Typology, transformations, and clustering of spatio-temporal data

Lecture given by
prof. Gennady Andrienko and
prof. Natalia Andrienko

Content

- Types of spatio-temporal data: spatial events, spatial time series, and trajectories of moving objects.
- Transformations between the data types.
- Partition-based and density-based clustering.
- Two possible perspectives for looking at spatial time series and two complementary ways of applying partition-based clustering to them.
- Real-time detection of event concentrations

Types of spatio-temporal data

Data with spatial and temporal components

Spatial events

- A **spatial event** is a physical or abstract entity that appears at a certain time at a certain location in space.
  - **Instant event**: appears but does not exist any longer
    - or the existence time is negligibly small or out of interest for analysis
  - **Durable event**: exists during some time interval
- Examples
  - Instant: lightning flash, photo taken, tweet posted
  - Durable: election, stormy weather, New Year celebration
- Data structure: < *event identifier**, spatial position, time of appearing, *time of disappearing or duration of existence***, *any attributes*>

  * May not be given explicitly; it is assumed that each data record describes one event

  ** Need to be specified only for durable events

Example: posted tweets

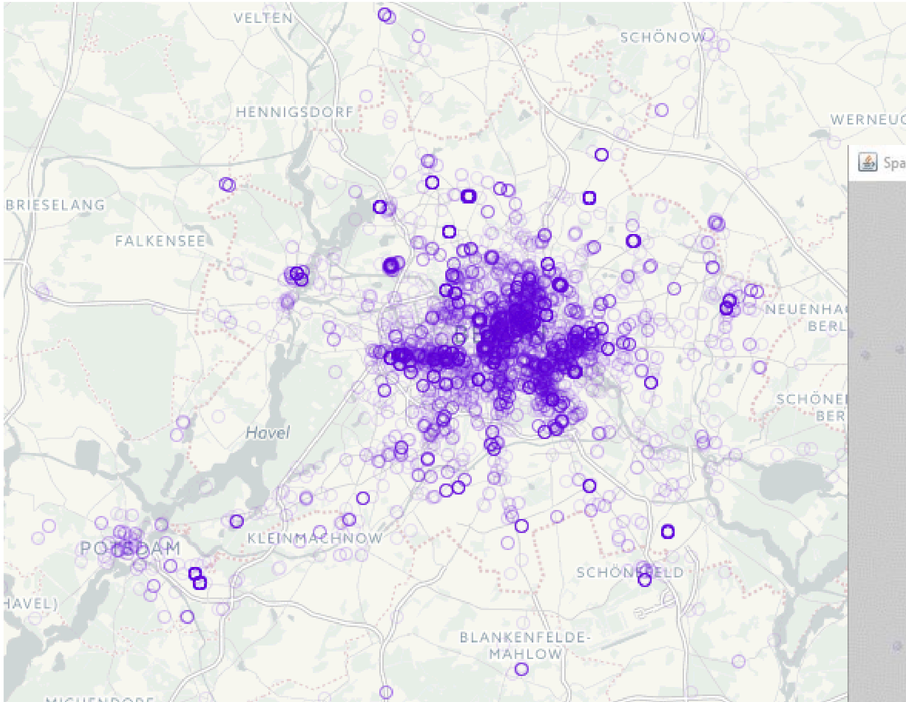
Spatial location

Time of appearance

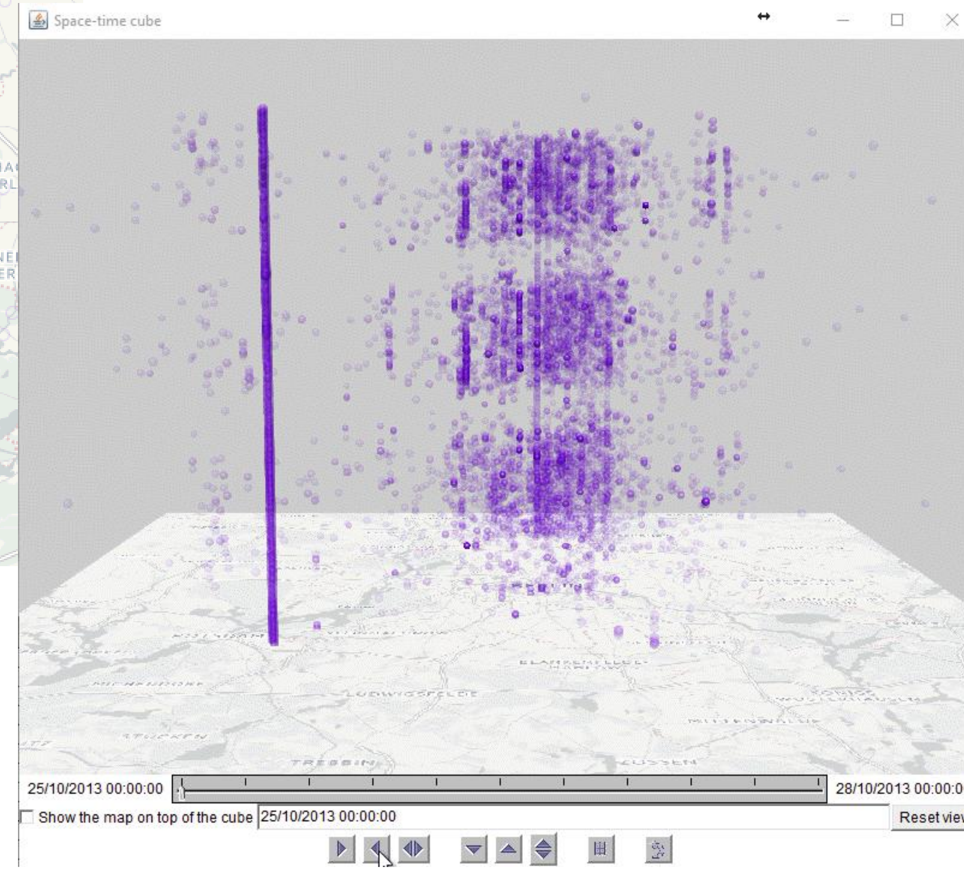
identifier	LONGITUDE	LATITUDE	MESSAGEDATE	USERID	USERSCREENNAME	MESSAGETEXT
3936903239	13.121572494506836	52.38300323486328	25/10/2013 12:47:42	5755692	matesl	Recruiting Session at
3936903880	13.382525444030762	52.53111267089844	25/10/2013 12:47:57	218215137	buschensemble	@MusiciansBFund @
3936905674	13.384698867797852	52.52968215942383	25/10/2013 12:48:40	14331417	WolfgangBremer	I'm at Bonfini (Berlin) h
3936906364	13.119993209838867	52.38190460205078	25/10/2013 12:48:56	160874621	_BB_RADIO_MUSIC	#nowplaying #stromae
3936907307	13.11646842956543	52.38713836669922	25/10/2013 12:49:19	161262801	RadioTeddyMusic	#nowplaying #silly ~ Si
3936907527	13.385029792785645	52.45845413208008	25/10/2013 12:49:24	88318861	ApfelVonSodom	@a_rabella eu te amo
3936908996	13.384437561035156	52.50713348388672	25/10/2013 12:49:59	173684513	a_abella	Just posted a photo @
3936909835	13.45739459991455	52.51235580444336	25/10/2013 12:50:19	14776696	ninalemos	Sabe o que ando pens
3936911312	13.3888578414917	52.52632522583008	25/10/2013 12:50:54	24479611	sebastianwaters	new Lunch Menu (@ R
3936911396	13.23953914642334	52.514625549316406	25/10/2013 12:50:56	182001862	SINNBUERO	be berlin. #relax @ Oly
3936913668	13.359097480773926	52.59198760986328	25/10/2013 12:51:50	1889365722	KingslandRoadDE	@KingslandRd hi; I lov
3936916785	13.460264205932617	52.4754753112793	25/10/2013 12:53:05	1929914154	HOWLmariagluck	Thank you so much for
3936919856	13.412495613098145	52.50828552246094	25/10/2013 12:54:18	176009999	ngamzegungor	@IrmakTanriover AMA
3936921490	13.455198287963867	52.53550720214844	25/10/2013 12:54:57	1077825853	thomasmatzka	Oh mein Gott!!! Der @
3936925416	13.423683166503906	52.491004943847656	25/10/2013 12:56:31	541618147	EniseCoruh	Özledi im lezzet :) @ La
3936925531	13.451854705810547	52.54849624633789	25/10/2013 12:56:33	137301856	mrmojiorisin	Mi è passata la voglia
3936926839	13.376999855041504	52.5161018371582	25/10/2013 12:57:04	1336218432	trendinaliaDE	#Snowden nur ein Tre
3936927184	13.394739151000977	52.47782516479492	25/10/2013 12:57:13	21513949	ralphcochrane	Getting ready for the cr
3936928716	13.388001441955566	52.49345016479492	25/10/2013 12:57:49	20170781	david_g_cooper	German bar staff need
3936929248	13.289820671081543	52.4578742980957	25/10/2013 12:58:02	92283308	x_elly_	So dreckig wie heute g
3936929706	13.414168357849121	52.50786590576172	25/10/2013 12:58:13	175051281	betuel07	I'm at Bahçe ehir Unive
3936931181	13.11768913269043	52.38417816162109	25/10/2013 12:58:48	119337117	quangVFX	I'm at Animago Confer

Sort by: MESSAGEDATE | Ascending | TableLens | Attribute...

Spatial positions of the tweet posting events



Spatio-temporal positions of the events



Notes:

1. Events may have non-zero spatial extents (not necessarily points).
2. Durable events can be represented in a space-time cube by vertical bars or prisms.

The idea of ***space-time cube*** comes from:
Hägerstrand. T. (1970).
"What about people in regional science?"
Papers of the Regional Science Association; 24:7-21.

Trajectories of moving objects

- A **trajectory** is a chronologically ordered sequence of time-referenced spatial positions of a moving object
- Examples: GPS tracks of vehicles or animals
- Data structure: <object identifier and/or trip identifier, time stamp, spatial position, *any attributes*>

Example: trajectories of vehicles

Trip identifier

Time stamp

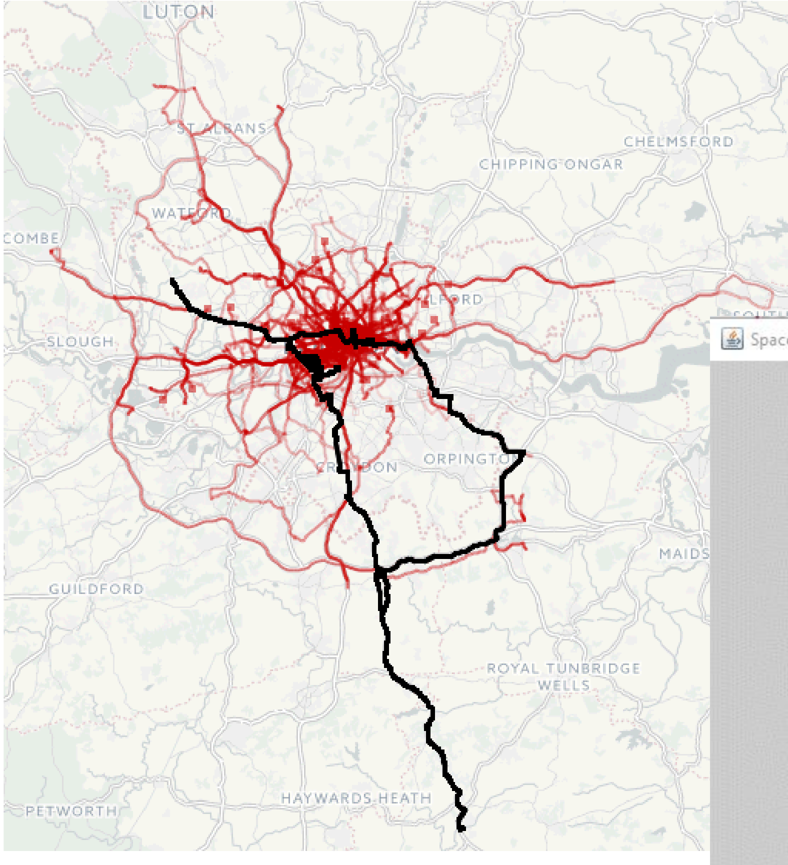
Spatial position

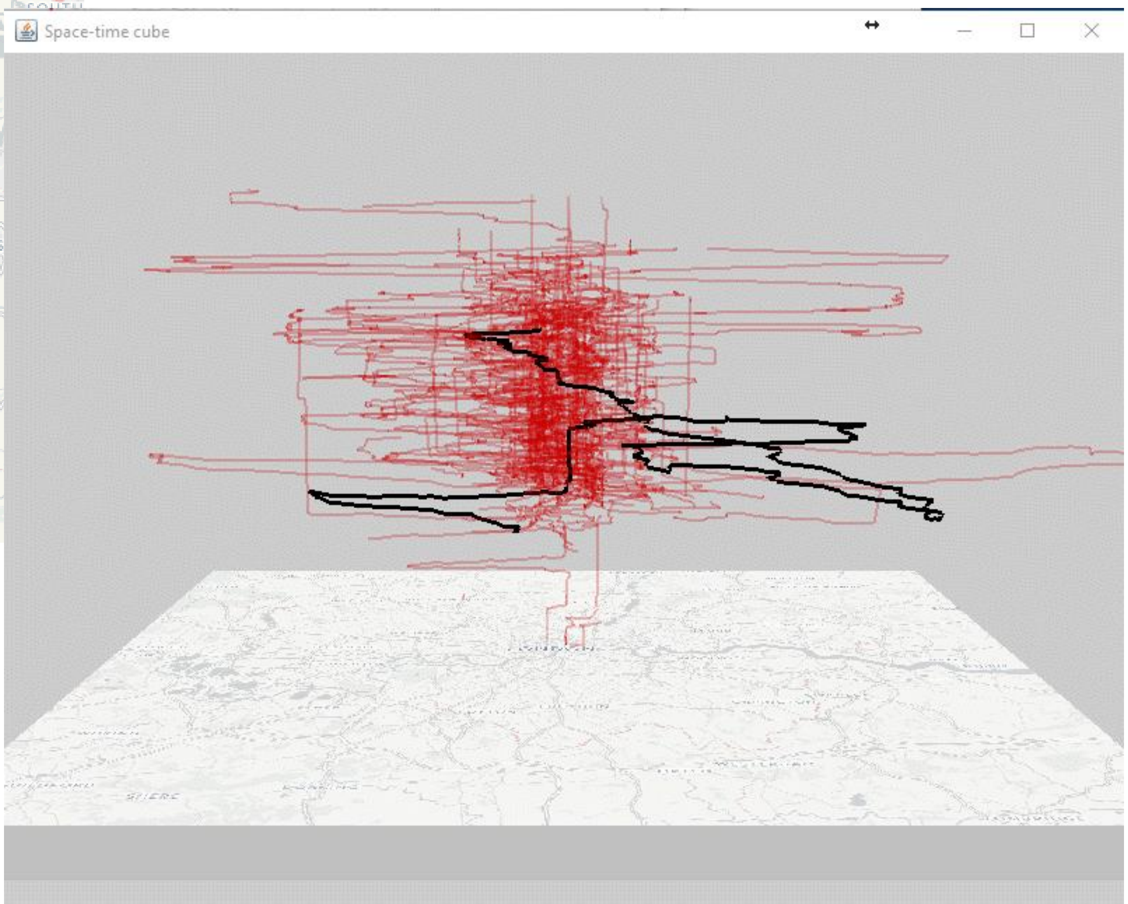
Object identifier

	Trajectory ID	time	longitude	latitude	Entity ID
339_900	339	06/03/2007 19:17:31	-0.10423000156879425	51.50750732421875	motorbike 7
339_901	339	06/03/2007 19:17:39	-0.10427333414554596	51.507408142089844	motorbike 7
339_902	339	06/03/2007 19:17:49	-0.10411166399717331	51.50752258300781	motorbike 7
339_903	339	06/03/2007 19:18:02	-0.10446833074092865	51.50712203979492	motorbike 7
339_904	339	06/03/2007 19:18:10	-0.1043633297085762	51.50709533691406	motorbike 7
339_905	339	06/03/2007 19:18:51	-0.10951333492994308	51.505882263183594	motorbike 7
339_906	339	06/03/2007 19:19:13	-0.11245166510343552	51.5056266784668	motorbike 7
339_907	339	06/03/2007 19:19:23	-0.11319166421890259	51.50448226928711	motorbike 7
339_908	339	06/03/2007 19:19:34	-0.1137833297252655	51.50422668457031	motorbike 7
339_909	339	06/03/2007 19:19:44	-0.1153983324766159	51.50351333618164	motorbike 7
345_0	345	06/03/2007 08:31:06	0.0583166666328907	51.52149200439453	motorbike 78
345_1	345	06/03/2007 08:31:17	0.058694999665021896	51.52083206176758	motorbike 78
345_2	345	06/03/2007 08:31:30	0.05913333222270012	51.52022171020508	motorbike 78
345_3	345	06/03/2007 08:31:38	0.05857166647911072	51.51984405517578	motorbike 78
345_4	345	06/03/2007 08:31:48	0.05670500174164772	51.519508361816406	motorbike 78
345_5	345	06/03/2007 08:31:58	0.05439166724681854	51.51934814453125	motorbike 78
345_6	345	06/03/2007 08:32:09	0.052186667919158936	51.519493103027344	motorbike 78
345_7	345	06/03/2007 08:32:19	0.04961499944329262	51.51976776123047	motorbike 78
345_8	345	06/03/2007 08:32:31	0.04677499830722809	51.52006149291992	motorbike 78
345_9	345	06/03/2007 08:32:39	0.044165000319480896	51.52008056640625	motorbike 78
345_10	345	06/03/2007 08:32:49	0.04174000024795532	51.519737243652344	motorbike 78
345_11	345	06/03/2007 08:33:02	0.04002833366394043	51.519317626953125	motorbike 78

Sort by: No selection | Ascending | TableLens | Attribute...

Spatial footprints of multiple trajectories





Spatial (= spatially referenced) time series

- A **time series** is a chronologically ordered sequence of *data items* that refer to different moments or intervals in time
  - Time series of attribute values
    - E.g., measured values of weather parameters at different times
  - Time series of satellite images
- A **spatial time series** is a set of time series of attribute values, where each time series refers to a distinct spatial location (point or region in space) or a spatial object.

Example: crime statistics

Reference 1: time

Reference 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 table columns

Reference 1: place

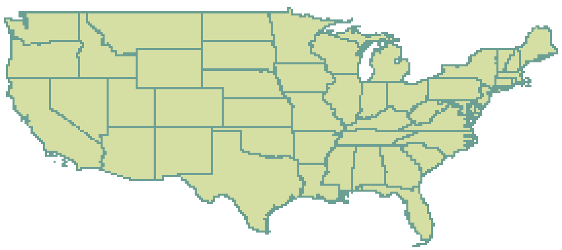
Reference 2: time

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

Slices of spatial time series

1)

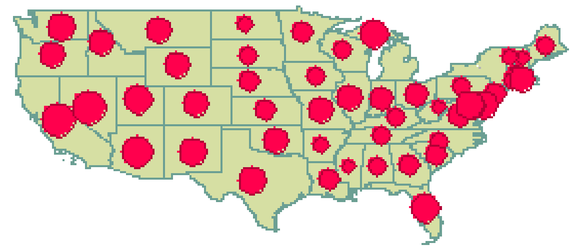
Space as a whole



Selected time

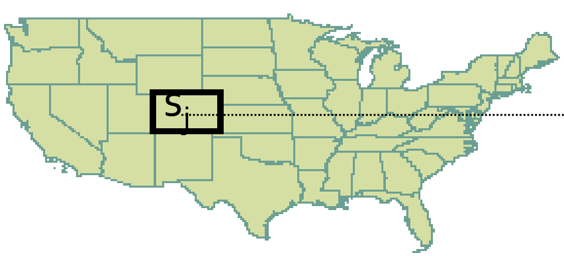
t_k

Slice: spatial **behaviour** (= distribution of attribute values over space) at this time



2)

Selected place

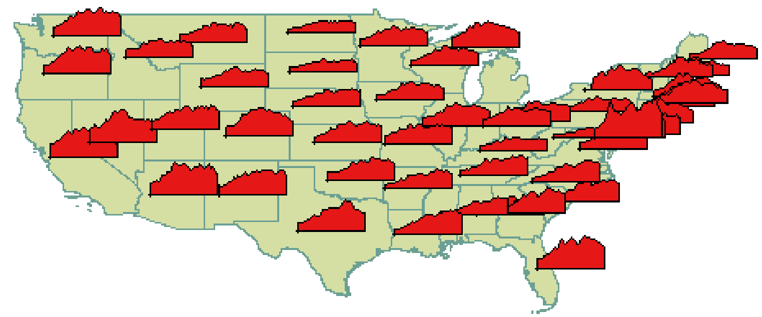


Time as a whole

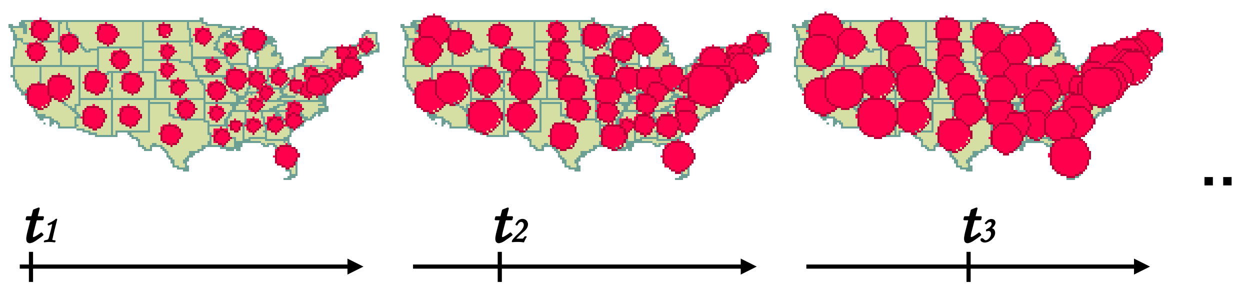
Slice: temporal **behaviour** (= variation of attribute values over time) in this place

Two complementary views of spatial time series

1. As a spatial distribution of local time series: a set of time series of attribute values in different locations



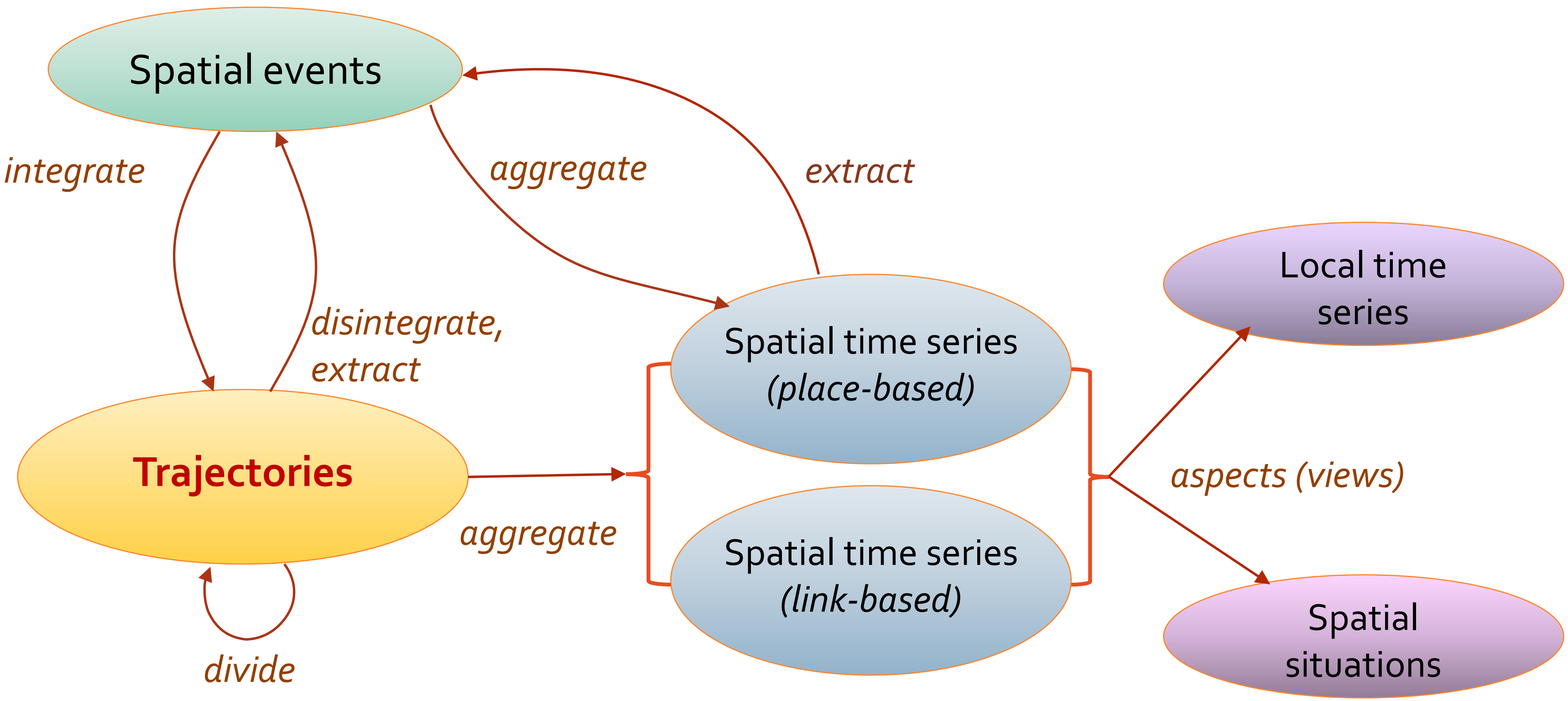
2. As time-varying spatial situations: sequence of spatial distributions of attribute values at different times



These views require different visualisation and analysis techniques.

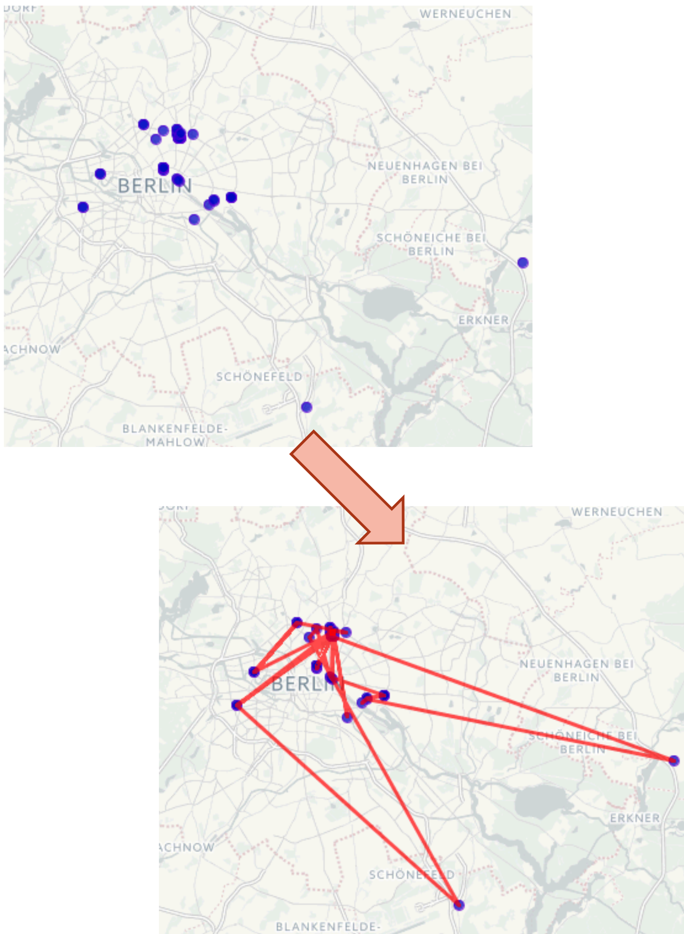
Transformations of spatio-temporal data

Transformations of spatio-temporal data

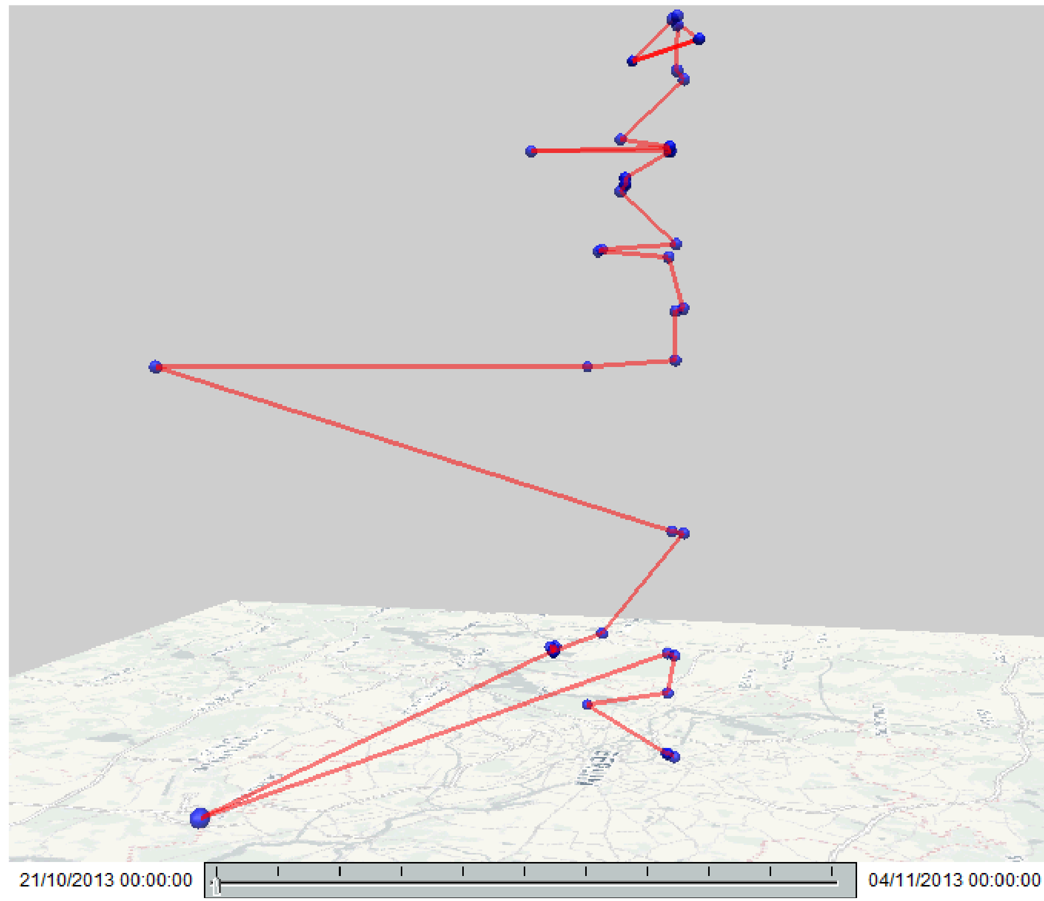


Spatial events → trajectories

Tweet events of one Twitter user

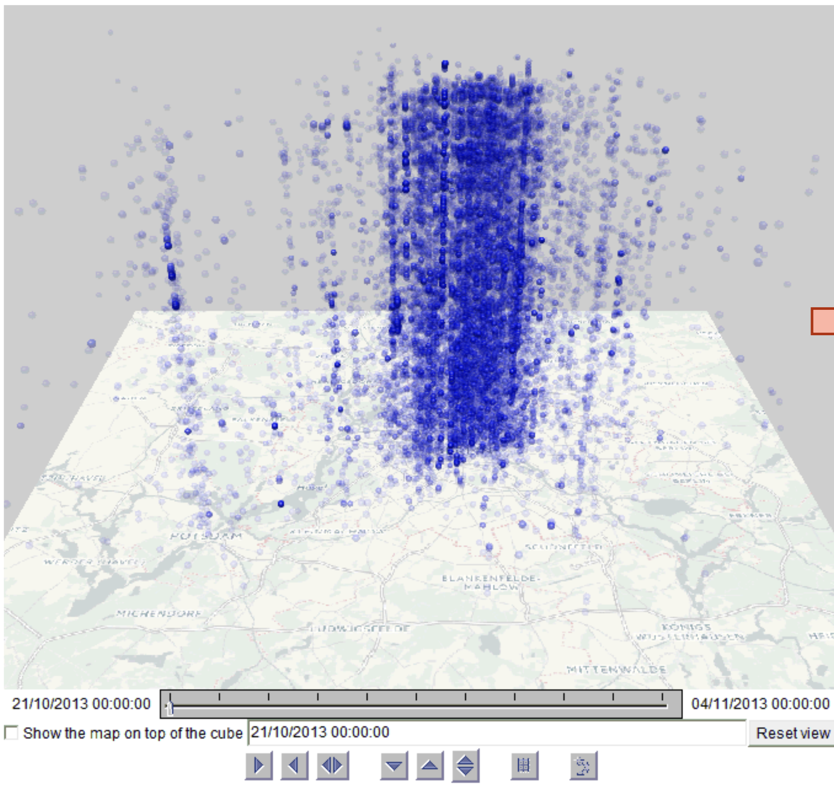


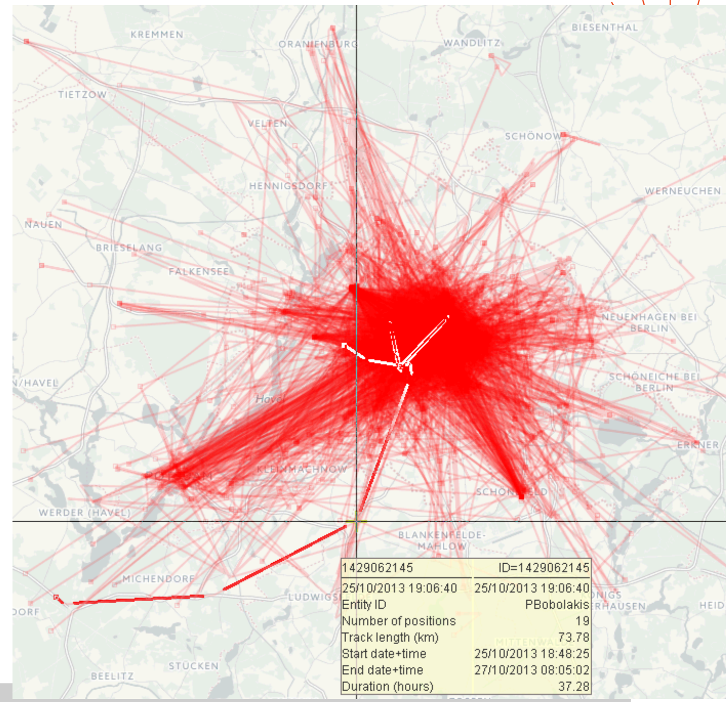
Trajectory of the Twitter user

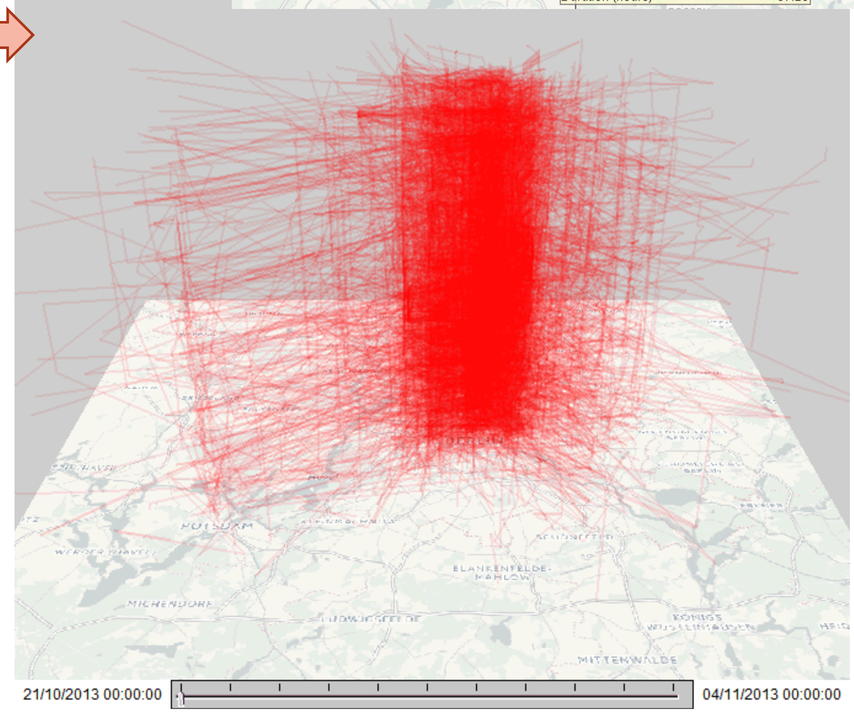


Spatial events → trajectories

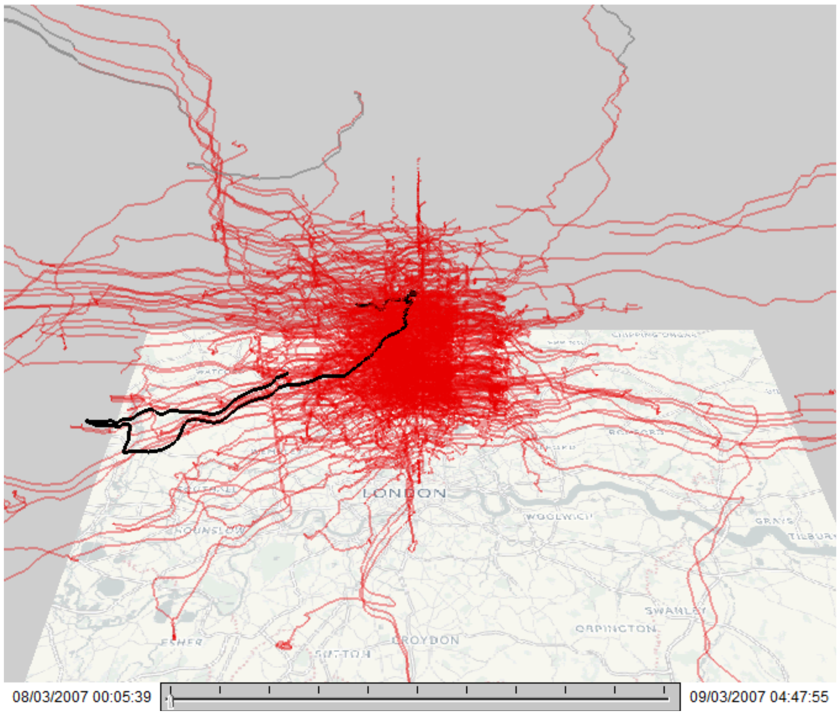
Tweet events of multiple users

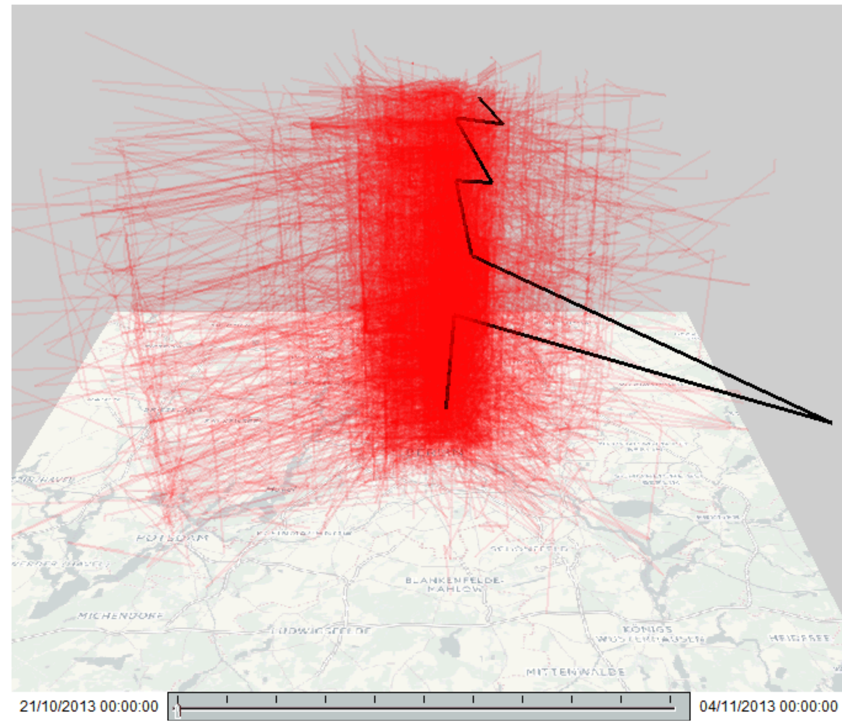






Quasi-continuous vs. episodic trajectories





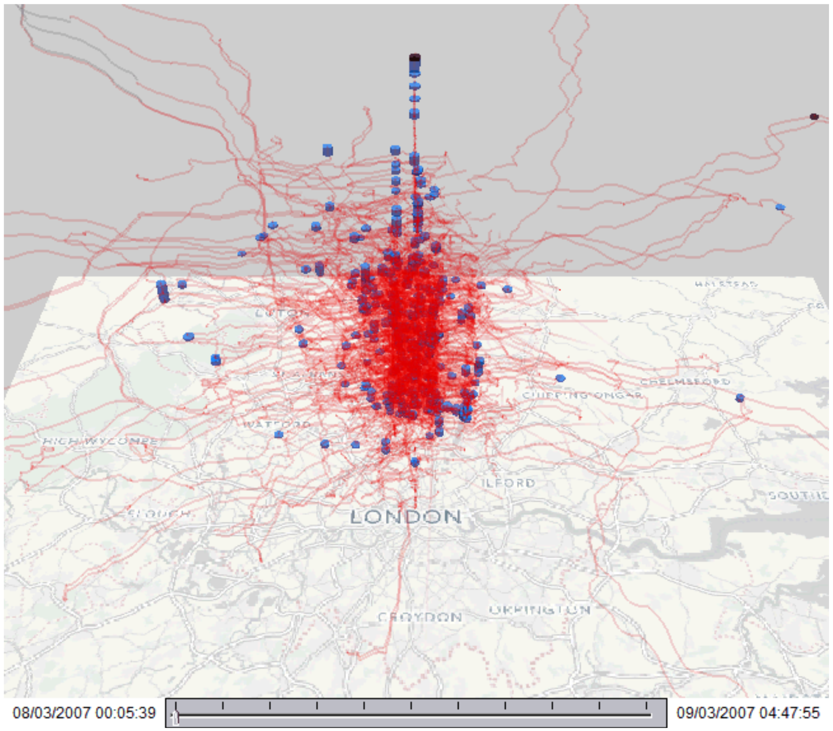
Quasi-continuous:

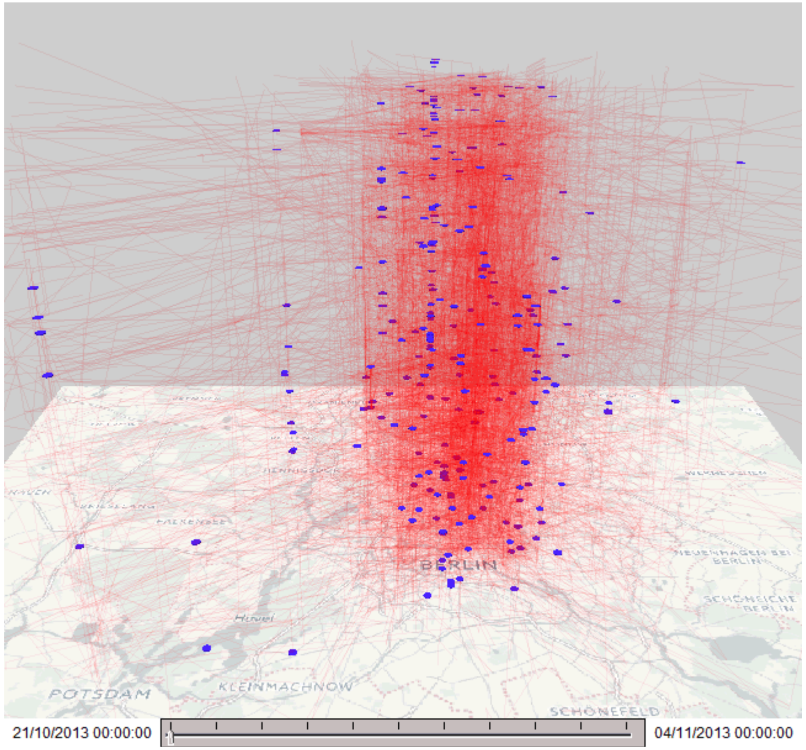
- Small spatial and temporal distances between consecutive positions
- Spatially smooth
- Permit interpolation between recorded positions.

Episodic:

- May have large spatial and temporal gaps between consecutive positions
- Spatially abrupt
- No valid interpolation between recorded positions is possible.

Trajectories → spatial events (of interest)





E.g., stops for at least 5 minutes

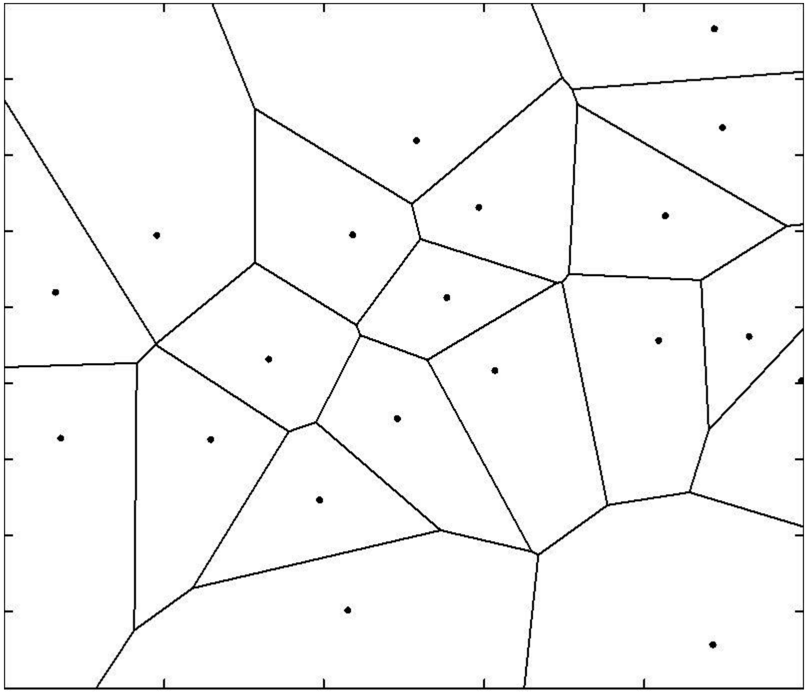
Spatio-temporal aggregation of events

- Spatial events and their thematic attributes can be aggregated spatially by spatial compartments and time intervals.
  - Partition the underlying space extent into suitable compartments (areas)
    - Create a regular or irregular grid
    - Use a pre-existing division (e.g., administrative)
  - Partition the time span of the data into intervals
  - For each compartment and time interval:
    - Count the events; possibly, normalize by compartment areas, resident population, …
    - Compute statistics of values of thematic attributes
- Resulting data type: **spatial time series** of
  - event counts, densities, counts per capita, …
  - statistical summaries of thematic attributes: mean, median, mode, minimum, maximum, quantiles, …

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

Used for building irregular grids

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



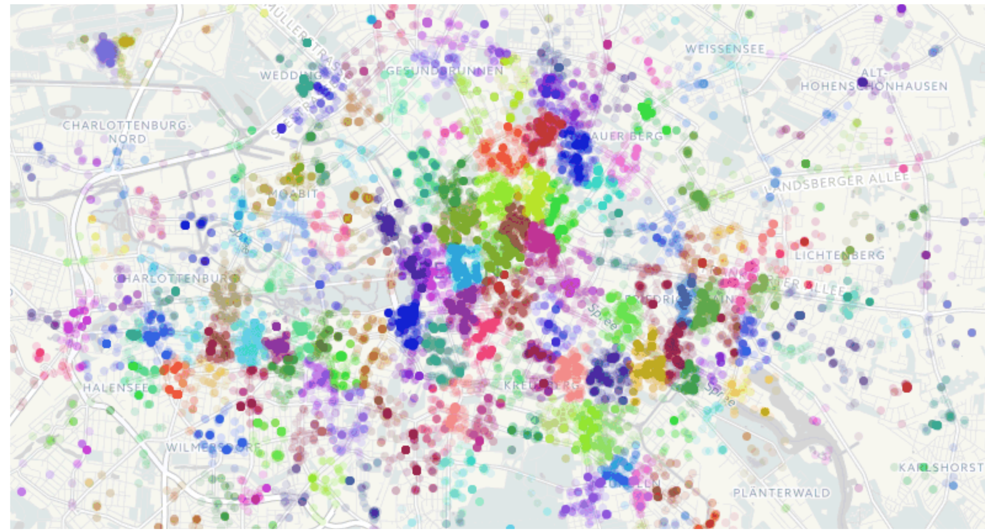
Data-driven tessellation

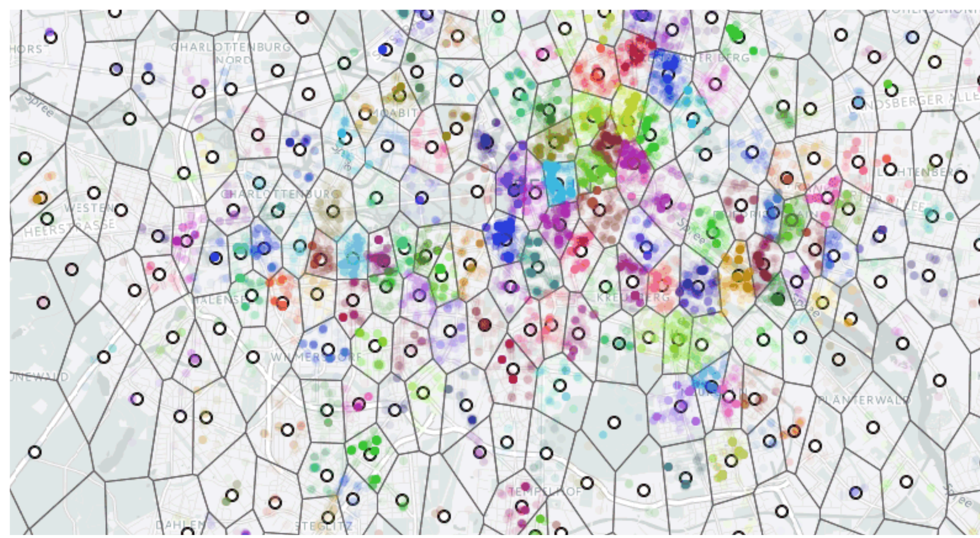
Step 1: apply a special algorithm for grouping of points based on their spatial proximity.

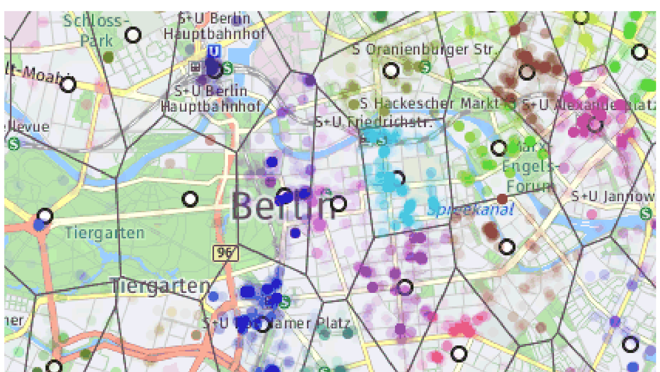
The main idea: put the points into circles with a given maximal radius R.

In this example, R = 1000 m.

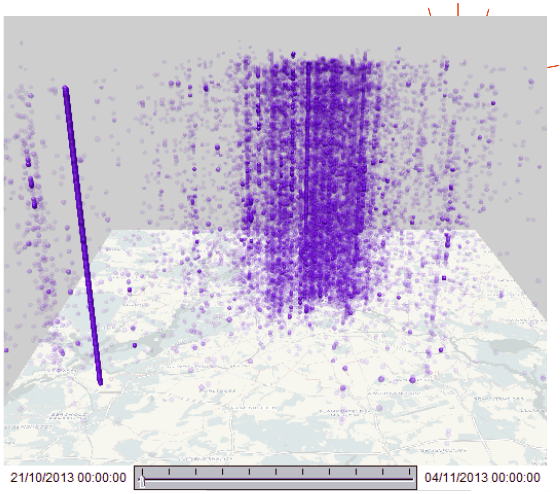
Step 2: use the centres of the point groups as the generating seeds for the Voronoi tessellation.

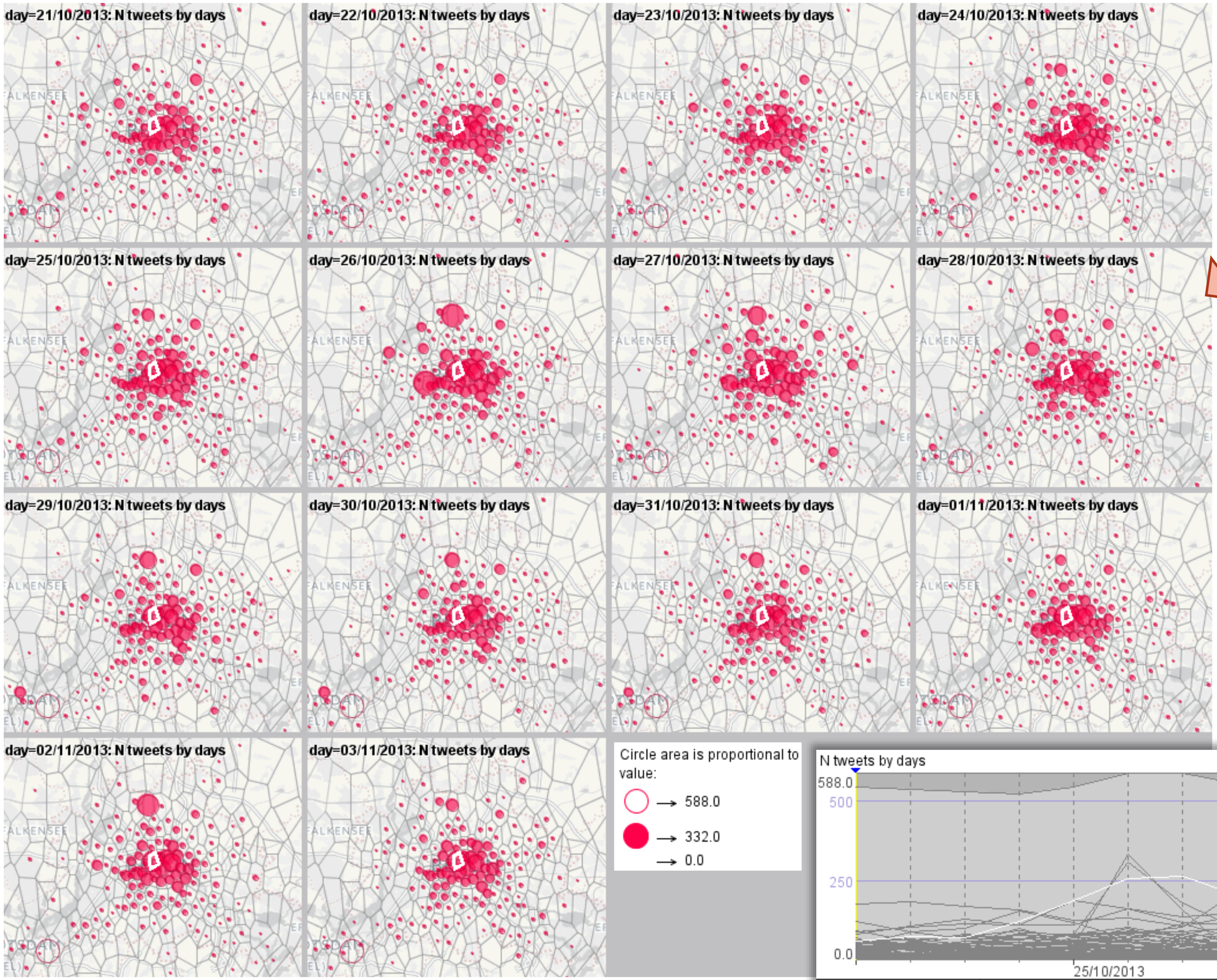


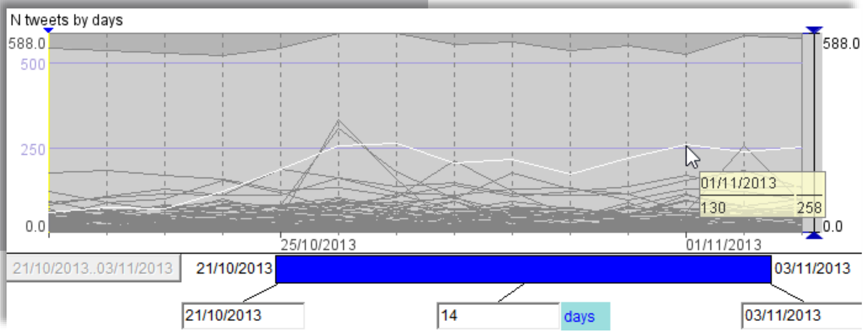




Spatial events → spatial time series







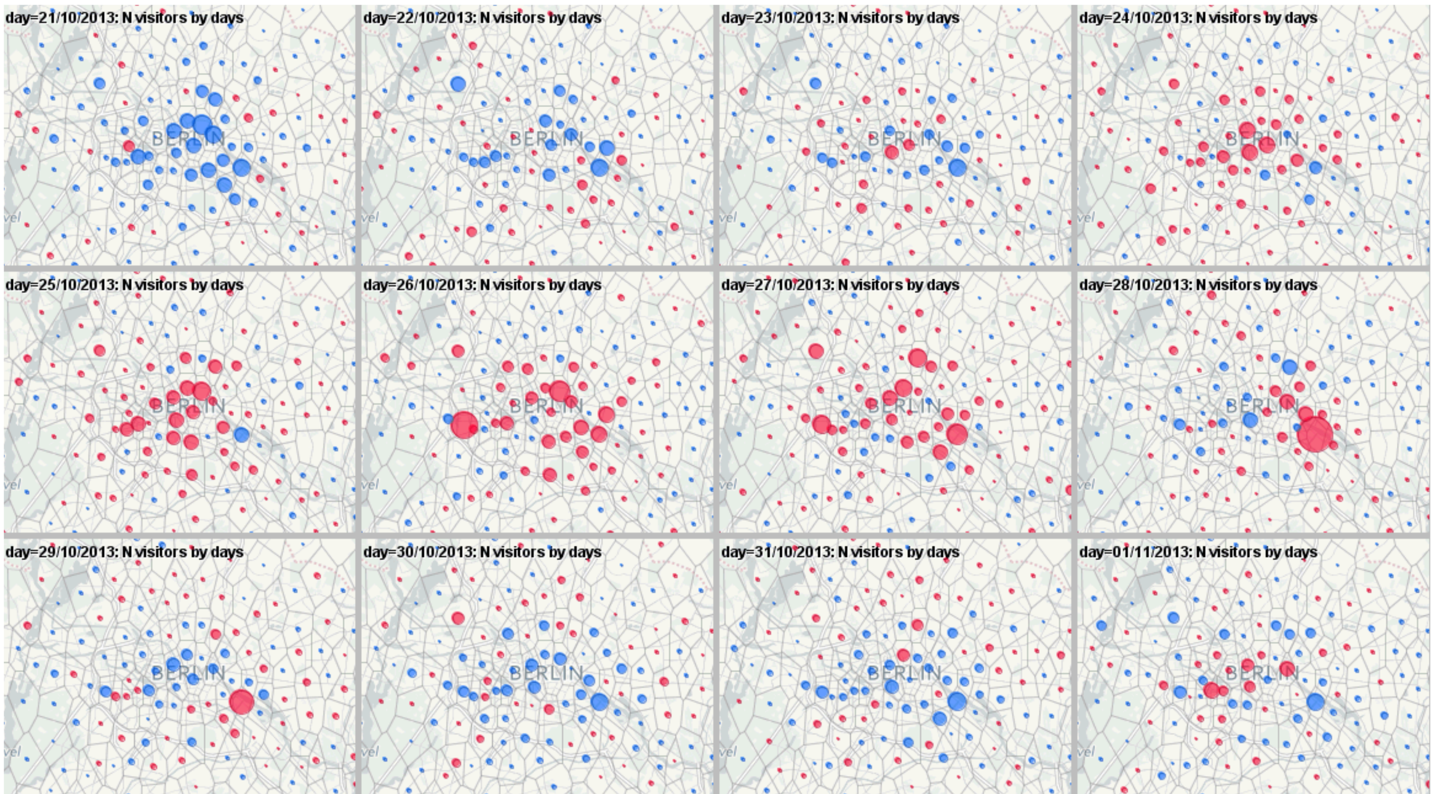
Spatio-temporal aggregation of trajectories

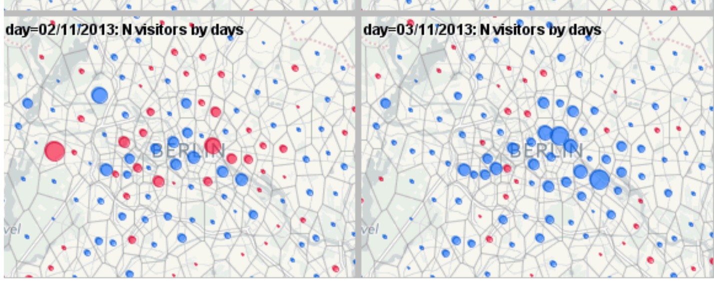
- Partition the underlying space extent into suitable compartments (areas)
- Partition the time span of the data into intervals

1. For each compartment C and time interval Δt, count the number of *visits* that occurred during Δt and the number of distinct *visitors*
   $\rightarrow$ place-based time series (time series of presence)

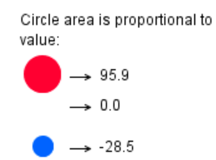
2. For each ***pair of compartments*** C_1 and C_2 and time interval Δt, count the number of *moves (transitions)* from C_1 to C_2 that occurred during Δt and the number of distinct *objects that moved*
   $\rightarrow$ link-based time series (time series of aggregate moves, called *flows*)

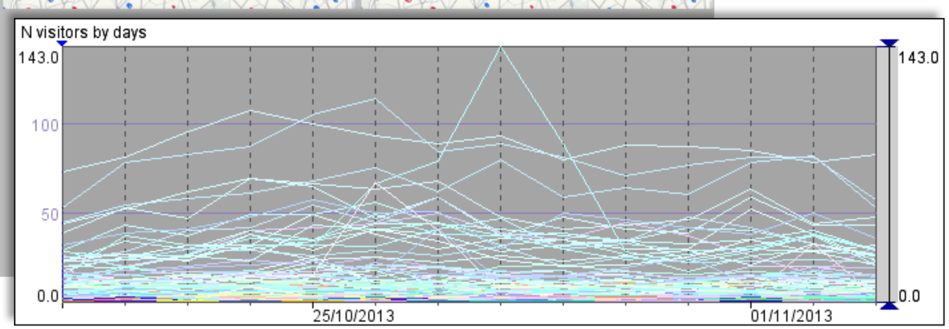
Trajectories → place-based time series (presence)



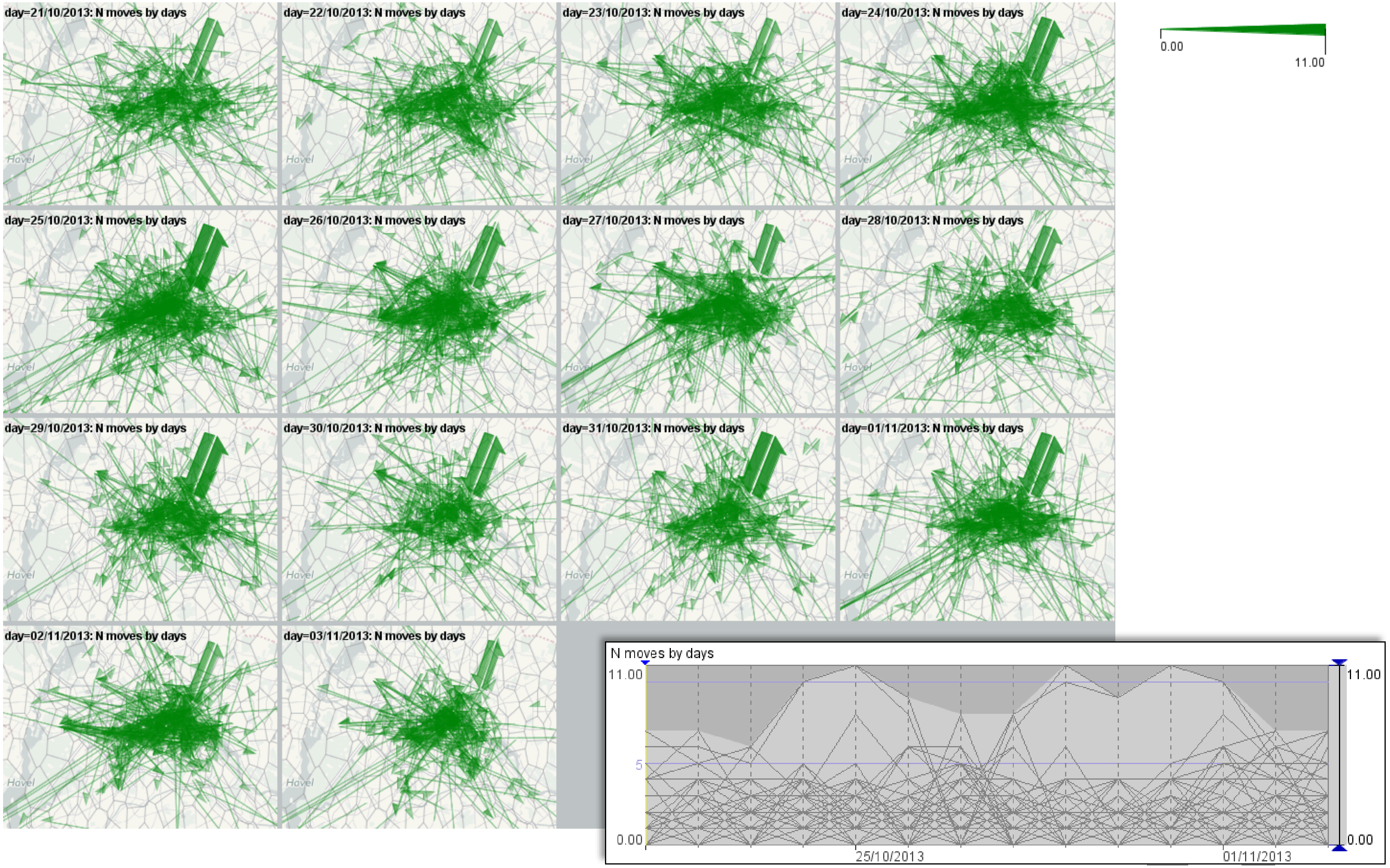


The map shows the counts of distinct visitors transformed to the differences from the means

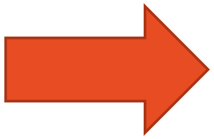




Trajectories → link-based time series (flows)



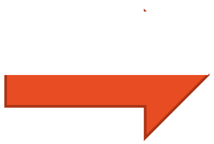
The flow symbols



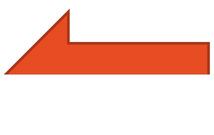
Arrow: shows the movement direction;
the width can represent a numeric attribute, such as the number of moving objects or the count of moves.

Problem: it is hard to represent movements in two opposite directions.

Solution: use halves of arrows!

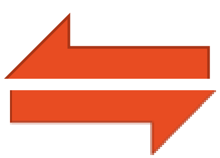


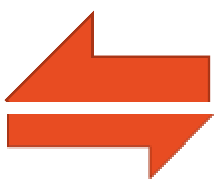
Movement to the right



Movement to the left

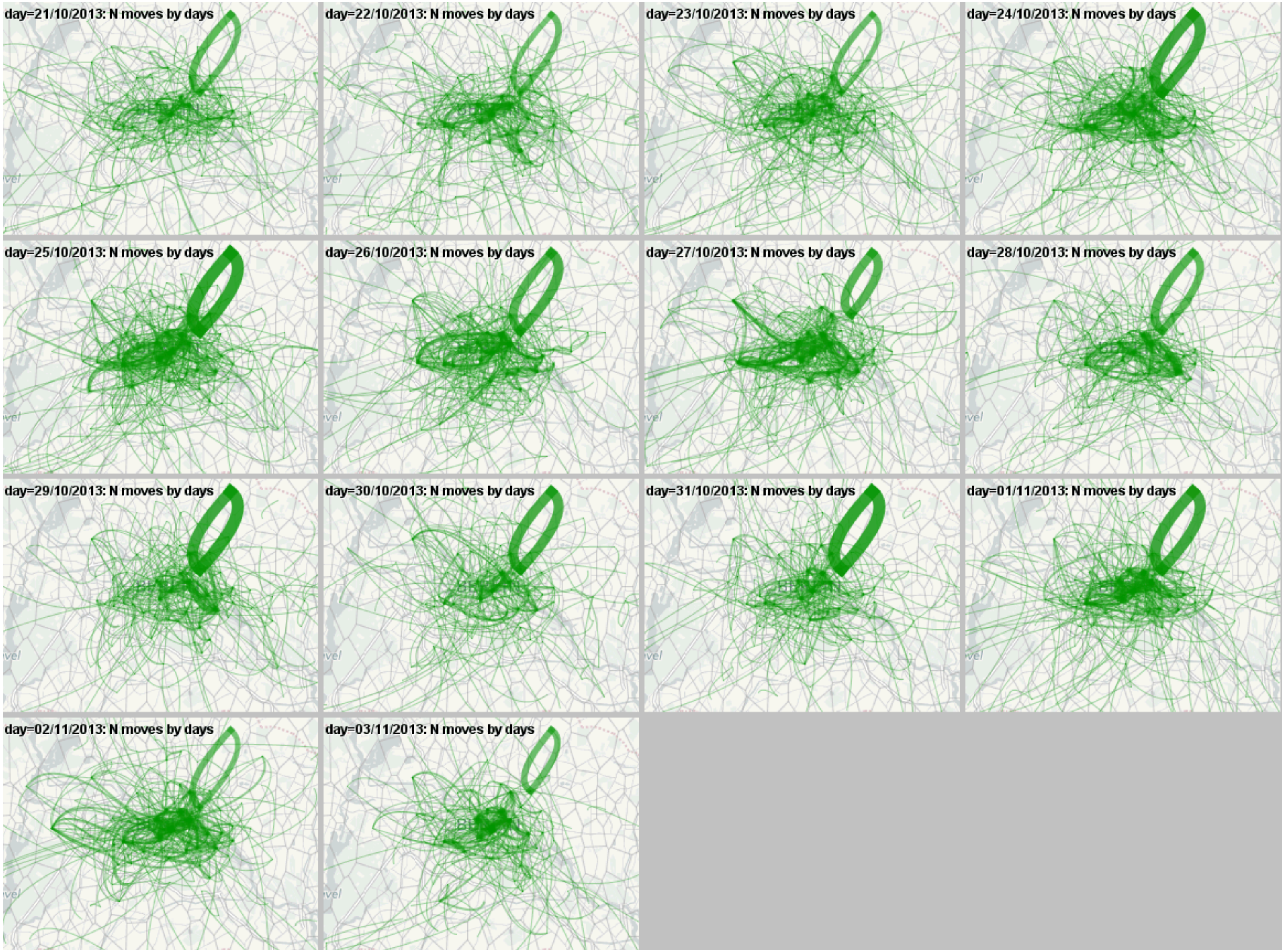
Easy to put together:

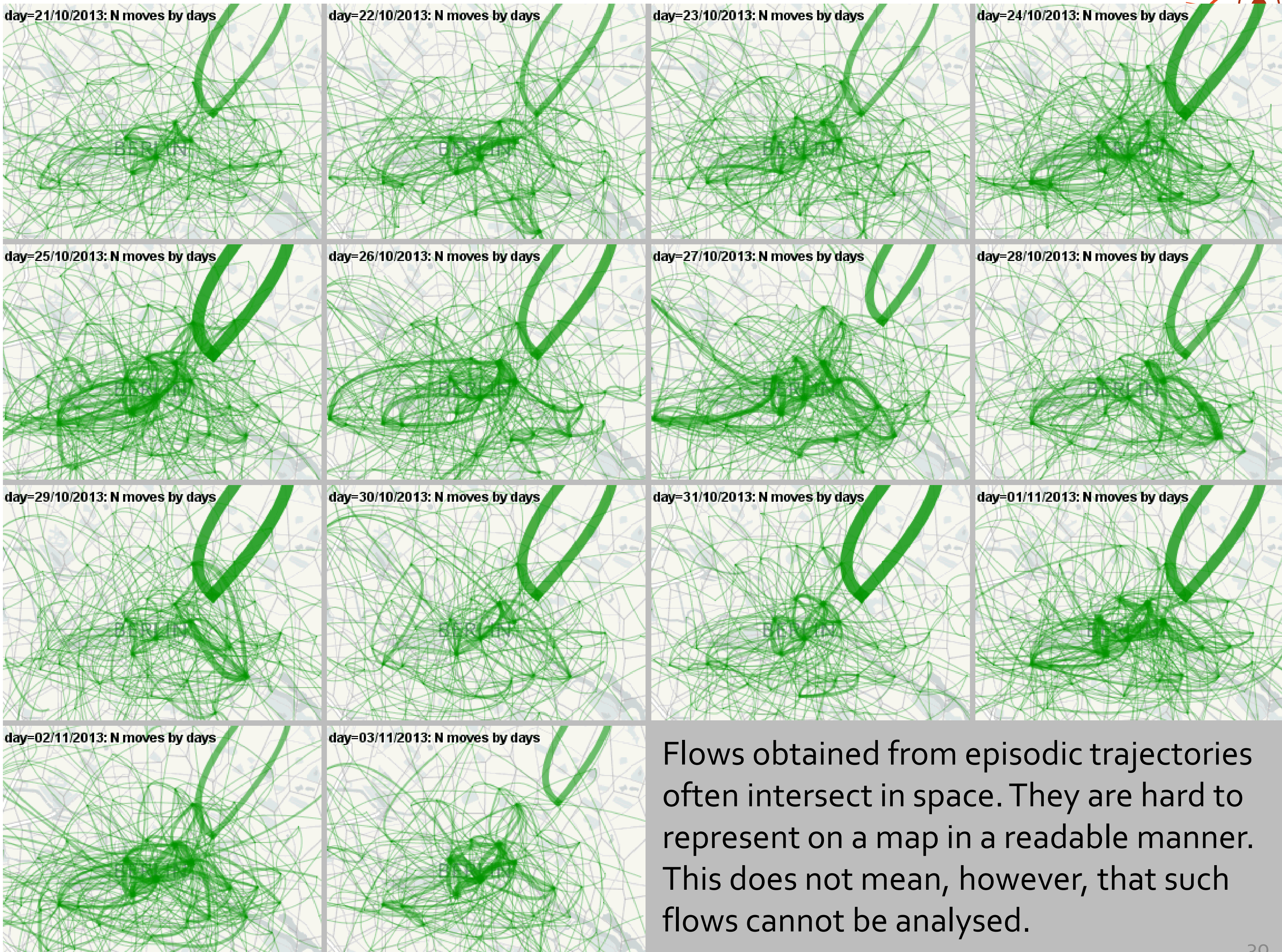




Can show asymmetric amounts of movement

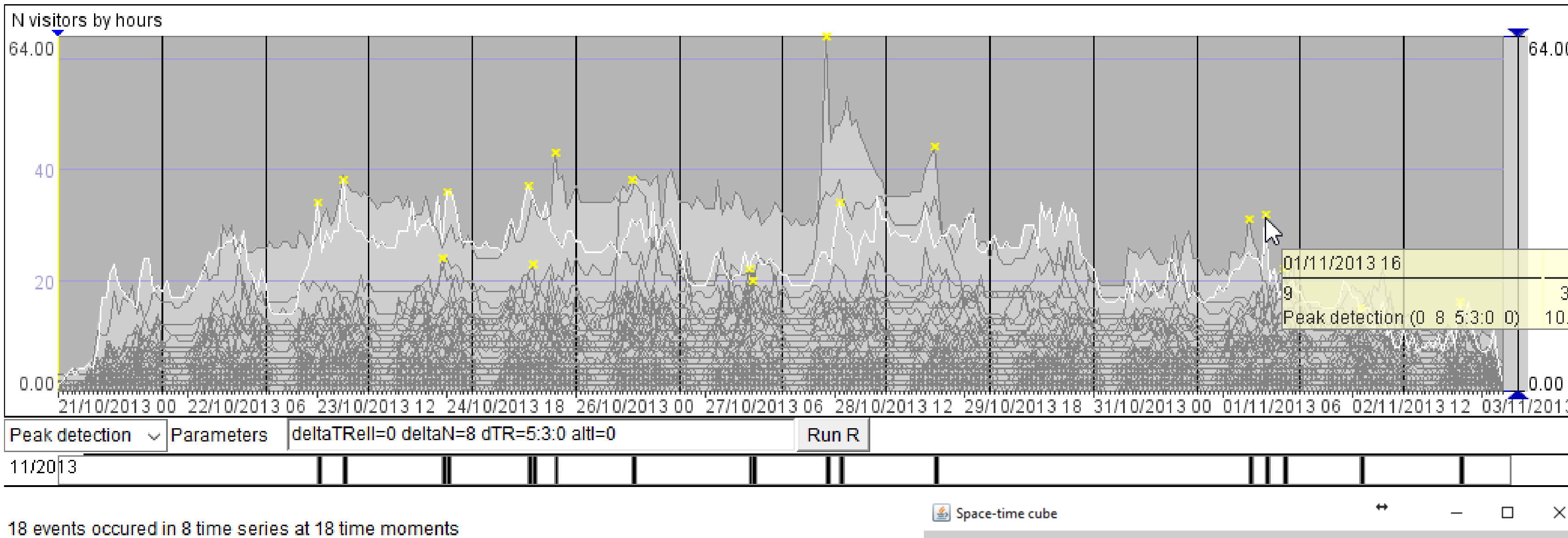
Another variant of flow symbols





Flows obtained from episodic trajectories often intersect in space. They are hard to represent on a map in a readable manner. This does not mean, however, that such flows cannot be analysed.

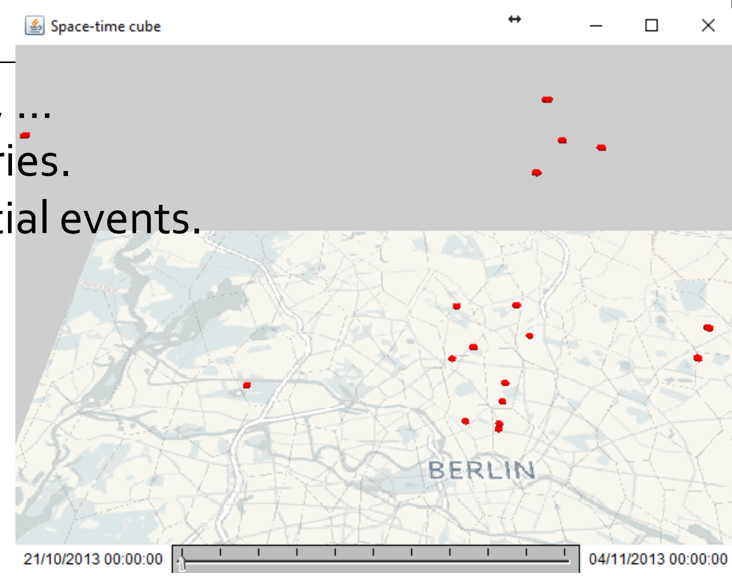
Spatial time series → spatial events



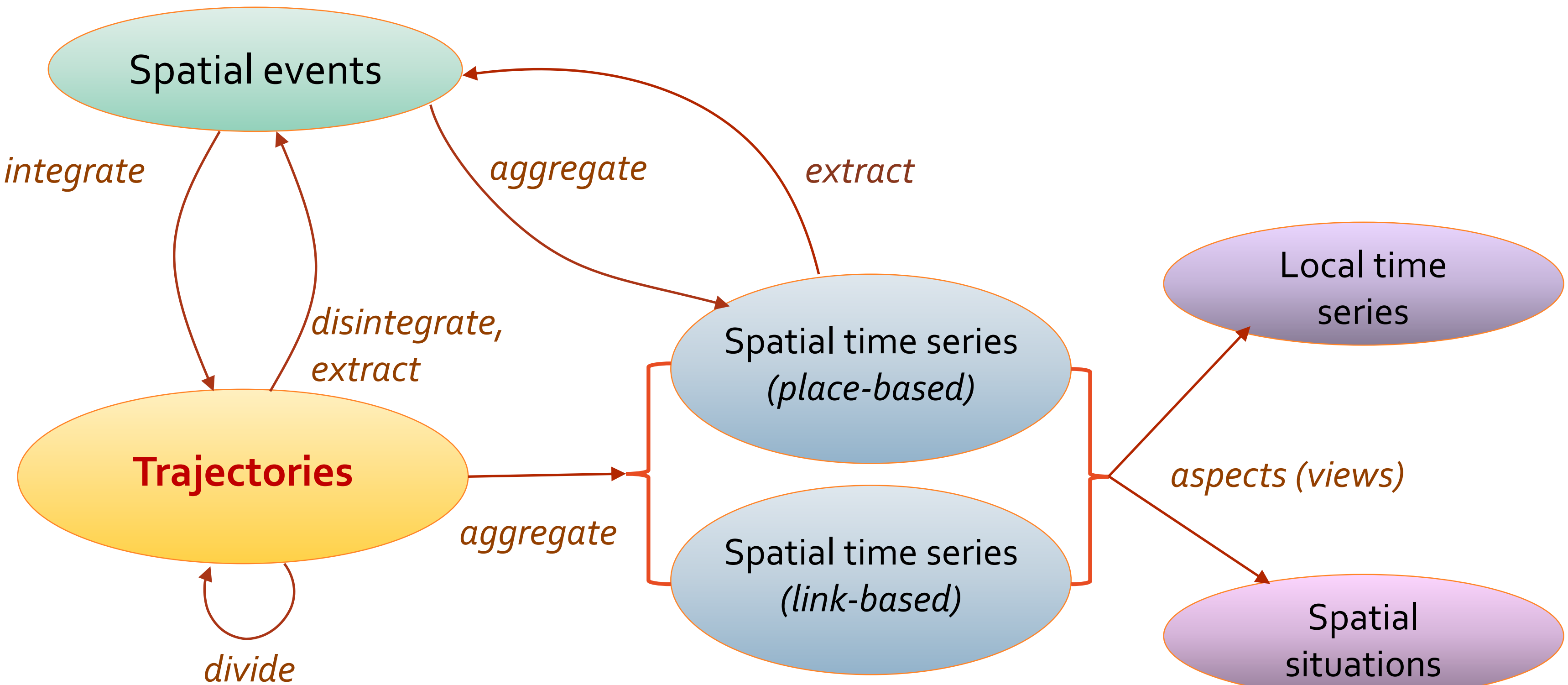
Events in time series: peaks, drops, trend change, …

Events of interest can be **extracted** from time series.

Events extracted from spatial time series are spatial events.



Transformations of spatio-temporal data



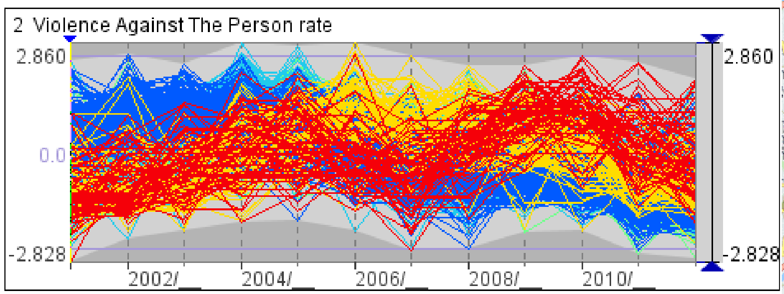
Transformations enable multi-perspective analyses of spatio-temporal data.

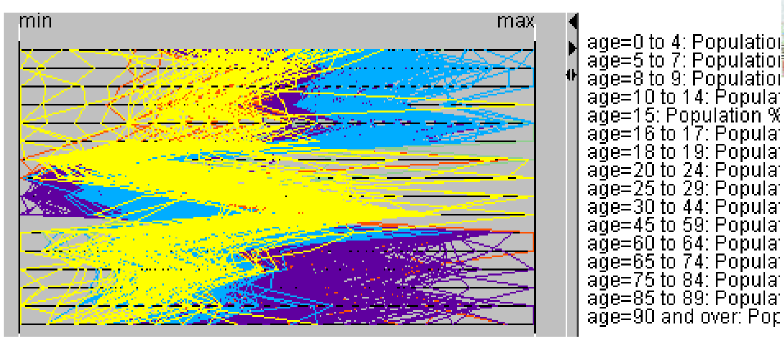
Introduction to clustering

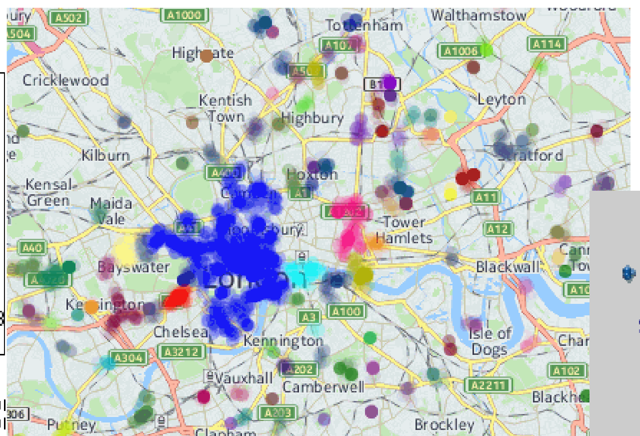
Partition-based and density-based clustering

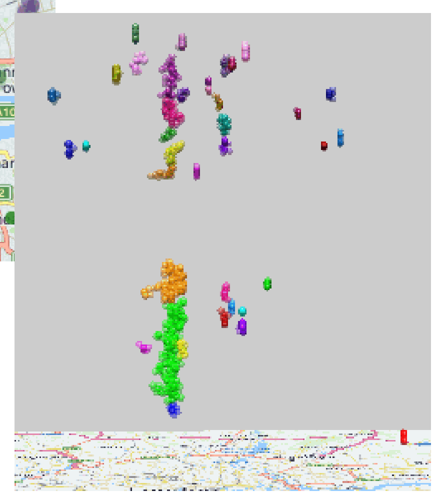
What is clustering?

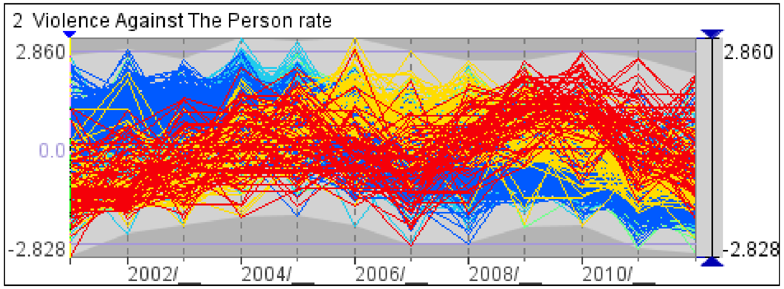
- Loose definition: clustering is the process of organising objects into groups whose members are close or similar in some way.
- A cluster is a group of objects which are "similar" or "close" between them and are "dissimilar" or "distant" to the objects belonging to other clusters.



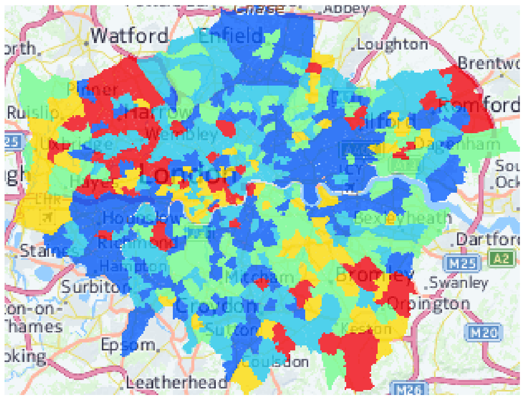


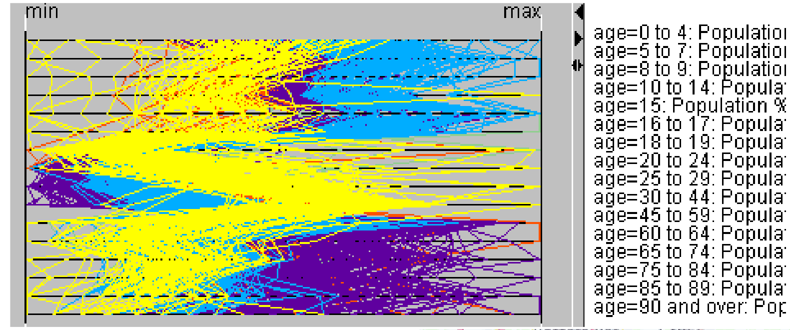




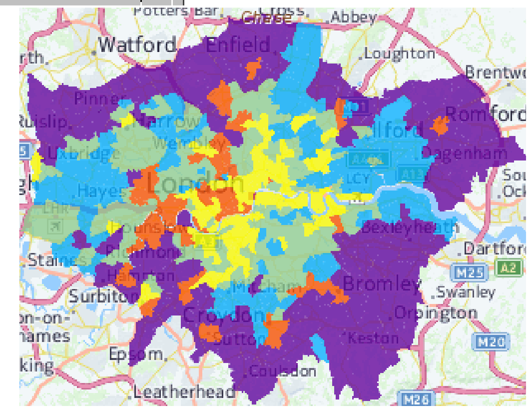


Clusters of similar time series

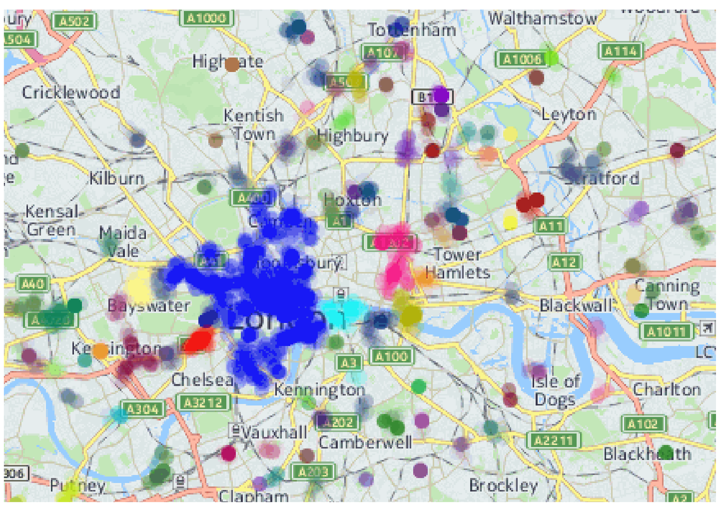




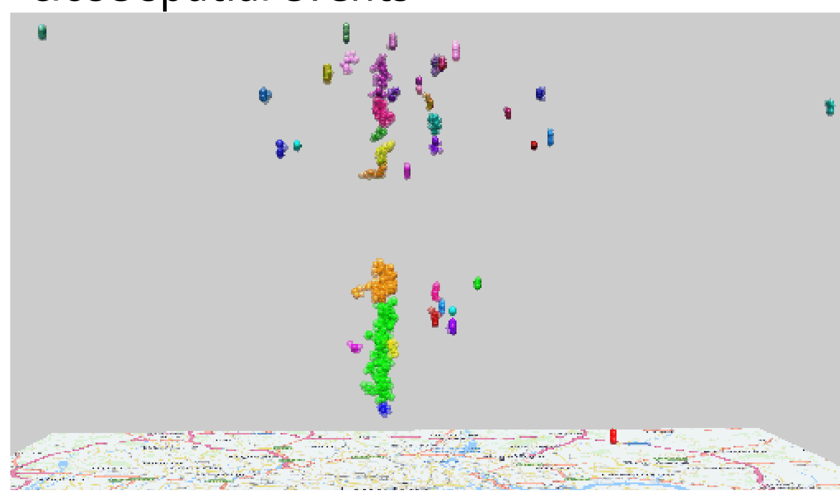
Clusters of similar multi-attribute value combinations



Clusters of spatially close spatial objects



Clusters of spatially and temporally close spatial events



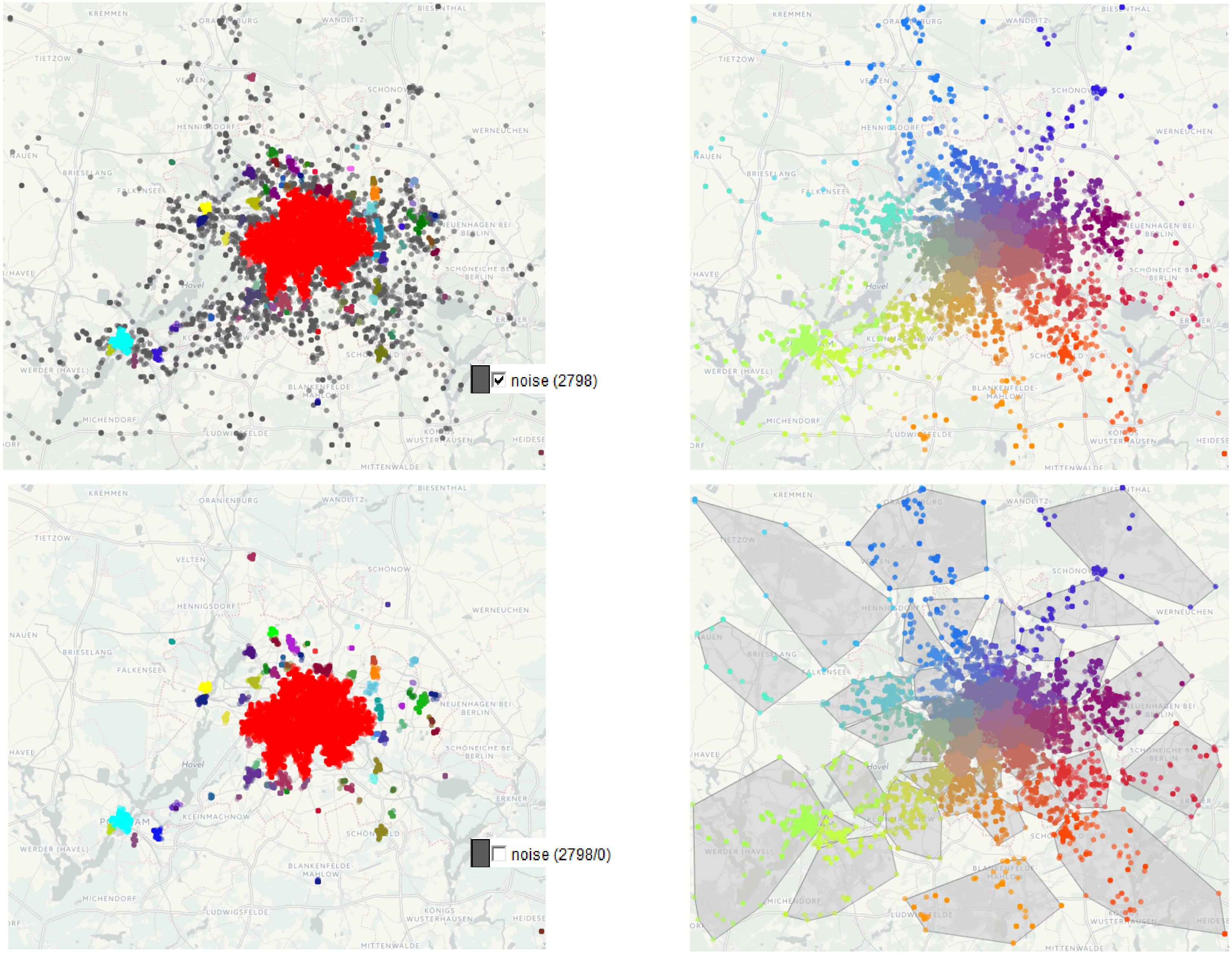
Role of clustering in visual analytics

- Grouping of similar or close items plays an essential role in VA
  - as a tool supporting abstraction: elements → subsets; the subsets may be considered as wholes
  - as a tool to manage large data volumes
  - as a tool to find specific features of interest, e.g., event concentrations
  - as a tool to deal with multiple attributes and multiple time series, which are hard to visualise

Two major types of clustering

- **Partition-based clustering (PBC)**: divide items into groups so that items within a group are similar (close) and items from different groups are less similar (more distant)
  - Examples: k-means, SOM (self-organizing map)
  - Property of the result: each item belongs to some group
- **Density-based clustering (DBC)**: find groups of highly similar (close) items and separate from them items that are less similar (more distant) to others
  - Examples: DBScan, OPTICS
  - Properties of the results: some items belong to groups, other items remain ungrouped and are treated as “noise”

Example: DBC (left) and PBC (right) of points according to their spatial positions



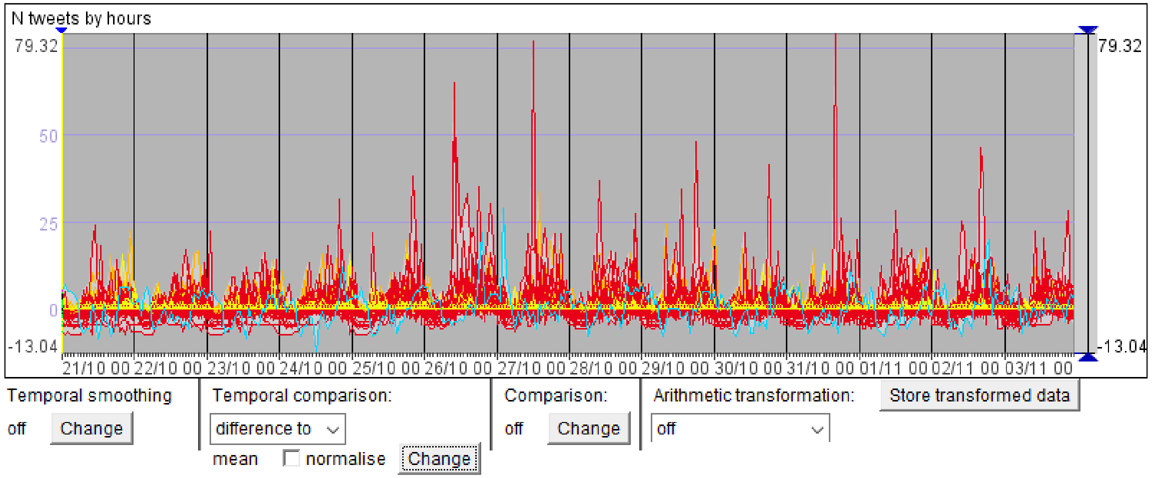
Use of the two types of clustering

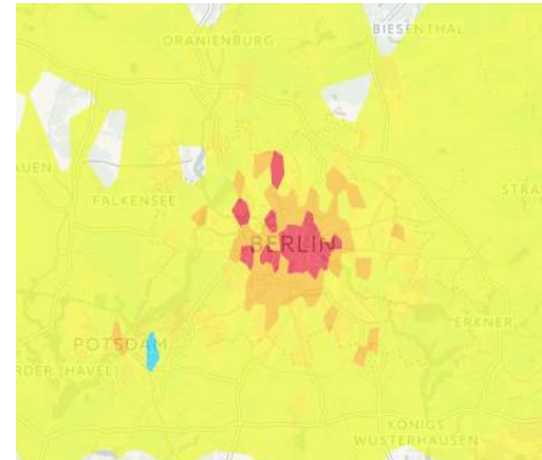
- **Partition-based:**
  - Typically applied to multiple thematic (non-spatial) attributes or to **time series** of thematic attributes
  - Objective: divide objects into groups such that objects within a group have similar attribute values and differ from the objects in the other groups
- **Density-based:**
  - Typically applied to spatial and temporal attributes of spatial or spatio-temporal objects
  - Objective: find concentrations of objects in space or in space and time (i.e., groups of objects with close spatial locations and existence times)
    - concentrations of objects may have special meanings;
      e.g., spatio-temporal cluster of low speed events $\Rightarrow$ traffic jam

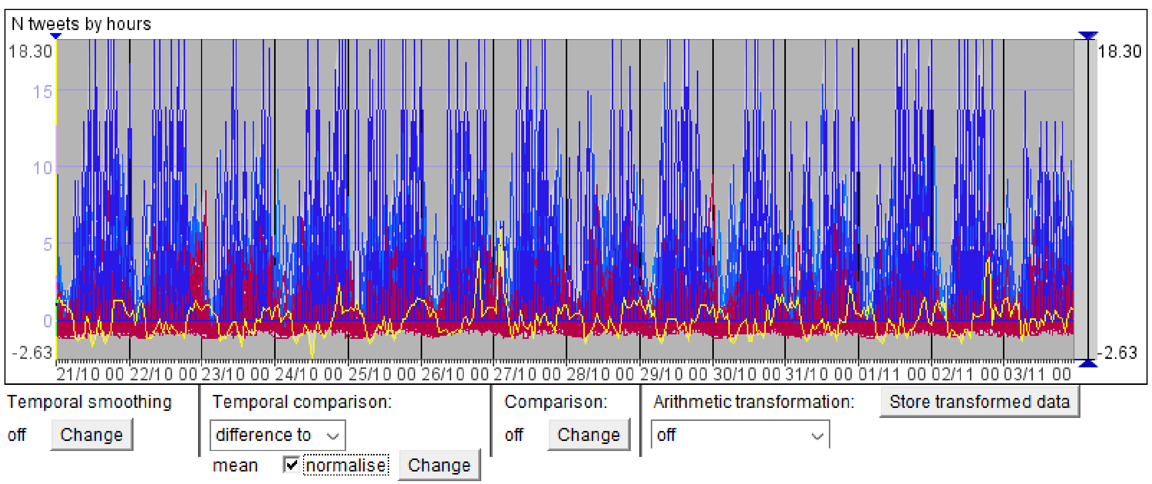
Partition-based clustering of spatial time series

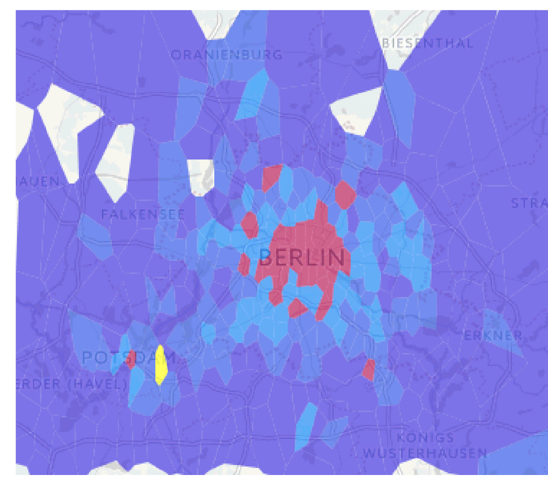
- as local time series
- as spatial situations

PBC of local time series (e.g., k-means)



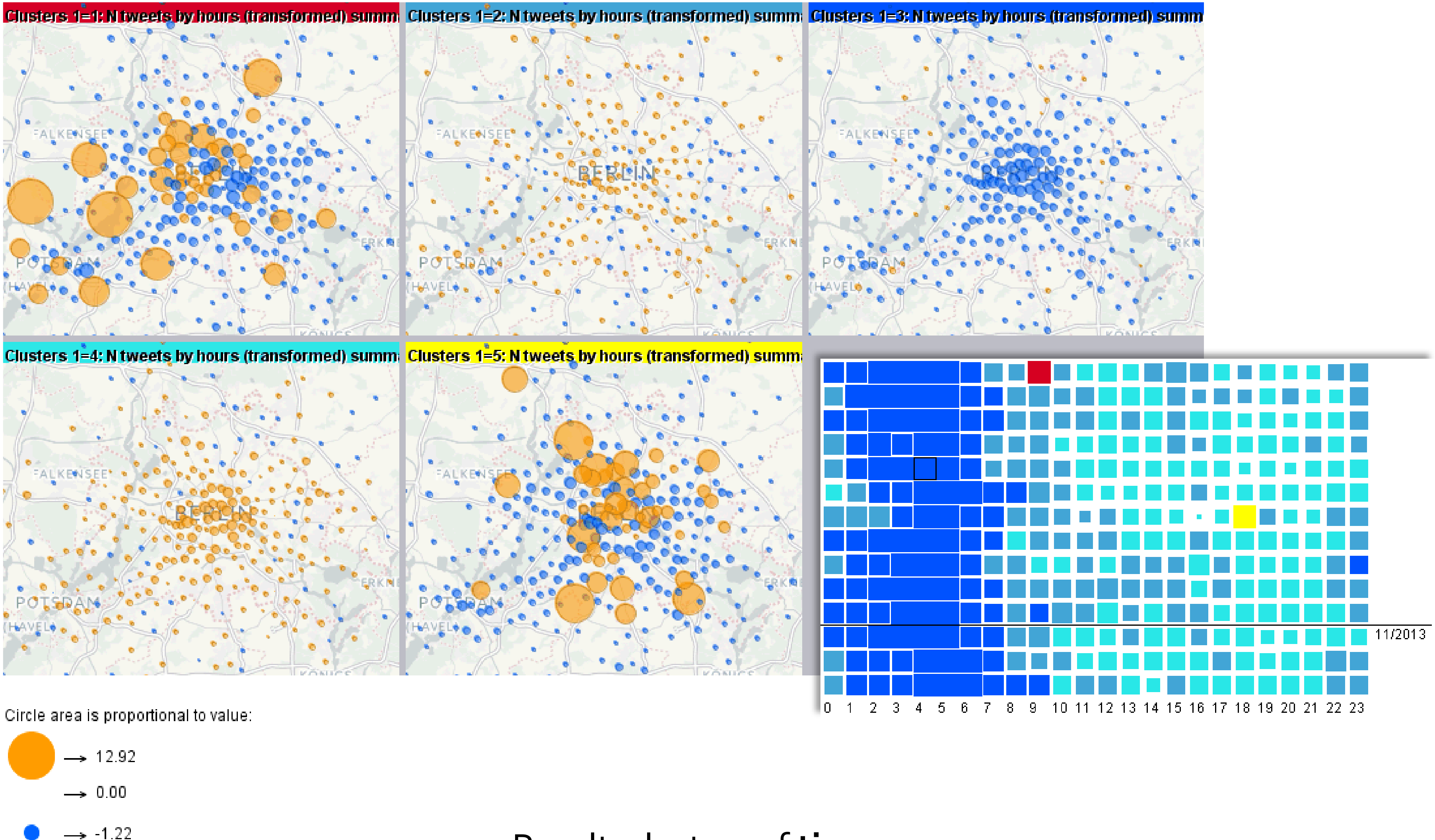






Result: clusters of **places**

PBC of spatial situations (e.g., k-means)



Result: clusters of **times**

Two-way PBC of spatial time series

times

places

	hour=21/10 00: N tweets by hours (transformed)	hour=21/10 01: N tweets by hours (transformed)	hour=21/10 02: N tweets by hours (transformed)	hour=21/10 03: N tweets by hours (transformed)	hour=21/10 04: N tweets by hours (transformed)	hour=21/10 05: N tweets by hours (transformed)	hour=21/10 06: N tweets by hours (transformed)	hour=21/10 07: N tweets by hours (transformed)	hour=21/10 08: N tweets by hours (transformed)	hour=21/10 09: N tweets by hours (transformed)	hour=21/10 10: N tweets by hours (transformed)	hour=21/10 11: N tweets by hours (transformed)
1	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09
2	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14
3	-0.50	0.61	-1.06	0.06	-1.06	0.61	0.06	0.06	-1.06	1.17	1.17	2.28
4	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11
5	-0.70	-0.70	-0.32	-0.70	-0.70	-0.70	-0.32	2.00	-0.32	1.61	-0.32	1.61
6	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09
7	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08
8	-1.04	-1.04	-0.84	-1.04	-1.04	-1.04	-1.04	-0.84	0.62	1.03	-0.01	-0.84
9	-0.72	-0.91	-0.72	-0.72	-1.09	-0.91	-1.09	-0.72	-1.09	0.76	3.34	4.45
10	-0.93	-0.93	-0.51	-0.93	-0.93	-0.93	-0.93	-0.51	-0.93	-0.09	-0.09	-0.09
11	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07
12	-0.46	-0.69	-0.69	-0.69	-0.69	-0.46	-0.69	-0.69	-0.69	-0.46	-0.69	-0.69
13	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09
14	-0.91	-0.53	-0.91	-0.91	-0.91	-0.91	-0.91	0.22	-0.53	-0.91	-0.16	-0.91
15	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11
16	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15
17	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42	-0.42
18	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	2.86	0.60	0.60
19	-0.58	-0.58	-0.58	-0.58	-0.29	-0.58	-0.58	-0.58	-0.58	-0.58	-0.58	-0.29
20	-0.92	-0.92	-0.56	-0.92	-0.92	-0.92	-0.92	0.18	-0.19	-0.19	-0.56	1.28
21	-0.58	-0.92	-0.92	-0.92	-0.92	-0.92	-0.92	-0.24	-0.58	-0.58	-0.58	0.10
22	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20
23	-0.96	-0.96	-0.96	-0.51	-0.51	-0.96	-0.96	-0.96	-0.96	-0.06	-0.51	-0.51
24	-0.51	-0.74	-1.19	-1.19	-1.19	-0.96	0.39	-0.74	-0.96	-0.51	-0.96	0.16
25	-0.40	-0.89	-0.40	-0.89	-0.89	-0.89	-0.89	-0.89	-0.89	-0.89	-0.40	-0.40

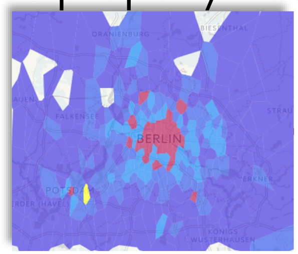
Sort by: No selection Ascending TableLens condensed Attribute...

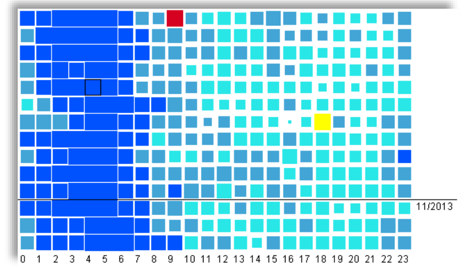
PBC can be applied to table rows or to columns

Problem: what value of k to choose?

Generally: for any computational tool, what parameter settings to choose?

- Typically not known in advance
- Computation results (such as clusters) need to be properly visualised and examined
  - Clustering results are often represented by colour-coding, which is applied to different visual objects, depending on the structure of the input data





- The analyst needs to run the tool with different settings and see how the results change
- The analyst then selects the settings bringing the “best” results:
  - easy to interpret (e.g., understandable spatial patterns)
  - internal variance within the clusters is sufficiently low
  - fit to the purpose (e.g., the intended analysis scale may require coarser or finer division)

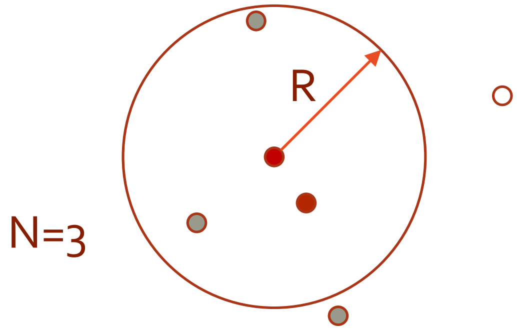
Density-based clustering

concept and parameters

Density-based clustering (DBC)

Goal: find dense groups of close or similar objects

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



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

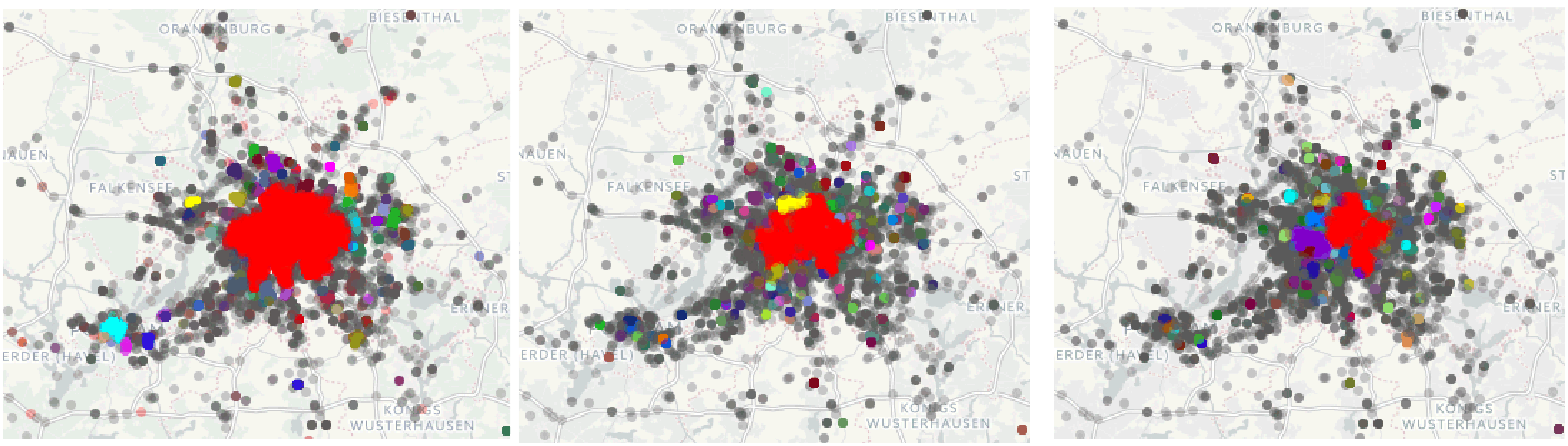
Density-based clustering

Parameters

- For DBC, the user needs to specify the neighbourhood radius (distance threshold) **R** and the minimum number of neighbours **N**.

  $\Rightarrow$ The use of DBC requires an understandable definition of **distance** between objects

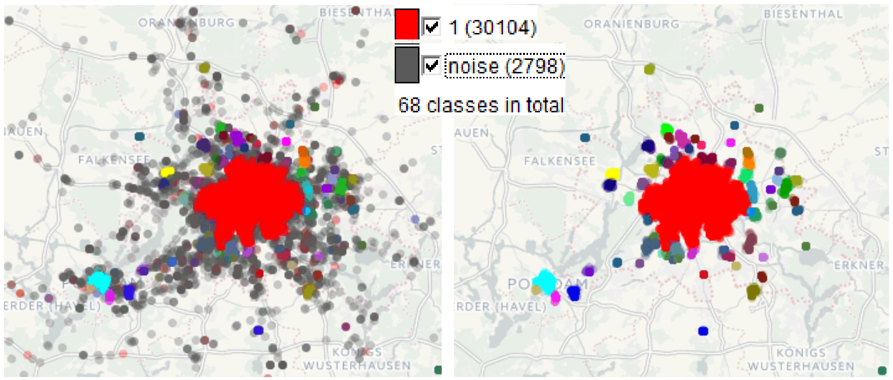
  - e.g., spatial distance or spatio-temporal distance.
- Results of DBC greatly depend on the parameter choice.
- Visualisation and interactive exploration help the analyst to find suitable values for R and N that lead to good results.



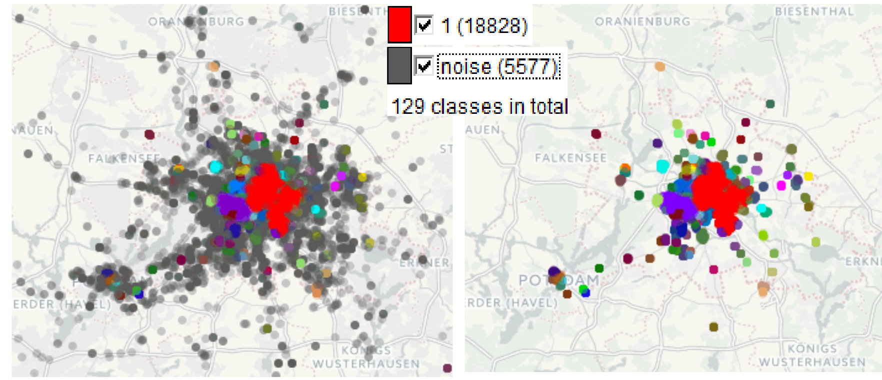
Grey: "noise"

Exploring the impact of the DBC parameters

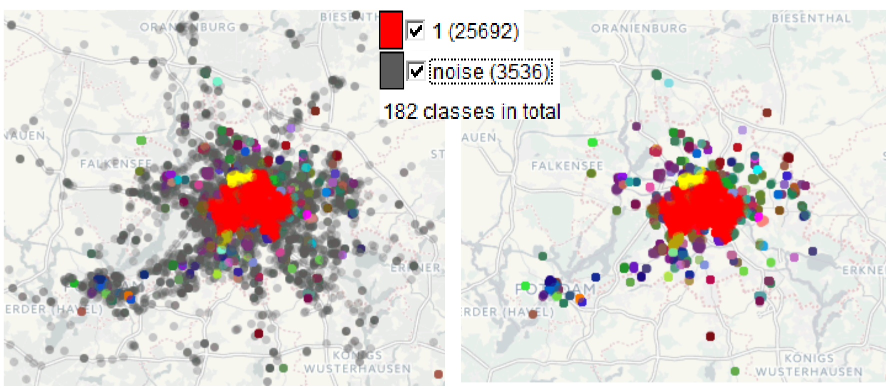
Example: spatial clusters of point objects



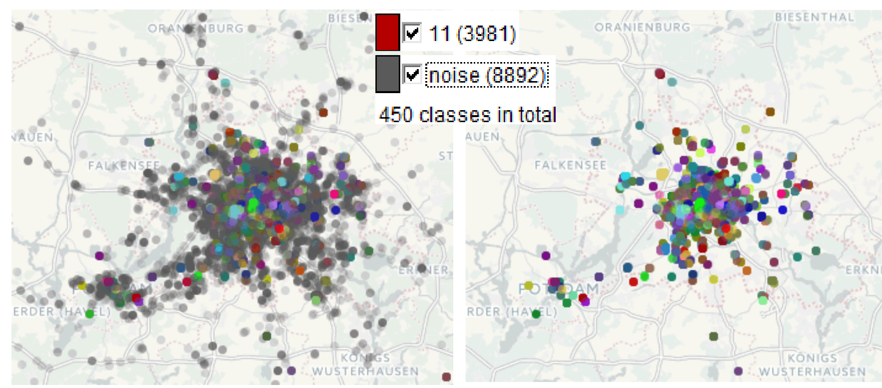
R = 500m; N = 20



R = 250m; N = 20



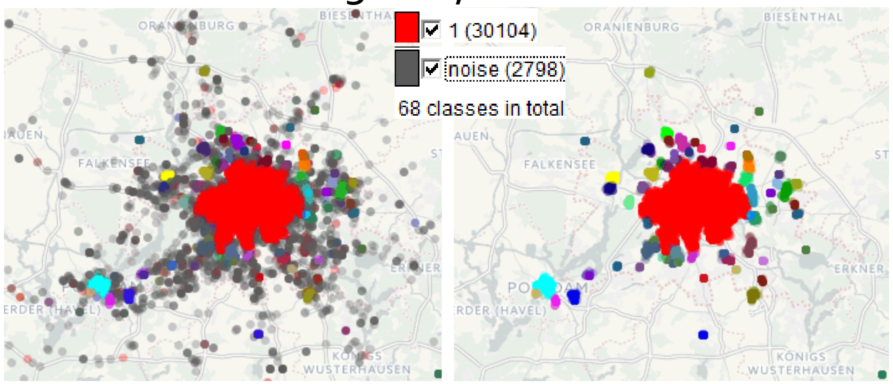
R = 250m; N = 10



R = 100m; N = 10

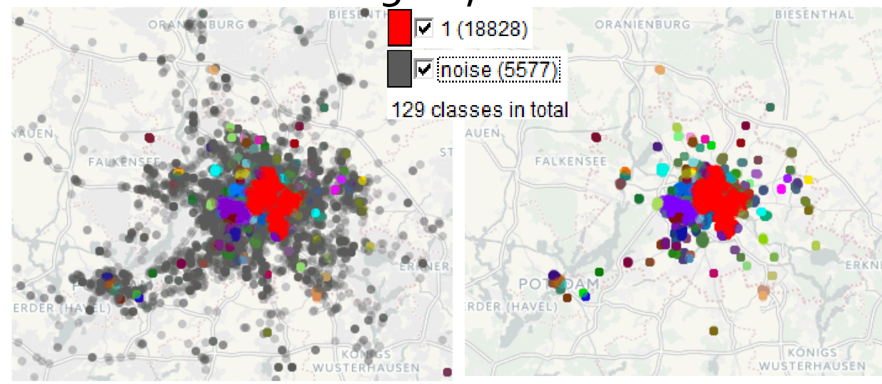
Good parameter settings?

R = 500m; N = 20



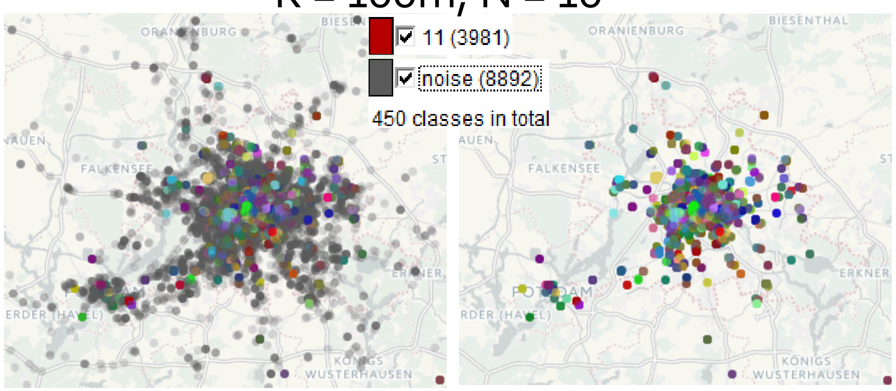
Some clusters are too loose and too extended in space.

R = 250m; N = 20



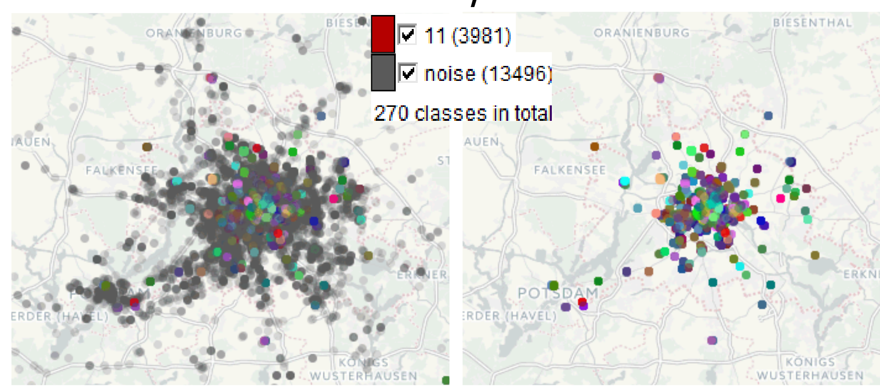
Some clusters are still too loose.

R = 100m; N = 10



Clusters are compact (but quite numerous).

R = 100m; N = 20



Clusters are compact and less numerous, but too many objects go to the noise.

Density-based clustering of spatial events

by positions in space and time

DBC by spatio-temporal distances

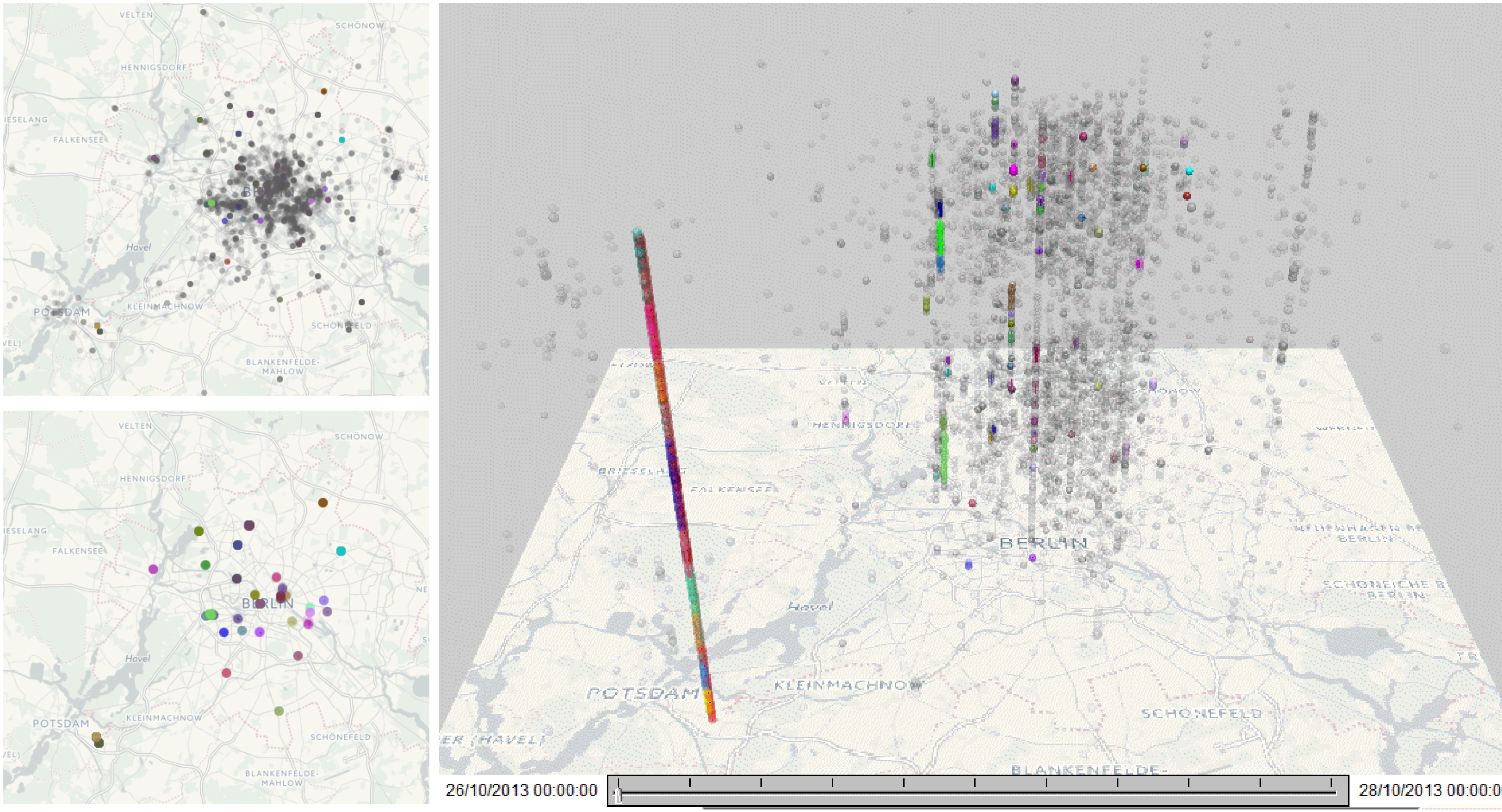
Used for finding spatio-temporal concentrations of spatial events

- For any two objects, there is a distance in space d_{space} and a distance in time d_{time}.
- To cluster the objects by their spatio-temporal proximity, the analyst may choose two neighbourhood radii R_{space} and R_{time}
  - e.g., R_{space}= 300 m and R_{time} = 30 minutes.
- However, the clustering algorithm requires a single distance and a single radius.

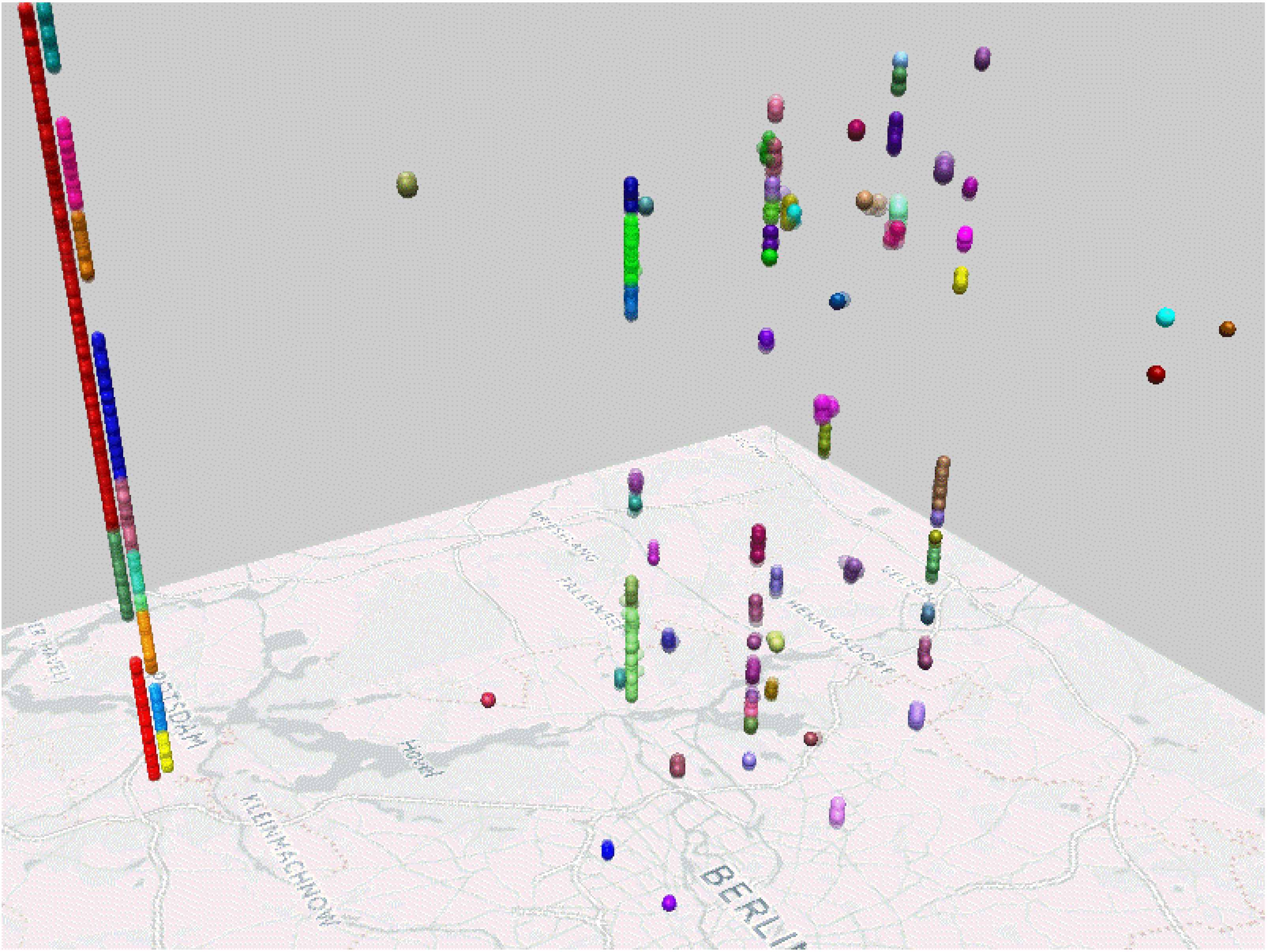
$\Rightarrow$ Spatial and temporal distances need to be combined together

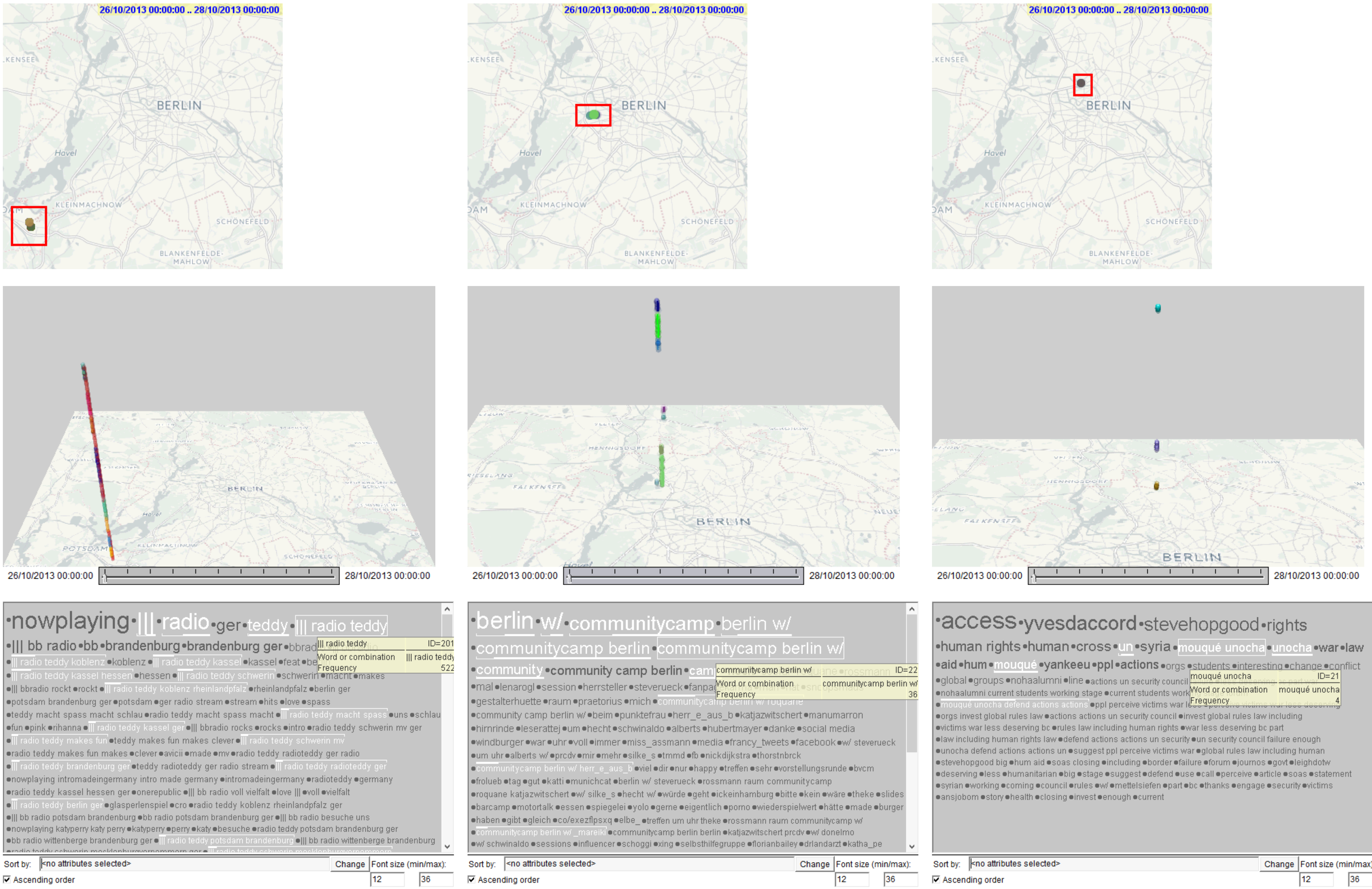
  - e.g., $d = max(d_{space} / R_{space}, d_{time} / R_{time}) * R_{space}$

Example: spatio-temporal clusters of tweet events



R_{space}= 250 m; R_{time} = 15 minutes; 8,476 events (2 days); 78 clusters with size >=10; largest cluster size: 370 events; noise: 5,897 events


POTSDAM
KLEINMACHNOW


26/10/2013 00:00:00 .. 28/10/2013 00:00:00
BERLIN
Havel
KLEINMACHNOW
SCHÖNEFELD
BLANKENFELDE-MAHLOW
26/10/2013 00:00:00
28/10/2013 00:00:00
nowplaying
radio
ger
teddy
||| radio teddy
||| bb radio
bb
brandenburg
brandenburg ger
berlin
w/
communitycamp
berlin w/
communitycamp berlin
communitycamp berlin w/
community
community camp berlin
access
yvesdaccord
stevehopgood
rights
human rights
human
cross
un
syria
mouqué unocha
unocha
war
law
aid
hum
mouqué
yankeeu
ppl
actions
Sort by:
<no attributes selected>
Change
Font size (min/max):
12
36
Ascending order

Density-based clustering: a summary

- Goal: find groups of highly similar (close) items and separate from them items that are less similar (more distant) to others.
- DBC is often applied to spatial and spatio-temporal objects
  - to find spatial and spatio-temporal concentrations of objects;
  - to find groups of objects with similar spatial or spatio-temporal properties
- Parameters:
  - distance threshold (neighbourhood radius) **R**
  - minimal number of neighbours of a cluster core object **N**
- The analyst needs to set a meaningful distance threshold

  ⇒ Well understandable distances between items must exist

  ✓ Spatial distance, temporal distance, difference between directions, …

Distance functions in DBC

- Elementary distances: spatial, temporal, difference of values of a single thematic attribute
- It may be necessary to group objects on the basis of two or more elementary distances, e.g., spatial and temporal

  $\Rightarrow$ A distance function integrating the elementary distances is needed
- General approach:
  1) Set a separate threshold for each elementary distance
  2) Transform the absolute elementary distances to relative w.r.t. the respective thresholds
  3) Combine the relative distances:
     - take their maximum or compute the Euclidean or Manhattan distance

Investigation of parameter impact

- The results of DBC greatly depend on the parameter settings (values of R and N)

$\Rightarrow$ It is necessary to run the clustering tool multiple times with different parameter settings

  - Choose clear, easily interpretable results
  - Results from different runs may complement each other and contribute to better understanding

- Interactive visual interfaces are used for investigating the results of different runs.

Two major types of clustering: a reminder

- **Partition-based clustering**: divide items into groups so that items within a group are similar (close) and items from different groups are less similar (more distant)
  - Examples: k-means, self-organizing map
  - Property of the result: each item belongs to some group
- **Density-based clustering**: find groups of highly similar (close) items and separate from them items that are less similar (more distant) to others
  - Examples: DBScan, OPTICS
  - Properties of the results: some items belong to groups, other items remain ungrouped and are treated as "noise"

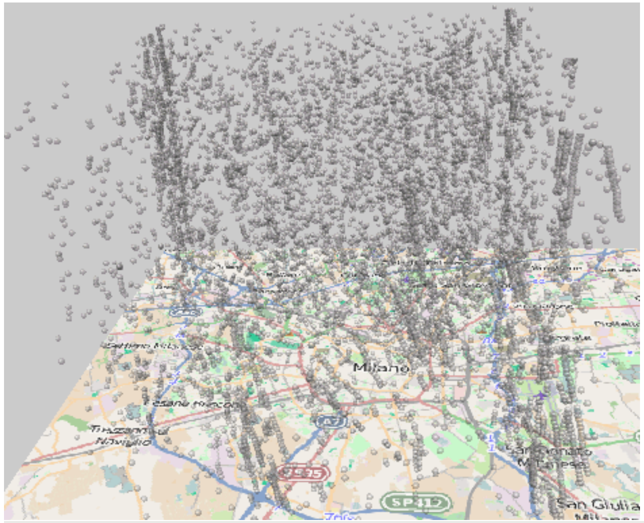
Density-based clustering of streaming events

Real-time detection and tracking of spatio-temporal concentrations of events

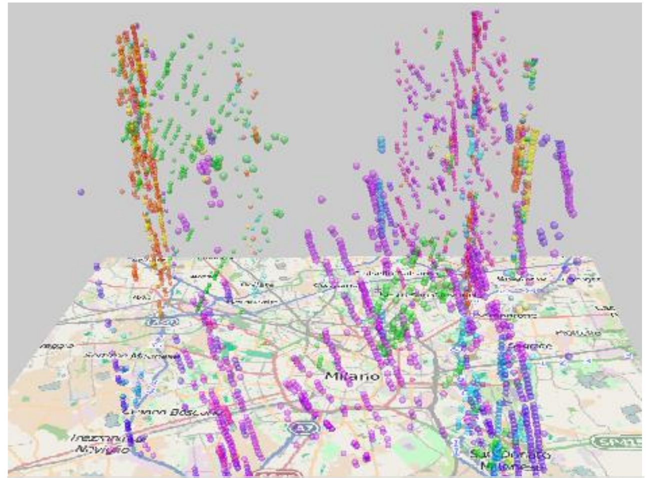
Problem setting

- Scenario:
  - Spatial events are registered in real time → stream of event data
  - Any individual event is not significant but <u>spatio-temporal concentrations (clusters)</u> of events are.
  - Monitoring task: detect emerging clusters of spatial events and <u>trace their further evolution</u>
- Our goal: support the observer
  - Automatically detect event clusters in the stream as soon as they emerge
  - Visually present the detected clusters and all their further changes to the observer

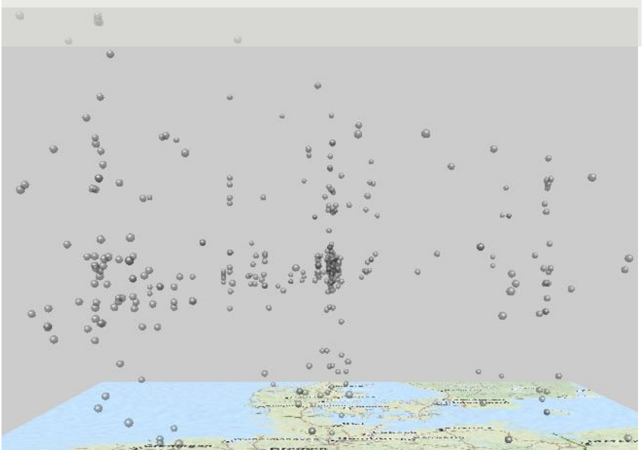
Motivating examples



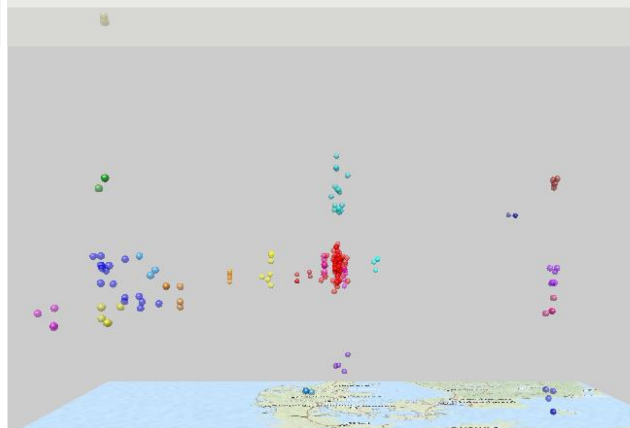
Low speed events are emitted by cars as the speed drops below a threshold.



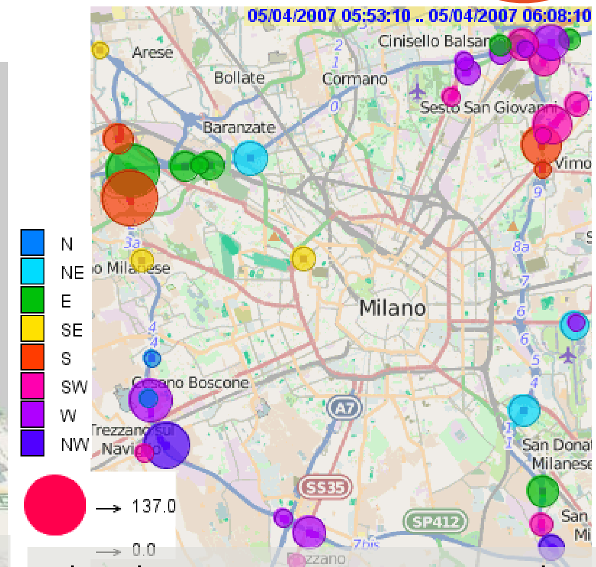
Spatio-temporal clusters of low speed events with similar movement directions may indicate traffic jams.



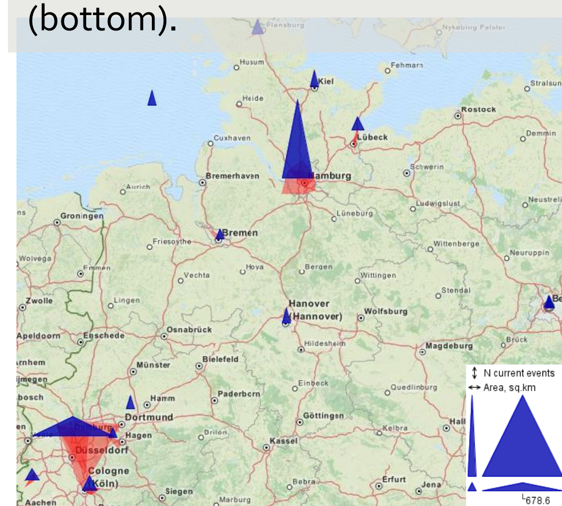
Among geo-located tweets, some may mention extreme weather conditions (e.g., a storm).



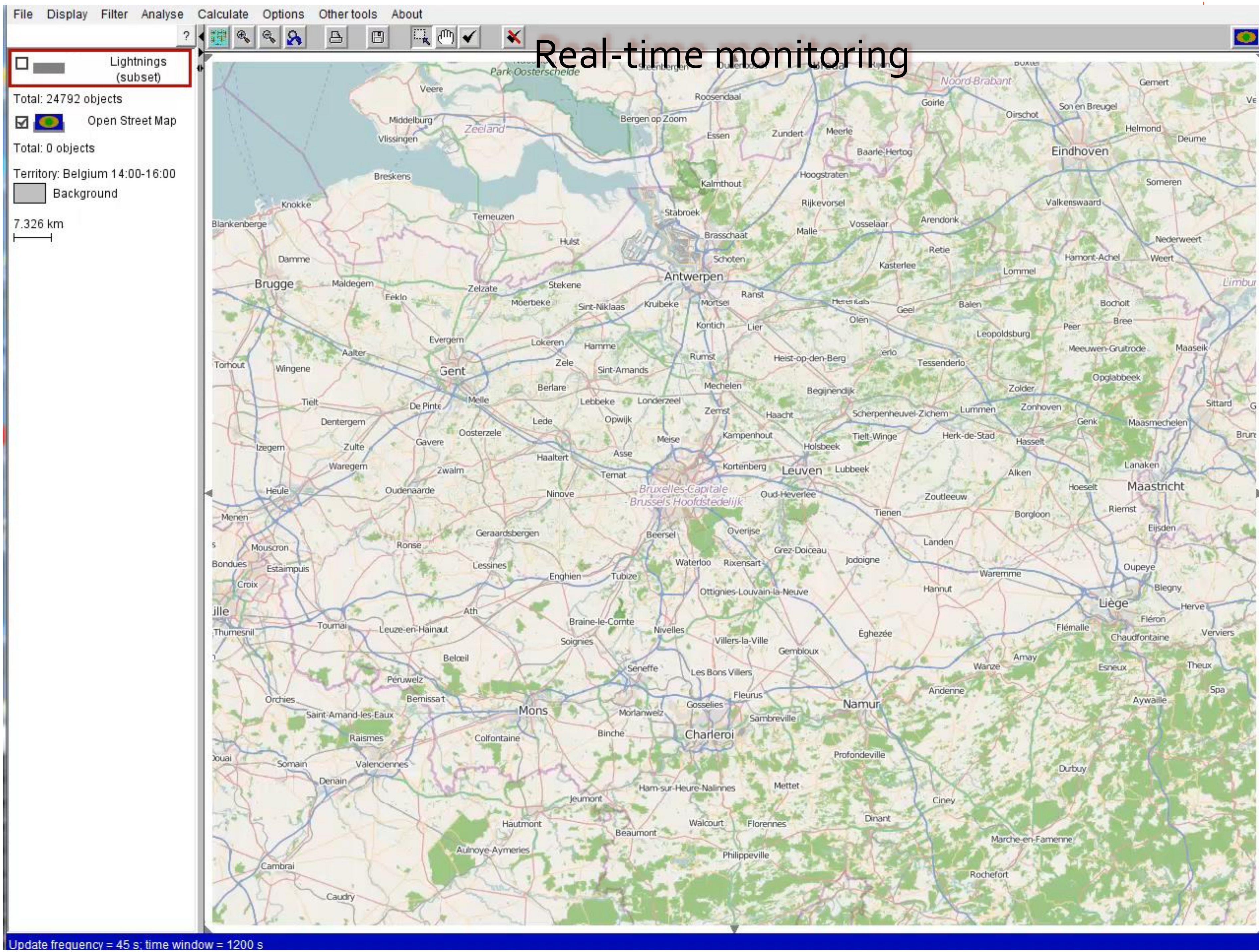
Spatio-temporal clusters of such tweets may indicate that certain regions are strongly affected by these conditions.



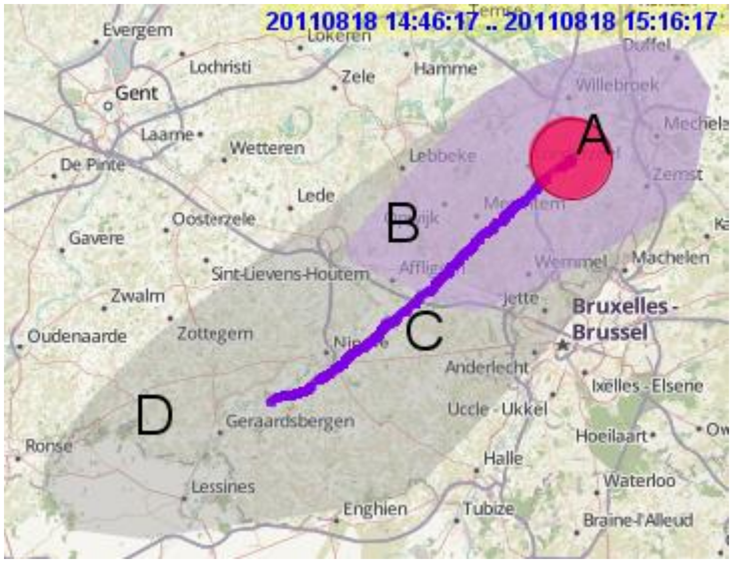
The observer wants to see not the individual events but the positions and sizes of the clusters as possible indicators of traffic jams (top) or storm-affected regions (bottom).

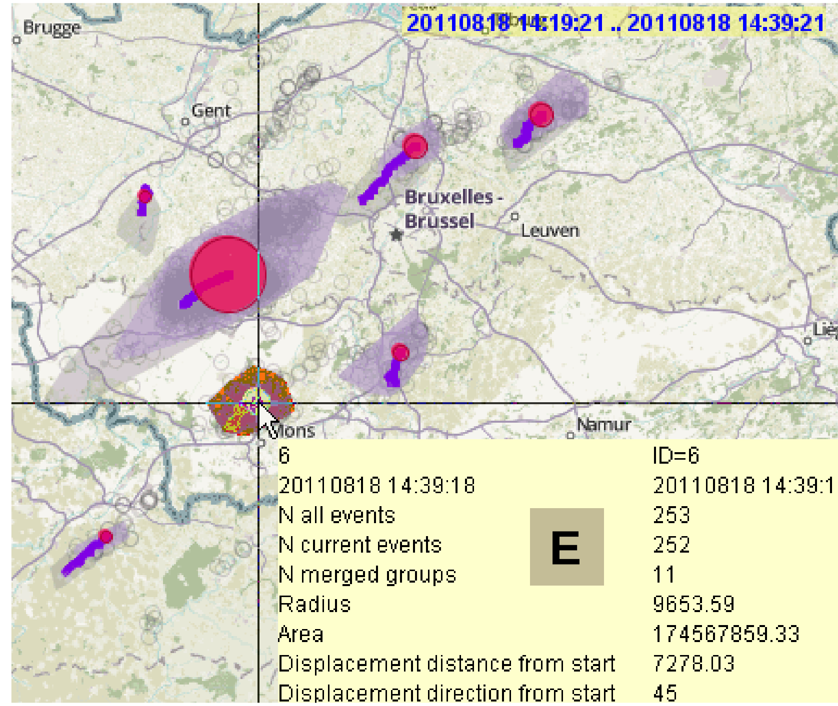


Real-time monitoring



Visualisation of a current cluster state and its recent history





A: the current position of the cluster centre;
the circle size represents the number of member events

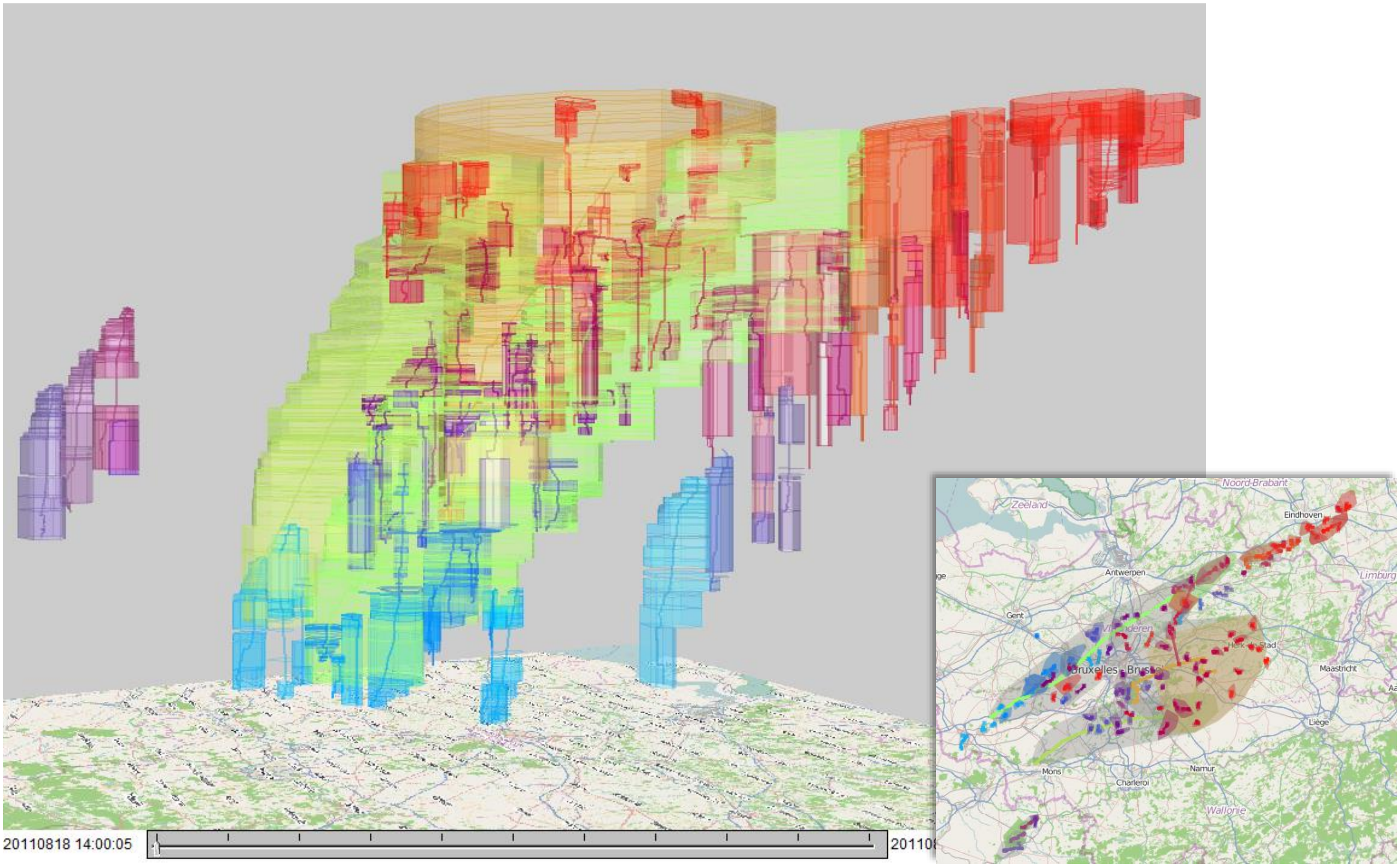
B: the spatial convex hull of the latest state of the cluster

C: the trajectory of the cluster centre made during the observation time window

D: the spatial convex hull of all cluster states attained during the observation time window

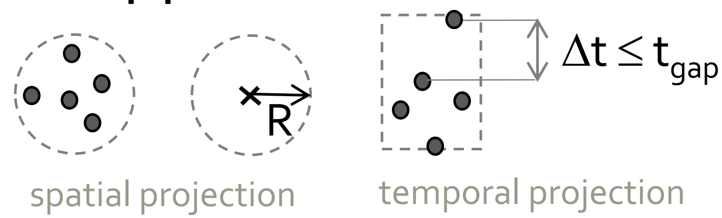
E: cluster state details are accessible through mouse pointing

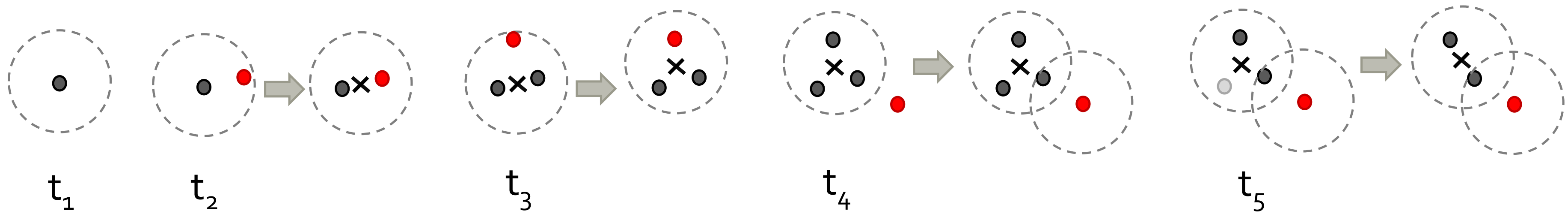
Visualisation of the cluster evolution



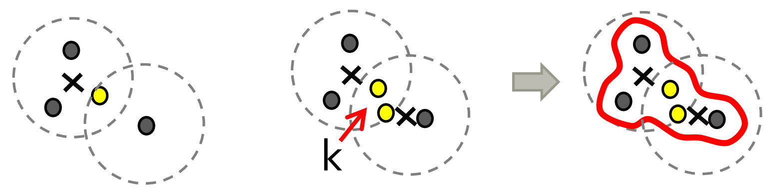
Our approach

1) Organize incoming events into micro-clusters





2) Join micro-clusters having $\geq k$ connecting events



3) Treat clusters* with $\geq N_{min}$ events as significant $\rightarrow$ present to the observer

* macro-clusters (= unions of micro-clusters) and isolated micro-clusters

4) Store cluster history (summarized cluster states)

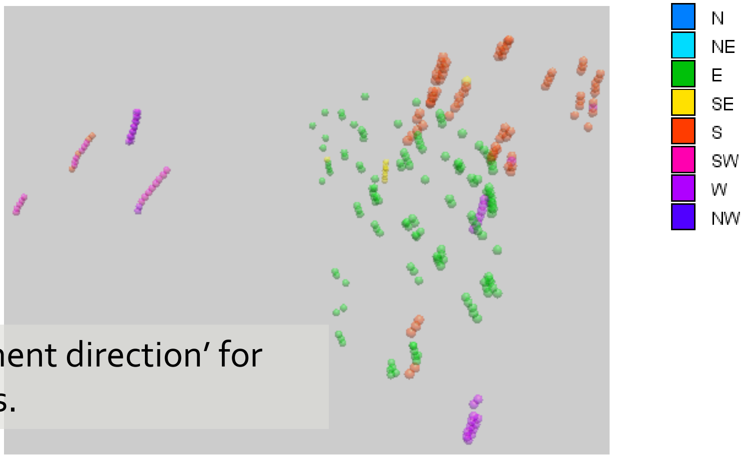
	Cluster N	Start time	Duration, minutes	N current events	N all events	N merged groups	Centre X	Centre Y	Radius	Area, sq. km.
14_0	14	28/10/2013 15:47:11	104.1	1	1	1	6.8126	51.5454	0.00	0.0
14_1	14	28/10/2013 17:31:15	25.0	2	2	1	6.7981	51.4712	8310.29	0.0
14_2	14	28/10/2013 17:56:13	9.0	3	3	2	6.7999	51.3915	17730.42	29.2
14_3	14	28/10/2013 18:05:14	19.3	2	3	2	6.7936	51.3145	9195.44	0.0
14_4	14	28/10/2013 18:24:32	0.2	3	4	2	6.7681	51.3487	13192.14	39.7
14_5	14	28/10/2013 18:24:44	3.3	5	6	3	6.8282	51.2441	29914.82	245.5
14_6	14	28/10/2013 18:28:04	8.2	6	7	3	6.8154	51.2598	31213.67	265.0
14_7	14	28/10/2013 18:36:15	56.7	5	7	3	6.8154	51.2598	24503.49	139.2
14_8	14	28/10/2013 19:33:00	31.0	6	8	3	6.9141	51.2748	26745.40	487.3
14_9	14	28/10/2013 20:04:04	2.8	7	9	4	6.7786	51.4137	30481.73	678.6
14_10	14	28/10/2013 20:06:56	0.0	8	10	4	6.7794	51.4134	29805.49	678.6
15_0	15	28/10/2013 16:42:23	54.7	1	1	1	7.4229	51.2899	0.00	0.0
15_1	15	28/10/2013 17:37:05	126.7	2	2	1	7.4418	51.3254	4150.76	0.0
15_2	15	28/10/2013 19:43:48	24.1	1	2	1	7.4607	51.3610	0.00	0.0
15_3	15	28/10/2013 20:07:55	0.0	2	3	1	7.3387	51.4001	9319.85	0.0

Extensions

1) Account for thematic attributes:

a) Add an event to a micro-cluster only if its attribute value differs from the values of the latest M events by $\leq D_{max}$.

b) Join two micro-clusters only if their connecting events satisfy condition (a) for both micro-clusters.



E.g., thematic attribute 'movement direction' for low speed car movement events.

2) Account for the number of distinct event sources:

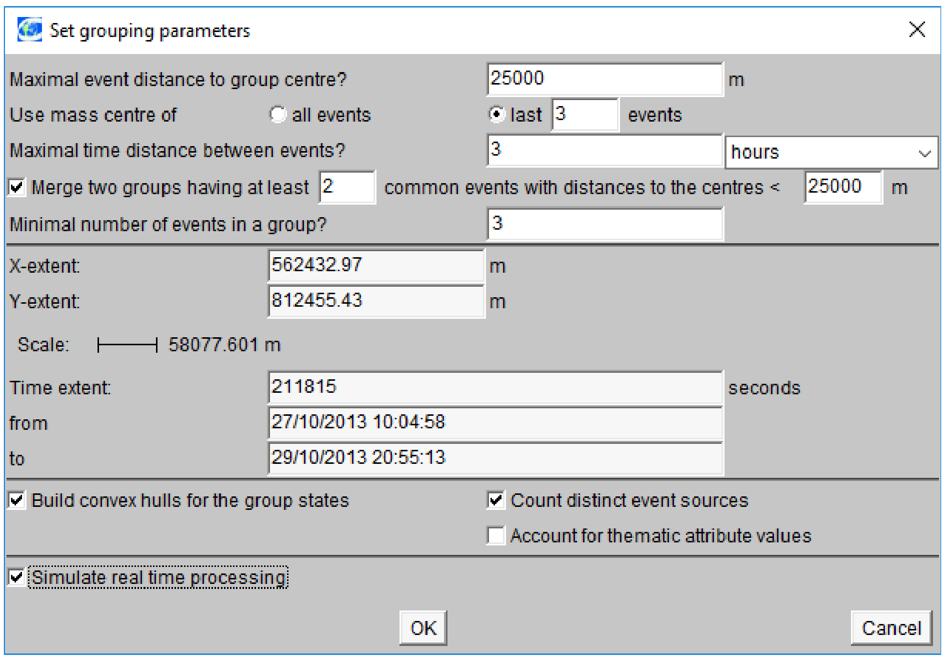
- The events have an attribute indicating the event source: car identifier, Twitter user identifier, …
- A list of distinct event sources is created and maintained for each micro and macro-cluster.
- A cluster is treated as significant only if the number of distinct event sources is $\geq$ S.

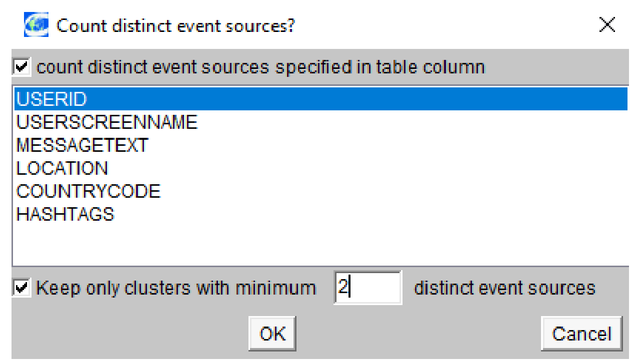
	Cluster N	Start time	Duration, minutes	Movement direction (degree)	Cardinal direction (text)	N distinct event sources	N current events	N all events	N merged groups	Radius
18_0	18	04/04/2007 05:17:22	2.30	108	E	1	1	1	1	0.00
18_1	18	04/04/2007 05:19:41	3.25	109	E	2	2	2	1	14.84
18_2	18	04/04/2007 05:22:57	1.82	110	E	2	1	2	1	0.00
18_3	18	04/04/2007 05:24:47	0.45	110	E	3	2	3	1	44.25
18_4	18	04/04/2007 05:25:15	0.03	109	E	3	1	3	1	0.00
18_5	18	04/04/2007 05:25:18	1.53	110	E	4	2	4	1	33.03
18_6	18	04/04/2007 05:26:51	0.53	109	E	5	3	5	1	35.16
18_7	18	04/04/2007 05:27:24	0.55	109	E	6	4	6	1	49.99
18_8	18	04/04/2007 05:27:58	0.57	109	E	6	5	7	1	48.14
18_9	18	04/04/2007 05:28:33	1.13	109	E	6	7	9	2	138.78
18_10	18	04/04/2007 05:29:42	0.67	107	E	8	13	15	3	232.25
18_11	18	04/04/2007 05:30:23	0.40	107	E	8	12	15	3	241.01
18_12	18	04/04/2007 05:30:48	0.05	107	E	8	13	16	3	228.74
18_13	18	04/04/2007 05:30:52	0.42	107	E	8	12	16	3	231.72
18_14	18	04/04/2007 05:31:18	0.03	109	E	8	11	16	3	215.56
18_15	18	04/04/2007 05:31:21	0.00	109	E	8	9	16	3	190.36
18_16	18	04/04/2007 05:31:22	0.00	109	E	8	1	17	1	0.00
19_0	19	04/04/2007 05:28:57	0.83	269	W	1	1	1	1	0.00
19_1	19	04/04/2007 05:29:48	0.48	269	W	2	2	2	1	27.85
19_2	19	04/04/2007 05:30:18	0.02	271	W	3	3	3	1	65.77
19_3	19	04/04/2007 05:30:20	0.50	271	W	3	4	4	1	83.13
19_4	19	04/04/2007 05:30:51	0.00	272	W	3	5	5	1	82.55

1 2

group by classes Sort by: No selection Ascending TableLens condensed Attribute...

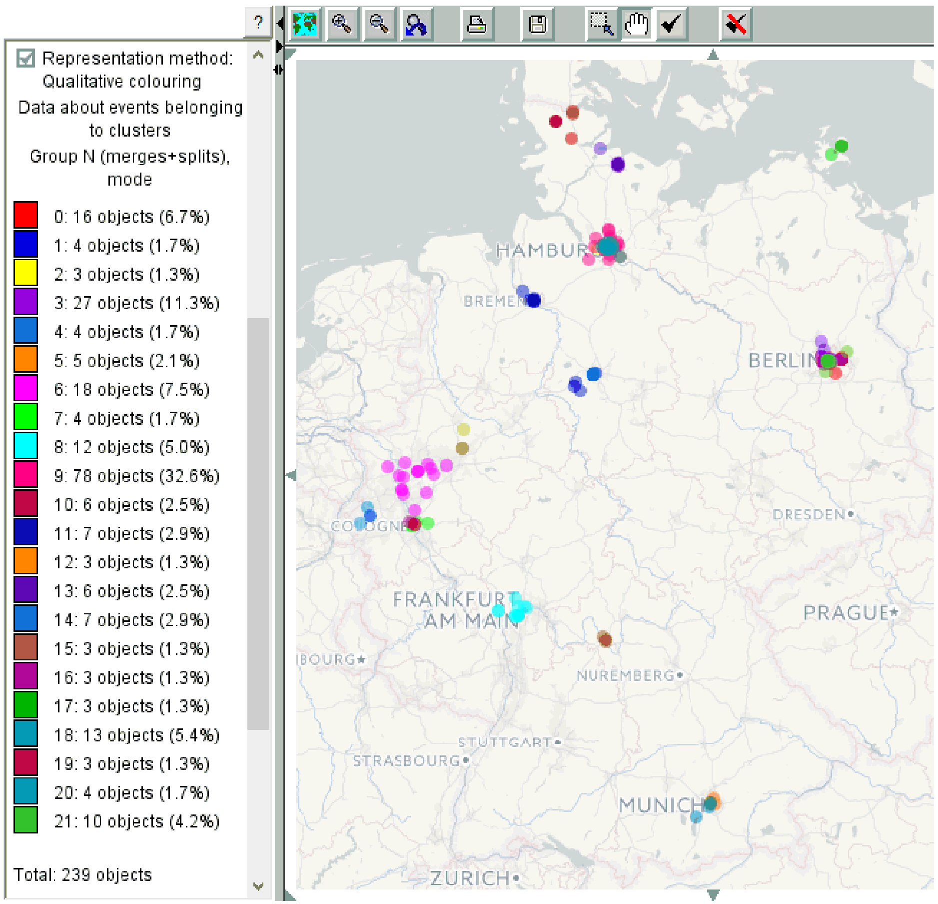
Example: simulated real time clustering of storm-related tweet events

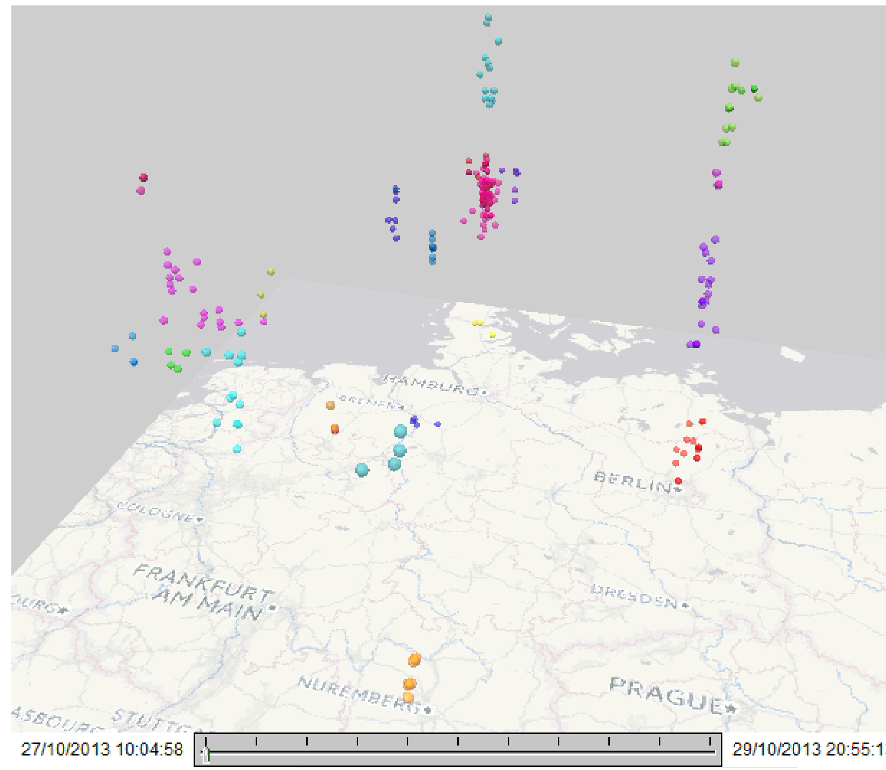




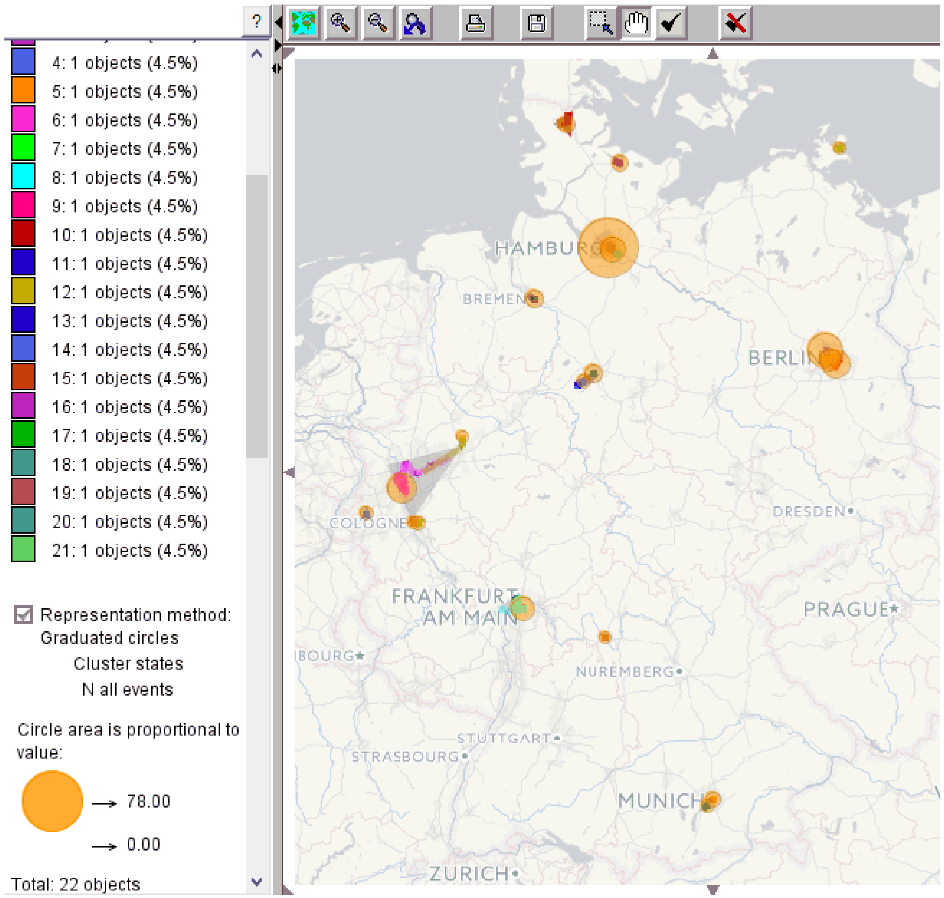
Example: simulated real time clustering of storm-related tweet events

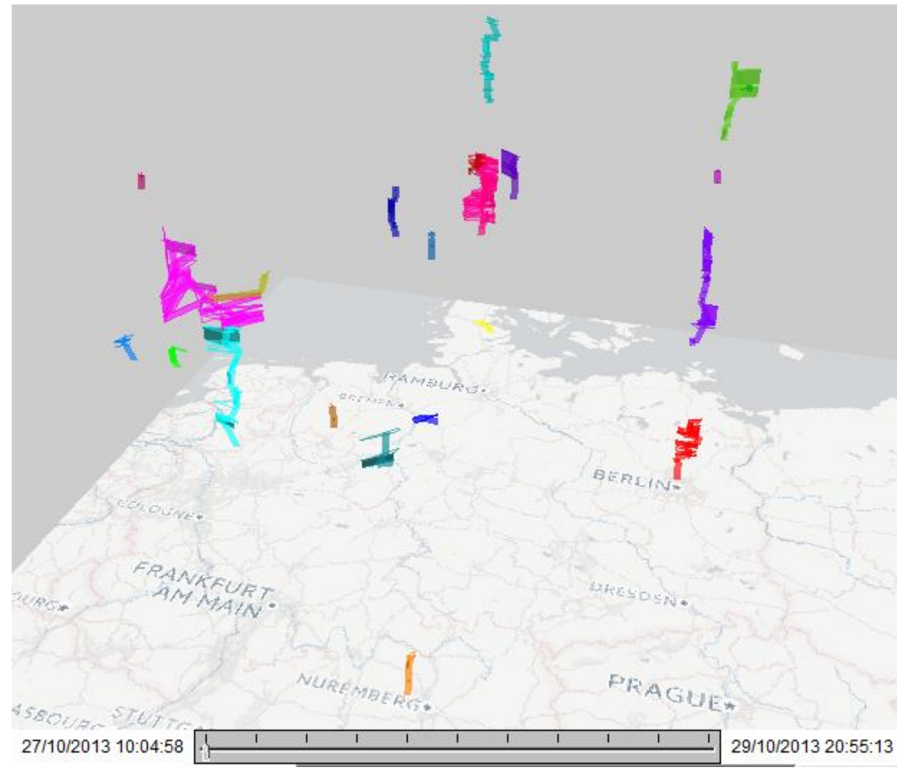
Resulting clusters



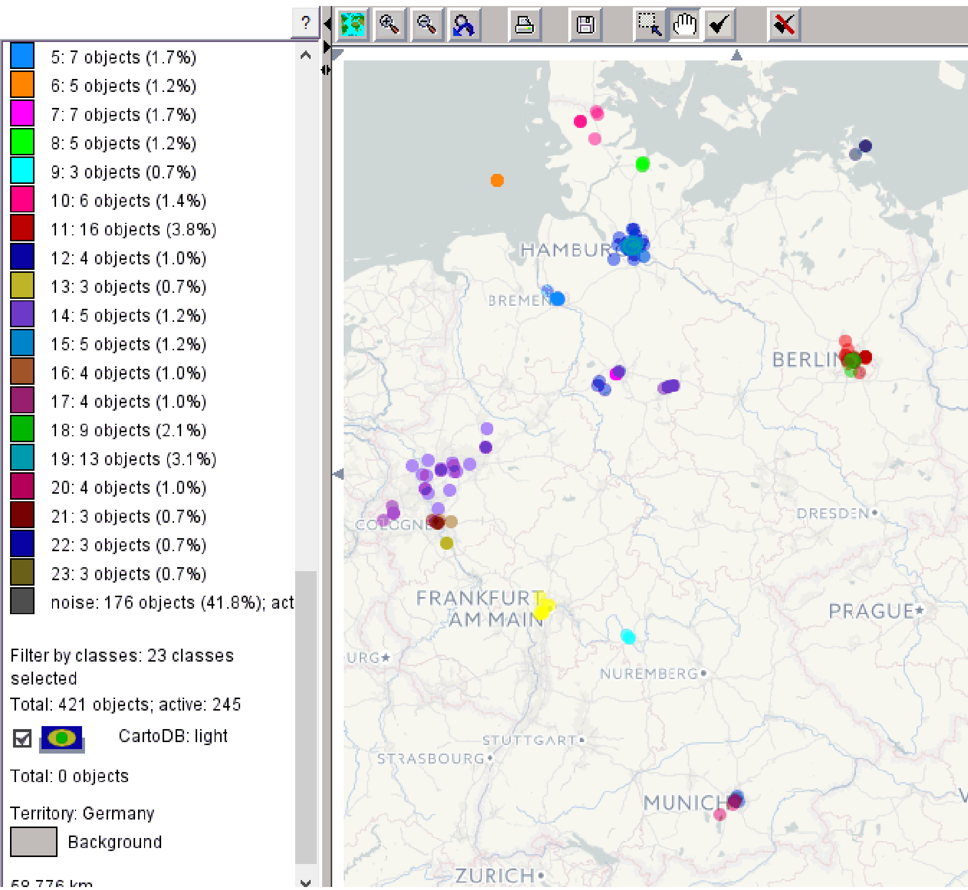


Cluster trajectories

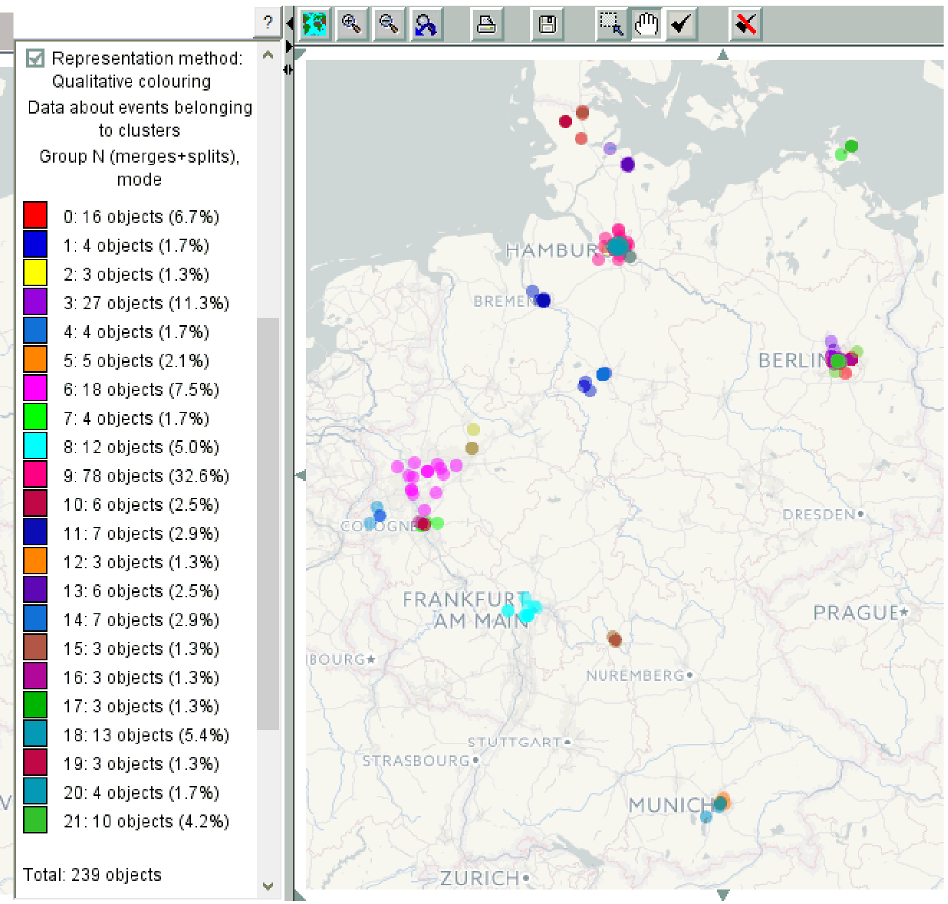




Comparison to "classical" density-based clustering (OPTICS; not accounting for N of distinct sources)

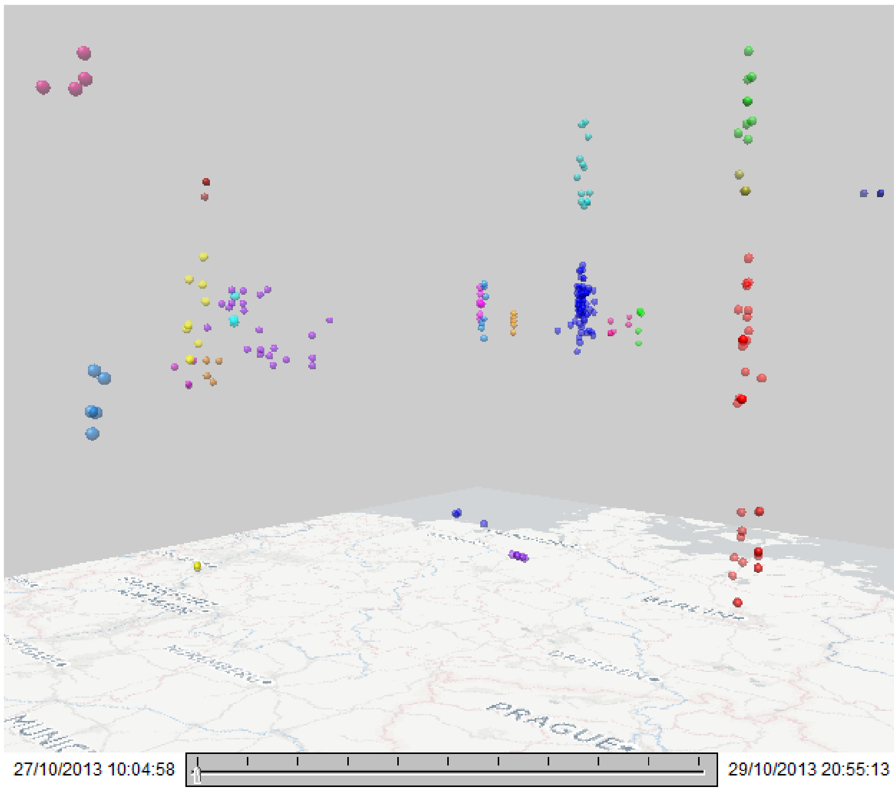


OPTICS

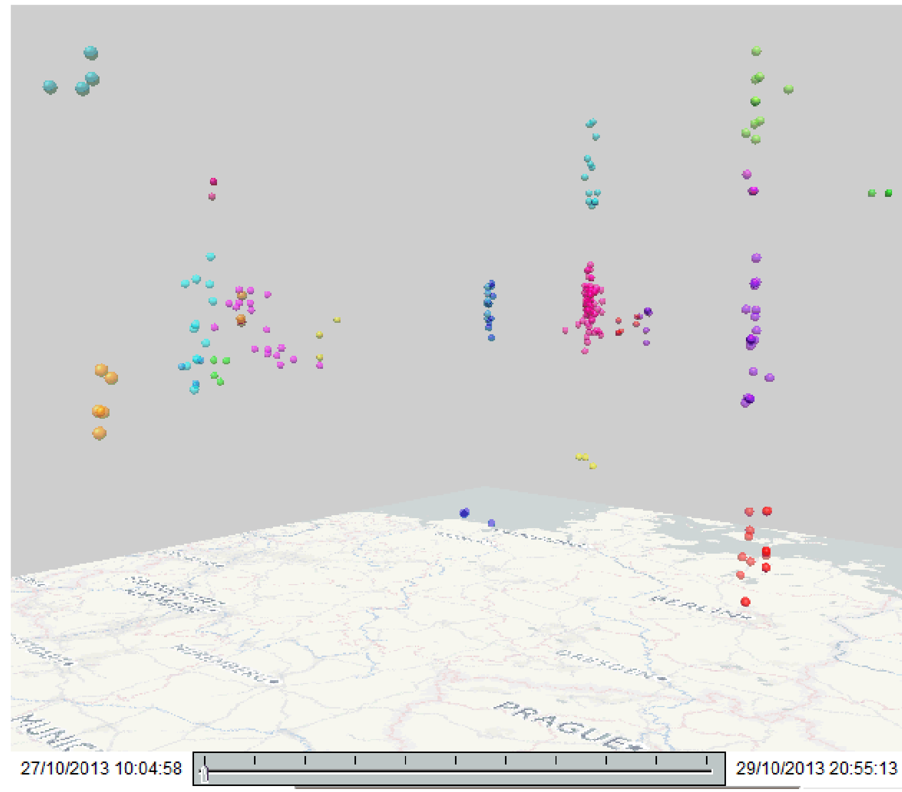


Real-time

Comparison to "classical" density-based clustering (OPTICS; not accounting for N of distinct sources)



OPTICS



Real-time

Summary

- Types of spatio-temporal data: spatial events, trajectories (quasi-continuous, episodic), spatial time series
- Two aspects of spatial time series: local time series referring to places vs. spatial situations referring to times
- Transformations of spatio-temporal data
- Partition-based vs. density-based clustering
- Two-way partition-based clustering of spatial time series:
  - clustering of places by similarity of local time series
  - clustering of times by similarity of spatial situations
- Density-based clustering of spatial events for detecting spatio-temporal concentrations
- Density-based clustering of streaming events