Land doesn't get cancer, people do: an experiment comparing the effectiveness of two displays

Stephanie Kobakian
Queensland University of Technology
Science and Engineering Faculty
Brisbane, Australia
stephanie.kobakian@qut.edu.au

Dianne Cook
Monash University
Econometrics and Business Statistics Faculty
Melbourne, Australia
dicook@monash.edu

***Abstract*—The abstract goes here. On multiple lines eventually.**

***Index Terms*—statistics; visual inference; geospatial; population**

INTRODUCTION

Geospatial statistics are often presented on the geographic map base. A choropleth map is the common display to present aggreagated statistics for geographic units, and they are often used to present statistics regarding the population. Creating a choropleth map involves drawing the administrative boundaries and filling them with colour to communicate the value of the statistic. In Australia, there are sets of administrative boundaries that define subdivisions of the population at various granularities. The set of Australia statistical areas presents an example of a heterogrenous distribution of area. The rural communicates on a much larger geographic space than small inner city communities. This has the negative effect of incorrectly showing the spatial distribution of the statistic, especially when a spatial distribution is related to the size of the areas, or the population density.

An alternative display can also be used to effectively communicate a spatial distribution for a set of heterogeneous areas. Viewers of spatial distributions may come to incorrect conclusions.

MOTIVATION

Australian Cancer Atlas

The Australian Cancer Atlas explores the burden of cancer on Australian communities. There are many cancer types presented, and they can be explored on an individual or aggregate level. The Australian communities are examined at the Statistical Areas at Level 2 (SA2)("Australian Statistical Geography Standard (ASGS)" 2018) used by the Australian Bureau of Statistics. Bayesian spatial smoothing has been applied to incorporate the statistics of neighbouring areas, for both privacy and stability of the estimates. The statistics that can be mapped are the diagnoses (Standardised Incidence Rates) and excess deaths for each SA2, communicated as the difference from the Australian average of the statistics. The values of the statistic for each are communicated using a diverging colour scheme.

The Australian Cancer Atlas communicates the trends in the distributions of cancer over geographic space. It uses a choropleth map display and diverging colour scheme to draw attention to relationships between neighbouring areas.

BACKGROUND

Methodology

Spatial visualisations - Choropleth However, the issue of using a choropleth map base becomes obvious when considering the distributions. Position is extremely important for analysis of a visualisation.

Population focussed displays

Map creators have the ability to present spatial statistics in alternative displays that can highlight the population. This work aims to show that a hexagon tile map display is a viable alternative to the geographic map base for presenting population statistics. The same data were shown on a choropleth map, and on a hexagon tile map. Comparing the results of participants who see the choropleth to those who see a hexagon tile map will show that population related distributions are spotted more frequently in a hexagon tile map display.

When presenting population statistics on a geographic map base, the size of the regions can allow errornous conclusions to be drawn about the state of the statistic over the entire population. This occurs as large regions filled with a consistent colour or pettern can draw the attention of map readers, and small regions are not paid equal attention. A choropleth map is not the only display that can be used for presenting geospatial data. Alternative maps include various cartograms, and tesselated tile maps. They allow other variables to be included in the display to highlight the staistical values of various geographic areas.

Visual Inference

- Communicating data through visualisations
- Effective displays for types of data
- Protocol for testing the effectiveness

Classical statistical inference involves hypothesis testing, the process of rejecting a null hypothesis in favour of an alternative. This approach relies on data, the appropriate distributions and their assumptions. Visual inference null hypothesis:

independence in the variables (absence of all features), alternative hypothesis: Relationship between the variables (presence of some feature).

The lineup protocol is used for visual inference testing. 1. simulate null plots 2. Insert data with structure into a random location 3. Ask uninvolved person to select the most different plot 4. If location is chosen correctly, the existence of a feature is significant at $\alpha = 1/N$.

> "In this framework, plots take on the role of test statistics, and human cognition the role of statistical tests." Buja et al. (2009)

The line up protocol involves placing a "guilty" data visualisation in a lineup of "innocents". Where the guilty data set contains structure, and the innocents are equivalent to a null data set. In a grid of visualisations, an observer is asked to pick the display that is most different, if they select the data set containing structure, they have identified the guilty hidden within the group innocents. The guilty data is identified as different from the innocent data with probability $1/m$, where m is the number of null plots plus 1 to account account for the guilty data set. When the guilty data set is chosen, the null hypothesis that it was innocent is rejected with a $1/m$ chance or type I error of being wrong.

The lineup protocol can be used in a variety of testing scenarios. The choropleth map is best used for testing spatial structure in a data set.

Study Design

This study aims to answer several questions around the presentation of spatial distributions:

1. Are spatial disease trends, that impact highly populated small areas, detected with higher accuracy when viewed in a hexagon tile map display?
2. Are people faster in detecting spatial disease trends, that impact highly populated small areas, when using a hexagon tile map display?

additional considerations when completing this experimental task included exploration of the difficulty experienced by participants

Experimental design

The most common display for spatial cancer data is the choropleth map. This will be the comparative visualisation for presenting the lineups (Majumder, Hofmann, and Cook 2013). Most geographic distributions will have some degree of spatial autcorrelation between neighbours. This feature will exist in all plots in the lineup displays, the plot that contains the trend feature shown in only one set of data will also be affected by spatial autocorrelation. A reasonable amount of null plots $N-1$ in the lineup was chosen to ensure data is well hidden. For the detailed choropleth of Australian SA2 areas, we set $N = 12$ to not overwhelm participants. A line up protocol was implemented to arrange 12 maps in each display. Individual displays were created by a combination of map type, and spatial trend model.

TABLE I
THE EXPERIMENTAL DESIGN

Trend	Map type	Replicates
NW-SE	Choropleth	2
	Hexagon tile	2
Three cities	Choropleth	2
	Hexagon tile	2
All cities	Choropleth	2
	Hexagon tile	2

The hypotheses for each lineup are H_0 : All plots look the same H_a : One plot looks different to the other plots

Recruited participants to be uninvolved judges with no prior knowledge of the data to avoid discrimination or advantages. The online crowdsource platform Figure-Eight was used to recruit participants.

The researchers contrasted the different plot designs, as hexagon tilemap and geography in the lineups were created using the same data, and same null positions within the lineup.

Let n be the number of independent observers and x_i the number of observers who picked plot i, $i = \{1, ..., m\}$

Then $x_i, x_2, ..., x_m$ follows a multinomial distribution$Mult_{\pi_1,\pi_2,...,\pi_m}(x_i, x_2, ..., x_m)$ with $\sum_i \pi_i = 1$, where π_i is the probability that plot i is picked by an observer, which we can estimate as $\hat{\pi}_i = x_i/n$. The researchers compared the length of time taken, and the accuracy of the participants choices. The power of a lineup can therefore be estimated as the ratio of correct identifications x out of n viewings.

The variables being manipulated and measured

The variables that were changed between groups were the type of plot shown and the trend model.

Each participant was randomly allocated to either Group A or Group B when they begun the survey. This resulted in 42 participants allocated to Group A, and 53 participants allocated to Group B.

The levels of the factors measured in the experiment were: - Map type: *Choropleth, Hexagon tile* - Trend: *Locations in three population centres, Locations in multiple population centres, South-East to North-West*

Factor combinations examined by each participant amount to 6 (2x3) lineup displays. A participant did see the same data for both map types. Four simulated sets of data were generated for each treatment. This will generate 24 lineups (12 were geographic maps, and 12 were hexagon tile maps). Participants will evaluate 12 lineups, 6 of each map type. Appendix A shows the experimental design visually. For each of the six geographic displays and six hexagon displays, two of each trend model were shown to participants.

The variables measured as a result of the changes were the probability of detection each display and the time taken to submit responses. To measure the accuracy of the detections, the plot chosen for each lineup evaluated was compared to the position of the real spatial trend plot in the lineup. A correct result occurs when the chosen plot matches the position of the real plot, this was recorded in an additional binary variable; 1 = correct; 0 = incorrect. High efficiency occurs when a small amount of time is taken to evaluate each lineup. This will be measured as the numeric variable measuring the length of time taken to submit the answers to the evaluation of each line up.

Simulation process

The underlying spatial correlation model was created to provide spatial autocorrelation between neighbouring areas using the longitude and latitude values for the Statistical Areas. formula = z ~ 1, locations = ~ longitude + latitude

Simulated spatially dependent data using the model on the centroids of each area, for 12 null plots in 12 lineups.

12 sets of data were created. In these 12 sets of data, each of the 144 maps were smoothed several times to replicate the spatial autocorrelation seen in cancer data sets presented in the Australian Cancer Atlas.

For each of the 144 individual maps, the values attributed to each geographic area are rescaled to show a similar colour scale from deep blue to dark red within each map.

A random location was selected for each set of lineup data. In this location, a trend model was overlaid on the null set of spatially correlated data. Each set of lineup data was used to produce a choropleth maps and hexagon tile maps. These matched pairs were split between Group A and Group B.

Participants

There were 95 participants involved in the study. We recruited participants using the Figure-Eight crowd source platform by advertising this survey to participants that fulfilled the following crtieria:

- level 2 or level 3 on the Figure-Eight Platform.
- at least 18 years old

Participants then selected our task from the list of tasks available to them.

Each participant was trained using three test displays orienting them to the evaluation task. Participants then proceeded to the survey, this involved evaluating 12 displays.

Experiment procudure and data collection

The participant answered demographic questions and provided consent before evaluating the lineups.

Demographics were collected regarding the study participants: - Gender (female / male / other), - Degree education level achieved (high school / bachelors / masters / doctorate / other), - Age range (18-24 / 25-34 / 35-44 / 45-54 / 55+ / other) - Lived at least for one year in Australia (Yes / No)

Participants then moved to the evaulation phase. The set of images differed for Group A and Group B. After being allocated to a group, each individual was shown the 12 displays in randomised order.

Three questions were asked regarding each display: - Plot choice - Reason - Difficulty

After completing the 12 evaluations, the participants were asked to submit their responses.

Data was collected through a web application containing the online survey. Each participant used the internet to access the survey. The data collection took place using a secure link between the survey web application and the googlesheet used to store results. The application would first connect to the googlesheet using the googlesheets (Bryan and Zhao 2018) R package, and interacted again at the completion of the survey by adding the participant's responses to the 12 displays as 12 rows of data in the googlesheet.

The methods of data analysis used

The data analysis methods used in order to analyse and collate the results included downloading the survey submissions and opening them into the analysis software R (R Core Team 2019).

For each of the 12 lineup displays the researchers calculated: - accuracy: the proportion of subjects who detected the data plot - efficiency: average time taken to respond

Visualisations: Side-by-side dot plots were made of accuracy (efficiency) against map type, facetted by trend model type.

Similar plots were made of the feedback and demographic variables - reason for choice, reported difficulty, gender, age, education, having lived in Australia - against the design variables.

Plots will be made in R (R Core Team 2019), with the ggplot2 package (Wickham 2016).

Modeling: The results will be analysed using a generalised linear model, with a subject random effect to account for differences in individuals. There will be two main effects: map type and trend model, which gives the fixed effects part of the model to be

$$\widehat{y_{ij}} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij}, \quad i = 1, 2;\ j = 1, 2, 3$$

where $y_{ij} = 0, 1$ whether the subject detected the data plot, μ is the overall mean, $\tau_i, i = 1, 2$ is the map type effect, δ_j is the trend model effect. We are allowing for an interaction between map type and trend model. Because the response is binary, a logistic model is used.

A similar model will be constructed for the efficiency, using a log time, and normal errors.

The feedback and demographic variables will possibly be incorporated as covariates.

Computation will be done using R (R Core Team 2019), with the `lme4` package (Bates et al. 2015).

Limitations of the data collection

This required internet connection for participants to access the survey

RESULTS

The survey responses from participants were kept only if the participant submitted answers for all 12 displays. This resulted in 95 participants.

The contributors who detected no plots correctly were analysed further. Three of these contribuotrs gave no choices

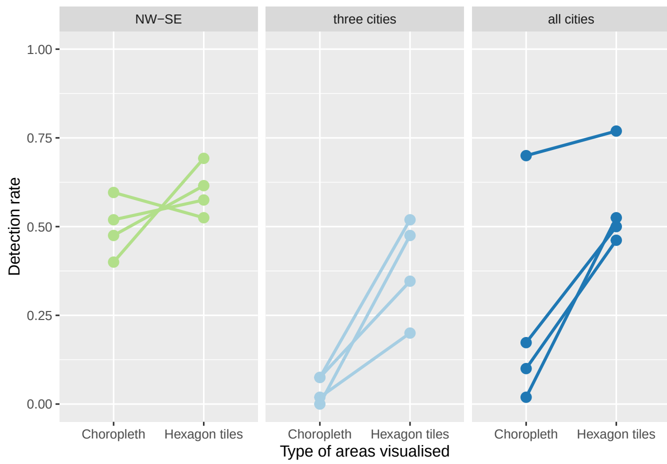


Fig. 1. Each point shows the probability of detection for the lineup display, separated by the trend model hidden in the lineup. The points for the same data set are linked to show the difference in the detection rate when the same data was seen in each display. 11 of the 12 real distribution plots were found more often in the hexagon display.

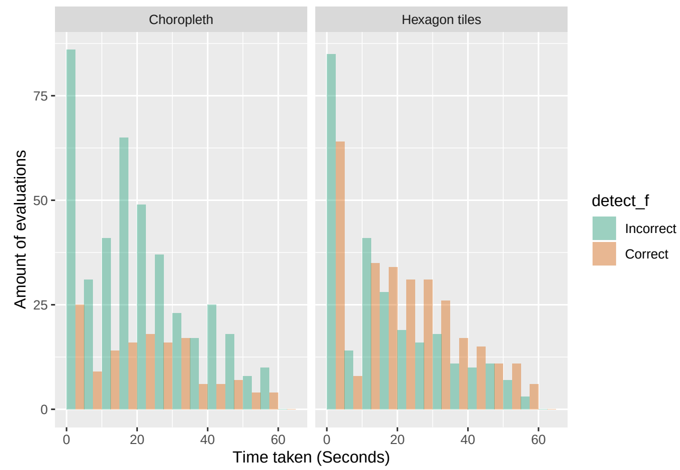


Fig. 2. The time taken to evaluate each display is broken into five second windows. The height of the histogram bars show how many evaulations were submitted within each time window. The distributions for the choropleth and hexagon tile maps are very similar. Both have a large peak at 0-5 seconds, and then a secondary peak at 10-20 seconds. No response took over one minute.

for any of the twelve displays. They were also removed for the rest of the analysis. The contributors who gave various plot choices and reasons for the twelve displays were kept.

Demographics:

```
## # A tibble: 6 x 2
##   age     participants
##   <chr>          <int>
## 1 18 - 24           14
## 2 25 - 34           37
## 3 35 - 44           21
## 4 45 - 54           11
## 5 55+                6
## 6 NA                 3
```

Gender	Bach	High School	Masters	Row Total
Col. Total	56	23	13	79
He	39	19	9	58
She	17	4	4	21

70 of the 95 participants were male, and 25 female and only two of the participants had lived in Australia before.

70 of the participants achieved a Bachelors or Masters degree.

Accuracy

The detection rate is used to find the accuracy for participants reporting the real data trend model. The accuracy can be seen from many views.

trend	type	replicate	mean	std.dev
NW-SE	Choropleth	1	25.00	16.41
NW-SE	Choropleth	2	21.12	15.28
NW-SE	Choropleth	3	21.12	17.05
NW-SE	Choropleth	4	21.20	14.52
NW-SE	Hexagon tiles	1	19.53	14.56
NW-SE	Hexagon tiles	2	21.95	16.42
NW-SE	Hexagon tiles	3	22.02	16.12
NW-SE	Hexagon tiles	4	18.17	14.51
three cities	Choropleth	1	15.30	13.88
three cities	Choropleth	2	25.22	14.36
three cities	Choropleth	3	22.49	15.87
three cities	Choropleth	4	18.68	13.74
three cities	Hexagon tiles	1	20.37	17.02
three cities	Hexagon tiles	2	19.95	16.59
three cities	Hexagon tiles	3	20.02	17.19
three cities	Hexagon tiles	4	18.02	16.93
all cities	Choropleth	1	20.14	16.56
all cities	Choropleth	2	22.75	14.72
all cities	Choropleth	3	19.88	15.03
all cities	Choropleth	4	19.33	16.15
all cities	Hexagon tiles	1	19.83	15.87
all cities	Hexagon tiles	2	16.96	14.65
all cities	Hexagon tiles	3	19.21	15.27
all cities	Hexagon tiles	4	20.03	15.11

A t-test shows the difference between the detection rates for the two types of displays. The value of 0.0041593 shows that it is very unlikely the difference is due to chance.

Speed

Certainty

Certainty levels are measured on a five point scale—they are subjective assessments by the participant 'how certain are you about your choice?'.

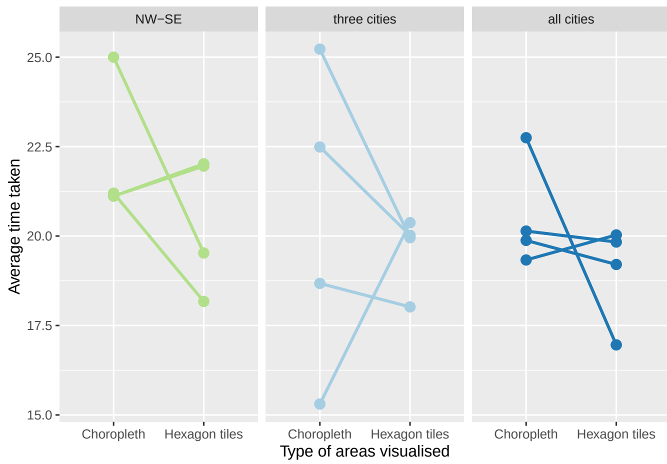


Fig. 3. Each point shows the average time taken for participants to evalute the lineup display, separated by the trend model hidden in the lineup. The points for the same data set are linked to show the difference in the avergae time taken when the same data was seen in each display. There is a lot of variation in the time taken. The shortest average time (15 seconds), and the longest average time (23 seconds) both occured when evaluating three cities

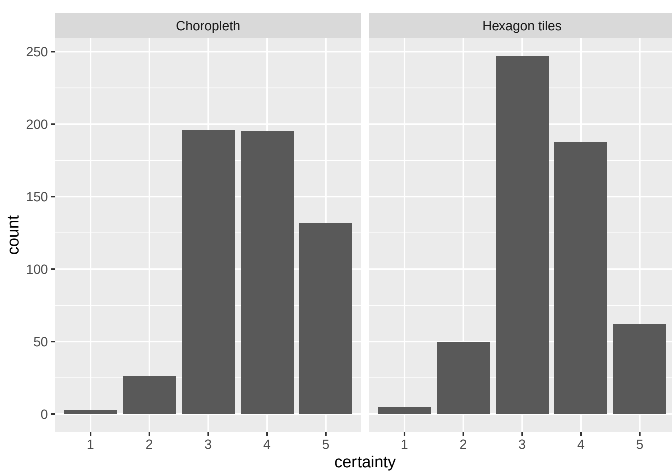
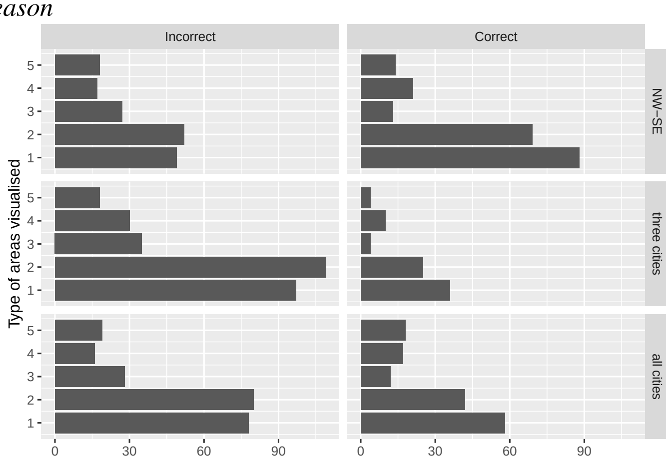


Fig. 4. The amount of times each level of certainty was chosen by participants when viewing hexagon tile map or choropleth displays. Participants were more likely to choose a high certainty when considering a Choropleth map. The default certainty of 3 was chosen most for the Hexagon tile map displays.

Reason

Contributors

Reason



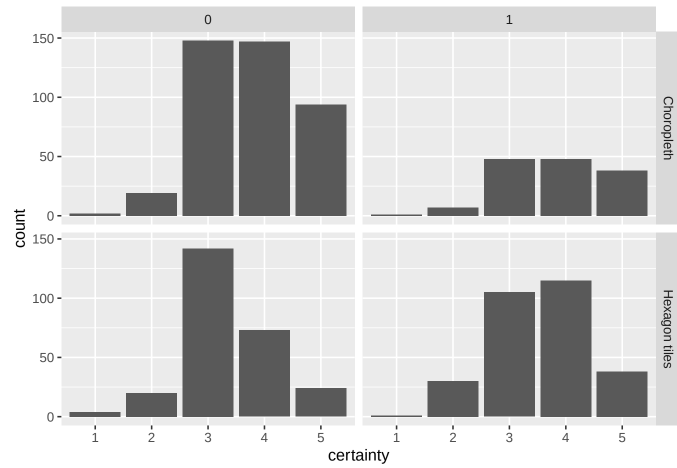


Fig. 5. The amount of times each level of certainty was chosen by participants. The columns shown whether a viewer correctly selected the real trend data plot when viewing hexagon tile map or choropleth displays.

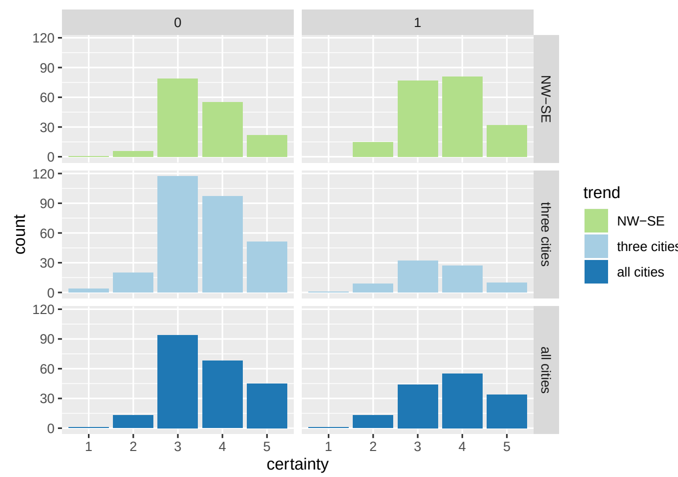


Fig. 6. The amount of times each level of certainty was chosen by participants. The columns shown whether a viewer correctly selected the real trend data plot when viewing hexagon tile map or choropleth displays. The rows show the type of the trend model added in the real data plot. The default certainty level of 3 was chosen most frequently when incorrect. Then shown a NW-SE or all cities trend participants felt more certain of their correct choice.

The choices made by participants are examined in Figure **??**. Participants were misled by the choropleth display, but not the hexagon display for all cities displays except (2). The maps with a North West to South East trend was chosen with much greater frequency in all displays. All of three cities displays, except (4), were detected in the hexagon display. All except one lineup had at least one participant select the correct map in the lineup as shown in Figure **??**.

Anomolies

Modeling the difference

A generalized linear mixed effects model can account for each individual participants' abilities as it includes a subject-specific random intercept. As each participant provides results from 12 lineups.

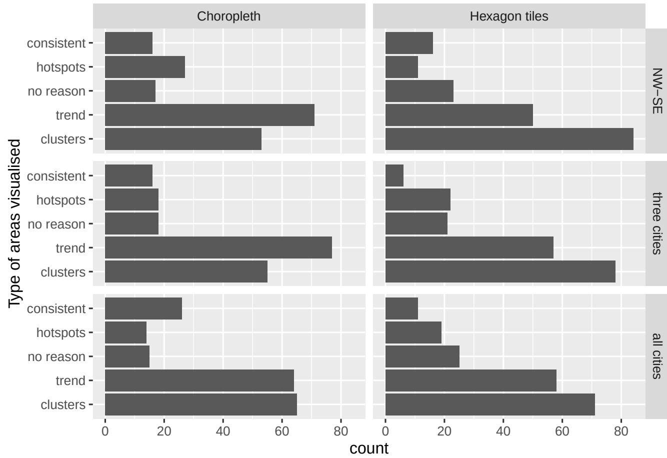


Fig. 7. The most common reason for choice of plot when looking at each trend model shown in Choropleth and Hexagon Tile maps. Clusters were the most common reason when viewing a Hexagon Tile map, trend was the most common choice for choropleth displays except for the all cities display.

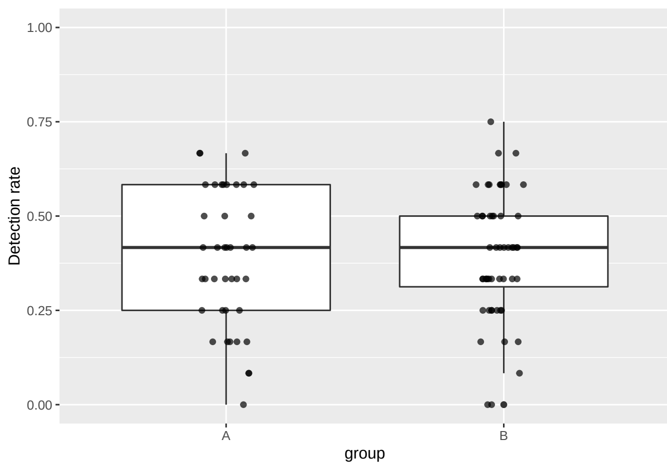


Fig. 8. The probablity of detection acheived by the contributors in each group is shown by the points. Group B has a larger range and a smaller inter-quartile range. Group A and both had 3 people who did not find any of the data maps in the displays.

Detection Rates:

```
## # A tibble: 6 x 5
##   term                                 estimate std.error statistic p.value
##   <chr>                                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)                            0.0217     0.147     0.147   0.883
## 2 typeHexagon tiles                      0.420      0.211     1.99    0.047
## 3 trendthree cities                     -3.25       0.412    -7.88    0
## 4 trendall cities                       -1.41       0.295  [illegible] 0
## 5 typeHexagon tiles:trendthree cities    2.37       0.465     5.10    0
## 6 typeHexagon tiles:trendall cities      1.08       0.312     3.47    0.001
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: detect_f
```

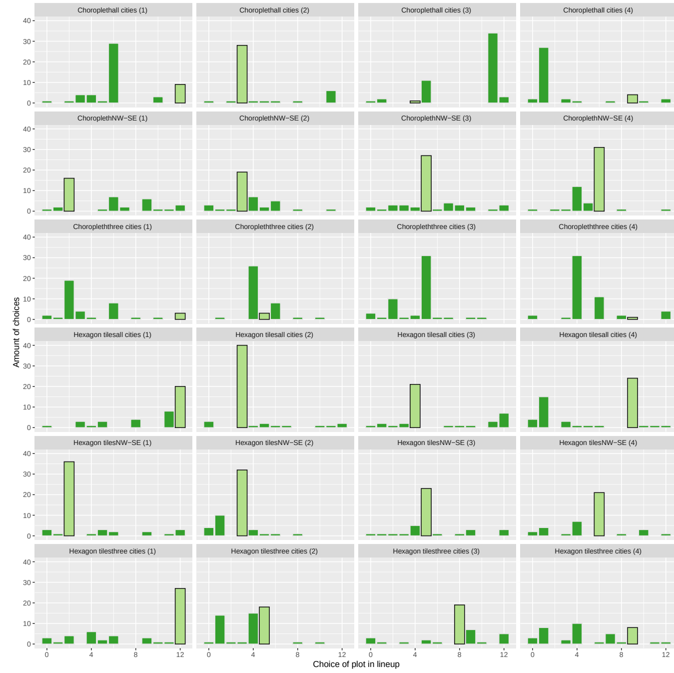


Fig. 9. Each facet is associated with one lineup, the height of the bars count the choices made by the participants considering each lineup. The bars coloured with black outlines show the map which contained a trend model, these are the correct choices. The numbers differentiate the replicates of each trend model and type of map display. Participants were able to select 0 to indicate they did not want to choose a map.

```
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr
## NULL                          1103     1477.0
## type          1   83.526      1102     1393.5 < 2
## trend         2  101.699      1100     1291.8 < 2
## type:trend    2   35.517      1098     1256.2 1.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
```

```
##
##               Incorrect Correct
##   Incorrect         431     121
##   Correct           242     310
```

```
## # A tibble: 1 x 6
##   sigma logLik   AIC   BIC deviance df.residual
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int>
## 1 0.430 [illegible]             1321.        1096
```

```
## # A tibble: 8 x 5
##   term                              estimate
##   <chr>                                <dbl>
## 1 (Intercept)                          0.505
## 2 typeHexagon tiles                    0.103
## 3 trendthree cities                   -0.467
## 4 trendall cities                     -0.277
```

```
## 5 typeHexagon tiles:trendthree cities    0.250    0.0633    3.95  fixed
## 6 typeHexagon tiles:trendall cities      0.239    0.0633    3.77  fixed
## 7 sd_(Intercept).contributor             0.118    NA        NA    contributor
## 8 sd_Observation.Residual                0.430    NA        NA    Residual
```

```
##
##              Incorrect Correct
##   Incorrect        49       1
##   Correct         624     430
```

For a base model of Choropleth map, using a NW-SE trend model. The detection rate for Hexagon tile maps using a NW-SE trend model changes the log odds of the detection by 0.42.

Certainty:

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: detect ~ type * trend + (1 | contributor)
##    Data: d
## REML criterion at convergence: 1350.719
## Random effects:
##  Groups      Name        Std.Dev.
##  contributor (Intercept) 0.1179
##  Residual                0.4296
## Number of obs: 1104, groups:  contributor, 92
## Fixed Effects:
##                       (Intercept)                    typeHexagon tiles
##                            0.5054                               0.1033
##                  trendthree cities                       trendall cities
##                           -0.4674                               0.2772
## typeHexagon tiles:trendthree cities    typeHexagon tiles:trendall cities
##                            0.2500                               0.2391
```

Discussion

Conclusion

how do the results found generalise to other work - Not just for Aus (Canada new Zealand could also use this effective display)

- For USA alternative methods can also be helpful

Supplementary Materials

Training

Survey application

Subject specific anomolies (0% detection)

Acknowledgment

The authors would like to thank…

Ethics approval for the online survey was granted by QUT's Ethics Committee (Ethics Application Number: 1900000991). All applicants provided informed consent in line with QUT regulations prior to participating in this research.

Bibliography styles

References

"Australian Statistical Geography Standard (ASGS)." 2018. *Australian Bureau of Statistics*. Australian Government. %7Bhttps://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian%20%20%20Statistical%20Geography%20Standard%20(ASGS)%7D.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bryan, Jennifer, and Joanna Zhao. 2018. *Googlesheets: Manage Google Spreadsheets from R*. https://CRAN.R-project.org/package=googlesheets.

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wlckham. 2009. "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 367 (1906): 4361–83. http://www.jstor.org/stable/40485732.

Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *Journal of the American Statistical Association* 108 (503): 942–56. https://doi.org/10.1080/01621459.2013.808157.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.