Response to Reviews

Stephanie Kobakian and Dianne Cook

Thank you for the careful reviews of our paper. Please find *our responses provided point-by-point with the comments in italics.*

Note that we have changed the title slightly to make it more general than cancer statistics, and we have added many more references in response to the review comments. This has made the paper longer by three pages. The revision has also been converted to a quarto document, which means that all numbers are automatically calculated from the analysis code in the document.

Comments to the Author:

While the contents and findings are interesting, and it is great to see more research done in the areas of visual inference and visualization of spatial areal data, all three reviewers expressed serious concerns about the writing and the lack of clarity throughout. I must say I agree: the manuscript is let down greatly by sloppiness and unpolished writing, lack of clarity in the mathematics and statistical modeling details (two reviewers pointed out potential basic mistakes in the maths), and occasional claims and statements which appear unsubstantiated.

I encourage the authors to respond to all the comments by the three reviewers in a substantive manner, and in doing so go through the manuscript with a very "fine pen" to ensure the writing and overall messaging is as clear as possible.

We are very sorry that the reviewers found the writing to be inadequate. We did not expect this! In this revision, we have tightened up the language substantially, provided additional references to support statements, and tidied the math and calculations.

Referee(s)' Comments to Author:

Referee: 1

Overall Comments:

This paper performs a user study to compare and contrast the ability of two visual displays, a choropleth map and a hexagon tile map. It uses a line up protocol to test the effectiveness of the two displays. The study showed that the hexagon tile map was significantly more effective than the choropleth map with the authors advocating for its use in the cancer atlas.

The authors do a nice job introducing each map, the benefits of both and outlined how each would be applied to communicate cancer statistics as part of the Cancer Atlas. I think that Section 3 could be tightened up in places, especially with regards to generating the lineups and the section that outlines the analysis (see below for specifics).

Overall, I see this as an important contribution not only to the visualisation community but also to the statistics community who often leave the communication of statistics and modelled outputs as an afterthought. I therefore recommend this paper be published after the authors consider my comments and revisions below, largely relating tao Section 3. I also believe the Cancer Atlas could benefit from such a visualisation and would encourage the authors to work with the authors of the Atlas to have this considered. As the acknowledgements suggest the Cancer Atlas team were involved in discussions, I suspect this would be an easy ask.

Specific Comments:

P7,L7 While the introduction focuses specifically on the maps communicating disease statistics, it would be good in the conclusion to extrapolate beyond that to gain a broader interest.

Done.

P8, L31-32. I would consider adding to the last sentence to write the following: "The results are summarized in Section 4, followed by a discussion about the broader implications for the use of this map style."

Done.

P9, L52. I'm on sure if the text in italics appears in the literature or if this is the author's view. If the author's view then I would replace "doesn't" with "does not". If it appears in the literature, please reference.

Attributed to sentiment suggested by Monmonier (2018) How to Lie with Maps

P9, L55. I would suggest replacing "better" with "adequately".

Done.

P9, L77-78. Consider the following: "Figure 1 shows the hexagon tile map, where the map is coloured from …"

Done.

P10, L97. Omit the words "To do".

Done.

P13, L167-170. Are you able to outline the smoothing procedure. This seems a little vague.

We've reduced this part to a single sentence. The exact smoothing procedures for processing the null data sets can be found in the code on the GitHub repo. It is not important enough to report in detail. The `gstat` *function produced null sets with insufficient spatial dependence, despite substantial adjusting parameters, and so further simple smoothing was done. These*

were compared with the Atlas plots, which had very complicated procedures to produce the cancer maps, that also ensured privacy of information.

P14, L178-183. The process here seems a little subjective. Can you provide some reasoning as to why these decisions were made?

Several sentences have been added explaining why these three trend models was used.

P14, L191-192. Why were the locations 1, 7, 10 and 11 not used?

Explained. It is because locations were randomly sampled.

P14, L193-195. This sentence does not make sense. Please revise.

Paragraph revised

P15, L226. Replace "will be" with "were".

Sentence removed because software is cited in the Acknowledgements.

P15, L234. y_ij = 0,1 represents whether the subject detected the data plot (1) or did not (0). Please fix.

Done.

P15, L239-240. This could be written better. Should read a "logit" link. I would consider writing the model as follows:

y_ij ~ Bernoulli(p_ij), where

logit(p_ij) = mu + tau_i + delta_j + (tau delta)_ij i=1,2. J=1,2,3.

Then discuss the random effects.

Done.

P16, L247. What is a "Figure Eight contributor ID"?:

This was used by the respondent to uniquely identify themselves and allow us to link them to the paid account we used for recruitment and payment. This reference has been replaced with a unique identifier created for each participant.

P11, Section 3.7. Could this information be captured more succinctly in a table perhaps? Line 290-293 just lists words. There are no questions relating to these 3 dotpoints. I'm also wondering why participants do not submit their responses as they go? Do they also have a chance to go back and change a response? It is difficult to determine what this process is.

Re-written. Responses were recorded on each lineup evaluation, but not saved in the Google sheet until the participant clicked Submit.

P18, L297-301. The math does not seem to work unless I'm missing something. Responses were collected from 92 participants. Five were removed which leaves 87 for evaluation. However you go on to say that Set A were evaluated by 42 participants and set B, 53 which adds up

to 95 participants. Then you go on to say this resulted in 1104 evaluations. Could you please revise and provide a more detailed outline of the evaluations you performed?

Fixed. Now generated automatically from the code in the qmd.

P22 Discussion through to Conclusions. As highlighted earlier, it would be good to extrapolate the results to other modelled outputs beyond cancer statistics to obtain broader coverage. There is considerable value in using a hexagon tile map for showing statistics and modelled output in statistical area levels across Australia or similar.

Done.

Referee: 2

Comments to the Author

SUMMARY

This paper contributes a lineup protocol study that assesses the utility of the hexagon tile map (as an alternative to the choropleth map) for communicating Australian cancer statistics. Unlike choropleth maps, hexagon tile maps represent each geographic/administrative area with the same amount of space on the map. In turn, densely populated areas of interest, which are small geographically (and thus hard to see), become more visible to the map reader. Ultimately, the study provides compelling evidence to support the use of the hexagon tile map as an alternative to the choropleth map, in the context of communicating cancer statistics in Australia.

WRITING

Please thoroughly fix the following writing issues, which appear throughout the manuscript and distract from its good content.

Done.

Inconsistent use of Australian vs. American spelling (e.g., visualise and visualize, neighbour and neighbor, colour and color, etc.). Please pick one (or check on ANZJS guidelines) and stick with it throughout the entire paper.

Australian English has now been consistently used in the writing of this article, except where it is necessary in the code to use American English.

Comma splices (e.g., lines 181, 239, 312, etc.).

Other random errors (e.g., uncapitalized start in line 347, unneeded apostrophe in line 212, missing verb in lines 337-338, etc.).

Fixed and checked.

FIGURES

Figure 1. This is a nice figure! My only suggestion is to adjust the caption to say "Thyroid cancer incidence" instead of "Thyroid incidence."

Done.

Figure 2. In the caption, do you want to say which of the twelve plots is the real data plot? Also, is it possible to include, in the supplementary materials, this same type of figure for a choropleth lineup? I am curious to see what that looks like.

The location has now been mentioned in the caption, and a sentence has been added in the text to direct readers to the supplementary materials.

Figure 3. Would it be clearer to use "Hexagon Tile" and "Choropleth" for the map display labels? In the caption, can you make a note of what "Location" refers to? (It took me a while to figure that out.)

This figure has been re-designed. Location is explained in the caption, also.

Figure 4. I think the trend legend is unnecessary, given that there is a separate panel in the plot for each trend model.

Done.

Figure 5. I think the green/orange dot legend for "Choro." And "Hex." is unnecessary.

Legend removed

Figure 6. Have you tried placing the bars next to each other instead of overlaying them? What does that look like?

We have changed this to be a mosaic plot, to focus on the proportion of times the participant chose the different certainty levels relative to both map types. This is better than the side-by-side bars which are hard to read because of different counts.

PAPER

Introduction.

A couple more sentences about the lineup protocol might be nice.

Done.

Background.

Lines 41-45. This part feels out of place, if you do not make it clear that visualizing communicable disease patterns is just one example use of a spatial data display. This is especially the case because the rest of the paper is about visualizing noncommunicable diseases (i.e., cancer).

Rephrased

Line 82. The word "perceived" here made me think about who is doing the perceiving and whether a discussion somewhere about the different audiences of spatial data displays is warranted / would be helpful. For example, who are the users of the Australian Cancer Atlas?

This is not especially about the Atlas, albeit motivated and grounded by the Atlas. We don't think it is relevant to discuss users of the Atlas. From our perspective, though, it would be ideal if the Atlas adopted this different display, and perhaps publication of this study might help to achieve this. We have modified the text in this paragraph to put the focus on the hexagon tile map vs the choropleth.

Methodology.

Lines 128-129. I am not sure this claim is legitimate: "The results should apply broadly to any other geographic area of interest." What about a choropleth map in which the geographic areas are more equally sized? (Not all choropleth maps are as extreme as the SA3 Australian one!)

Language has been modified.

Sections 3.5 - 3.7. Why are these sections, which cover data collection, located after Section 3.4, which covers the analysis? Would it make more sense to place them before Section 3.4?

Sections have been re-organised.

Section 3.7, lines 291-293. Please provide more information about the nature and structure of these three questions: plot choice, reason, and difficulty. You do not provide information until Section 4, which feels too late. Additionally, should the "certainty" question, referenced in Section 4.4, also be included here? I also wonder if Section 3.7 should be renamed, given that it is about more than just demographic data collection.

Done. The section was re-named to "Data collection". "Difficulty" was replaced by "Certainty", making terminology consistent.

Results.

Lines 361-368. This is an interesting finding!

Yes, but it is relatively weak. It's something found from other studies that people were often more confident with sub-optimal plots!

Discussion.

Please spend some time discussing limitations of the study. Please also spend some time discussing future work, related to this study.

Done.

Acknowledgment.

Please move the pointers to the supplementary and replication materials out of this section to a more appropriate place. Please also move the ethics approval out of this section to a more appropriate place.

Done.

CONCLUDING THOUGHTS

I think the conducted study is well-done, interesting, and an important contribution to the literature. However, the writing of the manuscript requires substantial revision/work. Please spend some time cleaning up the paper, so that the content can be better understood, assessed, appreciated, and utilized!

Thanks and we appreciate the careful reading.

Referee: 3

Comments to the Author

I appreciate the opportunity to review this interesting paper that describes an experimental study comparing the effectiveness of hexagon tile maps against choropleth maps using "the lineup protocol." While choropleth maps are a widely used visualization technique, newly proposed methods of displaying spatial data, such as hexagon tile maps, are not widely known to recipients of such data displays. In this context, the paper addresses the important question of the usability of this new plot type. With the exception of Kawakami et al., 2024, which analyzes a different type of hexagon maps, there are to my knowledge no findings on the effectiveness of hexagon tile maps so far.

The paper analyzes the effectiveness of different types of plots with respect to two aspects: First, whether a specific type of spatial distribution, namely disease trends that impact highly populated small areas, is detected with higher accuracy when viewed in a hexagon tile map compared to a choropleth map. Additionally, the paper aims to analyze whether recipients detect these disease trends faster in hexagon tile maps than in choropleth maps. The paper focuses on the first question and uses an innovative methodological approach. Findings from a lineup experiment using simulated data for Australia suggest that the probability of detecting this specific type of spatial distribution in hexagon tile maps is higher than in choropleth maps.

I enjoyed reading the manuscript and commend the authors for several strengths of their work, especially their research question, which covers an often neglected aspect in visualization practice, and their thoughtful experimental setup.

Thank you.

Considering these strengths, though, as I read the manuscript I found some areas in which I would have appreciated greater clarity. Starting with some major concerns, I'll list my comments in order of appearance:

1) The section “Visual Inference” (Page 4) covers the background of visual inference and the lineup protocol, focusing on its use as a tool for statistical inference. I would like to see the reasoning for using it as a method to compare effectiveness made more clear. From the later text, I understand the argumentation as follows: Because the data is simulated in a way that one location shows the same significant result in two different visualizations (hexagon and choropleth), the detection rate indicates how easily participants can make the detection with respect to the map type.

Added additional text.

2) I had some difficulties following the data simulation process in section 3.2 (Page 6). Specifically, the smoothing process and the reasoning behind this process should be explained more thoroughly.

This has been revised.

3) The analysis of detection accuracy and the presentation of the findings could be improved:

While there is clearly a difference between map types (Page 12, Lines 306-316), the overall level of detection rates is low. I would expect some discussion of these low results. Are these a result of the data simulation process (i.e., the “true” signal might not be different enough)? How do the results compare to the expected detection rates if selected by chance, as discussed in Section 2.2?

Sentence added to results explaining this.

The model definition in Section 3.4.3 on Page 9, as well as the reporting of the results (Page 12 and Table 2), need some work.

First: The model formulation on Page 9 is confusing. I would expect a model formulation of this form

hat(y_ij) = mu + beta1 * maptype + beta2*trend + beta3*maptype*trend + alpha_j + epsilon_ij with i=1,2, …,n decisions from j=1,2, …, 92 participants.

The model formulation has been revised.

Table 2 on Page 22 should include the variance of the random error.

Done, in the text.

Second: The interpretation of the interaction effects on Page 13 Line 329 is a as follows: “Allowing for the interaction effect, the difference in detection rate decreases for population-related displays for a choropleth map lineup but increases for a hexagon tile map display.”

I’m not sure if I understand that statement correctly. In my understanding, the interaction in Table 2 shows that the difference between choropleth and hexagon tiles differs between trend models, but it’s in favor of the hexagon models among all trends (full effect for hexagon with NW-SE trend: 1.63; full effect for hexagon all cities trend: 1.63 + 1.34 -1.16 = 1.81; full effect for hexagon tree cities trend 1.63 – 2.07 + 1.28 = 0.84).

Statement on interactions is removed, because the summary of the model estimates shows the different effects between the different treatments.

Additionally, I can't replicate the detection probabilities mentioned in the text, e.g., Page 13, Lines 333f.: "For the NW-SE distribution, the predicted detection rate for the hexagon tile map display increases the predicted probability of detection to 0.63 from 0.52 for choropleths, this is almost exactly the difference seen in the table of means and is significant only at the 0.05 level."

Are these fixed effects estimates? I can't replicate this and the following results from Table 2, but maybe I'm missing something?

P(detection =1 | choropleth; NWSE) = exp(-1.27)/(1 + exp(-1.27)) = 0.22

P(detection =1 | hex; NWSE) = exp(-1.27 + 1.63)/(1 + exp(-1.27 + 1.63)) = 0.59

P(detection =1 | choropleth; three cities) = exp(-1.27 -2.07)/(1 + exp(-1.27 -2.07)) = 0.03

P(detection =1 | hex; three cities) = exp(-1.27 + 1.63 -2.07 +1.28)/(1 + exp(-1.27 + 1.63 -2.07 + 1.28)) = 0.39

P(detection =1 | choropleth; all cities) = exp(-1.27 + 1.34)/(1 + exp(-1.27 +1.34)) = 0.52

P(detection =1 | hex; all cities) = exp(-1.27 + 1.63 +1.34 -1.16)/(1 + exp(-1.27 + 1.63 +1.34 -1.16)) = 0.63

There were errors in the hand-coded numbers. All of these are now computed using the emmeans package now.

4) My last major concern is the discussion of data quality from the experiment. While data from participants with fewer than three uniquely chosen locations were excluded from the results, the findings for detection speed on Page 14 made me curious about the participants who took only a few seconds. There is a clear gap between these participants and those who took longer to complete the experiment. Does the accuracy differ between these two groups? If the accuracy is significantly lower for the quicker participants, this might indicate that they chose a location by chance to finish the task as quickly as possible. What is the experience from the pilot study? Are these times taken plausible? If not, one might consider removing those cases with implausibly short decision times from the sample to eliminate randomly chosen locations. Regardless of how you deal with these cases, this detail should be discussed in the text.

The explanation of data cleaning has been revised to make it clearer. There is no difference in in groups or participant demographics. This is written in the text now bu results not included. The code for these checks is in the unevaluated chunks the qmd file.

Here are some additional comments, that might be useful to improve the paper:

Page 1 line 16: "None of the existing approaches for creating cartograms or hexagon tiling perform well for the Australian landscape"

and

Page 3 62 “For Australia, the transformations warp the country so that it is no longer recognizable” -> maybe cite Kobakian, Cook & Roberts 2020 to make clear why this statement is made?

Done.

Page 3 Line 78: what’s “substantially above”?

Language changed.

Figure 3: Explain term location in caption

Done.

Page 7 Line 164: “smoothed several times”; specify

Re-worded.

Page 7 Line 169: express smoothing process in math: 0.5 * value of area + 0.5 MEAN(value| area is neighbour)

Re-worded.

Page 9 Line 206: population-related displays; explain which displays

Re-worded

Page 9 Line 201: “The first step in the data cleaning process involved checking that survey responses collected for each participant were only included once in the data set.” Why is this necessary? Without context this sentence is confusing.

Explained.

Page 10 Lines 259 – 264 could be shortened

Done.

Page 11: Line 269 what’s the meaning of the levels of Figure Eight participants?

Explained.

Page 11 3.7 Demographic data collection -> section could be merged with 3.5 “Web application” the reduce redundancies

Kept in this section but section renamed to be “Data collection” because it concerns all the information recorded.

Page 12 Lines 297 – 301: 5 participants where remove because they provided not more than 3 unique choices -> please state a reason for that. The assumption is these participants just “clicked through” the experiment and choose a single location all the time? (see my major concern).

Explanation is improved, and numbers reported are now automatically calculated in the analysis script, which is part of the qmd file generating the paper.

Page 12 section 4.1: Are there differences between group A and B? This would be a good place to show that randomization is working. Underlying assumption: there a no group specific influences on test results.

There is no difference in results from group A and B. There is no difference in in groups or participant demographics. This is written in the text now bu results not included. The code for these checks is in the unevaluated chunks the qmd file.

Page 12 Line 322. Any explanation why the choropleth performed not better?

The difference is small, at least relative to the other two patterns. We expect it might be due to the large rural polygons might also play a role in obscuring large simple spatial trend, too. We don't want to speculate.

Page 18, section 4.2

It would be interesting to see results under control of socio demographics. A second model including those variables could give some hints (direct effects, reduction intra-participant variance). For reference see: (Vanderplas & Hofmann, 2016)

A reference to Majumder, Cook and Hofmann (2025) showing that demographic background has little effect on performance of participants reading lineups. The primary variation is individual. This study was actually done in 2012 but only just published. The Vanderplas paper reports on a different type of experiment where math ability is collected as part of a larger visual testing process, and the type of plots used were ones typically seen in statistics classes. Maps are a bit different from this. Anyway, we did not collect enough data to fit demographics, because it is not the main purpose of this experiment. Majumder et al suggests that as long as a suitable range of backgrounds is included in the study results are reliable and replicable. This has been explained in the text.

Figure 4: use color to distinguish between replicates (facets and color for trend models is redundant)

We disagree with this comment. Replicates are indicated by line segments. The primary factor, map type, is represented by a point, with paired measurements (replicates) connected by lines. This maps the primary information to position along a common axis, which is best graphical practice. Colour redundant with the facets draws attention to the secondary key element of the design, the different types of spatial patterns, and makes the plot more interesting to read.

Page 14: The section "Reason" is extremely short and might be merged with section 4.3 (Speed) on page 14.

We have kept is separate for clarity. Speed is qute different from reason.

There is no reference to table 3 on page 22 in the text.

Fixed.

Literature

Kawakami, Y., Yuniar, S., & Ma, K.-L. (2024). HexTiles and Semantic Icons for MAUP-Aware Multivariate Geospatial Visualizations. https://doi.org/10.48550/arXiv.2407.16897

Reference as broader lit:

Vanderplas, S., & Hofmann, H. (2016). Spatial Reasoning and Data Displays. IEEE Transactions on Visualization and Computer Graphics, 22(1), 459–468. IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2015.2469125

Thank you for the suggestions. The Kawakami reference is out of scope. The Vanderplas et al has been included in the Introduction to point to work on how ability to read a lineup it most related to a subject's visual aptitude. Many more references to the visual inference literature have been added.