Lecture 04: Use of clustering in visual data analysis
Exercise description

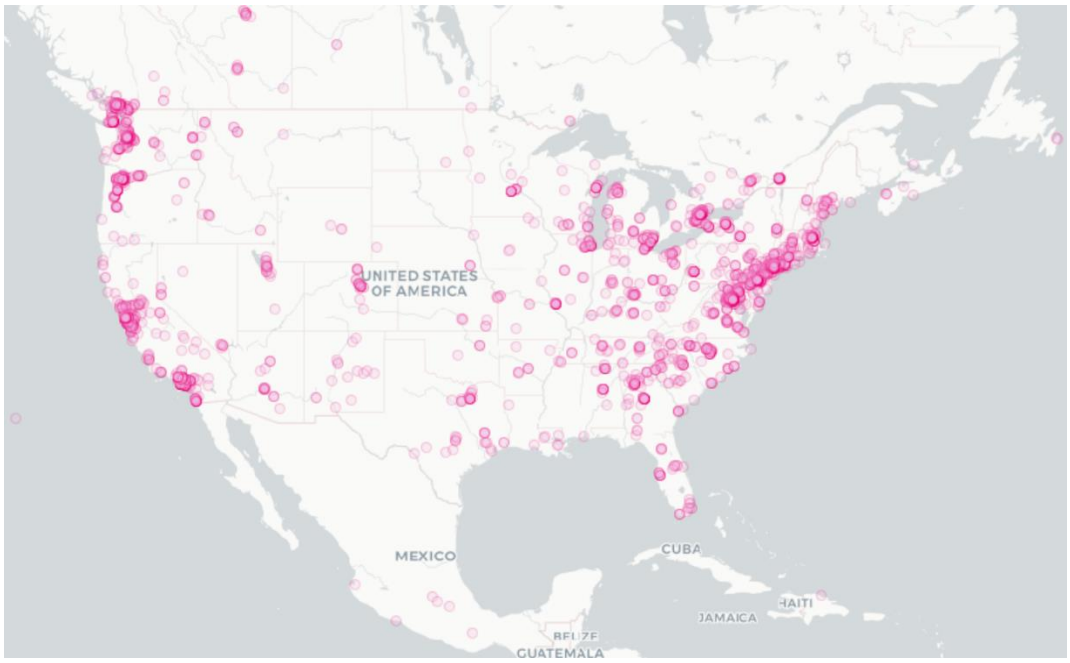
Goals

- Understand the concepts of density-based and partition-based clustering, role and impacts of their parameters.
- See the possibility of using different distance functions in density-based clustering.
- Understand the purposes of using density-based and partition-based clustering in data analysis.

Data

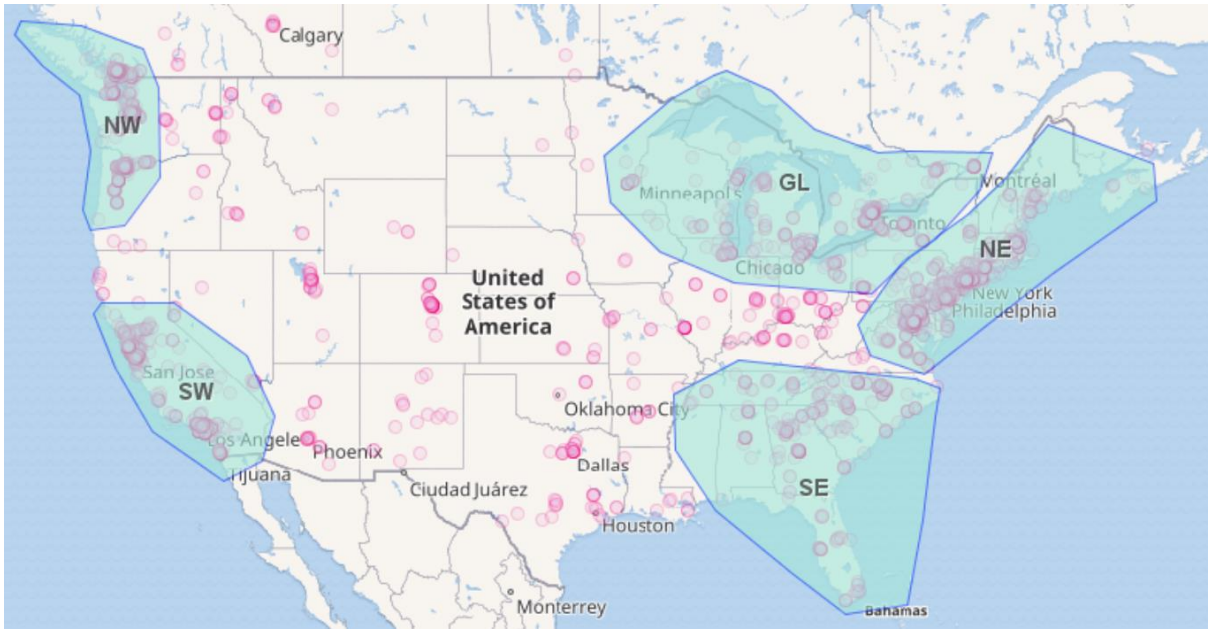
Photo taking events

The data to be used in the exercise are metadata records of geolocated photos taken on the territory of the North America in the years from 2007 till 2016 that were published online through the photo hosting service flickr and have keywords related to ***cherry blossoming*** in the titles or tags attached. The whole dataset containing 81,908 records was filtered to minimize the possible impact of some flickr users who tend to publish large sets of photos taken consecutively in the same place. To do the filtering, we (1) linked the records for each combination User+Date into a sequence and (2) took from each sequence the first record and, additionally, the records whose distances from the previous records were at least 1 km (i.e., the user moved by at least 1 km before taking the next photo). The resulting filtered dataset consists of 9,704 records, which include geographic coordinates (longitude and latitude), date and time when the photos were taken (field DateTaken), title (field NAME), tags, and some other fields. The geographic locations of the photos are shown on the map below. The original full dataset is also available and can be explored later.



Aggregated data: weekly counts by regions

For the partition-based clustering, we prepared a dataset where the photo taking events are aggregated by 5 regions and weekly time intervals. The outlines of the regions are shown in the map below. The data records (table rows) correspond to the weekly intervals, and there are 5 columns with the event counts in the regions.



Tasks

Tasks performed with density-based clustering

Data: records of individual photo taking events.

Analysis tasks:

1) find events of mass engagement in taking photos of cherry blossoms, i.e., areas and time periods in which many people took photos of cherry blossoms;
2) investigate long-term temporal patterns of the event occurrence in different areas: find out whether mass photo taking events re-occurred in the same areas in different years and, if so, compare the event times in different years.

Please note that mass photo taking events are expected to be extended in space and time; e.g., one event may be spread over a city and last for several days. Such events are manifested in the data as spatio-temporal clusters of elementary photo taking events, i.e., many photos taken at close locations and times.

The tasks are to be fulfilled using density-based clustering with different distance functions (namely, spatio-temporal distance and spatial distance) and different parameter settings.

Tasks performed with partition-based clustering

Data: per-region weekly counts of individual photo taking events.

Analysis tasks:

1) find and characterise patterns in the overall temporal variation of the photo taking across the regions;
2) investigate differences between the years in terms of the times and amounts of photo taking activities across the regions.

In fulfilling the tasks, you are supposed to experiment with partitioning the data into different numbers of clusters and observing the impacts on the observable patterns.

A possible analysis scenario involving density-based clustering

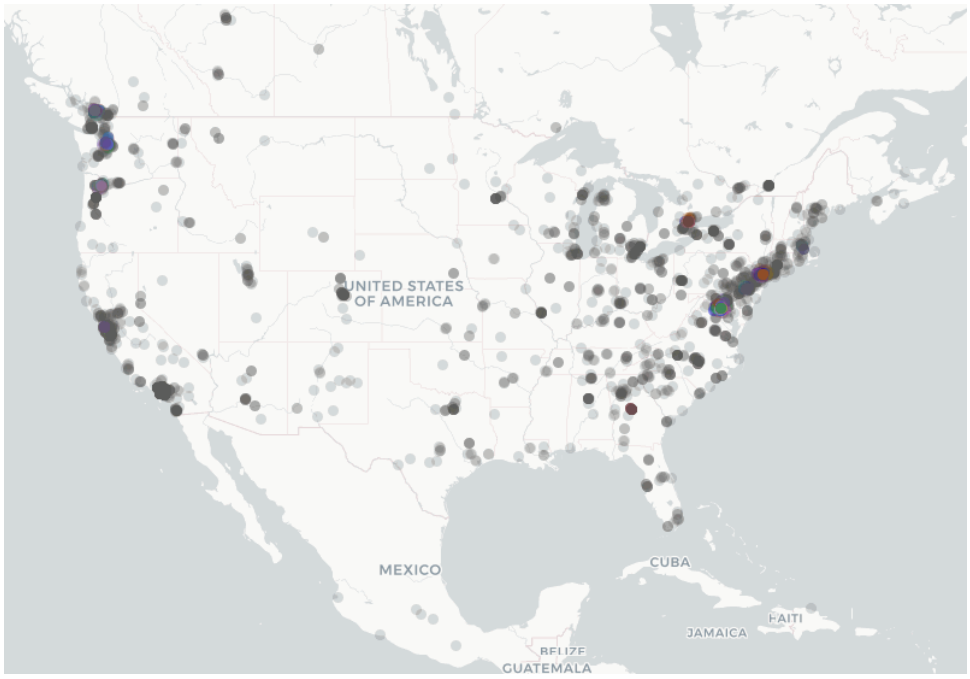
Finding mass photo taking events

To fulfil the first analysis task, you need to utilize density-based clustering with the spatio-temporal distance function. This requires setting 2 distance thresholds, spatial and temporal. According to the spatial scale of the analysis and the expected spatial and temporal extents of the mass photo taking events, both thresholds should be quite large, but it is unknown what specific values will work well enough. You therefore need to perform clustering several times with different settings and then decide which result to retain.

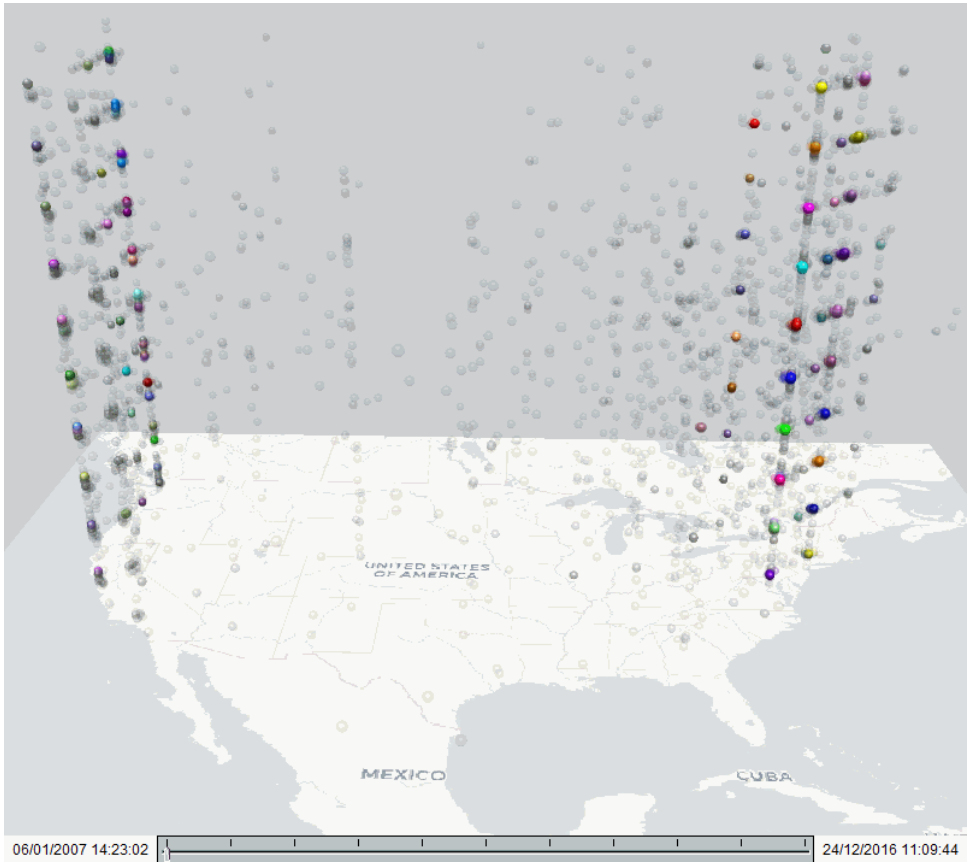
First clustering attempt

To begin with, we propose a combination of 30 km (spatial distance threshold) and 3 days (temporal distance threshold). Besides, you need to specify the minimal density threshold, i.e., the minimal number of neighbours required for a core object of a cluster. You can start with setting this parameter to 3, which is quite a loose constraint.

Density-based clustering puts some objects in clusters and labels other objects as “noise”, which means that these objects are not close enough to any cluster and do not have a sufficient number of neighbours for making another cluster. The maps below show how an outcome of density-based clustering may look like. In the first map, the events labelled as “noise” are in grey and the events put in clusters are painted in colours assigned to the clusters. In the second map, the “noise” is filtered out and only the clusters are shown. It can make sense to discard clusters with too few members. Thus, in the following examples, we discard clusters with less than 10 elementary events (i.e., the members of these clusters are treated as belonging to the “noise”).



The same clustering results are shown below in a space-time cube display:



It may be convenient to see clustering results in an aggregated form, e.g., as is shown in the table below:

identifie	N events	Begin time	End time	Duration, days	Area, sq.km
Cluster 15	99	17/02/2015 09:04:58	19/04/2015 11:48:42	61.11	273
Cluster 48	110	31/03/2011 17:00:52	12/05/2011 10:54:45	41.75	967
Cluster 11	414	09/03/2012 16:39:47	16/04/2012 21:37:05	38.21	1242
Cluster 16	84	02/04/2009 16:15:18	09/05/2009 10:52:06	36.78	2538
Cluster 32	99	16/03/2012 23:18:20	21/04/2012 14:57:21	35.65	837
Cluster 6	205	19/03/2015 10:49:06	22/04/2015 17:17:08	34.27	459
Cluster 38	100	07/04/2011 11:17:07	11/05/2011 08:20:49	33.88	675
Cluster 24	120	05/04/2013 12:33:49	08/05/2013 19:19:24	33.28	2047
Cluster 9	397	19/03/2013 12:06:58	21/04/2013 12:56:34	33.03	800
Cluster 2	471	12/03/2011 17:58:25	13/04/2011 20:08:15	32.09	1676
Cluster 42	79	13/04/2014 16:13:00	15/05/2014 08:45:44	31.69	1198
Cluster 49	48	17/03/2011 10:50:42	17/04/2011 16:05:51	31.22	135
Cluster 19	39	21/02/2010 12:15:00	24/03/2010 09:39:41	30.89	391
Cluster 8	393	16/03/2010 17:39:49	16/04/2010 08:25:42	30.62	1568
Cluster 44	275	17/03/2008 21:29:10	17/04/2008 07:43:26	30.43	1556
Cluster 10	341	14/03/2009 13:26:01	13/04/2009 12:39:32	29.97	1283
Cluster 21	54	04/04/2009 11:22:03	03/05/2009 21:49:56	29.44	590
Cluster 26	54	06/04/2010 15:52:26	05/05/2010 15:05:39	28.97	301
Cluster 45	62	20/03/2012 18:19:07	17/04/2012 19:25:00	28.05	516
Cluster 12	65	31/03/2010 11:54:35	27/04/2010 22:40:07	27.45	1341
Cluster 23	91	15/04/2015 16:42:05	12/05/2015 10:23:04	26.74	1209
Cluster 22	75	08/04/2008 17:30:06	04/05/2008 15:30:41	25.92	870

group by classes Sort by: Duration, days Descending TableLens condensed Attribute...

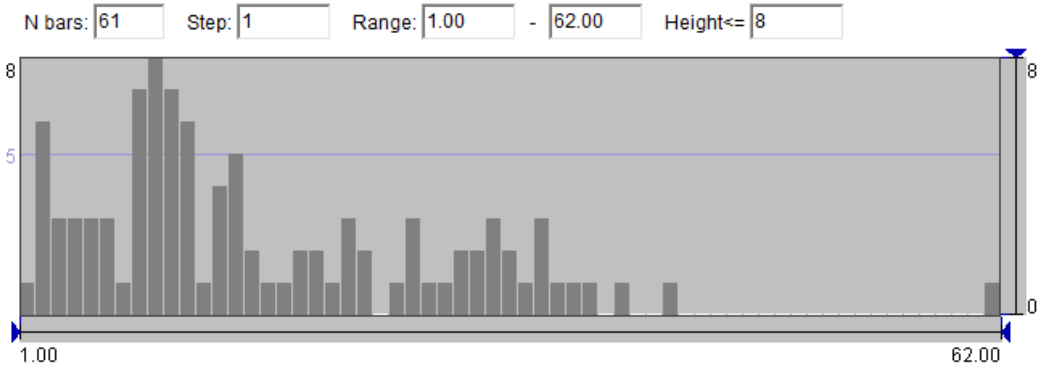
Each cluster is characterised by the number of the elementary events, the life time (i.e., the time from the first to the last event of this cluster), the duration, and the area covered in space. The area can be estimated by building a convex hull around each cluster and calculating the area of the convex hull.

The clusters in the table are sorted by the durations in the decreasing order. The image below shows the bottom part of the table, which describes clusters having short duration:

identifie	N events	Begin time	End time	Duration, days	Area, sq.km
Cluster 46	10	21/03/2014 13:03:49	26/03/2014 14:46:06	5.07	15
Cluster 67	13	10/04/2008 11:25:45	14/04/2008 23:55:56	4.52	74
Cluster 92	12	18/04/2014 21:04:50	23/04/2014 04:15:16	4.30	2
Cluster 61	13	22/03/2012 14:18:57	26/03/2012 17:38:28	4.14	172
Cluster 96	10	16/03/2007 17:12:24	20/03/2007 12:55:17	3.82	58
Cluster 17	34	18/04/2013 12:28:10	21/04/2013 16:31:05	3.17	23
Cluster 65	35	15/04/2011 17:24:01	18/04/2011 21:23:11	3.17	10
Cluster 47	12	10/04/2009 13:56:40	13/04/2009 12:21:41	2.93	125
Cluster 93	12	10/04/2008 17:58:48	13/04/2008 15:18:32	2.89	0
Cluster 82	19	18/04/2008 16:39:34	21/04/2008 06:04:56	2.56	0
Cluster 94	15	09/04/2011 11:08:21	11/04/2011 18:45:04	2.32	6
Cluster 86	13	21/04/2007 11:03:43	23/04/2007 16:53:02	2.24	1
Cluster 43	10	04/04/2009 10:41:36	06/04/2009 15:08:06	2.19	263
Cluster 85	13	14/04/2012 08:28:25	15/04/2012 15:33:32	1.30	0

group by classes Sort by: Duration, days Descending TableLens condensed Attribute...

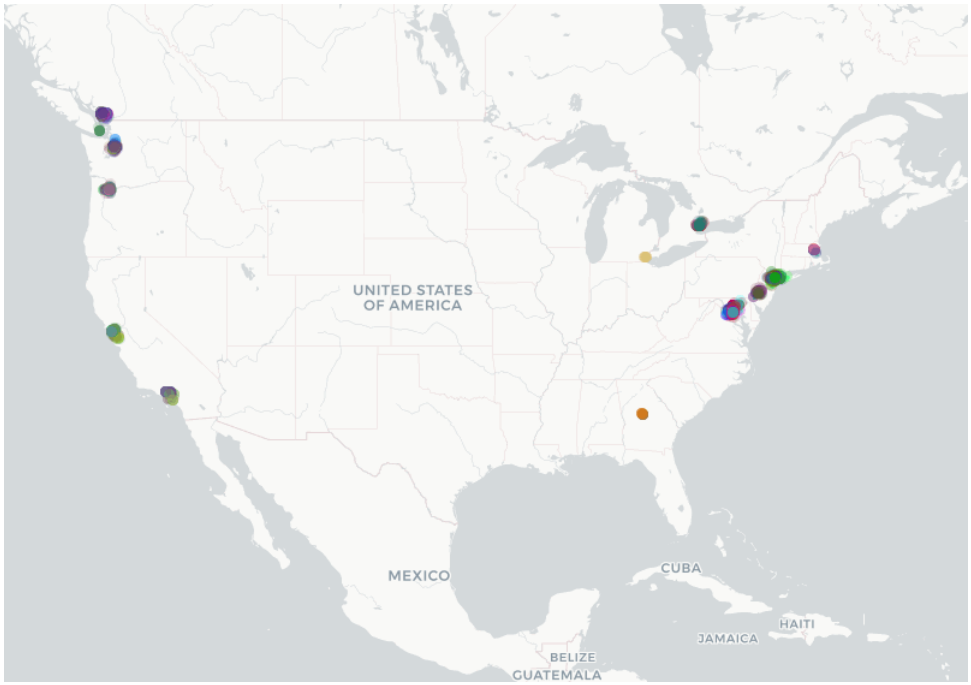
It may be useful to look at a frequency histogram of the cluster durations:



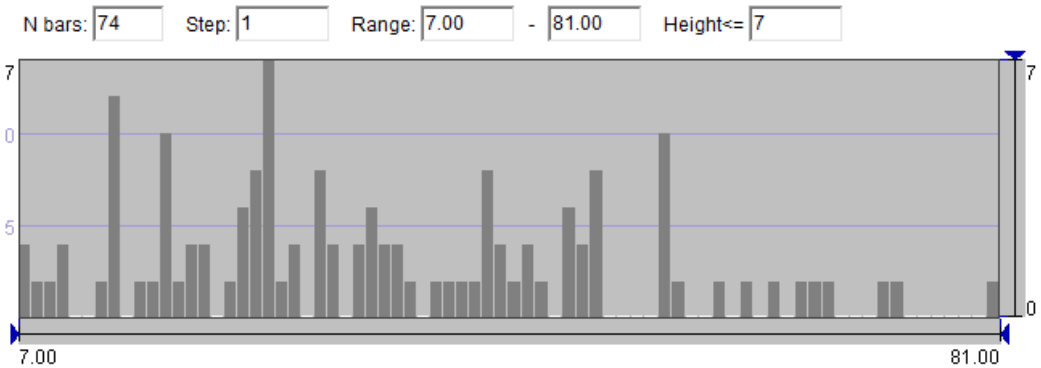
Modifying the temporal distance threshold

The statistical distribution of the cluster durations suggests that the events of mass interest to cherry blossoming may extend in time over weeks; at the same time, there are many clusters with quite short durations. A possible reason is interruptions in photo taking activities due to bad weather. There is a change that increasing the temporal distance threshold may help us to capture larger events that were occasionally interrupted. Another consideration to take into account is that many people might not have time for taking photos of cherry blossoms during their working days and could do this only on weekends. Therefore, it makes sense to increase the temporal distance threshold to, say, 7 days.

After performing the new run of clustering, compare the results to those of the previous run in terms of the number and cardinalities (i.e., counts of members) of the clusters and the amount of the noise. Also, look at the map and space-time cube, if possible, to see if there are clusters in places where no clusters were detected earlier.

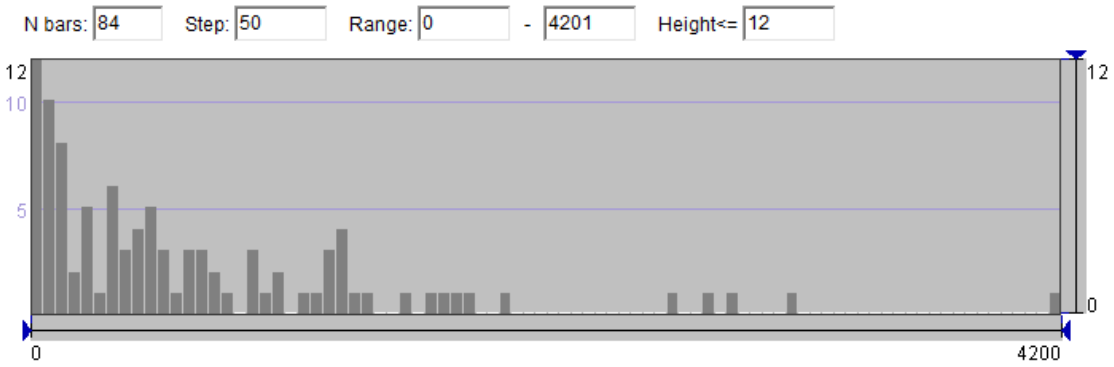


Compare also the durations of the clusters resulting from the first and the second runs.

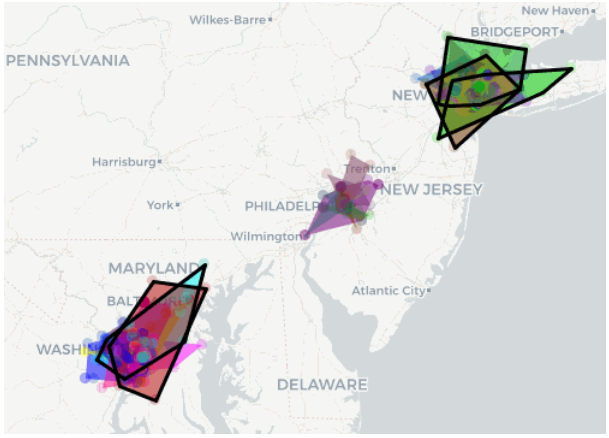


Modifying the spatial distance threshold

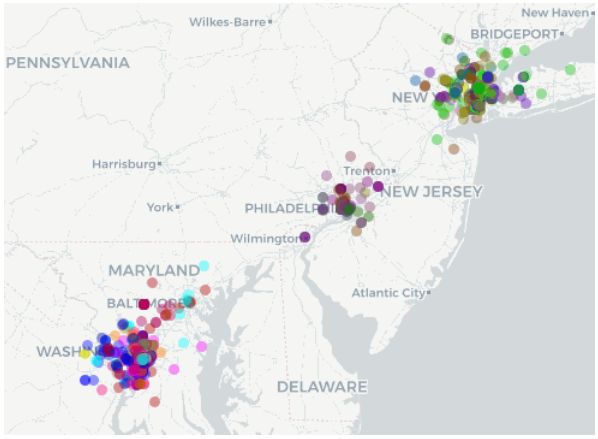
Examine the spatial extents of the clusters, e.g., by considering the areas of the convex hulls of the clusters. You can look at frequency histograms of cluster areas:



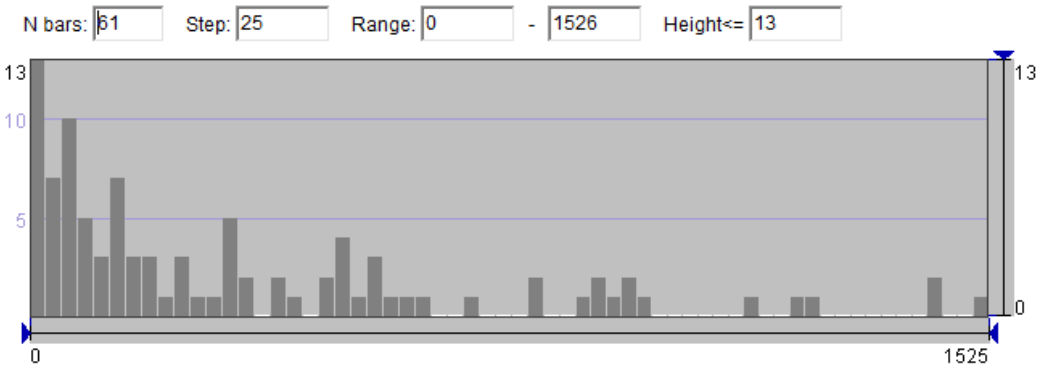
In the map below, the convex hulls of the 5 clusters with the largest areas are highlighted:

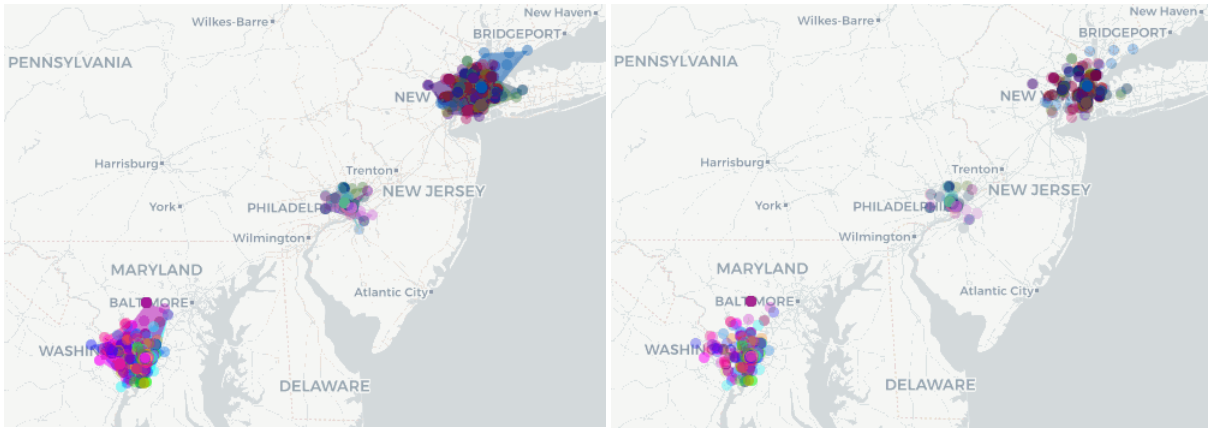


If the convex hulls and their areas are hard to obtain, zoom in the map to the areas containing clusters (the noise should be filtered out) and try to estimate visually how the clusters are spread in space, e.g., as in the map below:



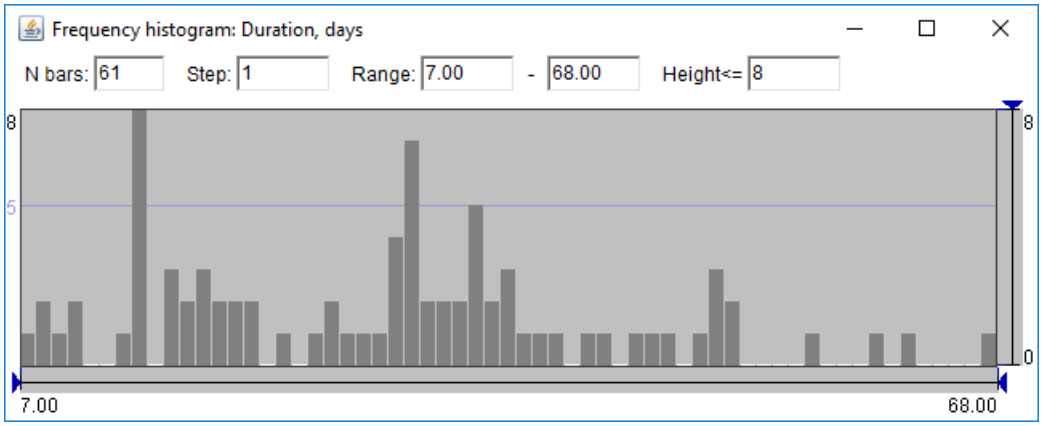
In our example, some clusters are quite too much extended in space, covering sometimes two or more big cities. It makes sense to decrease the spatial distance threshold, e.g., to 20 km. So, perform one more run of the density-based clustering with the smaller distance threshold and compare the results to those from the previous run. Again, consider the number and cardinalities of the clusters, amounts of noise, durations of the clusters, and their areas. Look at the shapes and spatial extents of the clusters on the map. Do you think that the results improved compared to the previous run?

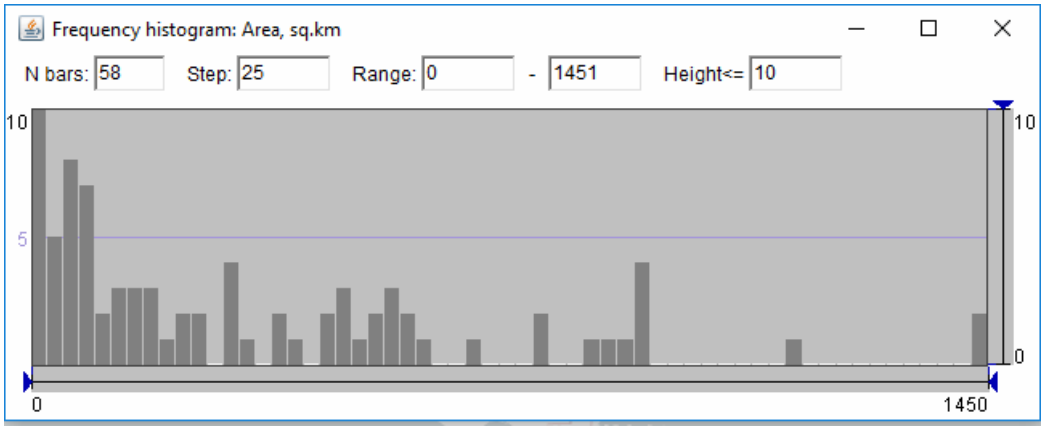




Modifying the density threshold

Perhaps, you have observed that some clusters look quite loose, and you may wish to obtain denser clusters, which better comply with the idea of mass engagement in taking photos. To obtain denser clusters, you need to increase the density threshold, which has been 3 so far. Try another run of clustering after increasing this threshold to, say, 10. Compare the new result to the previous one, as you did before.





Most probably, the number of clusters has decreased after increasing the density threshold. To examine the losses, you can select the elementary events that were in some clusters after the previous clustering run but were put in noise by the last run. In the map, you may find places where there are no clusters any more. It may also be useful to compute the counts of the members in the previous clusters remaining after filtering out the noise resulting from the last run. An example is shown in the table below.

identifier	N cluster members	N cluster members after filtering	Ratio (%)
42	22	10	45.45
57	21	13	61.90
83	16	12	75.00
67	16	0	0.00
49	16	10	62.50
44	16	0	0.00
80	15	0	0.00
47	15	0	0.00
90	14	0	0.00
50	14	0	0.00
87	13	0	0.00
54	13	0	0.00
96	12	0	0.00
95	12	0	0.00
93	11	0	0.00
94	10	0	0.00
85	10	0	0.00
84	10	0	0.00
60	10	0	0.00

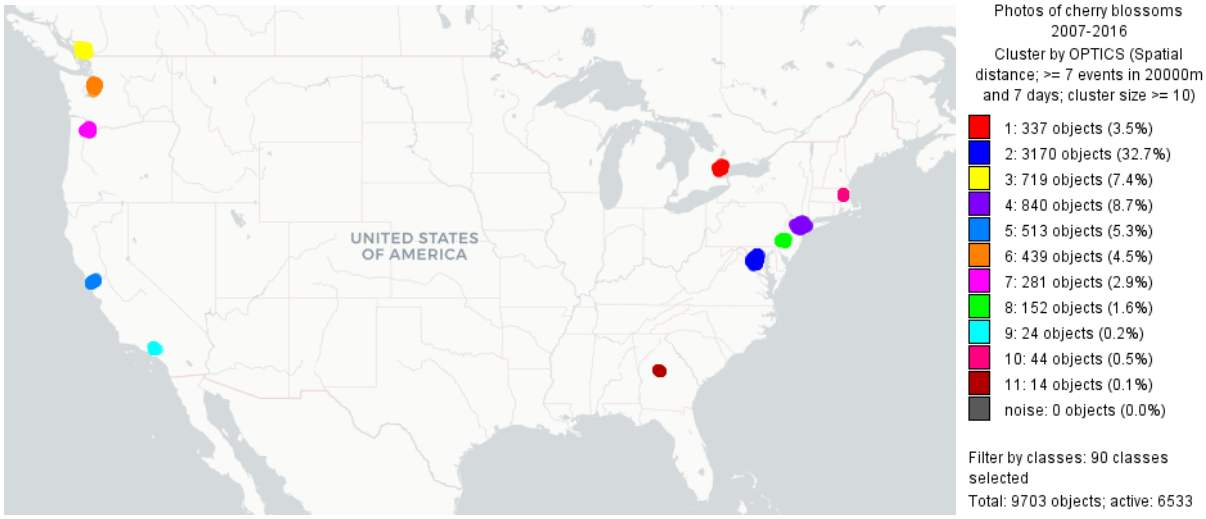
If you find that the losses are a bit too high and should rather be reduced, you can try out an intermediate value for the density threshold between 3 and 10, e.g., 7, and examine the effects.

identifier	N cluster members	N cluster members after filtering	Ratio (%)
44	16	10	62.50
50	14	10	71.43
96	12	0	0.00
95	12	0	0.00
93	11	0	0.00
94	10	0	0.00
85	10	0	0.00
84	10	0	0.00

After all clustering experiments, you need to make a final decision as to which result is the best in regard to your concept of a mass photo taking event.

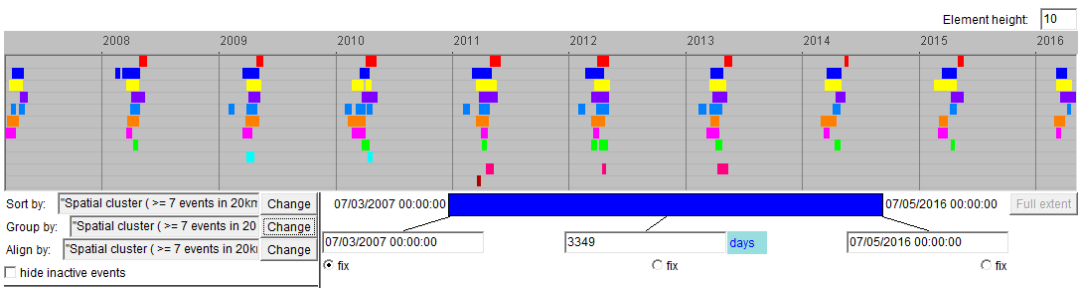
Investigating temporal patterns of events by areas

First of all, you need to define the areas that will be further considered. For this purpose, you need to put together the spatio-temporal clusters that overlap in space, i.e., that happened more or less on the same territory. This can be achieved by applying density-based clustering taking the spatial distance as the distance function, i.e., ignoring the time differences between the elementary events. Before performing the clustering by spatial distances, you choose the best result from all your runs of spatio-temporal clustering and filter out the noise. This means that the spatial clustering will be applied only to the members of the spatio-temporal clusters, and the noise will be ignored. You can use the same spatial distance threshold and the same density threshold as was used for obtaining the chosen best result of the spatio-temporal clustering. As a result of the spatial clustering, you are expected to receive a relatively small number of *spatial clusters* corresponding to the areas in which the spatio-temporal clusters were found. You aren't expected to get any noise.

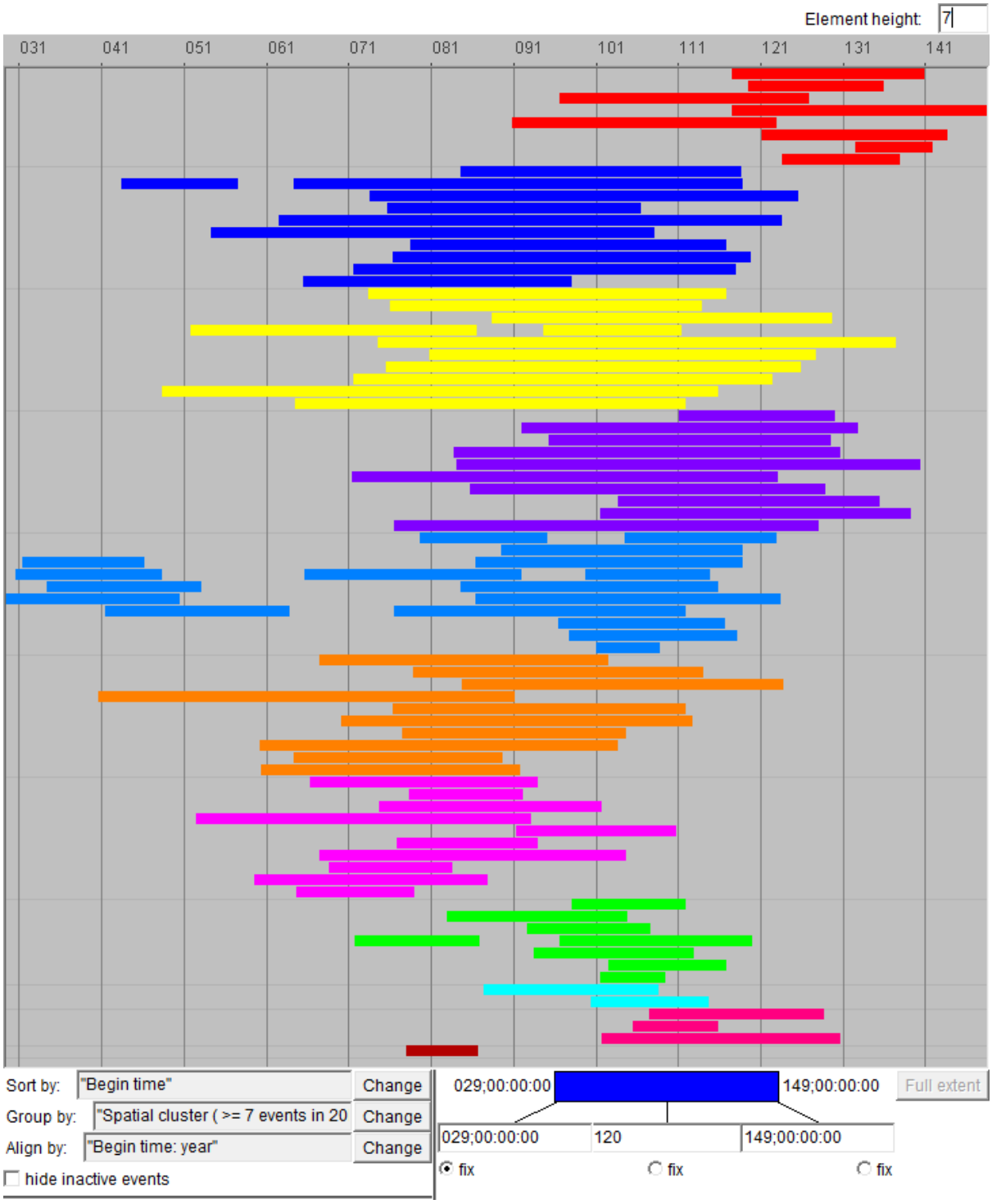


After obtaining the spatial clusters, which define the places where the mass photo taking events happened, you need to create a summary table describing the spatio-temporal clusters in which there must be columns specifying for each cluster its (1) start date, (2) end date, and (3) the label of the corresponding spatial cluster. If a summary table has been produced earlier, a column with the labels of the spatial clusters needs to be added to it. The values in this column are obtained by aggregation over the members of the spatio-temporal clusters.

Having the start and end date and the spatial cluster label for each spatio-temporal cluster, you can visualize these data on a temporal display, such as a timeline bar chart (a.k.a. Gantt chart), to compare the times and durations of the mass photo taking events in the different places. A display like the following one allows you to compare the event times in different places in each year. The bars represent the spatio-temporal clusters; their horizontal positions and lengths correspond to the times and durations of the clusters. The colours (which are the same as in the map above) differentiate the places, i.e., the spatial clusters. You can see in which places the events happened earlier and in which they happened later. You can also see whether the events in a place happened every year and note gaps, i.e., years without events.



In the following display, the bars representing the clusters are positioned horizontally according to the ordinal numbers of the days in the years (the days in a year are counted starting from January 1). The bars are grouped by the places; in each group, the bars are ordered chronologically; each row corresponds to one year. This display allows you to compare the start and end times of the events that happened in the same places in different years.



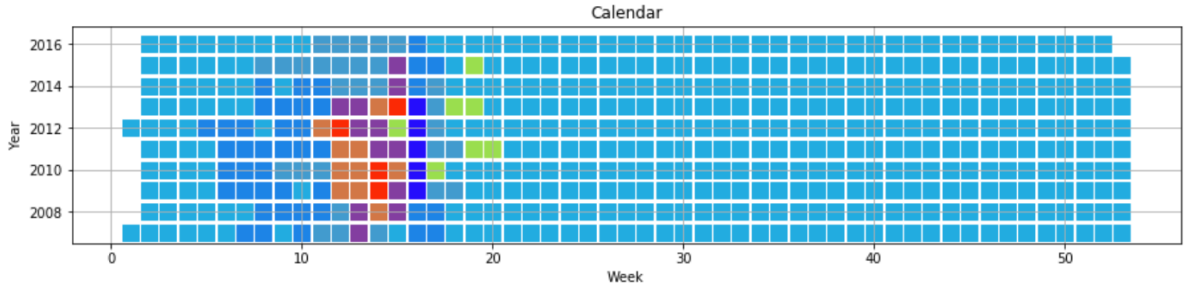
Questions

- In what places (cities) did mass photo taking events happen regularly (each year or almost each year)?
- In what placed did the events usually last longer than in the others?

- Were the relative timings of the events similar in different years?
- Are there places in which the mass photo taking events usually started earlier or later than in the others?
- Can you identify places and years where and when the events started unusually early or unusually late?
- Is there a general temporal trend of the events starting earlier or later from year to year? Are there local temporal trends in some places?

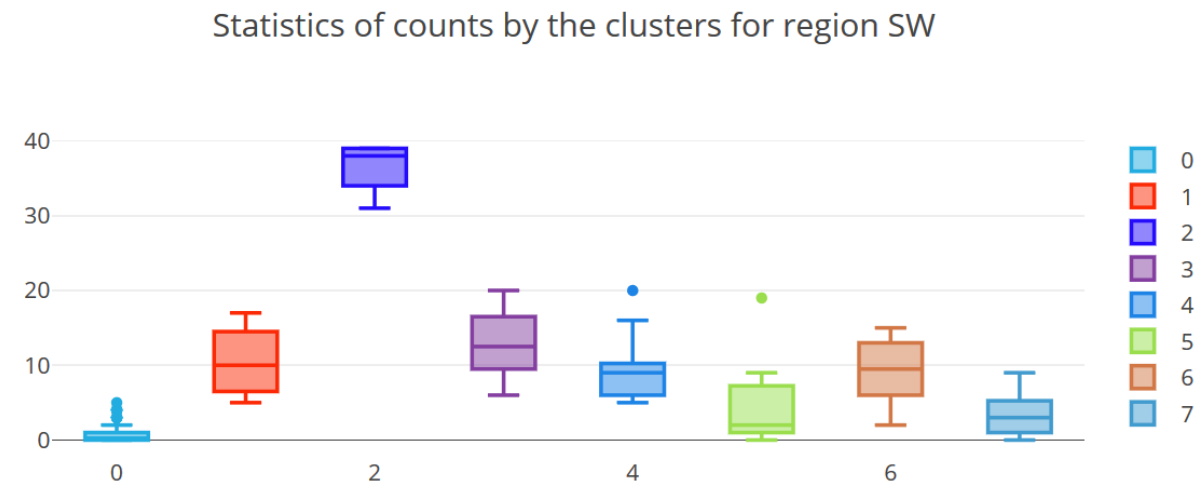
A possible analysis scenario involving partition-based clustering

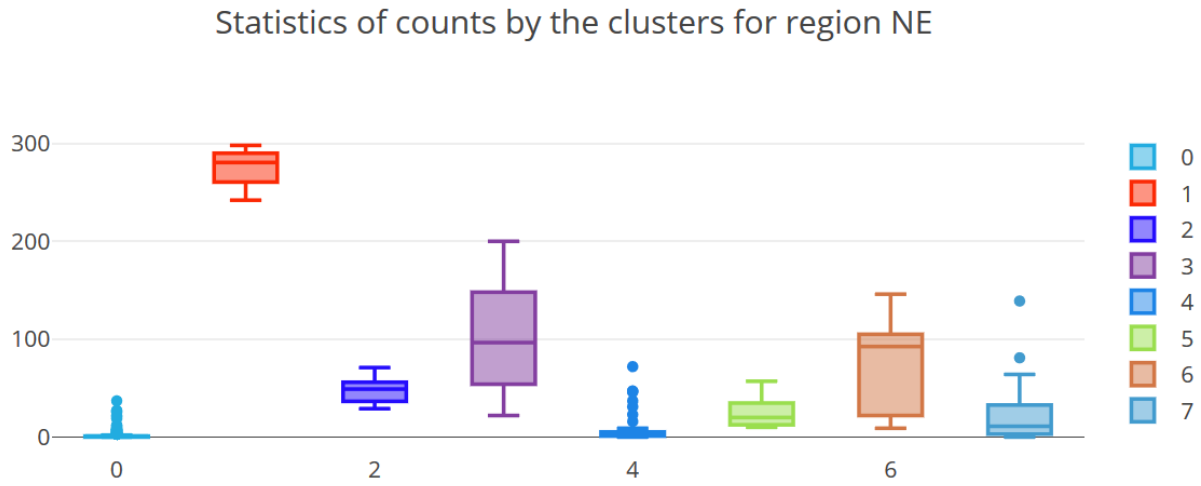
You apply partition-based clustering to the weekly time steps characterised by the counts of the photos taken in the different regions. In the result, each time step becomes a member of some cluster. You use this result to investigate how the distribution of the photo taking activities across the regions changed over time. For this purpose, you need a temporal display of the cluster membership. Since the photo taking activities are expected to vary seasonally (in accord with the seasonal character of the phenomenon of cherry blossoming), it is reasonable to create a temporal display where the yearly time cycle is represented. One possibility is a 2D plot where one dimension corresponds to the cycle of the weeks in a year and the other dimension to the sequence of the years, for example, like the following:

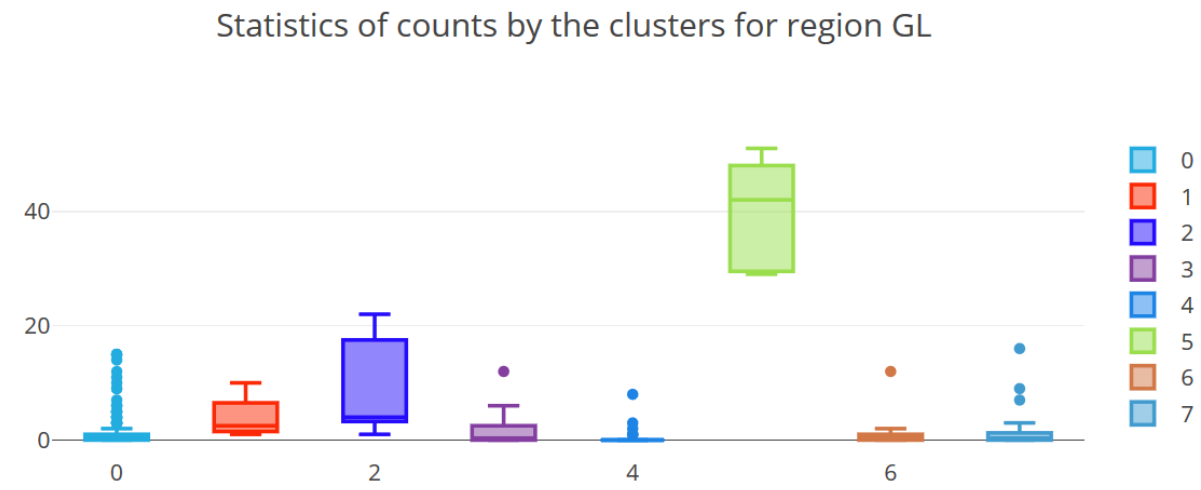


This display shows us that there is a single cluster (represented by the light blue colour in this example) containing most of the time steps, in particular, all weeks before week 4 and after week 20 of each year. There are also weeks, mostly between weeks 5 and 10, which obtained a slightly darker blue colour. The similarity of the colours indicates that the respective clusters are similar in terms of the per-region photo counts. The members of the remaining clusters occur from week 11 to week 20. Evidently, these are the times of cherry blossoming and taking photos of this phenomenon.

While the distribution of the colours across the display indicates periodic (i.e., seasonal) variation, there are differences between the years. To understand the differences, it is necessary to interpret the cluster colours in terms of the distribution of the photo taking activities across the regions. This can be achieved by visualising summary statistics for the clusters, for example, using box plots, as shown below:

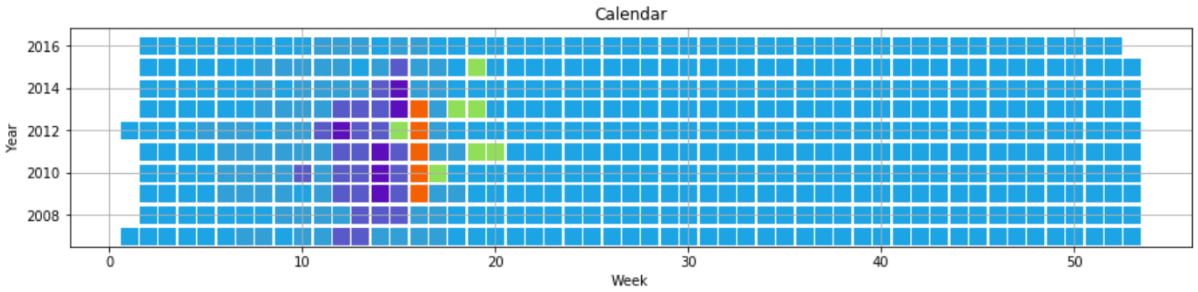




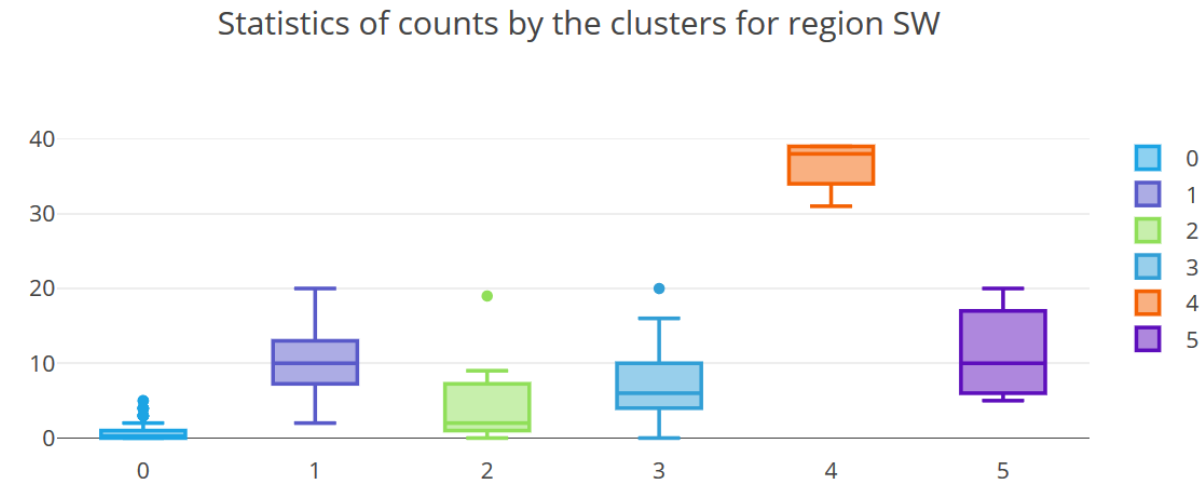


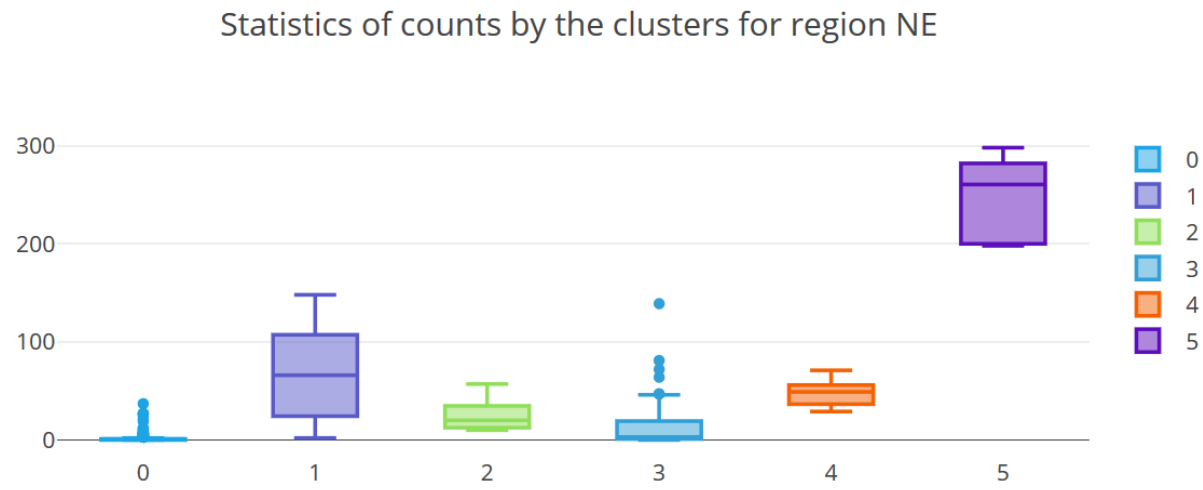
… and so on for the other regions.

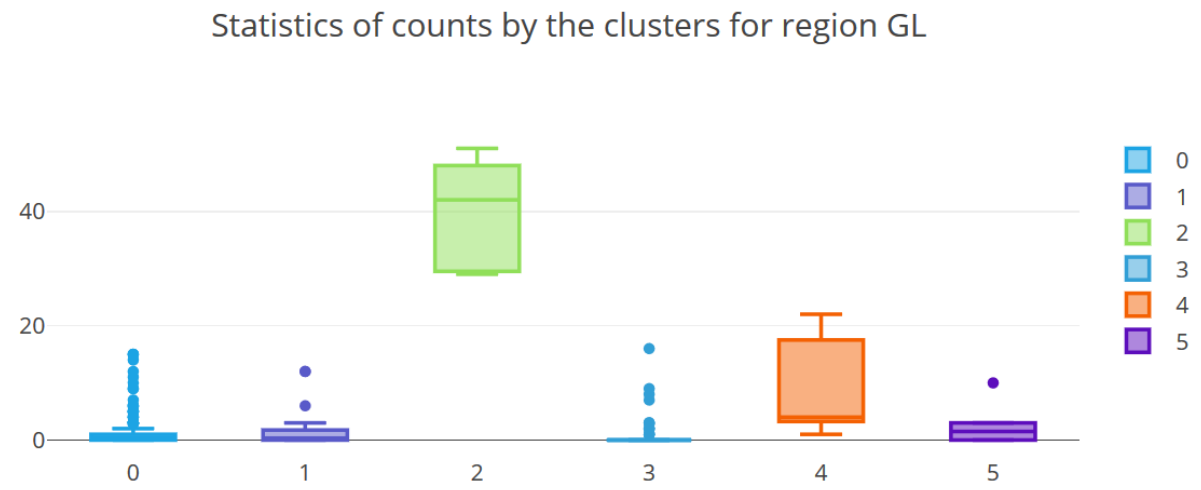
It seems that we, perhaps, have an excessive number of clusters. It is reasonable to look whether we can obtain clear temporal patterns and easily understandable cluster profiles with partitioning into fewer clusters. E.g., below is the temporal distribution for 6 clusters:



The patterns of the colour variation are consistent with those we had previously. However, the meanings of particular colours occurring in both displays may differ. It is necessary to interpret the colours again with the use of summary statistics displays.







By combining the information obtained from the calendar display and the summary statistics display, we can compare the times of the highest photo taking activities across the region s and across the years.

Draft Python scripts

We provide you with two separate draft Python scripts (notebooks) for the activities with the density-based and partition-based clustering.

The script focusing on the density-based clustering includes the following operations:

1. Loading of the photo data using a 3 years subset from year 2012 to 2014 (a larger data set requires much longer time for representing on a map and for clustering).
2. Representation of the data on a map and in a space-time cube.
3. Density-based clustering by the spatio-temporal distances. You are supposed to experiment with the clustering parameters.
4. Computing statistical summaries of the spatio-temporal clusters and obtaining colours for the clusters by means of spatialisation.
5. Showing the cluster membership of the events on a map and in a space-time cube.
6. Density-based clustering by the spatial distances, obtaining colours for the clusters, and visual representation of the results.
7. Summarization of the spatial clusters, in particular, summarization by the combination cluster + year for examining the year-to-year differences between and within the clusters.

The script focusing on the partition-based clustering includes the following actions:

1. Loading of the aggregated data, i.e., weekly counts by regions for the years from 2007 to 2016.
2. Application of a partition-based clustering algorithm (K-Means). You are supposed to experiment with the parameter K (number of clusters).
3. Generation of colours for the clusters by applying spatialisation to the cluster centroids.
4. Visualisation of the distribution of the cluster membership along and across the years.
5. Visualisation of statistical summaries of the clusters in terms of the per-region photo counts.

We suggest you to note your findings as comments in the notebook you are using and to share your notebooks with the changes and notes you have made in the moodle forum.

We wish you a successful and fruitful fulfilment of the exercise.