International Cartographic Association
Association Cartographique Internationale

Data classification methods for preserving spatial patterns

Jochen Schiewe [a,*]

[a] *HafenCity University of Hamburg, Lab for Geoinformatics and Geovisualization (g2lab), jochen.schiewe@hcu-hamburg.de*

* Corresponding author

Abstract: The primary purpose of choropleth maps is to display or even to emphasize special relationships or patterns in the spatial distribution of attribute values. However, because classification methods commonly used and implemented in software packages (such as equidistance, quantiles, Jenks, etc.) are data-driven, a preservation of such spatial patterns is not guaranteed. Instead of such a data-driven approach in the following a task-oriented procedure is pursued: For typical patterns (local and global extreme values, large value differences to neighbours, spatial clusters, hot/cold spots) specific algorithms have been developed, implemented and tested.

Keywords: choropleth maps, data classification, spatial patterns

1. Introduction

The primary purpose of thematic maps is to present statistical data as well as the results of spatiotemporal analyzes - be it for conveying information and knowledge (usually for lay people) or for exploration purposes (usually for expert users). More precisely, it is often a matter of pointing out special relationships or patterns in the spatial distribution of attribute values - these can be, for example, global or local extreme values of election results, hot spots of crime frequencies or clusters of industrial companies. It is obvious that there cannot be one and only one objective or even “correct” presentation. However, it is absolutely necessary that the design of cartographic representations, depending on the purpose of the map, should emphasize or at least receive certain information or “messages”. This idea will be followed within this article – with the focus on choropleth maps.

In order to achieve a better overview or faster readability in choropleth maps, the attribute values to be displayed are very often classified in advance. The classification methods commonly used and implemented in software packages (such as equidistance, quantiles, Jenks, etc.) are data-driven, i.e. the intervals are determined exclusively on the basis of the existing frequency distribution of the original values. However, the spatial context of the underlying data, which is important for many questions, is completely neglected in such a division along the number line. This means that information about spatial relationships or patterns and with that the proposed statements on a map can be lost (see example in Fig. 1).

Instead of such a data-driven approach in the following a task-oriented procedure is pursued. Firstly, spatial synoptic tasks like the representation of local and global extreme values (outliers, resp.), large value differences between neighboring polygons, spatial clusters, and hot/cold spots are specified. For each of them, specific new algorithms have been developed, implemented and tested in order to preserve the existing patterns.

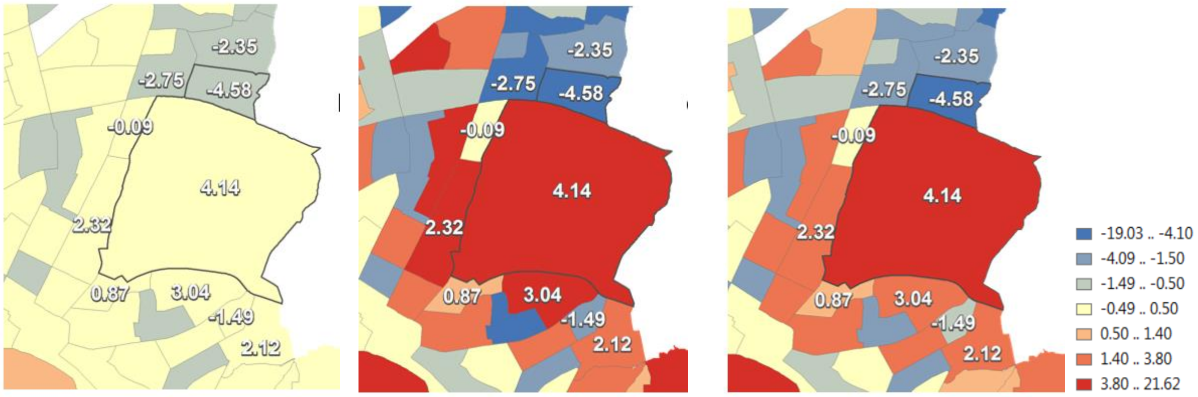


Fig. 1: Failed graphical preservation of local extreme value (“4.14”) with equal interval (left) and quantile classification methods (middle) – extreme value polygon shows same color like some neighboring polygons; successful preservation with new method (right; see also section 3.1). Note: Numerical values are typically not shown in choropleth maps.

Proceedings of the International Cartographic Association, 4, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-4-95-2021

After a short review of related work (section 2) the development of algorithms, which was carried out separately for the preservation of specific spatial relationships or patterns, is summarized in section 3. Implementation issues are given in section 4. Conclusions and outlook complete the contribution (section 5).

2. Related Work

The topic of data classification for cartographic purposes is dealt with extensively, for example in the overview articles by Cromley & Cromley (1996) or Coulsen (1987). The vast majority covers data-driven methods. Some empirical studies also compare classification methods for answering typical map usage tasks (e.g. Goldsberry & Battersby, 2009; Brewer & Pickle, 2002; Mersey, 1990). In addition, interactive tools have been developed to help find the "optimal" choice for a given application; e.g. by using linked views between the data histogram and the choropleth map (Andrienko et al., 2001). In view of this research focus, it is not surprising that only data-driven data classification methods are implemented in current GIS or cartography software.

On the other hand, relatively little has been published on the problem of neglecting the spatial context, or the use of a task-oriented approach, resp. Armstrong et al. (2003) provide an overview of this. Attempts to describe and maintain value differences in spatial neighbors have been published by Smith (1986), Monmonier (1972) or Jenks & Caspall (1971). Often, however, the aim is to simplify the displayed patterns so that the map user can grasp the rough tendencies in the display more quickly without being disturbed by a detailed and "spotty" impression (Andrienko et al., 2001; Cromley, 1996; MacEachren, 1994; MacDougall, 1992). With such an approach, however, it is not guaranteed that any significant variations or patterns that may be present will be retained.

3. Methods for spatial pattern preservation

3.1 Polygons with local and global extreme values

Local extreme values are defined as polygons that show larger (or smaller) values compared to all neighbor polygons. With conventional data classification methods, this information might be lost because the local extreme value is put into the same class as one or more of its neighbors (Fig. 1). The method for preserving this type of spatial patterns, which has already been presented in Schiewe (2017), has been further developed and tested.

In a first step, all adjacent polygons to all candidates have to be identified. Adjacent polygons are defined by sharing a common boundary or a common vertex. An R-Tree spatial index is used to accelerate this spatial search for all polygon relations within the dataset. In order to allow a fast access, the neighbor relations are stored in a *SQLite* database. Among all neighbors of an actual local extreme value candidate, the minimum of the absolute difference values is identified and stored together with the local extreme value.

Ideally, at least one class break should to be placed within this interval. For setting the new class breaks, a line sweeping algorithm is applied as follows (Fig. 2): A vertical line is swept from the global minimum to the maximum value of the interval ranges. The goal is to retrieve a maximum number of intersection counts for each sweep step. A class break is set within the corresponding value interval where the maximum number of intersections happens. After identifying the first class break, those segments which intersect this class break are excluded before the next sweep is performed. The algorithm repeats this plane sweep process until the desired number of classes has been met or all intervals have been intersected.

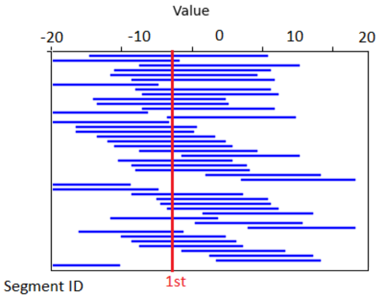


Fig. 2: Setting first class break (red line) using line seep algorithm

In order to prove the effectiveness of the proposed classification methods and to get an understanding of the parameter settings, a visual (see Fig. 1) and quantitative comparison to conventional methods (equidistant, quantiles, Jenks) has been performed for the local extreme value method. As Fig. 3 shows, there is the expected increase in preservation local extreme values with the number of classes for all methods – with the new method being the best methods in all instances.

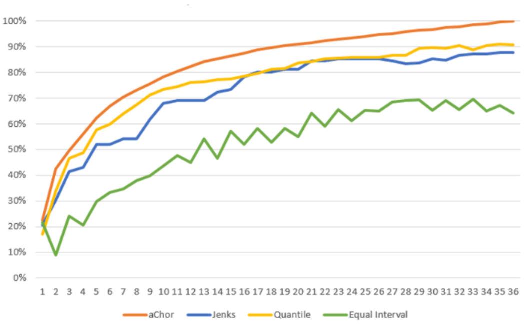


Fig. 3: Local extreme value preservation rates (upper axis) with different data classification methods (new method in orange) as function of number of classes (right axis)

A key parameter within this method is the sweep line interval. Default settings were derived from tests that took the preservation rate and execution time into account (with the latter being the critical parameter; Fig. 4).

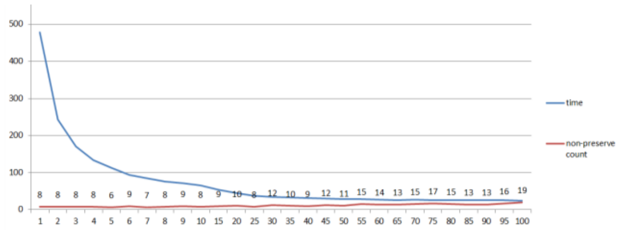


Fig. 4: Execution time (blue line) and count of not preserved local extreme values as function of sweep line interval (right axis)

The extraction of global extreme values requires the search for the global maximum and global minimum in a dataset using standard search algorithms. For both values, the most upper and lower classes are reserved, while the remaining classes are defined by the user (e.g. by applying equal intervals for the reduced interval).

3.2 Value differences between neighboured polygons

For preserving or even enhancing significant value differences between neighbored polygons (named x_l and x_r in the following), a class break between those polygons is desired (Fig. 5).

The first step in the new algorithm identifies all polygon neighbors. This step is similar to the one that has been described above for identifying local extreme values. At this point, a threshold can be applied that defines a minimum value difference.

For all candidates, the corresponding value intervals (bounded by the values of the neighboring polygons as derived in the previous step) are plotted and sorted according to interval widths, from largest to smallest (Fig. 6). Again, a line sweep algorithm is carried out in order to retrieve optimal intersections; a vertical line is swept from global minimum to global maximum value in order to retrieve intersection counts. The first class break is set where the value has the maximum number of intersections. Afterwards those segments are excluded that intersect the first class break and another sweep is performed. This procedure is repeated until the total number of classes has been met.

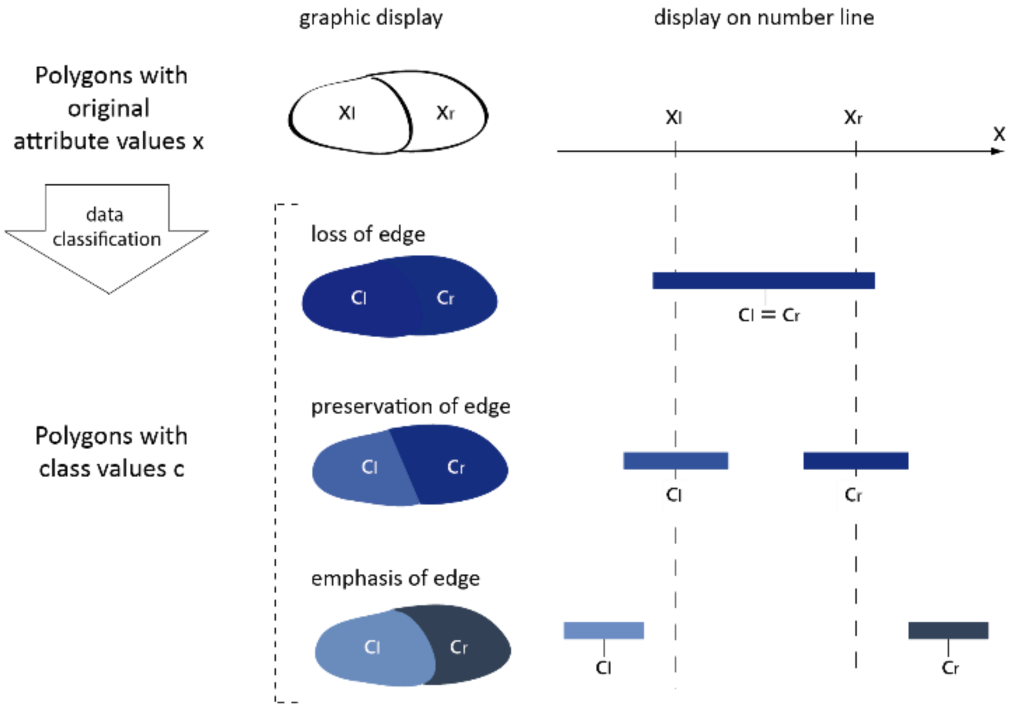


Fig. 5: Preservation or emphasis of “edges” between polygons with value differences (here: $x_r > x_l$) in graphic display and by setting class breaks along the number line

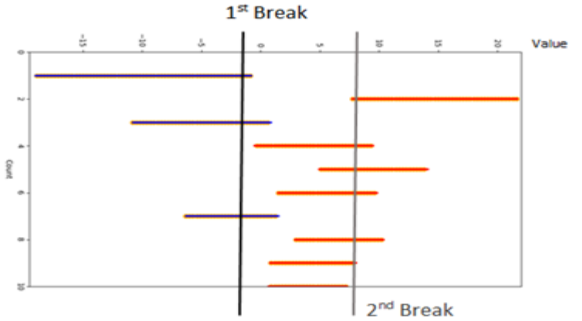


Fig. 6: Line sweep approach to determine class breaks x_l (1st break) and x_r (2nd break)

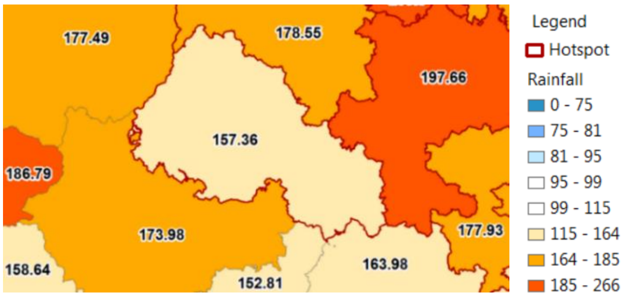


Fig. 7: Hot spot candidate which has a lower value ("157.36") compared to some of its surrounding neighbors

3.3 Spatial clusters

For the definition of spatial clusters, the Density Based Spatial Clustering of Applications with Noise (DBSCAN) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) are used as default methods. Based on the allocation of polygons to clusters, those neighbored polygon pairs are determined that belong to different clusters (and which shall be ideally divided with a class break). The setting of class breaks based on a sweep line algorithm is similar to the procedure as described before.

3.4 Hot and cold spots

For the determination of hot and cold spots the commonly used method proposed by *Getis & Ord* is applied (Getis & Ord, 1992). Applying thresholds (e.g., a z-score of > 2.58 and a p-value of < 0.01), a binary classification into hot and cold spot polygons and other polygons is possible. In a default setting, a fixed distance band is used for defining the neighborhood.

Based on the allocation of polygons to hot/cold spot clusters, those neighbored polygon pairs are determined that do not belong to one and the same hot/cold spot (and which shall be ideally divided with a class break).

For this pattern, the neighborhood identification is more complicated: By definition, a hot spot might also have relatively higher values compared to its neighbors (cold spots might have lower values). Fig. 7 gives an example – the hot spot polygon (showing the value "157.36") is surrounded by a couple of non-hot spot polygons with larger values. Hence, the algorithm takes both the closest positive and negative differences to non hot/cold spot polygons into account and stores the respective absolute values. Based on this, the aformentioned sweep line algorithm is performed in order to set the class breaks.

4. Implementation

The aforementioned methods have been embedded as a plug-in tool into the open source *QGIS* software. Several open source python modules such as *GDAL* (Geospatial Data Abstraction Library), *PySAL* (Python Spatial Analysis Library), *Fiona*, *Shapely* and *RTree* were used (Fig. 8). The user has to install these libraries before usage. A detailed user manual and its installation guide have been included. The plugin tool together with a detailed documentation has been published in the *GitLab* repository https://gitlab.com/g2lab/aChor

The usage of the tool is controlled by a graphical user interface (Fig. 9). First, the user selects the data set together with the attribute field under consideration. Currently, only polygon shape files with the *EPSG 3857* projection are allowed as input data. After this, one of the classification methods is selected. Further input variables are:

- In the case of preserving local extreme value polygons one can also handle minima and maxima separately, if desired.
- If the hot spot method is selected, the fixed distance band option is enabled with an automatically calculated distance setting that ensures that every element has at least one neighbor.
- With the next item the sweep interval is set. The corresponding default value is calculated as a function of the actual data range.
- Furthermore, the user has to define the number of classes, the default is set to 10 classes.
- Lastly, the user defines the type of color ramp that is used for final visualization.

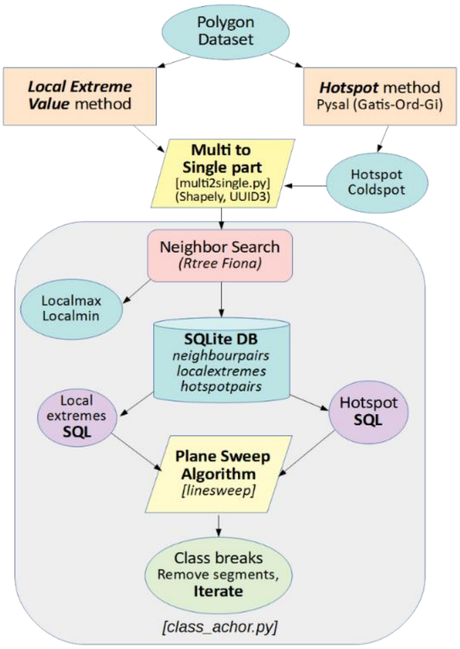
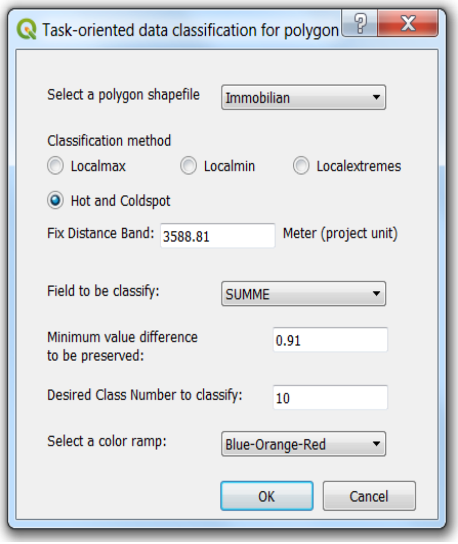

Polygon Dataset
Local Extreme Value method
Hotspot method Pysal (Gatis-Ord-Gi)
Multi to Single part [multi2single.py] (Shapely, UUID3)
Hotspot Coldspot
Neighbor Search (Rtree Fiona)
Localmax Localmin
SQLite DB neighbourpairs localextremes hotspotpairs
Local extremes SQL
Hotspot SQL
Plane Sweep Algorithm [linesweep]
Class breaks Remove segments, Iterate
[class_achor.py]

Fig. 8: Implementation diagram of the *QGIS* plug-in


Task-oriented data classification for polygon

Select a polygon shapefile: Immobilian

Classification method
○ Localmax ○ Localmin ○ Localextremes
● Hot and Coldspot

Fix Distance Band: 3588.81 Meter (project unit)

Field to be classify: SUMME

Minimum value difference to be preserved: 0.91

Desired Class Number to classify: 10

Select a color ramp: Blue-Orange-Red

Fig. 9: *QGIS* plugin user interface (older version)

5. Conclusions and Outlook

The problem of common methods for data classification in the course of designing choropleth maps - the disregard of spatial patterns - could be solved by the developing and testing new specific algorithms. However, a couple of aspects should be mentioned that lead to future research work.

First, it has to be noted that there is no algorithm that preserves all possible pattern at the same time. In fact, to some extent the pattern specific algorithms sometimes work counter-productively to one another. This application specific behavior of algorithms should be subject to further research.

Second, the algorithms presented cannot always lead to a perfect solution (i. e., preservation of every single pattern in the map). In many cases this is even not theoretically possible due to the limited number of pre-defined classes.

Third, it is necessary that the user specifies the pattern to be preserved (e.g. hot spots) so that the corresponding method can be selected. This process requires some knowledge and preliminary work. In particular, for laymen users an upstream analysis of geostatistical measures might help in a (semi-)automatic procedure.

Fourth, only uni-variate and static data sets have been used as input for spatial pattern detection. These methods have to be extended for detecting patterns in multi-variate or multi-temporal data sets. For this purpose, possible multi-dimensional patterns have to be defined. For example, there might be the task to find hot spots not only for a given date but for its existence for a certain, pre-defined time period.

For the perception of spatial patterns it is of course also mandatory that the user can correctly grasps the graphic coding of the class values. In most cases of the presented algorithms, the obtained data classification leads to intervals showing non-linear value differences that have to be ”translated” into corresponding color differences for the actual map visualization. At this point, the research question arises whether the transformation of non-linear value differences into non-linear color differences leads to a better, i.e. correct and faster perception by users. In literature, there is common understanding that color distances as perceived by humans) should expressed by the ΔE00 measure. For a reliable and fast distinction of two colors, a value of ΔE00 = 10 is recommended. The maximum range for black vs. white this is ΔE00 = 100, but for a sequential color scheme based on one color (here: blue) it is only ΔE00= 63.6 (from white = 0,0,0 to blue = 0,0,255). Avoiding pure white, we end up with a maximum range of ΔE00 = 50...55. This corresponds very well to Brewer color scales. Dividing this value by the number of classes one can derive that from a theoretical point of view a non-linear color scheme is only feasible for a very small number of classes (less or equal four). In other words, from a theoretical point of view there is very little potential to adapt color differences to non-linear class value differences.

At a later point in time, other influences (such as neighboring polygons of very different sizes or spatial separating topographical elements between polygons) on the visual interpretation will also be taken into account.

Despite of the above mentioned restrictions the presented algorithms are considered significant progress towards the improvement of the usability of choropleth maps by means of data classification in the course of performing synoptic tasks.

6. References

Andrienko, G., Andrienko, N. & Savinov, A. (2001): Choropleth maps: Classification revisited. Proceedings International Cartographic Conference, 9 S.

Armstrong, M.P., Xiao, N. & Bennett, D.A. (2003): Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps. Annals of the Association of American Geographers, 91(3): 595-623.

Brewer, C.A. & Pickle, L. (2002): Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. Annals of the Association of American Geographers, 92(4): 662-681.

Cromley, R.G. (1996): A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. International Journal of Geographical Information Systems, 10(4) 405-424.

Cromley, E.K. & Cromley, R.G. (1996): An analysis of alternative classification scheme for medical atlas mapping. European Journal of Cancer, 32A(9) 1551-1559.

Coulsen, M.R.C. (1987): In the matter of class intervals for choropleth maps: With particular reference to the work of George Jenks. Cartographica, 24(2): 16-39.

Getis, A. and Ord, J.K. (1992): The Analysis of Spatial Association by Use of Distance Statistics, *Geographical Analysis*, 24: 189-206.

Goldsberry, K. & Battersby, S. (2009): Issues of Change Detection in Animated Choropleth Maps. Cartographica, 44(3): 201-215.

Jenks, G. & Caspall, F. (1971): Error on choroplethic maps: definition, measurement, reduction. Annals of the Association of American Geographers, 61: 217-24.

MacDougall, E.B. (1992): Exploratory analysis, dynamic statistical visualization, and geographic information systems. Cartography and Geographic Information Systems, 19(4): 237-246.

MacEachren, A.M. (1994): Some truth with maps: A primer on symbolization and design. Association of American Geographers, Washington, DC.

Mersey, J.E. (1990) Choropleth Map Design – a Map User Study. Cartographica, 27(3): 33-50.

Monmonier, M. (1972): Contiguity-biased class-interval selection and location allocation models. Geographical Review, 62: 203-228.

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.

Smith, R.M. (1986): Comparing traditional methods for selecting class intervals on choropleth maps. The Professional Geographer, 38(1): 62-67.

Acknowledgements

The work presented here was carried out as part of the project “Task-oriented data classification and design of choropleth maps (aChor)”, funded by the German Research Association (DFG). Individual passages from the project report have been included in this article.

The following data sets have been used:

Figures 1 and 2:
http://suche.transparenz.hamburg.de/dataset/sozialmonitoring-integrierte-stadtteilentwicklung-bericht-2016-anhang

Figure 3:
https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Eheschliessungen-Ehescheidungen-Lebenspartnerschaften/Publikationen/Downloads-Eheschliessungen/eheschliessungen-geborene-gestorbene-5126001187004.html

Figure 7:
https://www.dwd.de/DE/klimaumwelt/cdc/cdc_node.html/

https://doi.org/10.5194/ica-proc-4-95-2021 |