Task-oriented Data Classification of Choropleth Maps for Preserving Local Extreme Values

JUIWEN CHANG[1] & JOCHEN SCHIEWE[1]

Abstract: It is well known that the selection of class breaks for choropleth maps, as a powerful tool widely used to illustrate values among polygons by geographers and cartographers, have a significant influence on user perception. Furthermore, conventional approaches do not take spatial context into consideration. This contribution presents an innovative algorithm for the task of preserving local extreme values and compares it to other classification methods. With that, we cannot only prove the enhanced visual effect but also the quantitative improvement in terms of preserved local extreme values in choropleth maps.

1 Introduction

In thematic maps, e.g. for representing population densities, the attribute values of single regions (like districts) can be categorized into representative classes and respective colors. The visual impact of these choropleth maps – and with that derived interpretations and decisions – is influenced by the actual value distribution of the input data, the data classification method and the selection of color ramps.

Conventional data classification methods (like equidistant, quantiles, or Jenks) work data-driven. With that they explicitly do not consider any spatial context within the dataset. If, for example, a state that contains a local extreme value is put into the same category as some of its neighbor states, this important information gets lost. Moreover, neglecting geographical relationships among adjacent polygons may also influence other patterns like hot spots or clusters.

Instead of the conventional data-driven approach, this contribution will present a new task-oriented classification approach that has been developed and implemented for the task of preserving local extreme values.

2 Method

For the task of preserving polygons showing local extreme values, the following three-step procedure has been implemented and tested.

2.1 Identification of local extreme values

A local extreme value is defined as a polygon that shows a larger (or smaller) value compared to all its neighboring polygons. Neighbors are defined as polygon pairs having a common boundary or sharing a common vertex. An R-Tree spatial index was used to accelerate the search for neighboring polygons. Among all neighbors of an actual extreme value candidate, the minimum of the absolute difference values is identified and stored. In fig. 1, for example, this difference

[1] HafenCity University Hamburg, Lab for Geoinformatics and Geovisualization (g2lab), Überseeallee 16, D-20457 Hamburg, Germany, E-Mail: [juiwen.chang, jochen.schiewe]@hcu-hamburg.de

value is 10.17. The reason why we select the minimum absolute difference is because we want to find one break value that separates this local extreme polygon from all its neighbors. In the case of fig. 1, any value between 7.52 and 17.69 can be used for such a separation.

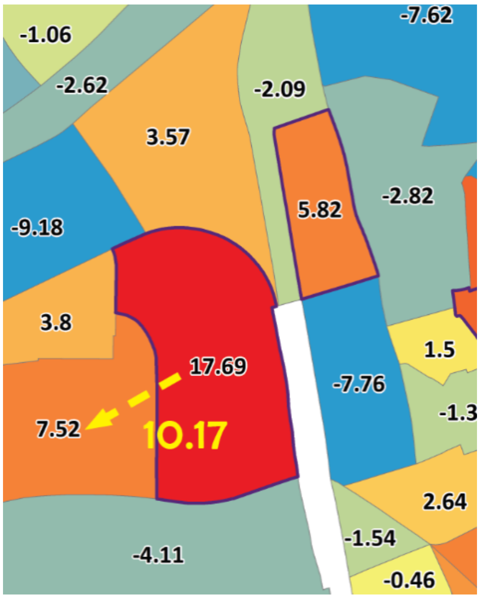

Fig. 1: Identified polygon with local extreme value (value 17.69) and the "closest" neighbor (value 7.52)

2.2 Sorting all local extreme value polygons according to significance

For all local extreme values the corresponding value intervals (bounded by the values of the extreme polygon and the numerically closest neighboring polygon as derived in step 2.1) are plotted and sorted from largest to smallest interval widths (Fig. 2). The larger the interval width (e.g., value 10.17 in Fig. 1) is, the more important or "significant" this local extreme value is in comparison to other candidates.

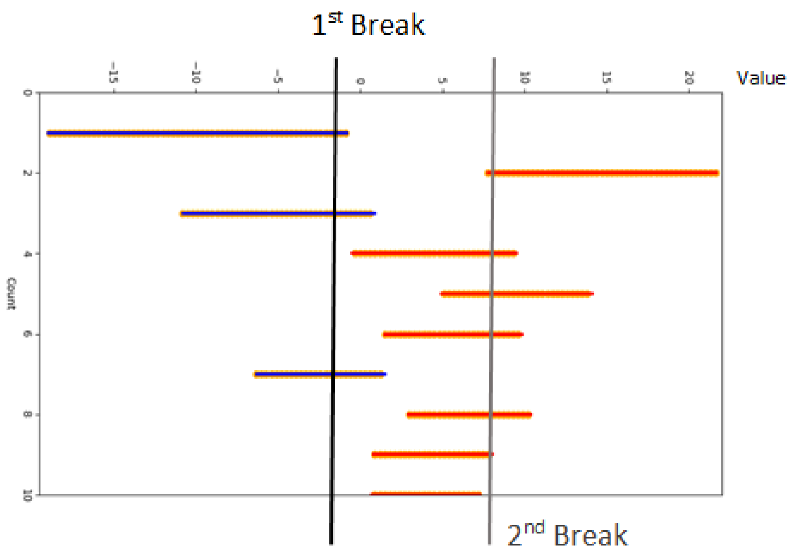

Fig. 2: Intervals between extreme value polygons and closest neighbors (blue: local minimum, red: local maximum) and setting of class breaks

2.3 Setting class breaks

In the last step, a Plane Sweep Algorithm is carried out in order to retrieve optimal intersections (i.e., class boundaries; Fig. 2). For this, a vertical line is swept from left (global minimum) to right (global maximum) in order to retrieve total number of intersection counts for each line. Starting with the most significant extreme value polygon (rf. step 2.2), a class break is set within the corresponding value interval where the number of intersections has its maximum. After identifying the first class break, those segments which intersect this one, are excluded and then the next sweep is performed. This algorithm repeats the above steps until the total number of classes (predefined by the user) have been met or all intervals have been integrated.

3 Results and Evaluation

In this chapter, the task-oriented local extreme method was implemented and tested through datasets with different characteristics. An evaluation was performed with QGIS in order to retrieve a better understanding regarding the relationship between the preservation rate of local extreme values and numbers of classes. In this example, the deprived index from the City of Hamburg at so-called statistic district level in the year 2016 is used (see section 7 for a description of this dataset). It comprises social monitoring indices including unemployment rate, children with immigration background and single parents, long-term social help applicants (SGB II recipients) and education level. The original dataset consists of 847 records. Indices were nominalized and ranged from -19.03 to 21.62 with a standard deviation of 2.926.

3.1 Visual comparison

For comparison purposes, Fig. 3 shows visualizations of this data set using different data classification approaches (with 7 classes) and the same color scheme.

It demonstrates how different classification approaches can influence the recognition of neighboring geographical elements. A local maximum (red) (value 4.14) is preserved with the task-oriented method (aChor), similarly for nature breaks (Jenks). Conversely, equal interval and quantile classification are not able to strictly preserve the local maximum value in this case. The local minimum (blue) (value -4.58) is well preserved with the task-oriented method whereas the other three methods are not able to distinguish this value from its surrounding elements.

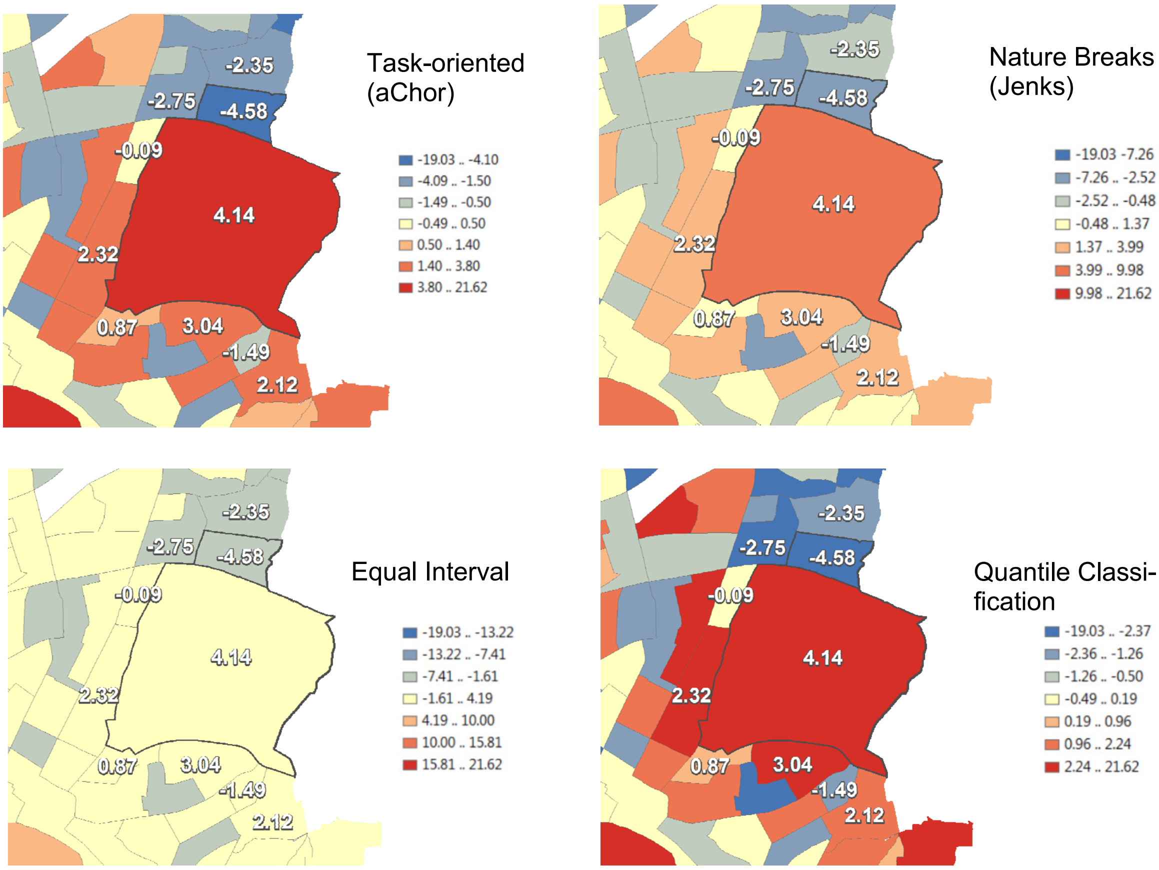

Fig. 3: Comparison of task-oriented (aChor) with conventional data classification methods

3.2 Numerical comparison

In order to get an overall understanding of the local extreme value preservation rate for different classification methods, Table 1 compares aChor with other methods for a fixed number of seven classes.

Tab. 1: Task-oriented data classification in comparison with conventional method (7 classes)

<table>
<tr><th></th><th>No. of local maximum values (Max)</th><th>No. of local min-imum val-ues (Min)</th><th>Preserved Max</th><th>Preserved Min</th><th>Total %</th></tr>
<tr><td>Task-oriented method (aChor)</td><td rowspan="5">136</td><td rowspan="5">135</td><td>91</td><td>90</td><td>66.8%</td></tr>
<tr><td>Quantile Classification</td><td>76</td><td>86</td><td>59.8%</td></tr>
<tr><td>Natural Breaks Jenks</td><td>73</td><td>68</td><td>52.0%</td></tr>
<tr><td>Geometrical Interval</td><td>70</td><td>65</td><td>49.8%</td></tr>
<tr><td>Equal Interval</td><td>37</td><td>53</td><td>33.2%</td></tr>
</table>

Extending this evaluation, Fig. 4 shows dependency of the preservation rate on the number of classes. In this example, aChor reaches a 100 % preservation rate when 36 classes are used. In general, the quantile classification method has a rather good performance with a smaller number of classes; nevertheless it shows no advantage anymore when compared with Jenks with class numbers larger than 20.

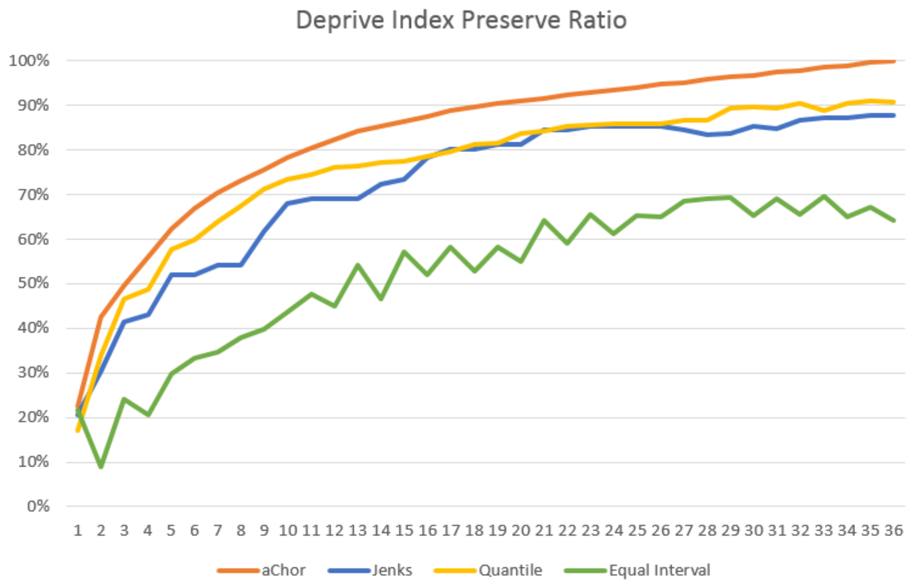

Fig. 4: Preservation rate within different data classification method with the same dataset

4 Conclusion and Discussion

We presented a method that preserves spatial relationships, in particular local extreme values. Our studies (beyond the aforementioned datasets, further examples have been tested (see section 7) and have proven visually and numerically that the rate of preservation for these features at any given number of classes is significantly better compared to conventional data classification methods. Also a 100 % preservation rate is reached with a smaller number of classes compared to all other methods.

Nevertheless, presently this method needs to go through further implementation and adjustment as well as an application of other datasets that show different statistical distributions.

5 Acknowledgements

The work presented here has been conducted within the project *"Aufgabenorientierte Datenklassifikation und Gestaltung von Choroplethenkarten (aChor)"*, funded by the German Research Foundation (DFG).

6 Literature

SCHIEWE, J., 2016a: Preserving attribute value differences of neighboring regions in classified choropleth maps. International Journal of Cartography, **2**(1), 6-19. DOI: 10.1080/23729333.2016.1184555

SCHIEWE, J., 2016b: Aktuelle Forschung und Entwicklung in der Kartographie – das Beispiel der aufgabenorientierten Gestaltung von Choroplethenkarten. In: Harzer, C. (Ed.): GIS-Report 2016/17, Harzer-Verlag: 9-16.

SCHIEWE, J., 2017: Data Classification for Highlighting Polygons with Local Extreme Values in Choropleth Maps. In: Peterson M.P. (ed.): Advances in Cartography and GIScience, Lecture Notes in Geoinformation and Cartography, Springer: 449-459.

7 Attachment: Dataset description for testing purposes

Deprived index:
http://suche.transparenz.hamburg.de/dataset/sozialmonitoring-integrierte-stadtteilentwicklung-bericht-2016-anhang

Yearly average age of the inhabitants of German communities in 2011:
https://www.destatis.de/DE/Methoden/Zensus_/Downloads/2F_BevoelkerungAlterGeschlecht.html

SPD second votes in the Hamburg districts in the federal elections 2017: https://www.statistik-nord.de/wahlen/wahlen-in-hamburg/bundestagswahlen/2017/