Lecture 01

Fundamentals of Visual Analytics

given by
prof. Gennady Andrienko and
prof. Natalia Andrienko

Content and objectives

- After defining what visual analytics is and why visualisation is important, the main principles of representing data visually will be introduced.
- Since the purpose of data visualisation is to support data analysis, we shall consider the general types of analysis tasks and the consequent requirements to analysis-supporting visual displays.
- After the lecture and following practical, you will be able to interpret the content of different types of data display and use various interactive operations designed to support data exploration.
- You will be acquainted with the software system V-Analytics, which will be used in the further study.

Definition of Visual Analytics

What did you learn earlier?
(Module INM430 – Introduction to Data Science)

- Visualisation is important!
- Visualisation is used to
  - understand whether data contain what you need
  - uncover data imperfections (strange values, strange gaps, …)
  - understand how to process the data to make them useful
  - check the results of data processing
  - compensate for limitations of descriptive statistics (recall Anscombe's Quartet http://en.wikipedia.org/wiki/Anscombe%27s_quartet)
  - understand outcomes of statistical analysis, machine learning, other computational analysis methods
- Shortly: visualisation enables understanding!

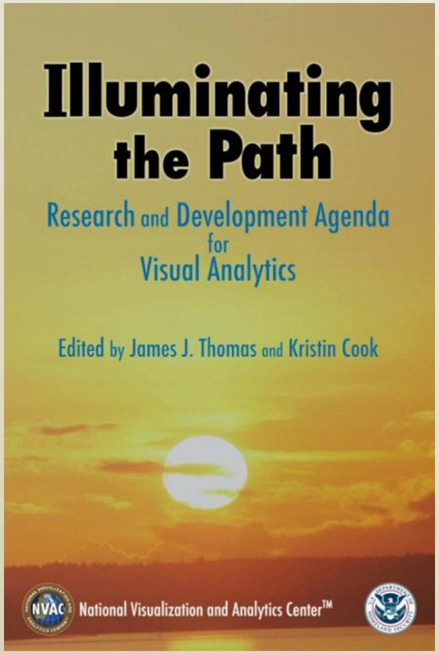
Visualisation is essential for thinking

- Visualise = make perceptible to human's mind
- "An estimated 50 percent of the brain's neurons are associated with vision. Visualisation <…> aims to put that neurological machinery to work."
  - B. McCormick, T. DeFanti, and M. Brown. Definition of Visualization. *ACM SIGGRAPH Computer Graphics*, 21(6), November 1987, p.3
- "An **abstractive grasp of structural features** is the very basis of perception and the beginning of all cognition."
  - R. Arnheim. *Visual Thinking*. University of California Press, Berkeley 1969, renewed 1997, p. 161

⇒ The act of seeing already includes analytical activities (abstraction and feature extraction) and triggers all further mental processes.

Visual Analytics definition (reminder)

- The science of analytical reasoning* facilitated by interactive visual interfaces
- People use visual analytics tools and techniques to
  - synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
  - detect the expected and discover the unexpected
  - provide timely, defensible, and understandable assessments
  - communicate assessment effectively for action



The book (IEEE Computer Society 2005) is available at http://nvac.pnl.gov/

***Analytical reasoning =**

data → information → knowledge → solution, decision, …

(interpreted data)

VA focuses on human reasoning

- VA deals with problems that cannot (yet) be solved algorithmically
  - ill-defined
  - involving incomplete and/or uncertain and/or conflicting data
- Human thinking is essential
  - pattern grasping, guessing, flexible use of previous knowledge and experience, novel approaches, trial and error
- Visualisation is essential for human thinking

➢ But: only human thinking is often insufficient

- Cannot effectively cope with massive data amounts, high dimensionality, …
- May be too slow

Principle of Visual Analytics:

Use the best of the humans and the computers

Computers

- can store and process great amounts of data
- are very fast in searching information
- are very fast in data processing
- can interlink to extend their capacities
- can efficiently render high quality graphics

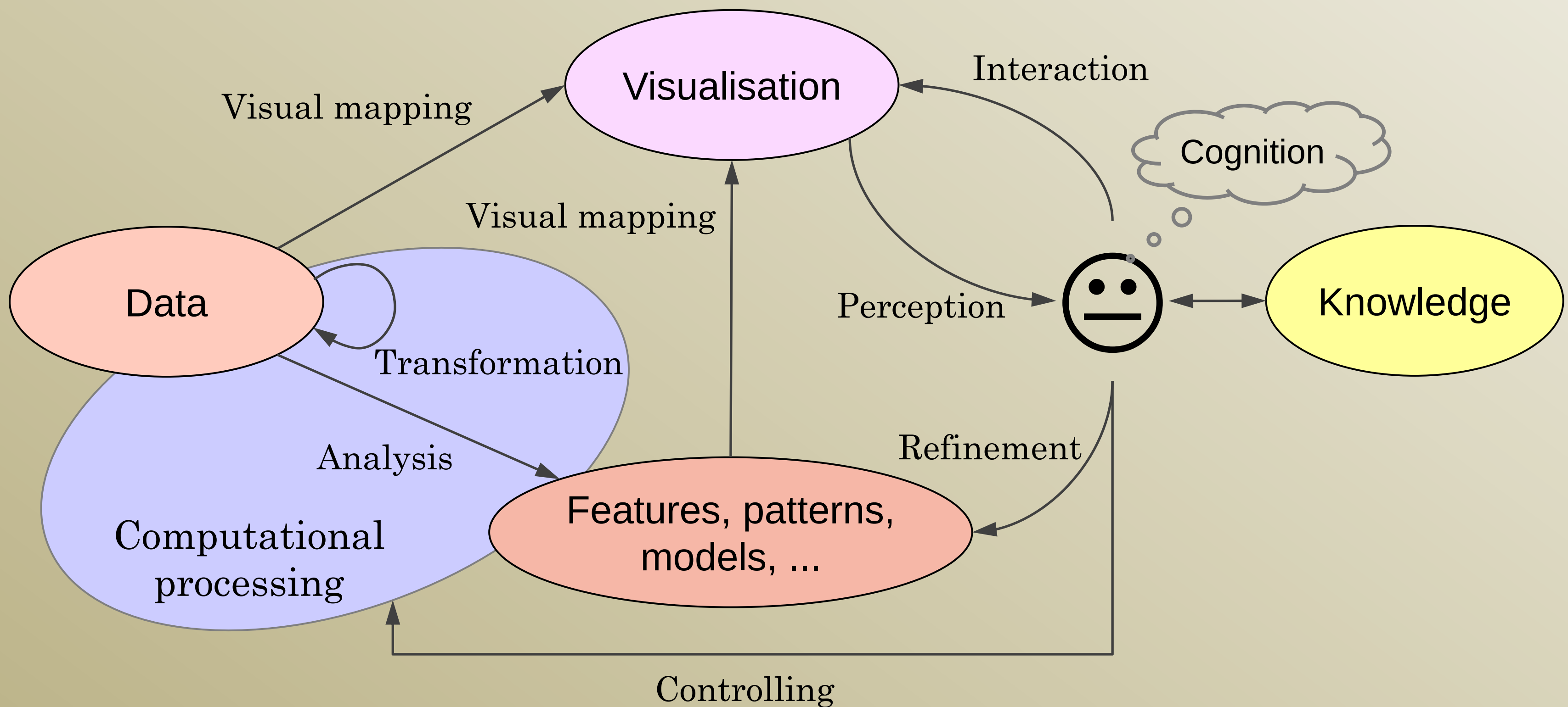
Humans

- are flexible and inventive, can deal with new situations and problems
- can solve problems that are hard to formalise
- can reasonably act in cases of incomplete and/ or inconsistent information
- can simply ***see*** things that are hard to compute

Visual Analytics technology:

- combine visual and computational analysis methods

Goal: divide the labour between humans and computers so as to enable their synergistic work.



Visual Analytics

a summary (1)

- VA develops methods and technologies for solving complex problems requiring joint efforts of humans and computers
  - Ill-defined, hardly formalisable problems; incomplete and/or inconsistent data; new situations → humans
  - Massive data amounts, high dimensionality, rapid growth → computers
- **VA** may thus be defined as
  the science of human-computer data analysis, knowledge building, and problem solving

Visual Analytics

a summary (2)

- VA technology combines interactive visualisations with computational processing
  - transformations, database querying, data mining algorithms, statistics, geographical analysis methods, …
- VA aims at effective division of labour between humans and machines
  - *Visual representations* are the most effective means to convey information to human's mind and prompt human cognition and reasoning
  - Computational power must amplify humans' inherent perceptual and cognitive capabilities

Questions?

Definition of Visual Analytics

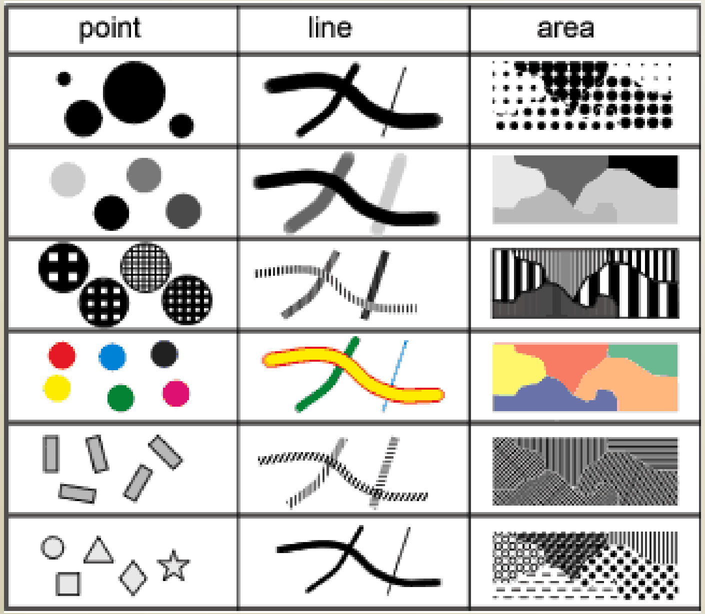
Visualisation primitives

Components of a visual display

- Display space (or visual space): set of positions
- Visual marks: points, geometric shapes, symbols positioned in the display space
- Types of marks:
  - Points
  - Lines
  - Areas
  - Surfaces
  - Volumes
- Visual properties of marks: colour, shape, orientation, size, texture

Visual variables

- Position in the display space (x,y)
- Size
- Value, or degree of darkness
- Texture
- Colour (hue)
- Orientation
- Shape



Jacques Bertin (1983): *Semiology of Graphics. Diagrams, Networks, Maps*. University of Wisconsin Press, Madison. Translated from Bertin, J.: *Sémiologie graphique*, Gauthier-Villars, Paris, 1967.

Perceptual properties of visual variables *(Bertin: level of organisation)*

- **Selection**: easy to locate marks with a given value of the visual variable and perceive them separately from others
- **Association**: several marks with the same value of this visual variable can be easily grouped (perceived all together) despite differences in other visual variables
- **Order**: the values of the variable can be put in an order, e.g., from small to big, from light to dark, ...
- **Quantity**: differences between two values of the variable can be interpreted numerically, e.g., how much bigger
- **Perceptual length**: number of different values of the variable that can be easily distinguished by an average human

Selectivity

- Can attention be focused on one value of the variable, excluding other variables and values?

- Find all letters on the left half of the image (position)
- Find all small letters (size)
- Find all red letters (colour)
- Find all letters ‘K’ (shape)

Easy to find ⇒ selective variable

Associativity

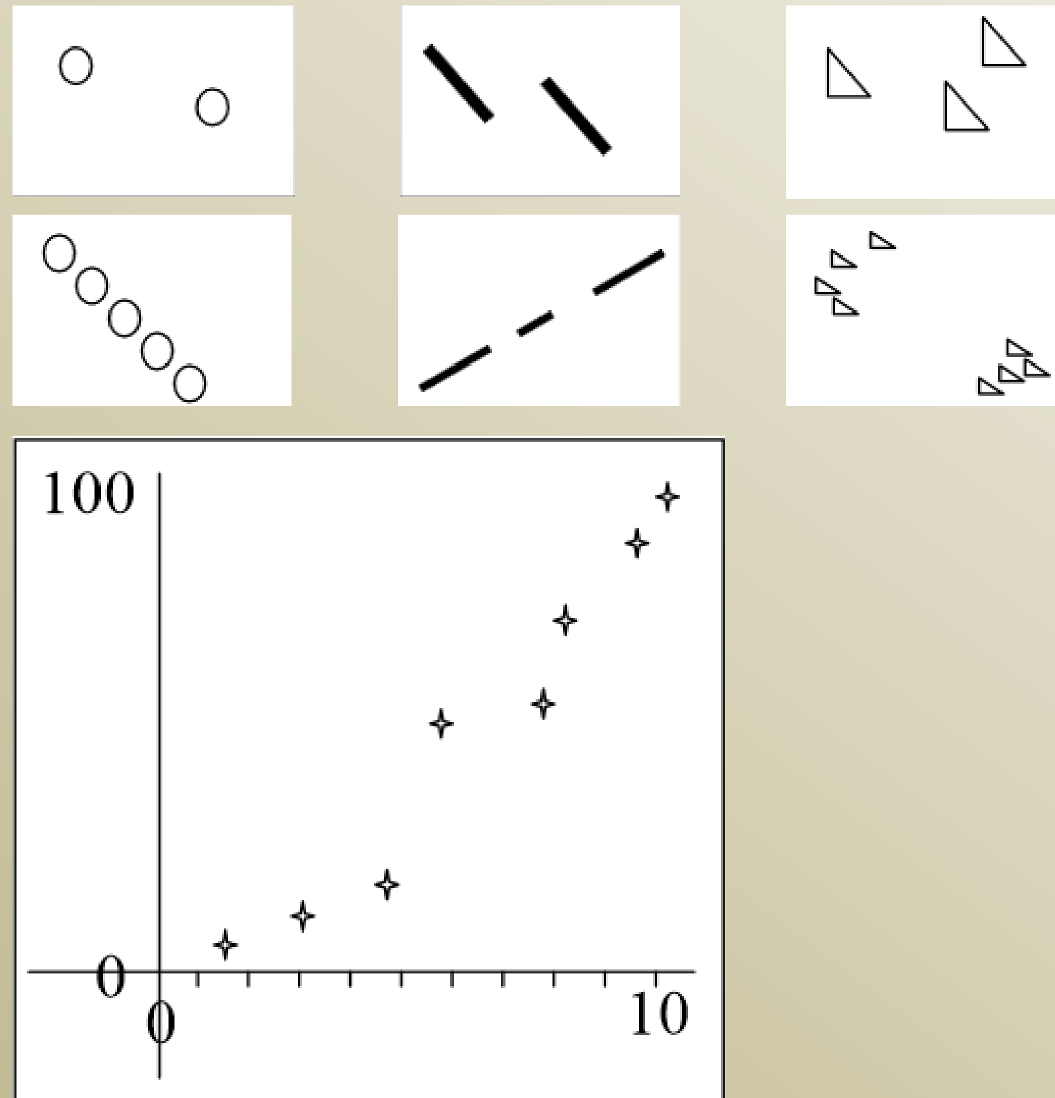
- Can marks with the same value of the visual variable be perceived simultaneously?

- Group all letters located on the left half of the image (position)
- Group all small letters (size)
- Group all red letters (colour)
- Group all letters ‘K’ (shape)

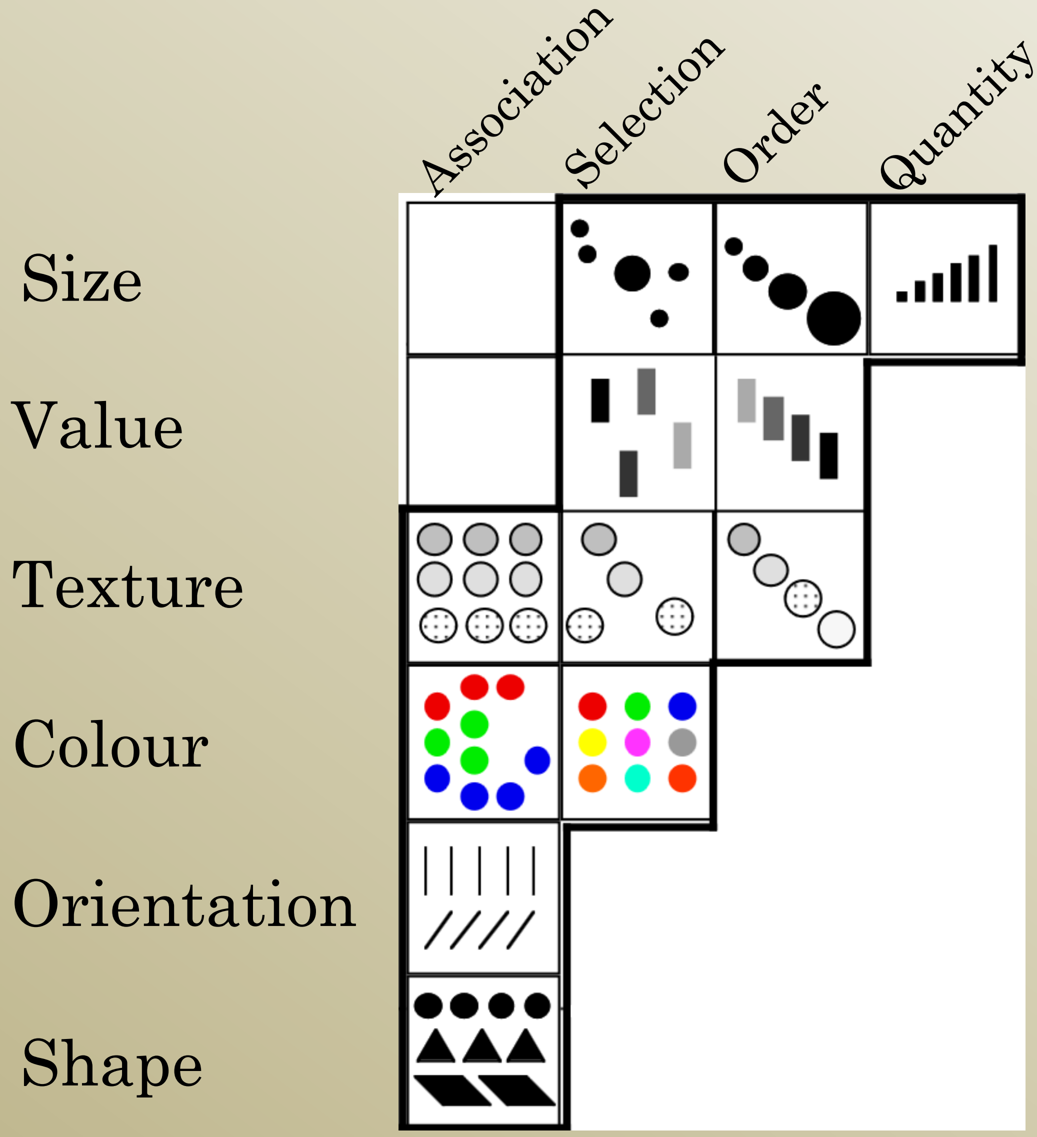
Easy to group ⇒ associative variable

Visual variable ‘position’

- ✓ Selection
- ✓ Association
- ✓ Order
- ✓ Quantity
- ✓ Length



"Retinal" visual variables



Perceptual properties of visual variables

	Association	Selection	Order	Quantity	Length
Position	+	+	+	+	limited by display size
Size	-	+	+	+	5-7
Value (darkness)	-	+	+	-	5-7
Texture	+	+	+/- *	-	5-7
Colour	+	+	-/+ **	-	7-8
Orientation	+	+	-	-	4
Shape	+	-	-	-	5-7

* + : grain density or size; - : grain shape
** + for parts of the spectrum (e.g., cold – hot)

The fundamental rule of visual mapping

- Visual mapping: the process of mapping data components to visual variables
- The fundamental rule: the perceptual properties of the visual variables must correspond to the properties of the data components they represent

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.

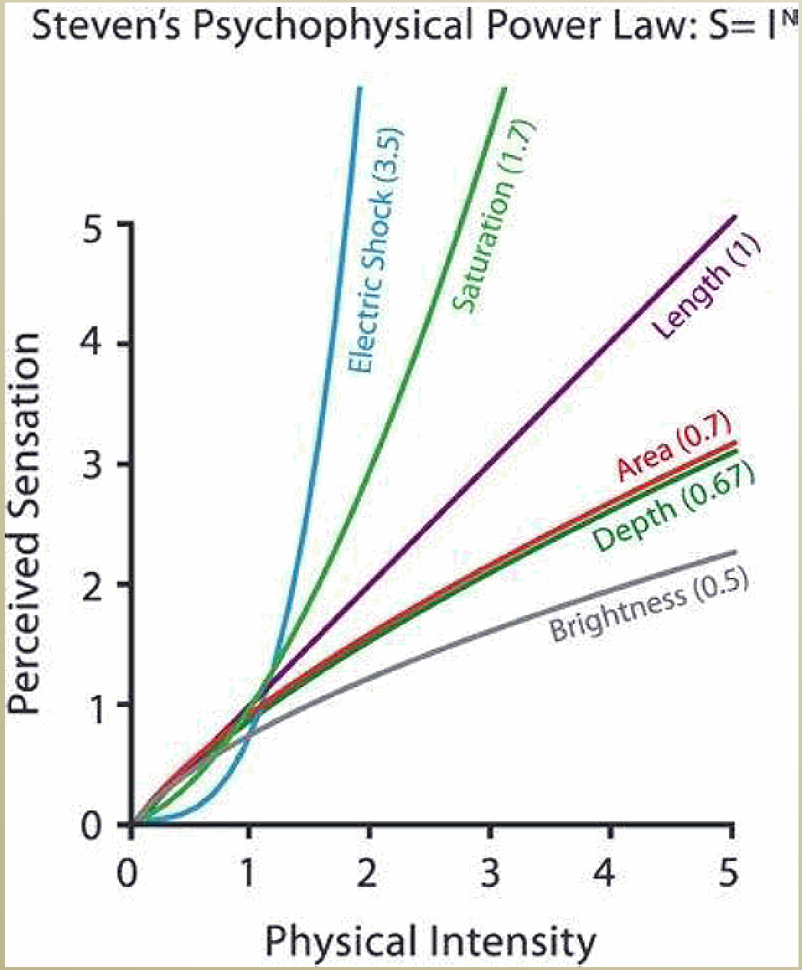
Correspondence of visual variables to types and scales of data components

	Nominal	Ordinal	Interval	Ratio	Spatial	Temporal
Position	+	+	+	+	+	+
Size	-	+	+	+	-	-
Value (darkness)	-	+	+	-	-	-
Texture	+	+ *	-	-	-	-
Colour	+	+ **	-	-	-	-
Orientation	+	+	-	-	-	-
Shape	+	-	-	-	-	-

* grain density or size

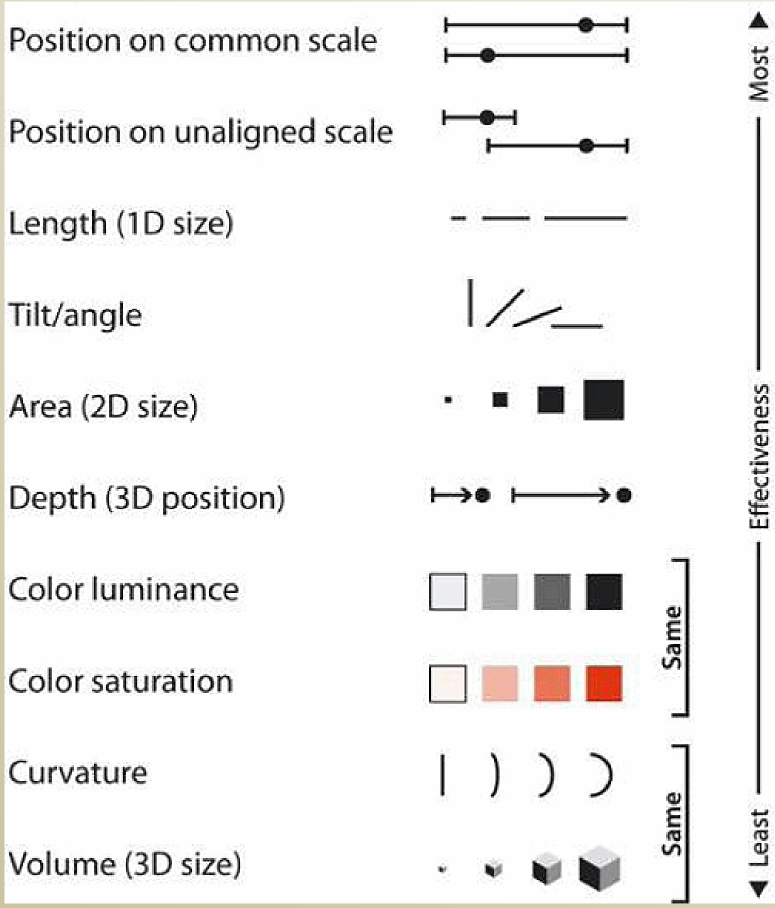
** parts of the spectrum (e.g., cold – hot)

Extensions and refinements of Bertin's theory



Stevens, S.S. (1957): On the psychophysical law. *Psychological Review* 64 (3): 153–181

Suitability for representing numbers



T. Munzner (2014): *Visualization Analysis and Design*. A K Peters Visualization Series, CRC Press

Questions?

Visualisation primitives

Data analysis tasks

(briefly)

Data analysis tasks

- Data reflect some pieces of reality. We analyse data to learn something about the reality.

data → information → knowledge → solution, decision …

- Analysis can be seen as a sequence of **tasks**.
- Each task is an attempt to obtain a certain kind of information or knowledge about the reality from the available data.
- It can be seen as finding an answer to some question about the reality.
- Tasks can be distinguished according to the kind of information or knowledge they seek to obtain.

Task types

- **Bertin**: There are as many possible question types as components in the data, i.e., each component may be a target of a question (task).
  - Who lives on Pine street? When was Paul born? What is Julia's school grade? Where does Mary live? How far is the school from Peter's home? How does Mary get to the school?

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

Task levels

- **Bertin**:
  - Elementary: question (task) addresses one or several *individual data items*, e.g., How do Peter and Julia get to the school?
  - Overall: question (task) addresses the *whole set* of data items, e.g., What are the methods of getting to the school? Which method is the most frequent?
  - Intermediate: question (task) addresses one or several *subsets of data items*, e.g., Which method of getting to school is more frequently used by girls?
- Synoptic tasks: overall + intermediate
  - Distinguishing property: a set or subset of data items is considered in its entirety; individual items are not of interest.

  ⇒ Synoptic tasks entail *abstraction* from individual items to sets

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

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	[illegible]	209	706	1711	[illegible]	8926	23207	4406
1960	5	Arkansas	1786272	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	899
1960	6	California	15717204	546069	37558	616	2859	15287	18796	508511	143102	311956	53453
1960	8	Colorado	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3750
1960	9	Connecticut	2535234	29321	928	41	103	236	548	28393	8452	16653	3288
1960	10	Delaware	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	122858	39966	[illegible]	9289
...													
1972	54	West Virgini	1781000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1953
1972	55	Wisconsin	4520000	[illegible]	4358	126	376	1661	2195	129024	28862	89642	10520
1972	56	Wyoming	345000	[illegible]	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	9456	1736
1973	4	Arizona	2058000	137966	9877	167	637	3031	6042	128089	40301	76560	11228
1973	5	Arkansas	2037000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2952
1973	6	California	20601000	1298872	116563	1862	8357	49531	56813	1182309	407824	643488	130997
1973	8	Colorado	2437000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	13951
1973	9	Connecticut	3076000	112717	[illegible]	102	[illegible]	[illegible]	3388	106296	[illegible]	58742	15893
...													
2000	44	Rhode Islan	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4665
2000	45	South Caroli	4012012	209482	32293	233	1511	5883	24666	177189	38888	123094	15207
2000	46	South Dakot	754844	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	798
2000	47	Tennessee	5689283	278218	40233	410	2186	9465	28172	237985	56344	154111	27530
2000	48	Texas	20851820	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	93161
2000	49	Utah	2233169	[illegible]	5711	[illegible]	[illegible]	1242	3563	[illegible]	14348	73438	6461
2000	50	Vermont	608827	18185	[illegible]	[illegible]	140	117	[illegible]	17494	3501	13184	809
2000	51	Virginia	7078515	[illegible]	[illegible]	401	[illegible]	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

Please note: **data structure ≠ table structure**!

It is not necessary that there is exactly one table column for each referrer.

1. Values from two or more table columns may specify the same reference in a complementary way, as in this example.
2. There may be no special column with references
   - Implicit references: when data describe objects with no identifiers assigned (e.g., lightning strokes), it is assumed that each table row refers to a different object.
3. Table columns may be associated with different references. E.g., the data from this table can be put in an equivalent table where columns correspond to different combinations (attribute, year). This does not change the structure of the data! *See the next slide.*

References are not always in columns

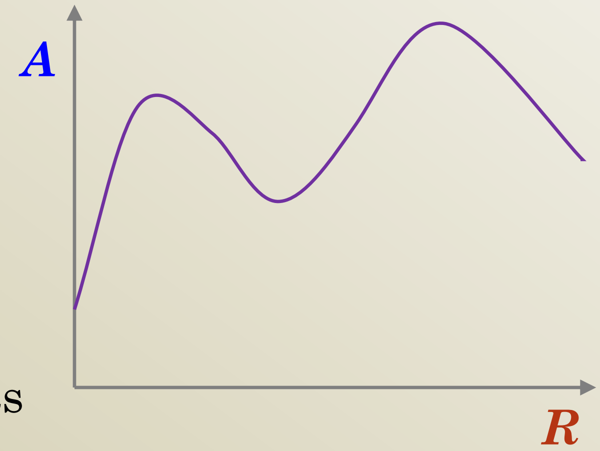
Referrer 1: place

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

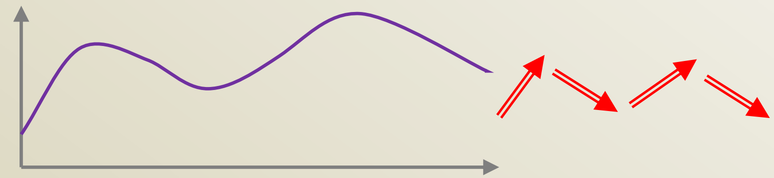
Data components viewed as variables

- 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

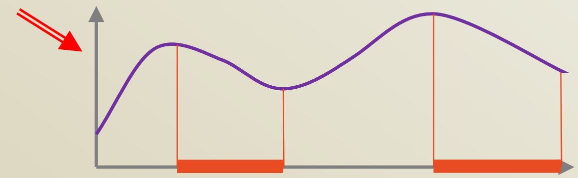


Synoptic tasks ~ tasks addressing <u>behaviours</u>

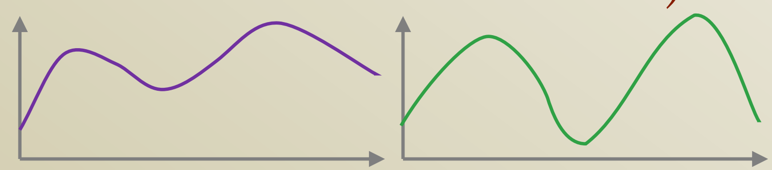
- Describe the behaviour of one or more attributes



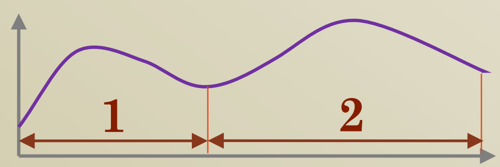
- Find subsets of references where attributes have particular behaviours



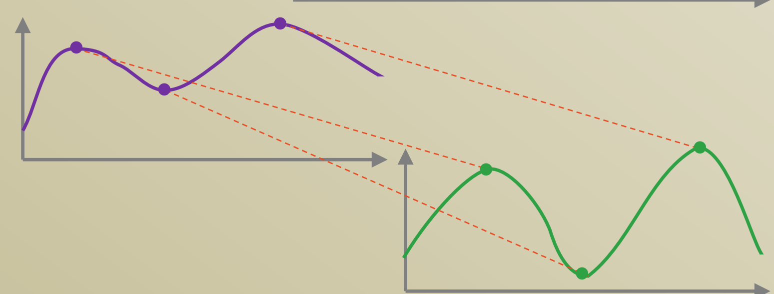
- Compare two or more behaviours (find similarities and differences)
  - Different attributes over the same set of references



  - 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

Elementary tasks ~ tasks addressing individual values

- Major classes of elementary tasks:
  - Determine the characteristics (attribute values) corresponding to particular references
    - What were the crime rates in Clerkenwell in the last year?
  - Find references having particular attribute values
    - Find districts and years where burglary rates were >50
  - Compare two or more characteristics
    - Different attributes for the same reference
      - Compare the burglary and robbery rates in Clerkenwell in the last year
    - Same attributes for different references
      - Compare the burglary rates in Clerkenwell and Bunhill; in 2013 and 2014

Synoptic tasks for data with >1 referrers

- When data have 2 or more referential components, there are tasks that are synoptic with regard to one component and elementary with regard to the other component(s)
  - ➢ "Semi-synoptic" tasks
- E.g. crime rates by districts and time
  - Describe the temporal variation of the crime rates in Clerkenwell – synoptic w.r.t. time and elementary w.r.t. space (the set of districts)
  - Describe the spatial distribution of the crime rates in 2012 – synoptic w.r.t. space and elementary w.r.t. time

Data analysis ~ system of tasks

- Data analysis consists of one or more synoptic tasks
- Complex synoptic tasks may be decomposed into simpler synoptic tasks (subtasks)
  - Describe the variation of different types of crime → describe the variation of each type of crime, compare the variations
  - Synoptic tasks may need to be decomposed into semi-synoptic tasks
  - Describe the overall spatio-temporal variation of the crime rates → describe the temporal variation in each district + describe the spatial variation in each year
- Elementary tasks play subordinate role: help in fulfilling synoptic tasks

Analysis tasks and data visualisation

- Purpose of data visualisation: enable a human analyst to accomplish analysis tasks

  ⇒ The visualisation must suit the analysis tasks

- Synoptic tasks entail *abstraction* from individual data items to sets and behaviours

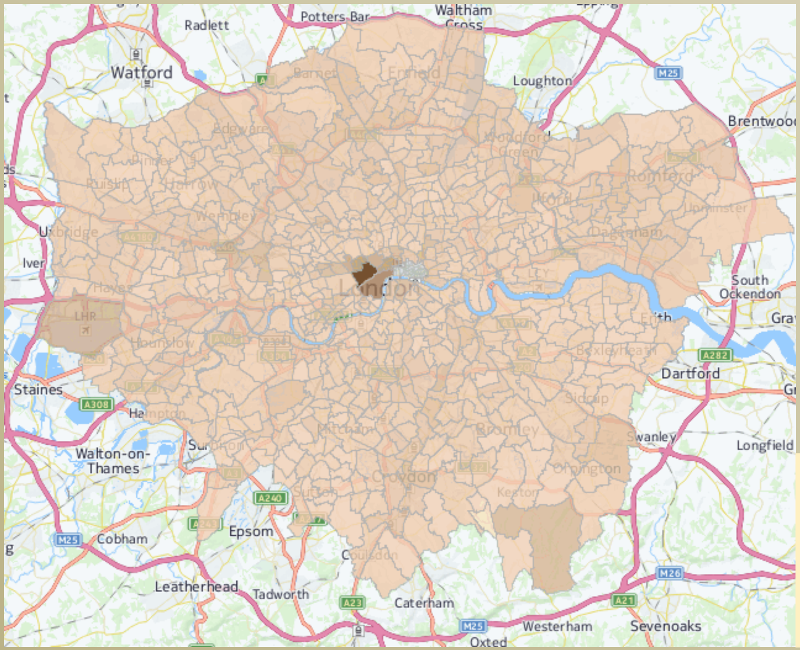
  ⇒ The visualisation should support *abstraction*

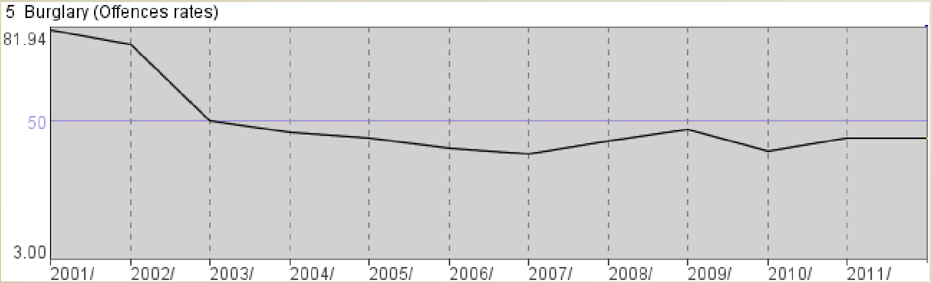
  ⇒ It should provide an *overall view* of the behaviour(s) of interest

  - Ideally, the analyst should be able to grasp the entire behaviour by a single sight

Bertin's "image"

- The meaningful visual form perceptible in the minimum instant of vision is called the **image**.
- The most efficient visualisations are those in which any question, whatever its type or level, can be answered in a single instant of perception, in a single image.





A behaviour over time is represented by a single line.

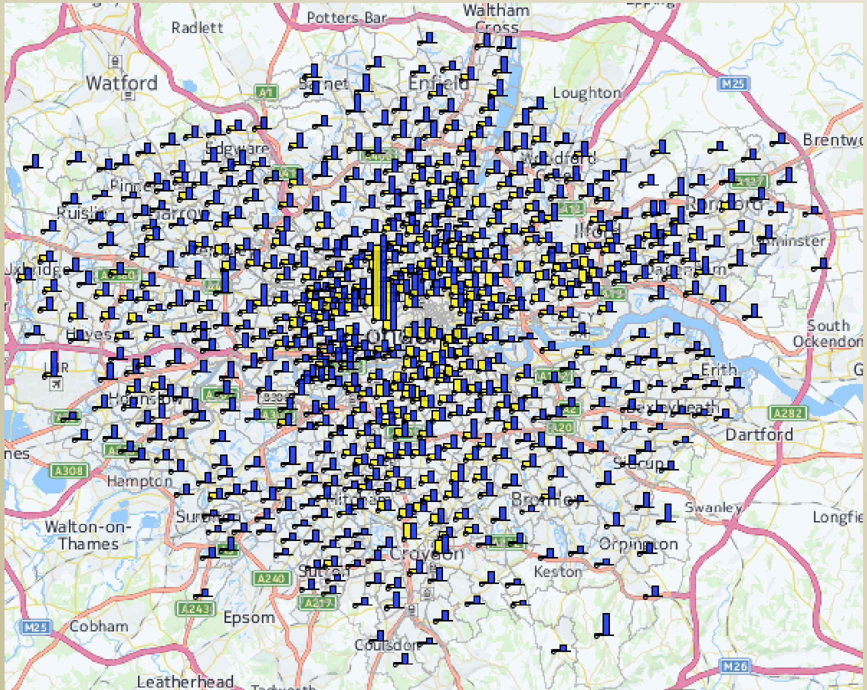
A behaviour over space is represented by colour value variation that can be perceived as a single figure.

Betrin’s “image” (continued)

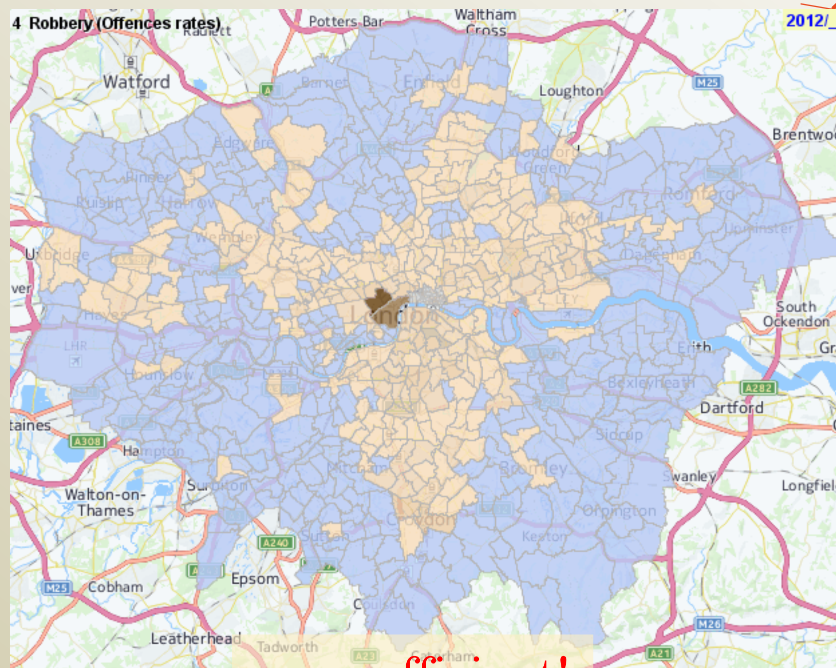
- An image can be composed using
  - two display dimensions
  - two display dimensions + single retinal variable

⇒ An image will not accommodate more than three data components

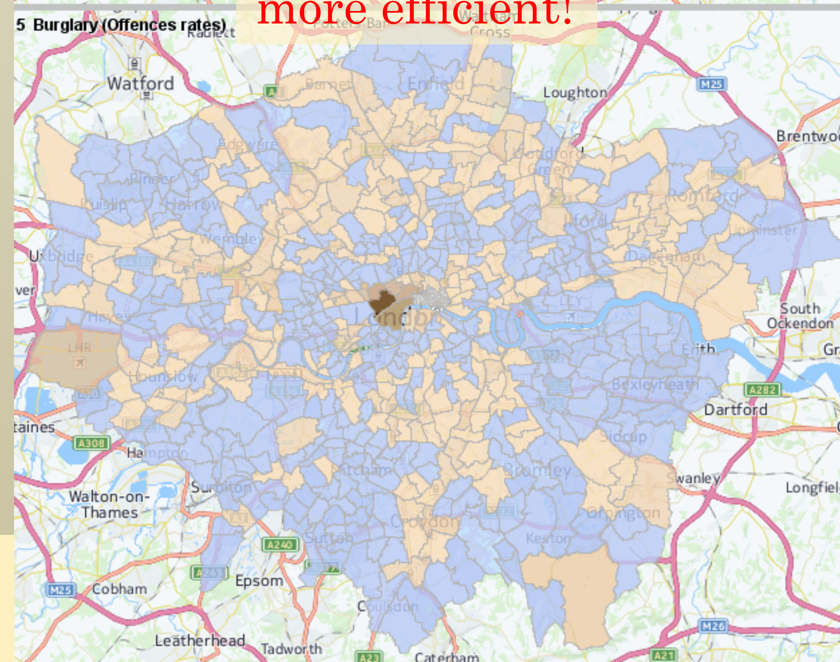
- For a given set of data components, a visualisation with fewer images is more efficient



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.

Support of synoptic and elementary tasks

- It is not always possible to construct a visualisation that is equally good for synoptic and elementary tasks

General approach:

- Visualisation should first and foremost support **synoptic tasks**
- Elementary tasks can be supported by *interactive operations* (to be considered later)

Questions?

Data analysis tasks

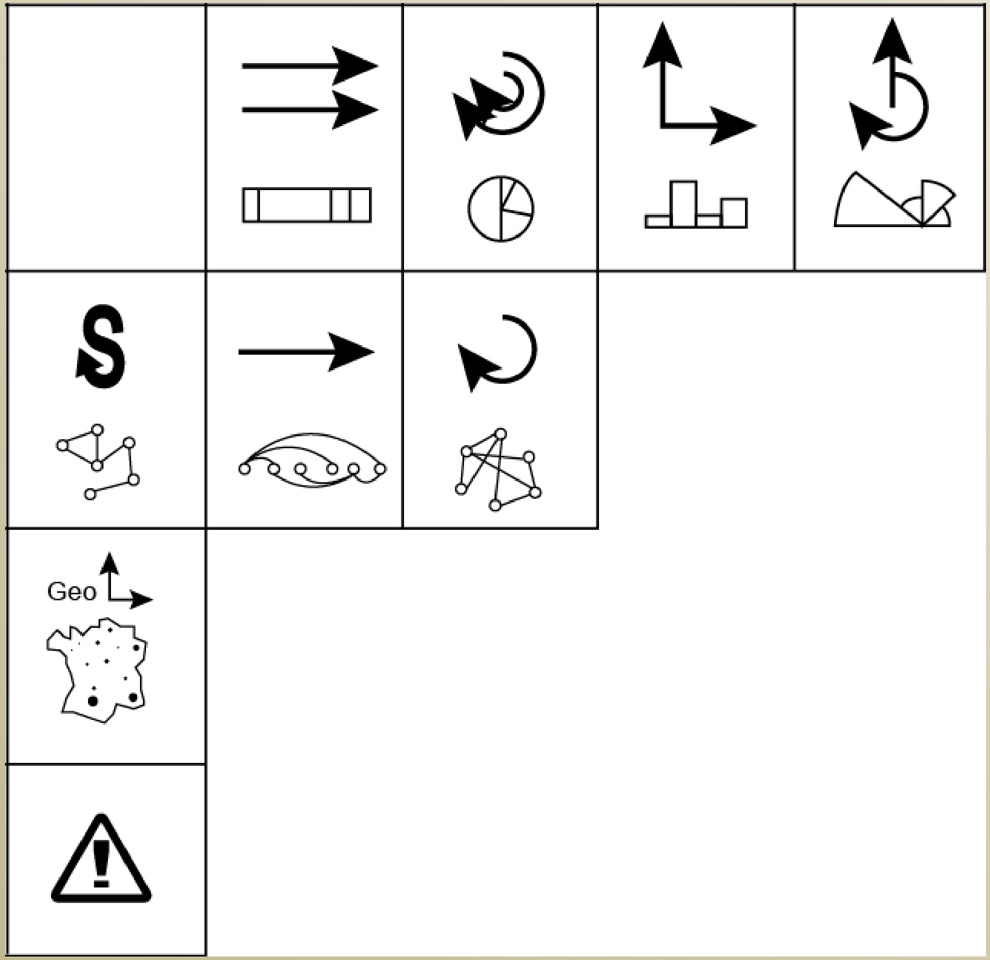
(briefly)

Types of visual display

Bertin’s “impositions”*

* Display types distinguished according to the way of utilising the plane dimensions

Arrangement Rectilinear Circular Orthogonal Polar



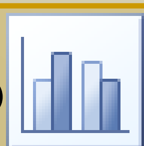
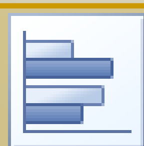
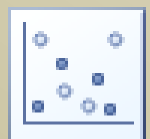
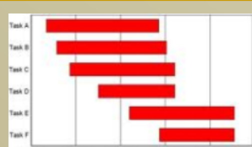
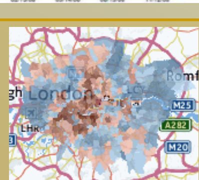
Diagrams
Coordinate system
(a.k.a. Graphs)

Networks
Layout
(a.k.a. Charts)

Maps
Isomorphism to physical space

Symbols
Arbitrary
(a.k.a. Icons)

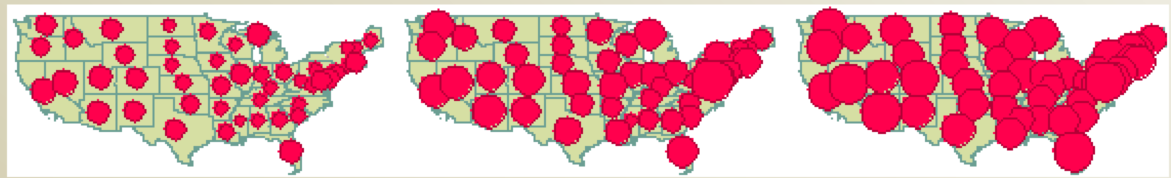
Common display types

	Display elements	Data components
Bar graph (bar chart)  	x- or y-position bar size (length)	references numeric attribute
Scatter plot 	marks (dots) x-position y-position	references numeric attribute numeric attribute
Line graph 	x-position y-position	ordered references, especially temporal numeric attribute
Gantt chart 	y-position x-position bar size (length)	temporal objects (events, processes …) existence time (start, end) duration
Map 	marks (x,y)-position other variables	spatial references (refer to spatial objects) spatial references or spatial attributes attributes

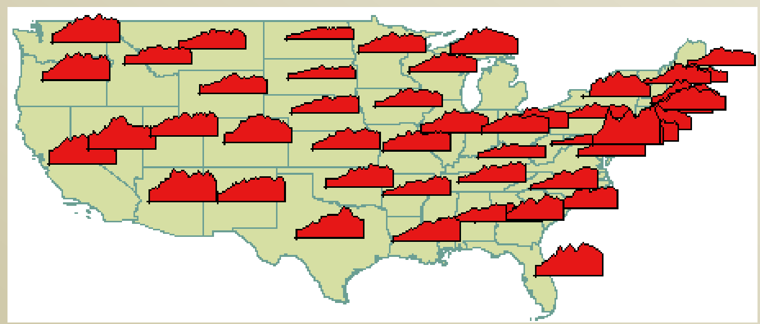
One display may be insufficient

Techniques to visualise complex data

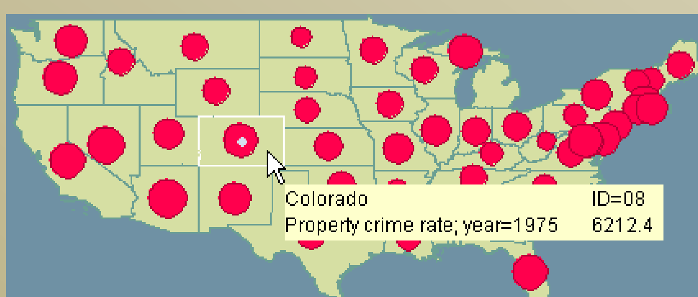
Display space division (small multiples)

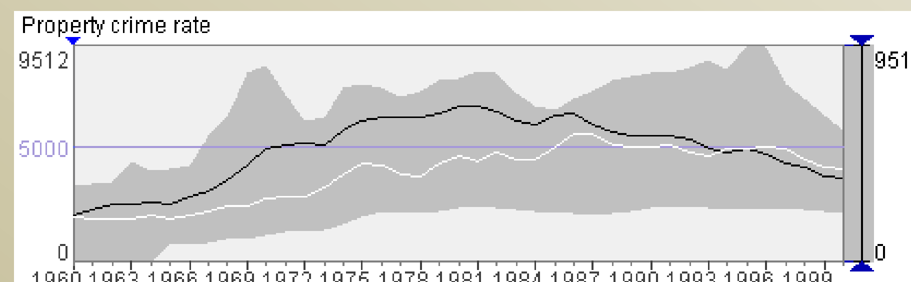


Space embedding (e.g., diagram spaces within map space)



Complementary displays

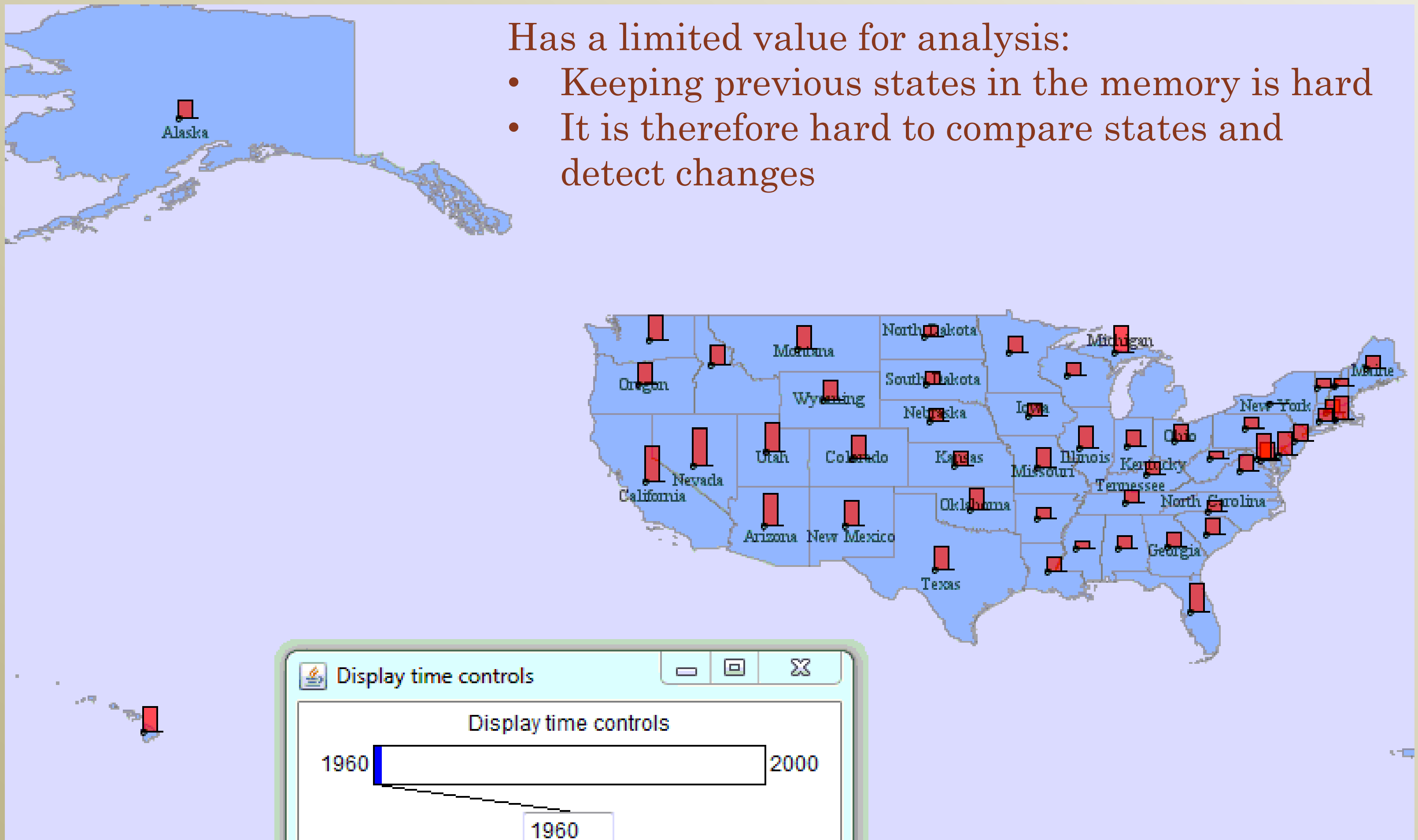




Additional visual variable: display time (animation)

Has a limited value for analysis:

- Keeping previous states in the memory is hard
- It is therefore hard to compare states and detect changes



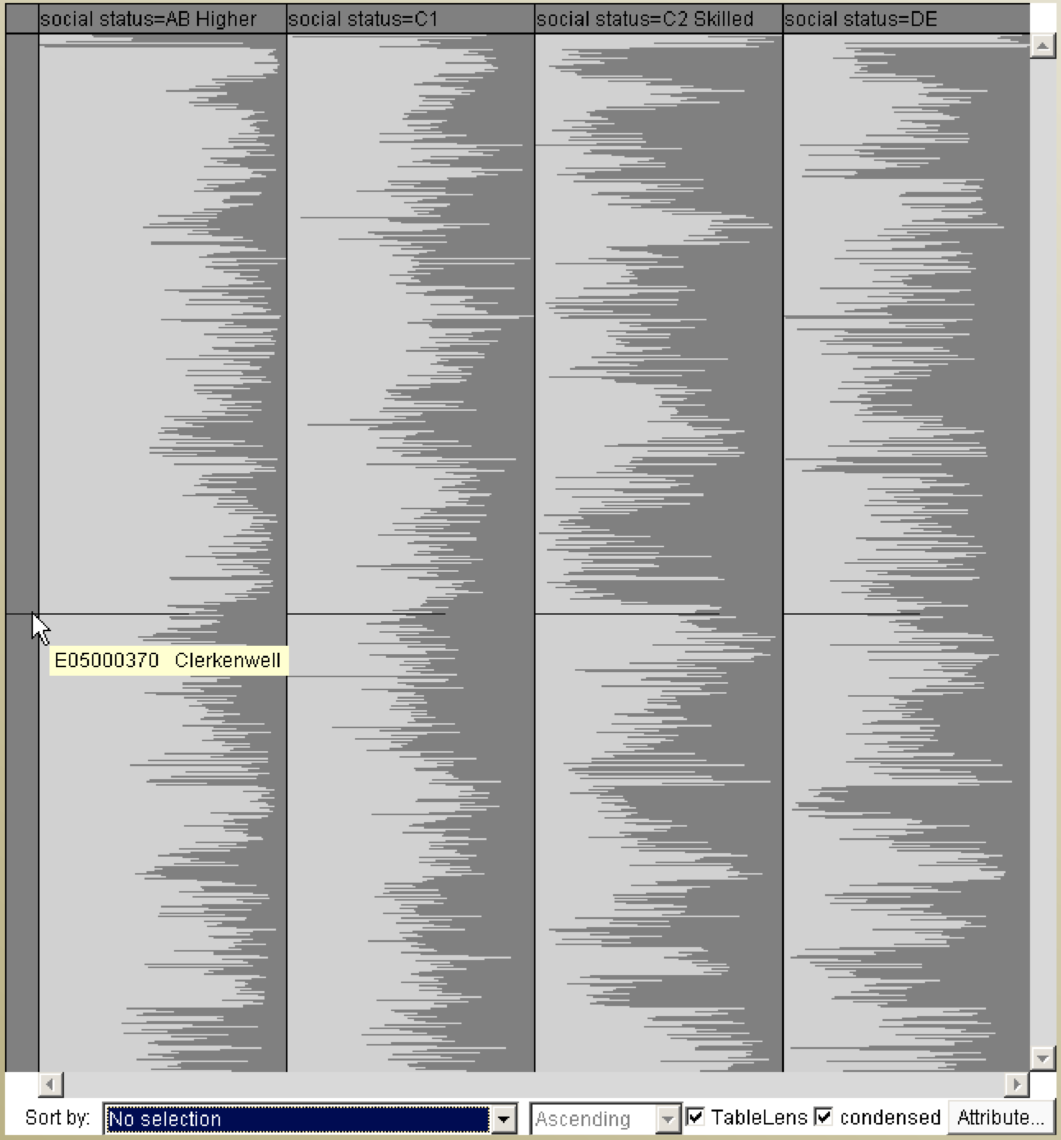
Bar diagram (bar chart)

- Here there is a bar diagram in each table column corresponding to a numeric attribute.
- Table rows (y-positions) correspond to references.
- The length of the darker grey bars represent numeric values.
- The number of simultaneously shown references is quite limited.
- Synoptic tasks are not supported.

☑ identifiers	social status=AB Higher and intermediate managerial/ administrative/ professional occupations: Population % by social status	social status=C1 Supervisory, clerical and junior managerial/ administrative/ professional occupations: Population % by social status	social status=C2 Skilled manual occupations: Population % by social status	social status=DE Semi-skilled and unskilled manual occupations; unemployed and lowest grade occupations: Population % by social status
E05000001 Aldersgate	52.8	17.7	1.8	1.5
E05000005 Bishopsgate	52.3	32.0	7.2	4.1
E05000015 Cripplegate	42.5	18.8	3.0	4.6
E05000017 Farringdon Within	42.8	22.8	4.3	2.9
E05000018 Farringdon Without	59.3	22.0	2.4	1.9
E05000021 Portsoken	14.9	21.1	11.2	23.2
E05000022 Queenhithe	66.1	22.6	4.4	0.6
E05000023 Tower	64.8	25.6	3.5	0.9
E05000026 Abbey	13.1	21.3	13.0	23.0
E05000027 Alibon	6.7	17.2	17.9	21.1
E05000028 Becontree	7.3	19.0	16.7	22.0
E05000029 Chadwell Heath	8.1	18.3	15.0	18.4
E05000030 Eastbrook	8.7	19.4	17.4	16.9
E05000031 Eastbury	7.1	17.9	16.5	22.0
E05000032 Gascoigne	7.1	17.0	13.1	24.6
E05000033 Goresbrook	6.1	17.8	17.9	21.5
E05000034 Heath	6.2	17.7	15.1	21.5
E05000035 Longbridge	11.8	20.8	14.8	17.9
E05000036 Mayesbrook	5.8	17.3	17.0	21.9
E05000037 Parsloes	6.2	17.2	18.9	20.3
E05000038 River	6.2	16.9	18.4	21.1
E05000039 Thames	7.4	17.8	16.4	21.3
E05000040 Valence	6.8	17.1	18.0	20.4
E05000041 Village	6.2	18.6	16.9	22.0
E05000042 Whalebone	9.7	21.2	16.8	16.9
E05000043 Brunswick Park	19.1	21.6	12.0	12.2

Sort by: No selection | Ascending | ☑ TableLens ☐ condensed | Attribute...

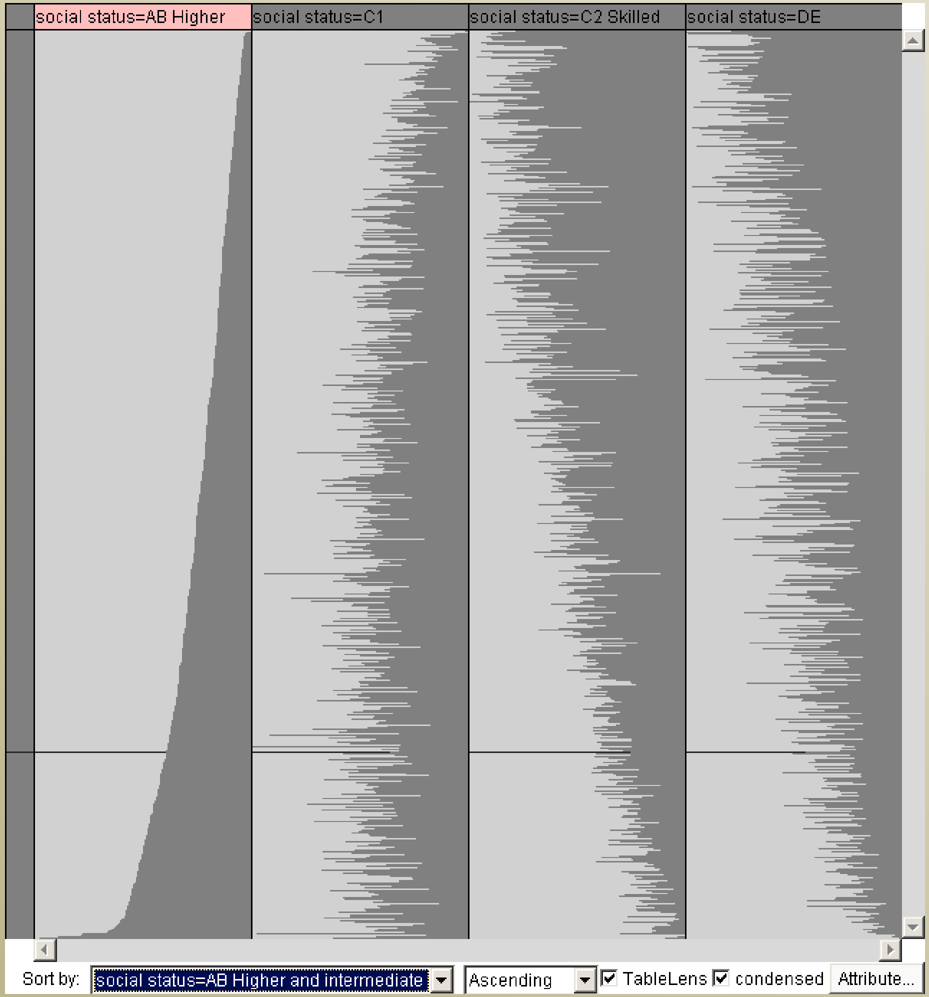
“Condensed” bar diagram

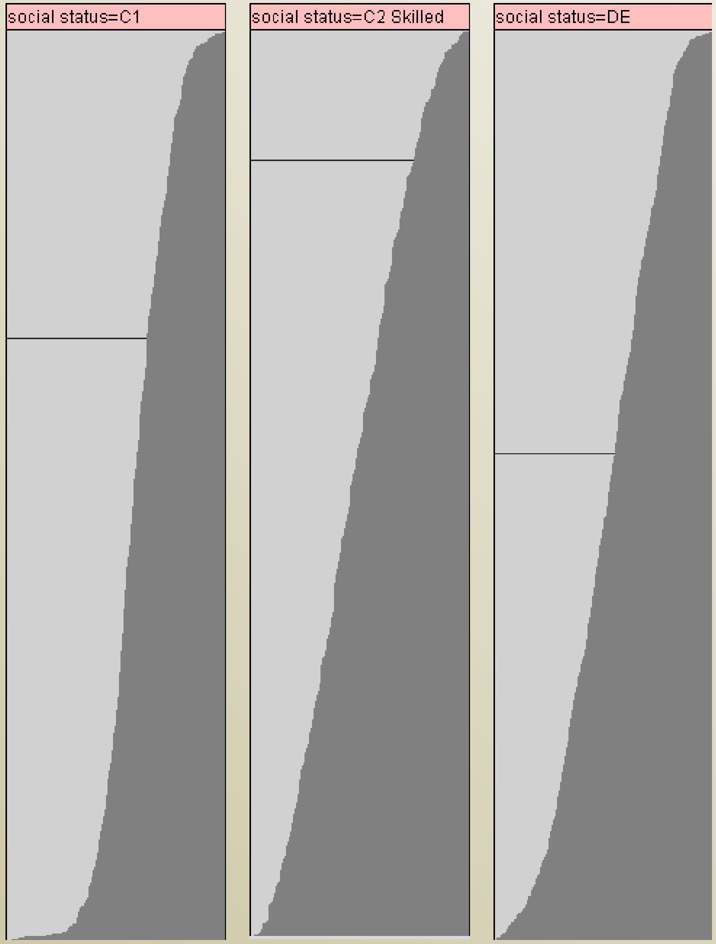


All references and corresponding attribute values are represented; however, individual references and attribute values can only be accessed through interaction.

Synoptic tasks are still not well supported.

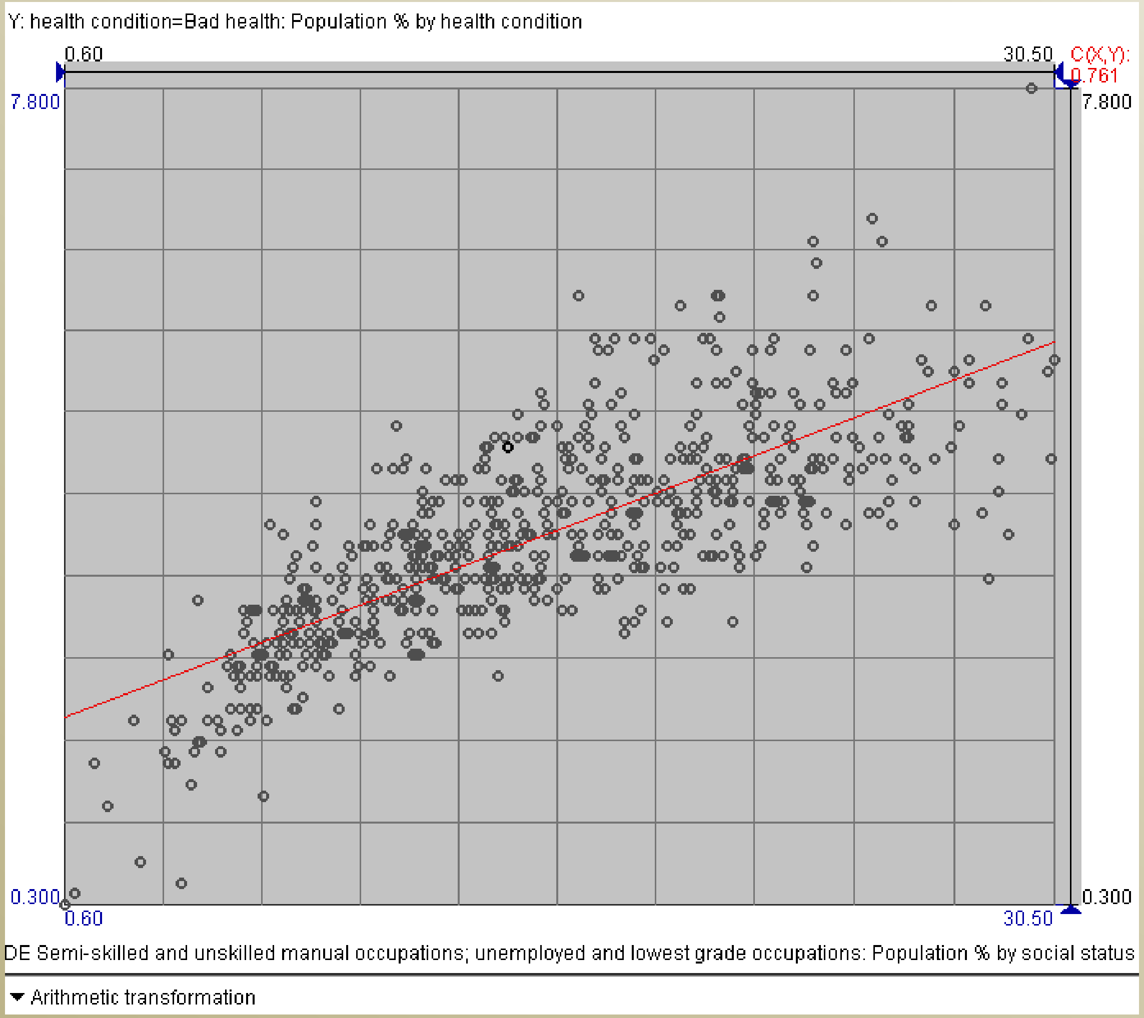
Bar diagram + sorting





We can grasp the properties of value distribution within the sorted column by a single instance of vision.

Scatter plot

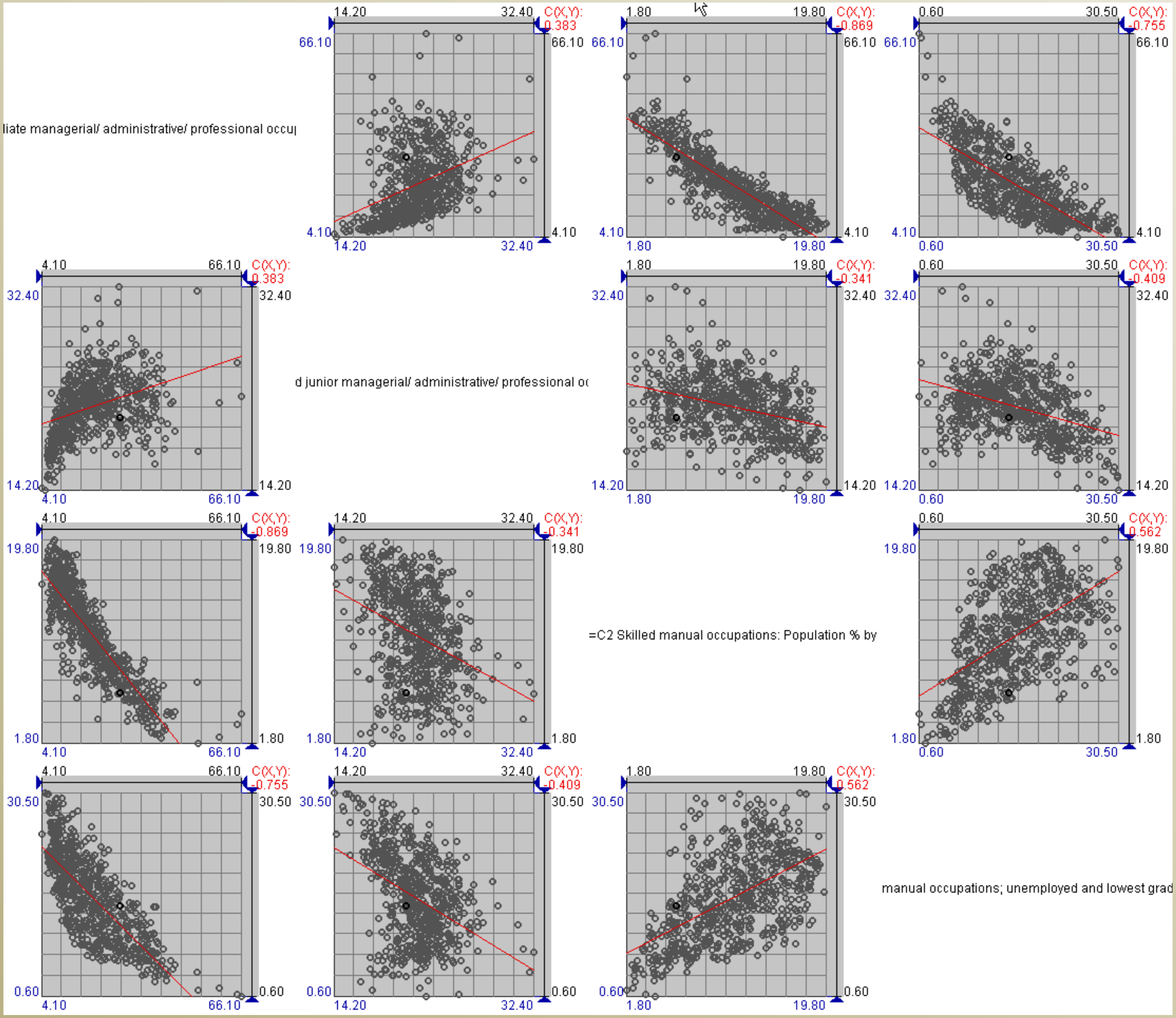


Represents the mutual behaviours of two numeric attributes ⇒ allows the analyst to relate these two attributes.

The marks form one or more figures (clouds) that can be perceived in a single instance of vision.

The relationship is inferred from the figure shape(s).

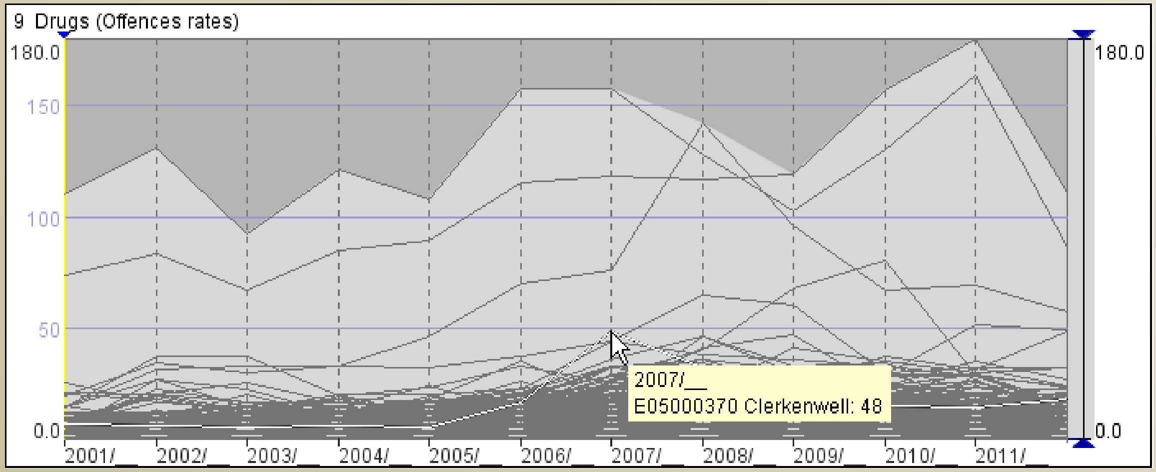
Scatterplot matrix

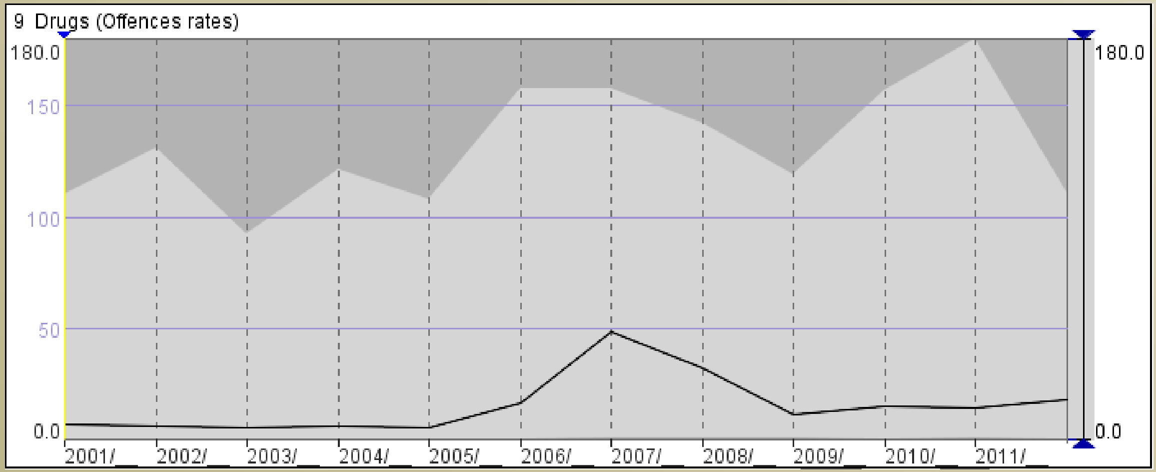


An application of the "small multiples" technique.

Allows comparison of mutual behaviours of different attribute pairs.

Line graph



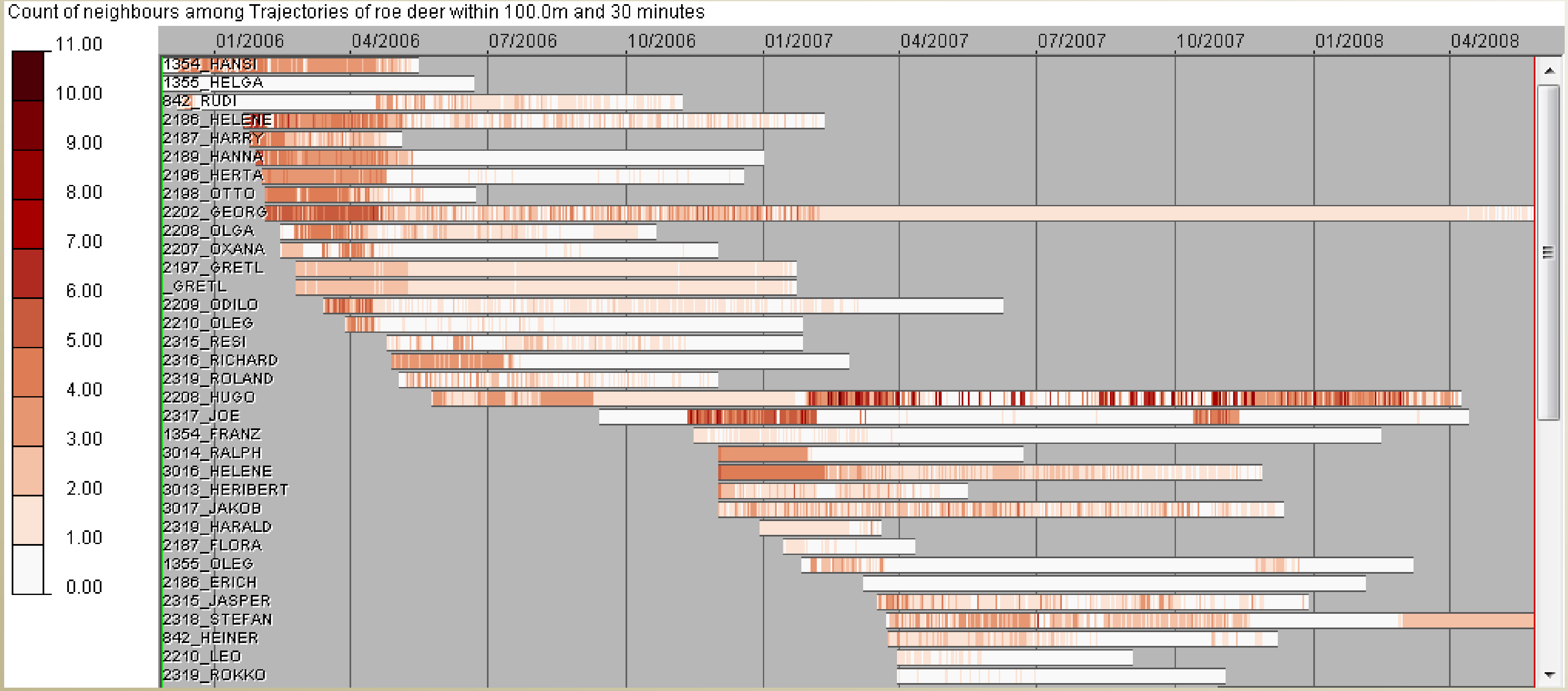


One curve (polygonal line) can represent the behaviour of a numeric attribute over an ordered set (sequence) of references, e.g., temporal.

In a case of two referrers (e.g., temporal and non-temporal), multiple curves corresponding to different non-temporal references can be drawn in the same display space. This, however, creates display clutter.

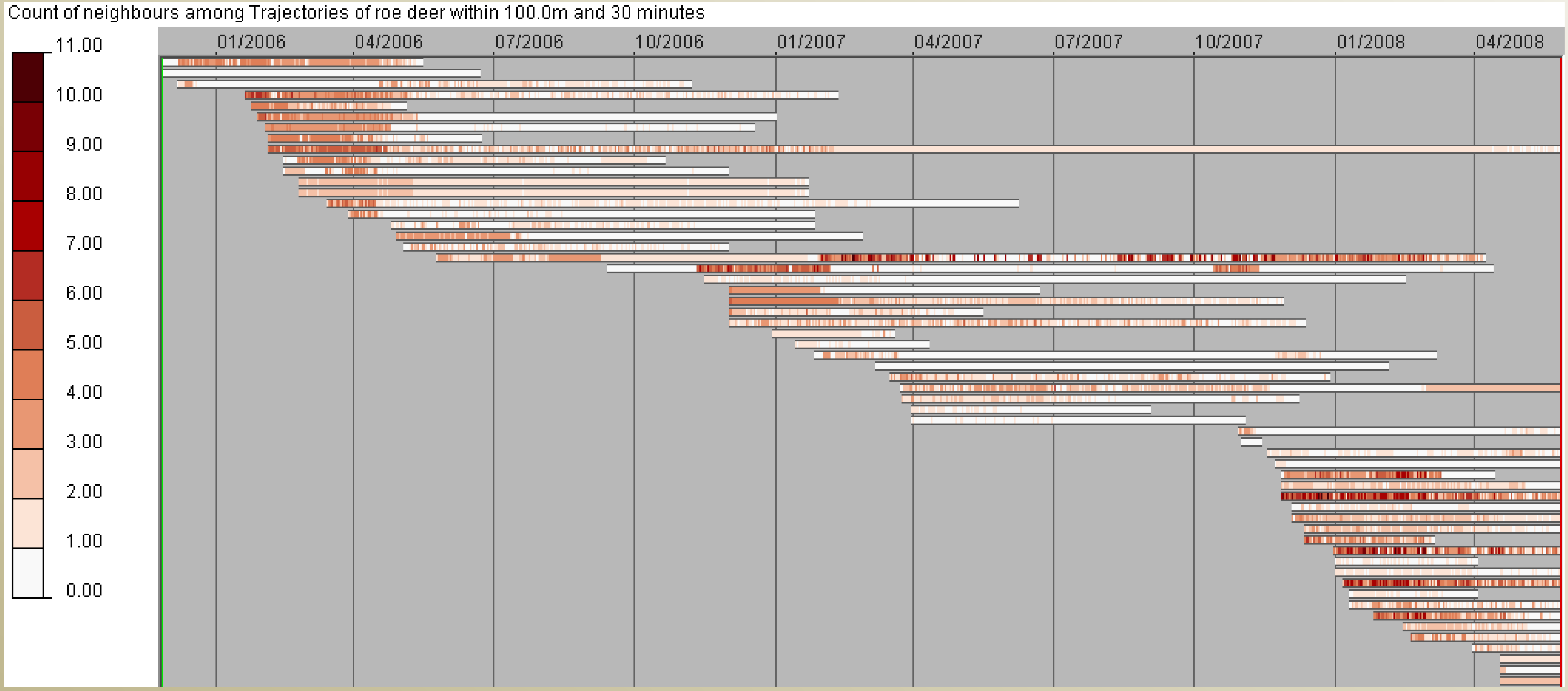
Interactive selection of singular curves supports semi-synoptic tasks.

Gantt chart



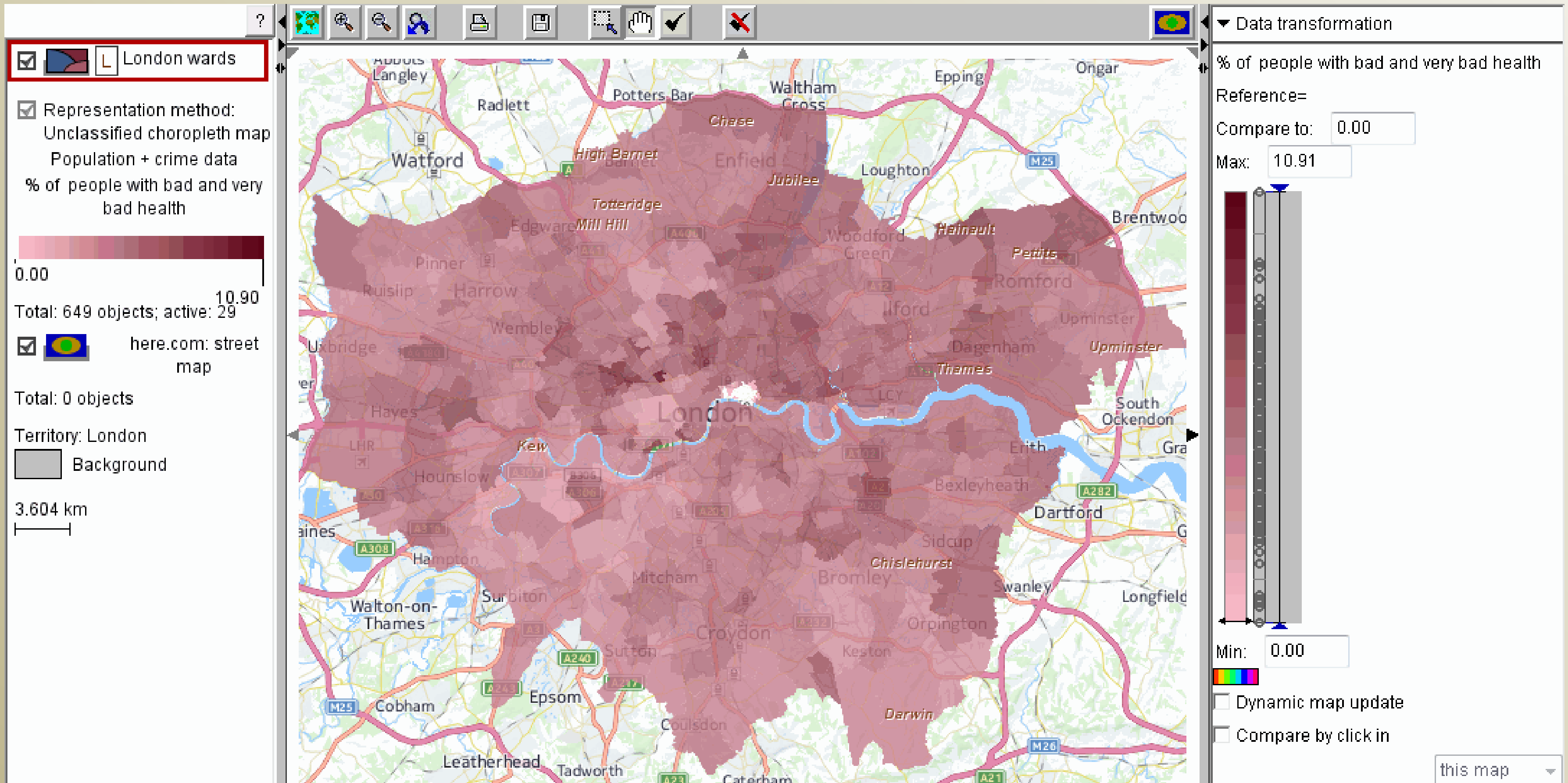
The horizontal dimension represents time. The horizontal positions and lengths of the bars show the existence times of objects, events, activities, or processes (here: life times of roe deer). Additionally, the bars can be painted to represent values of a time-variant attribute (here: number of neighbours).

Gantt chart



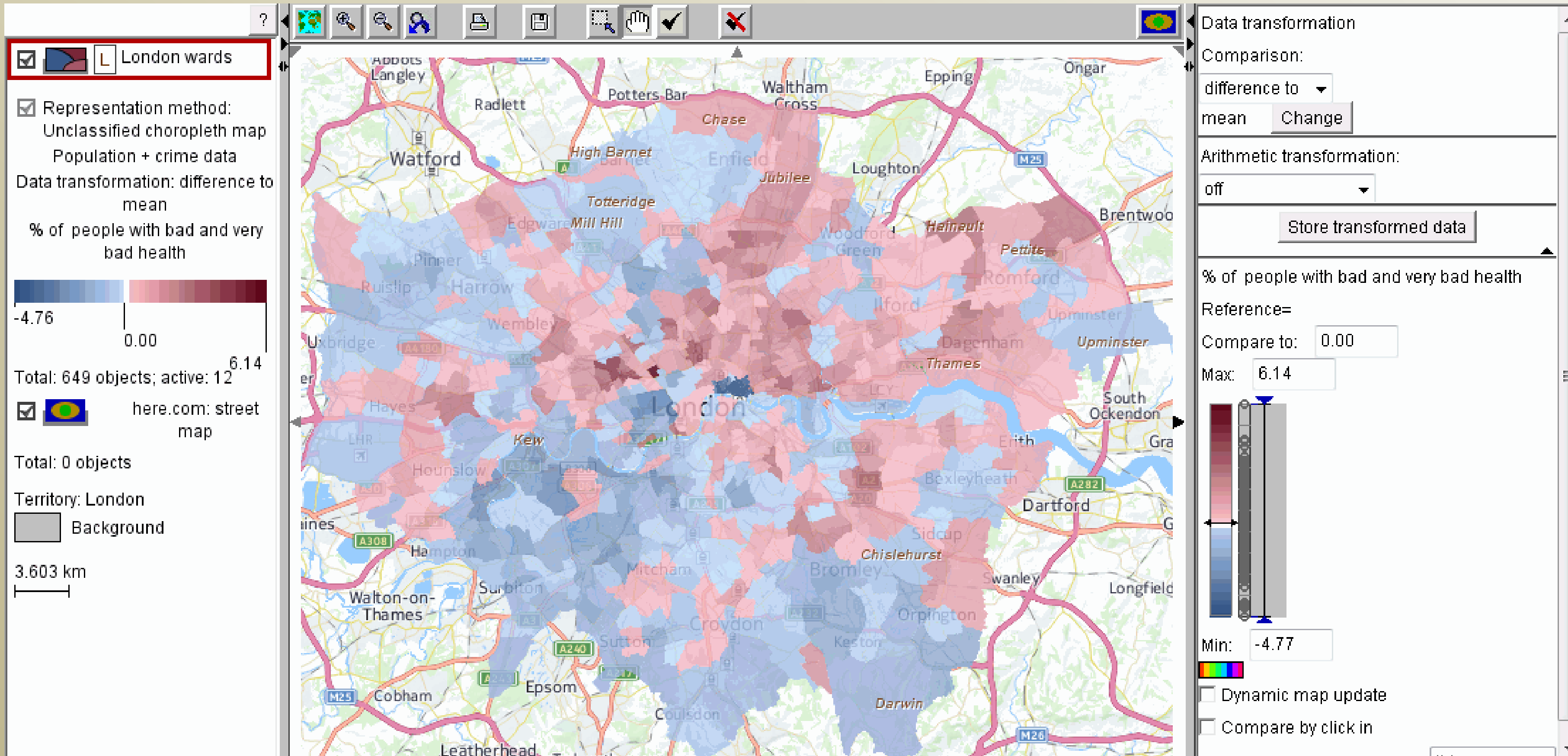
When all bars can be seen simultaneously, the display enables an overall view of the temporal distribution of the referents' existence times and can thereby support synoptic tasks. However, this is possible for a relatively small number of distinct references. The bars need to be sorted according to the existence times.

Choropleth map



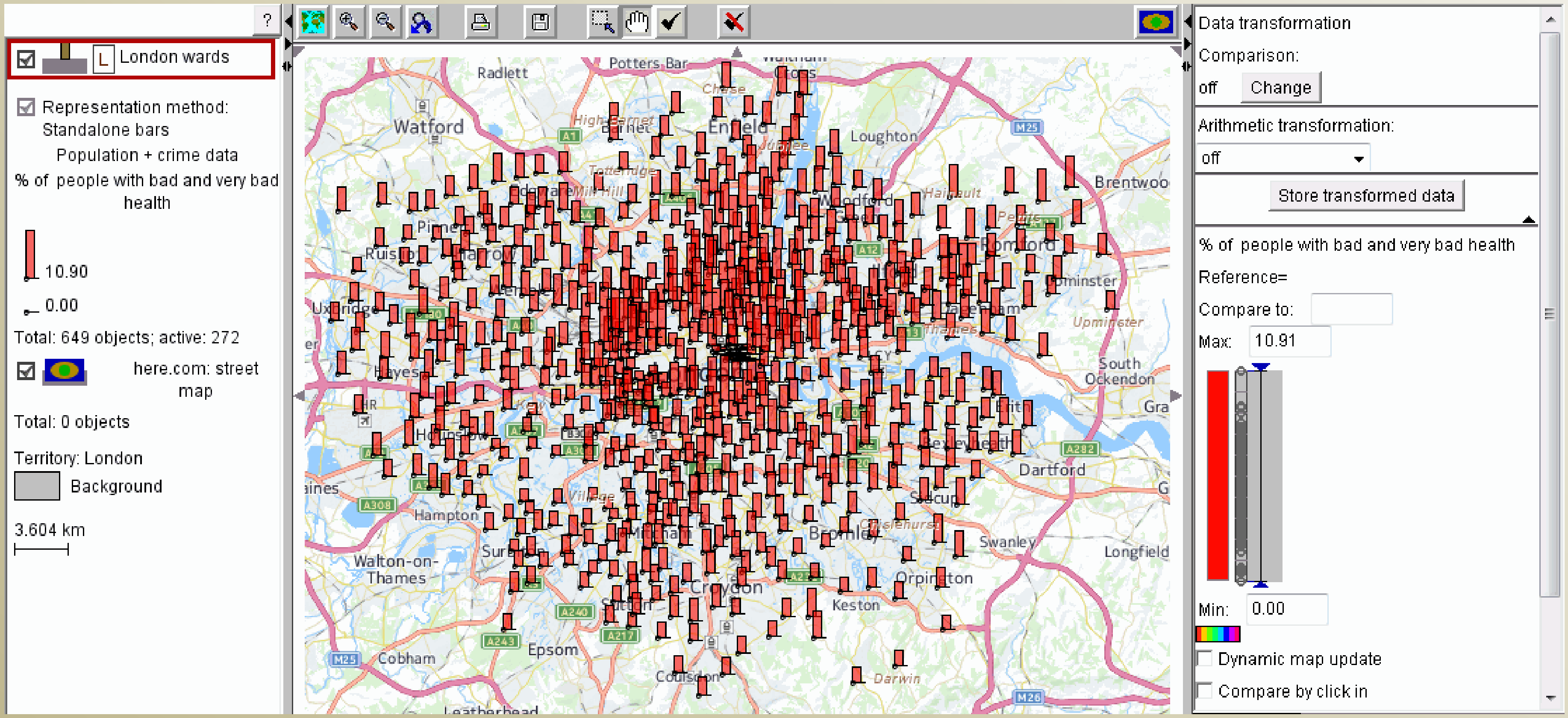
Values of a numeric attribute are represented by the visual variable ‘colour value’. Darker shades correspond to higher attribute values. The shades are used for painting areas or objects on the map. The map is perceived as a single image and thus supports synoptic tasks w.r.t. space.

Choropleth map: diverging colour scale

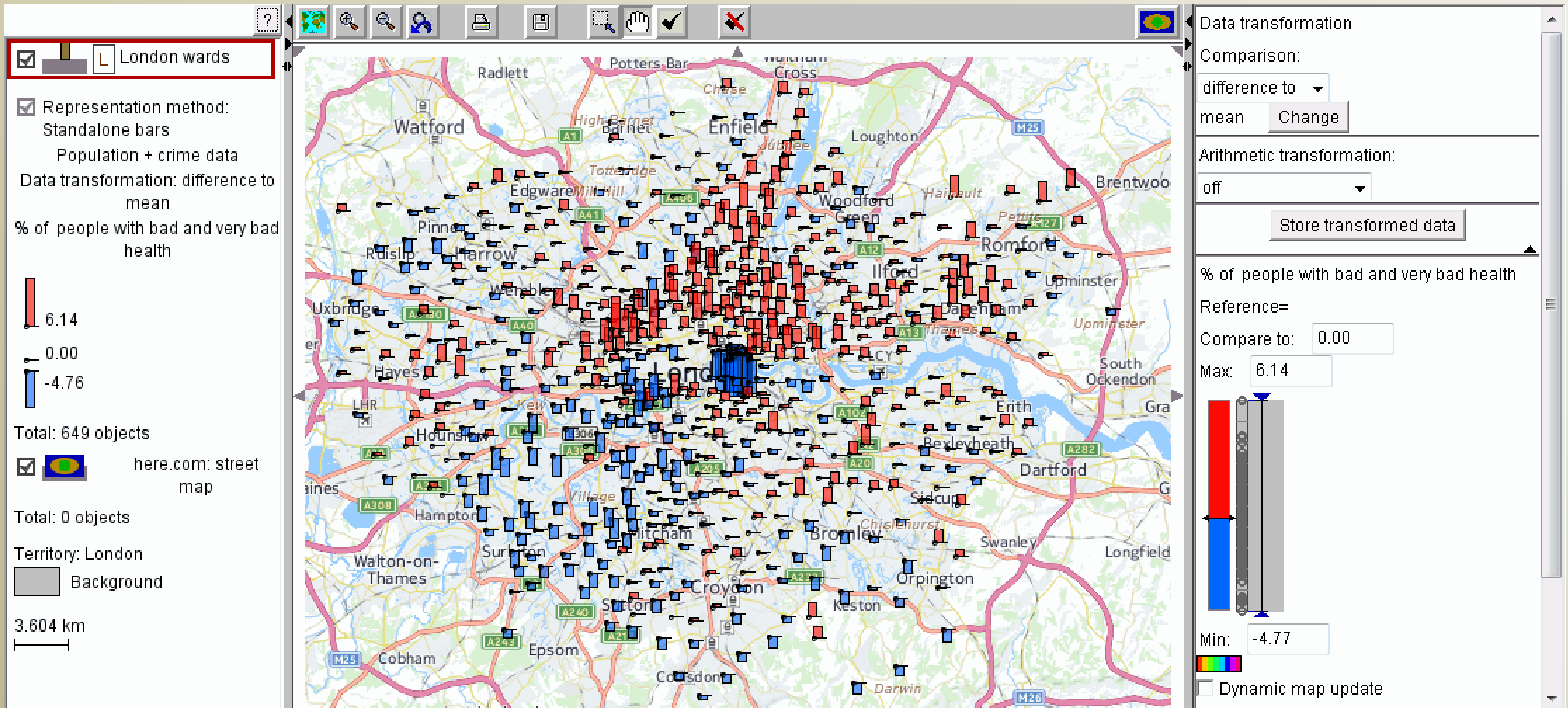


A diverging scale of colour values uses two distinct colour hues for representing positive and negative values or positive and negative differences from a chosen central value, such as the overall mean. Darker shades correspond to larger differences. The map is perceived as a single image and exposes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

Map with proportional symbols (bars)

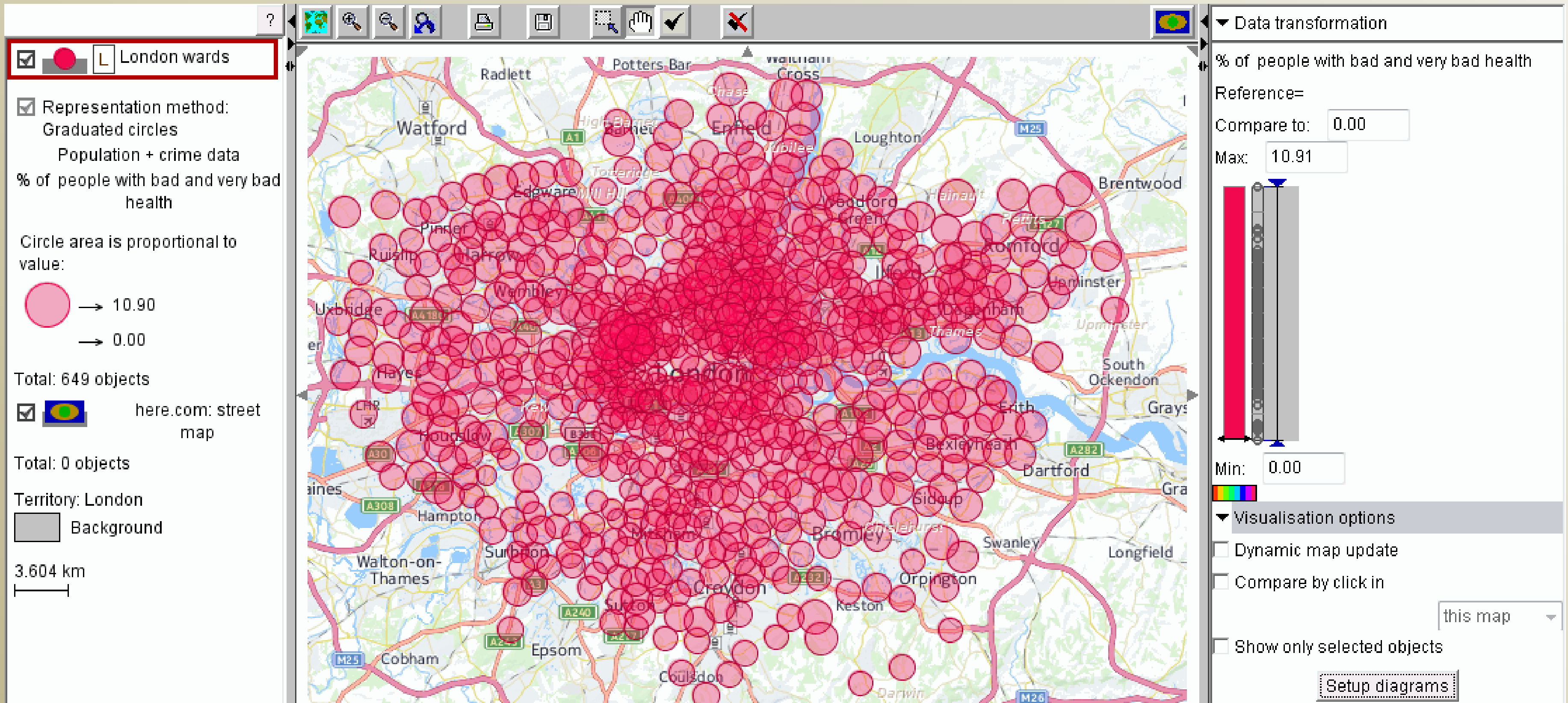


Map with proportional symbols (bars)



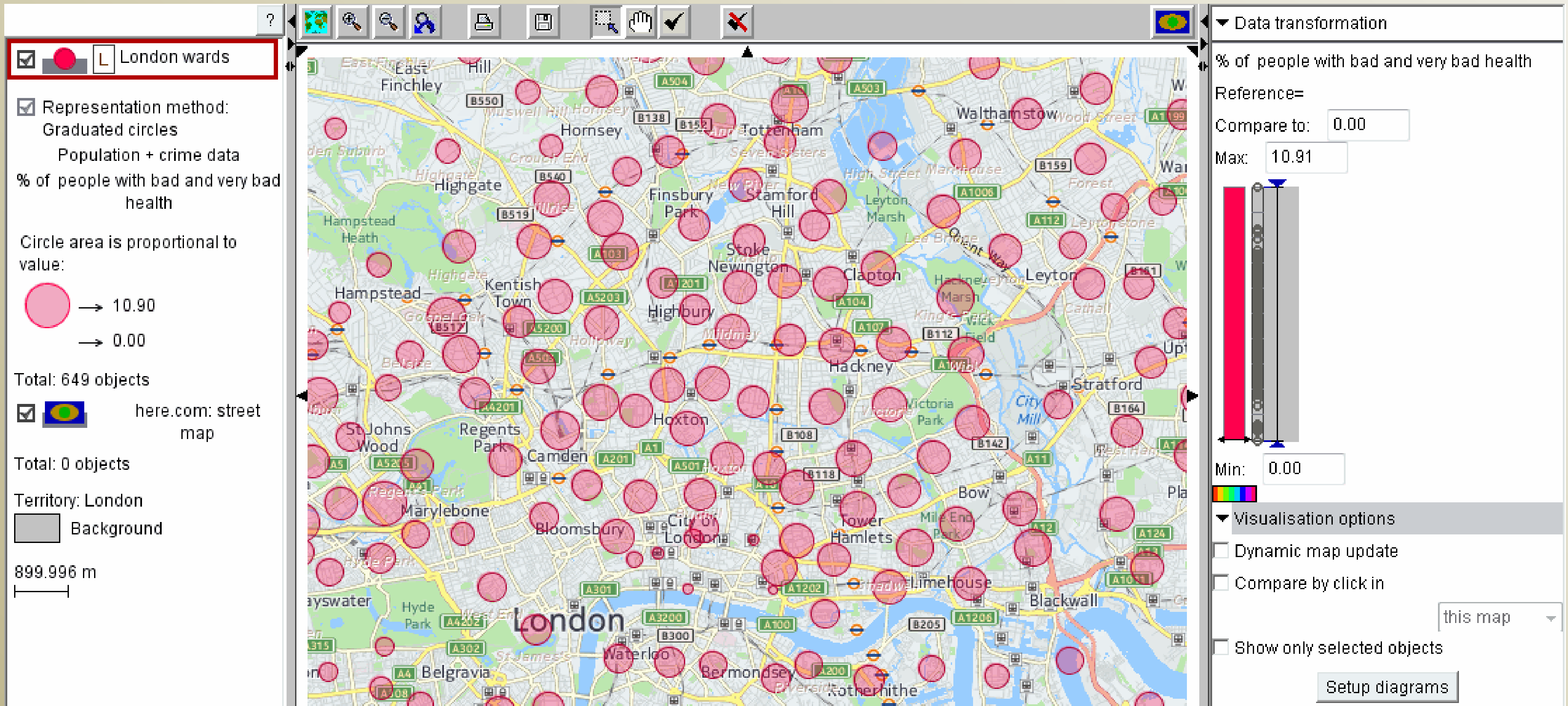
Diverging bars use orientation (upward and downward) and, complementarily, two distinct colour hues for showing positive and negative attribute values or differences from a selected central value, such as the overall mean.

Map with proportional symbols (circles)



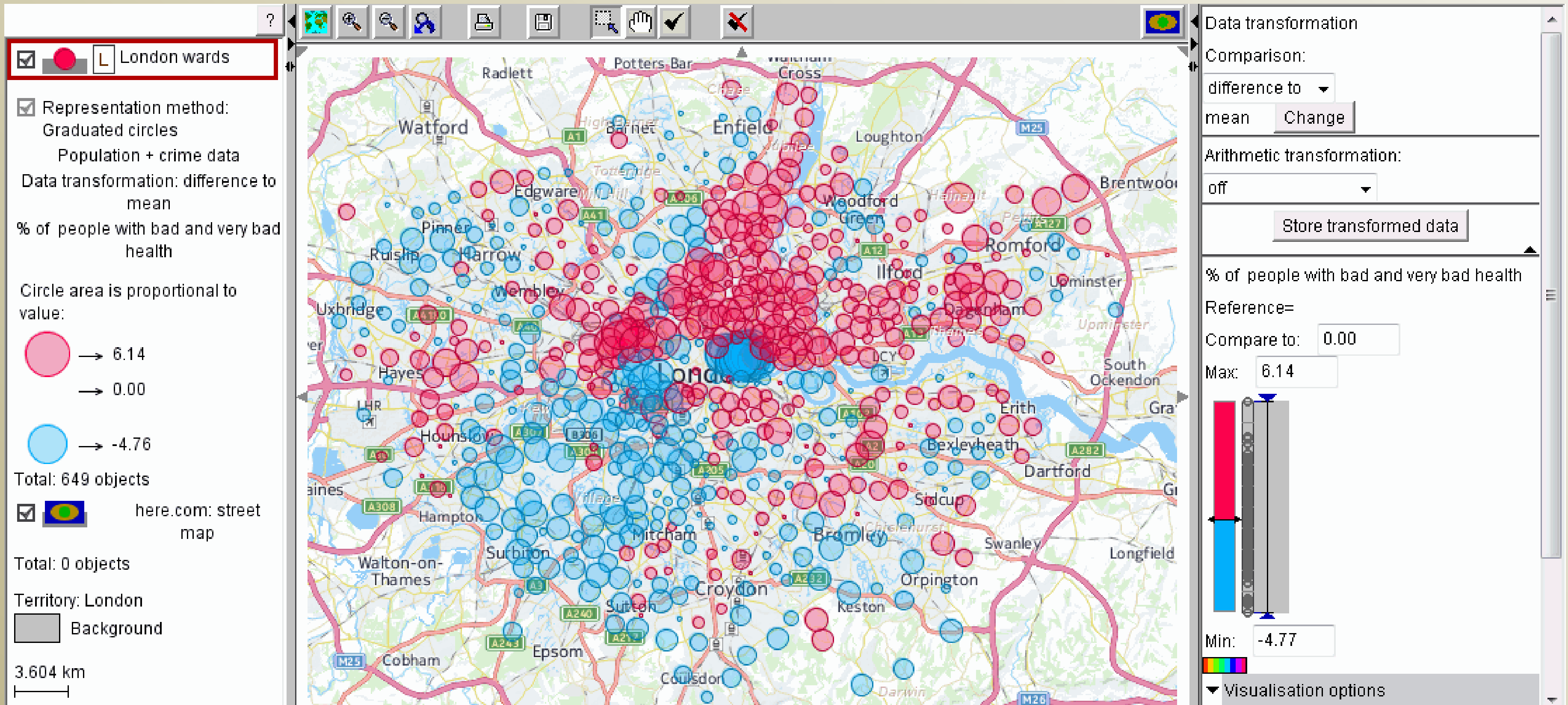
Problem: too much visual clutter; requires zooming for discerning the symbols.

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).

Maps with proportional symbols vs. choropleth maps

- Studies of human perception show that people can better estimate numeric values represented by symbol sizes than represented by colour shades.

  ⇒ It may seem that maps with proportional symbols are more effective than choropleth maps.

- However, a choropleth map has serious advantages:
  - 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.
- Accurate estimation of numeric values is only required for elementary tasks, which can be supported by interactive techniques.
- Both choropleth maps and proportional symbol maps become more expressive and effective when diverging scales are applied.

Area cartograms

Geographic regions are transformed so that their sizes become proportional to their population or some other demographic attribute, such as income or disease incidence.

from http://www-personal.umich.edu/~mejn/cartograms/population1024x512.png

Dorling, D. (1995):
A New Social Atlas of Britain.
London: John Wiley and Sons.

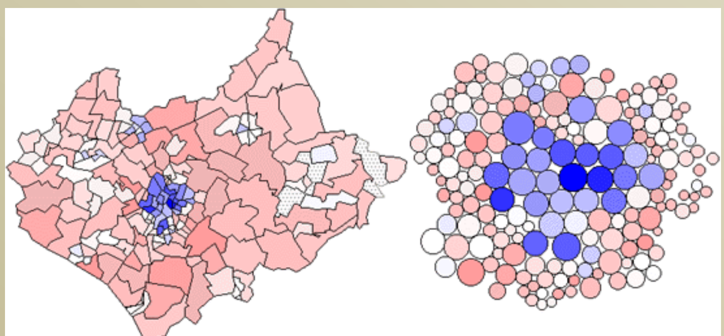
Michael T. Gastner and M. E. J. Newman (2004): Diffusion-based method for producing density-equalizing maps.
Proc. Nat. Acad. Sci. USA, 101, 7499-7504

Area cartograms: pros and cons

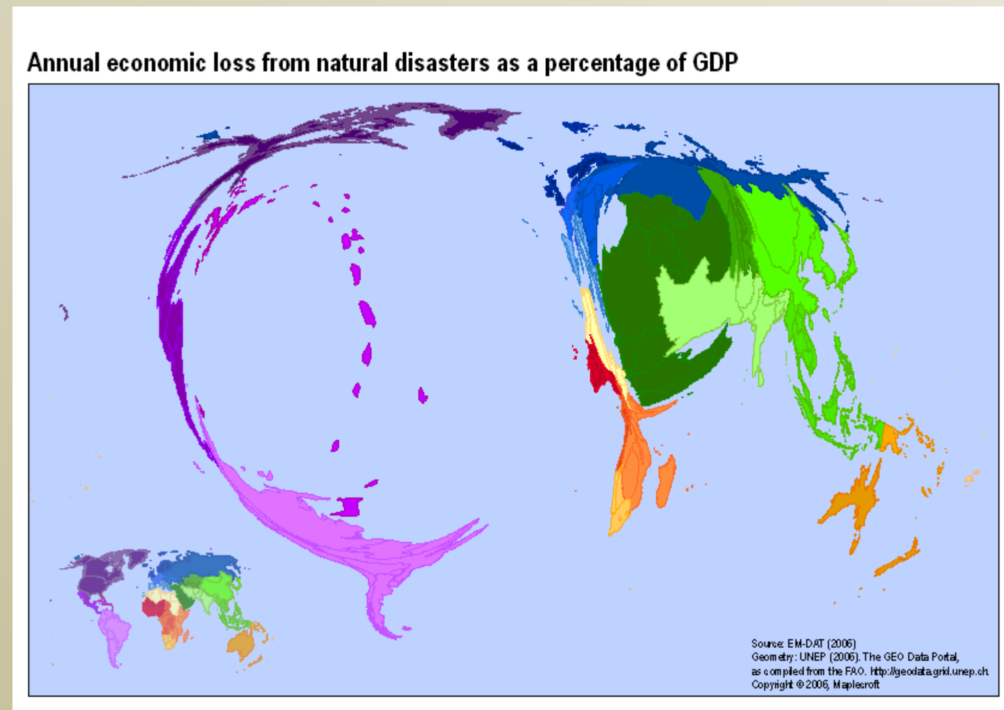
The principle of isomorphism of the display space to the physical space is violated.

+ Can be very impressive

– but only when you know well the true sizes and shapes



– Even well-known regions may be hardly recognisable due to the distortions

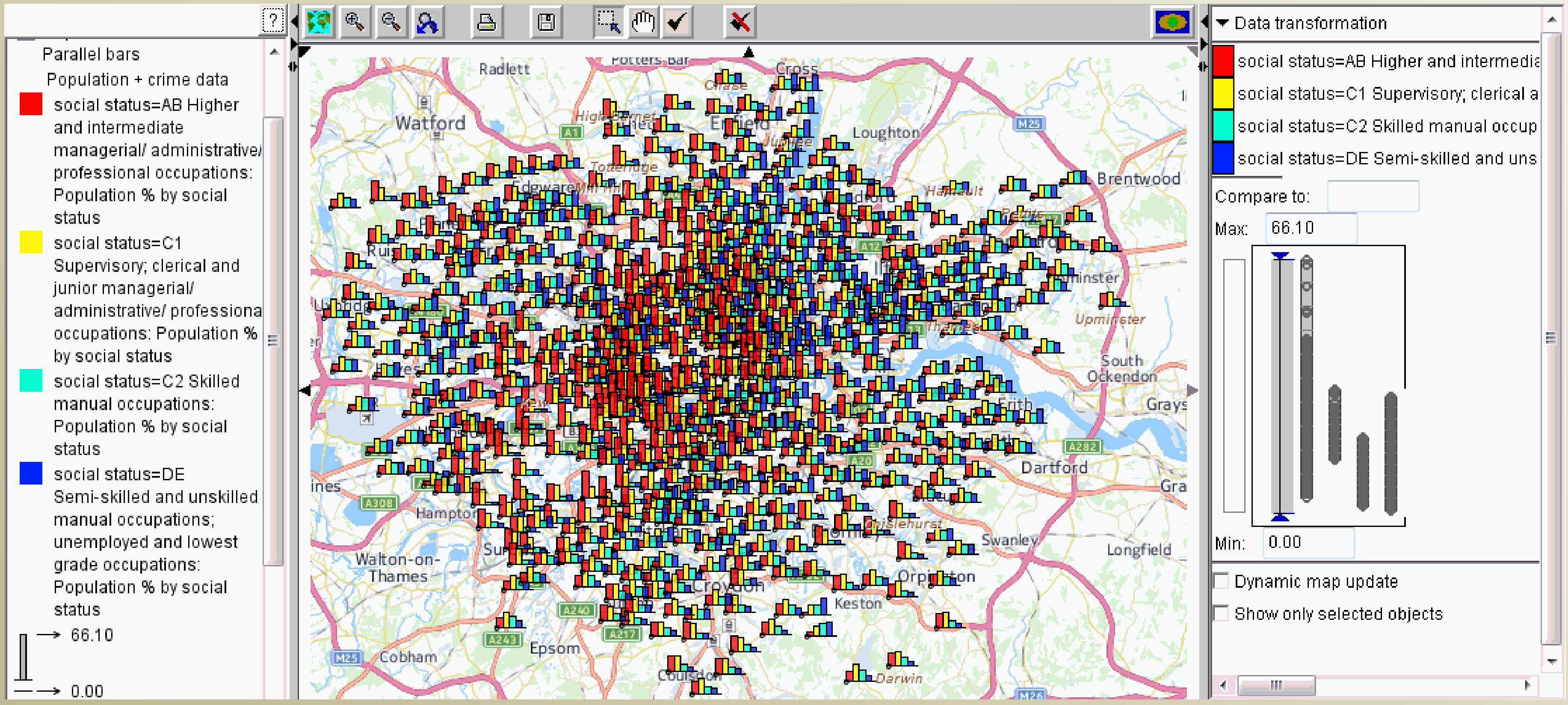


+ Can provide a background for visualising another attribute, as on a choropleth map, and enable the analyst to relate the two attributes.

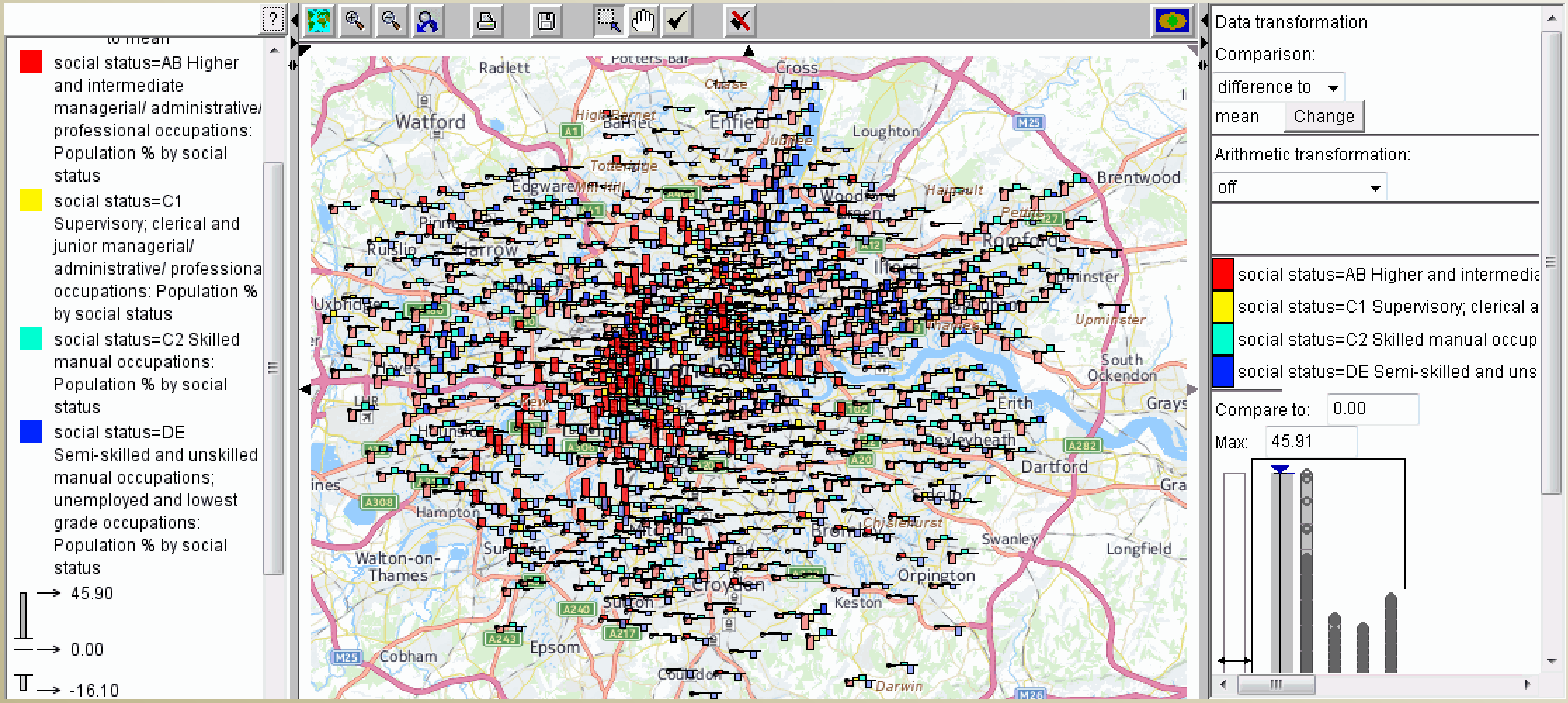
– It is hard to relate any of them to the true geographic space.

⇒ It may be a good means for expressive communication of some message but not a tool for analysis.

Maps with diagrams (bar diagrams)

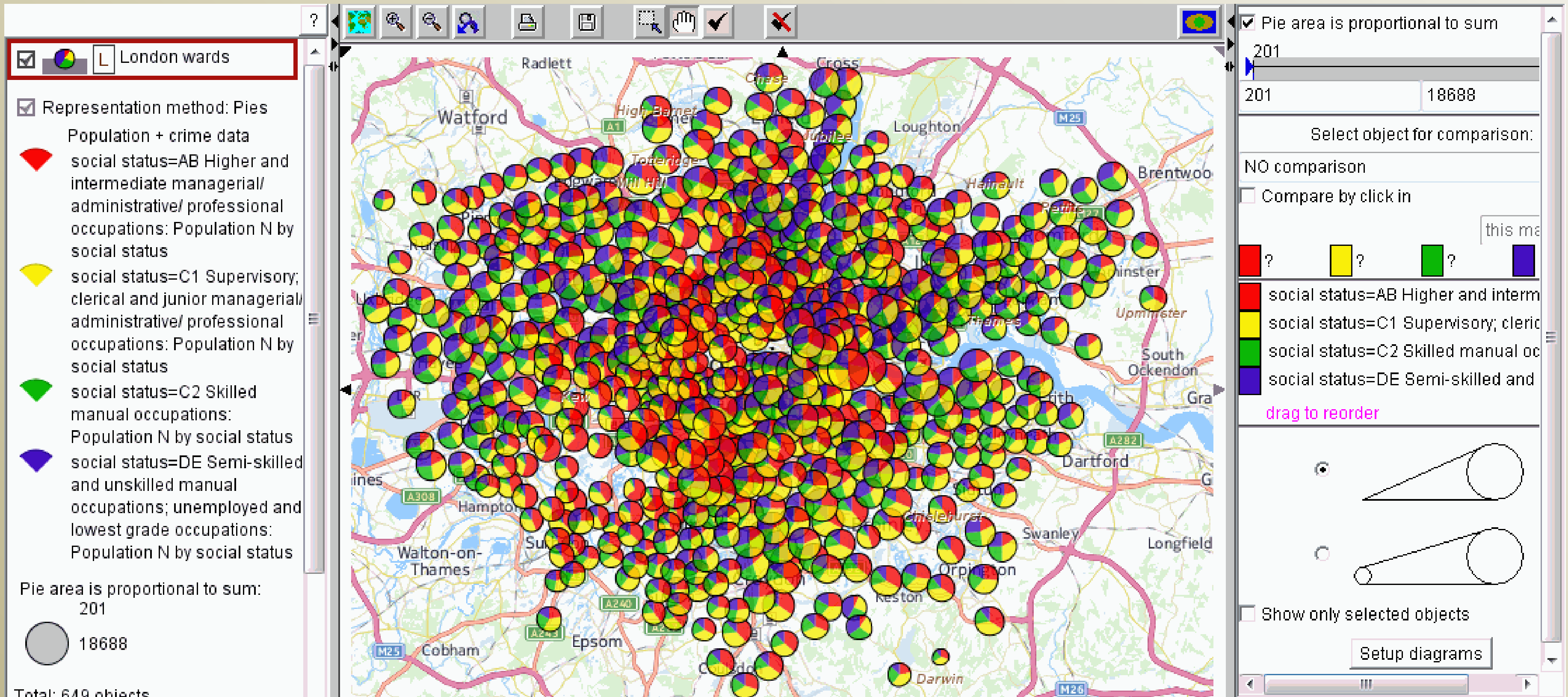


Map with bar diagrams



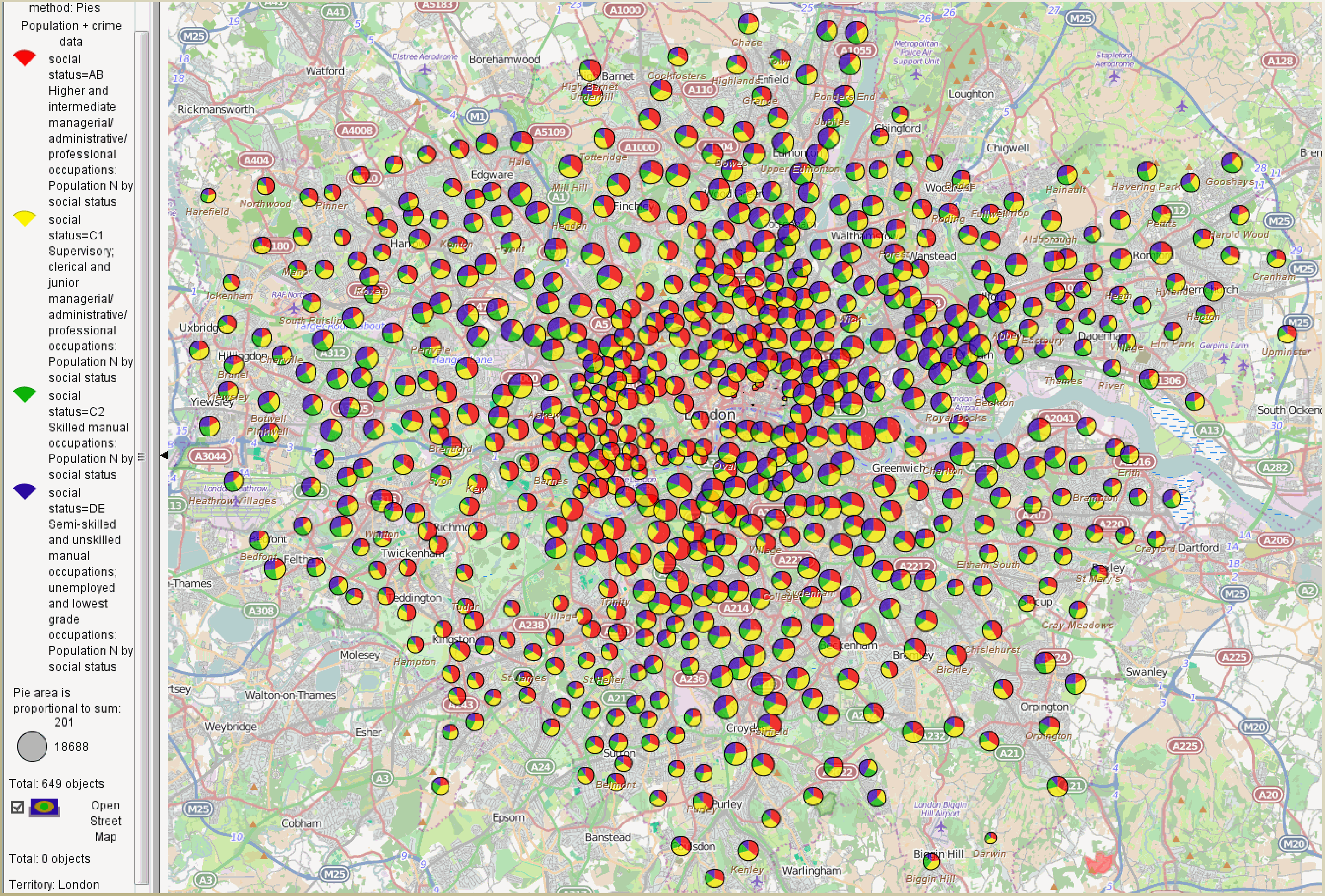
Diverging bars show the differences of the attribute values from the respective means. The bars showing negative differences are painted in less saturated colours.

Map with diagrams (pie charts)

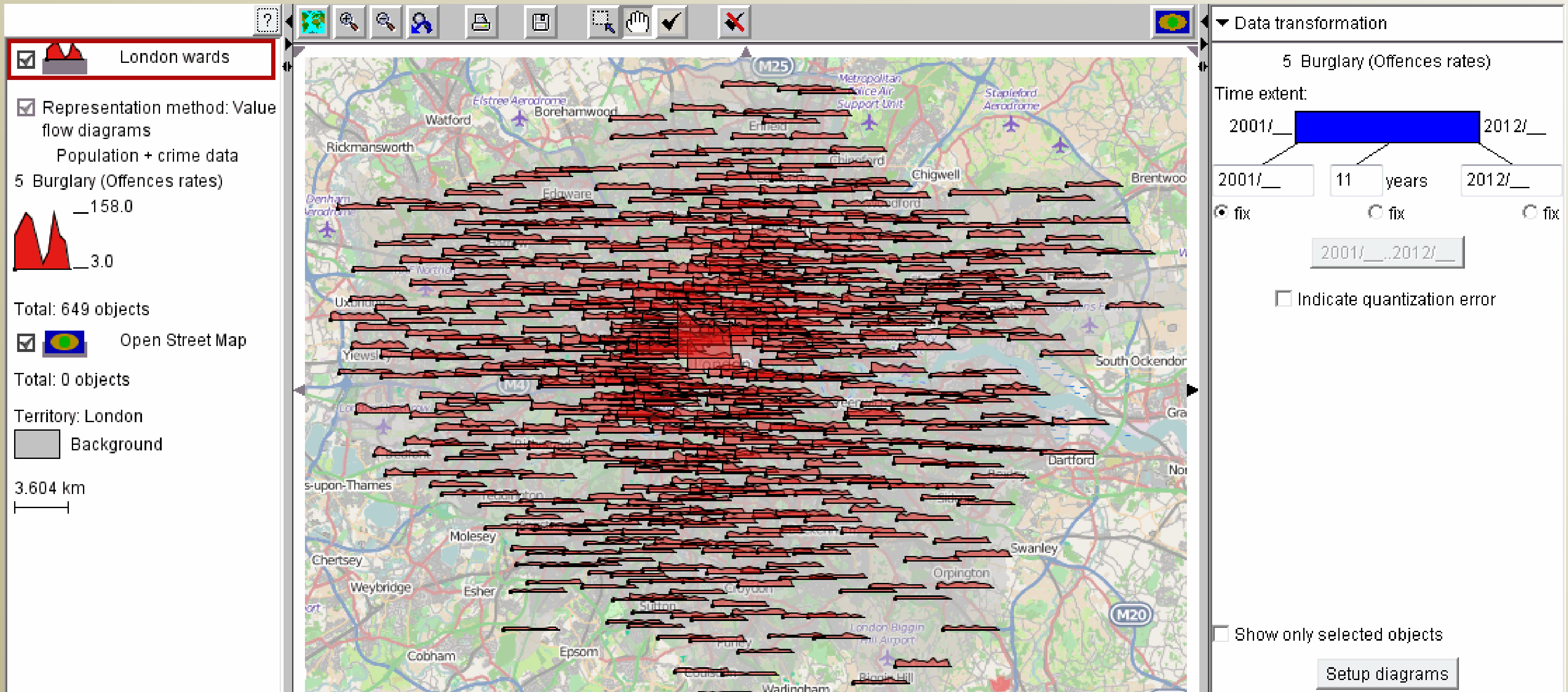


- Too much visual clutter.
+ Still, areas with the prevalence of particular colours are detectable.
• The method is applicable to attributes that make together a meaningful sum. Here: population N by social status.

Map zooming reduces the clutter

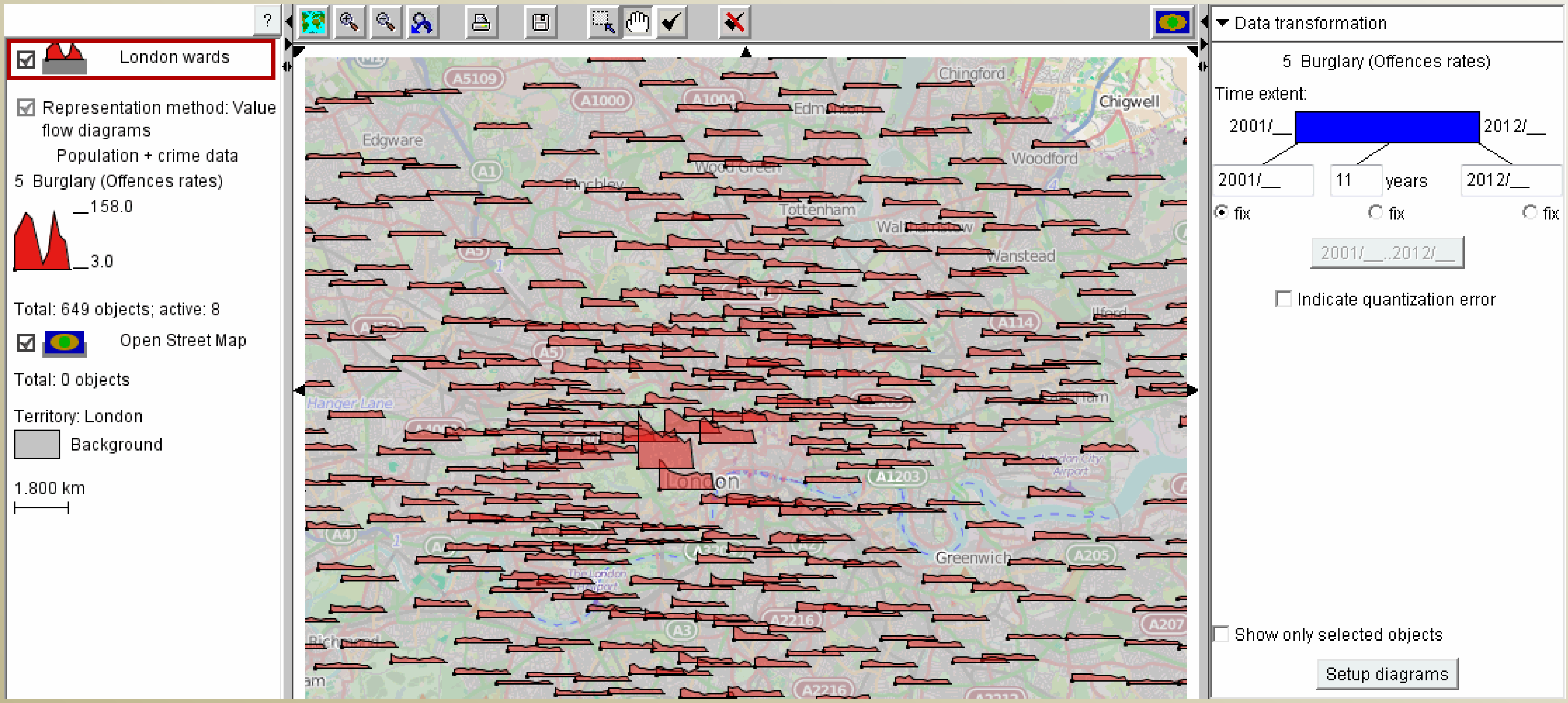


Map with "value flow" diagrams

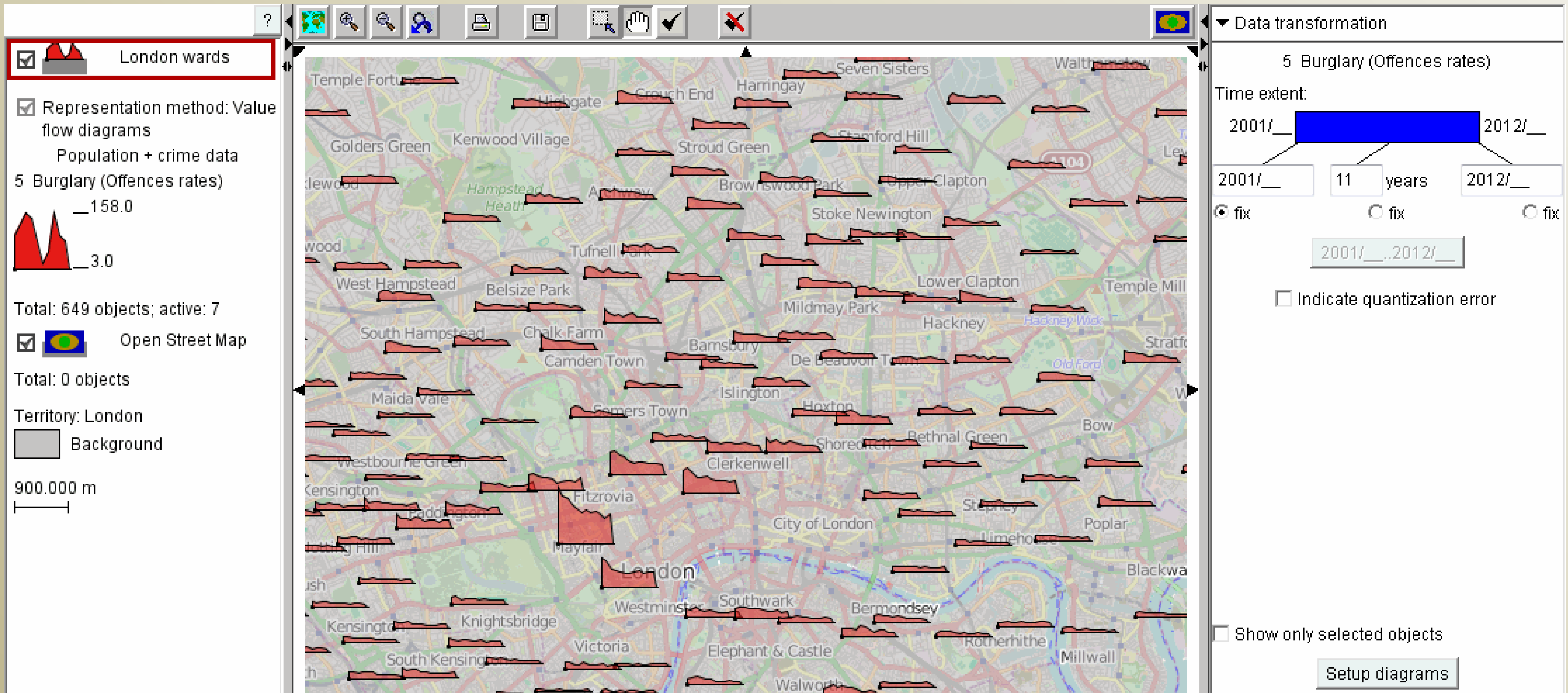


The diagrams represent the variation of attribute values over time. This method can be used for data with 2 referrers: space and time. However, such a map does not enable an overall view and therefore does not support synoptic tasks. It also suffers from visual clutter, which complicates the analysis.

Map with “value flow” diagrams

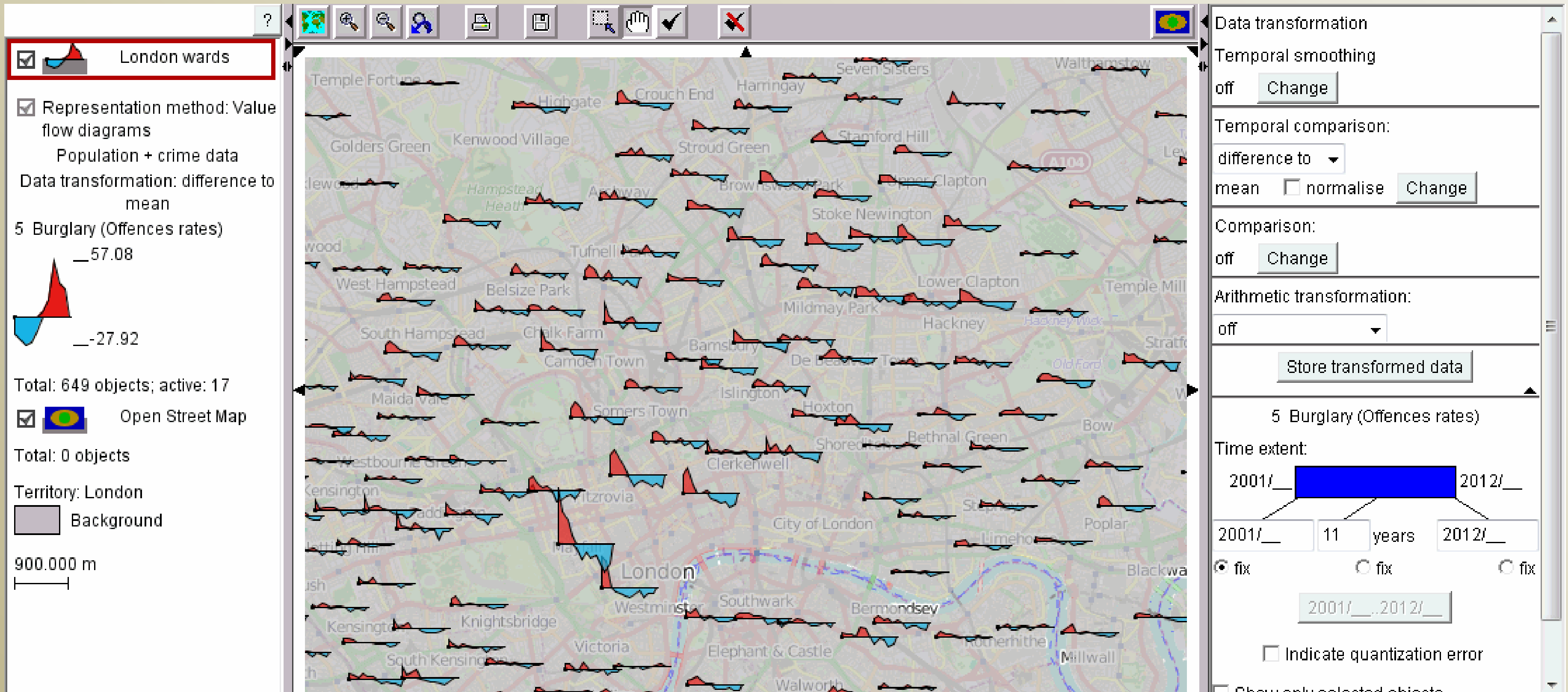


Map with “value flow” diagrams



Only at this zoom level, the diagrams can be easily discerned.

Map with “value flow” diagrams



The technique of diverging scale can also be applied to value flow diagrams. This can work well when the diagrams do not overlap.

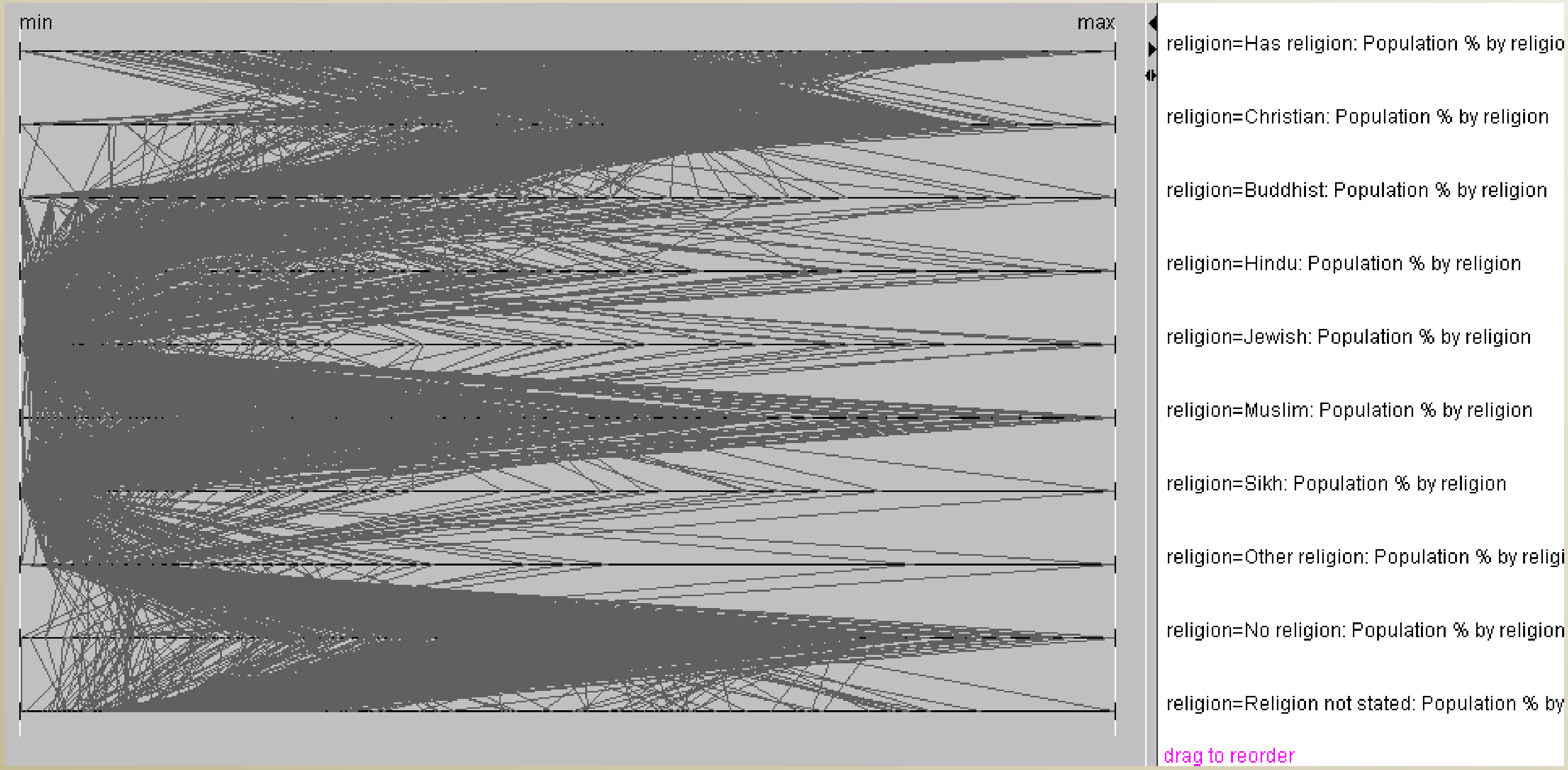
Maps with diagrams: general notes

- In principle, can be used for representing the spatial variation of multiple numeric attributes or spatio-temporal variation of a single numeric attribute.
- However, diagram maps are prone to visual clutter, which complicates the perception and analysis.

  ⇒ In practice, diagram maps can be used when data refer to a small number of places (requiring a small number of diagrams).
- Even in absence of visual clutter, a diagram map does not provide a single image (cannot be perceived all at once) and therefore does not support synoptic tasks.

Parallel coordinates plot

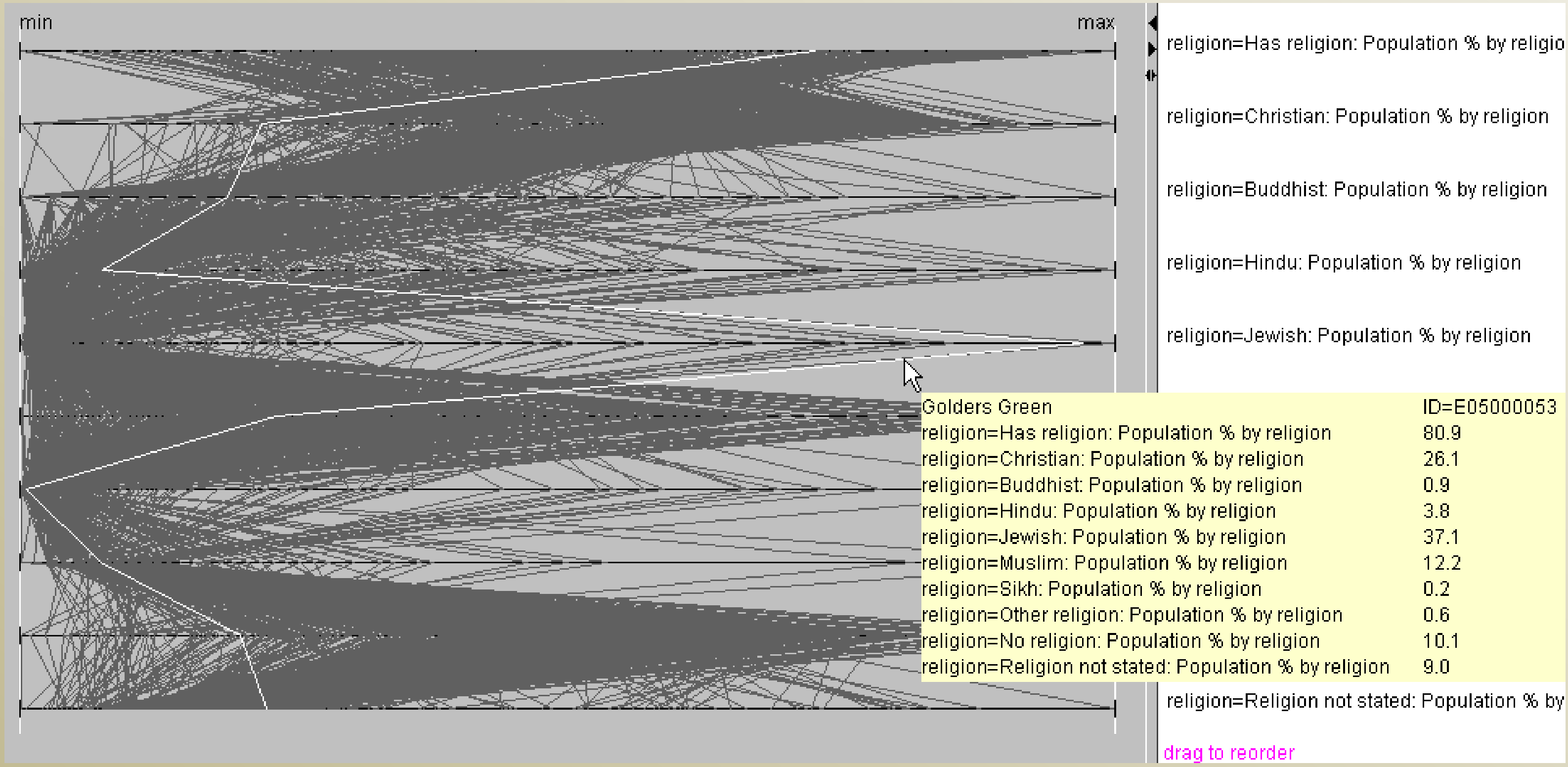
not so common but popular in visualisation research community



Multiple parallel axes are used for multiple attributes. Attribute values are represented by positions on the respective axes. For each reference, there is a polygonal line connecting the positions on neighbouring axes representing the corresponding attribute values.

Parallel coordinates plot

not so common but popular in visualisation research community



A parallel coordinate plot in its basic form supports neither synoptic nor elementary tasks. Various enhancements have been suggested by visualisation researchers (consideration would require a separate lecture).

Common display types: general notes

- All display types have their limitations.
- Display types that can support synoptic tasks:
  - ± Bar chart: for a relatively small number of references (all bars need to be visible simultaneously); sorting is essential
  - ✓ Scatter plot
  - ✓ Line graph (single curve or a few curves)
  - ± Gantt chart: same as bar chart
  - ✓ Choropleth map
  - ± Maps with proportional symbols: for a relatively small number of spatial references (such that visual clutter can be avoided); diverging scales are helpful.
- Complex data may require multiple complementary displays, interactive techniques, and computational processing.

Questions?

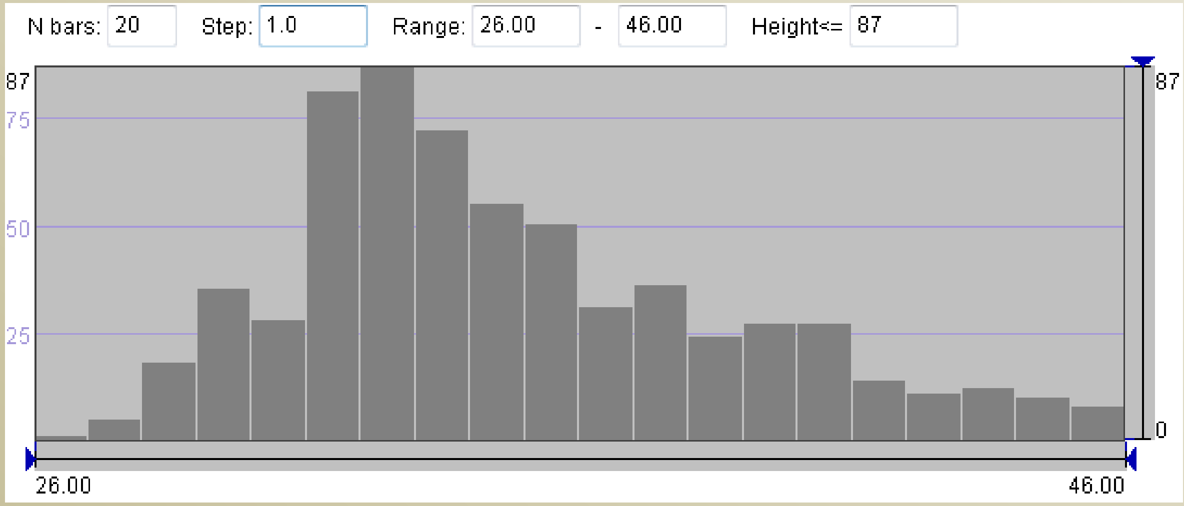
Types of visual display

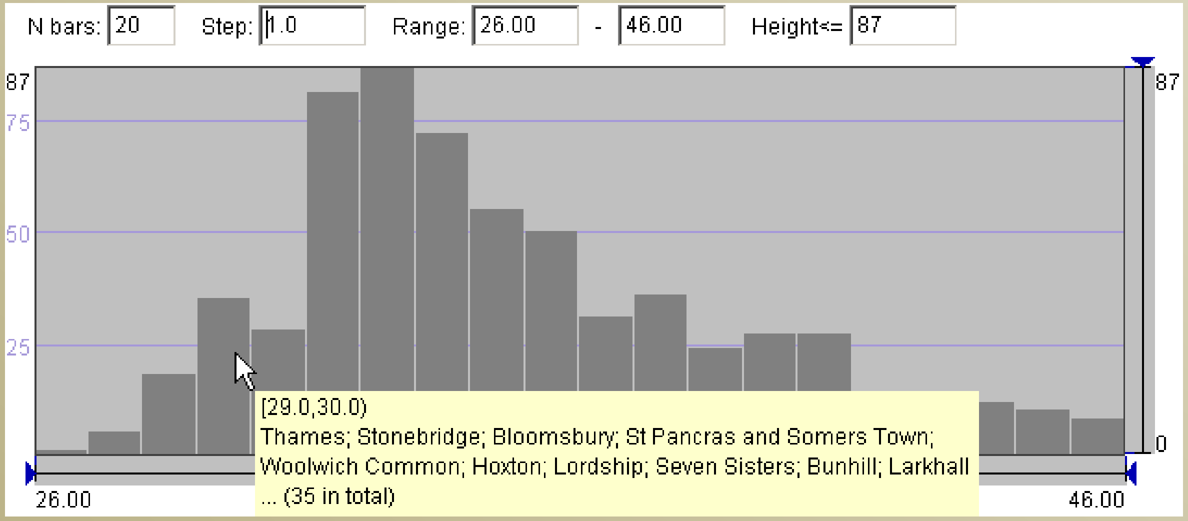
Displays of aggregated data

Aggregated vs. detailed data displays

- In detailed data displays (considered so far), marks correspond to individual references. This imposes limitations and creates problems for analysis:
  - Display size may be insufficient to include marks for all references ⇒ no overall view, no support to synoptic tasks.
  - Display may be cluttered and hardly readable.
  - Even when marks are not too numerous, some of them may overlap (e.g., in a scatter plot).
- In aggregated data displays, marks correspond to groups (subsets) of references.
  - ✓ Groups may be arbitrarily large ⇒ displays are scalable to large amounts of data.
  - ✓ Synoptic tasks can be well supported.
  - \- No support to elementary tasks.

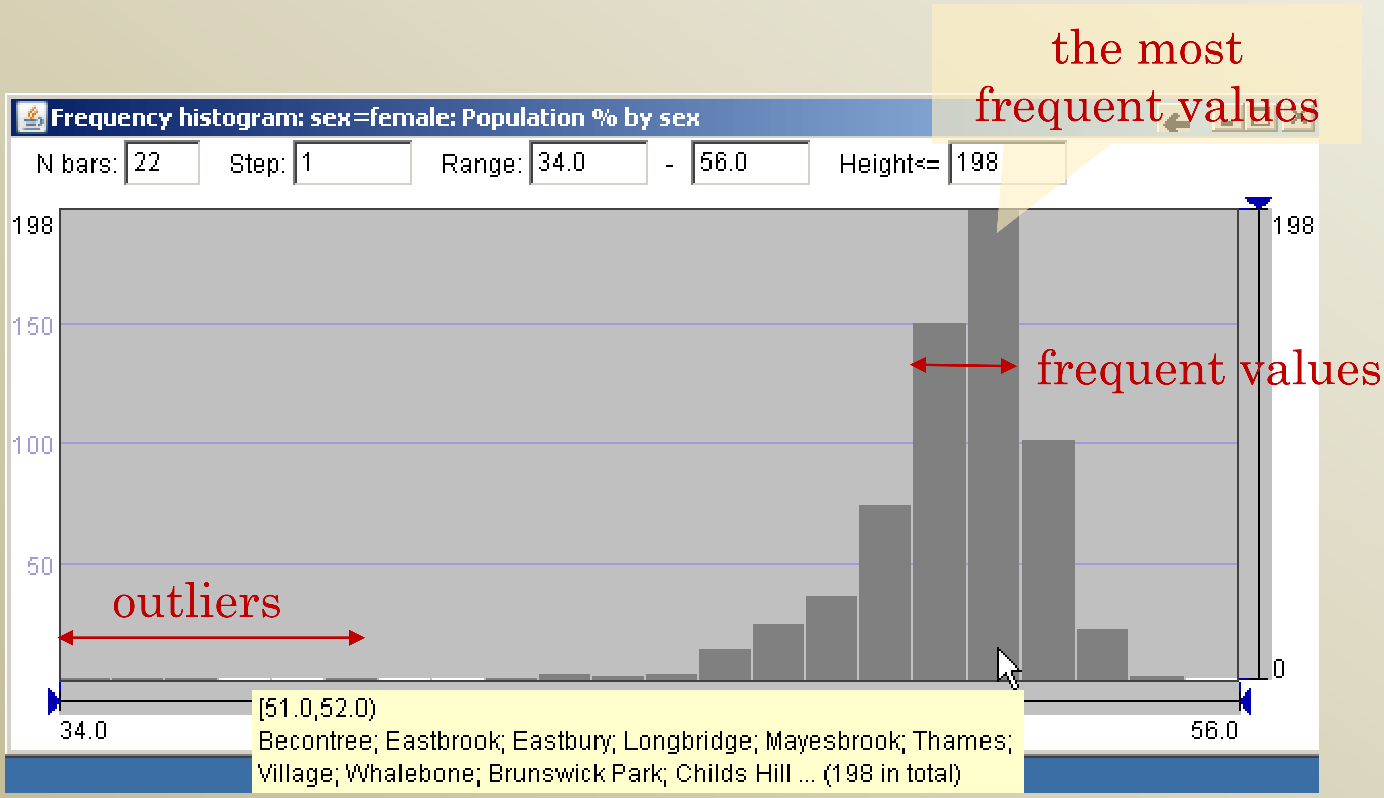
Frequency histogram



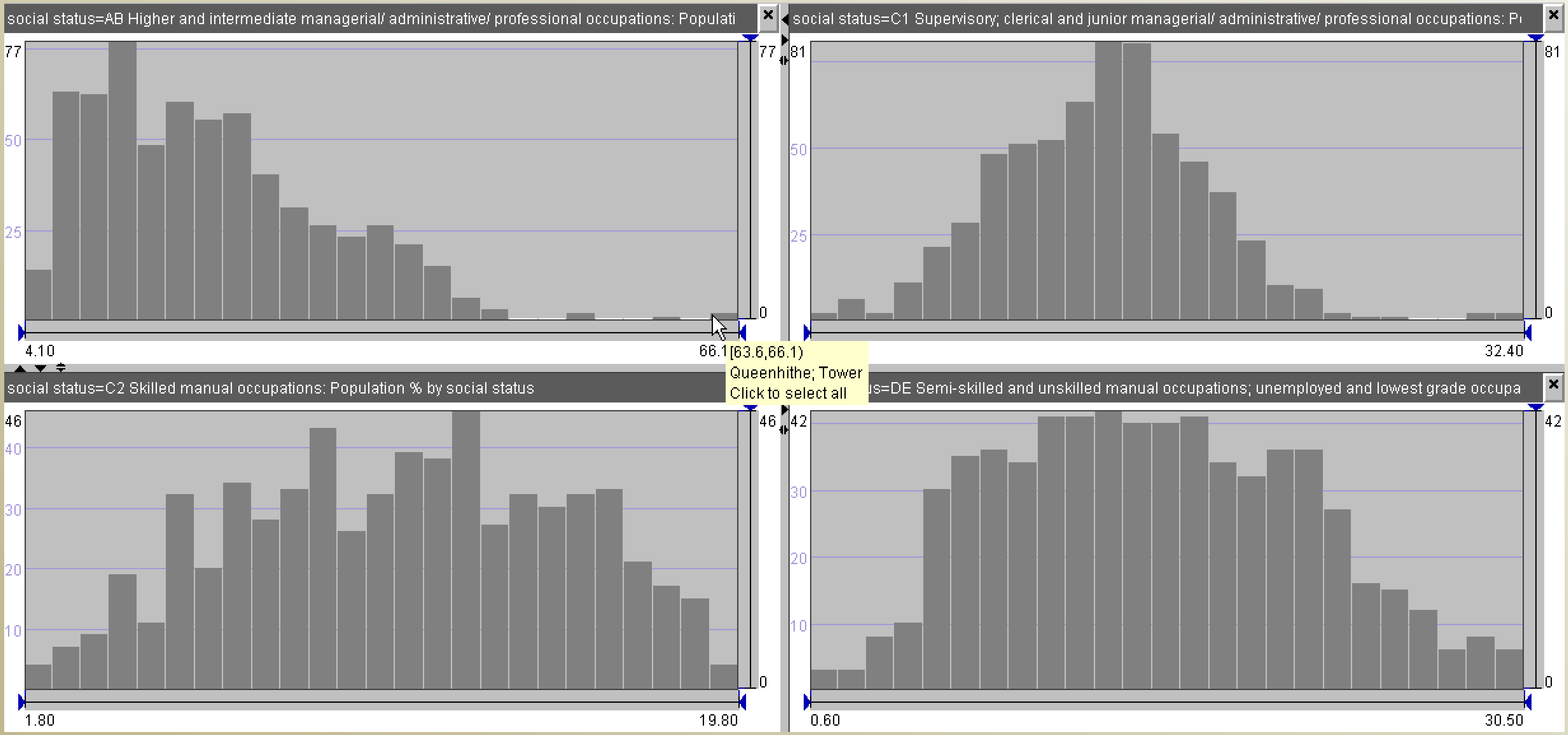


Shows the distribution of numeric attribute values over the set of references. Each bar represents a group of references such that the corresponding attribute values lie within a certain interval, which is represented by the horizontal position of the bar. The height of the bar is proportional to the group size (number of references).

Frequency histogram

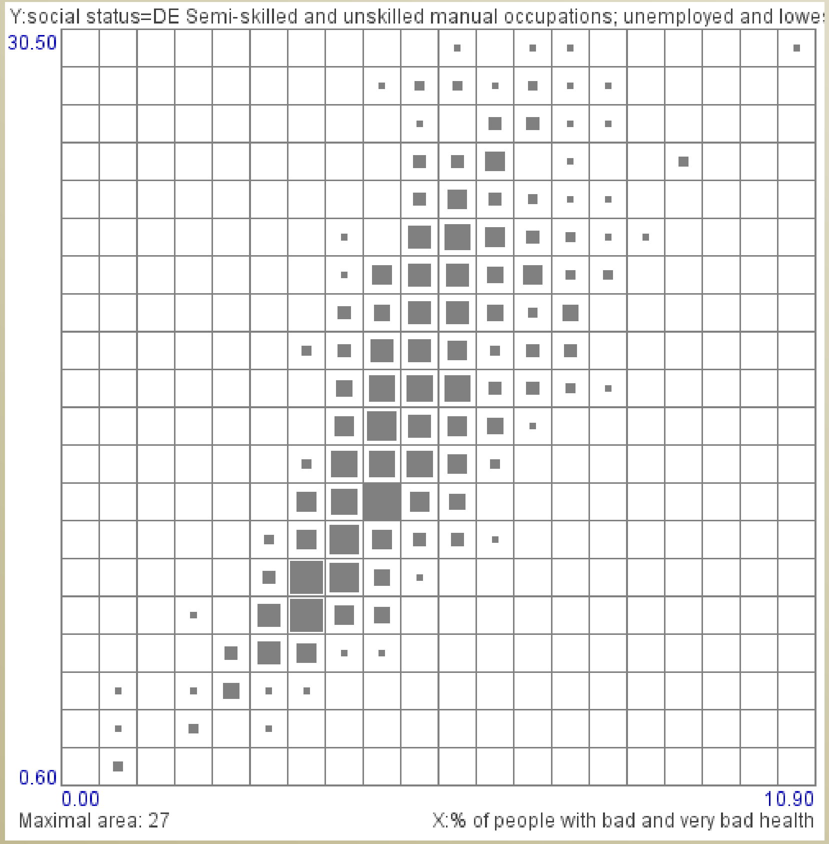


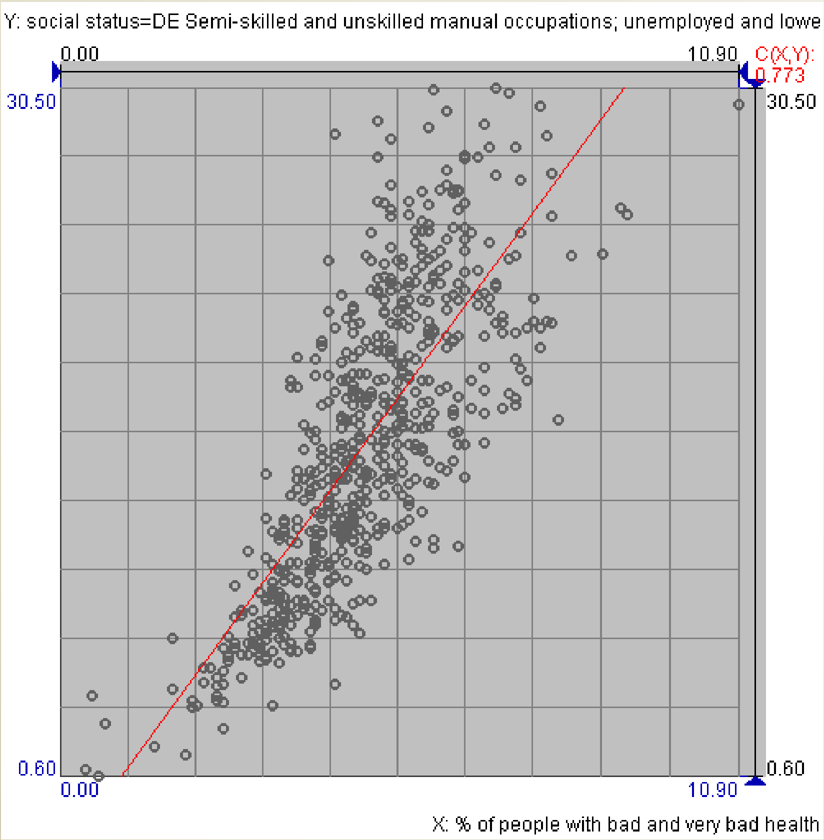
Multiple frequency histograms



Value distributions of several attributes can be compared using several frequency histograms. Here: the histograms represent the same attributes we tried to look at using bar diagrams (slide 50). The comparison of the distributions is easier with the histograms.

2D histogram (binned scatter plot)

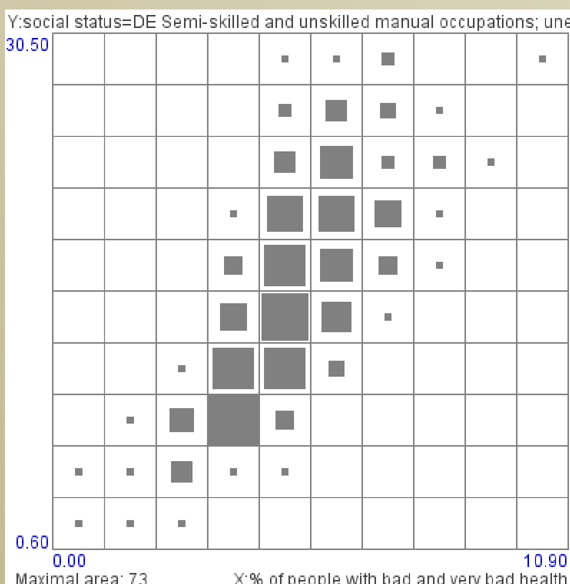


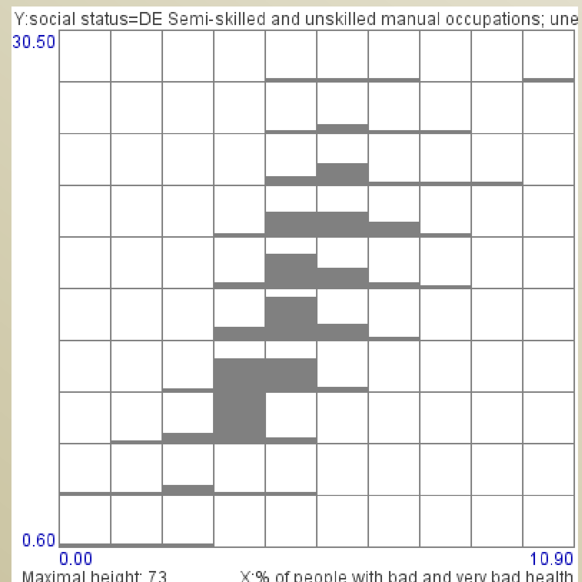


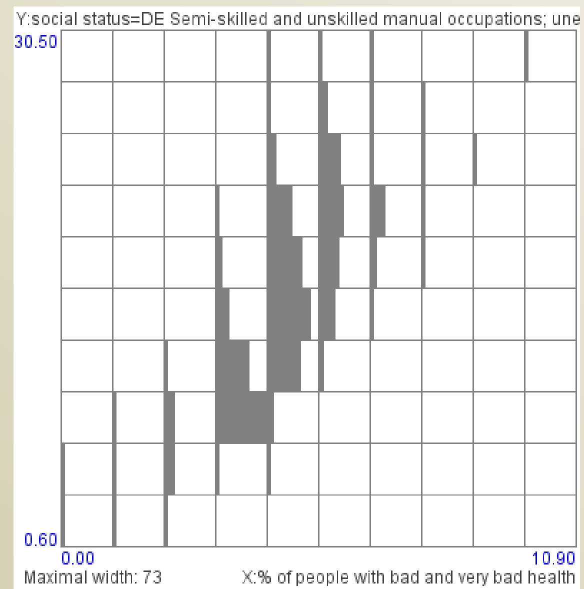
Like in a scatter plot, two dimensions represent two numeric attributes. The value ranges are divided into intervals, which creates a rectangular grid over the plot area. The sizes of the symbols within the grid cells are proportional to the numbers of references having the attribute values within the respective intervals.

What can a 2D histogram show us?

- Relationships (such as correlation) between the attributes
- Frequencies of value combinations

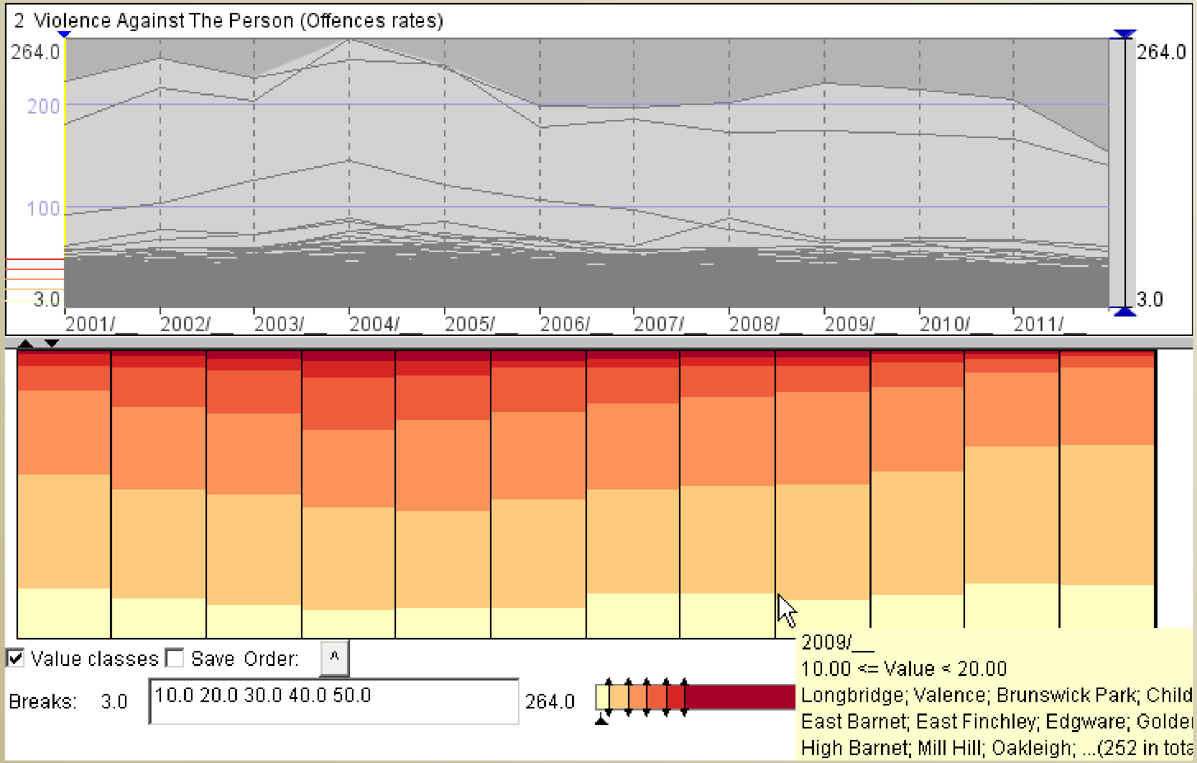






- For different value intervals of one attribute, how the values of the other attribute are distributed.
  - Multiple 1D histograms that are easy to compare

Segmented bars (time histogram)

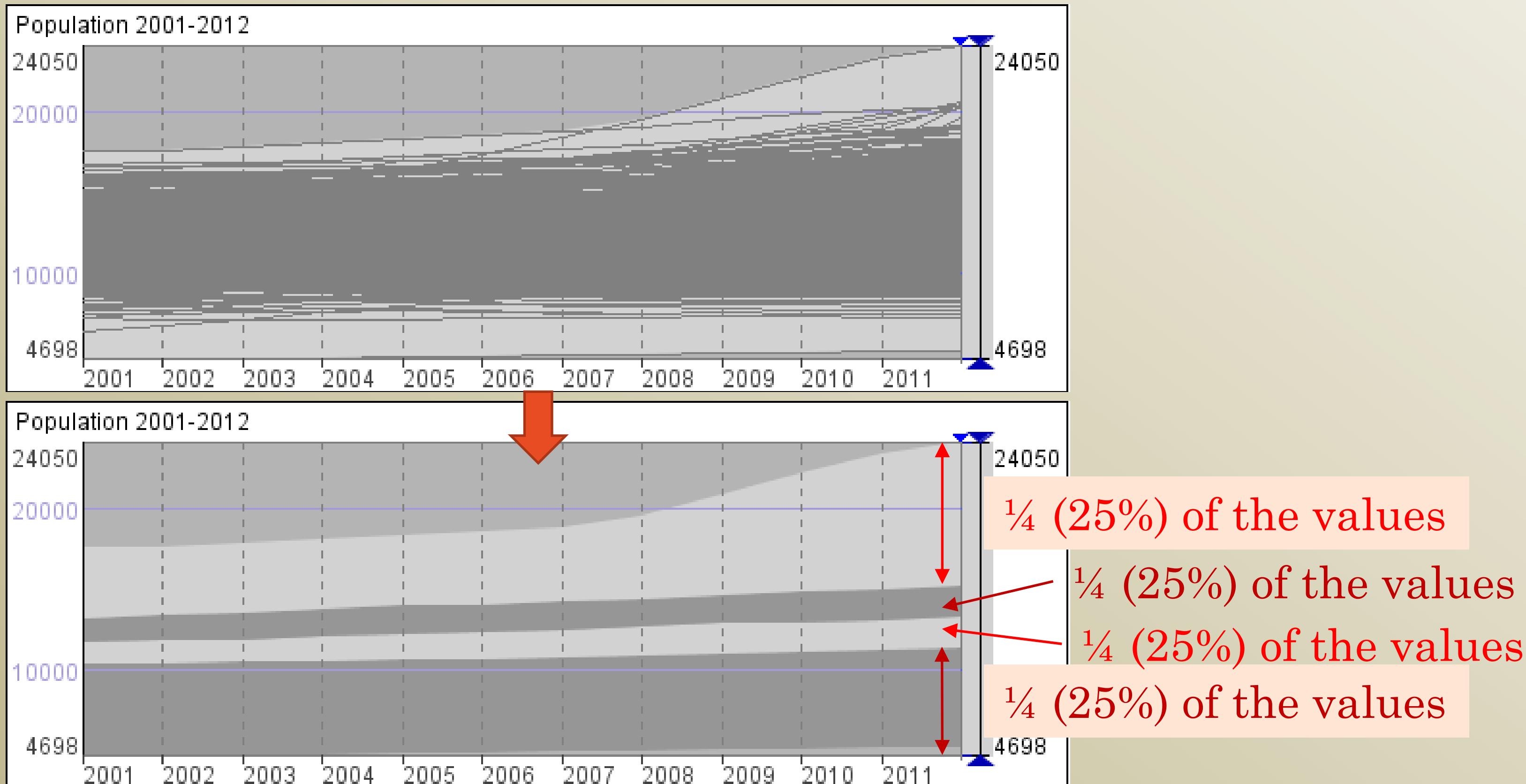


This technique is applicable to data with two referrers: one linearly ordered (such as time) and one arbitrary.

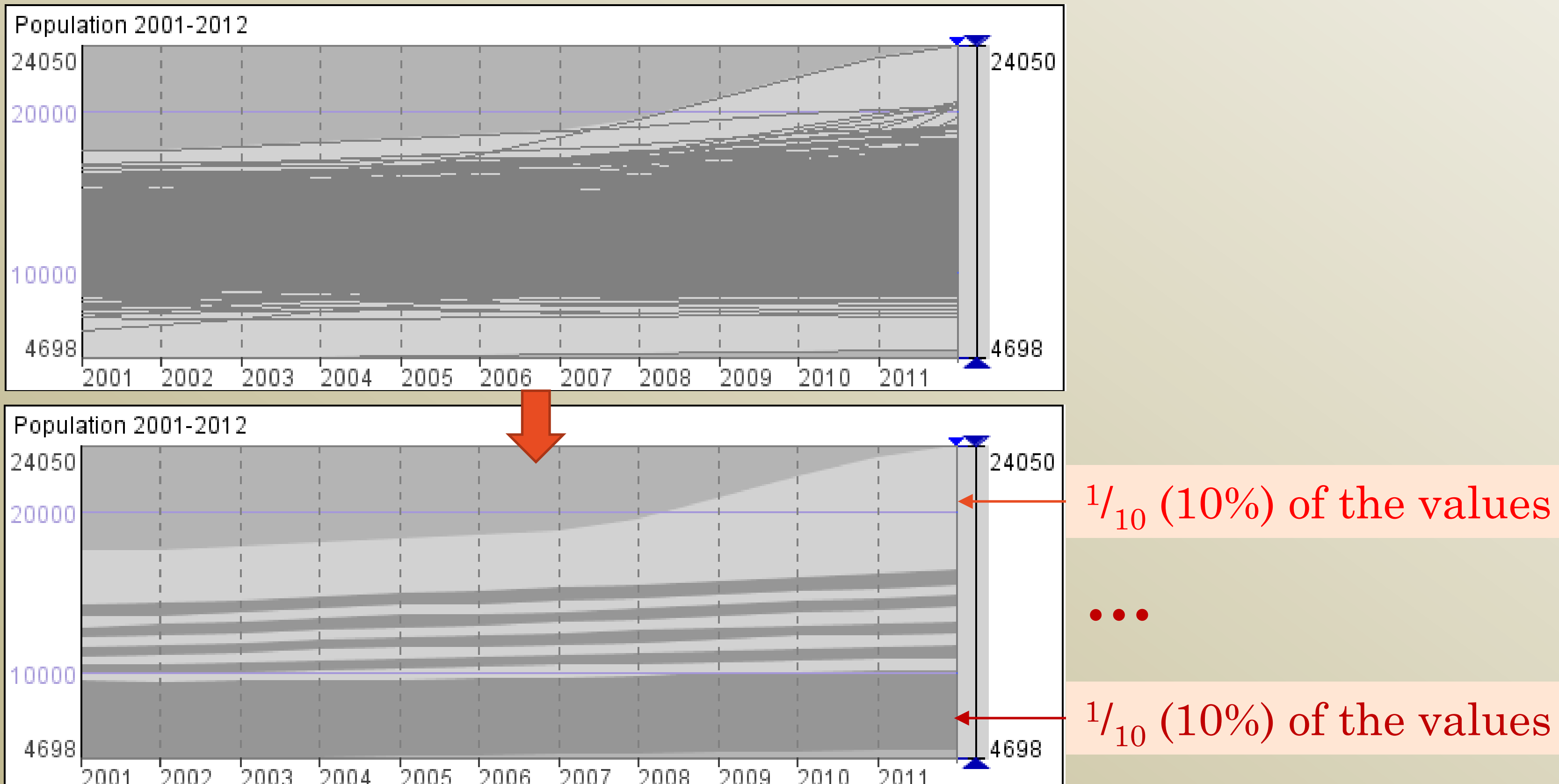
Note that a line graph with multiple curves may be unreadable due to line overlapping.

Each bar stands for one time step (generally, one value of referrer 1). The value range of the attribute is divided into intervals. The height of each segment is proportional to the number of values of referrer 2 for which the values in this time step lie within the respective interval. The intervals are represented by the segment colours.

Quantile graph

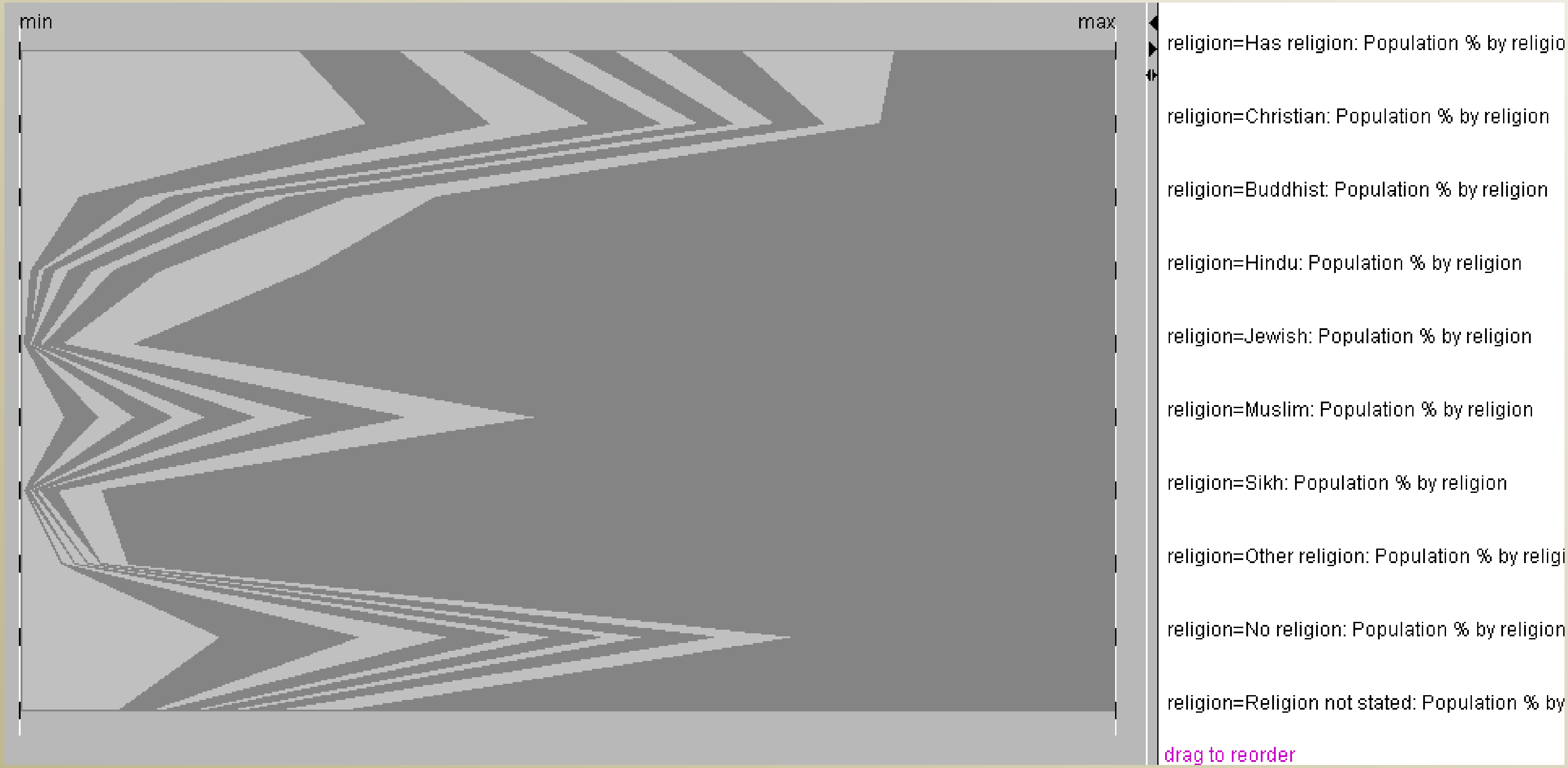


Quantile graph



Segmented bars and quantile graphs show how the value distribution w.r.t. one referrer (set of objects, places, etc.) varies over the range of linearly ordered values of another referrer (such as time).

Quantile parallel coordinates plot



The idea of representing value quantiles can also be applied on a parallel coordinate plot.

Aggregated data displays: general notes

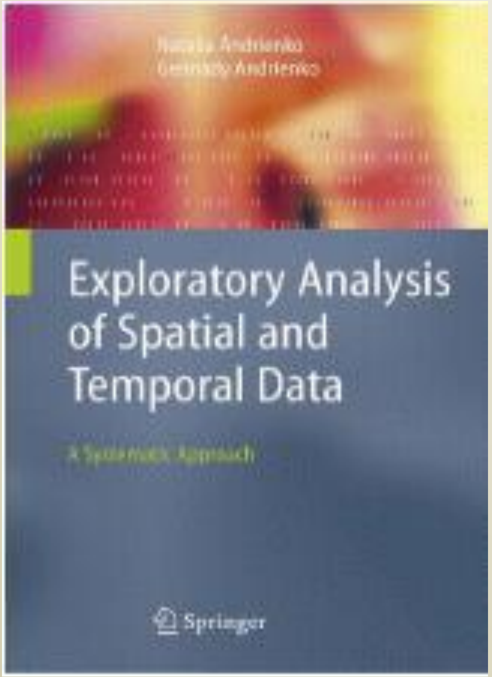
- Aggregated data displays solve particular problems of detailed data displays: insufficient display size for available data, visual clutter, and overlapping of marks.
- Aggregated data displays can better support synoptic tasks but do not support elementary tasks.
- Interactive operations are necessary for both classes of data display.
  - To be considered next.

Questions?

Displays of aggregated data

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

Section 4.3
Visualisation in a Nutshell