Comparing the Effectiveness of the Choropleth Map with a Hexagon Tile Map for Communicating Cancer Statistics

Stephanie Kobakian and Dianne Cook

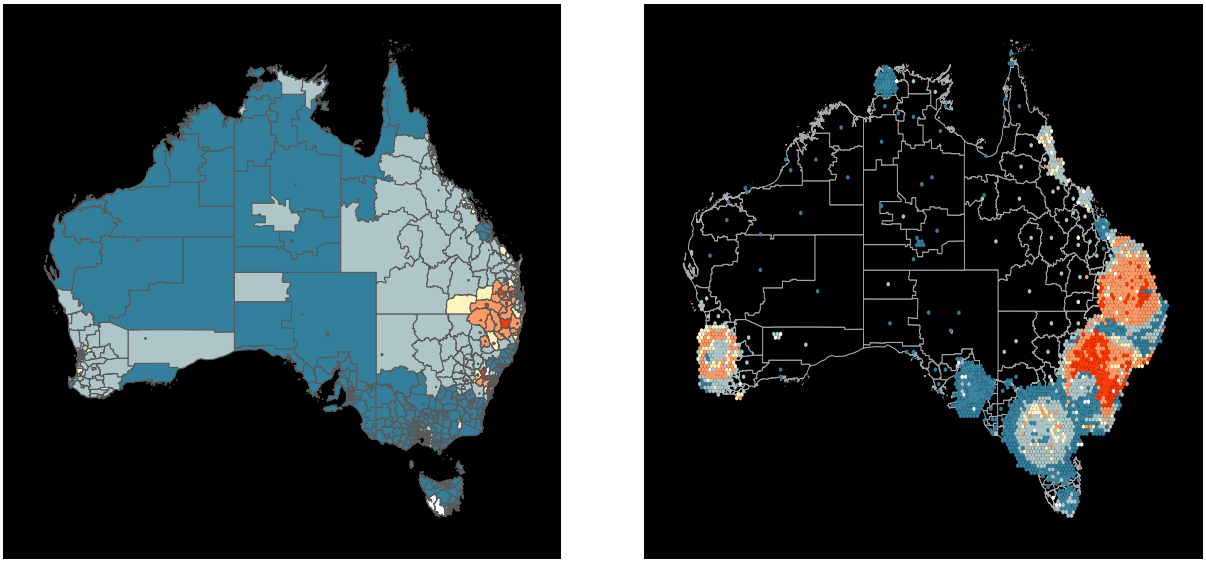

Fig. 1: Which plot type is better? Distribution of thyroid incidence among females across the Statistical Areas of Australia at Level 2, displayed as a choropleth map (left), and a hexagon tile map (right). Blue indicates lower than average and red indicates higher than average incidence. There is high incidence in several of the metropolitan regions (Brisbane, Sydney and Perth) that can now be seen in the hexagon tile map, along with numerous isolated spots, which were obscured in the choropleth.

Abstract—The choropleth map display is commonly used for communicating spatial distributions across geographic areas. However, when choropleths are used the size of the geographic units will influence the understanding of the distribution derived by map users. The hexagon tile map is presented as an alternative display for visualizing population-related distributions effectively. Visual inference is used to measure the power of the hexagon tile map design, in comparison to the choropleth. The hexagon tile map display is tested using a distribution that is directly related to the geography, with values monotonically increasing from the North-West to South-East areas of Australia. This study finds in a hexagon tile map lineup the single map that contains a population-related distribution is detected with greater probability than the same data displayed in a choropleth map. These findings should encourage map creators to implement alternative displays and consider a hexagon tile map when presenting spatial distributions of heterogeneous areas.

Index Terms—statistics; visual inference; geospatial; population distribution, cartogram, experiment, power comparison

1 Introduction

This study compares the effectiveness of a new spatial display, a hexagon tile map, against the standard, a choropleth map, for communicating information about disease statistics. The choropleth map is the traditional method for visualizing aggregated statistics across administrative boundaries. The hexagon tile map builds on existing displays, such as the cartogram, and tessellated hexagon displays. A hexagon tile map forgoes the familiar boundaries, in favor of representing each geographic unit as an equally sized hexagon, placed approximately in the correct spatial location. It differs in the relaxed requirement to have connected hexagons and allows sparsely located hexagons. This type of display may be useful for other countries, and other purposes. The algorithm to construct a hexagon tile map is available in the R package `sugarbag` [12].

The hexagon tile map was specifically designed for Australia to address the challenges of displaying spatial statistics for the Australian Cancer Atlas. Traditional approaches for creating cartograms or hexagon tiling were not suitable for the Australian landscape. The vast open spaces and small population clusters concentrated on the coastlines generate convergence problems for cartogram algorithms and even when they converge yield spatial displays that are no longer recognizable as Australia. There was a need for a new approach to displaying Australian geography to accurately represent the spatial distribution of cancer statistics.

The Australian Cancer Atlas [7] is an online interactive web tool created to explore the burden of cancer on Australian communities. There are many cancer types to be explored individually or aggregated. The Australian Cancer Atlas allows users to explore the patterns in the distributions of cancer statistics over the geographic space of Australia. It uses a choropleth map display and diverging color scheme to draw attention to relationships between neighboring areas. The hexagon tile map may be a useful alternative display to enhance the Atlas.

To objectively test the effectiveness of the two displays, this experiment was conducted using the lineup protocol, a visual inference

- *Stephanie Kobakian was with Queensland University of Technology. E-mail: stephanie.kobakian@gmail.com*
- *Dianne Cook is with Monash University E-mail: dicook@monash.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

procedure [29]. The design of the experiment includes several spatial patterns as seen in the cancer statistics data, several simulated sets of data to provide replication, and crowd-sourcing to collect responses from almost 100 people. The procedure follows that used in Hofmann et al. (2012) [11].

The paper is organized as follows. The next section discusses the background of geographic data displays and visual inference procedures. The Methodology section describes the methods for conducting the experiment and analyzing the results. The results are summarized in the Results section.

2 Background

2.1 Spatial data displays

Spatial visualizations communicate the distribution of statistics over geographic landscapes. The choropleth map [23, 24] is a traditional display. It is used to present statistics that have been aggregated on geographic units. Creating a choropleth map involves drawing polygons representing the administrative boundaries, and filling them with color mapped to the value of the statistic. The choropleth map places the statistic in the context of the spatial domain so that the reader can see whether there are spatial trends, clusters, or anomalies. This is important for digesting disease patterns. If there is a trend it may imply that the disease is spreading from one location to another. If there is a cluster, or an anomaly, there may be a localized outbreak of the disease. Aggregating the statistic on administrative units, provides a level of privacy to individuals while allowing the impact of the disease on the community to be analyzed.

The choropleth map is an effective spatial display if the size of the geographic units is relatively uniform. This is not the case for most countries. Size heterogeneity in administrative units is particularly extreme in Australia: most of the landscape of Australia is sparsely settled, with the population densely clustered into the narrow coastal strips. Fig. 1 shows the choropleth map of thyroid cancer rates in Australia. The choropleth map focuses attention on the geography, and for heterogeneously sized areas it presents a biased view of the population-related distribution of the statistic [15]. *Land doesn't get cancer, people do* – a more effective way to communicate the spatial distributions of cancer statistics is needed.

A cartogram is a general solution for better displaying a population-based statistic. It transforms the geographic map base to reflect the population in the geographic region while preserving some aspects of the geographic location. There are several cartogram algorithms [8, 15]; each involves shifting the boundaries of geographic units, using the value of the statistic to increase or decrease the area taken by the geographic unit on the map. The changes to the boundaries result in cartograms that accurately communicate the population by mapping the area for each of the geographic units but can result in losing the familiar geographic information. For Australia, the transformations warp the country so that it is no longer recognizable.

Alternative algorithms make various trade-offs between familiar shapes and the representation of geographic units. The non-contiguous cartogram method [20] keeps the shapes of geographic units intact and changes the size of the shape. This method disconnects areas creating empty space on the display losing the continuity of the spatial display of the statistic. The Dorling cartogram [8] represents each unit as a circle, sized according to the value of the statistic. Neighbor relationships are mostly maintained by how the circles touch. A similar approach was pioneered by Raisz [22], using rectangles that tile to align borders of neighbors [17]. There have been thorough reviews of the array of methods, suitable for cancer atlas displays [13, 23].

The hexagon tile map algorithm, automatically matches spatial regions to their nearest hexagon tile, from a grid of tiles. It has the effect of spreading out the inner city areas while maintaining the spatial locations or regions in remote areas. The algorithm is available in the R package, `sugarbag` [12]. Fig. 1 shows the hexagon tile map. Color maps from substantially below average (blue) to substantially above average (red) rates. The inner city areas have expanded, making it possible to see the cancer incidence in the small, densely populated areas. Remote regions are represented by isolated hexagons, which is not ideal but maintains the spatial location of these data values. It is of interest to know how well the spatial distribution is perceived for this display, in comparison to the choropleth.

2.2 Visual Inference

In order to assess the effectiveness of the hexagon tile map, the lineup protocol [6, 29] from visual inference procedures is employed. The approach mirrors classical statistical inference. The procedures for doing a power comparison of competing plot designs, outlined in Hofmann et al. (2012) [11], are followed.

In classical statistical inference hypothesis testing is conducted by comparing the value of a test statistic on a standard reference distribution, computed assuming the null hypothesis is true. If the value is extreme, the null hypothesis is rejected, because the test statistic value is unlikely to have been so extreme if it was true. In the lineup protocol, the plot plays the role of the test statistic, and the data plot is embedded in a field of null plots. Defining the plot using a grammar of graphics [26] makes it a functional mapping of the variables and thus, it can be considered to be a statistic. With the same data, two different plots can be considered to be competing statistics, one possibly a more powerful statistic than the other.

To do hypothesis testing with the lineup protocol requires human evaluation. The human judge is required to identify the most different plot among the field of plots. If this corresponds to the data plot – the test statistic – the null hypothesis is rejected. It means that the data plot is extreme relative to the reference distribution of null plots.

The null hypothesis is explicitly provided by the grammatical plot description. For example, if a histogram is the plot type being used, the null might be that the underlying distribution of the data is Gaussian. Null data would be generated by simulating from a normal model, with the same mean and standard deviation as the data. In practice, the null hypothesis used is generic, such as *there is NO structure or a pattern in the plot*, and contrasted to an alternative that there is structure.

The chance that an observer picks the data plot out of a lineup of size m plots accidentally, if the null hypothesis is true is $1/m$. With K observers, the probability of k randomly choosing the data plot, roughly follows a binomial distribution with $p = 1/m$. Fig. 2 shows a lineup of the hexagon tile map, of size $m = 12$. Plot 3 is the data plot, with data generated as described below, and the remaining 11 are plots of null data.

In order to determine the effectiveness of a type of display, this probability is less relevant than the overall proportion of observers who pick the data plot, k/K. The power of the test statistic (data plot) is provided by this proportion. Power in a statistical sense is the ability of the statistic to *produce a rejection* of the null hypothesis if it is indeed *not true*. With the same data plotted using two different displays, the display with the highest proportion of people who choose the data plot would be considered to be the most powerful statistic.

3 Methodology

This study aims to answer two key 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?
2. Are people faster in detecting spatial disease trends that impact highly populated small areas when using a hexagon tile map?

Additional considerations when completing this experimental task included the difficulty experienced by participants and the certainty they had in their decision.

Maps of Australia, with Statistical Area 3 (SA3) [3] as the geographic units are used for the study. The results should apply broadly to any other geographic regions of interest.

3.1 Experimental factors

The primary factor in the experiment is the plot type. The secondary factor is a trend model. Three trend models were developed: two

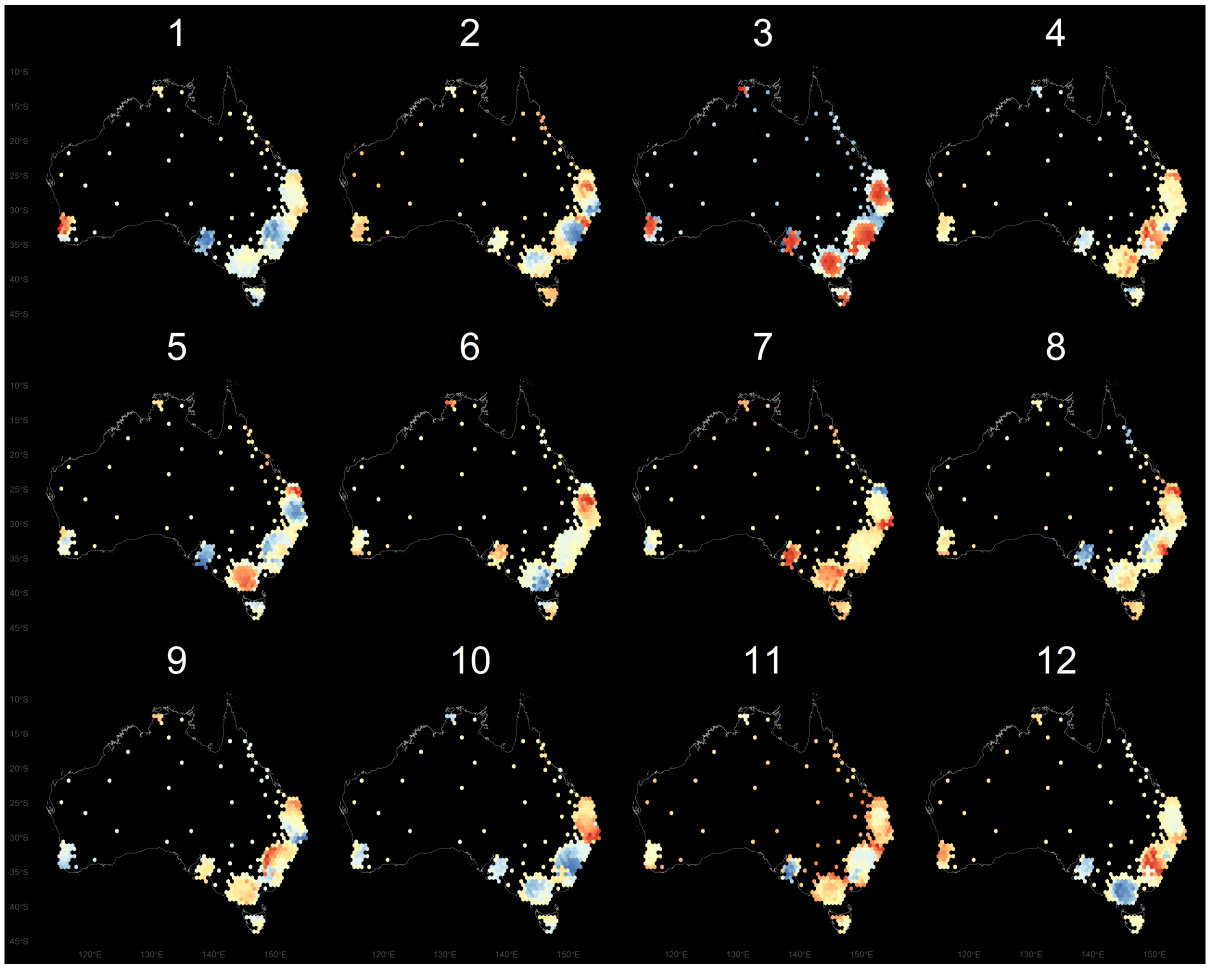

Fig. 2: An example of the lineups shown to participants. This lineup of twelve hexagon tile map displays contains one map with a simulated population-related structure and the remainder are null plots that contain only spatial correlation between neighbors. In this lineup, the population structure is the concentration in all cities, and is located in position 3.

with differing levels of inner city hot spots, and one mirroring a large spatial trend for which the choropleth would be expected to do well. Fig. 3 shows examples of the different trend models displayed using a hexagon tile map. The inner city hotspots trend model reflects the structure seen in the thyroid cancer data (Fig. 1). This produces six treatment levels:

- Map type: *Choropleth, Hexagon tile*
- Trend: *North West to South East; Locations in three population centres; Locations in multiple population centres,*

Data is generated for each of the trend models, with four replicates, and each is displayed both as a choropleth and as a hexagon tile map, which yields 12 data sets, and 24 data plots. This set of displays is divided in half, providing two sets of 12 displays, Group A and Group B. Participants were randomly allocated to Group A or B. Participants saw a data set only once, either as a choropleth or as a hexagon tile map. Fig. 4 summarises the design and the allocation of the displays.

3.2 Generating null data

Null data needs to be data with no (interesting) structure. In most scenarios, permutation is the main approach for generating null plots. It is used to break any associations between variables while maintaining marginal distributions. This is too simple for spatial data. In spatial data, a key feature is the spatial dependence or smoothness over the landscape. To do something simple, like permuting the values relative to the geographic location would produce null plots which are too chaotic, and the data plot will be recognizable for its smoothness rather than any structure of interest.

For spatial data, null data is stationary data, where the mean, variance, and spatial dependence are constant over the geographic units. Stationary data is specified by a variogram model [16] Simulating from a variogram model, where the spatial dependence is specified, generates the stationary spatial data used for the null plots. The parameters for the Gaussian model were sill=1, range=0.3 with the variance generated by a standard normal distribution.

The R package `gstat` [10] was used to simulate 144 null sets, 12 data sets for each plot in a lineup, and 12 sets for 12 lineups.

The null model imposed by our hypothesis suggests that neighbors are related. The randomness induced when generating the null data was smoothed to mirror the practices employed by the Australian Cancer Atlas statisticians. In these 12 sets of data, each of the 12 maps was smoothed several times to replicate the spatial autocorrelation seen in cancer data sets presented in the Australian Cancer Atlas, without implementing uncertainty via transparency.

A list of neighbors for each geographic unit was generated to use when smoothing the distributions. For each geographic unit, the same spatial smoother was applied in each layer of smoothing. It kept half of the units' previous value and derived the new half as the mean of the values of its neighbors at the previous layer of smoothing.

This smoothing allowed neighbors to be related to each other, but also allowed outliers, and showed distributions similar to the thyroid cancer distribution (Fig. 1).

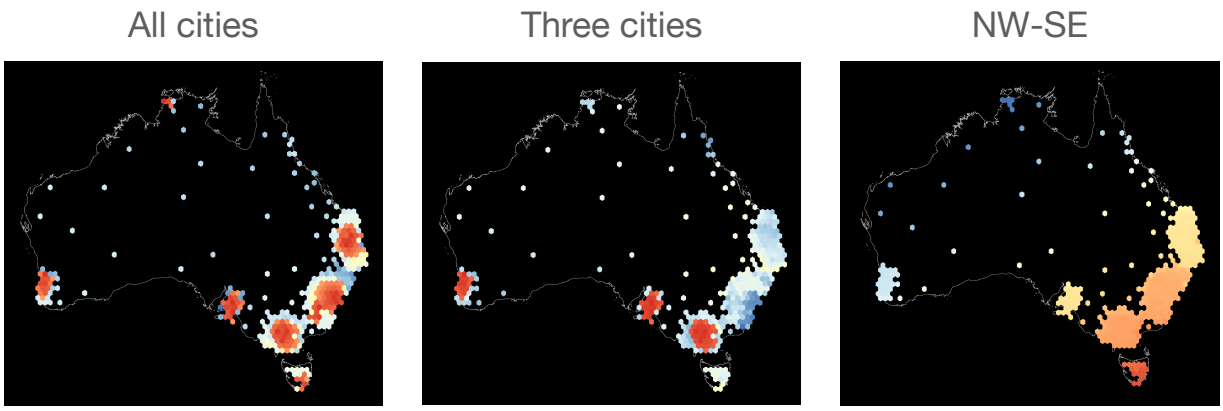

Fig. 3: The three trend models, displayed as a hexagon tile map. One would expect the choropleth to do best only on the "northwest to southeast" pattern, and the hexagon tile map to perform better with the "all cities" and "three cities".

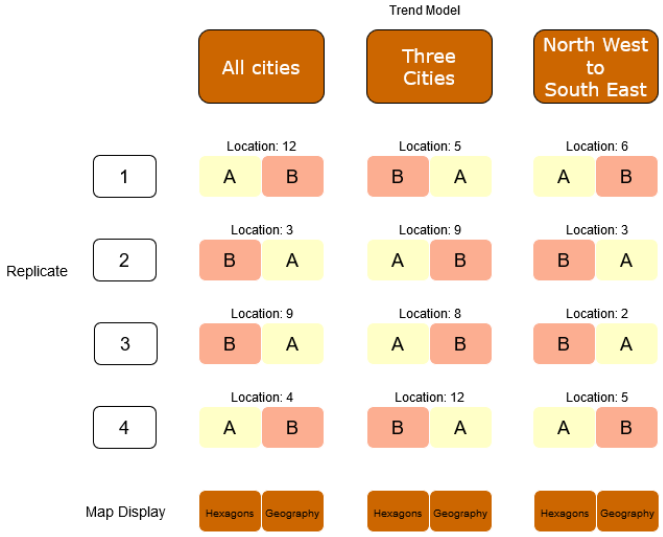

Fig. 4: Schematic illustrating the experimental design used in this visual inference study.

3.3 Generating lineups

For each trend model, four real data displays were created by manipulating the centroid values of each of the SA3 geographic units.

The North West to South East (NW-SE) distribution was created using a linear equation of the centroid longitude and latitude values.

The All Cities trend model was created using the distance from the centroid of each geographic unit to the closest capital city in Australia, calculated when creating the hexagon tile map using the `sugarbag` [12] package. 201 of the 336 SA3s were considered greater capital city areas, and the values of these areas were increased to create red clusters. The amount was chosen to make clusters around the cities visible in the choropleth display even if they were not overly noticeable.

A similar selection process was applied to the Three Cities' trend model. However, for each of the four replicates for the Three Cities trend, a random sample of capital cities was taken from Sydney, Brisbane, Melbourne, Adelaide, Perth, and Hobart. Only the values of the areas nearest to the three cities were increased to create clusters.

One of the lineup locations was chosen to embed the real trend model map, in each of the four replicates, for the three trend models. The location was chosen from a sub-sample of the 12 possible locations. The chance of repetition using resampling was introduced to prevent participants from inducing the location by elimination, the locations 1, 7, 10, and 11 were not used.

As seen in Fig 4, the choropleth and hexagon display used the same location for the real data display of the trend model and was added to the spatially correlated null values for each lineup. Each set of lineup data was used to produce a choropleth map lineup and hexagon tile map lineup. These matched pairs were split between Group A and Group B according to the 2 x 3-factor experimental design depicted in 4.

For each of the 144 individual maps, the values for each geographic area were rescaled to create a similar color scale from deep blue to dark red within each map. This meant at least one geographic unit was colored dark blue, and at least one was red, in every map display of every lineup.

For the geographic NW-SE distribution, this resulted in the smallest values of the trend model (blue) occurring in Western Australia, the North West of Australia, and the largest values of the trend model (red) occurring in the South East. This resulted in Tasmania being colored completely red.

For the population-related displays, the clusters in the cities appeared redder than the rest of Australia.

3.4 Analysis

3.4.1 Data Cleaning

The first step in the data cleaning process involved checking that survey responses collected for each participant was only included once in the data set. The data cleaning process also involved filtering out participants' who did not provide at least three unique choices when considering each of the twelve lineups. These participants achieved a detection rate of 0. If participants had made various plot choices for the 12 displays they saw they were still included in the dataset.

3.4.2 Descriptive statistics

Basic descriptive statistics were used to contrast the detection rate for the two types of displays. A comparison was also made across the trend models, contrasting the mean and standard detection rate for each group, which had seen the different map display types for each replicate.

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

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

Plots will be made in R [21], with the `ggplot2` package [26].

3.4.3 Modelling

The likelihood of detecting the data plot in the lineup can be modeled using a linear mixed-effects model. The R [21] `glmer()` function in the `lme4` [4] package implements generalized linear mixed effect

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 $y_{ij} = 0, 1$ is the log odds for 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 is binary, so a logistic model was used. As each participant provides results from 12 lineups, this model can account for each individual participant's abilities as it includes a subject-specific random intercept.

The model specifies a logistic link, this means the predicted values from the `glmer` model should be back-transformed to fit between 0 and 1. The predictions $\widehat{p}(\eta)$ are transformed to be probabilities between 0 and 1 with the link specified below:

$$\widehat{p}(\eta) = \frac{e^{\eta}}{1 + e^{\eta}}$$

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

3.5 Web application to collect responses

The `taipan` [14] package for R was used to create the survey web application. This structure was altered to collect responses regarding participants' demographics and their survey responses. The survey app contained three tabs. Participants were first asked for their demographics their Figure Eight contributor ID, and their consent to the responses being used for analysis. The demographics collected included participants' preferred pronouns, the highest level of education achieved, their age range, and whether they had lived in Australia.

After submitting these responses, the survey application switched to the tab of lineups and associated questions. This allowed participants to easily move through the twelve displays and provide their choice, the reason for their choice, and the level of certainty.

When participants completed the twelve evaluations the survey application triggered a data analysis script. This created a data set with one row per evaluation. Containing the responses to the three questions. The script also added the title of the image, which indicated the type of map display, the type of distribution hidden in the lineup, and the location of the data plot. It also calculated the time taken by participants to view each lineup.

Each participant used the internet to access the survey. The data transfer from the web application to the data set took place using a secure link to the google sheet used to store results. The application connected to the sheet using the `googlesheets` [5] R package when participants opened the application and interacted again when participants chose to submit the survey. At this time it added the participant's responses to the twelve lineup displays as twelve rows of data in the sheet.

3.6 Participants

Participants were recruited from the Figure Eight crowdsourcing platform [9] to evaluate lineups. The lineup protocol expects that the participants are uninvolved judges with no prior knowledge of the data, to avoid inadvertently affecting results. Potential participants need to have achieved level 2 or level 3 from prior work on the platform. All participants were at least 18 years old.

Participants were allocated to either group A or group B when they proceeded to the survey web application. There were 92 participants involved in the study. All participants read introductory materials and were trained using three test displays, to orient them to the evaluation task. All participants who completed the task were compensated $AUD5 for their time, via the Figure Eight payment system.

A pilot study was conducted in the working group of the Econometrics and Business Statistics Department of Monash University. This allowed us to estimate the effect size, and thus decide on a number of participants to collect responses from.

3.7 Demographic data collection

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

Demographics were collected regarding the study participants:

- Gender (female / male / other),
- 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 evaluation 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.

4 Results

Responses from 92 participants were collected. Five participants did not provide more than three unique choices for the twelve lineups, and their data was removed. Set A was evaluated by 42 participants, and 53 evaluated set B. This resulted in 1104 evaluations, corresponding to 92 subjects, each evaluating 12 lineups, that were analysed on accuracy and speed. The certainty and reasons of subjects in their answers is also examined.

4.1 Participant demographics

Of the 92 participants, 67 were male, and 25 female. Most participants (56) had a Bachelors degree, 13 had a Masters degree, and the remaining 23 had high school diplomas.

4.2 Accuracy

Fig. 5 displays the average detection rates for the two types of plots separately for each trend model. Each trend model was tested using four repetitions, and evaluations on the same data set were seen as either choropleths or hexagon tile maps by each group as specified in Fig. 4; the detection rates for each display are connected by a line segment. The Three Cities and All Cities trend models shown in the hexagon tile map allowed viewers to detect the data plot substantially more often than the choropleth counterparts. One replicate for the All Cities group had similar detection rates for both plot types, the rate of detection using the choropleth map was much higher than other replicates. Surprisingly, participants could also detect the gradual spatial trend in the NW-SE group from the hexagon tile map. We expected that the choropleth map would be superior for the type of spatial pattern, but the data suggests the hexagon tile map performs slightly better, or equally as well.

Table 1 shows the means and standard deviations of the detection rate for each type of plot and each trend model. This also gives the standard deviations, the smallest standard deviation for all sets of replicates was the Three Cities trend model shown in a Choropleth display. This group of displays had a very small detection rate of 0.04. The mean detection rate for the Three Cities trend model shown as choropleth map lineups was also the smallest at 0.40. The North-West to South-East (NW-SE) trend model unexpectedly had a higher mean detection rate for the hexagon tile map displays, but the difference in the means of detection rate was only 0.10.

Table 2 presents a summary of the generalized linear mixed effects model, testing the effect of plot type and trend model on the detection rate. The results support the summary from Fig. 5 and all parameters are statistically significant despite the large standard deviations observed in Table 1. Overall, the hexagon tile map performs marginally better than the choropleth for all trend models, which is a pleasant surprise. Allowing for the interaction effect, the difference in detection rate

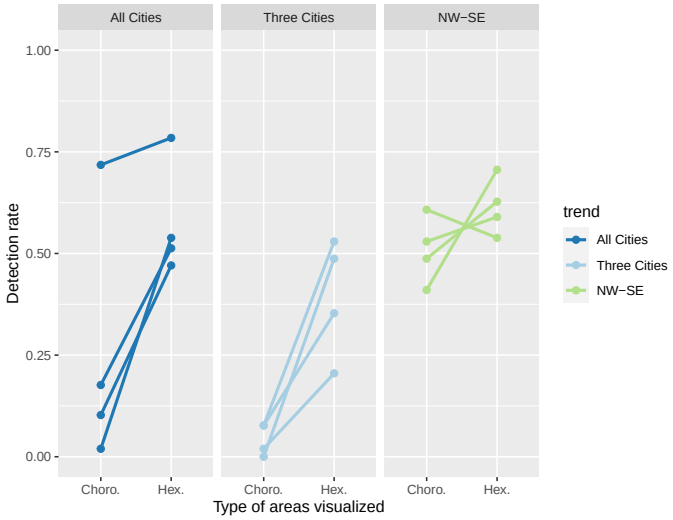

Fig. 5: The detection rates achieved by participants are contrasted when viewing the four replicates of the three trend models. Each point shows the probability of detection for the lineup display, the facets separate the trend models hidden in the lineup. The points for the same data set shown in a choropleth or hexagon tile map display are linked to show the difference in the detection rate.

Table 1: The mean and standard deviation of the rate of detection for each trend model, calculated for the choropleth and hexagon tile map displays.

Type	NW-SE	Three Cities	All Cities
Choro.	0.52 (0.50)	0.04 (0.19)	0.23 (0.42)
Hex.	0.62 (0.49)	0.40 (0.49)	0.58 (0.49)

decreases for population-related displays for a choropleth map lineup but increases for a hexagon tile map display. The log odds of detection are shown in Table 2 can be back-transformed after taking the sum of all terms for the trend and type of display that are of interest. 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.

When a choropleth map display is used, the predicted detection rate for the Three Cities trend, 0.03; this is extremely low, especially compared to the NW-SE trend of 0.52. When the All Cities trend is presented in a choropleth display the predicted probability of detection is 0.22. The hexagon tile map has a substantially high detection rate for the display of a Three Cities trend 0.39 and All Cities trend 0.59.

4.3 Speed

Fig. 6 shows horizontally jittered dot plots to contrast the time taken by participants to evaluate each lineup when viewing each type of display. The time is also separated by the trend model and whether the data plot was detected or not detected. The time taken to complete an evaluation ranges from milliseconds to 60 seconds. The average time taken for the type of display is shown as a large colored dot on each plot. when considering the heights of the green and orange dots, there is little difference in the average time taken to read a choropleth or hexagon tile map. Comparing the same colored dot across each trend model row, there is a slight increase in the time taken to correctly detected the data plot in the hexagon tile map lineup, but little difference in evaluation time for the choropleth display. However, there were substantially fewer correct detections for choropleth lineups for the Three cities and All Cities trends.

Table 2: The model output for the generalized linear mixed effect model for detection rate. This model considers the type of display, the trend model hidden in the data plot, and accounts for contributor performance.

Term	Est.	Sig.	Std. Error	P val
Intercept	0.07		0.16	0.67
Hex.	0.46	*	0.22	0.04
Three Cities	-3.41	***	0.42	0.00
All Cities	-1.34	***	0.24	0.00
Hex:Three Cities	2.44	***	0.47	0.00
Hex:All Cities	1.16	***	0.33	0.00

Table 3: The number of participants that selected each reason for their choice of a plot when looking at each trend model shown in Choropleth and Hexagon Tile maps. The facets show whether or not the choice was correct.

Trend	Detected	Choro.	Hex.
NW-SE	No	trend	clusters
	Yes	trend	clusters
Three Cities	No	trend	clusters
	Yes	consistent	clusters
All Cities	No	trend	clusters
	Yes	clusters, consistent	clusters

4.4 Certainty

Participants provided their level of certainty regarding their choice using a five-point scale. Unlike the accuracy and speed of responses that were derived during the data processing phase, this was a subjective assessment by the participant prompted by the question: 'How certain are you about your choice?'. Fig. 7 shows the number of times participants provided each level of certainty. This was separated for each combination of trend models and display type, and colored depending on whether a participant correctly detected the data plot in the lineup. Participants often chose 4 or 5 when viewing the population-related trends in the choropelth display, even though they were often incorrect when viewing an All Cities trend and overwhelmingly incorrect for the Three Cities trend. This shows overconfidence in their detection ability when using a choropleth map display. Participants were less likely to be certain when their choice was incorrect and they were viewing a hexagon tile map. For each trend model, participants were more likely to doubt their choice and choose 1 or 2 in the hexagon tile map displays, even though many had made the correct choice.

4.5 Reason

Participants were asked why they had made their plot choice and were able to select from a set of suggested reasons. "Color trend across the areas" was the most common selection for NW-SE trend displays.

The reasons chosen by participants from the list provided to them varied more when viewing choropleth displays than the hexagon tile map. The hexagon tile map displays resulted in "Clusters of color" as the most common choice made by participants.

The choice "None of these reasons" was used as the default value to minimize noise from participants who did not select a response.

5 Discussion

The objective of this study was to compare the effectiveness of the choropleth map and the hexagon tile map in displaying population-related data. To achieve this, we employed the visual inference lineup

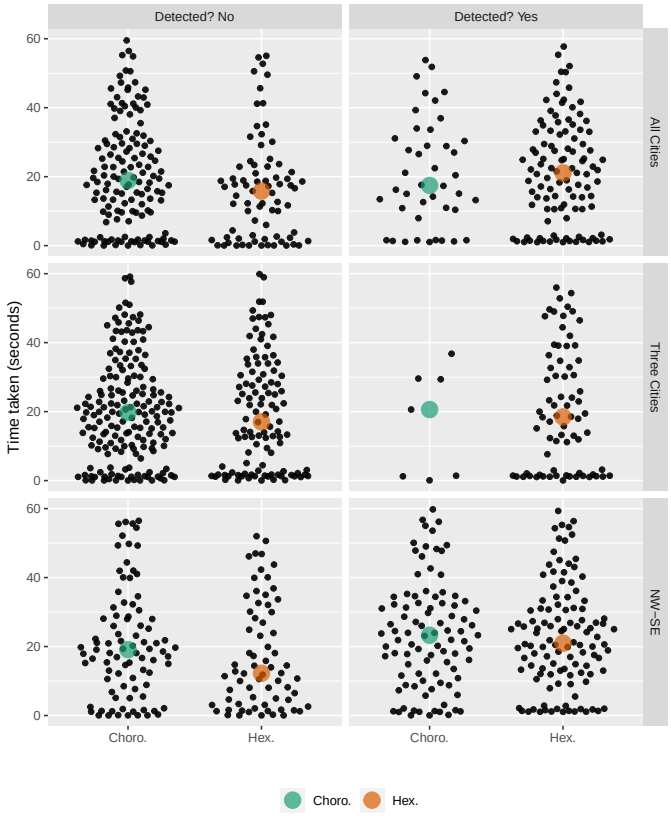

Fig. 6: The distribution of the time taken (seconds) to submit a response for each combination of trend, whether the data plot was detected, and type of display, shown using horizontally jittered dot plots. The colored point indicates the average time taken for each plot type. Although some participants take just a few seconds per evaluation, and some take as much as much as 60 seconds, there is very little difference in time taken between plot types.

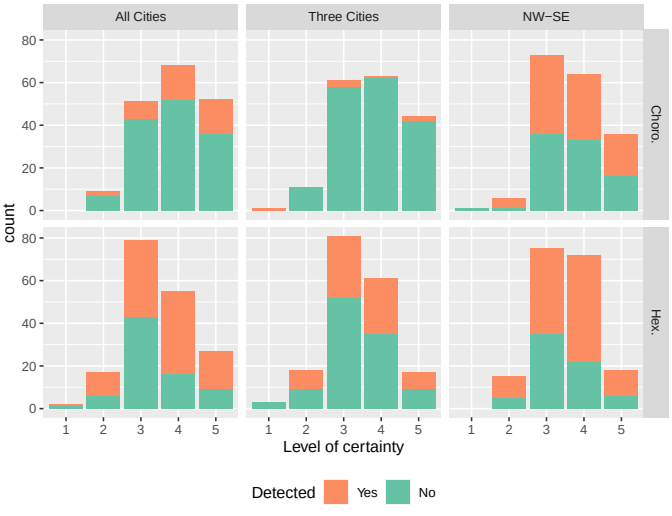

Fig. 7: The number of times each level of certainty was chosen by participants when viewing hexagon tile map or choropleth displays. Participants were more likely to choose a high certainty when considering a Choropleth map. The mid value of 3 was the default certainty, it was chosen most for the Hexagon tile map displays.

protocol, which helped us evaluate the efficacy of each display. Our results indicate that participants were more successful in identifying the data plot in the hexagon tile map display, making it a superior option for representing spatial distributions of population-related data. Thus, our study concludes that the hexagon tile map display is a valid and effective alternative to the choropleth map display for communicating this type of information.

Our initial expectation was that the hexagon tile map would be superior to the choropleth map for the two inner city hotspots pattern (All Cities, Three Cities), but that the choropleth would outperform the hexagon tile map for conveying a spatial dependence trend pattern across geography (NW-SE). We surprisingly found that the hexagon tile map was more effective in all three trend models. The detection rate was significantly higher for the hexagon tile map, although the margin was small for the NW-SE trend model. This was reflected in Fig. 5, the descriptive statistics in Table 5, and the linear model summary in Table 2.

While the significance of the difference in detection was the primary focus of this experiment, a secondary focus was the time taken by participants. It was expected that the participants may take longer to consider the hexagon tile map distribution but would be able to detect the data plot in the lineup. The bimodal distributions seen in Fig. 6 showed very little difference in the mean evaluation times. As the maximum time of all of the distributions approached 60 seconds it cannot be said that the participants took longer to evaluate the hexagon tile map displays.

The responses to the questions asked of participants included the reason for their choice and the certainty around their choice. Fig. 7 shows high levels of certainty of 4 and 5 were chosen by participants when looking at the population distributions in a choropleth map display showing that they were overconfident when attempting to find the real data plot in the choropleth map displays. Participants performed better on the NW-SE distribution shown in the choropleth display and were reasonably confident about their decisions. The high levels of the mid-range value of 3 could indicate that the participant did not want to provide a response, as this was the default value. Those who chose level 4 or 5 were equally likely to be correct for the three cities lineups, but more likely to be correct than incorrect for the other two trend models.

The color scaling applied in Three cities and All cities displays resulted in the rural areas of the real data plot appearing more blue or yellow than the other plots in the lineups. Due to the consistent coloring of rural areas in a choropleth display, the choice “All areas have similar colors” was the most common reason for a participant’s choice. The All Cities displays colored the inner-city areas of all capital cities redder, this was observable to participants and explains the equal choice of the city clusters or rural color consistency. Choosing “Clusters of color” was expected when participants viewed the Hexagon tile map display of the All Cities and Three Cities distributions. It was unexpected that it was also the most common reason for the NW-SE hexagon tile map displays. Due to the spatial covariance introduced in the smoothing, groups of similarly colored hexagons were present in all of the hexagon tile map displays. All Cities and Three Cities distributions of real data trends had distinctly different patterns or red inner-city areas, while some of the plots in each lineup may have shared similar features.

6 Acknowledgment

The authors would like to thank the Australian Cancer Atlas team for discussions regarding alternative spatial visualizations, and Professor Kerrie Mengersen and Dr. Earl Duncan for regular meetings filled with suggestions and comments. Mitchell O’Hara-Wild was a co-developer of the `taipan` [14] R package for image tagging, used as the base for the web app constructed to collect participant evaluations of lineups. We are thankful to the NUMBATs (Non-Uniform Monash Business Analytics Team) for participating in the pilot study that helped to assess the experimental design and determine an appropriate sample size for the study.

The source code to produce this document can be found at https://github.com/srkobakian/experiment/paper. Supplementary

materials have been included to discuss the survey procedures and the lineups that were used. The full set of images can be found here, too.

The supplementary material contains:

- Additional analysis of the experimental results
- Survey procedure including training materials for the participants
- 24 lineups as images, that were used in the experiment
- 12 data sets used to construct the lineups

The analysis of the work was completed in R [21] with the use of the following packages:

- For document creation: `rmarkdown` [32], `rticles` [1], `knitr` [31].
- For lineup creation and data analysis: `tidyverse` [27], `nullabor` [28], `ggthemes` [2], `RColorBrewer` [19].
- For image displays: `cowplot` [30], `png` [25], `grid` [18].
- For modelling and presentation of models: `gstat` [10], `lme4` [4], `kableExtra` [33].

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.

References

[1] J. Allaire, Y. Xie, R Foundation, H. Wickham, Journal of Statistical Software, R. Vaidyanathan, Association for Computing Machinery, C. Boettiger, Elsevier, K. Broman, K. Mueller, B. Quast, R. Pruim, B. Marwick, C. Wickham, O. Keyes, M. Yu, D. Emaasit, T. Onkelinx, A. Gasparini, M.-A. Desautels, D. Leutnant, MDPI, Taylor and Francis, O. Ögreden, D. Hance, D. Nüst, P. Uvesten, E. Campitelli, J. Muschelli, Z. N. Kamvar, N. Ross, and R. Cannoodt. *rticles: Article Formats for R Markdown*, 2019. R package version 0.13. 8

[2] J. B. Arnold. *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*, 2019. R package version 4.2.0. 8

[3] Australian Bureau of Statistics. Australian Statistical Geography Standard (ASGS), Jul 2018. 2

[4] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01 4, 8

[5] J. Bryan and J. Zhao. *googlesheets: Manage Google Spreadsheets from R*, 2018. R package version 0.3.0. 5

[6] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society, A (Invited)*, 367:4361–4383, 2009. doi: 10.1098/rsta.2009.0120 2

[7] Cancer Council Queensland. Australian Cancer Atlas, publisher = Queensland University of Technology, Cooperative Research Centre for Spatial Information, issue = Version 09, url=https://atlas.cancer.org.au, year = 2018, accessed = Jan 12 2020. 1

[8] D. Dorling. *Area Cartograms: Their Use and Creation*, vol. 59, pp. 252 – 260. University of East Anglia: Environmental Publications, 04 2011. doi: 10.1002/9780470979587.ch33 2

[9] Figure Eight Inc. The essential high-quality data annotation platform, 2019. 5

[10] B. Gräler, E. Pebesma, and G. Heuvelink. Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218, 2016. 3, 8

[11] H. Hofmann, L. Follett, M. Majumder, and D. Cook. Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18:2441–2448, 2012. 2

[12] S. Kobakian and D. Cook. *sugarbag: Create Tessellated Hexagon Maps*, 2019. https://srkobakian.github.io/sugarbag/, https://github.com/srkobakian/sugarbag. 1, 2, 4

[13] S. Kobakian, D. Cook, and J. Roberts. Mapping cancer: the potential of cartograms and alternative map displays. *Annals of Cancer Epidemiology*, 4(0), 2020. 2

[14] S. Kobakian and M. O'Hara-Wild. *taipan: Tool for Annotating Images in Preparation for Analysis*, 2018. R package version 0.1.2. 5, 7

[15] C. Kocmoud and D. House. A Constraint-based Approach to Constructing Continuous Cartograms. In *Proc. Symp. Spatial Data Handling*, pp. 236–246, 1998. 2

[16] G. Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963. 3

[17] M. Monmonier. Cartography: Distortions, World-views and Creative Solutions. *Progress in Human Geography*, 29(2):217–224, 2005. doi: 10.1191/0309132505ph540pr 2

[18] P. Murrell. The grid graphics package. *R News*, 2:14–19, 2002. https://journal.r-project.org/articles/RN-2002-010/. 8

[19] E. Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. R package version 1.1-2. 8

[20] J. M. Olson. Noncontiguous Area Cartograms. *The Professional Geographer*, 28(4):371–380, 1976. doi: 10.1111/j.0033-0124.1976.00371.x 2

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. 4, 8

[22] E. Raisz. Rectangular Statistical Cartograms of the World. *Journal of Geography*, 35:8–10, 1963. doi: 10.1080/00221343608987880 2

[23] A. Skowronnek. Beyond Choropleth Maps – A Review of Techniques to Visualize Quantitative Areal Geodata, 2016. 2

[24] E. R. Tufte. *Envisioning Information*. Graphics Press, 1990. 2

[25] S. Urbanek. *png: Read and write PNG images*, 2013. R package version 0.1-7. 8

[26] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. 2, 4

[27] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686 8

[28] H. Wickham, N. R. Chowdhury, D. Cook, and H. Hofmann. *nullabor: Tools for Graphical Inference*, 2018. R package version 0.3.5. 8

[29] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '10)*, 16(6):973–979, 2010. 2

[30] C. O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2019. R package version 1.0.0. 8

[31] Y. Xie. knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng, eds., *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595. 8

[32] Y. Xie, J. Allaire, and G. Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018. ISBN 9781138359338. 8

[33] H. Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2019. R package version 1.1.0. 8