?

Analysis of spatial events

Summary: data structures and types

- Data components are distinguished according to their semantic roles: referrers and attributes
- Data types are distinguished according to the types and structure of the references
  - Major types: discrete objects, space, time
  - Structure: 1D (single referrer), 2D (2 referrers), …
    - In case of 2+D, each reference is a combination of values of the referrers; the set of possible references is a Cartesian product $R_1 \times R_2$ (× …).
- Discrete objects are distinguished according to their attributes:
  - Generic objects: have arbitrary attributes
  - Spatial objects: have attribute ‘spatial location’
  - Temporal objects (events): have attribute ‘existence time’
  - Spatio-temporal objects: have attributes ‘spatial location’ and ‘existence time’

Summary: synoptic analysis tasks

- Data ~ a function (in mathematical sense), i.e., a mapping from the domain of possible references to the domain of possible characteristics (combinations of attribute values) $f: R \to A$
  - R (referrers): independent variables
  - A (attributes): dependent variables
- Synoptic analysis tasks address the *behaviour* of this function, i.e., how A varies over the value domain of R
  - I.e., how the characteristics (attribute values or combinations) are distributed over the set of references.
- Tasks: describe the behaviour, detect particular type of behaviour (pattern), compare behaviours, relate behaviours

Summary: analytical techniques

- References: generic objects
  - Visualisations: attribute displays (frequency histogram, scatter plot, parallel coordinates)
  - Transformations: aggregation, projection (dimensionality reduction)
  - Interactions: selection, classification, attribute-based filtering
- References: space (spatial locations) or spatial objects
  - Visualisations: maps + attribute displays
    - Linking through interactive selection and classification; 2 ways of propagating selections and classes: attribute displays → map and map → attribute displays

  \+ All other techniques suitable for generic objects

Summary: analytical techniques *(continued)*

- References: spatial events (= objects having spatial locations and existence times)
  - Space (S) and time (T) can be viewed as object containers
    - Separately or in combination S × T = space-time continuum

  ⇒Specific definition of behaviour::= distribution of the objects and their characteristics over S, T, and S × T

  - Specific visualisations:
    - Map: shows the distribution over S
    - Temporal displays (time-based histograms, Gantt chart): show the distribution over T
    - Space-time cube: shows the distribution over S × T
  - Specific interactive techniques:
    - Spatial filtering
    - Temporal filtering

  \+ All techniques suitable for generic objects or spatial objects/locations

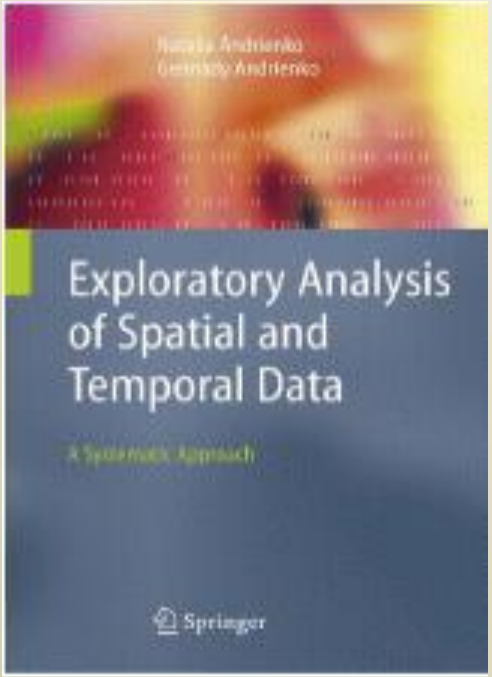
Reading:

http://0-dx.doi.org.wam.city.ac.uk/10.1007/3-540-31190-4

Natalia and Gennady Andrienko

Exploratory Analysis of Spatial and Temporal Data
A Systematic Approach

Springer-Verlag, 2005, ISBN 3-540-25994-5



Chapter 2: Data
Chapter 3: Tasks