Cartograms, hexograms and regular grids: minimising misrepresentation in spatial data visualisations

Samuel Langton¹ and Reka Solymosi²

¹Department of Sociology, Manchester Metropolitan University

²Centre for Criminology and Criminal Justice, University of Manchester

Corresponding author: Samuel Langton (samuel.langton@stu.mmu.ac.uk)

Funding: Samuel Langton's research contribution was completed under a Vice Chancellor's doctoral scholarship at Manchester Metropolitan University.

Introduction

Thematic maps are powerful, accessible and aesthetically appealing visualisations widely applied to represent spatial data (Barrozo et al., 2016). In urban analytics, spatial data visualisation is important to effectively communicate and engage with stakeholders (Billger, Thuvander and Wästberg, 2017) and can even serve to analyse geographical information (Rae, 2011). However, irregularly shaped polygons and large differences in the sizes of areas being mapped can introduce misrepresentation. The message researchers want to get across might be lost, or misunderstood by readers. To address this issue, methods have been developed to distort the shape and size of areas, either by turning irregular polygons (such as neighbourhoods) into regular or hexagonal grids (Bailey, 2018), or by using cartograms, where the distortions of size and shape are made explicit and communicate meaning (Dorling, 1996; Tobler, 2004). However, it is unclear how these different transformations can impact on viewers' interpretation of the map. Using a crowdsourced survey, we explore the extent to which alternative methods of visualising spatial data can improve communication of an intended message by testing people's understating of maps transformed using four different methods. We hope that these findings highlight the issue of misrepresentation in spatial data for the urban analytics community, but more specifically, we aim to provide some guidance as to which methods might be the most appropriate.

Thematic maps have various issues (see Dorling, 1996), however we address a specific problem common to traditional area-based choropleth maps, whereby variation in the size and shape of areas being visualised may affect map legibility (Stigmar and Harrie, 2011). In extreme cases, larger areas come to dominate the map and render smaller regions almost invisible. Census data in England and Wales, for instance, is published at spatial scales designed to be uniform by population (e.g. Lower Super Output Area). Consequently, sparsely populated areas dominate visualisations at the expense of those that are densely populated. In such cases, even the most well-intentioned researcher, using geographically accurate spatial data, may introduce a degree of misrepresentation in their visualisations or fail to communicate their message to readers as intended.

To date, a popular method for overcoming these obstacles has been the cartogram. Although there are numerous methods of operationalising cartograms (Dougenik, Chrisman and Niemeyer, 1985) the underlying premise is that areas are rescaled according to a variable (Nusrat and Kobourov, 2016). By rescaling areas by some uniform variable (such as population in the example of Lower Super Output Areas) an effort is made to minimise the misrepresentation that can be introduced by using raw area boundaries. Larger areas become smaller, and less dominant, and 'invisible' areas are expanded to become more visible. That said, this approach has come under some criticism for alleviating misrepresentation through invisibility at the expense of introducing misrepresentation through distortion (Harris et al., 2017a). Even well-specified scaling variables can cause alterations which result in some polygons appearing as lines, for instance (Coltekin, 2015).

A recent development is the 'balanced area' cartogram, which aims to minimise the distorting side-effects of cartograms (see Harris et al., 2017a; 2017b). The balance is achieved by predefining an

‘interpretability threshold’ which is the smallest legible unit size given the dimensions of the final published map. In producing the cartogram, any areas that fall below this areal threshold are ‘protected’ from the rescale, and instead are set as the minimum unit size. Harris and his colleagues demonstrated the benefits of this approach using Local Authority data on residential geography in England. The degree of error, defined as the percentage of non-overlap between the original map and the cartogram, was minimised with the balanced cartogram compared to a solely attribute-scaled (e.g. population) cartogram (Harris, 2017).

This approach has also been extended to include a ‘hexogram’, whereby an iterative binning algorithm assigns the centroid of polygons from the balanced cartogram to tessellated hexagons, each representing the original polygons. In doing so, the data is said to maintain spatial accuracy whilst also being uniform in shape and size (Harris, Charlton and Brunsdon, 2018a; 2018b). Comparable alternatives to this approach are tile maps which use a distance-based procedure (e.g. Hungarian algorithm) to assign original polygons to a grid of uniform shapes, such as a hexagons or squares, in a manner that minimises the distance between the original and the new synthetic boundaries (Bailey, 2018). In doing so, tile maps generate an aesthetically appealing contiguous grid of polygons which can introduce topological inaccuracies, such as previously separated polygons becoming neighbours. The hexogram prioritises the maintenance of the original topological links but is not contiguous. In each case, the stylised map retains the same number of observations as the original map, but the boundaries have been transformed into something more uniform and less distracting, which may be better suited for conveying the message of the researcher.

That said, little is known about how different methods of visualising spatial data impact on people’s interpretation of the information presented. This study aims to rectify this shortcoming through the use of a crowdsourced online survey questionnaire designed to measure the extent to which various alternatives to a traditional thematic map can more accurately convey geographic information. We begin by providing an outline of the survey design and methodology, followed by the reporting of results, and conclude with a discussion on our findings and suggestions for future research.

Survey design

Studies have made some attempt to gauge how people interpret different visualisations of the same data to draw conclusions (e.g. Borgo et al., 2012; Borkin et al., 2016; Skau and Kosara, 2016). Specific to maps, Coltekin *et al* (2015) asked respondents to complete a series of tasks using various different tools available in Google Maps (e.g. 2D default map, 3D satellite images, Street View) and found that the degree of accuracy with which people completed questions varied by the tool used. That said, “visualization researchers have been increasingly leveraging crowdsourcing approaches to overcome a number of limitations of controlled laboratory experiments, including small participant sample sizes and narrow demographic backgrounds of study participants” (Borgo et al., 2018: 573).

Here, we use a crowdsourced survey to assess the ability of different thematic mapping

techniques to visualise and communicate a situation where high values spatially cluster in small areas. Descriptive maps can play an important role in identifying and understanding spatial clusters in urban analytics, despite continued advances in more complex statistical methods (e.g. Jones et al., 2018). We used electoral result data from the 2016 European Union (EU) referendum at Local Authority level in England to create a map considered to be a good example of high value clustering which is obscured by significant differences in area sizes. Areas with a high proportion of Remain votes are concentrated in Greater London (Hobolt, 2016), which has geographically small Local Authorities compared to the rest of the country. On a traditional thematic map, using original boundaries as defined by the Office of National Statistics, strongly Leave areas dominate the visual at the expense of densely populated Remain areas, which became almost ‘invisible’ (see *Figure 1*).

Figure 1: proportion of Remain votes in 2016 EU referendum by Local Authority area in England using original boundaries.

Alternatives to this original map were then generated using four different techniques for transforming the Local Authority area polygons. Balanced area-based cartograms and hexograms were created in R (version 3.5.1) using the default minimum threshold options (see Harris, 2017). Uniform hexagonal and square tile grids were generated using the geogrid R package using the default options regarding the optimisation of cell sizes (Bailey, 2018). A decision was made to create the uniform grids from the balanced cartogram rather than from the original boundaries to produce a more optimal outcome and reduce computation time. A result of this was that the outputted boundaries were not completely contiguous, contrary to what was produced using the original boundaries. In total, five

visualisations were created: the original (see *Figure 1*), balanced cartogram, hexogram, hexagonal grid and square grid (see *Figure 2*). Polygons were shaded according to the percentage of Remain voters in each Local Authority.

These maps were then collated in a survey, and for each map, participants were asked to rate the extent of their agreement with a statement considered to be a true and accurate description of the data: “High values (in yellow) appear to be clustered near one another, with a handful of outliers elsewhere in the country”. Respondents viewed the five maps in isolation and reported their agreement with the statement using a 5-point Likert scale (Strongly agree, slightly agree, neither agree nor disagree, slightly disagree, strongly disagree). There was an additional option for ‘Don’t know’. If respondents agreed with the statement, the map was considered to be representing the data accurately and conveying the message of high value clustering as intended.

Figure 2: proportion of Remain votes in 2016 EU referendum by Local Authority area in England using (a) balanced cartogram, (b) hexogram, (c) square grid, (d) hexagonal grid.

The survey itself did not include any reference the referendum or the method used to generate each map. Options were shown in the same order for every respondent (starting from the balanced cartogram, hexogram, hexagonal grid, original and finally the square grid). In the interests of keeping the survey as short as possible, and to minimise dropout, no demographic information (e.g. gender, education) was asked of the respondents. The survey was created using Google Forms and distributed via social media platforms Twitter and Reddit (using the r/samplesize subreddit).

Results

The survey was deployed for three days, generating a sample size of 768 respondents, all of which had completed the questionnaire in its entirety. Overall, the map with the highest proportion of respondents agreeing with the statement was the balanced cartogram, with 94% of respondents reporting agreement (either ‘strongly agree’ or ‘slightly agree’). Only two respondents did not know and a total of 4% disagreed with the statement in some way (either ‘strongly disagree’ or ‘slightly disagree’). This was followed by the hexogram, for which 91% of respondents showed some degree of agreement, two respondents did not know and 6% either strongly or slightly disagreed with the statement. Although in general, respondents appeared to interpret the original map as intended, with 77% reporting some level of agreement with the statement, this was still lower than the balanced cartogram and hexogram. The original map also had the highest number of don’t know responses (12). The regular hexagonal and square grids had the highest level of disagreement, with only 53% and 51% either strongly or slightly agreeing with the statement respectively.

For further analysis, the 5-point Likert scale was recoded to a binary outcome variable to indicate whether respondents ‘agreed’ (strongly agreed, slightly agreed) or ‘disagreed’ (strongly disagreed, slightly disagreed) with the true statement for use in a logistic regression model. Whilst we acknowledge that the sample is not random, thus compromising generalisability to a wider population, this analysis was considered useful to gauging effect sizes. Respondents who answered “Don’t know” or “Neither agree nor disagree” were excluded from this analysis. The outcome variable (agree-disagree) was regressed on a single categorical variable indicating which map type respondents were observing when answering the Likert scale question, with the original map set as the reference category. Odds ratio estimates confirm the descriptive findings (see *Table 1*). The odds of respondents agreeing with the true statement were 3.8 and 2.6 times that of the original map for the balanced cartogram and hexogram respectively. When viewing the hexagonal or square grids, people were about 3 times less likely to agree with the statement compared to the reference category of the original map. These odds were significant at $p < 0.001$ alpha, noting the issue regarding generalisability of these results to a wider population due to self-selection in the crowdsourced sample.

Table 1: logistic regression results, response variable recorded from Likert-scale to binary agree (1) disagree (0); original map is reference category; 95% confidence intervals reported.

Map type	B	Se	Z ratio	Probability	Odds
Original (reference)	-	-	-	-	-
Balanced cartogram	1.3364	0.2047	6.529	< 0.001	3.80 (CIs: 2.57-5.76)
Hexogram	0.9478	0.1827	5.187	< 0.001	2.58 (CIs: 1.81-3.72)
Square grid	-1.3205	0.1320	-10.000	< 0.001	0.27 (CIs: 0.21-0.34)
Hexagonal grid	-1.3130	0.1314	-9.989	< 0.001	0.27 (CIs: 0.21-0.35)
Model Chi Square: 553.418 (p = < 0.001) Likelihood ratio R ² : 0.155 AIC: 3036.5					

Discussion

Overall, findings suggest that it is possible to improve communication of a message by transforming map boundaries, in particular using balanced cartogram and hexogram approaches. However, it is also possible to impair interpretation, as seen with transformations into tiled grids (square and hexagonal), where fewer people interpreted the visualisation as intended. In interpreting these results, there are a number of discussions points to consider.

Firstly, the research design assumes that the ‘true’ statement was both an accurate description of the raw data and a fair means of gauging the extent to which respondents were understanding each map. We acknowledge that the statement could have been written differently, with some respondents demonstrating lack of understanding (e.g. those that answered ‘Don’t know’) possibly as a product of the research design, rather than the visualisations themselves. Secondly, due to a lack of individual characteristic data, we were unable to control for other factors that might have impacted on how respondents interpreted the task. For instance, those familiar with cartography, or data visualisation more generally, may have been more sensitive to the issues that can arise when creating maps. The degree to which our sample was representative of a wider population, whether it be the general public or geographers, is unknown. Thirdly, although the survey contained no reference to the EU referendum, respondents familiar with England’s social geography may have drawn upon other personal information

to inform their decision. For instance, many other social phenomenon (e.g. income) are commonly known to cluster in the urban areas where Remain voters also clustered.

The findings from this survey mark a first attempt to empirically examine how people interpret different methods of visualising area-based thematic maps. As such, we highlight some key directions for future research. Firstly, as mentioned above, collecting individual characteristics on respondents could help control for confounding factors such as expertise and assess generalisability of findings. Secondly, questions could be added to test different dimensions of misrepresentation or improve objectivity. For instance, respondents could be asked to manually identify areas of high clustering by clicking or selecting on the map interactively, which could then be compared to the ‘true’ results from a Local Indicators of Spatial Autocorrelation analysis. Thirdly, there are other methods of generating thematic maps that we did not consider. It would be interesting to compare the balanced cartogram to standard cartograms scaled by a selection of variables, or alternatives such as the Dorling cartogram. Future research could consider making these advancements using synthetic data, to mitigate the impact of prior knowledge of the study area.

Software

For those interested in reproducing these methods in R for their own research, the balanced cartogram and hexogram used open source code (see Harris, 2017) which included the packages *cartogram* (Jeworutzki, 2018), *fMultivar* (Wuertz, Setz and Chalabi, 2017) and *sp* (Pebesma and Bivand, 2005). Spatial data handling and visualisations were carried out using *sf* (Pebesma, 2018) and *tidyverse* (Wickham, 2017).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Bailey, 2018. geogrid: Turn Geospatial Polygons into Regular or Hexagonal Grids. *R package version 0.1.0.1*. <https://CRAN.R-project.org/package=geogrid>
- Barrozo, L.V., Pérez-Machado, R.P., Small, C. and Cabral-Miranda, W., 2016. Changing spatial perception: dasymetric mapping to improve analysis of health outcomes in a megacity. *Journal of Maps*, 12(5), pp.1242-1247.
- Billger, M., Thuvander, L. and Wästberg, B.S., 2017. In search of visualization challenges: The development and implementation of visualization tools for supporting dialogue in urban planning processes. *Environment and Planning B: Urban Analytics and City Science*, 44(6), pp.1012-1035.
- Borgo, R., Abdul-Rahman, A., Mohamed, F., Grant, P.W., Reppa, I., Floridi, L. and Chen, M., 2012. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp.2759-2768.
- Borgo, R., Micallef, L., Bach, B., McGee, F. and Lee, B., 2018, June. Information visualization evaluation using crowdsourcing. In *Computer Graphics Forum* (Vol. 37, No. 3, pp. 573-595).
- Borkin, M.A., Bylinskii, Z., Kim, N.W., Bainbridge, C.M., Yeh, C.S., Borkin, D., Pfister, H. and Oliva, A., 2016. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1), pp.519-528.
- Coltekin, A., Lokka, I.E. and Boér, A., 2015, August. The utilization of publicly available map types by non-experts—A choice experiment. In *Proceedings of the 27th International Cartographic Conference (ICC2015), Rio de Janeiro, Brazil*(pp. 23-28).
- Dorling, D., 1996. Area cartograms: their use and creation. *Concepts and Techniques in Modern Geography (CATMOG)*.
- Dougenik, J.A., Chrisman, N.R. and Niemeyer, D.R., 1985. An algorithm to construct continuous area cartograms. *The Professional Geographer*, 37(1), pp.75-81.
- Garnier, S., 2018. viridis: Default Color Maps from 'matplotlib'. *R package version 0.5.1*. <https://CRAN.R-project.org/package=viridis>
- Harris, R., 2017. Hexograms: better maps of area based data. *Retrieved from https://rpubs.com/profrichharris/hexograms*
- Harris, R., Charlton, M., Brunsdon, C. and Manley, D., 2017a, January. Tackling the curse of cartograms: Addressing misrepresentation due to invisibility and to distortion. In *Proceedings of the 25th GIS Research UK Conference, Manchester*. Retrieved from http://huckg.is/gisruk2017/GISRUK_2017_paper_29.pdf.
- Harris, R., Charlton, M., Brunsdon, C. and Manley, D., 2017b. Balancing visibility and distortion: Remapping the results of the 2015 UK General Election. *Environment and Planning A*, 49(9), pp.1945-1947.
- Harris, R., Charlton, M. and Brunsdon, C., 2018a. Mapping the changing residential geography of White British secondary school children in England using visually balanced cartograms and hexograms. *Journal of Maps*, 14(1), pp.65-72.
- Harris, R., Charlton, M., and Brunsdon, C., 2018b. Using hexograms to map areal data. In *Proceedings of the 26th GIS Research UK Conference, Leicester*. Retrieved from

https://www122.lamp.le.ac.uk/download/GISRUK2018_Contribution_023.pdf

Hobolt, S.B., 2016. The Brexit vote: a divided nation, a divided continent. *Journal of European Public Policy*, 23(9), pp.1259-1277.

Jeworutzki, S., 2018. Cartogram: Create Cartograms with R. *R package version 0.1.1*. <https://CRAN.R-project.org/package=cartogram>

Jones, K., Manley, D., Johnston, R. and Owen, D., 2018. Modelling residential segregation as unevenness and clustering: A multilevel modelling approach incorporating spatial dependence and tackling the MAUP. *Environment and Planning B: Urban Analytics and City Science*, 45(6), pp.1122-1141.

Nusrat, S. and Kobourov, S., 2016, June. The state of the art in cartograms. In *Computer Graphics Forum* (Vol. 35, No. 3, pp. 619-642).

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, <https://journal.r-project.org/archive/2018/RJ-2018-009/>

Pebesma, E.J. and Bivand, R.S., 2005. Classes and methods for spatial data in R. *R News* 5 (2), <https://cran.r-project.org/doc/Rnews/>.

R Core Team, 2018. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. <https://www.R-project.org/>.

Rae, A., 2011. Flow-data analysis with geographical information systems: a visual approach. *Environment and Planning B: Planning and Design*, 38(5), pp.776-794.

Skau, D. and Kosara, R., 2016, June. Arcs, angles, or areas: Individual data encodings in pie and donut charts. In *Computer Graphics Forum* (Vol. 35, No. 3, pp. 121-130).

Stigmar, H. and Harrie, L., 2011. Evaluation of analytical measures of map legibility. *The Cartographic Journal*, 48(1), pp.41-53.

Tobler, W., 2004. Thirty five years of computer cartograms. *ANNALS of the Association of American Geographers*, 94(1), pp.58-73.

Wickham, H., 2017. tidyverse: Easily Install and Load the 'Tidyverse'. *R package version 1.2.1*. <https://CRAN.R-project.org/package=tidyverse>

Wuertz, D., Setz, T. and Chalabi, Y., 2017. fMultivar: Rmetrics - Analysing and Modeling Multivariate Financial Return Distributions. *R package version 3042.80*. <https://CRAN.R-project.org/package=fMultivar>