Micro Diagrams: A Multi-Scale Approach for Mapping Large Categorised Point Datasets

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

Location-based social media from different platforms such as Twitter and Flickr increasingly serve as data source for many diverse research projects with their point-geocoded content. For analyses and visualisation, it is necessary to show distributions of categories in different scales and resolutions. This article introduces a new mapping method for large geospatial point datasets aggregated in an adjustable grid of diagrams. For example, a pie chart shows the numerical proportion, and the size or transparency of the chart symbolises the number of records. The Geohash serves as a data structure behind the aggregation of the points and for the generation of different aggregation levels. As result, it is possible to access multiple scales with a separate content zoom interaction and to carry out scale-dependent pattern analysis of multivariate point datasets.
Keywords: volunteered geographic information, cartography, geovisualisation, geovisual analytics

1 Introduction and Motivation

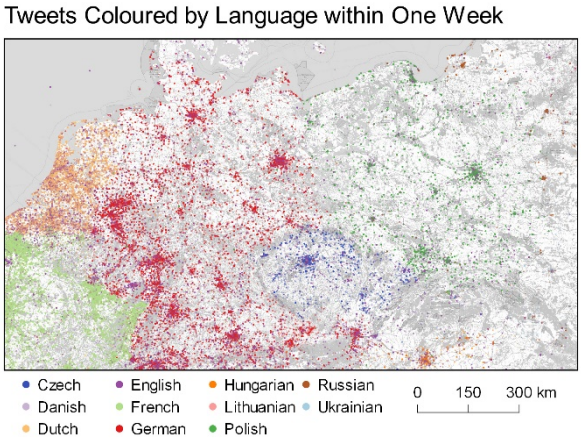
Within the last years, several new data sources and applications for mapping were established. With the rise of volunteered geographic information (VGI) and the consequence that open data became an important part of government and companies' business strategy. Social media platforms such as Twitter, Flickr or Instagram, reviews of restaurants or museums and position data of trains or airplanes are available via application programming interfaces (API).

This whole set of data is produced by people and contains many useful information, which are hidden in different spatial distributions. That is often associate with Big Data or, in our case, Big Geodata. Typical is a high data volume with millions of records, high velocity of data generation that can also quickly change and high variety of different events and topics. In most cases, the geometric reference is a geospatial point feature. It is important to describe this data as point collection and not as collection of points of interest (POI), because the distribution over all points is the subject of interest opposite to the concept of POIs where each point is important [1].

An often-used visualisation for point collections is the dot map in combination with Alpha Blending. That means every record is a transparent point. Overlapping points cause areas with more intensive colour as shown in Figure 1. One point represents one Tweet, and the colour shows the respective language. This causes overlapping effects and mix of different colours. In the map, it is possible to recognise borders between countries, because of differing most used languages and point colours. This issue is already well known [2].

Nevertheless, the presented solution can show distributions very detailed but also hide information. Particularly in potentially more interesting agglomeration regions as counterpart to rural areas, it is difficult to get detailed information. The reader cannot read out the distribution of qualities or quantities for these regions. It is only possible to describe the relative distribution over the map. Another aspect is that computation and drawing time is increasing with the number of geospatial points in a data set.

Figure 1: A dot map of geocoded Tweets coloured according to the language tag provided by Twitter.



2 Related work

Displaying and analysing large point dataset has been a research challenge for years. There are earlier publications about the Grid Fit Algorithm [3] and the ideas about "Visual Data Mining in Large Geospatial Point Sets" [4]. The proposed solution is to use empty grid cells to plot points instead of over plotting data points. Another approach to avoid the overplotting effects is to improve the dot map method: The reclassification of scatter plots and the similar dot map [5], colour blending for groups based on contour lines [6] or the computation of density maps with Bayesian Weighting [7].

The following publications demonstrates features that are a part of the Micro Diagrams visualisation method in context with similar visualisation tasks. An idea was to subdivide the space in regular cells and show the percentage of the categories by pixels within each cell [8]. One important feature was introduced in the VisCAT Application [9]: Many pie charts on a map to display the distribution of categorised point data. The possibility to use the spatial data structure quadtree for clustering based on cartographic aspects was already topic of an dissertation by Bereuter [10].

3 The Method

Our approach is not to plot every single point, but rather aggregate the points in a defined area and symbolise them by means of cartographic symbols. The following section describes each step of the method and our implementation to produce the shown maps.

3.1 Micro Diagram

The idea of the Micro Diagram method is to offer a simplified overview and detailed information on user's demand. On one hand the diagrams are small but while being large enough to recognise an individual feature, on the other hand mainly working with the overall impression. As minimal diagram diameter we propose a value greater or equal than 0.3 mm according to general cartographic minimal dimensions [11]. Furthermore, there should not be more than 10 diagrams per square centimetre to preserve readability [12]. Examples for suitable types of diagrams for this method are the pie chart, wing chart, polar area chart or the bar chart. Important is the fitting in a quadratic outline. Colours symbolise the categories and the diagram illustrate the numerical proportions. The size or transparency of the chart shows the number of data records for all categories. In this paper, we decided to use pie charts only to keep it simple and demonstrate the idea.

For displaying the different categories among the collated points, we use the slices of the pie chart. The diameter or the transparency of the diagram scales with the number of data records in this area. A possible visualisation result shows Figure 2 and Figure 3. The diagrams are relatively small that is not possible to recognise them at the first sight. On the second sight, it should be conceivable to recognise the diagrams. Areas with more Tweets attract more attention using diagrams with a bigger diameter or diagrams with less transparency. In this way, hotspots attract more attention than other zones with less visible diagrams. Behind this appearance stands a classification to symbolise different classes by size or transparency.

Figure 2: A Micro Diagram visualisation of geocoded Tweets coloured according to the language tag provided by Twitter and classified diameter according to the number of Tweets.

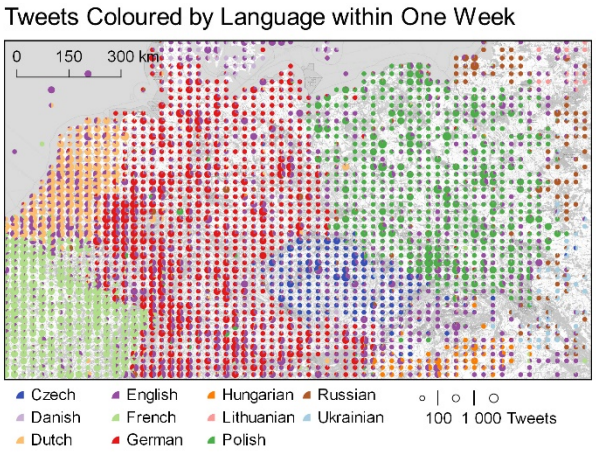
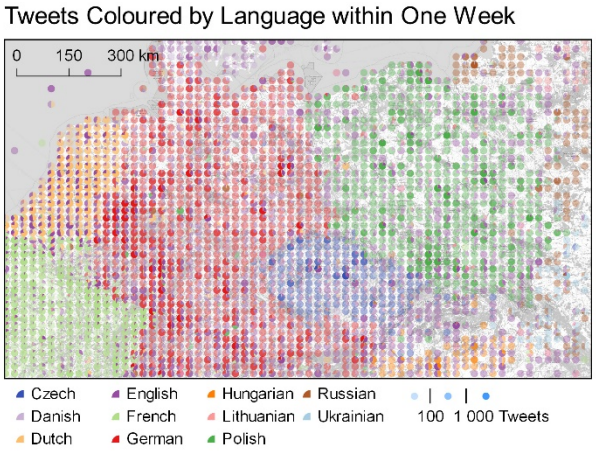


Figure 3: A Micro Diagram visualisation of geocoded Tweets coloured according to the language tag provided by Twitter and classified transparency according to the number of Tweets.

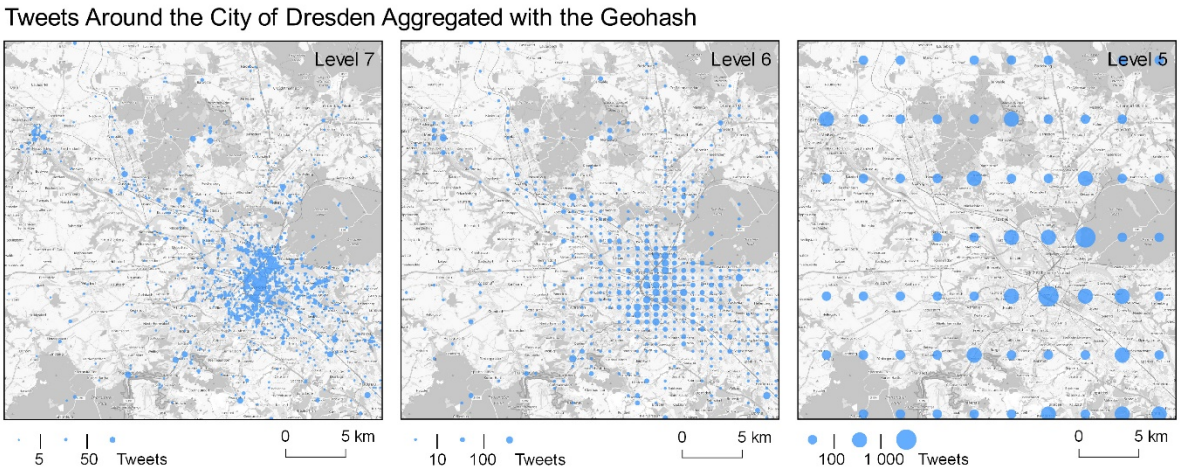


3.2 Aggregation

For displaying Micro Diagrams in a specific area, a spatial aggregation is required. There are two ways to cluster points in a hierarchical order: Clustering Algorithms are a suitable solution for a smaller amount of points with the advantage of an aggregation out of the data. The alternative is to use regular tessellation such as Raster [13], Hexagons [14] or data structures, i.e. Quadtree [10] or Geohash [15]. The benefits are a higher performance and the implementation in spatial databases. That is the reason we decided to use the Geohash for our demonstration scope.

The Geohash [16] is a hierarchical spatial data structure that may also serve as a geocoding system. It represents a pair of coordinates as a string. For instance, 57.64911 North and 10.40744 East results to "u4pruydqqvj". Removing characters from the end of the code reduces the precision. Recoding

Figure 4: Different spatial solutions produced by varying the length of the Geohash at the same scale: Left seven characters, middle six characters, right five characters. The size of the points is a measure for the number of aggregated points.



"u4pru"[1] back to geographic coordinate's results to 57.6 North and 10.4 East. That can also refer to 57.647 North and 10.401 East or another point starting with these numbers. In this way, a Geohash also describes a bounding box.

It is an efficient solution to compute the Geohash and afterwards execute a selection on the first characters of the resulting string, which is an aggregation, too. The length of the string regulates the precision and the spatial resolution. There is an example in Figure 4, which demonstrates the advantage of different aggregations steps. It is imaginable to separate spatial resolution and map scale. The results provide different perspectives on the same dataset. In addition, there might be an optimal spatial resolution for each scale to describe patterns such as the map in the middle of Figure 4. It is possible to identify hotspots in the data and areas with less or no data. Additionally, it looks tidy and clearer.

3.3 Interaction

The Micro Diagram method deals with three steps towards getting a visualisation: selection of the spatial resolution, classification of data and the adjustment of visual variables. All these phases can control the user, which makes this method highly interactive. In addition, the operator can adapt the map extent and the zoom level to compare the results with other regions. Nevertheless, there are some conceivable additional yet useful features such as a magnifying glass, a mouse over effect to view the diagrams in detail or smart legends [17].

Today, users expect those features in a visualisation to interact with it and retrieve more information. By adjusting the spatial resolution with the help of a spatial data structure such as the Geohash, we added a content zooming functionality to our method. The user is able to control the map scale by zooming in or out and can adjust the aggregation as second parameter. That enables the discovery of patterns on aggregation levels.

4 Comparison

A closer comparative look on Alpha Blending and Micro Diagrams presents Figure 5. These two maps show details of the previously discussed Figure 1 and Figure 2 in the southwest of Germany. In both visualisations, the reader can identify the border between Germany and France by German Tweets in red and French Tweets in light green. The dot map shows more details and makes it easier to recognise urban areas and the background map than on the Micro Diagram map.

An interesting observation is an area of majoritarian English Tweets (purple) in Germany visible on both visualisations. That might be due to a base of the U.S. army situated in this area. In the dot map visualisation, it is a bit more difficult to recognise this because the coloured area is noticeably smaller. An advantage of the Micro Diagram method is the possibility to get a clearer impression on distributions in hotspots such as cities. Not only is the amount of Tweets higher in urban areas, the number of languages on Twitter is significantly higher as well. That is visible in the lower right corner of the enhanced detail map (Swiss). On the dot map, the user only can identify a region with many Tweets in different languages. The Micro Diagrams show all languages and their portions.

5 Discussion

Mostly, the Micro Diagram solves the problem of the visualisation of large categorised point collections. There is no over plotting or overlapping on the map. That produces a clearer statement and allows using a background map to provide information in context to their position. The possibility of details on demand by using the Geohash as multi-resolution gird creates an efficient way to generate a variety of different perspectives on one data set. It aggregates point data on demand to provide a better overview alternative to a detailed view. It can be surely stated that, it is conceivable to read quantity and quality out of the map.

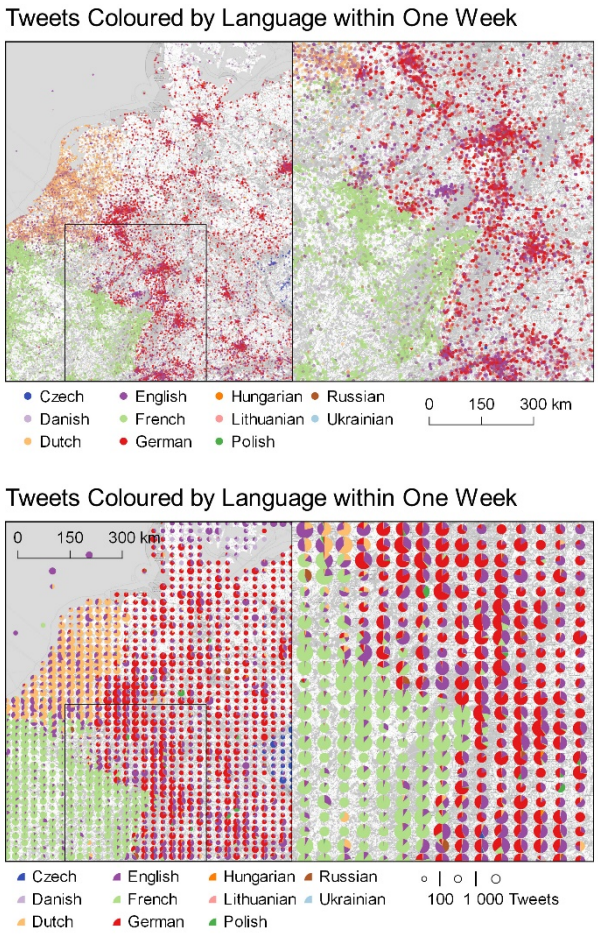
[1] http://geohash.org/u4pru

It is also imaginable to produce results such as the dot map with the Micro Diagrams using less aggregation by the Geohash, although this is not the initial intention. The left map within Figure 4 illustrates this exemplarily.

The Geohash as an optimised data storage and structure that is implemented in relational and non-relational data stores underpins the presented methods. That helps to generate the aggregation of millions of points in a few moments.

It is important to be sensible of possible drawbacks: Understanding the visualisation takes time. Furthermore, it is more complex to deal with the possibility of content zoom plus the parameters behind that, which might have a strong influence on the visualisation and their interpretation. Another aspect is the limited number of categories. This is a problem if we would expand the example for, e.g., all Europe. There are more than twelve different languages. The Colorbrewer[2] proposes only twelve distinguishable colours. However, this already allows the visualisation of a reasonable number of categories.

Figure 5: Comparison between a dot visualisation with Alpha Blending and Micro Diagram visualisation of Tweet languages for the southwest of Germany in detail.



6 Conclusion

The presented research introduces a new visualisation concept for point collections that offer a methodology for comparative mapping of quantity and quality for distribution and is adjustable in spatial resolution independently from the map scale. A first implementation uses pie charts as diagrams and the Geohash as data structure.

The development of an end-user suitable solution, usability test to get feedback from users and to overcome some general limitations are subject to future investigation: The method is more suitable for visual analysis than for consumers; the number of twelve identifiable colours restricts the maximum count of categories and the count of classes for size or transparency are also limited. The proposed solution is applicable in different use cases such as analysis of movement, sentiment or event-related data and solves the problem of overlapping features of the dot map method.

References

[1] P. S. Bereuter, "Quadtree-based Real-time Point Generalisation for Web and Mobile Mapping," Universität Zürich, Zürich, 2015.

[2] E. Fischer, "Language communities of Twitter (European detail)," 24-Oct-2011. [Online]. Available: https://www.flickr.com/photos/walkingsf/6276642489/in/album-72157631997324222/. [Accessed: 07-Sep-2015].

[3] D. A. Keim and A. Herrmann, "The Gridfit algorithm: an efficient and effective approach to visualizing large amounts of spatial data," in *Visualization '98. Proceedings*, 1998, pp. 181–188.

[4] D. A. Keim, C. Panse, M. Sips, and S. C. North, "Visual Data Mining in Large Geospatial Point Sets," *IEEE Comput. Graph. Appl.*, vol. 24, no. 5, pp. 36–44, Sep. 2004.

[5] H. Chen *et al.*, "Visual Abstraction and Exploration of Multi-class Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1683–1692, Dec. 2014.

[6] A. Mayorga and M. Gleicher, "Splatterplots: Overcoming Overdraw in Scatter Plots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 9, pp. 1526–1538, Sep. 2013.

[7] M. Correll and J. Heer, "Surprise! Bayesian Weighting for De-Biasing Thematic Maps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 651–660, Jan. 2017.

[2] http://colorbrewer2.org/

[8] J. R. Miller, “Attribute Blocks: Visualizing Multiple Continuously Defined Attributes,” *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 57–69, May 2007.

[9] T. M. Ghanem, A. Magdy, M. Musleh, S. Ghani, and M. F. Mokbel, “VisCAT: Spatio-temporal Visualization and Aggregation of Categorical Attributes in Twitter Data,” in *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2014, pp. 537–540.

[10] P. Bereuter and R. Weibel, “Real-time generalization of point data in mobile and web mapping using quadtrees,” *Cartography and Geographic Information Science*, vol. 40, no. 4, pp. 271–281, Sep. 2013.

[11] Ernst Spiess, Ulrich Baumgartner, Stefan Arn, and Claude Vez, *Kartengrafik und Generalisierung*. Schweizerische Gesellschaft für Kartographie, 2002.

[12] J. Bertin, *Graphische Semiologie*. Berlin, New York: DE GRUYTER, 1974.

[13] L. Ding, “Visual Analysis of Large Floating Car Data - A Bridge-Maker between Thematic Mapping and Scientific Visualization,” TU München, München, 2015.

[14] R. Feick and C. Robertson, “A multi-scale approach to exploring urban places in geotagged photographs,” *Computers, Environment and Urban Systems*, vol. 53, pp. 96–109, Sep. 2015.

[15] A. Fox, C. Eichelberger, J. Hughes, and S. Lyon, “Spatio-temporal indexing in non-relational distributed databases,” in *Big Data, 2013 IEEE International Conference on*, 2013, pp. 291–299.

[16] “Geohash,” *Wikipedia*, 10-Jan-2017. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Geohash&oldid=759257309. [Accessed: 26-Jan-2017].

[17] S. Wiesmann, “Smart Legend - Aufbau und Interaktivität digitaler Kartenlegenden,” Universität Zürich, Zürich, 2007.