Communicating human related spatial distributions: A comparison of spatial visualisations using visual inference.

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—**Choosing a visualisation method may seem to be a simple task. The choropleth map display is commonly used for communicating spatial data. The visualisation method used to present geospatial data will influence the information derived by map users. Choosing an alternative display will influence the communication of the distribution. The hexagon tile map is presented as an alternative display. Visual inference is used to measure the power of design. The choropleth is used as a comparison point. The hexagon tile map display is an effective visualisation for communicatin=g population related distributions. The hexagon tile map display also allowed identification of the geographic distribution with values monotonically decreasing from North West toward South East areas of Australia.**

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

INTRODUCTION

A geographic map base is commonly used to present geospatial statistics. The statistics presented may also be aggreagated statistics for geographic units, in this case the choropleth map is the common display used. Choropleth maps 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. The several sets of Australian statistical areas presents an example of a heterogrenous distribution. The rural communities in Australia operate 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 be used to effectively communicate a spatial distribution for a set of heterogeneous areas.

MOTIVATION

The Australian Cancer Atlas

The Australian Cancer Atlas is an online interactive web tool created to explore 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 value of the statistics of the neighbouring areas. Spatial smoothing allows for the protection privacy and gives stability to the estimates. The statistics 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 geographic area is communicated through the colour used to fill the areas, the colours are chosen for the 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

Population focussed displays

Spatial visualisations communicate the distribution of statistics over geographic space. The most common display for spatial data is the choropleth map. However, the issue of using a choropleth map base becomes obvious when considering population related distributions. 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. 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.

A cartogram transforms the geographic map base. The transformation begins with a choropleth map base, and the

associated populations. There are several algorithms for transforming, they all involve shifting boundaries to increase or descrease the area of the geographic units on the map depending on the statistics. The shifts of the boundaries result in cartograms with accurate map area for each of the geographic units. The transformation will keep boundaries of neighbouring areas connected, but the result is unfamiliar shapes for each geographic unit.

Alternative displays make the trade off between familiar shapes and representation of variables using the area of geographic units. The non-contiguous cartogoram method keep the shapes of geographic units intact, and change the size to communicate the values of statistics. This method breaks the spatial relationship between areas, and emphasises the difference between the statistic and the amount of land for each geographic unit. The Dorling cartogram is another method that does not retain the boundaries of geographic units. Each unit becomes one circle on the cartogram dispaly, these can be sized according to the value of a variable. Their neighbours can be illustrated by connecting the boundaries of the circles. These cartogram methods all allow two variables to be shown in the single display.

Australia is an extreme case of heterogenous geographic units. To communicate information regarding the small inner city geographic units of Australian cities an alternative display must be employed. This will emphasise the value of the statistic that is relating to areas with high population levels. Figure @ref(fig:liver) shows the distribution of the smoothed average amount of Liver cancer diagnoses for Australian Males from XXX years. The hexagon tile map (b) shows orange hexagons for the inner city Statistical Areas at Level 2 with higher than average levels of diagnoses in the capital cities of Melbourne and Sydney. These inner city areas are not visible on the choropleth map (a). There are still many light blue areas in both displays. The northern Queensland, and Northern Territory SA2's with higher than average rates are still visible but utilise less map space on the hexagon tile map.

Methodology

To test the

Visual Inference

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

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.

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:
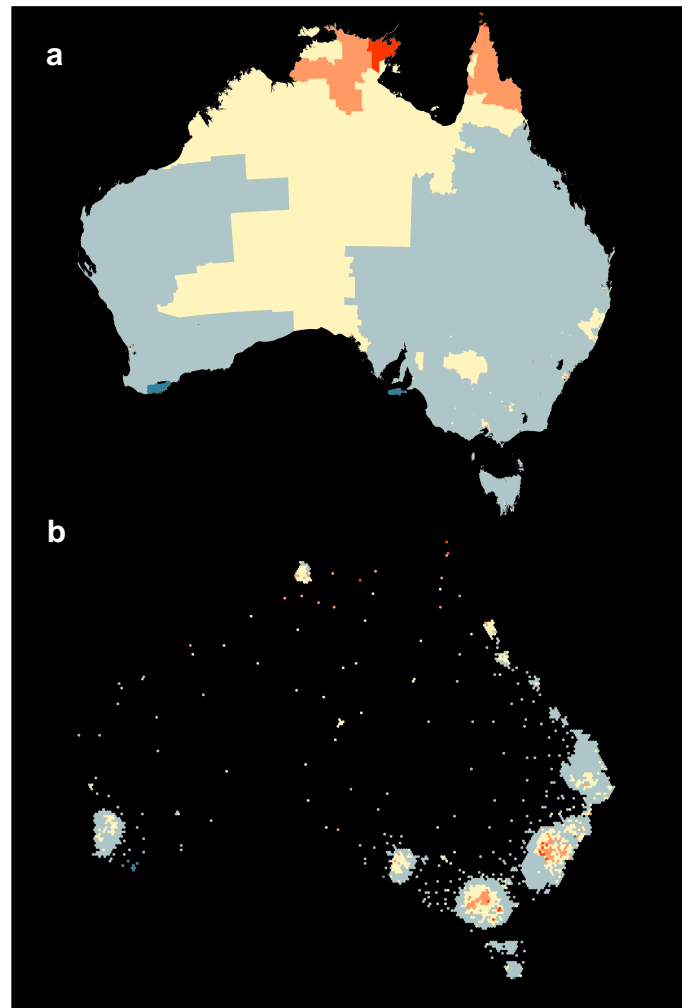

Fig. 1. The smoothed average of liver cancer diagnoses for Australian males. The divering colour scheme shows dark blue areas with much lower than average diagnoses, yellow areas with diagnoses around the Australian average, orange with above average and red shows diagnoses much higher than average.

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

The mean of the detection rate for choropleth map, denoted as μ_C, and the hexagon tile maps, μ_H will be contrasted. This leads to the following one sided hypothesis for this study:

$H_0 : \mu_H = \mu_C \ H_a : \mu_H > \mu_C$

The detection rate $\hat{\pi}$ is calculated as the amount of people that made the choice of plot that contained the real data, out of the participants who saw data plot in the lineup of the null data plots.

The difference in the detection rates for the two displays will be compared using the following:

$$\hat{\pi}_C - \hat{\pi}_H \pm t_{1-\alpha,n-1}\sqrt{\hat{\pi}_C(1-\hat{\pi}_C)/n_C + \hat{\pi}_H(1-\hat{\pi}_H)/n_H}$$

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 autocorrelation 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.

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.

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 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. As the response for whether the participant detected the data plot 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

The demographics of the participants showed a skew towards male participants. The randomness of the group allocation also resulted in more participants being allocated to Group B. Due to the allocation of lineup displays the participants all saw six Choropleth displays and six Hexagon Tile Map displays.

RESULTS

The survey responses from participants were kept only if the participant submitted answers for all 12 displays. This resulted in 92 participants.

The contributors who gave various plot choices and reasons for the twelve displays were kept. The contributors who detected no plots correctly were analysed further. Three of
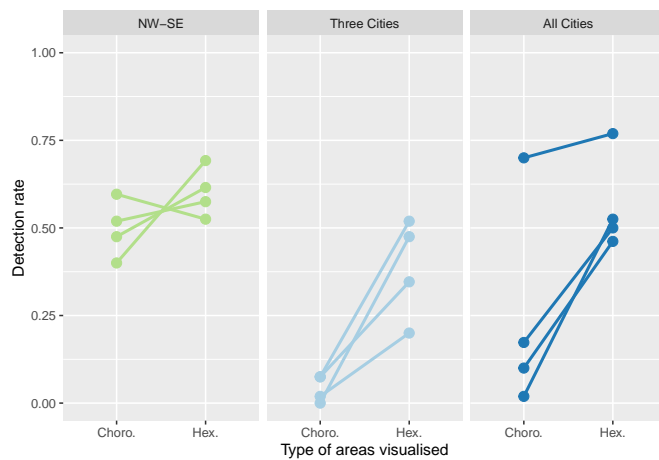
Fig. 2. The detection rates achieved by participants are contrasted when viewing the four replicates of the three trend models. Each point shows the probability of detection for the lineup display, the facets separate the trend models hidden in the lineup. The points for the same data set shown in a choroleth or haxgon tile map display are linked to show the difference in the detection rate.
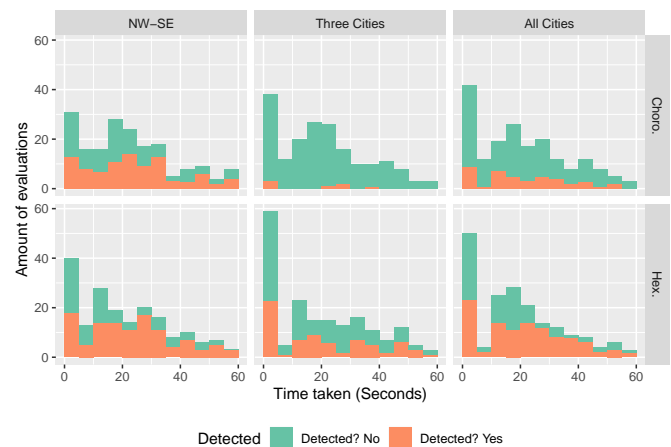

Fig. 3. Six histograms show the distribution of the time taken to submit a response for each combination of trend and type of display. The time taken to evaluate each display is broken into five second windows. The height of the histogram bars show how many evaulations were submitted within each time window. The orange regions of the bars show the amount of correct detections, and the green regions are the amount of incorrect detections.

these contribuotrs gave no choices for any of the twelve displays. They were also removed for the rest of the analysis.

Of the 92 participants, 67 were male, and 25 female. Only two of the participants had lived in Australia before. The education level achieved by the participants was 70 Bachelors or Masters degree.

ACCURACY

The detection rate for participants reporting the real data trend model is used to calculate the accuracy of the responses of participants.

The detection rates are usually higher when participants viewed the hexagon tile map displays. The difference was greater when viewing a three cities trend model. When the trend model affected all cities participants also detected the data display in the lineup of null plots more often. The North West to South East data plot was detected with similar rates when viewing the choropleth and hexagon tile map displays, though for three of the replicates the participants detected the data display more often.

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

The amount of time taken for a participants to submit a response is explored. The time considered was the amount of time it took for a participant to submit their choice of plot, the reason for their choice, and the perceived difficulty of the choice. The time was taken when each participant hit "Next".

The relationship between time taken to submit, and the correct detection by the participant of the real data display in a lineup is considered in Figure @ref(fig:hist_height). The bimodal distributions are similar across the type of display and

the trend model hidden in the real data plot. The hexagon map displays have a peak for the group that took up to 5 seconds to submit a response. When viewing a choropleth map display, the second peak occured at around 20 seconds. The longest a participant took to respond was 60 seconds.
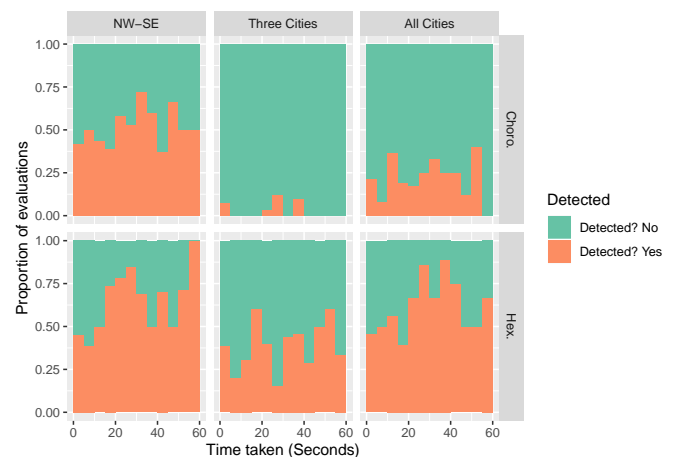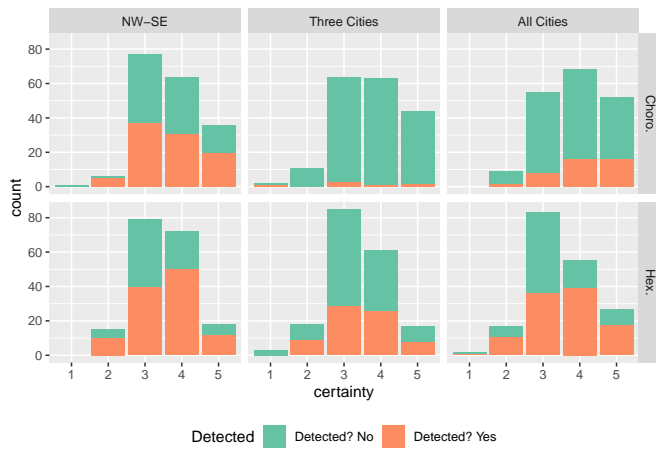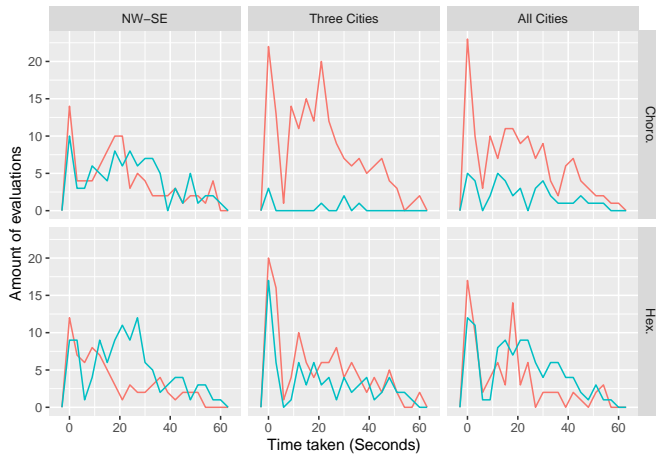
Fig. 4. The amount of times each level of certainty was chosen by participants when viewing hexagon tile map or choropleth displays. Participants were more likely to choose a high certainty when considering a Choropleth map. The mid value of 3 was the default certainty, it was chosen most for the Hexagon tile map displays.
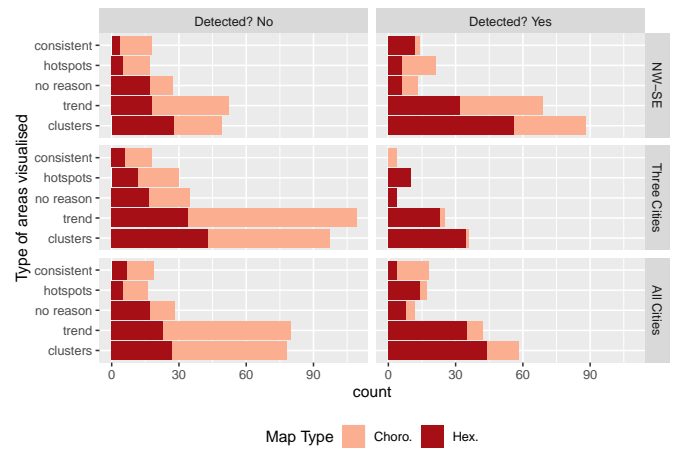

Fig. 5. The amount of participants that selected each reason for their choice of plot when looking at each trend model shown in Choropleth and Hexagon Tile maps. The facets show whether or not the choice was correct.
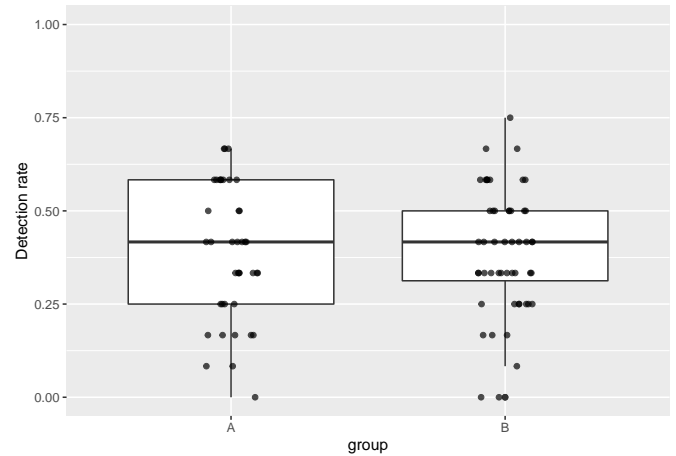

The proportion of correct and incorrect responses within each 5 second time window. For the NW-SE distribution, it is equally likely that quick responders will be correct or incorrect. It is more likely that participants picked the correct response if they took longer to respond. For the Choropelth

However, these quick responders were likely to be either correct or incorrect. The quick responders viewing a choropleth display were unlikely to choose correctly, unless viewing a NW-SE trend model.

The choropleth map displays that included a distribution of three cities was very difficult for participants.

Certainty

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

Reason

The most common choice made by participant was clusters, this response was expected when participants viewed the Hexagon display of the all cities and three cities distributions.

Fig. 6. 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.

Contributors

The choices made by participants are examined in Figure @ref(fig:choices). 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 @ref(fig:choices}.

Anomolies

Modeling the difference

The likelihood of detecting the data plot in the lineup can be modelled using a generalized linear mixed effects model. The base R (R Core Team 2019) glm() function implements
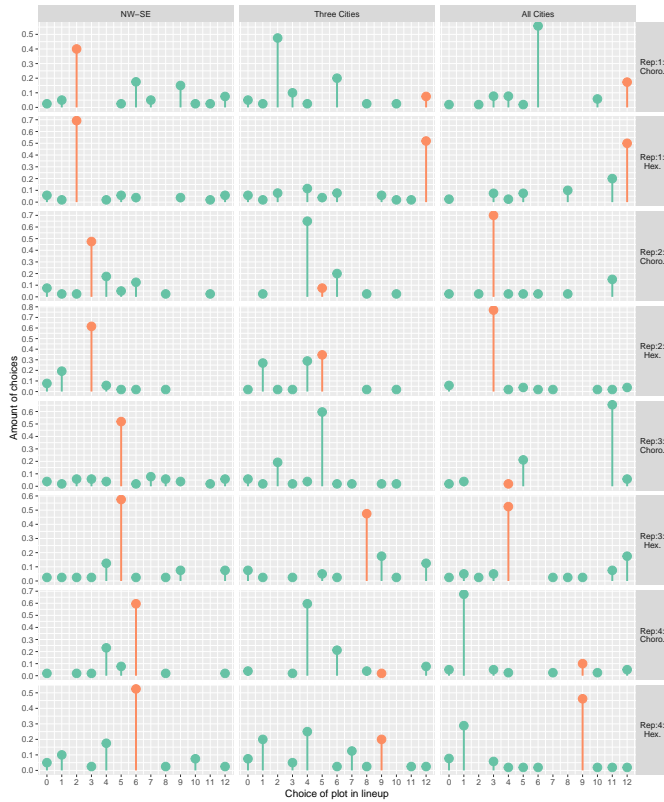
Fig. 7. Each facet is associated with one lineup, the height of the points show the proportion of the participants that made each choice when considering each lineup. The points coloured orange show the map which contained a trend model, these are the correct choices. The numbers differentiate the replicates 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.

generalised linear models. The model used includes the two main effects map type and trend model, which gives the fixed effects model to be:

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

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

Detection Rates: The model specifies a logistic link, this means the predicted values from the `glm` model should be back-transformed to fit between 0 and 1. They are transformed with the link specified below:

$$\mu = \frac{e^{\eta}}{1 + e^{\eta}}$$

$$\eta = f(\tau_i, \delta_j)$$

_	terms	Estimate	Error	
Design	Intercept	0.0217400	0.1474507	0.1474
	Hex	0.4200928	0.2110904	1.990
Trend	Three cities	-3.2519793	0.4124809	-7.8839
	All cities	-1.2398974	0.2293327	-5.4063
Interactions	Hex : Three cities	2.3683138	0.4645163	5.0984
	Hex : All cities	1.0825772	0.3123987	3.4653

```
## 
## 
##                 Detected? No Detected? Yes
##     Predicted: No              431           121
##     Predicted: Yes             242           310
```

This gives the model:

$$\widehat{y_{ij}} = 0.0217 + 0.42Hex - 3.25_{three} - 1.24_{all} + 2.37Hex.three + 1.08Hex.a$$

The estimated marginal mean is the fitted value from a model over the treatment effects. The treatment effects considered in this model are map type displayed, trend model included, and the replicate. The `emmeans` package is used to calculate the estimated marginal mean. The value, between 0 and 1, can be interpretted as the probability of detecting the real data plot in the lineup, for each combination of map type, trend model and replicate.
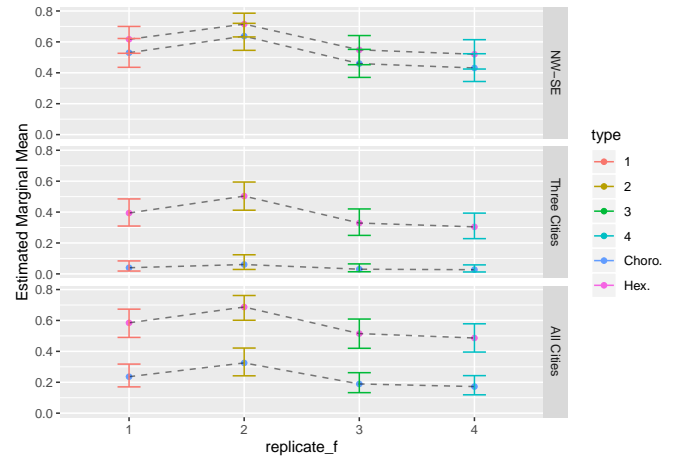

```
## # A tibble: 1 x 6
##    sigma logLik    AIC    BIC deviance df.residual
##    <dbl>  <dbl>  <dbl>  <dbl>    <dbl>        <int>
## 1  0.430  -675. 1367.  1407.    1321.         1096

## # A tibble: 8 x 5
##    term                          estimate std.error
##    <chr>                            <dbl>     <dbl>
## 1 (Intercept)                      0.505    0.0340
## 2 typeHex.                         0.103    0.0448
## 3 trendThree Cities               -0.467    0.0448
## 4 trendAll Cities                 -0.277    0.0448
## 5 typeHex.:trendThree Cities       0.250    0.0633
## 6 typeHex.:trendAll Cities         0.239    0.0633
## 7 sd_(Intercept).contributor       0.118    NA
## 8 sd_Observation.Residual          0.430    NA

## 
##          Detected? No Detected? Yes
```

```
##    No          49              1
##    Yes        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.

```
## # A tibble: 1 x 6
##   sigma logLik  AIC   BIC deviance df.resi
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>        <
## 1 0.445  -685. 1380. 1405.    1365.
```
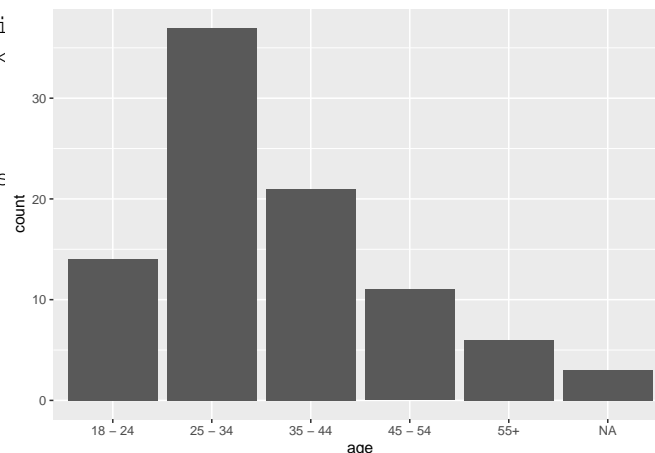
```
## # A tibble: 5 x 5
##   term                        estimate s
##   <chr>                          <dbl>
## 1 (Intercept)                    0.669
## 2 sd_(Intercept).trend           0.456
## 3 sd_typeHex..trend              0.287
## 4 cor_(Intercept).typeHex..trend -0.989
## 5 sd_Observation.Residual        0.445
```

```
##
##        Detected? No Detected? Yes
##    No           0           0
##    Yes        673         431
```

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)
##                  0.5054
##          trendThree Cities
##                 -0.4674
## typeHex.:trendThree Cities
##                  0.2500
```

Training

Survey application

Overall Performance

Subject specific anomolies (0% detection)

Demographics:


```
## # A tibble: 2 x 2
##   gender      n
##   <chr>   <int>
## 1 He         67
## 2 She        25
```

```
## # A tibble: 3 x 2
##   education       n
##   <chr>       <int>
## 1 Bach           56
## 2 High School    23
## 3 Masters        13
```

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

```
## typeHex.
## 0.1033

## trendAll Cities
## -0.2772

## typeHex.:trendAll Cities
## 0.2391
```

BIBLIOGRAPHY STYLES

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

References

"Australian Statistical Geography Standard (ASGS)." 2018. *Australian Bureau of Statistics*. Australian Government. {https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian Statistical Geography Standard (ASGS)}.

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/.