Assessing Dot-Map Aggregations

Wouter Meulemans* TU Eindhoven
Martijn Tennekes† Statistics Netherlands

ABSTRACT

Compositional geospatial data can be visualized as dot maps, where the color of each dot represents its class. For interactive dots maps, where it is possible to zoom out in order to see the global picture, it is often needed to aggregate the dots. Hence, we face the following aggregation problem: let M be an input matrix where each cell is assigned a class; find an aggregated matrix A in which each cell aligns with k by k cells of M such that A is a good summary of M. We distinguish three dimensions of "good summary": class balance, representation and presence. The first is holistic, whereas the other two capture spatial aspects. We propose a simple heuristic algorithm and explore the three quality dimensions with a visualization tool.

Index Terms: Human-centered computing—Visualization

1 INTRODUCTION

Dot maps are a geovisualization tool in which data elements are rendered as colored dots at their location on a map. Our scope is unordered categorical data, where the color represents the category of the data element. For example, individuals represented as dots colored by ethnic origin [1,2]. The purpose of dot maps is to acquire insights in the geospatial distribution as well as the composition of the data [3]. The ability to zoom in interactive dot maps is particularly useful to see the global picture and local details. At the highest resolution, each individual data element can be represented as a dot. However, at lower resolutions, so when zoomed out, dots have to be aggregated to maintain a legible map.

Aggregation of the data is typically done by blending the colors in the same pixel [3]. Though computationally and conceptually straightforward, blended colors become hard to distinguish and categorize. Moreover, in many applications the individual dots do not correspond to the exact location of the data elements due to privacy concerns; instead, dots are placed only roughly in the right neighborhood to avoid linking data to particular individuals or households. Often, it is more important to retain the distribution and composition of the dots after aggregation than their exact location.

Contributions We propose explicit aggregation as an alternative to color blending. In particular, we aggregate the original dots into larger dots and assign each such dot to a single category, present in the input – that is, we do not add new classes to represent a mix of classes. This further exploits the inherently desirable spatial anonymization, while keeping a clean and simple visualization. We propose three quality dimensions to assess aggregation quality. Using visualization combined with a simple greedy algorithm, we explore these dimensions on some initial results.

Preliminaries Throughout this abstract, we assume our dot maps to be aligned to a grid; that is, the data can be thought of as a matrix with rows and columns, each cell representing a single data element. We use M to denote the input matrix of sk rows and sk columns and A the aggregated matrix of s rows and s columns, such that each entry in A represents $k \times k$ entries of M. Each entry is assigned one of the data classes in a set $\mathscr{C}$. For a class $c \in \mathscr{C}$, we use $\mathrm{F}(X,c)$ to denote the cells in matrix X that have class c. We use $\mathscr{C}$ also as a function, mapping an entry to its assigned class in $\mathscr{C}$.

*e-mail: w.meulemans@tue.nl
†e-mail: m.tennekes@cbs.nl

2 MEASURING AGGREGATION QUALITY

We consider three dimensions to quantify aggregation quality: class balance, representation and presence.

Class balance To capture the overall composition, the relative number of dots of each color should remain the same. We can quantify deviation in various ways. We use a simple sum of squares on the differences, expressed in the formula below. We use $k^2\mathrm{F}(A,c)$ as each entry in A has the size of k^2 dots in the input map.

$$\sum_{c\in\mathscr{C}} \Big(|\mathrm{F}(M,c)| - k^2|\mathrm{F}(A,c)|\Big)^2$$

Representation Each aggregated dot a in A represents up to k^2 input dots in M of the same class. Each smaller dot can be represented only once. That is, we assume that each aggregated dot has a set S_a containing up to k^2 dots of the same class, such that all sets S_a are pairwise disjoint. To capture spacial aspects, we want an appropriate representation: that is, a should be a good summary of those k^2 smaller dots in S_a. The further away the smaller dots are from a, the worse a represents them. We propose to quantify this via the sum of squared Euclidean distances, to penalize large distances:

$$\sum_{a\in A} \Big((k^2 - |S_a|)D^2 + \sum_{s\in S_a} \|a-s\|^2\Big)$$

where $\|a-s\|$ indicates the Euclidean distance between the center of a and of s and D is a constant, penalizing aggregated dots that represent fewer that k^2 input dots. We can find the optimal sets S_a via a minimum-cost flow computation. The parameter D then ensures that the distance between a and a dot in S_a is at most D.

Presence Complementing representation, we ideally want that each input dot is present in the aggregated map in the form of a nearby aggregated dot. This particularly considers the effect of local minorities, which are not necessarily captured by the other dimensions. We measure the presence of a dot as its squared Euclidean distance to the nearest aggregated dot of the same class.

$$\sum_{m\in M} \min_{a\in\mathrm{F}(A,\mathscr{C}(m))} \|m-a\|^2$$

We explicitly choose to not use a matching between small and large dots, to keep this dimension independent from representation. Specifically, a large dot may be the closest for more than k^2 small dots.

3 A VISUAL-ANALYTICS APPROACH

To support our reasoning and validate our measures, we developed a simple visual-analytics tool that will allow us to view, compute, analyze, and interact with dot maps and their aggregations.

Algorithm Starting from an empty map A, our approach iteratively picks the class c that minimizes $\mathrm{F}(A,c)/\mathrm{F}(M,c)$, i.e., that has the largest relative class imbalance. For this class c, it then finds the empty entry $a \in A$ that achieves the best representation if a is assigned class c. It sets the class of a and S_a accordingly[1] and removes S_a from M such that they cannot be used in next iterations.

[1]These computed S_a are not optimal, but can be recomputed afterwards.

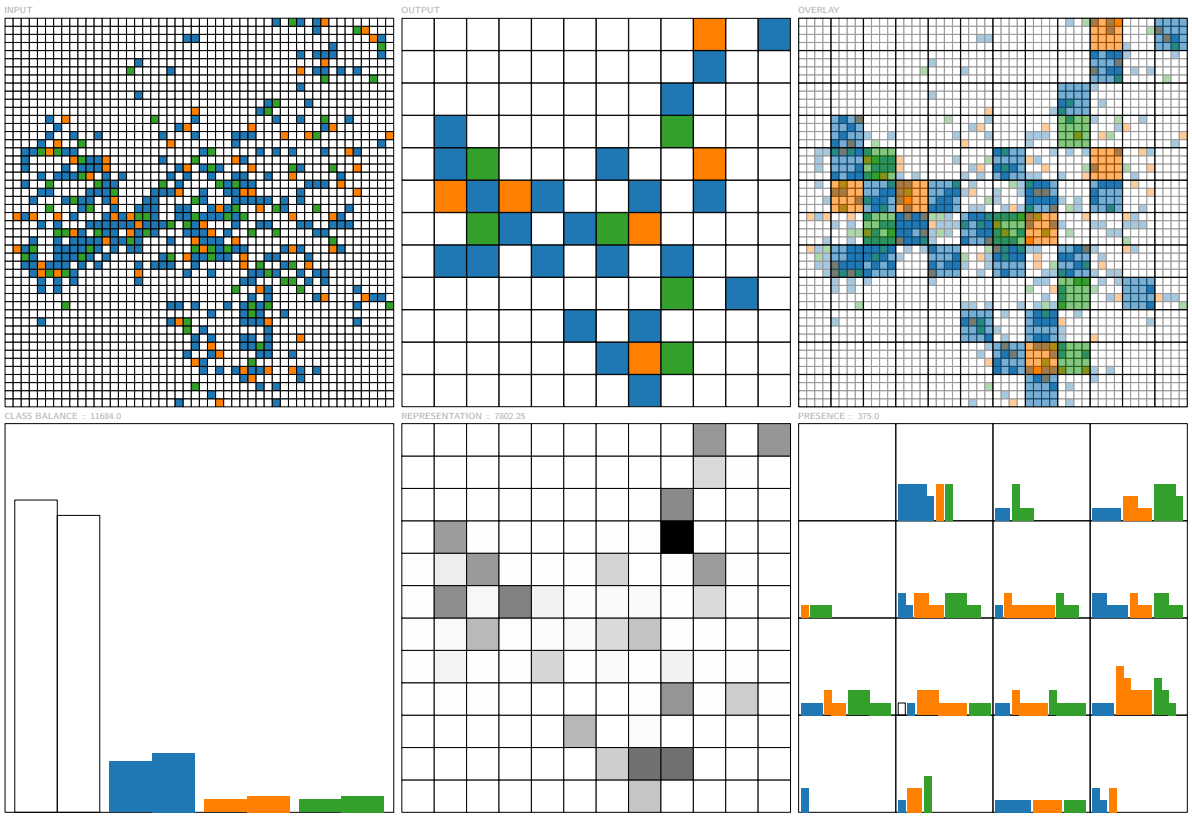

Figure 1: Main view of our tool to explore dot map aggregation. Top row shows the dot maps, whereas the bottom row shows analysis results.

The visualization tool The main view of our tool is shown in Fig. 1. At the top are three views showing input, output and their overlay. Hovering dots allows linking and exploring the maps according to the sets S_a used to compute representation.

At the bottom are views that allow visual assessment of the three dimensions of aggregation quality. Each such view is an interactive small multiples, showing analogous graphics in an array of frames. The detail level (i.e., the number of small frames) can be configured, starting from a frame for each output dot to a single frame for the entire map. Each view is also annotated with the overall score in that dimension. From left to right, the views are used as follows:

A bar chart to show class balance (for the entire map): the left bar of a color indicates the fraction of dots in the input map of that class, the right bar shows the fraction of dots in the output map.

A heat map to show representation (each output dot individually): darker dots show worse representation. Values are scaled such that the worst representation is black.

A chart to show presence (frames match the underlying 3×3 output submatrices): the 12 input dots with the worst score are each shown as a bar, its height indicating the presence score for that dot.

4 Discussion

A first observation is the special role of blank cells. We seem to intuitively focus on the colored dots that represent data, whereas the blanks are easily overlooked. While this is certainly desirable to some degree, it does cause a seemingly odd effect: in a region with many blanks, but also some colored dots, the best choice for class balance and representation (which are mostly driving our heuristic approach) will be to keep the blank as aggregated dot; for presence, however, it would be beneficial to inject more color into such places. We used this also in our algorithm, which can be configured to give a structurally lower priority to the blank class – this also is the reason why the other three classes are overrepresented in the result of Fig. 1 (bottom left).

Generally, our three dimensions cannot always be optimized simultaneously as they measure conflicting aspects. Beyond validating our measures, we need to understand which dimensions are more important and to what degree.

Algorithms that optimize for different dimensions are necessary to further explore these trade-offs. Eventually, it would be worthwhile to design algorithms that use parameters that reflect these trade-offs. The question is how (if at all) we can efficaciously combine the dimensions into a single judgment on aggregation quality. Realistically, there is not going to be a single unique answer for all use cases. Rather, we need to understand the above in context of various conditions: how does one pick the appropriate algorithm to match the needs of the use case?

So far, we focused on just one input and output map to assess the aggregation quality. Another interesting dimension may reveal itself when we consider a sequence of maps, aggregated in increasingly higher levels: how do we balance the quality of all the aggregations in the sequence? Moreover, it will be important to consider the stability of such a sequence, to avoid visual artifacts for interactive dot maps where the aggregation level adapts according to zoom level.

References

[1] D. Cable. The racial dot map. Published online http://www.coopercenter.org/demographics/Racial-Dot-Map, 2013.

[2] Statistics Netherlands. CBS experimenting with dot maps. Online publication https://www.cbs.nl/en-gb/our-services/innovation/project/cbs-experimenting-with-dot-maps, 2017.

[3] M. Tennekes and E. de Jonge. Coloring Interactive Compositional Dot Maps. In *EuroVis 2016 - Posters*, 2016.