Resiliency: A Consensus Data Binning Method

Arpit Narechania
Georgia Institute of Technology, Atlanta, GA, USA

Alex Endert
Georgia Institute of Technology, Atlanta, GA, USA

Clio Andris
Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Data binning, or data classification, involves grouping quantitative data points into bins (or classes) to represent spatial patterns and show variation in choropleth maps. There are many methods for binning data (e.g., natural breaks, quantile) that may make the same data appear very different on a map. Some of these methods may be more or less appropriate for certain types of data distributions and map purposes. Thus, when designing a map, novice users may be overwhelmed by the number of choices for binning methods and experts may find comparing results from different binning methods challenging. We present **resiliency**, a new data binning method that assigns areal units to their most agreed-upon, consensus bin as it persists across multiple chosen binning methods. We show how this "smart average" can effectively communicate spatial patterns that are agreed-upon across binning methods. We also measure the variety of bins a single areal unit can be placed in under different binning methods showing fuzziness and uncertainty on a map. We implement resiliency and other binning methods via an open-source JavaScript library, **BinGuru**.

2012 ACM Subject Classification Human-centered computing → Geographic visualization

Keywords and phrases data binning, data classification, choropleth maps, geovisualization, geographic information systems, geographic information science, cartography

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.55

Category Short Paper

Supplementary Material
Software (Source Code): `https://github.com/arpitnarechania/binguru`
InteractiveResource (Observable Notebook): `https://observablehq.com/@arpitnarechania/binguru-demo`

1 Introduction

Data binning (or classification) is the process of grouping quantitative data values into bins (or classes), that are then represented by different colors, shades, textures, or sharpness to show spatial patterns or variations in choropleth maps [4]. A classic example of a choropleth map may be a country's states shaded light or dark according to their low or high population, respectively. To create such maps, the population values may be placed into groups such as *high*, *medium* and *low population* using a binning method that assigns each state to a group. Using one binning method, a single state may be classified as *high population*, but using another binning method, it may be classified as *low population*. This affects the reader's interpretation of the state and how resources may be allocated to the state.

Many binning methods exist [1, 9, 3, 14] and are built into popular GIS tools and libraries [5, 15, 12, 18]. They have strengths and weaknesses that make them (un)suitable for certain types of data distributions and map purposes. For example, *standard deviation* emphasizes normality and regions of high and low deviation from the mean; *quantile* evenly distributes data values into bins irrespective of the data distribution, highlighting regional

differences; *pretty breaks* rounds off bin extents, making them visually appealing and easy-to-interpret; and *natural breaks* can capture organic data groupings and reveal outliers. Choosing an appropriate binning method is important to ensure that the map effectively represents the data and communicates information to the reader [11]. However, this determination can be overwhelming for users, particularly those who are not well-versed in statistical or cartographic concepts. Even experts may find it challenging to compare and contrast results of different binning methods for data-driven decision making.

In this paper, we present **resiliency**, a new data binning method that assigns areal units to their most agreed-upon, consensus bin as it persists across multiple methods. For example, if a county is placed into bin #2 (i.e., second to lowest group of values) across a majority of binning methods, *resiliency* will attempt to place it in that bin. We also show how *resiliency* allows for spatial patterns to be communicated with(out) outliers and how it promotes reflection on the fuzziness that binning imposes during mapping. Note that *resiliency* is neither a panacea nor a prescriptive measure, but provides more insight into a dataset. It detects which areal units are likely to "hop" across bins upon switching to a different binning approach. As a consensus method, *resiliency* can help the user be more confident in choosing a familiar binning method (e.g., natural breaks) if its result resembles resiliency. We share *resiliency* and 17 other established binning methods via **BinGuru**, an open-source JavaScript library for developers to create custom geospatial applications, offering more variety than existing GIS tools[1] and libraries[2].

2 Related Work

While *continuous* or *unclassed* maps are valuable for maintaining exact numeric data relationships to the visual variable [19], it is more common to group areal units into bins to reveal patterns across geographic space. In terms of methods, *equal interval*, *natural breaks*, *standard deviation*, *quantiles*, and *pretty breaks* are particularly common [3]. Genetic algorithms [1] and proximity-based [9] binning methods, which promote spatially compact and homogeneous regionalization on maps, have also been explored but are not widely used in practice. OSCAR is a human-centered binning method that leverages usage information from visualization dashboards to suggest common bin sizes for an attribute [14].

Determining an appropriate binning method often depends on several factors such as the data distribution (e.g., *standard deviation* for normally distributed data), tacit and domain specific knowledge (e.g., *manual interval* or *diverging bins* centered around a certain meaningful baseline value) [16], or a desire to have the same number of data points in each group (e.g., *quantiles*). According to Brewer and Pickle [3], *quantile* and *minimum boundary error* are best suited for general reading of epidemiological rate maps followed by *natural breaks* and a hybrid version of *equal interval*. According to Smith [17], *quartile*, *equal interval*, *standard deviation*, and *natural breaks* are accurate for data sets with specific distributional characteristics, but none of them accurately bin all types of distributions. Prior work has also explored diverse approaches and measures for assessing map complexity, emphasizing their impact on cognitive load, readability, and visual effectiveness [7, 2]. For example, Monmonier [10] found that round-number bin breaks, which are easy to read and remember, can constrain the outputs of optimization algorithms that have more significant digits than the map user would prefer or that the precision of the data warrants.

[1] ArcGIS [5] (9 binning methods), QGIS [15] (6)
[2] ArcGIS Maps SDK [6] (6 binning methods), Python's PySAL [12] (10), and R's tmap [18] (9)

Algorithm 1 Resiliency.

```
1  input  : data values V, binning methods M, binning options O
2  output : resiliency bin breaks RB
3  // Compute bin breaks for all M
4  bin breaks B ← { }
5  for method m in M do
6  |  B[m] = ComputeBins(V, O, m)
7  // Determine bins for all V across all M
8  bin ids ID ← { }
9  for value v in V do
10 |  for method m in M do
11 |  |  ID[v][m] = AssignBin(v, B[m])
12 // Compute the frequency of each bin among all M
13 bin frequencies BF ← { }
14 for value v in V do
15 |  BF[v] = ComputeBinFrequency(ID[v])
16 // Place values in their most frequent bins
17 most frequent bins MFB ← { }
18 for value v in V do
19 |  MFB[v] = ComputeMostFrequentBin(BF[v])
20 // Compute Resiliency
21 resiliency bin breaks RB ← { }
22 working bin assignments WFB ← MFB
23 while ValidateBins(RB) do
24 |  RB, WFB = ResolveConflicts(WFB, RB)
25 return RB
```

3 Resiliency

Given the diversity and complexity of established binning methods, we propose a new method, *resiliency*, that assigns areal units to their most agreed-upon, consensus bin across multiple methods. Algorithm 1 illustrates the pseudo code for this method.

First, we compute the bin breaks for multiple *comparable*[3] binning methods (Lines 3 - 6). For each areal unit, we track the ID (or index) of the bin (*binID*) that it was assigned to across the binning methods (Lines 7 - 11). Next, we compute the frequency of the assigned *binID*s (Lines 12 - 15), i.e., the number of times it is placed across different *binID*s. For each areal unit, we then compute the frequency of the most frequently assigned *binID*s (Lines 16 - 19). The output is the *resiliency* **bin count** (number of bins), *interval* (high and low bounding values), and **size** (number of data points in each bin). The output at the data point (areal unit) level is its assigned bin and the number of times it has fallen into this bin (and also other bins). Next, we place each areal unit in its most frequent *binID*, and subsequently detect and resolve conflicts (Lines 20 - 24).

[3] Binning methods are considered comparable if they have the same resultant (or specified) bin count, e.g., if the specified bin count is five, then we can compare *equal interval*, *quantile*, *maximum breaks*, *natural breaks*, *ck-means*, and *geometric interval*. If the desired bin count is six, then *box-plot* and *percentile* may also be considered (as they generally output six bins). Other methods may also be considered on a case-by-case basis, e.g., *defined interval* if the specified bin interval results in the desired bin count.

We note three possible conflicts. First, if there is a tie for the most frequent bin assignment, we break the tie on a first-come-first-serve basis and use the smaller *binID*. Next, resultant bin extents could overlap, e.g., $binID = 1$'s maximum extent is 50 and $binID = 2$'s minimum extent is 45 (smaller instead of greater); in response, we skip the most frequent *binID* assignment and iterate on the next (i.e., second) most frequently assigned bin and so on. Third, some bins (e.g., the middle bin) may have no areal units; one could output fewer bins or equally split bins until the desired bin count is reached; *resiliency* supports both modes.

3.1 Usage Scenario

Imagine Kiran is an Indian government official who wants to map the "Total Fertility Rate" (number of children per women) across India to educate the general public and inform future birth-related policies. The data contains 28 states and 7 union territories, and the total fertility rate ranges from 1.1 (Sikkim) to 3.0 (Bihar) with an average of 1.8 children per woman. They are not sure which binning method to use.

Kiran uploads their dataset to an interactive notebook we developed powered by the *BinGuru* JavaScript library. They choose *six* bins and a divergent *purple to orange* color scheme, and inspect the output of eight (out of 18 supported) binning methods: *equal interval*, *geometric interval*, *quantile*, *maximum breaks*, *percentile*, *box plot*, *standard deviation*, and *natural breaks* (Figure 1**A-H**). They find that states of Bihar and Meghalaya (north east) have the largest fertility rates with *equal interval*, *geometric interval*, *maximum breaks*, and *standard deviation*. However, when using the *quantile*, *box plot*, and *natural breaks* methods, Uttar Pradesh (west of Bihar) is also placed in the same bin. *Maximum breaks* groups most

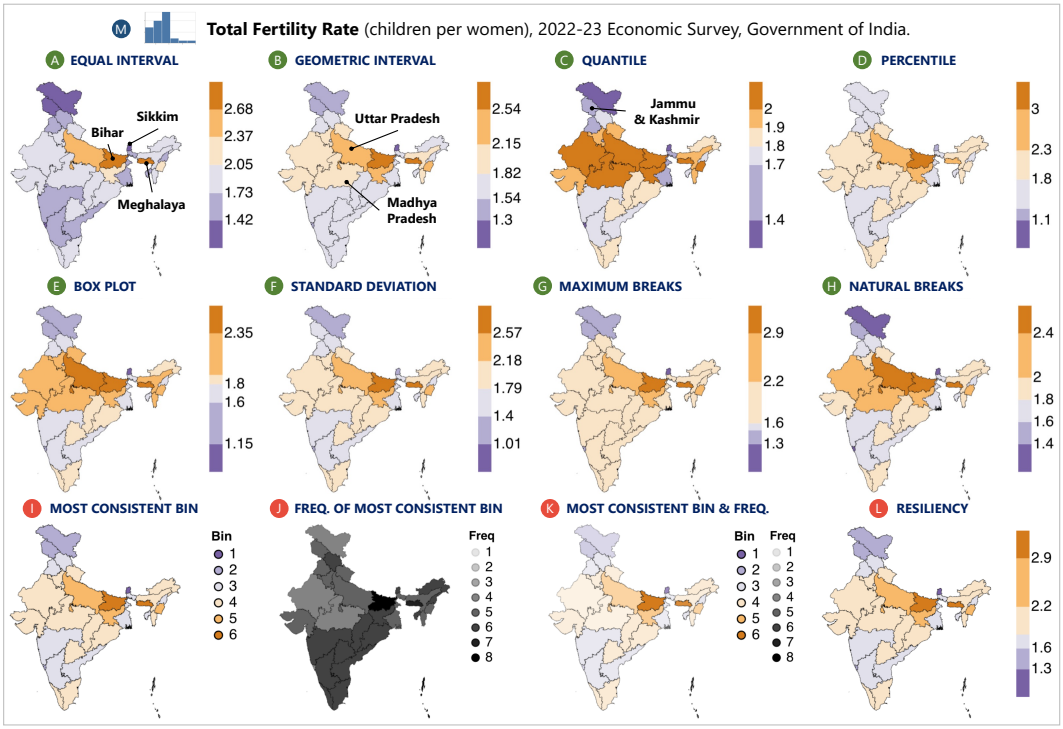

Figure 1 Small multiples of choropleth maps showing "Total Fertility Rate (children per women)" (**M**) in India [8] using established binning methods (**A-H**) and *resiliency* (**I-L**).

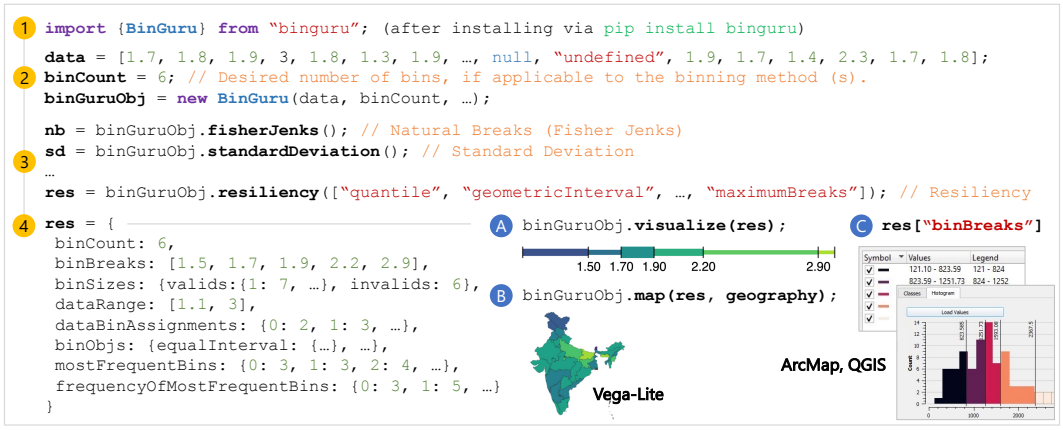

Figure 2 Usage scenario demonstrating how developers can (1) import the **"binguru"** library, (2) initialize a **BinGuru** class instance with the input data and binning parameters (e.g., **binCount** – desired number of bins), (3) explore different binning methods (e.g., **.fisherJenks()**), and (4) inspect their output comprising resultant *binBreaks* (bin boundaries), *binSizes* (number of points in each bin), *dataBinAssignments* (binID corresponding to each point), *binObjs* (applicable for *resiliency*, with intermediate binning outputs of constituent binning methods), *mostFrequentBins*, and *frequencyOfMostFrequentBins*. Developers can also visualize the output on (A) a legend, (B) a map – if underlying geography is available, and/or (C) copy-paste into commercial GIS software.

of western, central, and southern regions in the same bin ($binID = 4$). *Quantile* shows a more even distribution but does not distinguish extreme values. Kiran is uncertain which method to use and experiments with *resiliency*, visualizing the results using three maps (Figure 1**I-L**):

Most Consistent Bin. This map shows the most *frequent* bin across binning methods for each areal unit (Figure 1**I**). For example, Bihar (*Total Fertility Rate* equals 3.0) is colored dark orange, implying it is most frequently placed in $binID = 6$–with a high fertility rate.

Frequency of Most Consistent Bin. This map shows the *frequency of an areal unit's most frequent (or consistent) bin* assignment across binning methods (Figure 1**J**). Higher numbers mean that the areal unit consistently fell into the same bin. For example, Bihar is colored the darkest shade of black ($frequency = 8$), implying it is consistently placed in the same bin ($binID = 6$); whereas, Madhya Pradesh (center) is colored much lighter, implying it is inconsistent across bins. Kiran understands that their binning decisions will affect how Madhya Pradesh is classified and will discuss this "fuzziness" at future meetings.

Most Consistent Bin and its Frequency. This bivariate map combines the previous two maps into a value-by-alpha map [13]. The hue corresponds to the *most frequent bin* and the opacity corresponds to the *frequency of the most frequent bin* (Figure 1**K**), where higher opacity implies higher frequency. Here, Bihar is an opaque orange color, as it consistently falls in $binID = 6$. More transparency indicates low frequency, less certainty, and inconsistency.

Resiliency. This map (Figure 1**L**) often (but not always) resembles Figure 1**I**, but now includes a legend that reflects the actual data values as the bin breaks. Kiran observes that *Resiliency* retained the regions of Bihar and Meghalaya as the regions with the largest and Sikkim with the smallest (outlier) values, while also showing variance among other northern, southern, and western states. They also note that this result resembles the *standard deviation*

method (Figure 1**F**) except for the bin assignment of the regions of Jammu & Kashmir (north) and Sikkim (north east). They now decide to either use the output of *resiliency* as-is or use *standard deviation*, which they value as a more familiar, easy-to-understand method.

4 Implementation, Future Work and Conclusion

Resiliency and 17 other binning methods are available through an open-source JavaScript library, **BinGuru** (Figure 2). We next plan to make *resiliency* more robust with weighting (e.g., *equal interval* has more weight). We then plan to better guide users by recognizing the distribution of their data and suggesting binning methods that are appropriate for that distribution. We also hope to capture how cartographers and GIS experts might use *resiliency* to learn about its benefits and drawbacks, ease of use, and uptake to drive future iterations.

In conclusion, we presented *resiliency*, a new data binning method that assigns areal units to their most agreed-upon, consensus bin as it persists across multiple binning methods. We showed how *resiliency* can enable spatial patterns to be communicated with(out) outliers and promote reflection on the fuzziness often imposed during binning.

References

1 Marc Armstrong, Ningchuan Xiao, and David Bennett. Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps. *Annals of the Association of American Geographers*, 93:595–623, September 2003.

2 Arnold Bregt and Marco CS Wopereis. Comparison of complexity measures for choropleth maps. *The Cartographic Journal*, 27(2):85–91, 1990.

3 Cynthia A. Brewer and Linda Pickle. Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals of the Association of American Geographers*, 92(4):662–681, 2002.

4 Michael John De Smith, Michael F Goodchild, and Paul Longley. *Geospatial Analysis: A Comprehensive Guide To Principles, Techniques And Software Tools*. Troubador Publishing Ltd., 2007.

5 ESRI. ArcGIS, 2023. URL: `https://www.esri.com/en-us/arcgis/about-arcgis/overview`.

6 ESRI. ArcGIS Maps SDK, 2023. URL: `https://developers.arcgis.com/javascript/latest/api-reference/`.

7 Alan M MacEachren. Map complexity: Comparison and measurement. *The American Cartographer*, 9(1):31–46, 1982.

8 Ministry of Finance, Government of India. Economic Survey of India, 2023. URL: `https://www.indiabudget.gov.in/economicsurvey/doc/Statistical-Appendix-in-English.pdf`.

9 Mark Monmonier. Maximum-Difference Barriers: An Alternative Numerical Regionalization Method. *Geographical Analysis*, 5(3):245–261, 1973.

10 Mark Monmonier. Flat laxity, optimization, and rounding in the selection of class intervals. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 19(1):16–27, 1982.

11 Mark Monmonier. *How To Lie With Maps*. University of Chicago Press, 2018.

12 Serge Rey and Luc Anselin. PySAL, 2005. URL: `https://pysal.org/`.

13 Robert E Roth, Andrew W Woodruff, and Zachary F Johnson. Value-By-Alpha Maps: An Alternative Technique To The Cartogram. *The Cartographic Journal*, 47(2):130–140, 2010.

14 Vidya Setlur, Michael Correll, and Sarah Battersby. Oscar: A semantic-based data binning approach. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 100–104, Los Alamitos, CA, USA, October 2022. IEEE Computer Society.

15 Gary Sherman. QGIS, 2002. URL: `https://qgis.org/`.

16 Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. *Thematic Cartography And Geovisualization*. CRC Press, 2022.

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

18 tmap. Tmap, 2023. URL: `https://cran.r-project.org/web/packages/tmap`.

19 Waldo R Tobler. Choropleth Maps Without Class Intervals? *Geographical Analysis*, 1973.