International Cartographic Association
Association Cartographique Internationale

Significant locations in auxiliary data as seeds for typical use cases of point clustering

Johannes Kröger[a]

[a] *HafenCity Universität Hamburg, Lab for Geoinformatics and Geovisualization, Hamburg, Germany; johannes.kroeger@hcu-hamburg.de*

Abstract: Random greedy clustering and grid-based clustering are highly susceptible by their initial parameters. When used for point data clustering in maps they often change the apparent distribution of the underlying data. We propose a pro-cess that uses precomputed weighted seed points for the initialization of clusters, for example from local maxima in population density data. Exemplary results from the clustering of a dataset of petrol stations are presented.

Keywords: web mapping, cluster initialization, cluster preprocessing, significant locations, population density

1. Introduction

Clustering methods are often used on point data in interactive maps to avoid data crowding, overlapping of symbols or to minimize load on the client's computer. Here, instead of displaying a symbol for each point, groups of them are aggregated into clusters and their display is limited to one symbol for the whole group. Popular web map-ping libraries and services offer point clustering as a base feature, easily enabled by users without requiring any additional knowledge.

Due to their low computational cost and general applicability usually random greedy (e.g. Leaflet 2017) or grid-based clustering (e.g. Google 2017) algorithms are used. In random greedy clustering the clusters are initialized by randomly chosen points and aggregated by a fixed radius until all points are assigned to a cluster. In grid-based clustering a polygonal grid, usually made out of squares, is placed over the area of interest and points are aggregated per cell.

We argue that the resulting cluster distributions of these standard algorithms often mislead, at least in maps where the spatial distribution of the points follows an underlying pattern that the user knows or infers from the map background like population density. How the random points are chosen or how the grid is based and shaped deter-mines the outcome and can highly skew the representation. A typical store locator map for example, commonly displayed on company websites, should allow potential customers to determine if any store is available at a certain location. Here such clustering can drastically change the validity of the map: A group of stores, located naturally in a city, might be torn apart into separate clusters, resulting in the city not being visibly covered by any cluster. The city might sit just between two randomly chosen cluster centers of a greedy clustering approach or right on the border between cells in a grid-based clustering. If this happens, the map failed its purpose. Figure 1 shows the map that initially motivated this research as an example.

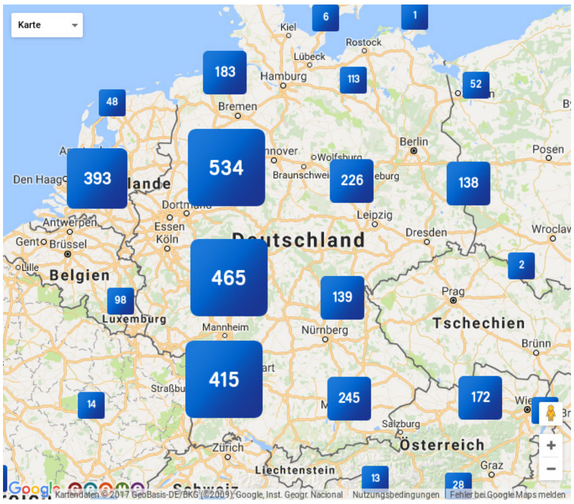

Figure 1: Map showing petrol stations of Aral in and around Germany using Google's (Google 2017) grid-based clustering (Aral Aktiengesellschaft 2017). Places with an high actual density of petrol stations like Hamburg or Berlin are not visible as such due to the nature of the grid.

We propose a process using seed points derived from significant locations in auxiliary data to initialize the clustering of datasets of a similar distribution as the auxiliary data. Following this process the cartographer gets fast, deterministic and practical clustering, viable for web mapping. Our clustering process is meant to be straightforward to implement and fast to perform as to ease implementation and compatibility.

2. Related Works

In addition to the aforementioned random greedy and grid-based clustering many other clustering algorithms exist for general purpose applications.

Density based clustering algorithms like DBSCAN (Ester et al. 1996), OPTICS (Ankerst et al. 1999) and their many variants might very well be applicable to the general problem and lead to good results but usually require fixed parameters or are computationally expensive.

Approaches based on k-means use a fixed number of desired clusters as start parameter which requires prior

Proceedings of the International Cartographic Association, 1, 2017. This contribution underwent single-blind peer review based on submitted abstracts | https://doi.org/10.5194/ica-proc-1-63-2017 | © Authors 2017. CC BY 4.0 License.

knowledge about the clustering of the data (which we would not have in that regard) or an iterative approach (which might get computationally expensive). There are countless approaches to improve the initialization of k-means clusters, for example: Bradley and Fayyad (1998) use estimation of the data distribution's modes to compute appropriate starting conditions for iterative clustering. Khan and Ahmad (2004) use "prominent attributes" to define deterministic cluster centers by clustering multiple times. Basu et al. (2002) initialize clusters from seed data with and without constraining the subsequent clustering to keep this initial data. A detailed discussion and evalua-tion of earlier approaches is found in Meilă and Heckerman (1998). Apart from initialization, also the method of determining cluster assignments has been the topic of many publications. Wagstaff et al. (2001) for example use domain-specific constraints on contiguity and spatial separation to assist clustering-based lane finding in GPS tracks.

Grid-based clustering is also a very active research topic with many approaches that could make its results more appealing in a cartographic context. Akodjènou et al. (2007) for example use a locally adaptive grid with randomly oriented borders to minimize the influence of the grid geometry to the clustering result. Bereuter and Weibel's (2010, 2011) research on point data generalization includes assessment and suggestions for the selection of suitable algorithms. They use a quadtree index to support e.g. aggregation of point data in real-time on mobile devices (Bereuter and Weibel 2013).

As Jain (2009) notes "thousands of clustering algorithms have been proposed in the literature in many different scientific disciplines", so what we are proposing might very well have been already suggested in a more abstract way and implemented before. But as it does not appear to be in practical cartographic use so far, we consider it a worthy contribution in any case.

3. Method

We propose a trivial and fast, semi-supervised procedure that improves upon both grid-based and random greedy clustering in terms of visual appeal and, if applied appropriately, spatial correctness. By preprocessing a common dataset like population density and finding its locally significant maxima, this data can be used as seeds for initialization of clusters in any dataset known to be correlated or very similarly distributed as the original dataset.

Our approach is not a general purpose clustering algorithm. It is meant for geographic data points of a single kind, as only their spatial component is considered when clustering. The user needs to have prior knowledge about the kind of their data, i.e. in this case if it is closely related to population and thus distributed similarly. In the following, population density data is used but the concept itself is applicable for other spatial phenomena where previous knowledge exists.

In our approach the cluster seeding points are located on the local population maxima. These local maxima can be calculated on varying scales using grid-based population maps or derived from datasets of populated places like settlements or metropolitan areas. Each seed's weight is used to calculate the extents of its catchment area. Neighboring points of the dataset-to-be-clustered are then aggregated into clusters per catchment area. Each seed's weight and its influence on the catchment area could be specified dependent on the scale or other cartographic measures, allowing a granular and dynamic control over suitable locations.

The main innovation of this approach is in using precomputed, weighted seed points instead of an iterative, random selection for distance-based clustering. The process is structured in three separate steps: (1) The generation of weighted seed points, (2) the clustering of point data utilizing the seed points and (3) visualization of the resulting clusters in an appropriate way. In the future these steps could, and probably should, be treated more closely coupled.

3.1 Generation of weighted seed points

The basis for our approach is weighted seed point data. To be able to seed the clustering process on densely populated regions, the seed points need to be placed and weighted accordingly. As this step is done offline and just once, any kind of calculations, even computationally intense, could be used to acquire a suitable end result. There-fore the clustering algorithms initially discarded by us might very well be used in this step. The educated, intentional choice of an appropriate method is crucial.

A most ideal approach to generate their seed point locations and weights would utilize detailed population distribution data, for example via a fully automated, raster-based local/regional maxima search, using the values in the neighborhood as weight. Another option might be using the spread of populated places by area rather than the ac-cumulated population within. Centroid point data and number of inhabitants of cities and other populated places could also serve as an appropriate basis. Of course a diligent cartographer could manually place and weight points by manual means as well. Some manual oversight is advisable in any case.

The result of this step are points centered on densely populated regions with an appropriate weighting value rep-resenting their importance or "influence". Their locations are ready to be used as seed points for cluster initialization and their values available to serve as weights in the determination of neighborhood relationships of the data points.

3.2 Clustering data based on seed points

The weighted seed points can now be used to cluster other point data accordingly. The choice of method for this is up to the user. The general idea of using precomputed seed points and weights is applicable for a variety of clustering algorithms.

We argue that the simple assignment of data points to clusters by their minimum weighted distance to seed points is "good enough" and very well appropriate for the intended use case:

```
for point in points:
        for seed, weight in seeds:
              calculate spatial distance between point and seed
              weight distance by weight
        assign point to cluster around seed with smallest weighted distance
count points per cluster [optional]
```

The result are the assignments of the data points to their clusters or, if that would be enough for the intended use case, the count of assigned data points per cluster.

Due to the nested nature of the algorithm its complexity is O(n*m), where n is the number of data points and m is the number of seed points. As the number of seed points should rarely exceed the lower double digits, this is acceptable even for a reasonably high number of data points. Spatial indexing or an hierarchical data structure could be utilized to more quickly partition the data points if necessary.

3.3 Symbology

The last step in the process is to visualize the generated clusters on the map. Again, the user is free to choose an appropriate method. We suggest the use of proportionally scaled circles as they represent the method from the previous step. The locations of the clusters are determined by their seeds' positions. Their spread results from the number of data points they represent. Generally, aesthetics, legibility and user acceptance are the most important aspects here. Reasonable amounts of overlap and clutter might be justifiable depending on the map's purpose.

4. Experimental Results

As a proof of concept we now describe the details of our implementation for one selected dataset at its full ex-tent and at a specific scale. A dataset of ~14,000 petrol stations in Germany at a scale of approximately 1:4 million (which corresponds closely to the zoom level 7 in a standard web map) was clustered. At this scale the whole of Germany fits vertically on a common computer screen. Figure 2 shows how closely the distribution of the stations resembles that of the population.

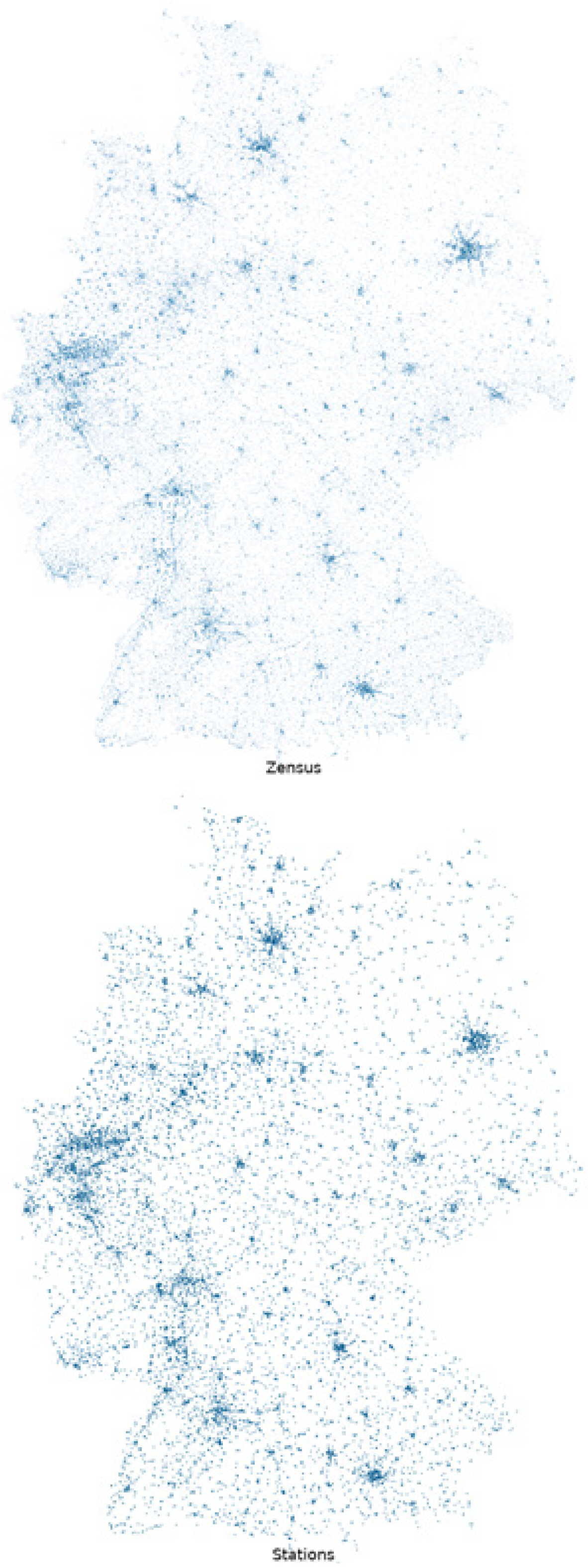

Figure 2: Left: Population density in Germany (data by Statistisches Bundesamt, 2015), Right: Petrol Stations in Germany (data by 1-2-3 Tanken, 2016)

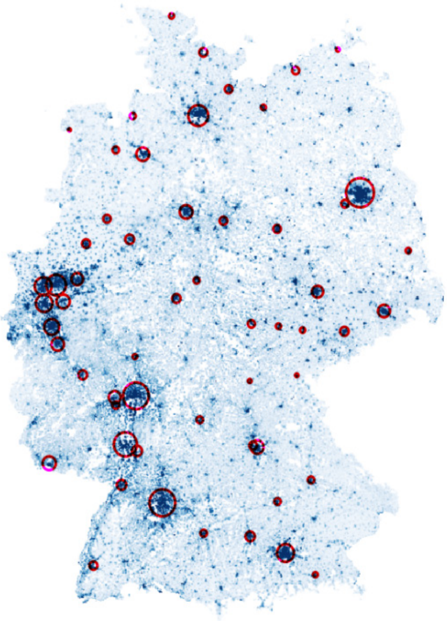

Figure 3: Population density according to census with a darker shade of blue indicating higher density in the background, populated places from the Natural Earth dataset as red circles proportionally scaled by the population value in the foreground

4.1 Natural Earth's Populated Places as proxy for population distribution

As proxy for more sophisticated population data the Natural Earth's Populated Places dataset in the 1:10 million scale version[1] was used. This dataset consists of point features of populated places around the world with a variety of attributes including population data. For example there is a point at the location of the city of Hamburg, Germany with a pop_max population value of 1,757,000.

Natural Earth's authors say they "favor regional significance over population census in determining our selection of places" (Natural Earth 2017) which on the one hand contradicts our initial plan to use pure population data but on the other hand is a welcome curation that improves the quality of the data. The general distribution of populated places in the dataset naturally corresponds to the distribution of population density (compare Figure 3).

As the stations dataset covers only Germany the places dataset was filtered accordingly. The dataset was further enhanced for our use-case in a supervised process by combining nearby points and aggregating their population values. The approach was as follows:

- Sort the populated places dataset in descending order by the population value
- Starting with the now highest ranked place, create a buffer polygon with a radius determined by its population value divided by a factor
- Recursively aggregate the population values of all populated places inside the buffer and update the buffer radius accordingly
- Once no more not-yet-aggregated populated places lie inside the buffer, take the next highest, not yet aggregated populated places and repeat the same process

We explored the results of different factors in this aggregation process and ended up using the factor 80 as an appealing compromise between visual aesthetic and appropriate aggregation (see Figure 4). This is a "magic" value and only acceptable as such in a prototyping stage. The data was then further reduced by removing places with a population of less than 500,000 people. The result is a dataset of 13 places resp. seed points and their aggregated population values.

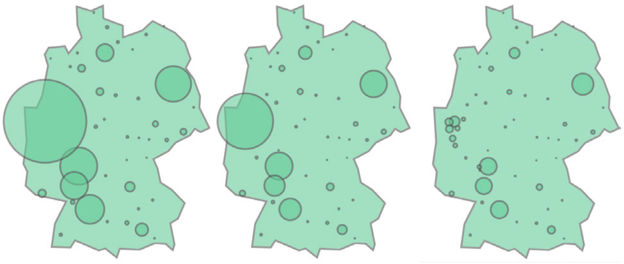

Figure 4: Aggregating populated places with the buffer radius being a fraction of the population and the factor 60, 80 (as used) and 100 from left to right.

4.2 Assigning stations to clusters by weighted distance

The seed points' population values were normalized to a range of 0 to 1. For each petrol station the distance to each seed point was calculated and weighted by the population of the seed point. An inverse, squared weighting of the distance was used:

$$distance_{weighted} = distance - \frac{population}{distance^2}$$

Each petrol station was then assigned to the closest seed point according to the normalized, weighted distance (see Figure 5).

4.3 Styling the clusters

For styling we used proportionally scaled circles with sizes adjusted after Flannery's perceptual model. The choices for minimum and maximum radius were made subjectively, aiming for an aesthetically pleasing outcome with appropriate overlap. Figure 4 shows the resulting circles on a plain map. As intended, the resulting image represents the expected distribution quite well as the circles are centered on the densely populated areas.

[1] http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-populated-places/

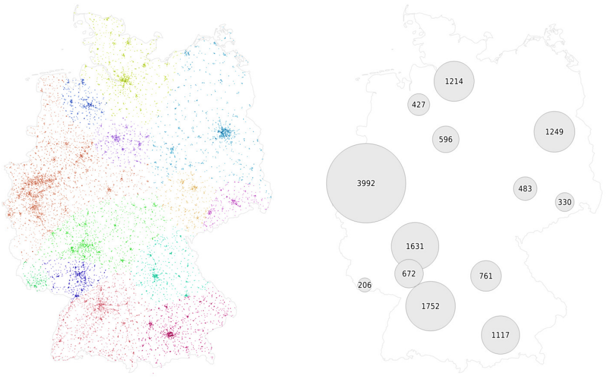

Figure 5: Left: Stations assigned to clusters, the colors denote the cluster, Right: Proportionally scaled circles on the seed points

5. Conclusions & Outlook

We demonstrated that our approach can lead to an intuitively informative map, if the dataset in which the seed points are created closely matches the distribution of the clustered dataset. The resulting clusters show a reasonable representation of the petrol stations' distribution.

While the resulting clusters appear appealing to us, the process and its results need to be rigorously evaluated. Both results and computational cost need to be compared against established clustering algorithms. The results need to be empirically tested and verified on user acceptance and legibility against the contesting algorithms.

As the seed points are fixed anchors, clusters for which only the aggregated values have to be determined, our approach will find clusters no matter the actual distribution of the data points. This makes it subject to user error if an inappropriate dataset is used. This is a major concern and requires suitable measures on the quality of the result-ing clusters.

Our approach finds how many data points can reasonably be clustered at the locations proposed by the seed points rather than partition the data points into intrinsic cluster structures. Thus the evaluation of the results' quality can not necessarily follow common clustering evaluation procedures. The cumulative distance of all data points to their assigned clusters might be a good, simple indicator of overall clustering quality.

We plan to further research and evaluate different approaches for each of the process' steps so that appropriate methods can be recommended. A tight coupling of all three steps could allow for a vertical integration of methods and parameters.

We hope to process population data to a global, hierarchical dataset that offers appropriate seed points for any scale.

6. References

Akodjènou-Jeannin, M.-I., Salamatian, K., Gallinari, P. (2007). Flexible Grid-Based Clustering. PKDD 2007 Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases, 350–357.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. ACM Sigmod Record, 49–60. https://doi.org/10.1145/304182.304187

Aral Aktiengesellschaft (2017). Tankstellenfinder und Routenplaner. http://www.aral.de/de/retail/online-services/tankstellenfinder-und-routenplaner.html. Accessed 25 Feb 2017.

Basu, S., Banerjee, A., & Mooney, R. (2002). Semi-supervised Clustering by Seeding. Proceedings of the 19th International Confer-ence on Machine Learning (ICML-2002), (July), 19–26.

Bereuter, P., & Weibel, R. (2010). Generalisation of point data for mobile devices: A problem-oriented approach. 13th Workshop of the ICA commission on Generalisation and Multiple Representation, Zürich, Switzerland, 12 September 2010 - 13 September 2010, 1-8. https://doi.org/10.5167/uzh-39623

Bereuter, P., & Weibel, R. (2011). A diagnostic toolbox for assessing point data generalisation algorithms. 25th International Carto-graphic Conference, Paris, 3-8 July 2011, 12.

Bereuter, P., & Weibel, R. (2013). Real-time generalization of point data in mobile and web mapping using quadtrees. Cartography and Geographic Information Science, 40(4), 271–281. https://doi.org/10.1080/15230406.2013.779779

Bradley, P. S., & Bradley, P. S. (1998). Refining Initial Points for K-Means Clustering. Microsoft Research, 91–99. https://doi.org/10.1.1.44.5872

Google (2017). Marker Clustering. https://developers.google.com/maps/documentation/javascript/marker-clustering. Accessed 25 Feb 2017.

Leaflet (2017). Leaflet.markercluster. https://github.com/Leaflet/Leaflet.markercluster. Accessed 25 Feb 2017.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. 19th International Conference in Pattern Recognition (ICPR), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Khan, S. S., & Ahmad, A. (2013). Cluster center initialization algorithm for K-modes clustering. Expert Systems with Applications, 40(18), 7444–7456. https://doi.org/10.1016/j.eswa.2013.07.002

Meilă, M., & Heckerman, D. (1998). An Experimental Comparison of Several Clustering and Initialization Methods. Proceedings of the Fourteenth Conference on

Uncertainty in Artificial Intelligence (UAI1998). https://arxiv.org/abs/1301.7401

Natural Earth (2017). Populated Places. http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-populated-places/. Accessed 25 Feb 2017.

Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S. (2001). Constrained K-means clustering with background knowledge. ICML'01: pro-ceedings of 18th international conference on machine learning, 577–584.