A Comparison of Spatial Aggregation Methods for the Derivation of Proportional Symbol Maps

Mathias Gröbe
TU Dresden
Institute of Cartography
01062 Dresden, Germany
mathias.groebe@tu-dresden.de

Dirk Burghardt
TU Dresden
Institute of Cartography
01062 Dresden, Germany
dirk.burghardt@tu-dresden.de

Abstract

With the rise of Volunteered Geographic Information, a huge amount of point data sets is available for a wide range of application and research questions. As consequence results the requirement of new analysis and visualisation tools to utilisable the contained information. One often used visualisation method is the rendering of all points as dot map. This can show a detailed distribution of the points, but not the number of records. Therefor is an aggregation necessary to map the number of values in one cluster to a visual variable like size. With an example we demonstrate the possibilities of data aggregation for visual analysis and the effects of different aggregation methods on the resulting patterns. As visualisation constraint the area that is used for the visualisation of the values should be always the same. Finally, we give some advices for the selection of a suitable data aggregation method and the production of meaningful maps.

Keywords: Cartography, Geovisualisation, Visual Analysis, Clustering, Generalisation

1 Motivation

Nowadays a huge amount of data is available with a great application potential that can be made accessible due new methods. The rise of Volunteered Geographic Information (Goodchild, 2007) have unlocked many new data sources. There is a lot of point data available, which were created by people and sensors. In this context the point map is an often-used method for visualisation. Depending on the investigation area with a high data density might occur over-plotting of points. Thus, simple point maps can show very detailed pattern, but also hide information. There is no possibility so get information about the quantities. Spatial aggregation can solve this problem, by counting the points in one aggregation group and mappings the number to a visual variable like size. The result is called proportional symbol map. We tried four different approaches and want to compare the visual results.

2 Categorisation of Aggregation Methods and Visualisation Constraint

The section subdivides aggregation methods into clustering, utilsation of spatial data struces and usage of predefined units. Because we want to compare the variable possibilities by their results, we have determined that in every map should covered the same area by the aggregated values. The last section provides the for the realisation necessary technique for the implementation of this constraint.

2.1 Clustering

Spatial clustering is the process of grouping objects into classes, which are usually called “cluster”. Often the distance is an important measure in the cluster generation. We decided to use a density-bases method, which regards cluster as regions of a high number of objects. Based on regions with a high density the DBSCAN-algorithm (Martin Ester et al., 1996) generate cluster after a defined distance and reject points which are too far away from the cluster.

2.2 Aggregation through Spatial Data Structures

To enable quick access on geospatial data within databases specialised data structures can be applied. There are used to build indices over the geometry to enable faster queries. Another often applied index it the Geohash (Anon, 2017). A string representation is calculated for each coordinate tuple. Removing characters from the end of the string reduce the accuracy. This means points starting with similar numbers were summarised to on point.

2.3 Aggregation by Regular Units

A simple approach is the usage of computed areas for the aggregation. The features that are contained in the area will be summarised. In former days often the grid lines were used as regular reference units (Bollmann, 2001). This way squares, rectangles or trapezes were generated, but it is also possible to construct triangles or hexagons. Their shape often looks more native and less man-made (Arnberger, 1993). This method is suitable for varying applications, because of the adjustable size and shape of the units and often used for statistic issues.

2.4 Aggregation by Administrative or Functional Units

A well-known method is to count values in administrative units like states or city districts. The same is possible with functional areas such as national parks for example. In the then often used choropleth maps should be shown the relative values. For absolute values are proportional symbols suitable.

2.5 Visualisation Constraint

For comparing the resulting maps, it is useful to define a scale for the symbol (legends of proportional symbols). That makes sure that the size of symbol for the aggregated values scale with the number of aggregated values and every feature in all maps get the same area and importance. We have used the formula below to calculate the diameter of each circle. The area of a circle that stands for the number of aggregated points N_i that is scaled on the base of the f_0 area (Töpfer, 1974):

$$d = \sqrt{\frac{4}{\pi} \cdot f_o \cdot N_i}$$

d ... Diameter of the resulting circle
f_0 ... Area for one value
N_i ... Number of values

3 Implementation and Case Study

As example to demonstrate the different aggregation methods in our case study we extracted all points which tagged as amenity in the OpenStreetMap database around Dresden. For the creation of a meaningful map we decided to use the sustenance amenity subgroup only (Anon, 2018). This give us the simple example, that we can map regions in the city their people go out for dinning, clubbing and so on. The data was extracted with the Overpass API and stored into a PostgreSQL 10 database with PostGIS 2.4 as spatial extension. That offered us the possibilities to work in a simple reproducible environment with all necessary tools. As front end and for the map creation we used QGIS 2.18.

A first example for the aggregation methods shows Figure 1 with the clustering of the points. The original points are blue, the clusters green. The green framed areas behind the clusters show the areas which can be constructed from the points features in one cluster. The DBSCAN algorithm was parametrised with a distance of 250 meters and a minimum cluster size of one point. On the one hand, that prevents us from losing some points and their position. On the other hand, there are some very small points and clusters that overlays single blue points.

Figure 2 in comparison is very different from Figure 1 with it regular grid like position of the symbols. This time less of the original points are covered by the clusters. We have used the Geohash with the length of six characters to create map.

In Figure 3 the reference area is very clearly visible with hexagon-structure. The size of the cells is approximately 1x1.2 kilometres. Like in Figure 2 the distribution of aggregated groups is regular. At least example Figure 4 uses the administrative districts of the city as area for the aggregation of points like in the map before and place the symbol in the centroid of the area. The polygons for the districts are from OpenStreetMap and are also visible as reference area. On the first look the distribution is very similar to Figure 1 in the city centre. In other parts of the map it is very different. In some examples the clusters are far away from the original data points.

4 Discussion

The clustering is very customisable with the parameters and the resulting pattern very similar to the given data distribution. The aggregation comes out of the data as well as the parameters and is not driven by any external structure.

The patterns of the spatial data structures and the reference units are present in the resulting maps of Figure 3 and Figure 4. Of course, it is easier to recognise a regular reference area than an irregular. In addition, the regular reference area in Figure 3 and the data structure in Figure 2 splits some clusters. That is not a good solution but shows more the real distribution. In contrast the administrative reference area disturbs the data massive as such as in Figure 4 that some clusters are far away from the data points.

The possibility of caching is not relevant for the less than 1,500 points in our example. There is no problem to compute the clustering on the fly. The precomputation of the Geohash can help in this point. Also, the point in area test can be prepared. But is not so flexible like the adjustable spatial resolution of the Geohash.

An interesting point is how adaptable the methods are to a modified map scale. The administrative and functional units are optimal for a small scale-range. For another scale-range the city districts in Figure 4 may not so suitable. The clustering can be adjusted also like the spatial data structures. It is also possible to construct new regular reference units with a suitable size. But this more complicated, than adjusting only a parameter.

From a cartographic point of view clustering is a very suitable method to aggregated data. It looks very natural. It is probably the best solution in our case study. The often-used administrative units emphasise in our case study the man-made borders more than for the thematic content interesting circumstances. The results of the spatial data structures and regular reference looks very generic. There are easy to compare over a whole map and fit also for this application.

5 Conclusion

We think clustering is the best solution for small data sets, where the computing time is appropriate. It does not disturb the data and is very flexible. The spatial data structures are suitable for time-critical use cases as well as higher number of features and very similar in the results to the regular reference areas. Social media analysis and statistical issues are the common examples of this solutions. The administrative units should be only used if the data depends on the area like elections or is no other area available. If applicable functional units are the better alternative. There can be more adjusted to the application. Otherwise should better applied the clustering, spatial data structure based or regular unit methods to aggregate the data and visualise the results.

Figure 1: Points aggregated with clustering (DBSCAN)

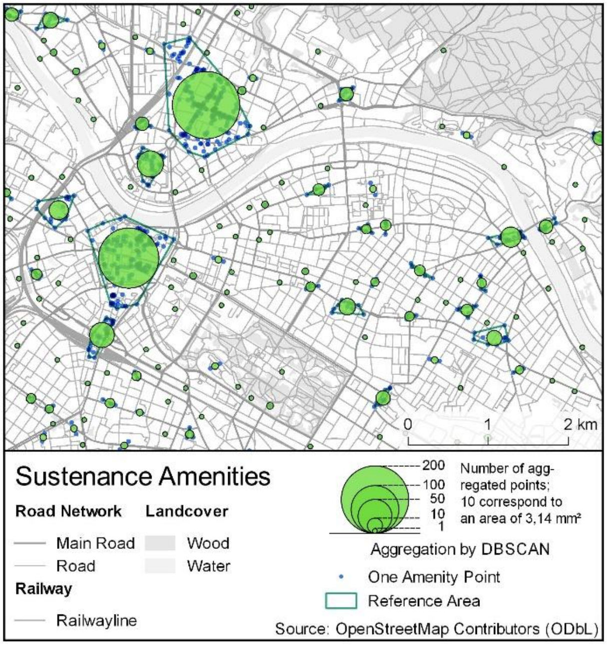


Figure 2: Points aggregated through a spatial data structure

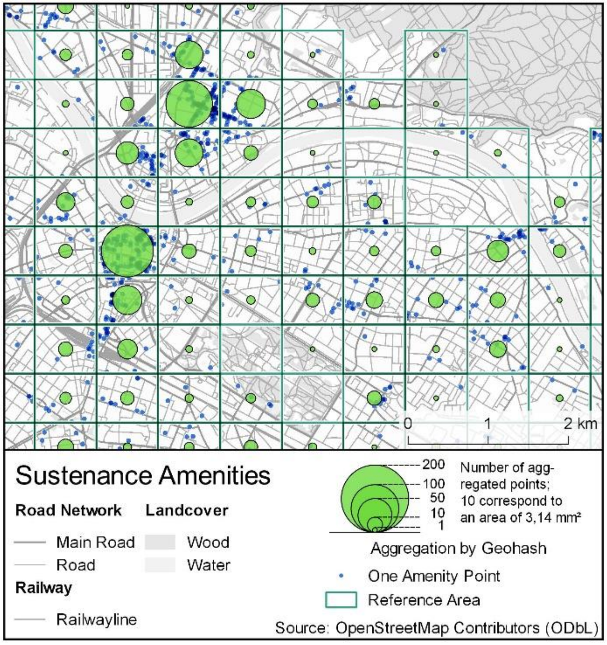


Figure 3: Points aggregated in regular units (Hexagons)

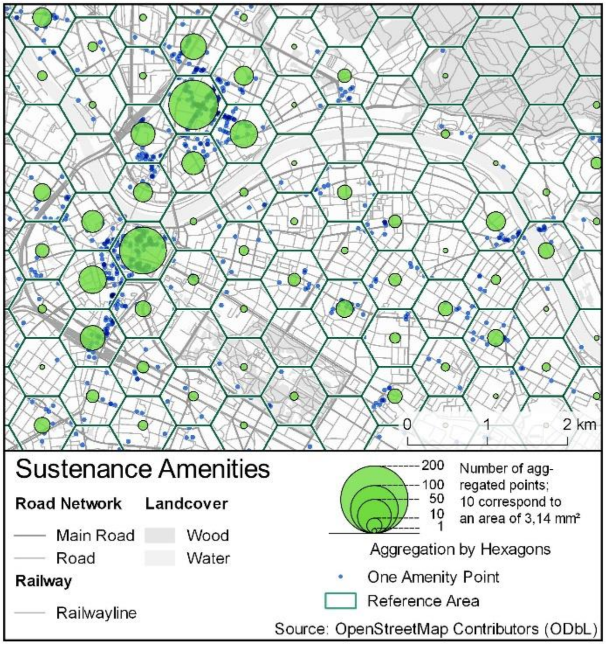
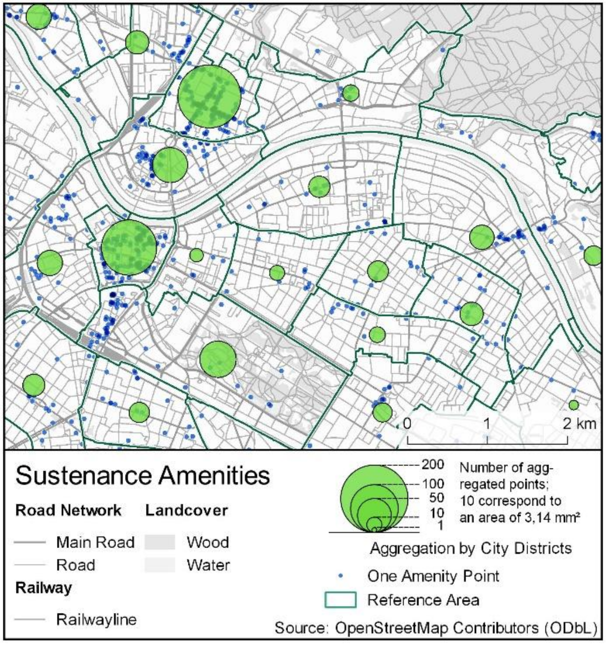


Figure 4: Points aggregated per administrative units


0
1
2 km
Sustenance Amenities
Road Network
Landcover
Main Road
Road
Wood
Water
Railway
Railwayline
200
100
50
10
1
Number of aggregated points; 10 correspond to an area of 3,14 mm²
Aggregation by City Districts
One Amenity Point
Reference Area
Source: OpenStreetMap Contributors (ODbL)

References

Anon: Geohash, Wikipedia [online] Available from: https://en.wikipedia.org/w/index.php?title=Geohash&oldid=759257309 (Accessed 26 January 2017), 2017.

Anon: Key:amenity – OpenStreetMap Wiki, Key:amenity – OpenStreetMap Wiki [online] Available from: https://wiki.openstreetmap.org/w/index.php?title=Key:amenity&oldid=1560322 (Accessed 30 January 2018), 2018.

Arnberger, E.: Thematische Kartographie: mit einer Kurzeinführung über EDV-unterstützte Kartographie mit Quellen der Fernerkundung, 3rd ed., Westermann, Braunschweig., 1993.

Bollmann, J. [Hrsg. .: Lexikon der Kartographie und Geomatik in zwei Bänden, Spektrum Akad. Verl., Spektrum Akad. Verl. [online] Available from: http://swbplus.bsz-bw.de/bsz089023811rez.htm, 2001.

Goodchild, M. F.: Citizens as sensors: the world of volunteered geography, GeoJournal, 69(4), 211–221, doi:10.1007/s10708-007-9111-y, 2007.

Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in KDD-96: proceedings, Portland Oregon. [online] Available from: http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf, 1996.

Töpfer, F.: Kartographische Generalisierung, Haack, Geograph.-Kartograph. Anst., Gotha [u.a.] : [online] Available from: http://slubdd.de/katalog?TN_libero_mab21163026, 1974.