New algorithms for effectively visualising Australian spatio-temporal disease data.

A thesis submitted in fulfilment of the

requirement for the degree of'
Master of Philosophy (Statistics)

by

Stephanie Rose Kobakian

B.Comm. and B.Eco., Monash University

School of Mathematical Sciences

Science and Engineering Faculty

Queensland University of Technology

Australia

2020

Contents

Copyright notice

Keywords

Keywords: cartogram; choropleth; geospatial statistics; maps; statistics; tile maps; visual inference

Abstract

Relationships between geographic areas can be communicated using maps full stop the speed of visual processes allow quick comparisons to be made between neighbouring areas. However, when presenting population-related statistics on the geographic map base, large areas that draw readers' attention can allow too much emphasis to be placed on rural areas. This problem can dramatically impact the interpretation of statistics across Australia, due to the large difference in the land area when the population is divided at various granularities.

As part of this thesis, an algorithm is presented which will take geospatial areas in the form of polygons and create an alternative graphical display of a spatial distribution. This algorithm takes a set of polygons and creates a map of tessellated hexagons, representing a single geographical area with a single hexagon. It arranges them to replicate spatial relationships of geographic areas in each city. The hexagon tile map visualisation produced by the algorithm is contrasted with the traditional choropleth map. The package `sugarbag` (Kobakian and Cook, 2019) implements the algorithm for the statistical software `R` (R Core Team, 2019a).

Using animations will allow us to control how people transform a recognisable map of Australia, or the cities within, into a more reliable map for inference. Animation is gaining popularity as access to computing power is increasing the amount of applications.

List of Figures

List of Tables

List of Publications

1. The literature review exploring current practice for visualising spatial data in Chapter 2 has been submitted to the journal *Annals of Cancer Epidemiology* for publication.

Kobakian S. and Cook, D. and Roberts, J. (2019). Cancer Applications of Choropleth Maps, and the Potential of Cartograms and Alternative Map Displays. Manuscript submitted for publication.

2. The details of the algorithm documented in Chapter 3 have been submitted to the *Journal of Statistical Software*.

Kobakian S. and Cook, D. (2020). An Algorithm For Spatial Mapping Using a Hexagon Tile Map, With Application to Australian Maps. Manuscript submitted for publication.

3. The details of the visual inference testing is documented in Chapter 4 has been submitted to the *IEEE Transactions of Visualisation and Computer Graphics* under the title "Comparing the Effectiveness of the Choropleth Map with a Hexagon Tile Map for Communicating Cancer"

Kobakian S. and Cook, D. (2020). Comparing the Effectiveness of the Choropleth Map with a Hexagon Tile Map for Communicating Cancer Manuscript submitted for publication.

4. The code for the algorithm documented in Chapter 3 is currently hosted on *CRAN* as the package sugarbag.

Kobakian, S. and Cook, D. (2019). sugarbag: Create Tessellated Hexagon Maps. `https://srkobakian.github.io/sugarbag/`, `https://github.com/srkobakian/sugarbag`.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three original papers submitted to peer reviewed journals. The core theme of the thesis is spatial visualisations. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the Faculty of Science and Engineering under the supervision of Distinguished Professor Kerrie Mengersen and Dr. Earl Duncan. It was also created under the supervision of the external supervisor Professor Dianne Cook.

The papers in Chapters 2, 3, and 4 have been individually formatted for journal submission however, I have renumbered the pages of the submitted papers for cohesion across this thesis.

Acknowledgements

This thesis would not be possible without the opportunity provided by Queensland University of Technology, Cancer Council Queensland and the ARC Centre of Excellence for Mathematical & Statistical Frontiers.

I would like to acknowledge my supervisors for their tireless work in directing, encouraging and supporting this work.

Professor Kerrie Mengersen (Professor of Statistics, Science and Engineering Faculty, QUT) provided the opportunity to study at Queensland University of Technology, organising and supervising this project.

Dr Earl Duncan (Research Associate at ARC Centre of Excellence for Mathematical & Statistical Frontiers, QUT) provided time and effort for to discussion, editing and commenting throughout this degree.

Professor Dianne Cook (Professor, Department of Econometrics and Business Statistics, Monash University) provided constant support, encouragement and recommendations throughout this degree. Working closely with Prof.. Di Cook enabled the rapid development of my research and writing skills.

0.1 Literature review

We would like to acknowledge Dr Susanna Cramb (Spatial Modeller, Cancer Council Queensland) and Dr Peter Baade (Senior Research Fellow, Cancer Council Queensland) for providing the opportunity and time to discuss alternative map solutions for presentation

in the Australian Cancer Atlas. It was in the development of this online cancer atlas that methods for disease map displays, and visual communication strategies were explored.

0.2 Visual inference study

Stephanie would like to thank Mitchell O'Hara-Wild was a co-developer of the taipan (Kobakian and O'Hara-Wild, 2018) R package for the web app constructed to collect participant evaluations of lineups.

We are thankful for 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.

0.3 R packages

Several `R` (R Core Team, 2019a) packages were used to produce this thesis:

- `absmapsdata` (Mackey, W. F., 2019)
- `cartogram` (Jeworutzki, 2018)
- `cowplot` (Wilke, 2019)
- `eechidna` (Cook et al., 2019)
- `ggplot2` (Wickham, 2016)
- `ggthemes` (Arnold, 2019)
- `grid` (R Core Team, 2019b)
- `gstat` (Pebesma and Graeler, 2019)
- `kableExtra` (Zhu, 2019).
- `knitr` (Xie, 2014)
- `lme4` (Bates et al., 2019)
- `nullabor` (Wickham et al., 2018)
- `png` (Urbanek, 2013)
- `RColorBrewer` (Neuwirth, 2014)
- `rmarkdown` (Allaire et al., 2019a)
- `rticles` (Allaire et al., 2019b)

- `sf` (Pebesma, 2018)
- `spData` (Bivand, Nowosad, and Lovelace, 2019)
- `sugarbag` (Kobakian and Cook, 2019)
- `tidyverse` (Wickham, 2017)

Chapter 1

Introduction

There are many visualisation methods used to present geospatial data. The design of the visualisation chosen can hinder or improve the communication of the spatial distribution. A choropleth map is the most common display used to present geographical data. Maps contribute to understanding spatial distributions of disease occurrence, and locating disease clusters. Disease data is often aggregated by political areas. One reason for this is privacy and another is the responsibility on the political entity to respond. The typical visualisation for aggregated spatial data is a choropleth map, where areas are coloured by the numerical value.

Choropleth maps do a disservice to the map reader, as the attention of the map user is distributed according to the size of the area. Using a choropleth map to get a broad perspective of Australia can be misleading, when the use of geographical areas misrepresents the spatial distribution of a dataset. This is not practical if each area is considered equally important. In Australia, the population is not equally dispersed across the geographic map base. Instead, the communities are densely populated in the inner city areas, especially around the capital cities. There are several visualisation methods that have been developed to emphasise the population dense areas. These alternatives should be considered when planning the communication of geospatial statistics, visualisations should be chosen to best represent the spatial distribution. The work is motivated by the Australian Cancer Atlas, which presents the spatial patterns of many cancers in Australia.

The aim of this thesis is to contribute an algorithm that creates effective visualisations for the communication of geospatial population statistics.

1.1 The Australian Cancer Atlas

This work was motivated by the Australian Cancer Atlas (ACA), an online, interactive web tool for exploring the impact of cancer on Australian communities. The prominent display used by the ACA is a visualisation of incidence rates or excess death rates. The set of geographic units used is Australian Statistical Areas, at Level 2 (SA2s). There are almost 2,200 individual SA2s.

The choropleth map used in the ACA is familiar to the general public of Australia. It is appropriate to use this display as users can orient themselves on the map base and find geographic areas relevant to them. However, when the intention of the map user is to convey the whole spatial distribution the information derived visually from the colours can be misleading. The rural areas are over emphasised, and the densely populated inner city areas are not given enough attention.

1.2 Visual Inference

Visual inference will be used to determine if the communication of population geospatial statistics is more effective when using an alternative display. Buja et al (Wickham et al., 2010) provide the 'lineup' protocol as a formal framework for testing visual statistical methods. Implementing this framework will allow new alternative visualisation method to be tested.

The lineup protocol will be used to test if a visualisation is effective, a visualisation displaying a real population based distribution can be hidden in a collection of visualisations that display null distributions (Roy Chowdhury, 2014). It takes inspiration from a police lineup. The witness in this regard is a participant recruited from a crowdsource platform, such as Figure-Eight. The visualisation containing a real distribution is considered the 'accused'. It is put in a lineup of innocent displays that do not show a real population based distribution. If the 'witness' chooses the 'accused' as different from the innocent plots, it can be considered that there is a specific pattern displayed that is not present in the others. In this protocol, the null hypothesis can be rejected in favour of the alternative when it is chosen in the lineup. The null hypothesis fails to be rejected when it is not selected in the lineup.

1.3 Aims and Objectives

This work aims to provide a solution to presenting geospatial data regarding populations. It considers the visualisation methods developed over the past two centuries that shift the focus from the geographic map base.

1. *Devising an Algorithm for creating hexagon tile maps of Australia:* The algorithm will take geospatial areas and create an alternative visualisation of the spatial distribution.

2. *Test the effectiveness of the hexagon tile map relative to the choropleth map:* The hexagon tile map produced by the algorithm will be contrasted with the traditional choropleth map, applying the same colour methods to represent the data. The maps will be used in an experiment to test the effectiveness by asking for users to spot spatial distributions.

3. *Communicating the relationship between the hexmap and choropleth map through animation:* Maximise the benefits of both displays when communicating to the public. The use of animations may control how people follow a recognisable map of Australia into an alternative visualisation for inference.

1.4 Research Contributions

This research contributes a new algorithm for creating hexagon tile map displays. It contributes an R (R Core Team, 2019a) package which implements the algorithm and allows R users to create their own visualisations. It presents a case study that contributes to a growing field of visual inference studies, applied to spatial data by comparing a choropleth map to a hexagon tile map display. It also shows how it can be used in practice to effectively communicate cancer distributions.

1.5 Thesis Structure

The thesis is structured as follows: Chapter two contains a literature review. The literature reviews considers the current peer reviewed literature and published books that explore spatial distributions of cancer across the globe. It also considers how to evaluate the visualisation methods used for spatial data.

Chapter three explores the algorithm to create hexagon tile maps and the code used to create a small example of Tasmania in Australia. Chapter four is a visual inference study that contains the methods and results that compare the use of a choropleth map and a hexagon tile map on the same data sets. Chapter five provides a conclusion of the results of the visual inference study and how the hexagon tile map may be used in practice.

Chapter 2

Literature Review

The following chapters in this thesis each contain an introductory section to introduce the relevant literature. This literature review chapter provides an overview of the key areas of interest in the literature that are relevant to this research. This chapter is organised as follows. Section 1 outlines the traditional spatial mapping technique, the choropleth map, and provides design inspiration using examples of online cancer atlases. Section 2 outlines contemporary mapping approaches, suggested as alternatives to the choropleth. Section 3 compares and critiques these alternative displays in light of the strengths and weaknesses of the choropleth method. section 4 considers how users interact with mapping displays in online cancer atlases, and how map creators can direct the attention of uses through animation.

This chapter was submitted for publication to the journal *Annals of Cancer Epidemiology* for publication. This was intended for an audience of cancer atlas creators to be inspired by current atlases, and to encourage the pursuit of alternative displays.

RESEARCH STUDENTS CENTRE
Examination Enquiries: 07 3138 1839
Email: research.examination@qut.edu.au

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

Cancer Applications of Choropleth Maps, and the Potential of Cartograms and Alternative Map Displays has been submitted to the journal Annals of Cancer Epidemiology for the focused issue -- Spatial Patterns in Cancer Epidemiology.

Contributor	Statement of contribution*
Stephanie Kobakian S R Kobakian 27/12/2019	Stephanie researched current methods for presenting geospatial data, wrote the initial draft and revised the drafts after suggestions were made by reviewers.
Prof. Dianne Cook	Prof. Dianne informed the structure of the review and provided heavy editing.
Jessie Roberts	Jessie researched and collated current web atlas examples.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all Co-authors confirming their certifying authorship. (If the Co-authors are not able to sign the form please forward their email or other correspondence confirming the certifying authorship to the RSC).

Name Signature Date

Cancer Applications of Choropleth Maps, and the Potential of Cartograms and Alternative Map Displays

Stephanie Kobakian[*], Dianne Cook[†] and Jessie Roberts[‡]

Abstract

Cancer atlases communicate cancer statistics over geographic domains, typically with a choropleth map. They subdivide these domains into administrative regions such as countries, states, or suburbs. When communicating human-related statistics, the choropleth has a disadvantage in that it draws attention to sparsely populated rural areas to the neglect of small inner city areas. The smaller geographic areas are important to consider if they are densely populated. Alternative map displays, such as a cartogram or a hexagon tile map, can shift the attention of map users from the large rural areas by decreasing their size on the map display. This means alternative displays can be more effective at accurately communicating spatial patterns across spatial areas. It is recommended that alternative displays are included in cancer atlases. In addition, with the ease of today's technology, user interaction with the displays is encouraged. Users should also be able to interactively display different statistics, such as incidence rate or relative incidence, or filtered by demographic variables.

[*]Science and Engineering Faculty, Queensland University of Technology.
[†]Faculty of Econometrics and Business Statistics, Monash University.
[‡]Queensland University of Technology.

1 Introduction

Researchers, health authorities, governments, not-for-profits and the media are common communicators of cancer statistics. They often present statistics to the public as aggregated values for geopolitical areas. Presenting these statistics requires aggregating individual observations for the geographical units, especially for privacy protection, but also for political and policy purposes. Examples of typical geographical units include states, provinces, local government areas, and post/zip codes. It is easy to provide counts or incidence rates of the diagnoses of these areas. This type of data is routinely collected for public health reasons and may be made available to the general public as a service to the community.

To visualize and communicate geospatial cancer statistics over geographic domain, a choropleth map is the common display. Choropleth maps show polygons representing the geographic units, where each polygon is shaded with a color according to the area-specific values of the statistic being conveyed. Visualizing this data is helpful as geographic patterns of disease may be obscured when reported in a table [1]. Providing a visual representation of cancer outcomes allows identification of geographic patterns of the disease that can then be addressed with public health policy and actions. The spatial distribution of the disease incidence can be examined using a choropleth and may reveal a trend in longitude or latitude, or rural vs urban, or coastal vs inland, or even specific hot spots of the disease. One of the key challenges with mapping spatial patterns of disease is the design of visualizations [2]. It is important to consider the strengths and weaknesses of designs, as visualizing diseases on maps is often the first step in exploratory spatial data analysis and helps in the formulation of hypotheses. This paper considers the current visualization techniques to communicate statistics to the public and their applications to cancer statistics. Alternative approaches are posed because they may be more effective than contemporary techniques. The limitations of the visualization methods, highlighting the differences and historic use of these displays is discussed.

The paper is structured as follows. The next section describes the choropleth map, which is the common approach to disease maps and presents examples of atlases in use today and discusses the limitations of the choropleth map. Section 3 describes alternative displays, including the cartogram, which is useful when the map has heterogeneously sized geographic units. Section 4 presents the limitations of the production and use of alternative displays. Disease maps are more useful when made interactive, and common options are described in Section 5, along with a discussion of benefits and disadvantages.

2 Traditional approaches for cancer map displays

A choropleth map displays the geographical distribution of data over a set of spatial units by shading areas of a map [3], [4]. Faithful rendering of the geography, when combined with an appropriate color scheme, can reveal spatial patterns among data values. Identifying and explaining spatial structures, patterns, and processes involve considering the individuals and organizing them into representable units of communities [1]. Early versions of choropleth maps used symbols or patterns instead of color. Choropleth maps can be used for displaying disease data [5], including cancer data [6]. In epidemiology, choropleth maps are often used as a tool to study the spatial distribution of cancer incidence and mortality.

Displaying familiar state boundaries can make a map easier to read [7] and allow viewers to infer the spatial relationships visually in the data using their mental model of the geography. The map users of disease displays may include researchers, the public, policymakers, and the media [6]. For these users, the familiarity of the geography is a worthy consideration when presenting results of spatial analysis.

2.1 Cancer atlases

A cancer atlas is a map, or collection of maps, representing cancer incidence and mortality for a country, or group of countries. Atlases are key to developing hypotheses regarding areas with unusually high rates, and geographic correlations [8]. The data collection methods across regions and the administrative control within regions lends itself to choropleth visualization. Cancer maps and atlases date back to Haviland's maps in 1875, and early work in US cancer atlases appearing in 1971 [9]. The presentation of cancer statistics has increased with greater access to computational power and the availability of geographic information systems software [2].

Cancer maps are effective tools for communicating incidence, survival, and mortality to a wide range of audiences, including the public and others not trained in statistical analysis. These visualizations enable non-expert audiences to interpret the outputs of sophisticated statistical analysis. Cruickshank (1947) as cited by S. D. Walter [5], discusses using visuals as a 'formal statistical assessment of the spatial pattern'. Overwhelmingly, choropleth maps are visualisations chosen to communicate cancer statistics to members of the public and other non-expert audiences.

Table 1: A selection of choropleth cancer maps from online atlases.

Fig.	Atlas	Statistic	Data source
1a	The Environment and Health Atlas of England and Wales	relative risk for women developing lung cancer in England and Wales in 2010 [10]	Office for National Statistics (ONS) (England) and from the Welsh Cancer Intelligence and Surveillance Unit (WCISU)
1b	Globocan 2018: Estimated Cancer Incidence, Mortality and Prevalence Worldwide	age standardized incidence rates (per 100,000) for all invasive cancers for both men and women, aggregated at a national level for 2018 [11]	World Health Organization's International Agency for Research on Cancer.

Fig.	Atlas	Statistic	Data source
1c	Atlas of Cancer in Queensland	the relative incidence ratio of lung cancer in males in the state of QLD within Australia based on data from 1998 to 2007, Queensland Cancer Council[12]	Queensland Cancer Registry
1d	Bowel Cancer Australia Atlas	the percentage of Australian males between 50 - 54 years of age diagnosed with bowel cancer in 2016 in Australia [13].	Bowel Cancer Australia
1e	United States Cancer Statistics: An Interactive Cancer Statistics Website	the incidence rate per 100,000, of all cancer types for men and women in the United States in 2016, aggregated at the state level [14].	*Centers for Disease Control and Prevention*, with data from state cancer registries.
1f	Map of Cancer Mortality Rates in Spain	side by side maps of relative risk of lung cancer for men vs women for 2004 to 2008 [15].	Map of cancer mortality rates in Spain
1g	Atlas of Childhood Cancer in Ontario	the incidence rate of childhood cancers per 100,000 (by census division) for children aged 0-14, in Ontario from 1995 to 2004 [16].	The Pediatric Oncology Group of Ontario Networked Information System

Epidemiologists and statisticians have developed the statistics used to communicate the burden of cancer over several decades. Table 2 summarizes the measures commonly presented in published cancer atlases. Mortality rates are commonly presented as relative rates of risk across the population and age-adjusted to correct for the higher prevalence of cancers in

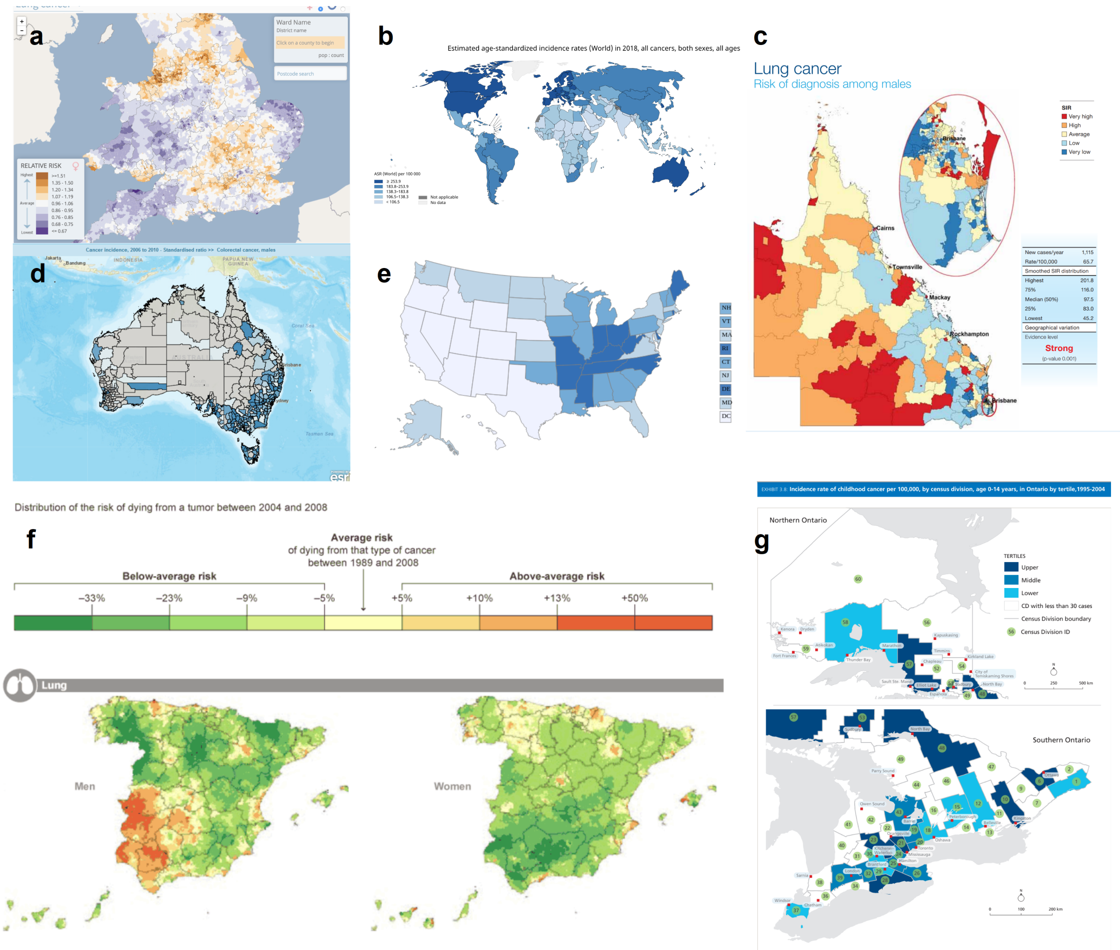


Figure 1: A selection of choropleth cancer maps from online atlases that are publicly available. Maps of various countries were chosen: United Kingdom, Australia, Spain, USA, Canada, and display several different colour palettes and legends. These atllases are described in Table 1.

older populations. As described in Howe [17], Englishman P. Stocks advanced the field of mortality statistics by introducing the standardized mortality ratios in the 1930s, which is an improvement on crude death rates.

Table 2: Common measures for reporting cancer information.

Measure	Details
1. Count	Crude cancer counts
2. Rate per 100,000	Cancer incidence per 100,000 population
3. IR (Incidence Ratio)	$(IR)_i = \frac{(Incidence\ Rate)_i}{Average\ Incidence\ Rate}$, The cancer incidence rate in region i over the average cancer incidence rate for all of the regions
4. Age-Adjusted Rate per 100,000	Standardized by age structure or region
5. Age-Adjusted Relative Risk	Standardized by age structure in each region i
6. SIR (Standardized Incidence Ratio)	Incidence standardized by population at risk in each region i
7. Below or above Expected	An alternative expression of the SIR
8. RER (Relative Excess Risk)	$RER = \frac{(Cancer\ related\ mortality)_i}{Average\ cancer\ related\ mortality}$ Represents the estimate of cancer-related mortality within five years of diagnosis. Also referred to as 'excess hazard ratio'

Roberts [18] identified 33 cancer atlases published between 2010 and 2018. Each of these online atlases uses choropleth maps. All except one of these were published by non-commercial organizations, including not-for-profits, government, research organizations, advocacy groups or government-funded partnerships. Figure 1 displays a subset of maps from these atlases, the selection varies in the geographies explored. Figure 1b shows Globocan 2018 [11] which explores Estimated Cancer Incidence, Mortality and Prevalence Worldwide using data sourced from cancer registries of each country. The Bowel Cancer Australia Atlas in Figure 1d presents an example of a cancer specific atlas – it shows the average Standardized Incidence Ratio of colorectal cancer for Australian males from 2006 to 2010 [13]. Like many of the atlases examined, there is a choice of gender displayed in the Bowel Cancer Atlas. Gender is displayed in side-by-side maps in the Map of Cancer Mortality Rates in Spain (Figure 1f) [15].

Resolution of the maps varies greatly. Figure 1b shows global information at a national level. The United States Cancer Statistics [14] shows data aggregated at the state level. The Environment and Health Atlas of England and Wales [10] (Figure 1a) shows the relative risk for women developing lung cancer at a neighborhood (small-area) scale. The Atlas of Cancer in Queensland (Figure 1c) shows the relative incidence ratio of lung cancer in males for each Statistical Area at Level 2 [19] in the state of Queensland within Australia [12].

Age-specific atlases are less common. Figure 1g displays Atlas of Childhood Cancer in Ontario, this communicates the incidence rate of childhood cancers per 100,000 (by census division) for children aged 0-14, in Ontario from 1995 to 2004 [16].

2.2 Additional considerations

Cancer atlases often display supplementary graphs and plots to add more information. Additional materials such as tables, graphs, and text explanations support understanding and inference derived from maps, ensuring the message communicated will be consistent across a range of viewers [6]. The many displays of statistical summaries, including dot plots, bar plots, box plots, cumulative distribution plots, scatter plots, and normal probability plots, can provide alternative views of the cancer statistics. These can also display supporting statistics such as error, confidence intervals, distributions, sample or population sizes, and standard deviation.

The statistics communicated in atlases are often used to describe differences between areas.

This can occur at different levels of aggregation. Aggregation of global health statistics occurs within administrative and arbitrarily defined regions, such as those used by the World Health Organization and the United Nations [20]. World atlases can allow for displays of data aggregated into continents, countries, states, provinces and congressional districts [14]. Each population area will probably have a different number of people, which is typically used to calibrate the statistic. Cancer atlases may also communicate the distribution of the population living in all areas in a table or histogram display [21]. Atlases can connect the population to the land available to them by communicating population density.

Maps can also be used to focus on demographic strata, such as age and sex. Some of the digital atlases surveyed allow subsets such as males, females, or those aged over 65, to be selected for display. Similarly, socioeconomic indicators, such as unemployment rates, poverty rates, remoteness, and education levels, can be used to filter data, in order to communicate how cancer prevalence varies for different members of society. Few atlases provide this level of detail.

Introducing population and demographic information helps to interpret the rates in areas effectively, but there will still be uncertainty around the rates. To address this, a cancer atlas often communicates uncertainty about the value of a statistic. There are several potential sources of uncertainty: sampling error, errors arising from the disease reporting process (or data collection), and errors arising from the statistical modelling or simulation process. The most common measures used to present uncertainty are credible or confidence intervals (CIs). Displaying the uncertainty associated with reported statistics is a vital feature of a cancer map, but it is difficult to display effectively. The map focuses on displaying the statistic and lacks additional space to represent the uncertainty. Providing an adjacent map or overlaying maps with symbols [22] are two common solutions.

2.3 Limitations of choropleth displays

Australia presents an extreme case of an urban rural divide. The land mass occupied by urban electoral districts is only 10% of Australia, yet 90% of the population live in these urban areas

[23].

Choropleth maps provide a familiar display, which shows data in a geographically recognisable way. A disadvantage is that the different population and geographical sizes of administrative areas can attract attention to the shades of the underpopulated but large areas [3]. Skowronnek also [4] discusses how choropleth maps suffer from area-size bias, as they give a 'stronger visual weight' to large administrative units. The administrative boundaries used to define regions may limit a choropleth display, as this display unfaithfully represents the disease distribution across the region by obscuring small geographic areas. Sparsely populated rural areas are emphasized, whereas the areas representing inner city communities are very small. This is especially true for Australia.

Choropleth maps colour each geographic unit to allow map users to measure the value of the statistic [3]. Map users contrast the colours in neighbouring areas to understand the spatial distribution. The ColorBrewer system [24] and viridis [25] palettes provide effective colour schemes for qualitative, sequential and diverging data. When communicating information using colour, a map creator should use a scheme that has a linear color gradient, with perceptually uniform color spaces that match equal steps in data space with equal steps in the colour space [26]. The use of borders and backgrounds, and their colours, can also change the appearance of the colors representing the value of the statistics [24]. These supports can be used to implement a reference point in the colour scheme as well as orient users to the geographic regions.

Inset maps like in Brisbane city in Figure 1c of the state of Queensland are commonly used to reduce distorted interpretations, but it is a bandaid remedy. For Australia, many, many inset maps would be needed.

3 Contemporary alternatives to choropleth maps

3.1 Cartograms

Choropleth maps imply uniformity of data across the geographic space but population densities are unlikely to be uniform [4]. Cartographers developed the cartogram to draw the attention to the population by transforming the map [27]. The resulting display can communicate the impact of the disease more accurately across the population, as recorded by the statistic, at the sacrifice of geographic accuracy.

When a map creator desires a uniform population density of the map base, the purposeful distortion of the map space is beneficial. The “population distribution is often extremely uneven”, making a distortion necessary so that population is more faithfully represented as a uniformly distributed background for the statistic to be presented [23] [28] [29]. An area cartogram [30], or population-by-area cartogram [31] is produced from the distortion of the geographical shape according to population. Event cartograms [22] change the area of regions on a map depending on the amount of disease-related events, rather than population.

Cartograms provide an alternative visualization method for statistical and geographical information. Monmonier [32] suggests that map creators can use white lies to create useful spatial displays. It is easy for the reader to disregard the impact of transformations used to create cartograms, for the benefit of reading the statistical distribution more accurately with approximate geographic information. The spatial transformation of map regions relative to the data emphasizes the data distribution instead of land size [33]. When visualizing population statistics, Dorling considers this design ‘more socially just’ [23], or honest [34], giving equitable representation and attention to all members of the population and reducing the visual impact of large areas with small populations [5]. Howe [17] suggests that ‘cancer occurs in people, not in geographical areas’ and that spatial socio-economic data, like cancer rates, are best presented on a cartogram for urban areas as the population map base avoids allocating ‘undue prominence’ to rural areas [28].

The creation of cartograms was historically in the hands of professional cartographers [35]. Early approaches by John Hunter and Jonathan Young (1968) and Durham's wooden tile method, Skoda and Robertson's (1972) steel ball-bearing approach and Tobler's (1973) computer programs [23]. Howe [17] discusses the impact of electronic computer-assisted techniques. Geographical information systems allow map creators to produce cartograms and they use these systems depending on 'the effectiveness, efficiency, and satisfaction of the map products' [35].

There are two key issues to consider when creating alternative map displays, (1) the intended audience of the map, and (2) its purpose. Nusrat and Kobourov [36] provided a framework to investigate implementations of the many algorithms presented, and the "statistical accuracy, geographical accuracy, and topological accuracy".

Table 3: Maps used to present statistics for the United States of America. The colour of each state communicates the average age-adjusted rate of incidence for lung and bronchus for females and males in the United States 2012-2016.

Map display	Details
a. Contiguous	It has distorted each state's shape according to the population of the state in 2015. The state of California has become much larger because of the large population density. This draws attention to the densely populated North-East region and detracts from the less populated Mid West.
b. Non-contiguous	It maintains the geographic shape of the states, but the size has altered according to the population of the state in 2015. The state of California has remained closer to its original size than its surrounding states. The North-East states have remained closer to their geographical size, for Massachusetts and Connecticut. This draws attention to the densely populated North-East region and the sparse Mid West.

Map display	Details
c. Dorling	Circles are used to represent each state, but the population of the state determines the size in 2015. The North-East states remain closer to their neighbors and are slightly displaced from their geographic location. It highlights the sparsity of the population in the Mid West by the distance between the circles at the geographic centroids.
d. Hexagon Tessellation	A hexagon of equal size represents each state. It is easy to contrast the neighboring states however the North-East regions have been displaced from their geographic location. It highlights the sparsity of the population in the Mid West by the light yellow color, the Age-Adjusted rate in Kentucky is the darkest and its neighbors are similar.

Figure 2 shows four different cartograms for the same data. The information in Table 3 summarizes what can be observed in the four types of cartograms.

3.1.1 Contiguous

A contiguous cartogram alters the choropleth according to a statistic and maintains connectivity of the map regions. Min Ouyang and Revesz [37] present three algorithms for creating value-by-area cartograms. They implement 'map deformation' to account for the value assigned to each area. Other methods include Tobler's Pseudo-Cartogram Method, Dorling's Cellular Automaton Method [23], Radial Expansion Method, Rubber Sheet Method, Line Integral Method, Constraint-Based Method [33].

Figure 2a shows a population contiguous cartogram of the United States. All states are visible and the shape of the United States overall is still recognizable. In contrast, Figure 3 a shows an Australian contiguous cartogram also based on population. The south east is enlarged,

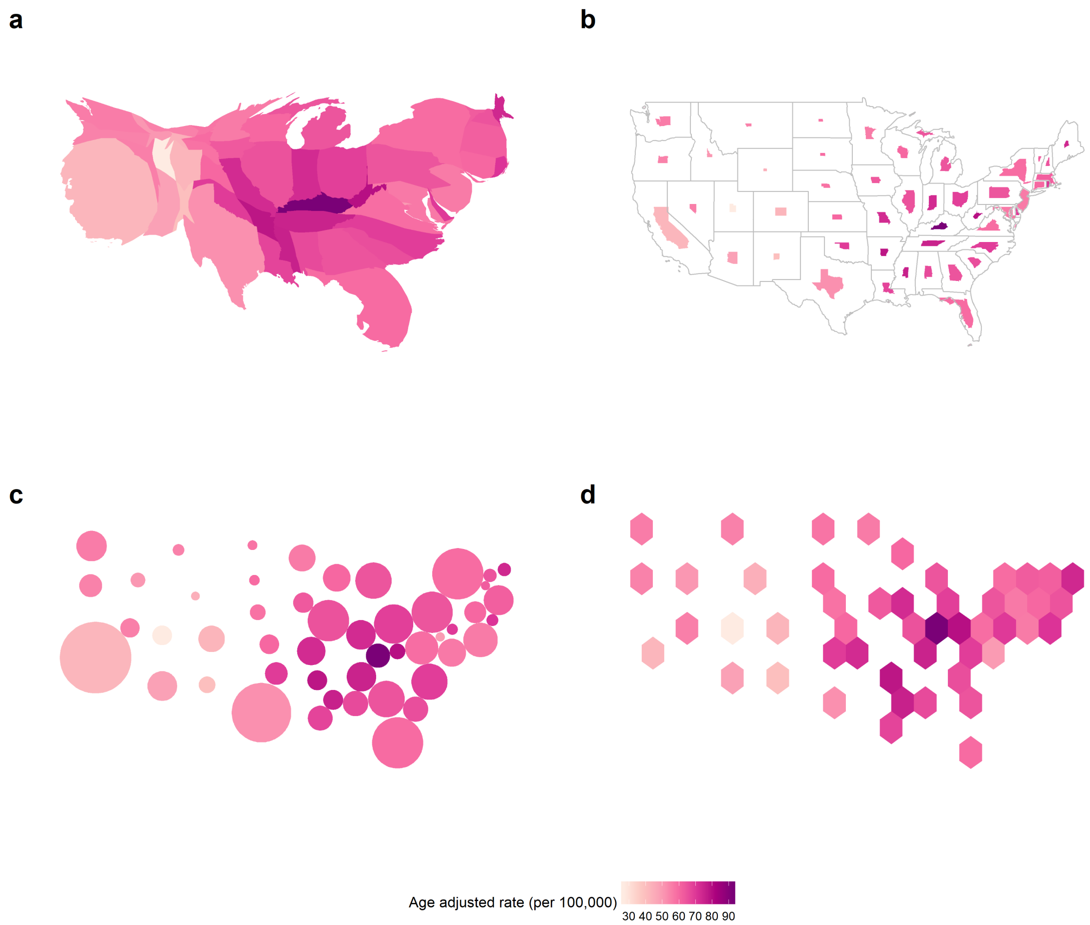


Figure 2: Common alternatives to maps, showing the same information for the United States of America: (a) contiguous cartogram, (b) non-contiguous, shape-preserved cartogram, (c) Dorling cartogram (non-contiguous), (d) hexagon tilemap (non-contiguous). Maps (a) - (c) are created by resizing and reshaping the states of the USA to match the 2015 population of the state. This provides a better sense of the extent of disease relative to the population in the country and can help ease losing information about physically small but population-dense states. Map creators give each state equal size and thus equal emphasis in (d) the hexagon tile map.

but high population areas are still small, and low population areas are still large on the map. The algorithm doesn't fully reach an optimal configuration where area matches population – Australia is too heterogeneous for the algorithm to handle.

To be able to recognize the significant changes, a reader will usually have to know the initial geography to find the differences in the new cartogram layout [30]. The shapes of small areas on a choropleth map and a cartogram are preserved using Tobler's Conformal mapping method. Koccmoud and House [33] present this issue as conflicting tasks or aims, to adjust region sizes and retain region shapes.

3.1.2 Non-contiguous

Non-contiguous cartograms prioritize the shapes of the areas instead of connectivity. Each area stays in a similar position to its location on a choropleth map. Displaying the choropleth map base allows map users to make comparisons regarding the change in the area. The addition is the gap between areas, created as each area shrinks or grows according to the associated value of the statistic. Olson [30] discusses the creation of these maps and the significance of the empty areas left between the geographic boundaries and the new shape.

The white space presents the meaningful empty-space property [38] [30] but it also distracts the reader from the data, with a low data density [39].

3.1.3 Dorling

Daniel Dorling presents an alternative display engineered to highlight the spatial distribution and neighborhood relationships without complex distortions of borders and boundaries [23]:

> "If, for instance, it is desirable that areas on a map have boundaries which are as simple as possible, why not draw the areas as simple shapes in the first place?"

He acknowledged the sophistication of contiguous cartograms but critiqued their 'very complex shapes,' he answers this with his implementation of maps created using 'the simplest of all shapes'. Circular cartograms use the same circle shape for every region represented, resized according to the statistic represented or the population. This simple shape may be more effective for understanding the spatial distribution than contiguous cartograms. Contiguous cartograms create 'nonsense' shapes that have 'no meaning' [34]. Both methods applies a gravity model to produce a layout, that avoids overlaps and keep spatial relationships with neighboring areas over many iterations. The circular cartogram is relatively fast to compute.

Raisz [40] laid the groundwork for this approach in the mid-1930s, drawing rectangular cartograms that provide simple comparisons, effective for correcting misconceptions communicated by geographic maps. Tobler [41] names and defines these as Value-Area Cartograms. This rectangular display may sacrifice contiguity but allows for tiling where geographic neighbors placed in suitable relative positions also share borders [42]. Rectangular cartograms communicate bivariate displays of the population by the size of each rectangular, and they use color to communicate a second variable [43].

3.2 Tile Map

A tile map provides a tessellated display of consistent shapes. A similar method to a rectangular cartogram, represents each geographic area using a square. The squares are tessellated to create a grid. Each area is represented by a square of the same dimensions, each tile is usually one unit of measurement, this could be geographic regions such as states or population-based that use a consistent measure of population for each tile. Regions with over four neighbors require some necessary displacement. The tile map uses color to represent a value of a statistic for each area. A similar method to a rectangular cartogram represents each geographic area using a square of the same dimensions. There are online media sources using this method, these include [44], [45], [46], [47]. Tile maps may be difficult to create as they are best created manually, they require additional time and care as the number of geographic areas to include increases.

Cano and others [48] define the term 'mosaic cartograms' for hexagonal tile displays, where the number of tiles for each area or the color of them can communicate the statistic of regions. When using several tiles per region, map makers can adjust the complexity of the boundaries in the resulting display. They can also make a trade-off between boundary complexity and simplicity by the size of the tiles used.

3.3 Geofacet

Hafen [49] introduces the term geofacet to describe a grid display of small plots. The arrangement of tiles mimics the geographic topology. Geofaceting has the functionality that a statistical plot can be constructed in each facet for each geographic area. A tile map can communicate only one value per region in a visualization, while geofaceting is a more flexible visualization for communication as it increases the amount of information displayed. Virtually any type of plot can be shown in the tile, allowing displays of multiple variables or values per geographic entity. Creating the layout of a geofacet is manual, but once created can be used for any data on that geographic base.

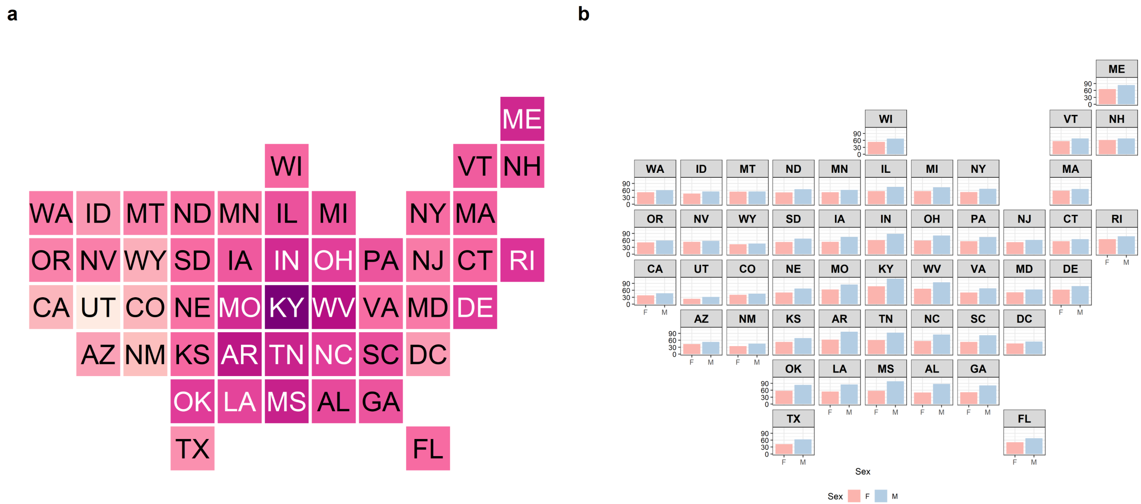


Figure 3: Two alternative displays, tile map (left) and geofaceted map (right), showing state age-adjusted rate of incidence for lung and bronchus in the USA. In the tile map, the layout approximates spatial location, with each state being an equal box filled with color representing cancer incidence. The geo-faceted map shows bar charts laid out in a grid approximating the spatial location of the state. The maps show age-adjusted rates for males and females. This display allows the presentation of multiple variables for each geographic area.

3.4 Multivariate displays

Pickle and others [50] present linked micromap plots to match geographic and statistical data visually, this serves as a solution to multi-dimensionality issues. These maps group areas based on their value for one variable, and additional columns provide displays that contrast the areas in each group by other variables. The display juxtaposes choropleth maps and statistical plots; it shows one map per group of the key separating variable, in a row with each additional statistical plot. Linked micromaps predominantly use the choropleth map for displays of spatial relationships. These maps show spatial relationships by allotting spatial neighbors to the same group. It is one of several alternative displays that allow maps to become bivariate displays, commonly used to present both an estimate and the associated uncertainty.

Lucchesi and Wikle [51] present bivariate choropleth maps blend color schemes to convey the intersection of categorized levels of an estimate and the associated uncertainty for each spatial area. They also suggest map pixilation, which breaks each region into small pixels, and allocates values to the individual pixels to create texture. This reflects the uncertainty around the area's estimate by randomly sampling from the confidence interval of the estimate of the area. Animating these displays involves resampling the pixels for each frame. Areas with uncertain values will flicker more dramatically than areas with more certain values.

4 Comparison and critique of alternative displays

4.1 Neither choropleth maps or cartograms perform well for Australia

Figure 4 shows four main types of cartograms using melanoma incidence on Australian Statistical Areas at Level 3 [19]. The version of a contiguous cartogram (a) has expanded the highly populated areas while preserving the full shapes of rural areas. It has not fully resolved the population transformation of areas, and if it had accurately sized areas by population, the country would be unrecognizable. The shape-preserved cartogram is unreadable, and it has reduced all areas to tiny spots on the map. Zooming in on a high-resolution output

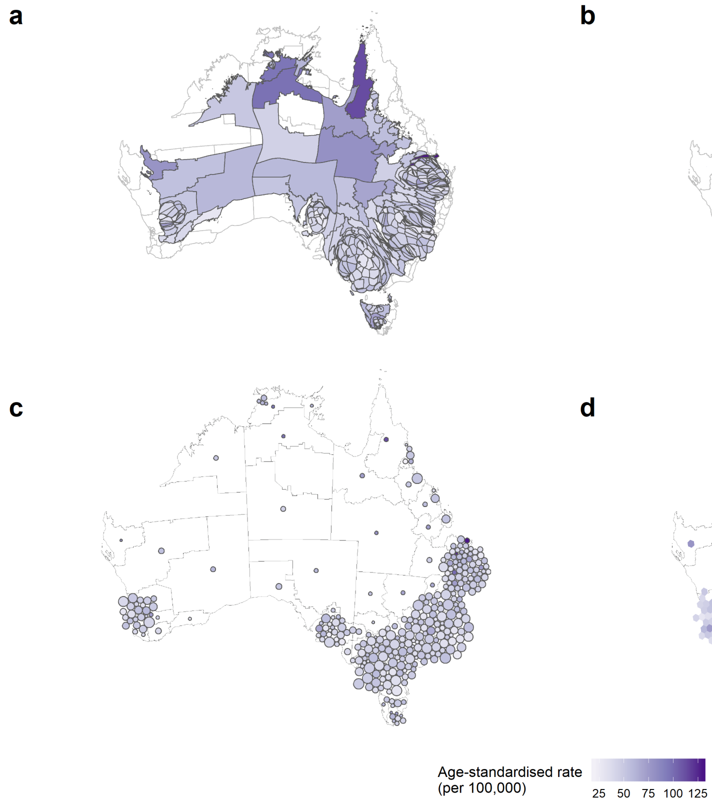


Figure 4: Cartograms showing melanoma incidence in Australia: (a) contiguous, partially population transformed, (b) non-contiguous shape preserved, (c) Dorling, (d) hexagon tile map. The contiguous cartogram has expanded the highly populated areas while preserving the full shapes of rural areas. If it accurately sized areas by population, the country would be unrecognizable. The shape-preserved is unreadable due to the small area sizes. The Dorling cartogram presents all areas but many are difficult to compare. The hexagon tile map provides a reasonable spatial distribution despite having isolated hexagons in the outback areas.

shows it does preserve the shapes. The Dorling cartogram and the hexagon tile map provide reasonable displays of the spatial distribution, despite having too much white-space in the outback areas.

4.2 Limitations of alternative displays

Cartograms provide the spatial distortion to more accurately convey the statistical distribution, focusing on the human impact of the disease. However, the transformation of contiguous cartograms often occurs at the expense of the shape of areas [33]. When the population density of the geographic units is highly dissonant with geographic density, the cartogram will lose all spatial context. Dorling [23] has a cartogram showing the 1966 general election results, which looked very little like the geographical shape of Australia.

Some mix of tiling, faceting or even micromaps, which allow some spatial continuity while also zooming into small areas, are good solutions for difficult geographies. Table 3 summarizes the key criteria for testing maps and alternative displays. Moore and Carpenter [1] and Bell et al. [6] provide suggestions and comments to help map creators best communicate their health data and spatial analysis.

Table 4: Summary of features and constraints of common mapping methods used to display cancer statistics (Y=Yes, N=No, S=Sometimes).

	Choropleth	Contiguous	Non-contig	Dorling	Tile maps	Geofacets
Spatial distortion	N	Y	Y	Y	Y	Y
Preserves neighbors	Y	Y	Y	S	S	S
Conceals small areas	Y	S	N	N	N	N
Uniform shape	N	N	N	Y	Y	Y
Univariate only	Y	Y	Y	S	S	N
Manual construction	N	N	N	N	Y	Y

5 User interaction

One of the concerns of adding too much information to a map is the fear of cognitive overload [52] in which the user reaches an information threshold, beyond which they become confused. It can be a juggling act for a diverse audience, with experts probably preferring more detail [53] while a simpler display is more broadly readable. Interactivity is a design feature within modern mapping methods that can be used to incorporate additional information and complexity without overloading the user. Effective user-centred interactive actions produce rapid, incremental, and reversible changes to the display [54].

Monmonier [32] says that interactivity can be used to allow users to explore the map for more information and provides flexibility for the display. The user can toggle between different variables, map views or even multiple realizations of future scenarios [55]. This provides additional mechanisms for the users to digest the uncertainty of the available information [56]. When the needs of the audience are changeable and are also the priority, the map creator can allow interactivity for map users to explore a data set through dynamic interactions. This can allow inspection of the data from many views [58]. User interaction with maps helps to understand and interpret the spatial distribution of disease, to validate, explain or explore the presented statistics and their relationships to each other [59].

Interactivity enables supplementary information to be incorporated into online atlases without cluttering the display. Interactive design features, found in online cancer maps, include tool tips, drop-down menus, data selection, zooming, and panning to allow users to explore the map as they want more information and allow flexibility in the display [32]. The use of these supports can be found in various online cancer maps and are shown in Figure 5 [18].

Animation, in contrast to interactivity, usually involves pre-computing views and showing these in a sequence. Lin Pedersen [60] provides an overview of animation for maps using the R package `gganimate` [61]. Animations are used to communicate a message by capturing and directing users' attention. It is most often employed to show changes over time. The controls for basic animation are usually placed outside of the plot space [60], and the map image is updated/replaced as the animation progresses.

Weather maps are a thoroughly developed examples of animation of spatial displays to communicate information to the general public [6]. The movement of a weather system will follow a forecasted path. All map users can follow the animated path of the weather system across the geography over a specified period.

The Australian Cancer Atlas [62] provides tours that change the display to draw users' attention to areas on the map that are relevant to the story. This implementation of animation gives users tools to plan their exploration.

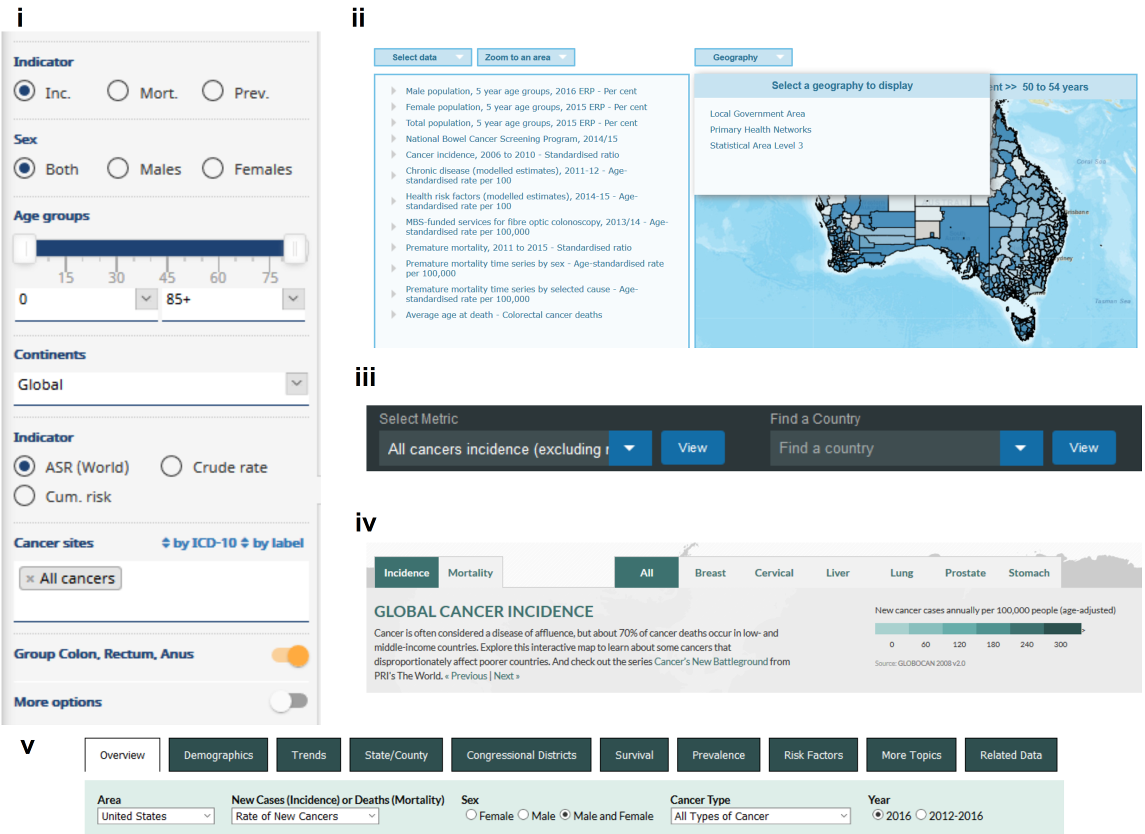


Figure 5: Interactive controls of displays in publicly available choropleth cancer maps: (i) GUI controls for statistic, sex, age groups, continents, and cancer types for Globocan 2018 [@Globocan], (ii) Menus for variable selection and zooming on Bowel Cancer Australia Atlas, (iii) Menus for choosing variables and countries in The Cancer Atlas, (iv) Tabs for different indicators and cancer types in Global Cancer Map, (v) Menus and toggles for variable and subset selection in United States Cancer Statistics: Data Visualizations.

Figure 6 shows two examples of more sophisticated interactive maps. The Spanish Cancer map (left) contains a linked display between a choropleth map and time series plots of cancer change. In linked plots, changing values in one display will trigger changes of corresponding elements in another display. Here, the temporal change in the choropleth map can be played out as an animation. Mousing over the time series plots will highlight the line for a particular region. The Canadian Breast Cancer Mortality map (right) has a magnifying glass that allows the user to zoom into small areas. It is easy to control and shows precise details in small areas.

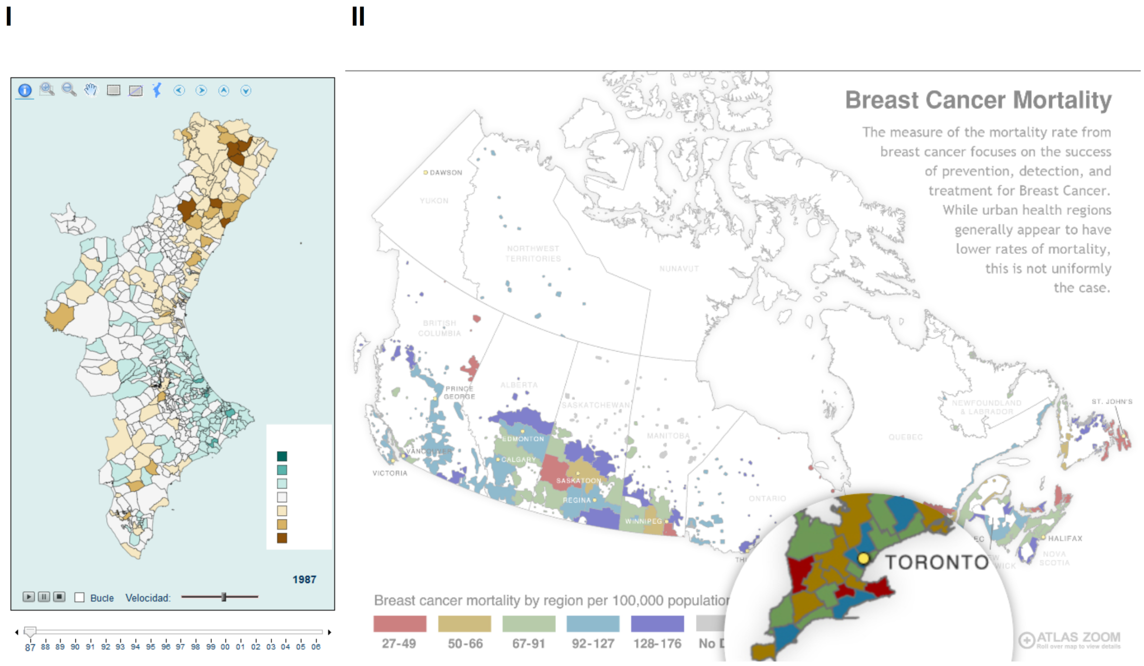


Figure 6: Two examples of advanced interactivity (and animation) in publicly available choropleth cancer maps: a. Linked maps and time-series line plots, with temporal animation in Map of Cancer Mortality Rates in Spain, b. A highly responsive magnifying glass on a map of Breast Cancer Mortality in Canada.

6 Conclusions

This paper provides an overview of mapping practices as commonly used for cancer atlases and recommends new approaches, such as cartograms and hexagon tile maps that should be adopted going forward. The conventional approach is the choropleth map, and it is widely used. The choropleth map suffers when there are small geographic units, as occurs in Australia where the population is concentrated on the coast, the information about the burden of cancer on those communities can be hidden. Making an inset can clarify congested regions but this breaks the viewers' attention as they shift focus from the map to the inset, and if there are many congested areas, many insets would be needed. The map alternatives implement trade-offs between the familiar shapes, and the importance of the geographic areas in the context of the areas. Given the population or a cancer statistic for each area, the geographic size or shape will change. Alternative displays allow the spatial distribution of cancer data to be digested by map users.

Many statistics are commonly used in cancer displays. The most basic is the incidence rate. It is common to see relative rates which measure how far a region is above or below the average. The purpose of using a relative rate is, perhaps the desire to pinpoint the areas that need attention because they have higher than expected rates. A region might be much higher than average, but it may not be close to a health concern, because all regions have a low incidence. Supplementary materials can allow map users to recognise when this occurs.

Interaction with maps is an important component of public atlases, and is easy to add with today's technology. The purpose is to provide access to more information than is possible to display in a single map, without overwhelming the viewer. Too many choices can similarly overwhelm a viewer, and thus decisions do need to be made about content to provide for accurate and comprehensive communication of information. Similarly, providing ways for users to interact with the display encourages engagement, and creative, efficient, elegant, interactive tools elicit curiosity about the data.

7 Acknowledgements

The authors would like to thank Dr. Earl Duncan for his contributions in editing and refining the drafts of this article. They would also like to thank Professor Kerrie Mengersen, Dr. Susanna Cramb and Dr. Peter Baade for conversations on the content of this article.

The following R [63] packages were used to produce this paper: tidyverse [64], RColorBrewer [65], ggthemes [66], png [67], cowplot [68], sf [69], spData [70], cartogram [71], sugarbag [72], knitr [73], rmarkdown [74] and absmapsdata [75].

Files to reproduce the paper, and code to reproduce the plots, are available at https://github.com/srkobakian/review.

References

1. Moore DA, Carpenter TE (1999) Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology. Epidemiologic Reviews 21:143–161

2. Exeter DJ (2016) Spatial Epidemiology. International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology 1–4

3. Tufte ER (1990) Envisioning Information. Graphics Press

4. Skowronnek A (2016) Beyond Choropleth Maps – A Review of Techniques to Visualize Quantitative Areal Geodata. In: Infovis Reading Group WS 2015/16. https://alsino.io/static/papers/BeyondChoropleths_AlsinoSkowronnek.pdf.

5. Walter SD (2001) Disease Mapping: A Historical Perspective. https://doi.org/https://dx.doi.org/10.1093/acprof:oso/9780198515326.003.0012

6. Bell BS, Hoskins RE, Pickle LW, Wartenberg D (2006) Current Practices in Spatial Analysis of Cancer Data: Mapping Health Statistics to Inform Policymakers and the Public. International Journal of Health Geographics 5:49

7. Brewster MB, Subramanian SV (2010) Cartographic Insights into the Burden of Mortality in the United Kingdom: A Review of "The Grim Reaper's Road Map". International Journal of Epidemiology 39:1120–1122

8. d'Onofrio A, Mazzetta C, Robertson C, Smans M, Boyle P, Boniol M (2016) Maps and Atlases of Cancer Mortality: A Review of a Useful Tool to Trigger New Questions. Ecancermedicalscience 10:670–670

9. Burbank F (1971) Patterns in Cancer Mortality in the United States 1950-67. National Cancer Institute Monograph Vol. 33, NCI, Washington DC

10. Emperial College London - Small Area Health Statistics Unit (2010) The environmental and health atlas of england and wales: National male lung cancer rate. http://www.envhealthatlas.co.uk/eha/Breast/. Accessed 26 Sep 2019

11. World Health Organization's International Agency for Research on Cancer (2018) Globocan 2018: Estimated cancer incidence, mortality and prevalence. http://globocan.iarc.fr/Pages/Map.aspx. Accessed 26 Sep 2019

12. Queensland Cancer Registry (2011) The Atlas of Cancer in Queensland (1998 - 2007). https://cancerqld.org.au/research/queensland-cancer-statistics/queensland-cancer-atlas/. Accessed 26 Sep 2019

13. Bowel Cancer Australia (2016) Bowel Cancer Australia Atlas. http://www.bowelcanceratlas.org/. Accessed 26 Sep 2019

14. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute - Cancer Statistics Working Group (2019) U.S. Cancer Statistics Data Visualizations Tool (data 1999-2016). http://www.cdc.gov/cancer/dataviz. Accessed 26 Sep 2019

15. El Pais (2014) Map of Cancer Mortality Rates in Spain. http://elpais.com/elpais/2014/10/06/media/1412612722_141933.html. Accessed 26 Sep 2019

16. Pediatric Oncology Group of Ontario (2015) Incidence Rate of Childhood Cancers, Atlas of Childhood Cancer in Ontario (1985-2004). https://www.pogo.ca/wp-content/uploads/2015/02/POGO_CC-Atlas-3-Incidence_Feb-2015.pdf. Accessed 26 Sep 2019

17. Howe G (1989) Historical Evolution of Disease Mapping in General and Specifically of Cancer Mapping. In: Cancer mapping. Springer, pp 1–21

18. Roberts J (2019) Communication of Statistical Uncertainty to Non-expert Audiences. https://doi.org/10.5204/thesis.eprints.130786

19. Statistics AB of (2018)https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS).

20. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2018) Global Cancer Observatory: Cancer Today. https://gco.iarc.fr/today.

21. Northern Ireland Cancer Registry (2011) All-Ireland Cancer Atlas (1995-2007). http://www.ncri.ie/publications/cancer-atlases.

22. Kronenfeld BJ, Wong DWS (2017) Visualizing Statistical Significance of Disease Clusters Using Cartograms. International Journal of Health Geographics 16:19

23. Dorling D (2011) Area Cartograms: Their Use and Creation. In: Concepts and techniques in modern geography (catmog). pp 252–260

24. Harrower M, Brewer CA (2003) ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. The Cartographic Journal 40:27–37

25. van der Walt, S. and Smith, N (2015) mpl colormaps. https://bids.github.io/colormap/.

26. Madsen R (2019) Programming Design Systems. https://programmingdesignsystems.com/.

27. Dougenik JA, Chrisman NR, Niemeyer DR (1985) An Algorithm to Construct Continuous Area Cartograms. The Professional Geographer 37:75–81

28. Griffin TL (1980) Cartographic Transformation of the Thematic Map Base. Cartography 11:163–174

29. Berry BJL, Morrill RL, Tobler WR (1964) Geographic Ordering of Information: New Opportunities. The Professional Geographer 16:39–44

30. Olson JM (1976) Noncontiguous Area Cartograms. The Professional Geographer 28:371–380

31. Levison ME, Haddon Jr W (1965) The Area Adjusted Map. An Epidemiologic Device. Public Health Reports 80:55–59

32. Monmonier M (2018) How to Lie with Maps (Third Edition). https://doi.org/10.1191/0309132505ph540pr

33. Kocmoud C, House D (1998) A Constraint-based Approach to Constructing Continuous Cartograms. In: Proc. Symp. Spatial data handling. pp 236–246

34. Dent BD (1972) A Note on the Importance of Shape in Cartogram Communication. Journal of Geography 71:393–401

35. Kraak MJ (2017) Cartographic Design. In: The International Encyclopedia of Geography: People, the Earth, Environment, and Technology. Wiley, United States, pp 1–16

36. Nusrat S, Kobourov SG (2016) The State of the Art in Cartograms. Computer Graphics Forum 35:619–642

37. Min Ouyang, Revesz P (2000) Algorithms for Cartogram Animation. In: Proceedings 2000 International Database Engineering and Applications Symposium (Cat. No.PR00789). pp 231–235

38. Keim DA, North SC, Panse C, Schneidewind J (2002) Efficient Cartogram Generation: A Comparison. In: IEEE Symposium on Information Visualization, 2002. INFOVIS 2002. IEEE, pp 33–36

39. Tufte ER (2001) The visual display of quantitative information. Graphics press Cheshire, CT

40. Raisz E (1963) Rectangular Statistical Cartograms of the World. Journal of Geography 35:8–10

41. Tobler W (2004) Thirty Five Years of Computer Cartograms. Annals of the Association

of American Geographers 94:58–73

42. Monmonier M (2005) Cartography: Distortions, World-views and Creative Solutions. Progress in Human Geography 29:217–224

43. Kreveld M van, Speckmann B (2007) On rectangular cartograms. Computational Geometry 37:175–187

44. Montanaro D (2016) NPR Battleground Map: Hillary Clinton Is Winning — And It's Not Close.

45. Kanjana J, Mehta D (2016) Who will win the presidency?

46. Zitner A, Yeip R, Wolfe J (2016) Draw the 2016 Electoral College Map.

47. Gamio L, D. C (2016) Poll: Redrawing the electoral map.

48. Cano RG, Buchin K, Castermans T, Pieterse A, Sonke W, Speckmann B (2015) Mosaic Drawings and Cartograms. In: Computer graphics forum. Wiley Online Library, pp 361–370

49. Hafen R (2019) Geofacet: 'Ggplot2' faceting utilities for geographical data.

50. W. PL, Carr DB, Pearson JB (2015) micromapST: Exploring and Communicating Geospatial Patterns in US State Data. Journal of Statistical Software 63:1–25

51. Lucchesi LR, K. WC (2017) Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation and Glyph Rotation. Stat. https://doi.org/10.1002/sta4.150

52. McGranaghan M (1993) A Cartographic View of Spatial Data Quality. Cartographica: The International Journal for Geographic Information and Geovisualization 30:8–19

53. Cliburn DC, Feddema JJ, Miller JR, Slocum TA (2002) Design and Evaluation of a Decision Support System in a Water Balance Application. Computers & Graphics 26:931–949

54. Perin C (2014) Direct Manipulation for Information Visualization. Theses, Université Paris Sud - Paris XI

55. Goodchild M, Buttenfield B, Wood J (1994) On Introduction to Visualizing Data Validity.

Visualization in geographical information systems 141–149

56. MacEachren AM (1992) Visualizing Uncertain Information. Cartographic Perspectives 10–19

57. Van der Wel FJ, Hootsmans RM, Ormeling F (1994) Visualization of Data Quality. In: Modern cartography series. Elsevier, pp 313–331

58. Dang G, North C, Shneiderman B (2001) Dynamic Queries and Brushing on Choropleth Maps. In: Proceedings Fifth International Conference on Information Visualisation. pp 757–764

59. Carr DB, Wallin JF, Carr DA (2000) Two New Templates for Epidemiology Applications: Linked Micromap Plots and Conditioned Choropleth Maps. Statistics in Medicine 19:2521–2538

60. Pedersen TL (2018) The Grammar of Animation. https://youtu.be/21ZWDrTukEs. Accessed 16 Nov 2018

61. Pedersen TL, Robinson D (2019) gganimate: A Grammar of Animated Graphics.

62. Cancer Council Queensland, Queensland University of Technology, and Cooperative Research Centre for Spatial Information (2018) Australian Cancer Atlas. https://atlas.cancer.org.au.

63. R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

64. Wickham H (2017) tidyverse: R packages for data science. https://CRAN.R-project.org/package=tidyverse.

65. Neuwirth E (2014) RColorBrewer: ColorBrewer palettes. https://CRAN.R-project.org/package=RColorBrewer.

66. Arnold JB (2019) ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. https://CRAN.R-project.org/package=ggthemes.

67. Urbanek S (2013) png: Read and write PNG images. https://CRAN.R-project.org/package=png.

68. Wilke CO (2019) cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. https://CRAN.R-project.org/package=cowplot.

69. Pebesma E (2018) Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10:439–446

70. Bivand R, Nowosad J, Lovelace R (2019) spData: Datasets for Spatial Analysis. https://CRAN.R-project.org/package=spData.

71. Jeworutzki S (2018) cartogram: Create Cartograms with R. https://CRAN.R-project.org/package=cartogram.

72. Kobakian S, Cook D (2019) sugarbag: Create Tessellated Hexagon Maps. https://CRAN.R-project.org/package=sugarbag.

73. Xie Y (2019) knitr: A General-Purpose Package for Dynamic Report Generation in R.

74. Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2019) rmarkdown: Dynamic Documents for R.

75. Mackey, W. F. (2019) Absmapsdata: A catalogue of ready-to-use asgs mapping data.

Chapter 3

An algorithm for creating a tesselated hexagon display

This chapter relates to the first research aim as stated in 1.3. The chapter introduces the step devised for the algorithm. The chapter frames the steps in light of the `sugarbag` (Kobakian and Cook, 2019) package functions that allow users to enact the algorithm in `R` (R Core Team, 2019a). It uses the Statisical Areas of Australia at Level 2, taking a subset and considering only those located within the island of Tasmania. It also provides an example of how to animate between the choropleth map display and the hexagon tile map.

This chapter was submitted for publication to the *Journal of Statistical Software* for publication. The steps in this algorithm are implemented in the `sugarbag` (Kobakian and Cook, 2019) package for `R` (R Core Team, 2019a).

RESEARCH STUDENTS CENTRE
Examination Enquiries: 07 3138 1839
Email: research.examination@qut.edu.au

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

An Algorithm For Spatial Mapping Using a Hexagon Tile Map, With Application to Australian Maps has been submitted to the Journal of Statistical Software.

Contributor	Statement of contribution*
Stephanie Kobakian S R Kobakian 27/12/2019	Stephanie researched current methods for presenting geospatial data, wrote the initial draft and revised the drafts after suggestions were made by reviewers.
Prof. Dianne Cook	Prof. Dianne helped to create and refine the steps of the algorithm that is implemented in the sugarbag package and discussed in this paper

Principal Supervisor Confirmation

I have sighted email or other correspondence from all Co-authors confirming their certifying authorship. (If the Co-authors are not able to sign the form please forward their email or other correspondence confirming the certifying authorship to the RSC).

Name Signature Date

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

An Algorithm For Spatial Mapping Using a Hexagon Tile Map, With Application to Australian Maps

Stephanie Kobakian
Queensland University of Technology

Dianne Cook
Monash University

Earl Duncan
Queensland University of Technology

Abstract

This algorithm creates a tesselated hexagon display to represent each of the spatial polygons. It allocates these hexagon in a manner that preserves the spatial relationship of the geographic units. It showcases spatial distributions, by emphasising the small geographical regions that are often difficult to locate on geographic maps. Spatial distributions have been presented on alternative representations of geography for many years. In modern times, interactivity and animation have begun to play a larger role, as alternative representations have been popularised by online news sites, and atlas websites with a focus on public consumption. Applications are increasingly widespread, especially in the areas of disease mapping, and election results.

Keywords: spatial, statistics, cartogram.

1. Introduction

The current practice for presenting geospatial data involves a choropleth map display. These maps highlight the geographic patterns in geospatially related statistics (Moore and Carpenter 1999). The land on the map space is divided into geographic units, these boundaries are usually administrative, such as states or counties. The units are filled with colour to represent the value of the statistic (Tufte 1990).

Australian residents are increasingly congregating around major cities, the vast rural areas are often sparsely populated in comparison to the urban centres. In Australia, government bodies such as the Australian Bureau of Statistics (ABS), and the Australian Electoral Commission (AEC) hold the responsibility for the division of geographic units. If it absolutely necessary, the AEC may adjust as the population increases. The division of the population into approximately equal population areas results in dramatically different square meterage of the geographic areas. This results in an unequal attention given to the statistic of each area, this can allow misrepresentation of the spatial distributions of human related statistics in geographic maps.

The solution to this visualisation problem begins with the geography. Cartograms apply a transformation to the geographic boundaries based on the value of the statistic of interest. These displays result in a distortion of the map space to represent differences in the statistic across the areas (Dougenik, Chrisman, and Niemeyer 1985). The statistic of interest is used to determine the layout. When using the Australian population, the result is a population cartogram that fails to preserve a recognisable display due to the difference in size of metropolitan and rural areas (Dorling 2011), (Berry, Morrill, and Tobler 1964). Contiguous cartograms change the shape of an area, preserving boundary relationships of neighbours. Non-contiguous cartograms maintain the geographic shape of each area, but lose the connection to neighbours as each areas shrinks or grows.

Alternative maps shift the focus from land area and shape, to the value of the statistics in a group of areas. Alternative mapping methods allow increased understanding of the spatial distribution of a variable across the population, by fairly representing each administrative area. This acknowledges that the amount of residents can be different but recognises that each area, or person is equally important.

Tilegrams, Rectangular cartograms (van Kreveld and Speckmann 2007) and Dorling cartograms (Dorling 2011), all use one simple shape to represent each area. This allows preservation of spatial relationships and decreases the emphasis from the size of the geographic areas. These maps focus on the relationship between neighbours attempting to preserve connections, and disregard the unique shapes of the administrative boundaries.

The **sugarbag** package provides a new method to create tesselated hexagon tilegrams. Extending the tilegram to Australian applications required preserving the spatial relationships. It emphasises the capital cities as population hubs, and emphasises the distances rather than size of large, rural geographic units.

2. Algorithm

This solution operates on a set of simple feature geometry objects, also known as **sf** (Pebesma 2018) polygons.

There are four steps to create a tesselated hexagon tilegram. These steps can be executed by the main function, `create_hexmap`, or can be implemented separately for more flexibility. There are parameters used in the process that can be provided, if they are not, they will be automatically derived.

1. Create the set of centroids to allocate
2. Create the grid of hexagons locations to use
3. Allocate each centroid to an available hexagon
4. Transform the data for plotting

2.1. Parameters

The `create_hexmap` function requires several parameters, if they are not provided, the information will be derived from the simple features (**sf**) set of shapes used. Users may choose to only use the `allocate` function when they wish to use a set of centroids, rather than **sf** polygons.

The following parameters must be provided to `create_hexmap`:

- *shp:* an sf object containing the polygon information
- *sf_id:* name of a column that distinguishes unique areas
- *focal_points:* a data frame of reference locations used to allocate hexagons

2.2. Polygon set

The polygon set of Statistical Areas at Level 2 (SA2) (Australian Bureau of Statistics 2018) of Tasmania in 2016 is provided with the **sugarbag** package as `tas_sa2`. A single column of the data set must be used to identify the unique areas. In this case, the unique SA2 names for each SA2 have been used.

The longitude and latitude centre of the capital cities of Australia are used to allocate areas around the closest capital city. Hobart will be the common focal point, as this example uses only the areas in the state of Tasmania.

```
R> data(capital_cities)
```

The following parameters will be determined within `create_hexmap` if they are not provided. They are created throughout the following example:

- *buffer_dist:* a float value for distance in degrees to extend beyond the geometry provided
- *hex_size:* a float value in degrees for the diameter of the hexagons
- *hex_filter:* amount of hexagons around centroid to consider for allocation
- *width:* the angle used to filter the grid points around a centroid

2.3. Create the set of centroid points

Individual **sugarbag** functions can be used outside of the main function. The set of polygons should be provided as an **sf** object, this is a data frame containing a `geometry` column. The `read_shape` function can assist in creating this object.

The centroids can be derived from the set of polygons using the `create_centroids` function:

```
R> centroids <- create_centroids(shp_sf = tas_sa2, sf_id = "SA2_NAME16")
```

2.4. Create the hexagon grid points

A tilegram presents areas on a tellesated set of tiles. The grid is created to ensure tessellation between the hexagons.

The grid of possible hexagon centroids is made using the `create_grid` function. The grid creation requires several steps. It uses the centroids, the hexagon size and the buffer distance.

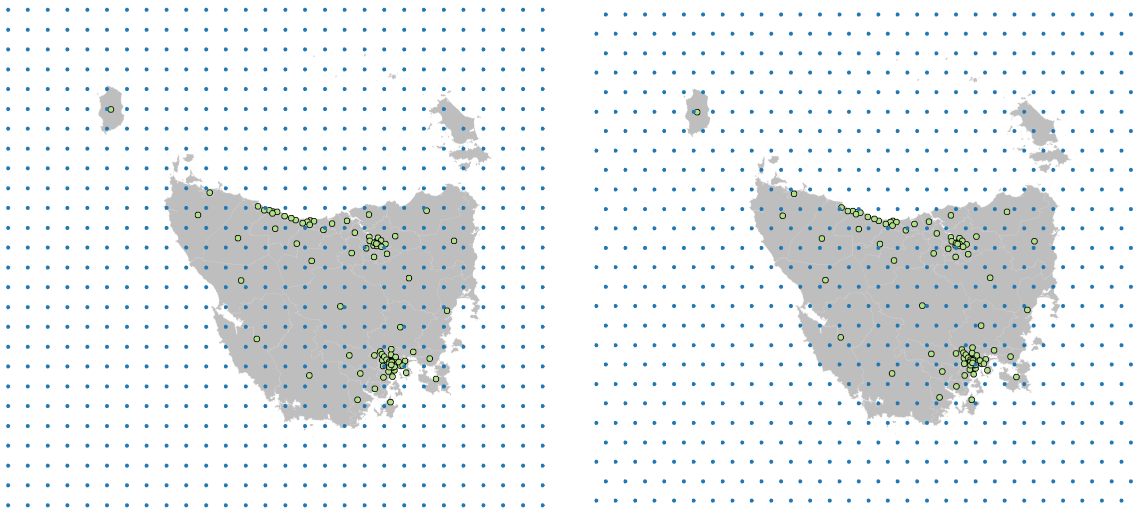


Figure 1: Grid points to create a tilegram.

```
R> grid <- create_grid(centroids = centroids, hex_size = 0.2, buffer_dist = 1.2)
```

Step 1: Creating a tesselated grid

A set of longitude columns, and latitude rows are created to define the locations of the hexagons. The distance between each row and column is defined by the size specified as `hex_size`. The minimum and maximum, longitude and latitude values of the centroid locations are found. Equally spaced columns are created from the minimum longitude minus the buffer distance, up to the maximum longitude plus the buffer distance. The rows are then created from the latitude values and the buffer distance. An individual hexagon location is created from all intersections of the longitude columns and latitude rows.

A square grid could be used for sqaure tiles, but it will not facilitate tessellated hexagons. Figure 1 allows for hexagons, as every second latitude row on the grid is shifted right, by half of the hexagon size.

Step 2: Rolling windows

Not all of the grid points will be used, especially if islands result in a large grid space. To filter the grid for appropriate points for allocation, the `create_buffer` function is called within `create_grid`. It finds the grid points needed to best capture the set of centroids on a hexagon tile map.

For each centroid location, the closest latitude row and longitude column are found. Then rows and columns of centroids are divided into 20 groups. The amount of rows in each latitude group and the amount of columns in each longitude group are used as the width of rolling windows. The rolling windows can be seen on the bottom and right of Fig. 2. This will tailor the available grid points to those most likely to be used. It also helps reduce the amount of time taken, as it decreases the amount of points considered for each centroid allocation.

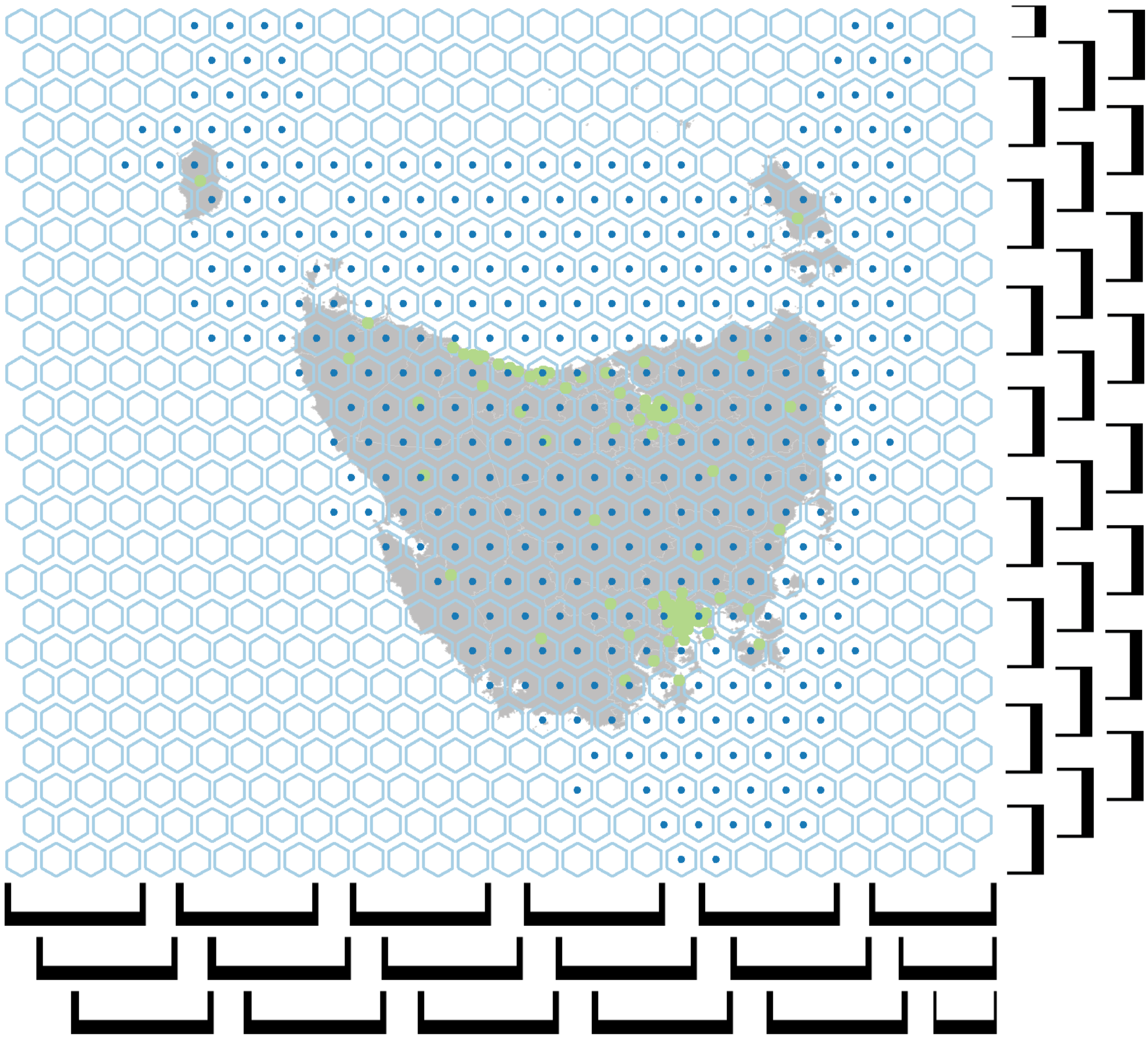


Figure 2: All possible hexagon locations from the initial grid are shown with blue outlines. The blue dots shown the grid points left as to choose from after the buffer step. The rolling windows to the right show the rows used to filter the hexagon locations.

The first rolling window function finds the minimum and maximum centroid values for the sliding window groups of longitude columns and the groups of latitude rows.

The second rolling window function finds the average of the rolling minimum and maximum centroid values, for the longitude columns and latitude rows.

Step 3: Filtering the grid

Only the grid points between the rolling average of the minimum and maximum centroid values are kept, for each row and column of the grid.

2.5. Centroid to focal point distance

The distance between each centroid in the set, and each of the focal points provided is calculated. The name of the closest focal point, and the distance and angle from focal point to polygon centroid is joined to polygon data set. To minimise time taken for this step, only Tasmania's capital city Hobart is provided. The order for allocation is determined by the distance between the polygon centroid and it's closest focal point. The points are arranged from the centroid closest to the focal points, to the furthest.

2.6. Allocate each centroid to a hexagon grid point

Allocation of all centroids takes place using the set of polygon centroids and the hexagon map grid. Centroid allocation begins with the closest centroid to a focal point. This will preserve spatial relationships with the focal point, as the inner city areas are allocated first, they will be placed closest to the capital, and the areas that are further will then be accommodated. Only the hexagon grid points that have not yet been allocated are considered.

The possible hexagon locations consider for a centroid location are determined by the `hex_filter`. This is the maximum amount of hexagons between the centroid and the furthest considered hexagon. It is used to subset possible grid points to only those surrounding the polygon centroid within an appropriate range. A smaller distance will increase speed, but can decrease accuracy if the angle width increases.

```
R> hexmap_allocation <- allocate(
R>   centroids = centroids %>% select(SA2_NAME16, longitude, latitude),
R>   sf_id = "SA2_NAME16",
R>   hex_grid = grid,
R>   hex_size = 0.2, # same size used in create_grid
R>   hex_filter = 10,
R>   width = 35,
R>   focal_points = capital_cities,
R>   verbose = TRUE)
```

The following example considers one of the Statistical Areas at Level 2. Within the algorithm, these steps are repeated for each polygon.

Step 1: Filter the grid for unassigned hexagon points

Keep only the available hexagon points, this will prevent multiple areas being allocated to the same hexagon.

Step 2: Filter the grid points for those closest to the centroid

This will allow only the closest points that are not yet assigned, to be considered.

A box of possible hexagon locations around the centroid. The corners of the box may not look square as the buffer has already removed unnecessary points from over the ocean.

The algorithm then removes the outer corners of the square, creating a circle of points, by only keeping points within a certain radial distance around the original centroid location.

The `width` parameter is used to take a slice of the remaining points. This uses the angle from the closest capital city, to the current centroid. This allows the spatial relationship to

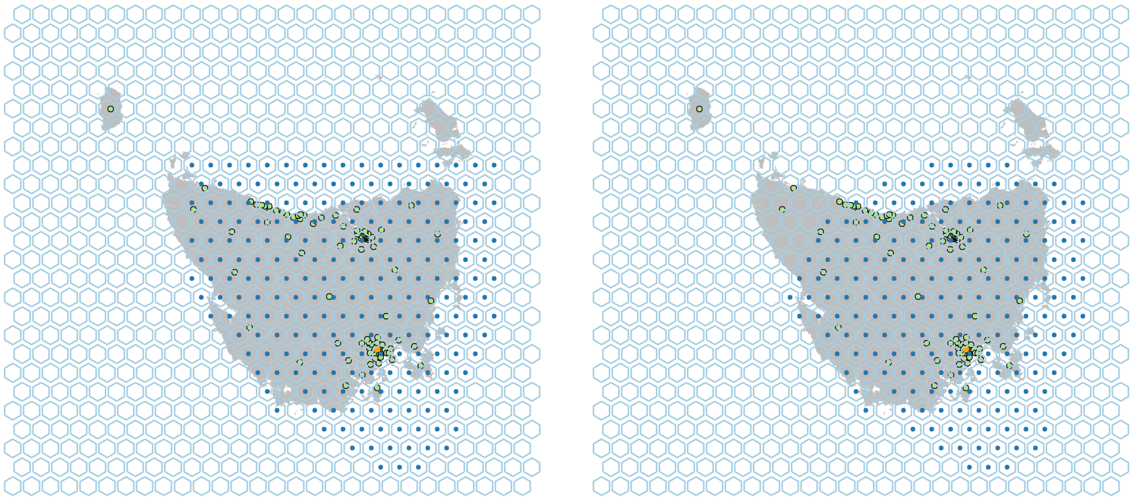


Figure 3: Filter for grid points within a square, then circular, distance for those closest to the centroid.

be preserved, even when it is allocated to a hexagon that is further from the focal point then the original centroid location.

If no available hexagon grid point is found within the original filter distance and angle, the distance is expanded, only when a maximum distance is reached will the angle expand to accommodate more possible grid points.
By default the angle filter to hexagon grid points that fall within the bounds of the angle from the focal point to the geographic centroid, plus and minus 30 degrees. This will increase if no points can be found within the `hex_filter` distance. The angle of 30 was chosen to allow the algorithm to choose hexagons that best maintained the spatial relationship between the focal point and geographic centroid. The allocation is returned and combined with the data relating to each polygon.

A complete hexagon tile map of Tasmania is created by applying the algorithm steps to each centroid. The hexagon tile map visualisation is used below to visualise the Australian Cancer Atlas data. Two views of the same data are produced by filling according to the Lung Cancer Standardised Incidence Rates (SIRs) downloaded from the Australian Cancer Atlas site. This small example in Figure 5 shows the group of blue areas in the Hobart CBD more prominently in the hexagon tile map (b). The small red areas visible in the choropleth map (a) along the north coast are much larger in the hexagon tile maps. The hexagon tile map shows less yellow, this no longer overwhelms the map space with the information regarding the more rural areas.

2.7. Neighbour relationships

It is possible to incorporate the neighbouring areas for each SA2, for stronger preservation of the spatial distribution.

An additional step can be included to allow the neighbours that have alredy been allocated to influence the placement of the current centroid. This requires respecifying the `sf` object as the

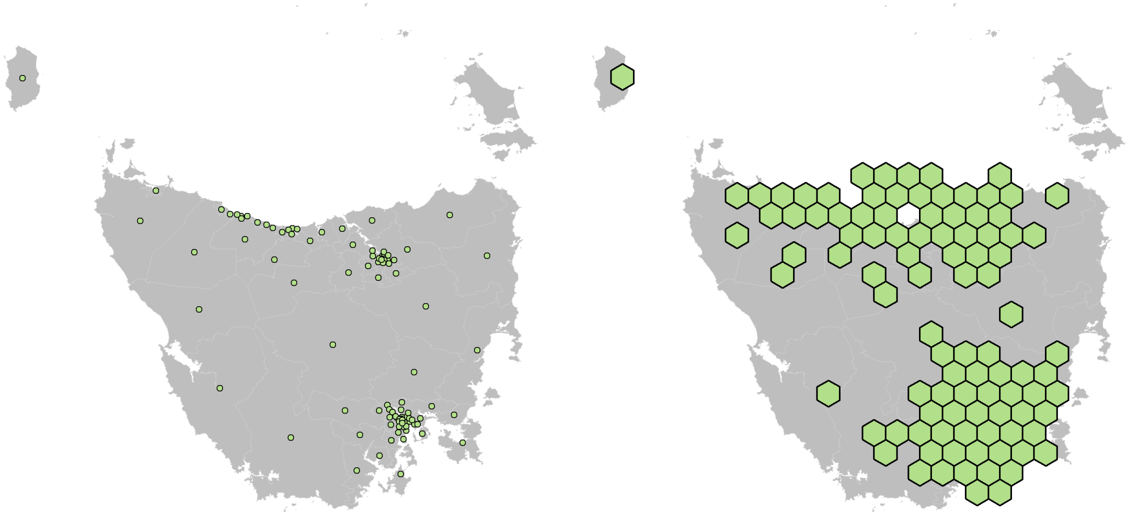


Figure 4: A complete hexagon tile map of Tasmania.

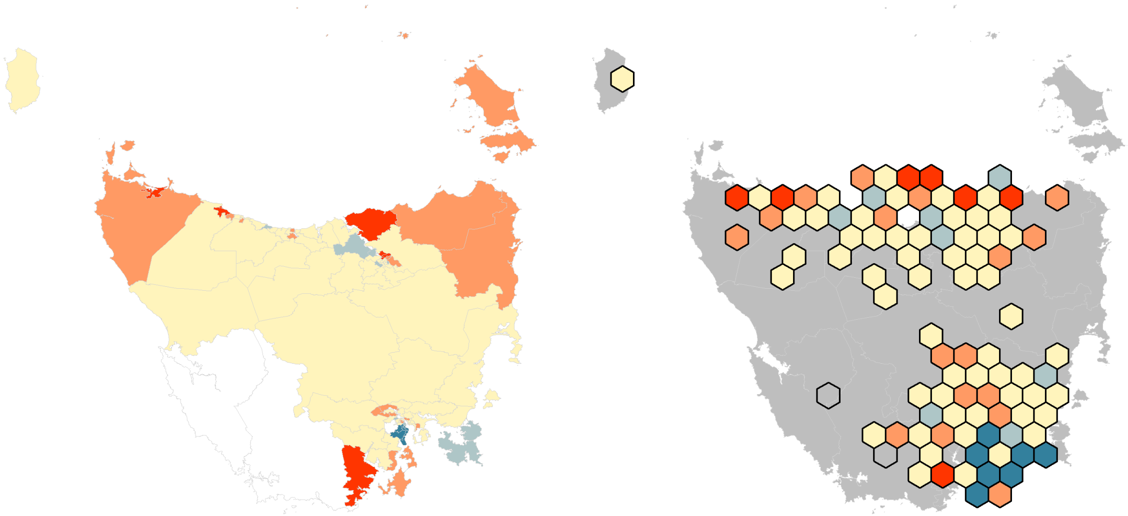


Figure 5: The Australian Cancer Atlas data has determined the colour of each Statistical Area of Australian at Level 2. A choropleth map (a) of Standardised Incidence Rates (SIRs) is paired with a hexagon tile map (b) to contrast the colours that are made obvious when every SA2 is equally represented.

argument for the `use_neighbours` parameter. This calculates neighbours using intersections of their polygons. This occurs for all areas before any allocations begin.

For the current centroid, the list of neighbours is consulted. If any neighbour was already allocated, the surrounding hexagons on the grid are prioritised. For multiple neighbours, the neighbouring hexagon grid points are aggregated and considered in order of distance from the original centroid.

3. Using sugarbag

3.1. Installation

The package can be installed from CRAN:

```
R> install.packages("sugarbag")
```

and the development version can be install from the GitHub repository:

```
R> devtools::install_github("srkobakian","sugarbag")
```

Load the library into your R session with:

```
R> library(sugarbag)
```

3.2. Creating a hexagon tile map

The following code creates the hexagon tile map for all the Statistical Areas at Level 2 in Tasmania.

```
R> # Load data
R> data(tas_sa2)
R>
R> # Create centroids set
R> centroids <- create_centroids(tas_sa2, "SA2_NAME16")
R>
R> # Create hexagon grid
R> grid <- create_grid(centroids = centroids,
R>     hex_size = 0.2,
R>     buffer_dist = 1.2)
R>
R> # Allocate polygon centroids to hexagon grid points
R> hex_allocated <- allocate(
R>   centroids = centroids,
R>   hex_grid = grid,
R>   sf_id = "SA2_NAME16",
```

```
R>   # same column used in create_centroids
R>   hex_size = 0.2,
R>   # same size used in create_grid
R>   hex_filter = 10,
R>   use_neighbours = tas_sa2,
R>   focal_points = capital_cities %>% filter(points == "Hobart"),
R>   width = 35,
R>   verbose = FALSE)
R>
R> # Prepare to plot
R> fort_hex <- fortify_hexagon(data = hex_allocated, sf_id = "SA2_NAME16", hex_size = 0.2)
R>
R> # Make a plot
R> library(ggplot2)
R> ggplot(fort_hex) +
R>   geom_polygon(aes(x=long, y=lat, group=hex_id, fill = lat)) +
R>   scale_fill_distiller("", palette="PRGn")
```

4. Applications

4.1. Australian Cancer Atlas

The Australian Cancer Atlas (Cancer Council Queensland, Queensland University of Technology, and Cooperative Research Centre for Spatial Information 2018) allows estimates derived from the models of Standardised Incidence Rates and excess deaths to be downloaded. Figure 6 is a choropleth map that uses colour to display the estimated Standardised Incidence Rate of melanoma cancer for all persons for each SA2. The Australian choropleth map display draws attention to the expanse of light blue areas across the rural communities in all states. The SA2s around Brisbane stand out as more orange and red. Comparatively, the hexagon tile map display in Figure 7 draws attention to contrast of the blue areas in Sydney and Melbourne and the capital city of Brisbane. In both Sydney and Melbourne, the hexagons that represent the SA2 areas in the inner-city areas have lower than average Incidence Rates.

With careful consideration of the choropleth map, the small geographic inner city areas may have been noticed by viewers, but the hexagon tile map display emphasises them. The communities in northern Queensland and the Northern territory do not draw attention because of their size as in the choropleth, but their colour is still noticably below average when contrasted with the heaxgons further south.

To create this choropleth map the SA2 polygons for 2011 from the ABS. The Standardised Incidence Ratios for each geographic unit are joined to the appropriate polygons.

To create the hexagon tile map display, the same steps are followed as outlined above:

- Create the set of centroid points
- Create the hexagon grid points
- Allocate each centroid to a hexagon grid point

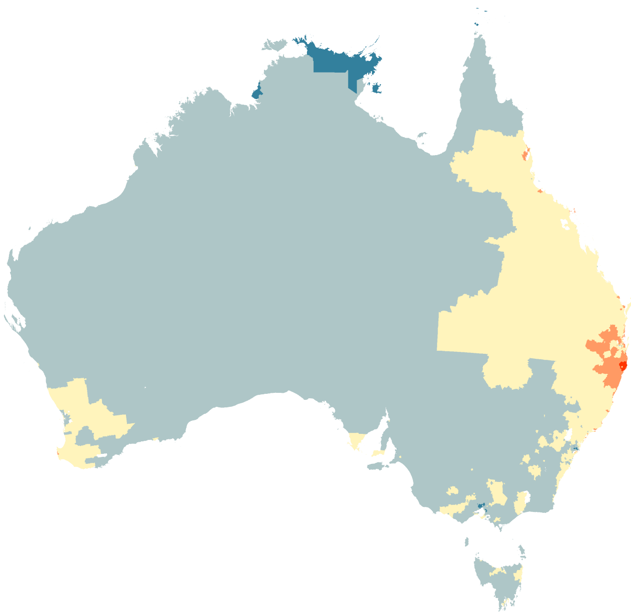


Figure 6: A choropleth map of the Statistical Areas of Australia at Level 2. The colours communicate the value of the estimated Standardised Incidence Rate of Melanoma for all persons, they range from much lower than average (blue) to much higher than average (red)

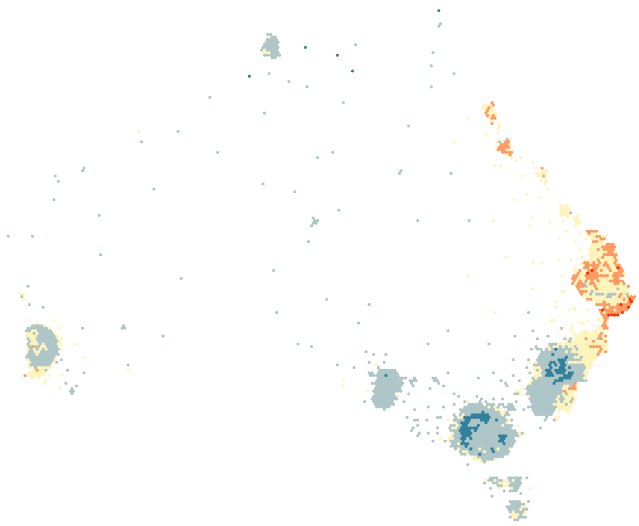


Figure 7: A hexagon tile map of the Statistical Areas of Australia at Level 2. The colours communicate the value of the estimated Standardised Incidence Rate, they range from much lower than average (blue) to much higher than average (red)

5. Animation

The **gganimate** package can be used to make an animation. It requires connecting the polygons for each area in two displays two displays, which can be done using the `sf_id` variable, such as the SA2 name. The animation[1] connecting these two displays will highlight the rapid growth of the inner-city areas, and will decrease the large rural areas. The hexagons that move the furthest will move rapidly in the animation.

6. Conclusion

It is possible to use alternative maps to communicate spatial distributions.While a choropleth map display is the current practice spatial visualisation of geographical data. Current methods do not always work for Australia due to the large gaps between densely populated capital cities. The administrative boundaries may distract from the statistics, communicated using colour.

Alternative maps highlight the value of the statistics across the geographic units. Alternative mapping methods allow increased understanding of the spatial distribution of a variable across the population, by fairly representing each administrative area. This acknowledges that the amount of residents can be different but recognises that each population area is equally important. The solution to this visualisation problem has equally sized areas, with neighbourhood boundary connections. This map algorithm is implemented in the **sugarbag** Kobakian and Cook (2019) package written for `R` R Core Team (2012). The **sugarbag** package creates tesselated hexagon tilegrams. The Australian application preserves the spatial relationships, emphasising capital cities. The hexagon tile map is a visualisation solution that highlights spatial distributions.

These hexagons equally represent each area. However, the tesselation does not allow the size of the hexagons to represent another variable, similar to the choropleth maps. The algorithm is heavily dependent on the focal points used, as this determines the order of allocation. It works on the assumption that viewers can use directional assumptions to identify their neighbourhoods, this can be aided by the animation.

Future work will include refining the algorithm. It would be possible to take a logarithmic function rather than a direct angle to help choose a closer hexagon to the original centroid location, before increasing the width of the angle used to filter the hexagons.

This algorithm has only been tested using single countries, and does not consider definite borders of countries. While the buffer allows extention beyond the furthest centroids, there is no mechanism to protect the borders and ensure centroids are placed within the geographic borders of a country.

This algorithm is an effective start to creating hexagon tile maps for many geographic units.

7. Acknowledgements

The authors would like to thank the Australian Cancer Atlas team for discussions regarding alternative spatial visualizations. They would also like to thank Professor Kerrie Mengersen

[1] This animation can be viewed at: `https://sugarbagjss.netlify.com/`

for suggestions and comments. Mitch O'Hara-Wild provided assistance with some parts of the algorithm and code. Sayani Gupta provided helpful advice on parts of this paper.

The code for **sugarbag** (Kobakian and Cook 2019) can be found on CRAN, along with vignettes on installation and usage.

Australian Electoral Data can be found in the **eechidna** package for R.

References

Australian Bureau of Statistics (2018). URL https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS).

Berry BJL, Morrill RL, Tobler WR (1964). "Geographic Ordering of Information: New Opportunities." *The Professional Geographer*, **16**, 39–44. doi:10.1111/j.0033-0124.1964.039_q.x.

Cancer Council Queensland, Queensland University of Technology, and Cooperative Research Centre for Spatial Information (2018). "Australian Cancer Atlas." URL https://atlas.cancer.org.au.

Dorling D (2011). *Area Cartograms: Their Use and Creation*, volume 59, pp. 252 – 260. ISBN 9780470979587. doi:10.1002/9780470979587.ch33.

Dougenik JA, Chrisman NR, Niemeyer DR (1985). "AN ALGORITHM TO CONSTRUCT CONTINUOUS AREA CARTOGRAMS." *The Professional Geographer*, **37**(1), 75–81. doi:10.1111/j.0033-0124.1985.00075.x. https://doi.org/10.1111/j.0033-0124.1985.00075.x, URL https://doi.org/10.1111/j.0033-0124.1985.00075.x.

Kobakian S, Cook D (2019). *sugarbag: Create Tessellated Hexagon Maps*. Https://srkobakian.github.io/sugarbag/, https://github.com/srkobakian/sugarbag.

Moore DA, Carpenter TE (1999). "Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology." *Epidemiologic Reviews*, **21**(2), 143–161. ISSN 1478-6729. doi:10.1093/oxfordjournals.epirev.a017993. http://oup.prod.sis.lan/epirev/article-pdf/21/2/143/6727658/21-2-143.pdf, URL https://doi.org/10.1093/oxfordjournals.epirev.a017993.

Pebesma E (2018). "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal*, **10**(1), 439–446. doi:10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Tufte ER (1990). *Envisioning Information*. Graphics Press.

van Kreveld M, Speckmann B (2007). "On rectangular cartograms." *Computational Geometry*, **37**(3), 175 – 187. ISSN 0925-7721. doi:10.1016/j.comgeo.2006.06.002. Special Issue on the 20th European Workshop on Computational Geometry, URL http://www.sciencedirect.com/science/article/pii/S0925772106000770.

Affiliation:

Stephanie Kobakian
Queensland University of Technology
School of Mathematical Sciences, Science and Engineering Faculty, Brisbane, QLD, Australia
E-mail: stephanie.kobakian@qut.edu.au

Dianne Cook
Monash University
Department of Econometrics and Business Statistics, Melbourne, VIC, Australia
E-mail: dicook@monash.edu

Earl Duncan
Queensland University of Technology
School of Mathematical Sciences, Science and Engineering Faculty, Brisbane, QLD, Australia
E-mail: earl.duncan@qut.edu.au

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

Chapter 4

Visual Inference Study

This chapter tests the performance of the hexagon tile map display created using the algorithm discussed in 3. It outlines the lineup protocol method of visual inference used to test information visualisations. Using a two factor experimental design, the experiment contrasts the performance of participants when they viewed a choropleth map, and a hexagon tile map. The contrast also involved three types of spatial trends, one geographic trend, and two population related distributions. The results showed that participants did in fact more accurately find the population related distributions.

This chapter has been submitted to the journal *IEEE Transactions of Visualisation and Computer Graphics*.

RESEARCH STUDENTS CENTRE
Examination Enquiries: 07 3138 1839
Email: research.examination@qut.edu.au

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

Comparing the Effectiveness of the Choropleth Map with a Hexagon Tile Map for Communicating Cancer Statistics has been submitted to the IEEE Transactions on Visualization and Computer Graphics.

Contributor	Statement of contribution*
Stephanie Kobakian S R Kobakian 27/12/2019	Stephanie created the lineup displays used in the experiment, and used the template provided by the taipan package to create the survey app to create the web application for the survey that collected data. She also drafted and formatted for submission.
Prof. Dianne Cook	Prof. Dianne planned the experimental design used in the experiment. She was instrumental in the development of the lineup protocol for visual inference and its application to spatial data. Dianne edited and refined the drafts for publication.
Dr. Earl Duncan	Dr. Earl contributed significant ideas and code to produce the simulated null data sets for the lineups used in the study.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all Co-authors confirming their certifying authorship. (If the Co-authors are not able to sign the form please forward their email or other correspondence confirming the certifying authorship to the RSC).

Name Signature Date

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

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 choropleth map display is commonly used for communicating spatial distributions across geographic areas. However, when choropleths are used, the sizes of areas can lead to the misinterpretation of the distribution. The visualization method used to present geospatial data will influence the understanding of the distribution derived by map users. Choosing an effective alternative display could positively influence the communication of the spatial distribution. 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 design, and the choropleth is used as a comparison. The hexagon tile map display is also tested using a distribution that is directly related to the geography, with values monotonically decreasing from the North-West to South-East areas of Australia. This study finds the single map in a hexagon tile map lineup 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**

I. Introduction

This study compares the effectiveness of a new type of display, a hexagon tile map, against the standard, a choropleth map, for communicating information about disease statistics. The choropleth map is the traditional approach for visualizing aggregated statistics across administrative boundaries. 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. The hexagon tile map builds on existing displays, such as the cartogram, and tessellated hexagon displays. It differs in that it relaxes the requirement to have connected hexagons, and allows sparsely located hexagons. The algorithm to construct a hexagon tile map is available in the R package sugarbag [1]. This type of display may be useful for other countries, and other purposes.

The hexagon tile map was designed for Australia, motivated by a need to display spatial statistics for a new Australian Cancer Atlas. None of the existing approaches for creating cartograms or hexagon tiling perform well for the Australian landscape, which has vast open spaces and concentrations of population in small regions clustered on the coastlines.

The Australian Cancer Atlas [2] 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 for the atlas.

The experiment was conducted using the lineup protocol, a visual inference procedure [3], to objectively test the effectiveness of the two displays.

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

II. Background

A. Spatial data displays

Spatial visualisations communicate the distribution of statistics over geographic landscapes. The choropleth map [4], [5] 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 with colour 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.

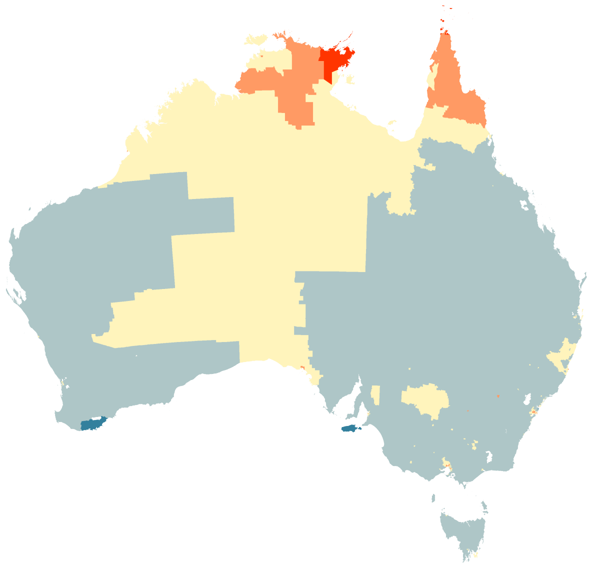


Fig. 1. A choropleth map of the smoothed average of liver cancer diagnoses for Australian males. The diverging colour scheme uses dark blue areas for much lower than average diagnoses, yellow areas with diagnoses around the Australian average, red shows diagnoses much higher than average. The hexagon tile map shows concentrations of higher than expected liver cancer rates in the cities of Melbourne and Sydney, which is not visible from the choropleth.

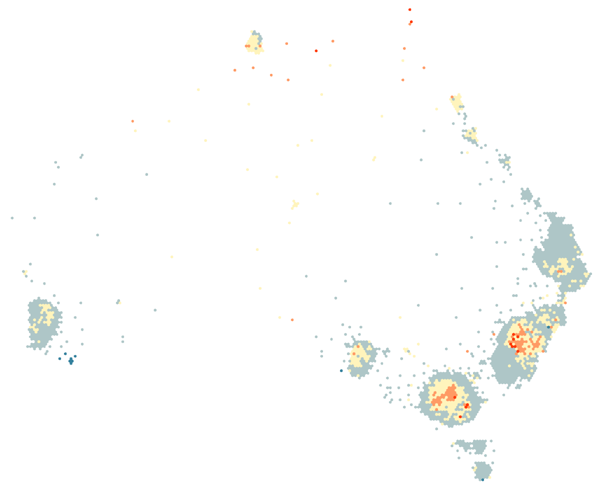


Fig. 2. A hexagon tile map of the smoothed average of liver cancer diagnoses for Australian males. The diverging colour scheme uses dark blue areas for much lower than average diagnoses, yellow areas with diagnoses around the Australian average, red shows diagnoses much higher than average. The hexagon tile map shows concentrations of higher than expected liver cancer rates in the cities of Melbourne and Sydney, which is not visible from the choropleth.

The choropleth map is effective 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. A 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 [6]. *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 [7], [6]; 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 population by map 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 representation of geographic units. The non-contiguous cartogram method [8] 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 [7] represents each unit as a circle, sized according to the value of the statistic. The neighbour relationships are mostly maintained by how the circles touch. A similar approach was pioneered by Raisz [9], using rectangles that tile to align borders of neighbours [10]. There have been thorough reviews of the array of methods, as suitable for cancer atlas displays [11], [5].

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 [1]. Figure 1 shows the hexagon tile map, along with the choropleth map of liver cancer rates in Australia. Colour maps from substantially below average (blue) to substantially above average (red) rates. The inner city areas are expanded out, 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.

B. Visual Inference

In order to assess the effectiveness of the hexagon tile map, the lineup protocol [3],[12] from visual inference procedures is employed. The procedures for doing a power comparison of competing plot designed, outlined in [13], are followed. The approach mirrors classical statistical inference.

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 [14] 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 a 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. 3 shows a lineup of the hexagon tile map, of size $m = 12$. Plot 3 is the data plot, 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.

III. 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?

Fig. 3. This lineup of twelve hexagon tile map displays contains one map with a real population related structure. The rest are null plots that contain spatial correlation between neighbours.

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.

Australia is used for the study, with Statistical Area 2 (SA2) [15] as the geographic units. The results should apply broadly to any other geographic area of interest.

A. Experimental factors

The primary factor in the experiment is the plot type. The secondary factor is a trend model. Three trend models were developed, one mirroring a large spatial trend for which the choropleth would be expected to do well, and two with differing level of inner city hot spots. These latter two reflect the structure seen in the liver cancer data (Fig. 1). This produces six treatment levels:

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

Data is generated for each of the trend models, with four replicates, and each 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, with as a choropleth or as a hexagon tile map. Table 4 summarises the design.

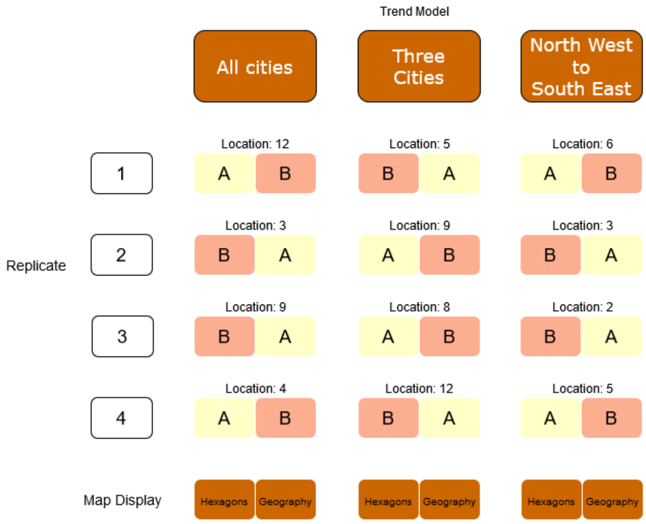


Fig. 4. The experimental design used in the visual inference study.

B. 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 association between variables, while maintaining marginal distributions. This is too simple for spatial data. In spatial data, a key feature is smoothness over the landscape corresponding spatial dependence. To do something simple, like permute the values relative to the geographic location would produce null plots which are too chaotic, and the data plot will be recognisable 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` [17] 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 of the Australian Cancer Atlas statisticians. 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, 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 allowed outliers, and showed distributions similar to the Liver cancer distribution shown in 1.

C. Generating lineups

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 trend model, the real data display was created by manipulating the centroid values of each geographic unit.

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. The distance to the closest capital city for each geographic unit was found when creating the hexagon tile map using the sugarbag [1] package. 201 Statistical Areas were considered greater capital city areas, the values of these areas were increased to create red clusters. They were chosen to make clusters around the cities visible, even if they were not overtly noticeable in the choropleth display.

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 values of the areas nearest to the three cities were increased to create clusters.

One of the lineup map locations was chosen to embed the real trend model data, in each of the four replicates, for the three trend models. In this location, real data display of the trend model was added to the null values of the spatially correlated data for each lineup. This location, the associated trend values and the null data were used to create both a choropleth and hexagon tile map display. The choropleth and hexagon lineups were split between group A and B, according to the experimental design shown in 4. The location was chosen from a sub sample of the 12 possible locations. The chance of reptition using resampling was introduced to prevent participants from inducing the location by elimination, the locations 1, 7, 10 and 11 were not used.

For each of the 144 individual maps, the values attributed to each geographic area were rescaled to create a similar color scale from deep blue to dark red within each map. This allowed at least one geographic unit to be colour dark blue, and at least one to be 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 more red than the rest of Australia.

D. Analysis

1) Data Cleaning: The first step in the data cleaning process involved checking that survey responses collected for each participants were only included once in the data set. The data cleaning process also involved filtering out participants' who did not provide any 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.

2) Descriptive statistics: Basic descriptive statistics were used to contrast the detection rate for the two types of displays. This was extended to consider the trend models, contrasting the mean and standard detection rate for each group, who had seen the different map display type for each trend model.

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 - reason for choice, reported difficulty, gender, age, education, having lived in Australia - against the design variables.

Plots will be made in R [18], with the `ggplot2` package [14].

3) Modelling: The likelihood of detecting the data plot in the lineup can be modelled using a linear mixed effects model. The R [18] `glmer()` function in the `lme4` [19] package implements generalised 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 $y_{ij} = 0, 1$ 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.

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. They are transformed with the link specified below:

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

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

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

Computation was performed using R [18], with the `lme4` package [19].

E. Web application to collect responses

The taipan [20] 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 shown the demographics and consent page. This asked participants for their Figure-Eight contributor ID, and their consent to the responses being used for analysis. The demographics collected included participants' preferred pronoun, 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, reason for their choice, and 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. Along with the responses to the three questions, the script 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 participant 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 googlesheet used to store results. The application connected to the googlesheet using the googlesheets [21] R package when participants opened the application, and interacted again when participants chosen to submit the survey and indicate their completion. At this time it added the participant's responses to the twelve lineup displays as twelve rows of data in the googlesheet.

F. Participants

Participants were recruited to evaluate lineups from the Figure Eight crowdsourcing platform [22]. The lineup protocol expects that the participants are uninvolved judges with no prior knowledge of the data, to avoid inadvertently affecting results. Potential participants needed to have achieved level 2 or level 3 from prior work. 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, which was hosted externally to the Figure Eight website. There were 95 participants involved in the study. All participants read introductory materials, and was trained using three test displays, to orient them to the evaluation task. All participants who completed the task were compensated $USD5 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 number of participants to collect responses from. From the pilot study amount of participants needed to acheieve a 0.05 power level was determined to be

G. Demographic 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 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.

IV. Results

Responses from 95 participants were collected. Five participants did not provide more than three answers 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 are analysed on accuracy and speed. The certainty and reasons of subjects in their answers is also examined.

A. 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.

B. Accuracy

Fig. 5 displays the average detection rates for the two types of plot separately for each trend model. Each trend model was tested using four repetitions, evaluations on the same data set were seen as either choropleths or hexagon tile maps by each group as specified in Table. 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 type. 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. I shows the means and standard deviations of the detection rate for each type of plot and each trend model. This also gives the standard errors, the smallest standard deviation for all sets of replicates was the Three Cities trend model

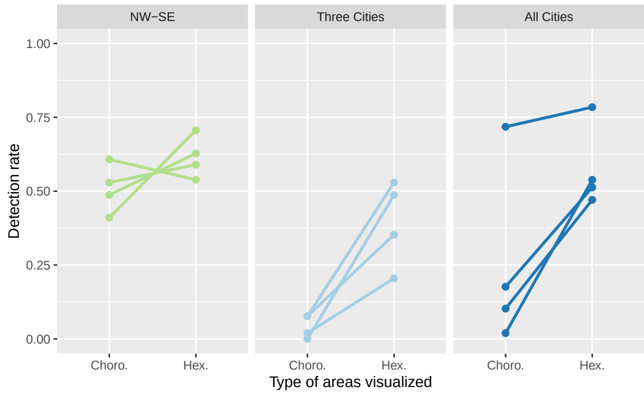


Fig. 5. The detection rates achieved by participants are contrasted when viewing the four replicates of the three trend models. Each point shows the probability of detection for the lineup display, the facets separate the trend models hidden in the lineup. The points for the same data set shown in a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

TABLE I
THE RATE OF DETECTION FOR EACH TREND MODEL HAS BEEN CALCULATE FOR THE CHOROPLETH AND HEXAGON TILE MAP DISPLAYS. THE ASSOCIATED STANDARD ERRORS ARE ALSO INCLUDED.

Type	NW-SE	Three Cities	All Cities
Choro.	((0.50)) (0.50)	((0.19)) (0.19)	((0.42)) (0.42)
Hex.	((0.49)) (0.49)	(0.40) (0.49)	((0.49)) (0.49)

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. II presents a summary of the generalised 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. I. 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 decreases for population related displays for a choropleth map lineup, but increases for a hexagon tile map display. The log odds of detection show in Table. II 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.

TABLE II
THE MODEL OUTPUT FOR THE GENERALISED 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

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.

C. 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 are also separated by trend model and whether the data plot was detected or not detected. The time taken to complete an evaluation ranged from milliseconds to 60 seconds. The average time taken for 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 less correct detections for choropleth lineups for the Three cities and All Cities trends.

D. 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?’. Figure. IV-D shows the amount 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 chose 4 or 5 often 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

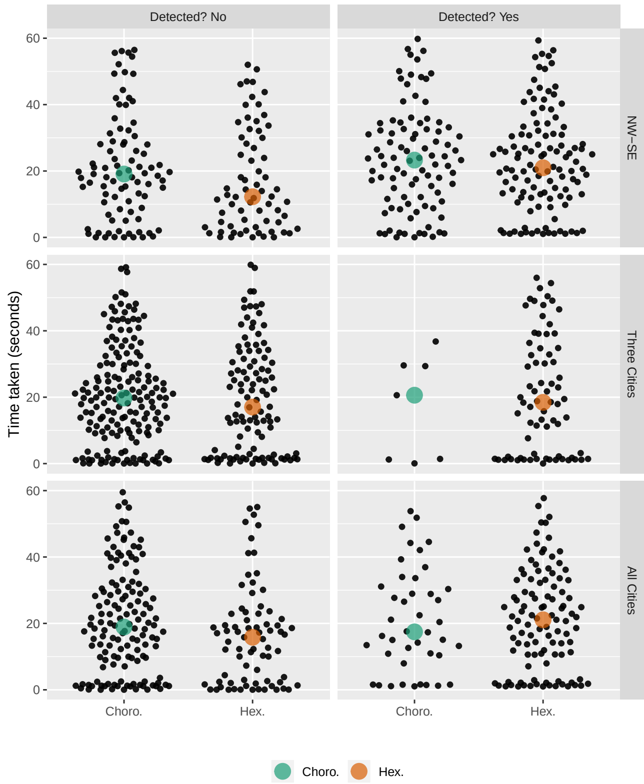


Fig. 6. The distribution of the time taken (seconds) to submit a response for each combination of trend, whether the data plot was detected, and type of display, shown using horizontally jittered dotplots. The colored point indicates average time taken for each plot type. Although some participants take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

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.

E. 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 minimise noise from participants who did not select a response.

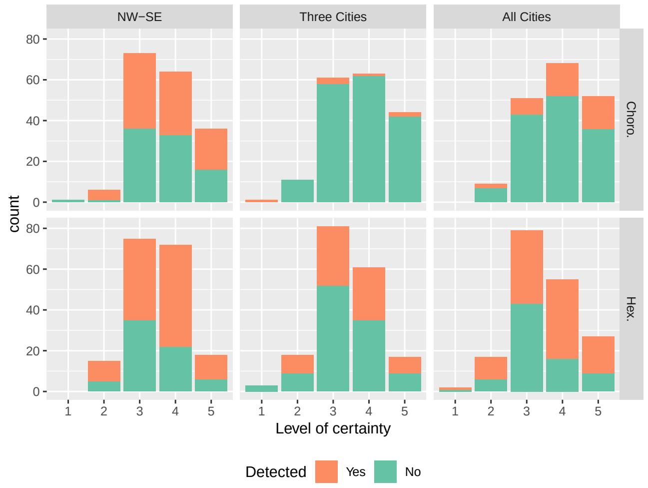


Fig. 7. The amount of times each level of certainty was chosen by participants when viewing hexagon tile map or choropleth displays. Participants were more likely to choose a high certainty when considering a Choropleth map. The mid value of 3 was the default certainty, it was chosen most for the Hexagon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

TABLE III
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.

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

V. Discussion

The intention of this study was to contrast the use of the choropleth map and the hexagon tile map. The visual inference lineup protocol was employed to contrast the effectiveness of the displays. The results have shown that overall the use of the hexagon tile map display allows participants to find the data plot in the lineup more often. Using the visual inference protocol this result can be extended to show that it is a valid alternative display to communicate spatial distributions of population related data.

We expected that the choropleth map would be superior for communicating the spatial pattern of geographic distributions. The data suggest that the participants perform slightly better or equally as well for each replicate in each trend model across the two displays. Table II shows that the difference in the mean detection rate for the two trend models was 0.10.

The differences seen in the compare detect plot and Table. I are reflected in the model results. Surprisingly the difference scene for the geographic distribution was significant at the 0.05 level. It also showed that the hexagon tile map display performs marginally better than the choropleth for all trend models. Unexpectedly the detection rate suffers when using a choropleth map to display population related distributions.

While the significance of the difference in detection was the key focus of this experiment, the secondary focus was the time taken by participants. it was expected that the disciplines 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 the Fig. 6 display showed very little difference in the mean evaluation times. As the ranges 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. **??** shows high levels of certainty of 4 and 5 were chosen by participants when looking at the population distributions in a choropleth map display show that they were over confident 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 most common reason for a participants choice. The All Cities displays colored the inner-city areas of all capital cities more red, this was observable to participants and explains the equal choice of the city clusters or rural color consistency. Choosing "Clusters of colour" 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.

VI. Conclusion

The choropleth map display and the tessellated hexagon tile map have been contrasted using the lineup protocol. The hexagon tile map was significantly more effective for spotting a real population related data trend model hidden in a lineup.

The hexagon tile map display should be considered as an alternative visualization method when communicating distributions that relate to the population across a set of geographic units. As an additional display to the familiar choropleth map, Cancer Atlas products may benefit from the opportunity to allow exploration via an alternative display. The spatial distributions used to test these displays were inspired by the real spatially smoothed estimates of the cancer burden on Australian communities. However, this technique may be extended to other population related distributions, such as other diseases.

The increasing population densities of capital cities despite large land area exacerbates the difference in the smallest and largest communities. The population density structure of Australia can be considered similar to that of Canada, New Zealand and many other countries. Therefore, this display is not only relevant to Australia, but all nations or population distributions that experience densely populated cities separated by vast rural expanses.

VII. 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 [20] R package for image tagging, used as the base for the web app constructed to collect participant evaluations of lineups. We are thankful for 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 on GitHub. Supplementary materials have been included to discuss the survey procedures and the lineups that were used. The full set of images can also be found at: https://github.com/srkobakian/experiment/tree/master/figures/final

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 [18] with the use of the following packages:

- For document creation: rmarkdown [23], rticles [24], knitr [25].
- For lineup creation and data analysis: tidyverse [26], nullabor [27], ggthemes [28], RColorBrewer [29].
- For image displays: cowplot [30], png [31], grid [32].
- For modelling and presentation of models: gstat [17], lme4 [19], 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. Kobakian S, Cook D (2019) Sugarbag: Create tessellated hexagon maps.

2. Queensland CC Australian Cancer Atlas (https://atlas.cancer.org.au).

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

4. Tufte ER (1990) Envisioning Information. Graphics Press

5. Skowronnek A (2016) Beyond Choropleth Maps – A Review of Techniques to Visualize Quantitative Areal Geodata. In: Infovis Reading Group WS 2015/16. https://alsino.io/static/papers/BeyondChoropleths_AlsinoSkowronnek.pdf.

6. Kocmoud C, House D (1998) A Constraint-based Approach to Constructing Continuous Cartograms. In: Proc. Symp. Spatial data handling. pp 236–246

7. Dorling D (2011) Area Cartograms: Their Use and Creation. In: Concepts and techniques in modern geography (catmog). pp 252–260

8. Olson JM (1976) Noncontiguous Area Cartograms. The Professional Geographer 28:371–380

9. Raisz E (1963) Rectangular Statistical Cartograms of the World. Journal of Geography 35:8–10

10. Monmonier M (2005) Cartography: Distortions, Worldviews and Creative Solutions. Progress in Human Geography 29:217–224

11. (2019) Cancer applications of choropleth maps, and the potential of cartograms and alternative map displays. Manuscript submitted for publication.

12. Buja A, Cook D, Hofmann H, Lawrence M, Lee E-K, Swayne DF, Wickham H (2009) Statistical inference for exploratory data analysis and model diagnostics. Philosophical Transactions of the Royal Society, A (Invited) 367:4361–4383

13. Hofmann H, Follett L, Majumder M, Cook D (2012) Graphical tests for power comparison of competing designs. IEEE Transactions on Visualization and Computer Graphics 18:2441–2448

14. Wickham H (2009) Ggplot2: Elegant graphics for data analysis. Springer New York

15. (2018) Australian Statistical Geography Standard (ASGS). Australian Bureau of Statistics

16. Matheron G (1963) Principles of geostatistics. Economic Geology 58:1246–1266

17. Gräler B, Pebesma E, Heuvelink G (2016) Spatio-temporal interpolation using gstat. The R Journal 8:204–218

18. R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

19. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67:1–48

20. Kobakian S, O'Hara-Wild M (2018) Taipan: Tool for annotating images in preparation for analysis.

21. Bryan J, Zhao J (2018) Googlesheets: Manage google spreadsheets from r.

22. (2019) The essential high-quality data annotation platform. Figure Eight

23. Xie Y, Allaire JJ, Grolemund G (2018) R markdown: The definitive guide. Chapman; Hall/CRC, Boca Raton, Florida

24. Allaire J, Xie Y, R Foundation, et al (2019) Rticles: Article formats for r markdown.

25. Xie Y (2014) Knitr: A comprehensive tool for reproducible research in R. Implementing reproducible computational research

26. Wickham H, Averick M, Bryan J, et al (2019) Welcome to the tidyverse. Journal of Open Source Software 4:1686

27. Wickham H, Chowdhury NR, Cook D, Hofmann H (2018) Nullabor: Tools for graphical inference.

28. Arnold JB (2019) Ggthemes: Extra themes, scales and geoms for 'ggplot2'.

29. Neuwirth E (2014) RColorBrewer: ColorBrewer palettes.

30. Wilke CO (2019) Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'.

31. Urbanek S (2013) Png: Read and write png images.

32. R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

33. Zhu H (2019) KableExtra: Construct complex table with 'kable' and pipe syntax.

Chapter 5

Discussion

This chapter discusses the use of the choropleth map in presenting spatial data. It discusses the hexagon tile map algorithm contributed by this thesis and discusses its similarities and differences from other alternative visualisation methods. It discusses the use of the visual inference study for evaluating the displays produced by the hexagon tile map algorithm for Australia. Finally, it discusses how animating between a choropleth and hexagon display can allow map creators to direct the attention of users to the overlooked, small but densely populated geographic areas.

5.1 Presenting spatial data

Choropleth map displays are the traditional, and current practice method for presenting geospatial data. The use of familiar geographic polygons allows users to readily find areas of interest on the map display (Tufte, 1990). The spatial patterns among the areas are revealed via the colour used to fill each geographic area.

Cancer Atlases are used to develop hypotheses about spatial distributions of cancer statistics (Bell et al., 2006). Hot spots with high values can be detected by dramatic changes in colour between neighbours (d'Onofrio et al., 2016).

However, when used as a tool communicate spatial distributions the use of the choropleth map may lead to misinterpretation of the overall distribution. This is because of the overemphasis on the large geographic areas, and the lack of visibility for the small inner-city communities (Dorling, 2011).

Visualisation methods have improved in iterations over many years. Cartograms (Dougenik, Chrisman, and Niemeyer, 1985) showed great promise and several algorithms were presented to create cartograms in the 1960s and 1970s. Cartogram techniques highlighted the population by expanding the map area of densely populated communities. The introduction of computer assisted cartogram techniques were developed in the 1970s and 1980s. Dorling (Dorling, 2012), (Dorling, 2011) introduced several alternative visualisation methods and their use has had a profound impact in the communication of data with population related distributions. These displays enabled viewers to physically see the densely populated areas and avoid spending attention on less populated, but geographically large map areas (Griffin, 1980).

5.2 The hexagon tile map algorithm

A key aim of this thesis was to present an algorithm for displaying geospatial data. The algorithm presented in chapter algorithm is the result of many iterations in the process necessary to transform geographic areas into a tessellated hexagon tile map.

The tessellation employed in the hexagon tile map algorithm maintains connectedness between neighbouring areas, this draws inspiration from contiguous cartograms (Min Ouyang and Revesz, 2000), rectangular cartograms (Raisz, 1963) and Dorling's circular cartograms (Dorling, 2011). However, the hexagon tile map algorithm does not employ the gravitation pull mathematics that is used to create contiguous cartograms. It also does not iterate on the placement of hexagons.
The choice of the hexagon, a consistent shape used for all areas, also draws from rectangular and Dorling cartograms. This encourages map readers to focus on the similarities or difference in the colour between geographic neighbours, and does not distract them with unfamiliar boundaries after a contiguous cartogram transformation.

The hexagon tile map display is least like the non-contiguous cartogram (Olson, 1976) which maintains the familiar geographic shapes. Some similarity in the final results may occur as densely populated regions maintain connectivity with neighbours, and sparsely populated areas will use the underlying white map space to separate neighbours.

The implementation of the `sugarbag` (Kobakian and Cook, 2019) package for R (R Core Team, 2019a) provides map creators the opportunity to apply the algorithm to any set of geospatial polygons. This does not require manual creation of grids, but the displays are reusable for any populated related data set that uses the same set of geographic units.

5.3 Visual inference study

Conducting a visual inference study allowed a contrast of the traditional choropleth display to the alternative hexagon tile map.

The visual inference study showed that the hexagon tile map display is a valid alternative as participants achieved higher detection rates for the hexagon tile map when presenting the geographic spatial distribution and the population related distributions.

5.4 Animating between displays

Interactivity can allow greater detail to be drawn about small neighbourhoods in a choropleth map. By zooming in, the same shapes shown on the initial map become much larger and comparisons between immediate neighbours becomes simpler. Interacting can also allow the map user to drive their exploration. In the case of the Australian Cancer Atlas, the interactive tools allow quick comparisons between areas far from each other on the map. It also allows quick comparisons of the estimated cancer statistics by showing the colours for each cancer type in a table-like display. These additions allow user driven exploration, but do not guarantee that the spatial distribution across the country is digested accurately.

Animating between a choropleth and a hexagon tile map will allows map users to understand how the small communities of a whole country are affected simultaneously. It also teaches map users how to find areas of interest as their attention is drawn to the capital cities, that may not have caught their attention in the display of the choropleth map. When communicating cancer statistics, there should be a balance between providing people a familiar landscape and ensuring the interpret the spatial distribution correctly. A directed exploration of a collection of inner city neighbourhoods will take the distribution out of context, the Australian Cancer Atlas provides some visual tools to combat this by highlighting that the areas displayed are only subsets However, allowing the values of an inner-city neighbourhood to be visible in the context of all other areas will enhance the understanding of the spatial distribution. Animations will communicate a specific message through the capture and direction of users' attention.

The animations created of the Australian Statistical Areas at Level 2, used in the Australian Cancer Atlas, highlight just how many SA2 areas are hidden due to their size in the choropleth display. The rapid shrinking of the rural areas allows users to recognise the emphasis that had been placed on the large geographic regions. Their attention is then drawn to the expanding capital cities, and the coastal towns that were not previously visible.

Chapter 6

Conclusion

The first aim of this thesis was to present an alternative visualisation method for spatial data. This thesis has provided a new algorithm to present spatial distributions of disease data, and includes an R code (R Core Team, 2019a) implementation. The spatial data sets that will be effectively communicated by this display will likely have population related distributions. The hexagon tile map display will represent each area equally on the map space to effectively convey the spatial distribution of the set.

The hexagon tile map visualisation method solves the misrepresentation problem of choropleth displays of geographic data sets that contain a substantial amounts of areas. This algorithm is accessible to all R users, in a set of simple functions. It can be applied to any set of areas in an `sf` (Pebesma, 2018) object.

The tile map allocation provided by the `allocate` function in the `sugarbag` package can easily be used to create animations between a choropleth and hexagon tile map display. Linking the familiar geography to the effective display for understanding the distribution across many heterogeneous geographic regions.

The effectiveness of the hexagon tile map has been proved by the visual inference study. It showed that participants could recognise the data display in the set of null distributions more frequently when viewing a hexagon tile map display. The choropleth map display is still effective for distributions that are directly related to the geography, such as the North-West to South-East distribution used in the study. This has expanded the applications of visual inference studies in a spatial data context.

Future work will include expanding on criteria to evaluate the hexagon tile maps produced by the algorithm. The methods to evaluate the alternative displays have not been thoroughly explored in this thesis. This framework will be used to create relevant tests that contrast the use of the map area, and changes in the visual statistics when the parameter of the hexagon tile map algorithm are altered.

The current hexagon tile map creates a template map that can be used to visualise any data set that contains the areas used to create the map. There is the possibility of allowing a bivariate display to incorporate uncertainty by using a colour scheme that operates in two directions as suggested by Lucchesi and Wikle (Lucchesi and C.K., 2017). The animation methods that allow colours filling the hexagons to flicker communicate uncertainty could also be employed.

With large hexagons, there is a potential to incorporate geofacets (Hafen, 2019) to create a tessellated display of small visualisations for each geographic unit. These displays become increasingly complex if the visualisation becomes more detailed, or the hexagons become smaller.

This work has contributed a new visualisation for spatial data sets. The spatial distributions of cancer burden for different types of cancers largely relates to the population rather than the geography. The alternative visualisation method highlights the communities, and the hexagon tile map may be implemented in future iterations of the Australian Cancer Atlas to improve the communication of spatial patterns of cancer burden on Australian communities. For wide use by map creators and those interested in alternative visual displays, the code implementation has been provided to any R user with examples and documentation. This work has also contributed to the literature of visual inference studies, by using the “lineup” protocol developed by Buja et al. and used by Wickham et al. (2010), and Hofmann et al. (2012). This example showed there was a difference in the rate of pattern recognition when participants saw competing spatial map displays.

To communicate human related spatial patterns of disease, map creators should consider the use of alternative displays. The hexagon tile map display has proven effective in this thesis for communicating spatial distributions in sets of heterogeneous geographic units. This thesis provides a practical guide for map creators communicating spatial displays of cancer data in Australia.

Appendix A

Visual inference study lineups

```
library(tidyverse)
library(readxl)
library(cowplot)
library(png)
library(grid)
library(lme4)
library(ggthemes)
library(RColorBrewer)
library(knitr)
library(kableExtra)

trend_colours <- c(
  "NW-SE" = "#B2DF8A",
  "Three Cities" = "#A6CEE3",
  "All Cities" = "#1F78B4")

type_colours <- c(
  "Choro." = "#fcae91",
  "Hex." = "#a50f15")
```

```
detect_f_colours <- c(
  "No" = "#66C2A5",
  "Yes" = "#FC8D62")

detect_colours <- c(
  "Detected? No" = "#66C2A5",
  "Detected? Yes" = "#FC8D62")

  # Downloaded data
d <- read_xlsx("data/experiment-export.xlsx", sheet=2) %>%
  filter(!is.na(contributor)) %>%
  mutate(contributor = factor(contributor))

# Check data set
# Need to clean multiple entries, 48, 24
# remove duplicated entries due to submit button
d <- d %>% group_by(group, contributor, image_name) %>%
  slice(1) %>% ungroup() %>%
  arrange(group, contributor, plot_order)

# Remove contributors who did not provide answers to most questions
keep <- d %>% count(contributor, sort = TRUE) %>% filter(n > 10)
d <- d %>%
  filter(contributor %in% keep$contributor) %>%
  filter(contributor != "1234567890")

# Remove contributors who did not provide any choices
bad_contribs <- d %>% group_by(contributor) %>%
  summarise(sum0 = sum(choice)) %>%
  filter(sum0 == 0) %>%
```

```r
  pull(contributor)


d <- d %>%
  filter(!(contributor %in% bad_contribs))



n_contributors <- d %>% count(contributor, sort=TRUE) %>%
  summarise(n_contributors = length(contributor))


d <- d %>% mutate(certainty = factor(as.character(certainty),
  levels = c("1", "2", "3", "4","5"), ordered=TRUE))
```

```r
replicate <- tibble(image_name = c("aus_cities_12_geo.png", "aus_cities_12_hex.png",
                                   "aus_cities_3_geo.png", "aus_cities_3_hex.png",
                                   "aus_cities_4_geo.png", "aus_cities_4_hex.png",
                                   "aus_cities_9_geo.png", "aus_cities_9_hex.png",
                                   "aus_nwse_2_geo.png", "aus_nwse_2_hex.png",
                                   "aus_nwse_3_geo.png", "aus_nwse_3_hex.png",
                                   "aus_nwse_5_geo.png", "aus_nwse_5_hex.png",
                                   "aus_nwse_6_geo.png", "aus_nwse_6_hex.png",
                                   "aus_three_12_geo.png", "aus_three_12_hex.png",
                                   "aus_three_5_geo.png", "aus_three_5_hex.png",
                                   "aus_three_8_geo.png", "aus_three_8_hex.png",
                                   "aus_three_9_geo.png", "aus_three_9_hex.png"),
                    replicate = c(1, 1, 2, 2, 3, 3, 4, 4,
                                  1, 1, 2, 2, 3, 3, 4, 4,
                                  1, 1, 2, 2, 3, 3, 4, 4))
# Add rep info to data
d <- d %>% left_join(., replicate, by = "image_name")
```

```
# Tidy for analysis
d <- d %>%
  separate(image_name, c("nothing", "trend", "location", "type", "extra"), remove = FALSE
  select(-nothing, -extra) %>%
  mutate(location = as.numeric(location),
    # detect measures the accuracy of the choice
        detect = ifelse(location == choice, 1, 0)) %>%
  mutate(trend = case_when(
    trend == "nwse" ~ "NW-SE",
    trend == "cities" ~ "All Cities",
    trend == "three" ~ "Three Cities")) %>%
  mutate(trend = fct_relevel(trend, "NW-SE","Three Cities","All Cities")) %>%
  mutate(type = case_when(
    type == "hex" ~"Hex.",
    TRUE~"Choro.")) %>%
    mutate(detect_f = factor(detect, levels = c(0,1), labels = c("Detected? No", "Detecte

plots <- d %>% group_by(group, trend, type, location) %>%
  # pdetect measures the aggregated accuracy of the choices
  summarise(pdetect = length(detect[detect == 1])/length(detect))
```

A.1 Overall Performance

The detection rate is considered for each lineup. The detection rates for group A were less varied than the detection rates for the lineups seen by group B. Fig. A.1 shows the distribution using a boxplot. This shows the median value for detection rate was extremely similar.

```
plots %>% ggplot(aes(x = group, y = pdetect)) +
  geom_boxplot() +
  geom_jitter(width = 0.1) +
```

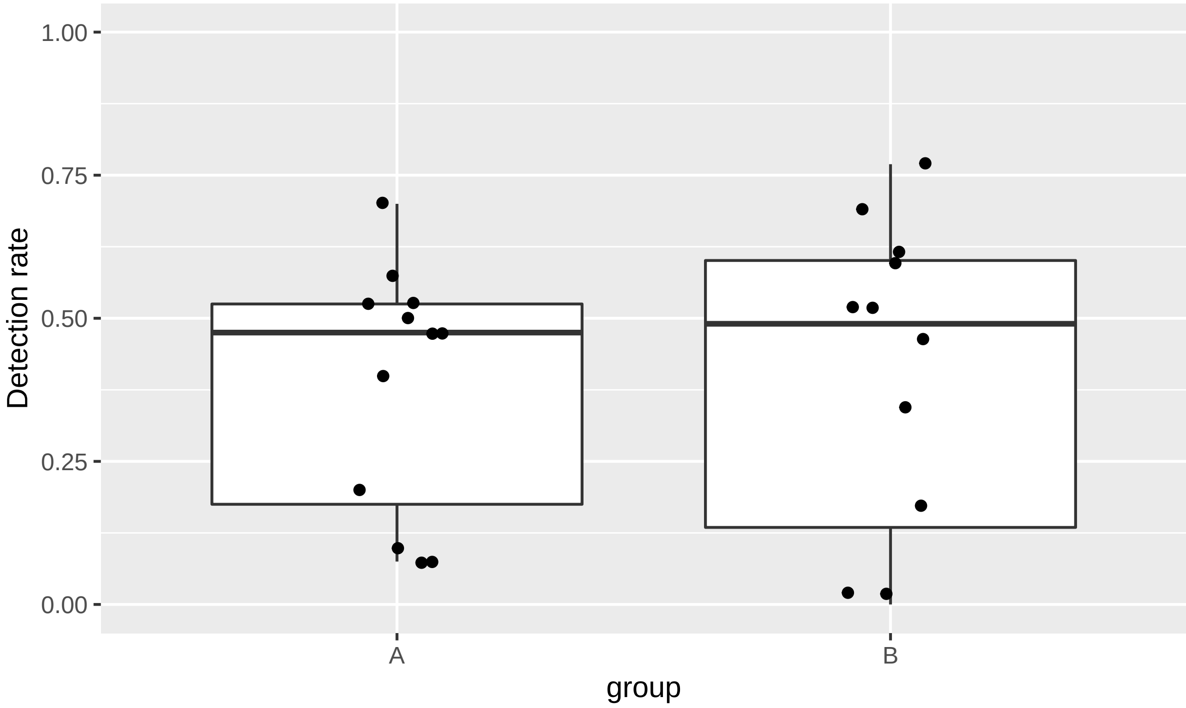


Figure A.1: *Boxplots of the distribution of detection rates for each line up, separated by group.*

```
  ylab("Detection rate") +
  ylim(c(0,1))
```

The overall detection rate is considered for each trend model in Fig. A.2. The detection rates for the NW-SE trend model was less varied than the detection rates for the Three Cities and All Cities trends. Fig. A.1 shows the distribution using a boxplot. This shows the distributions of the rates do not overlap for NW-SE and Three Cities trends, the Three cities range was larger, but the median was much higher for the NW-SE trend. The All Cities trend model distribution overlaps with the NW-SE and All Cities trends.

```
plots %>% ggplot(aes(x = trend, y = pdetect, fill = trend)) +
  geom_boxplot() +
  scale_fill_manual(values = trend_colours) +
  geom_jitter(width = 0.1) +
  ylab("Detection rate") +
  ylim(c(0,1))
```

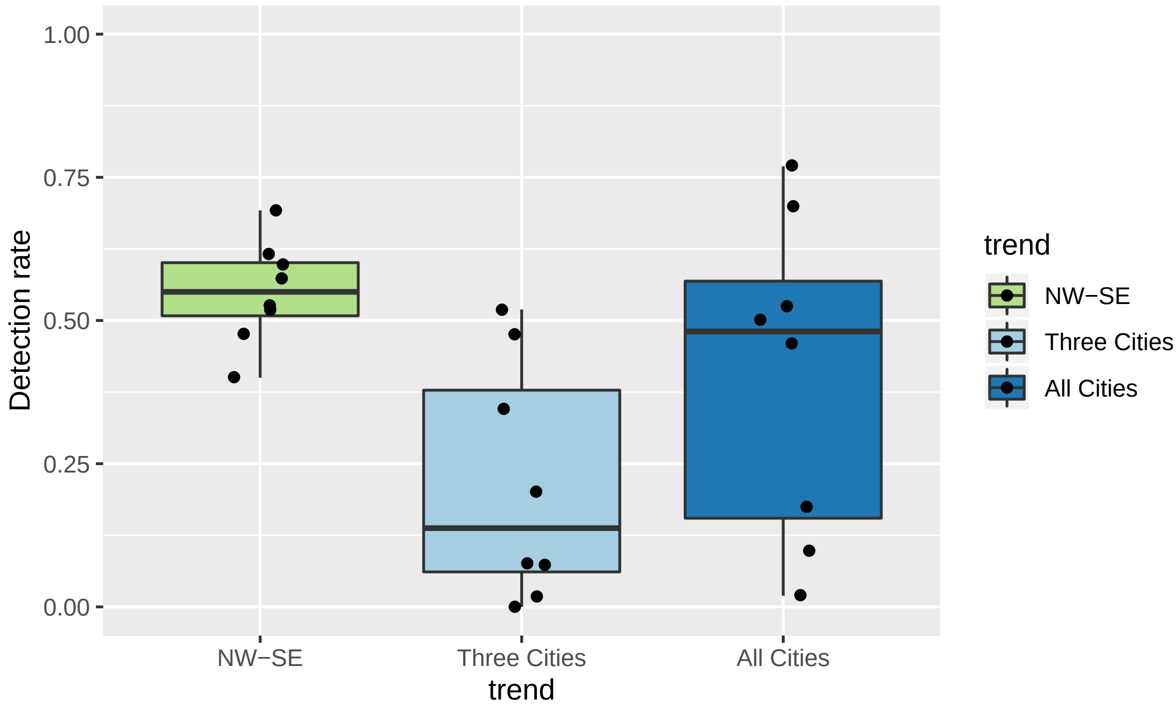


Figure A.2: *Boxplots of the distribution of detection rates for each line up, separated by trend model.*

The boxplots in Fig. A.3 contrast the distribution of the detection rates for eah type of display. The detection rates across the lineups was less varied for the hexagon display. There was a large difference in the medians for the types of displays. Without considering the relationship for each lineup, the hexagon lineup display allowed the participants to achieve higher detection rates.

```
plots %>% ggplot(aes(x = type, y = pdetect, fill = type)) +
  geom_boxplot() +
  geom_jitter(width = 0.1) +
  scale_fill_manual(values = type_colours) +
  ylab("Detection rate") +
  ylim(c(0,1))
```

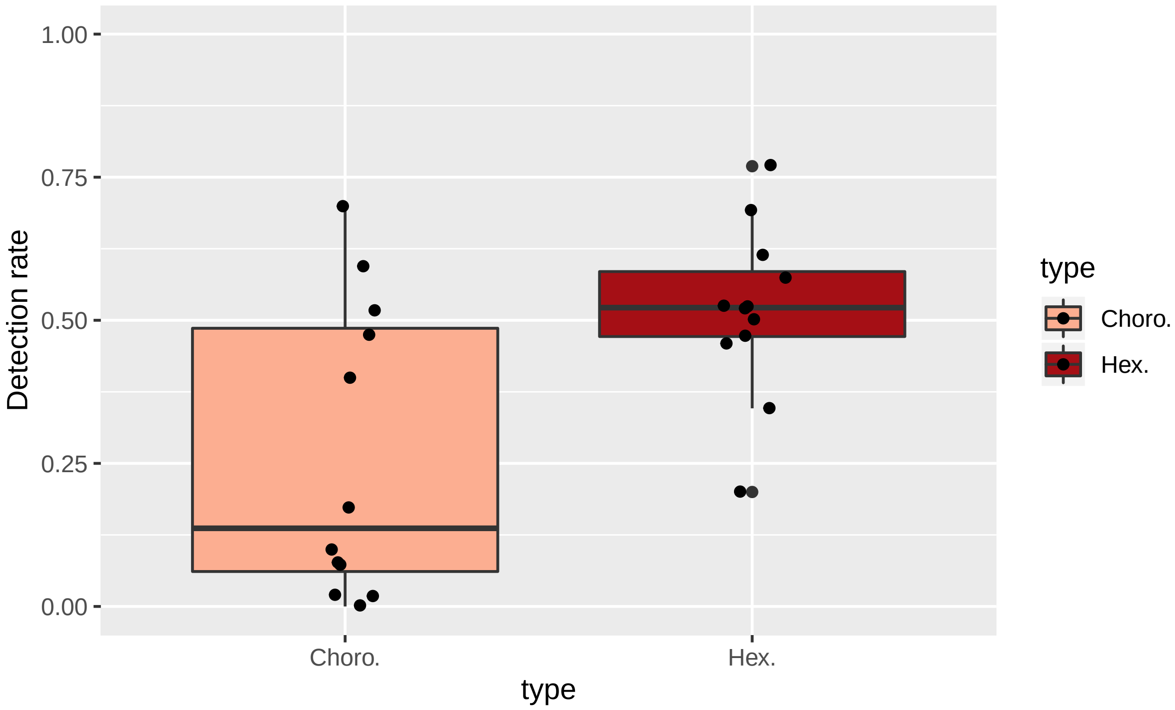


Figure A.3: *Boxplots of the distribution of detection rates for each line up, separated by type of display.*

A.2 Lineups

Each lineup had twelve map displays, this gave participants the choice of any plot, and the choice to not provide a response. this non-reponse is indicated by 0. The choices made by participants are displayed in Fig. A.4. The height of each orange lollipop indiates the proportion of participants that selected the map display of real data, they represent the correct choices. The green lollipops show the proportion of participants that selected the incorrect displays in each lineup.

The proportion of choices are also presented separately for each trend model in Table. A.1, Table. A.2, and Table. A.3. The correct map display in lineups with a North West to South East trend was chosen correctly with much greater frequency. In the lineups of All Cities displays, participants were misled by the choropleth display, but not the hexagon display for all except (2). 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 Fig. A.4.

```
d %>%

  count(type, trend, choice, replicate, choice, detect_f) %>%

  group_by(type, trend, replicate) %>%

  mutate(prop = n/sum(n)) %>%

  mutate(repl = paste("Rep:", replicate, ":\n", type,  sep = "")) %>%

  mutate(bottom = 0) %>%

  ggplot() +

  geom_point(aes(x = choice, y = prop, color = detect_f), size = 3) +

  geom_segment(aes(x = choice, xend = choice,y = bottom, yend = prop, colour = detect_f),

  facet_grid(repl ~ trend,

    drop = TRUE, as.table = TRUE, scales = "free_y") +

  labs(x = "Choice of plot in lineup", y = "Amount of choices") +

  scale_colour_manual(values = detect_colours) +

  scale_x_continuous(breaks = seq(from = 0, to = 12)) +

  scale_y_continuous(breaks = seq(from = 0.0, to = 1.0, by = 0.1)) +

  theme(legend.position = "bottom") +

  guides(colour = FALSE, fill = FALSE) +

  theme(strip.text.y = element_text(angle = 0))
```

```
choice_tabs <- d %>%

  count(trend, type, choice, replicate, choice, detect) %>%

  group_by(trend, type, replicate) %>%

  mutate(prop = as.character(round(n/sum(n), 2))) %>%

  rowwise() %>%

  mutate(prop = ifelse(nchar(prop)==3, paste0(prop,0), prop)) %>%

  mutate(repl = paste("Rep:", replicate, ":\n", type,  sep = "")) %>%

  mutate(bottom = 0) %>%

  select(Trend = trend, Rep = replicate, Type = type, prop, choice) %>%

  spread(choice, prop, fill = "0.00") %>%
```

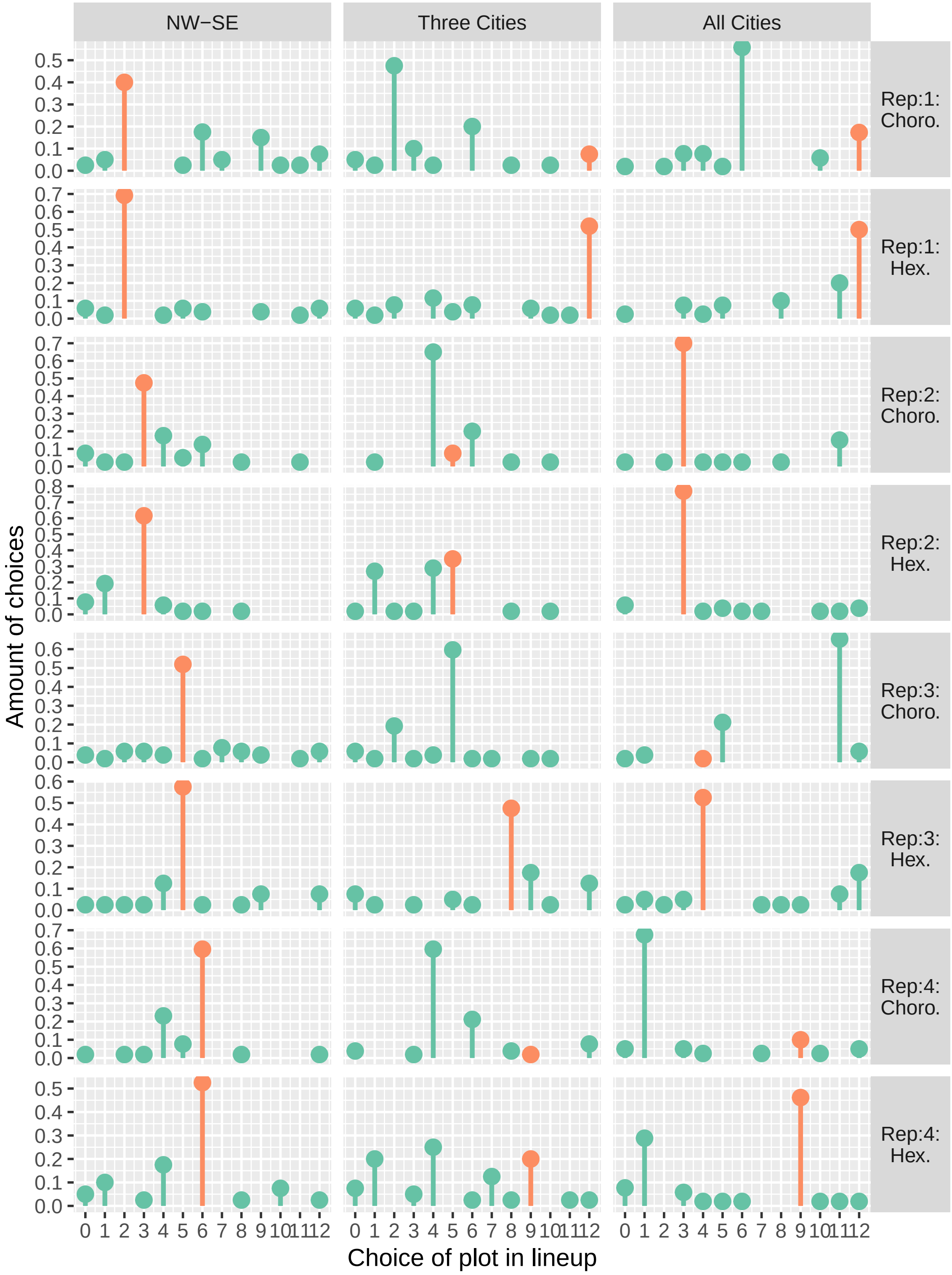


Figure A.4: *Pin plots of the propotion of choices made for each lineup location. Each facet is associated with one lineup, the height of the points show the proportion of the participants that made each choice when considering each lineup. The points coloured orange show the map which contained a trend 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.*

Table A.1: *The proportion of participants who selected each of the twelve map choices in each lineup for NW-SE displays.*

Rep	Type	0	1	2	3	4	5	6	7	8	9	10	11	12
1	Choro.	0.02	0.05	0.4	0	0	0.02	0.18	0.05	0	0.15	0.02	0.02	0.08
	Hex.	0.06	0.02	0.69	0	0.02	0.06	0.04	0	0	0.04	0	0.02	0.06
2	Choro.	0.08	0.02	0.02	0.48	0.18	0.05	0.12	0	0.02	0	0	0.02	0
	Hex.	0.08	0.19	0	0.62	0.06	0.02	0.02	0	0.02	0	0	0	0
3	Choro.	0.04	0.02	0.06	0.06	0.04	0.52	0.02	0.08	0.06	0.04	0	0.02	0.06
	Hex.	0.02	0.02	0.02	0.02	0.12	0.57	0.02	0	0.02	0.08	0	0	0.08
4	Choro.	0.02	0	0.02	0.02	0.23	0.08	0.6	0	0.02	0	0	0	0.02
	Hex.	0.05	0.1	0	0.02	0.18	0	0.52	0	0.02	0	0.08	0	0.02

```
  nest(data = c(Rep, Type, `0`, `1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`,
    `10`, `11`, `12`))
```

```
knitr::kable(choice_tabs$data[[1]], format = "latex",
  align = "rrrrrrrrrrrrrr", escape = FALSE, booktabs = TRUE,
  caption = "The proportion of participants who selected each of the twelve map choices i
  collapse_rows(., columns = c(1,2))
```

```
knitr::kable(choice_tabs$data[[2]], format = "latex",
  align = "rrrrrrrrrrrrrr", escape = FALSE, booktabs = TRUE,
  caption = "The proportion of participants who selected each of the twelve map choices i
  collapse_rows(., columns = c(1,2))
```

```
knitr::kable(choice_tabs$data[[3]], format = "latex",
  align = "rrrrrrrrrrrrrr", escape = FALSE, booktabs = TRUE,
  caption = "The proportion of participants who selected each of the twelve map choices i
  collapse_rows(., columns = c(1,2))
```

Table A.2: *The proportion of participants who selected each of the twelve map choices in each lineup for Three Cities displays.*

Rep	Type	0	1	2	3	4	5	6	7	8	9	10	11	12
1	Choro.	0.05	0.02	0.48	0.1	0.02	0	0.2	0	0.02	0	0.02	0	0.08
	Hex.	0.06	0.02	0.08	0	0.12	0.04	0.08	0	0	0.06	0.02	0.02	0.52
2	Choro.	0	0.02	0	0	0.65	0.08	0.2	0	0.02	0	0.02	0	0
	Hex.	0.02	0.27	0.02	0.02	0.29	0.35	0	0	0.02	0	0.02	0	0
3	Choro.	0.06	0.02	0.19	0.02	0.04	0.6	0.02	0.02	0	0.02	0.02	0	0
	Hex.	0.08	0.02	0	0.02	0	0.05	0.02	0	0.48	0.18	0.02	0	0.12
4	Choro.	0.04	0	0	0.02	0.6	0	0.21	0	0.04	0.02	0	0	0.08
	Hex.	0.08	0.2	0	0.05	0.25	0	0.02	0.12	0.02	0.2	0	0.02	0.02

Table A.3: *The proportion of participants who selected each of the twelve map choices in each lineup for All Cities displays.*

Rep	Type	0	1	2	3	4	5	6	7	8	9	10	11	12
1	Choro.	0.02	0	0.02	0.08	0.08	0.02	0.56	0	0	0	0.06	0	0.17
	Hex.	0.02	0	0	0.08	0.02	0.08	0	0	0.1	0	0	0.2	0.5
2	Choro.	0.02	0	0.02	0.7	0.02	0.02	0.02	0	0.02	0	0	0.15	0
	Hex.	0.06	0	0	0.77	0.02	0.04	0.02	0.02	0	0	0.02	0.02	0.04
3	Choro.	0.02	0.04	0	0	0.02	0.21	0	0	0	0	0	0.65	0.06
	Hex.	0.02	0.05	0.02	0.05	0.52	0	0	0.02	0.02	0.02	0	0.08	0.18
4	Choro.	0.05	0.68	0	0.05	0.02	0	0	0.02	0	0.1	0.02	0	0.05
	Hex.	0.08	0.29	0	0.06	0.02	0.02	0.02	0	0	0.46	0.02	0.02	0.02

A.3 North West - South East

The same data set is visualised in both displays, the difference is the land area of each SA3 is coloured in the choropleth and the hexagon representing the SA3 is coloured in the Hexagon Tile Map.

Lineups were created using the Australian Statistical Areas at Level 3.

A.4 All Cities

A.4.1 Replicate 1

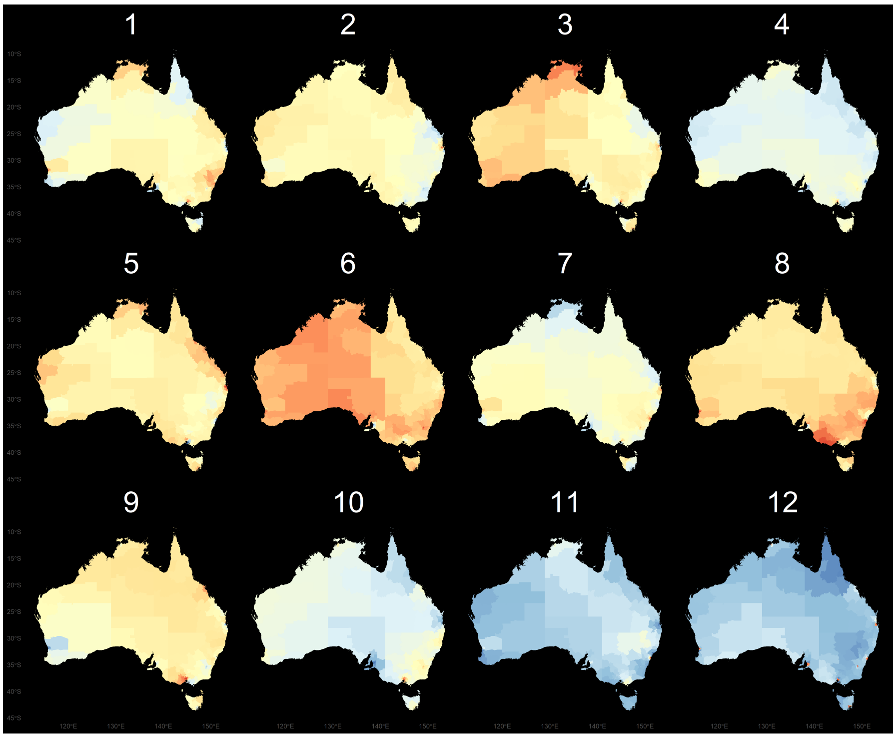


Figure A.5: *The lineup of choropleth map displays, location 12 contains a distribution that affects all capital cities.*

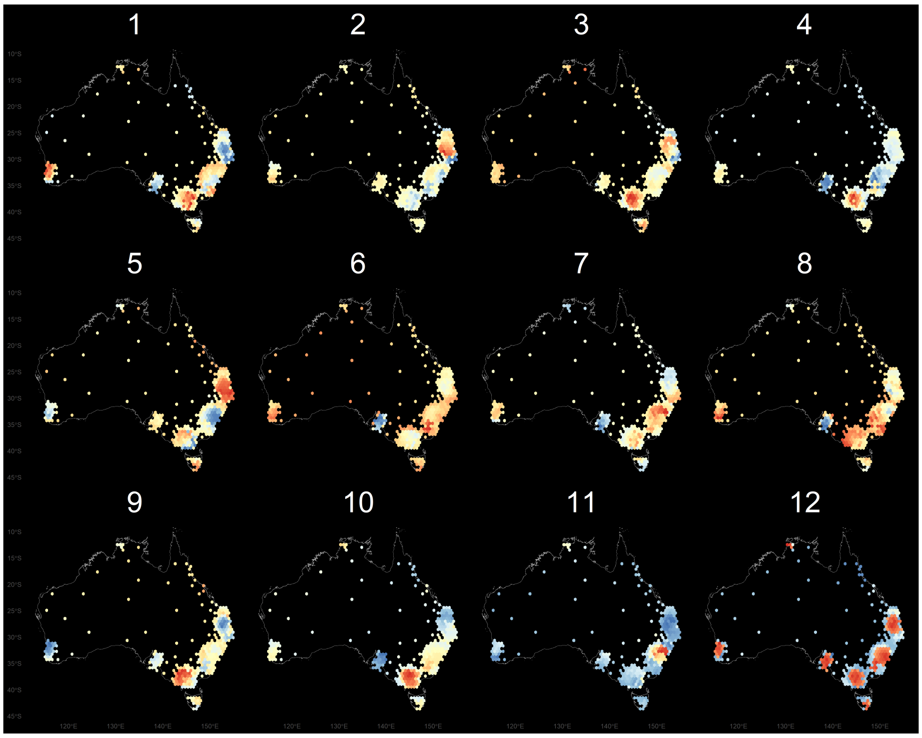


Figure A.6: *The lineup of hexagon tile map displays, location 12 contains a distribution that affects all capital cities.*

A.4.2 Replicate 2

Figure A.7: *The lineup of choropleth map displays, location 3 contains a distribution that affects all capital cities.*

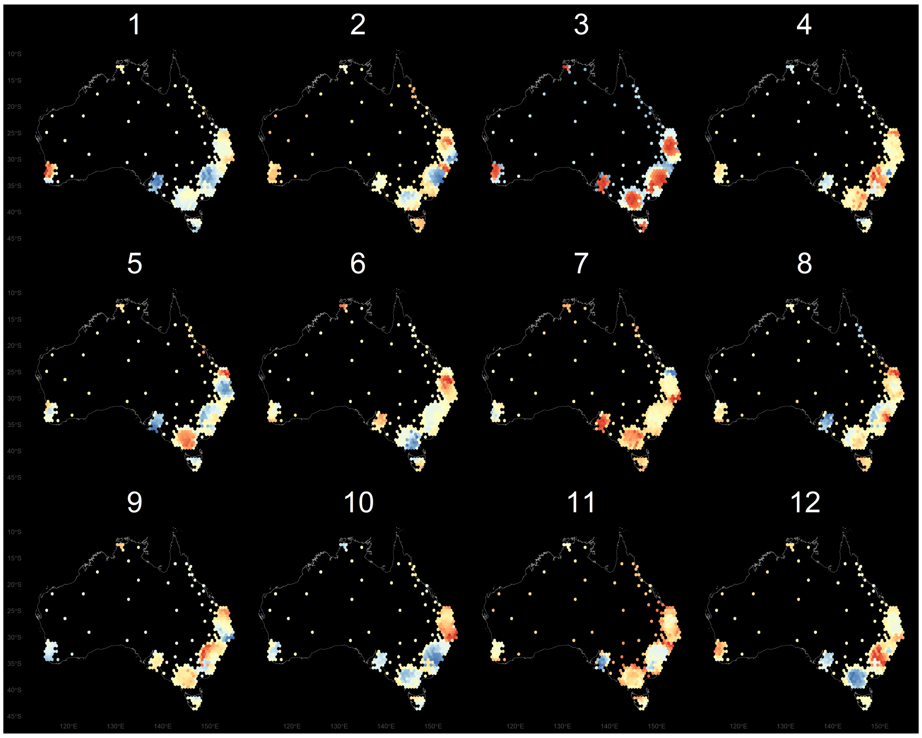


Figure A.8: *The lineup of hexagon tile map displays, location 3 contains a distribution that affects all capital cities.*

A.4.3 Replicate 3

Figure A.9: *The lineup of choropleth map displays, location 4 contains a distribution that affects all capital cities.*

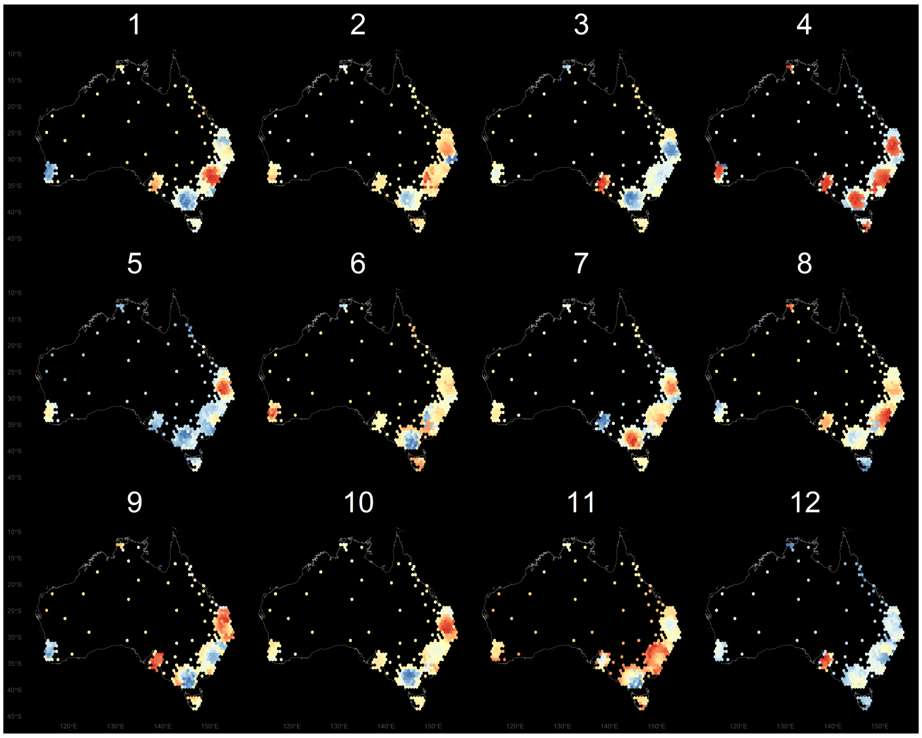


Figure A.10: *The lineup of hexagon tile map displays, location 4 contains a distribution that affects all capital cities.*

A.4.4 Replicate 4

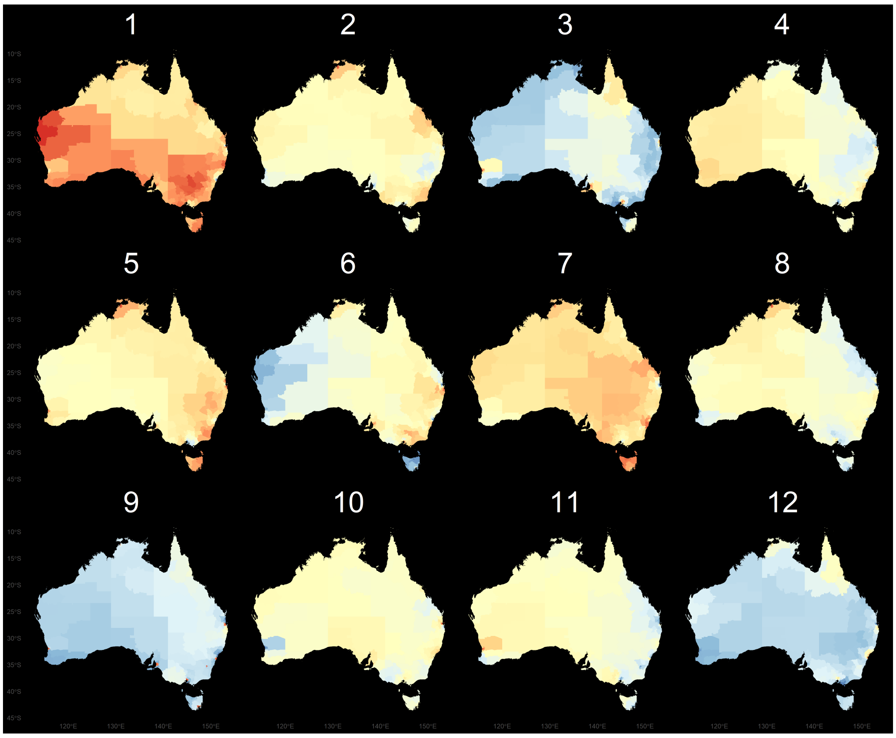


Figure A.11: *The lineup of choropleth map displays, location 9 contains a distribution that affects all capital cities.*

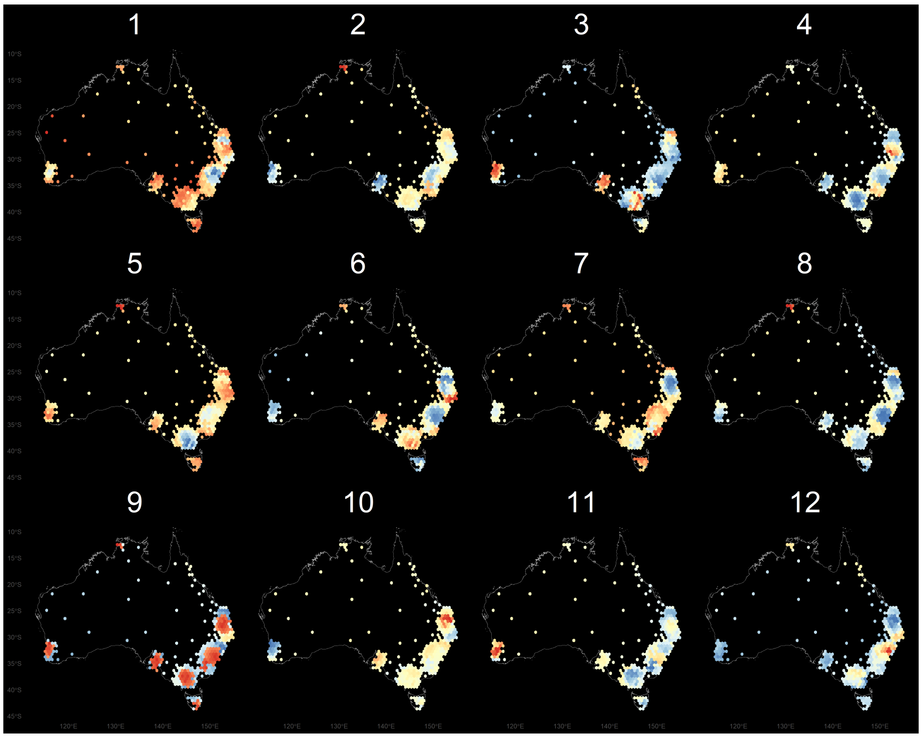


Figure A.12: *The lineup of hexagon tile map displays, location 9 contains a distribution that affects all capital cities.*

A.5 Three Cities

A.5.1 Replicate 1

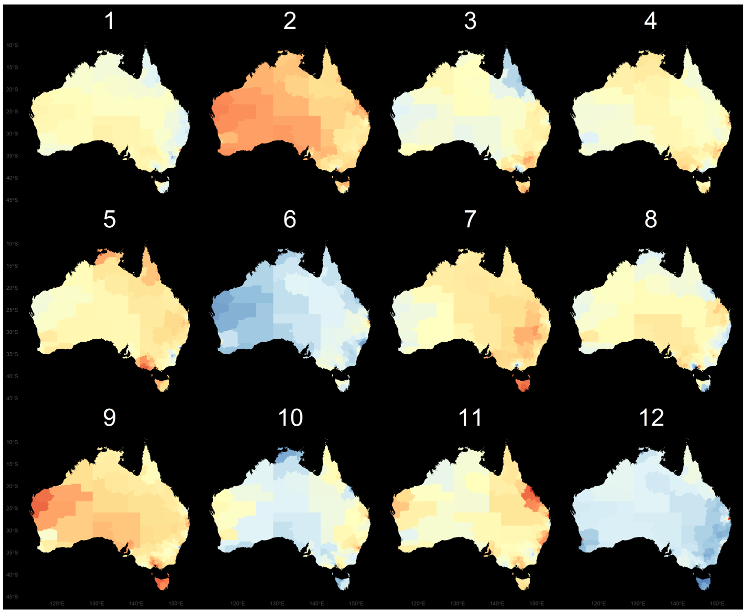


Figure A.13: *The lineup of choropleth map displays, location 12 contains a distribution that affects three of the Australian capital cities.*

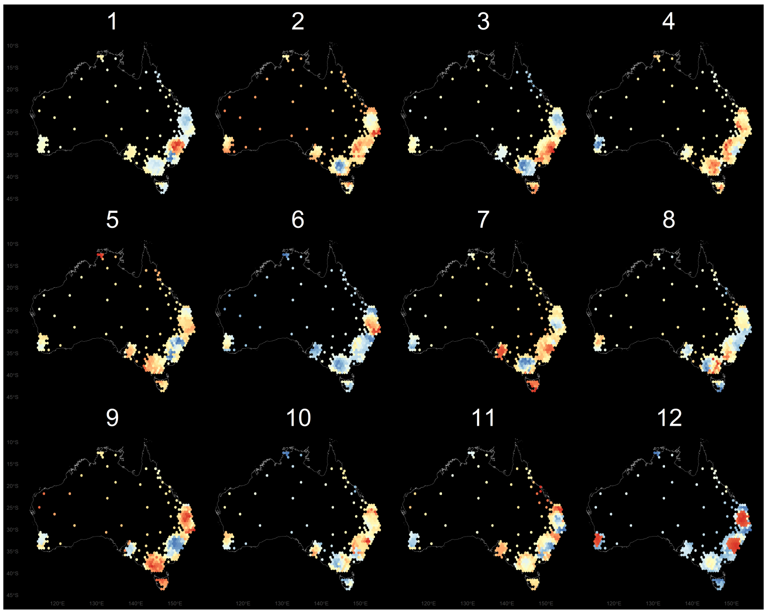


Figure A.14: *The lineup of hexagon tile map displays, location 12 contains a distribution that affects three of the Australian capital cities.*

A.5.2 Replicate 2

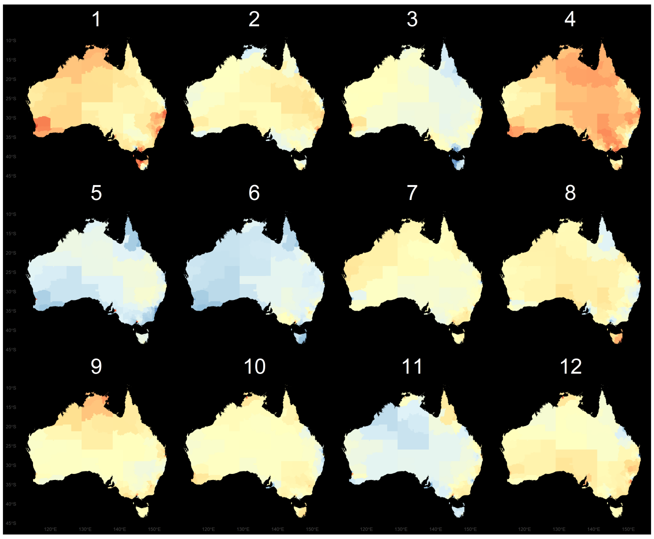


Figure A.15: *The lineup of choropleth map displays, location 3 contains a distribution that affects three of the Australian capital cities.*

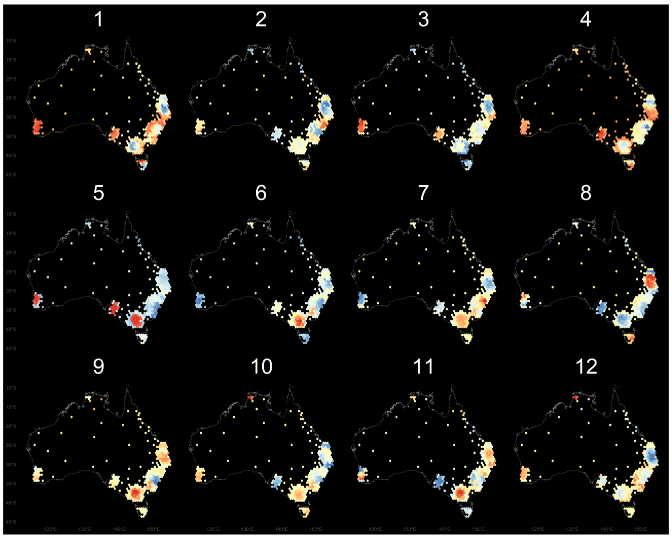


Figure A.16: *The lineup of hexagon tile map displays, location 3 contains a distribution that affects three of the Australian capital cities.*

A.5.3 Replicate 3

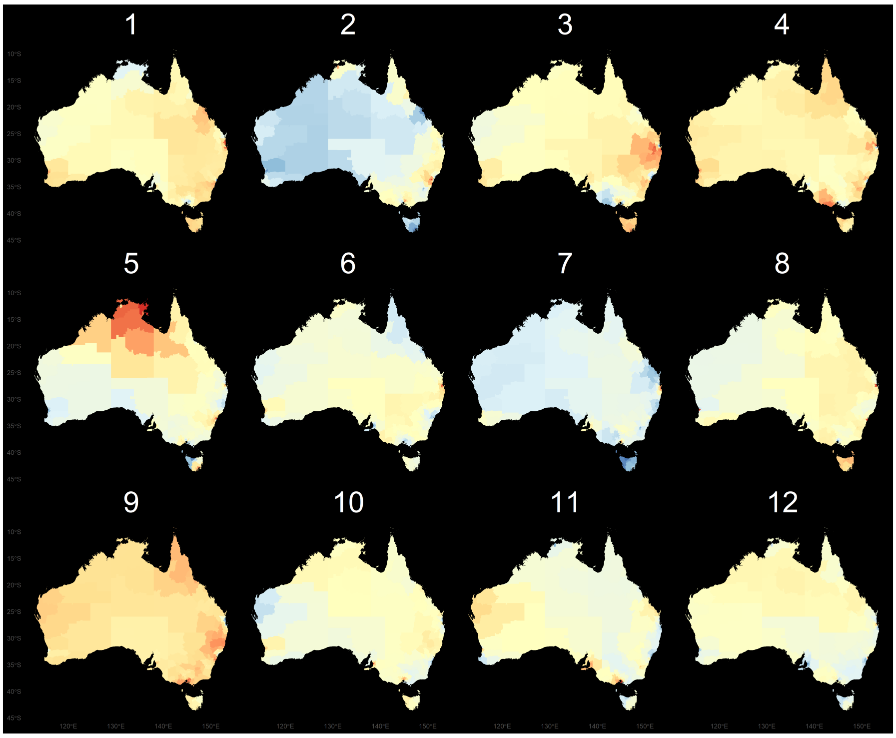


Figure A.17: *The lineup of choropleth map displays, location 4 contains a distribution that affects three of the Australian capital cities.*

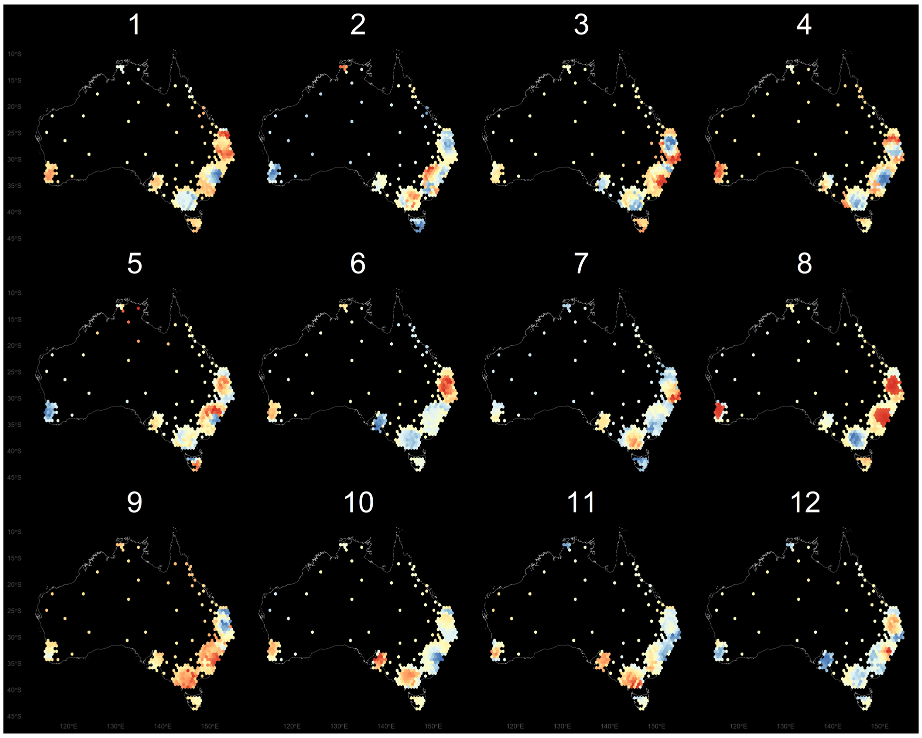


Figure A.18: *The lineup of hexagon tile map displays, location 4 contains a distribution that affects three of the Australian capital cities.*

A.5.4 Replicate 4

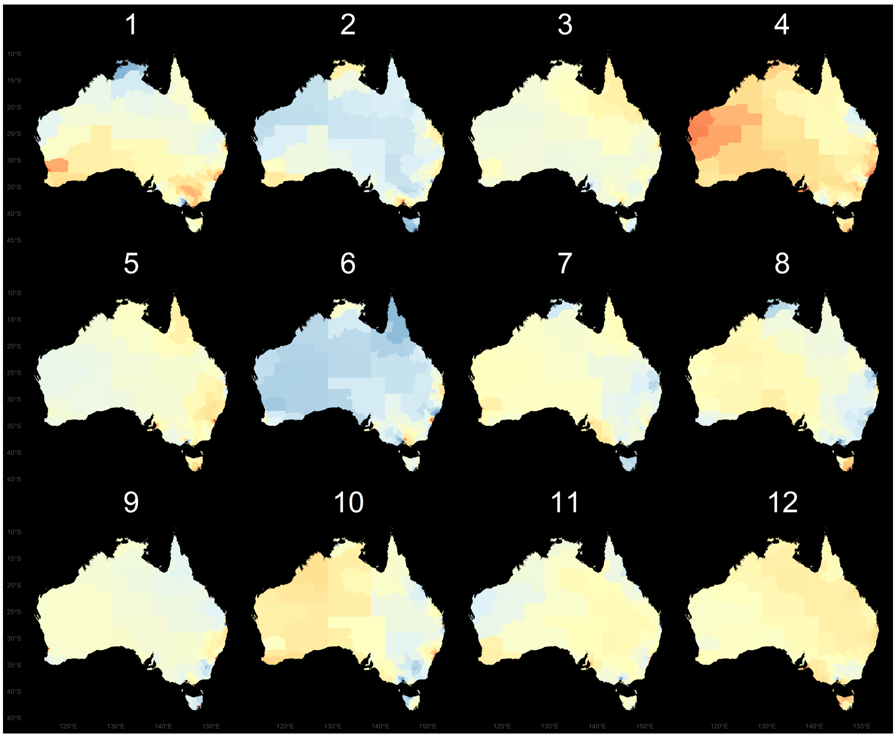


Figure A.19: *The lineup of choropleth map displays, location 9 contains a distribution that affects three of the Australian capital cities.*

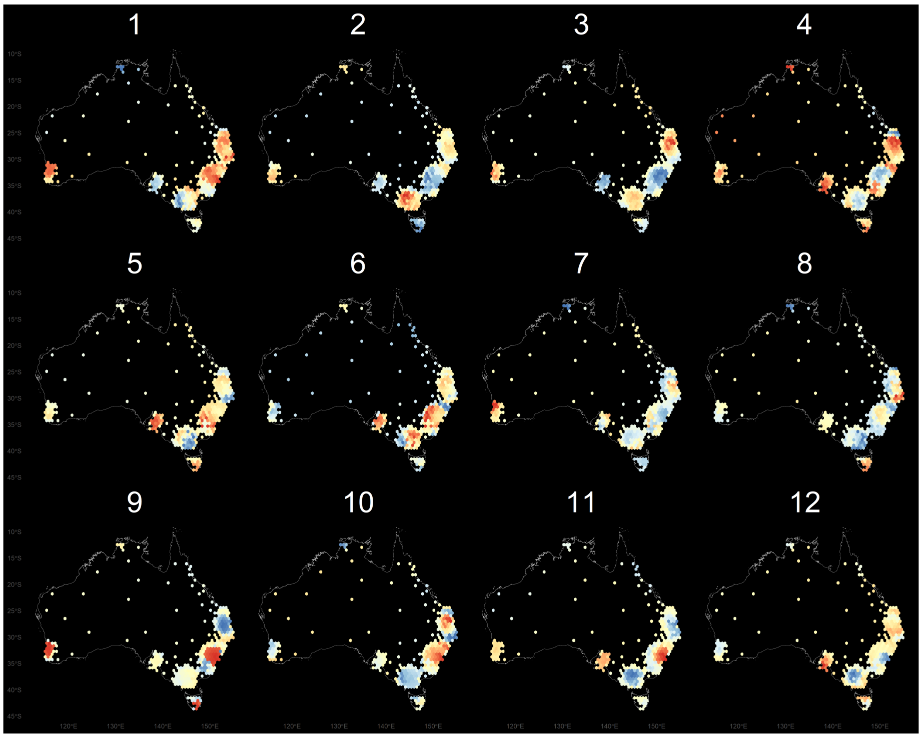


Figure A.20: *The lineup of hexagon tile map displays, location 9 contains a distribution that affects three of the Australian capital cities.*

A.6 North West to South East Geographic Trend

A.6.1 Replicate 1

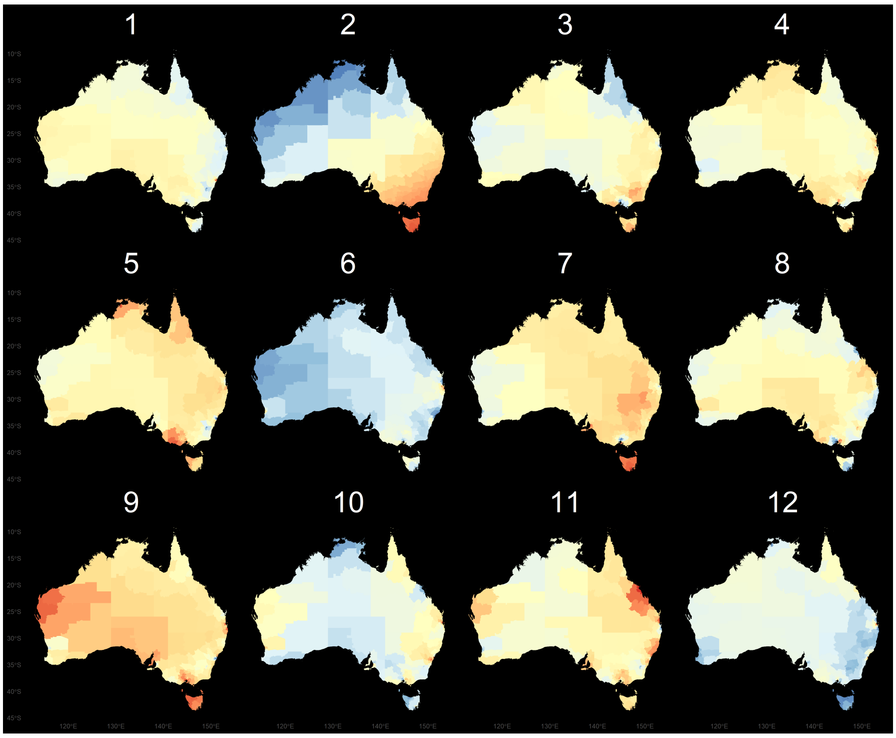


Figure A.21: *The lineup of choropleth map displays, location 12 contains a distribution that affects all capital nwse.*

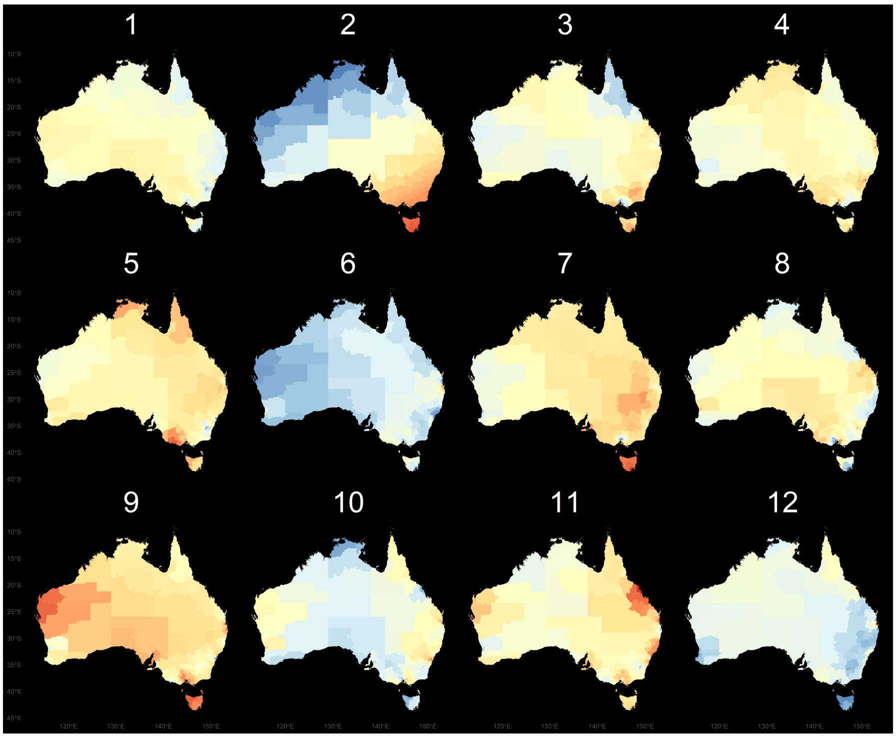


Figure A.22: *The lineup of hexagon tile map displays, location 12 contains a distribution that affects all capital nwse.*

A.6.2 Replicate 2

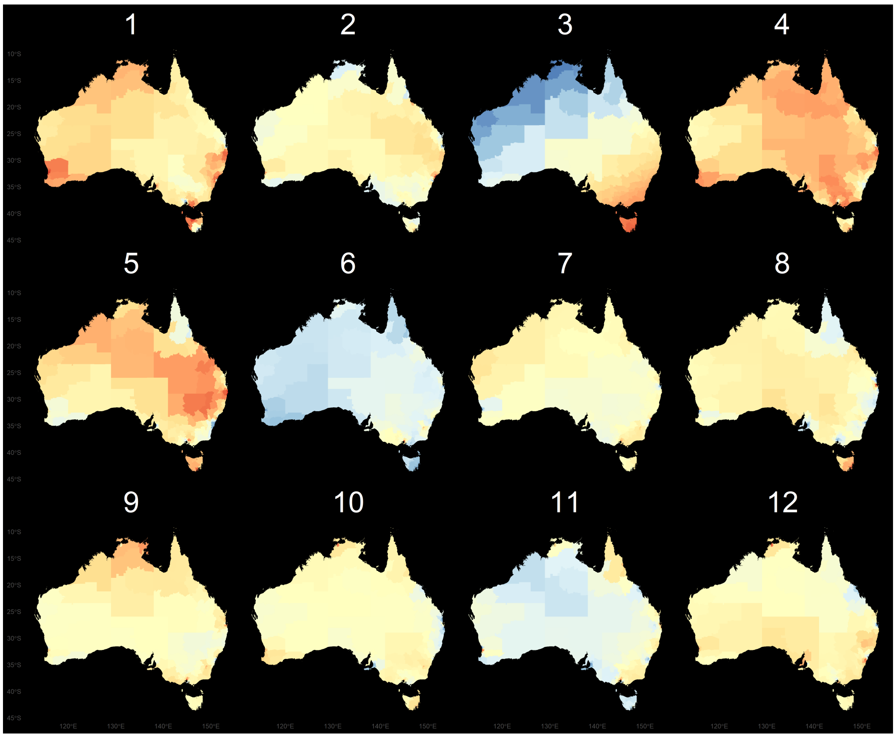


Figure A.23: *The lineup of choropleth map displays, location 3 contains a distribution that affects all capital nwse.*

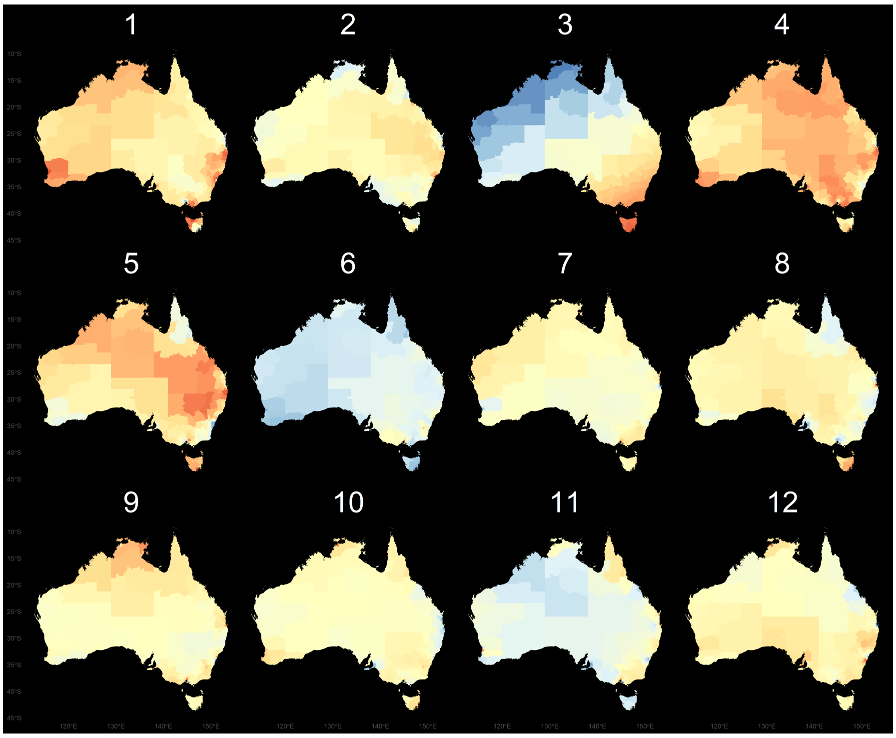


Figure A.24: *The lineup of hexagon tile map displays, location 3 contains a distribution that affects all capital nwse.*

A.6.3 Replicate 3

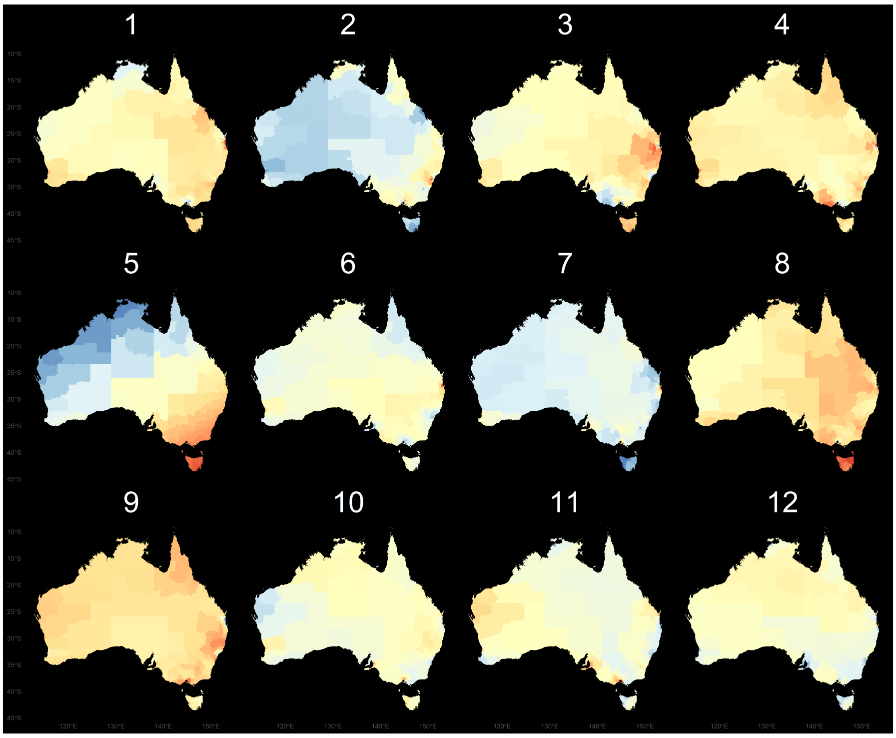


Figure A.25: *The lineup of choropleth map displays, location 4 contains a distribution that affects all capital nwse.*

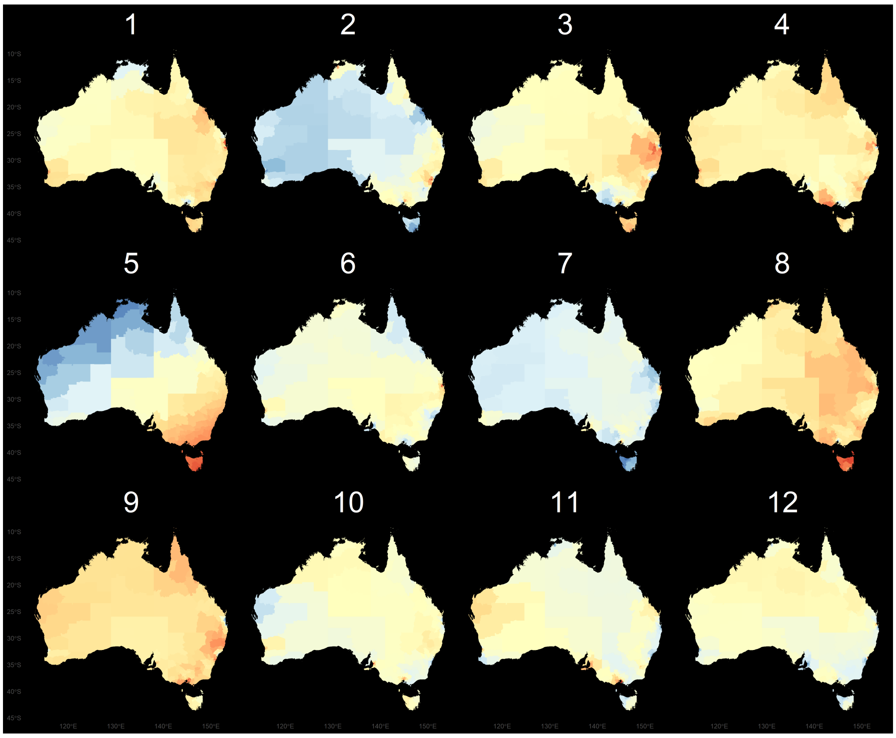


Figure A.26: *The lineup of hexagon tile map displays, location 4 contains a distribution that affects all capital nwse.*

A.6.4 Replicate 4

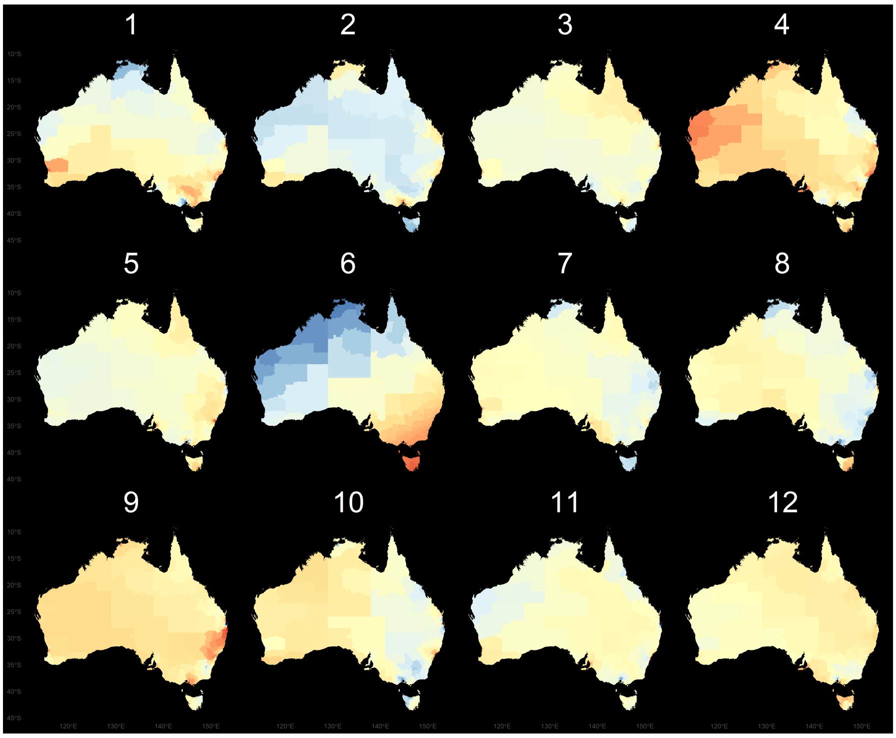


Figure A.27: *The lineup of choropleth map displays, location 9 contains a distribution that affects all capital nwse.*

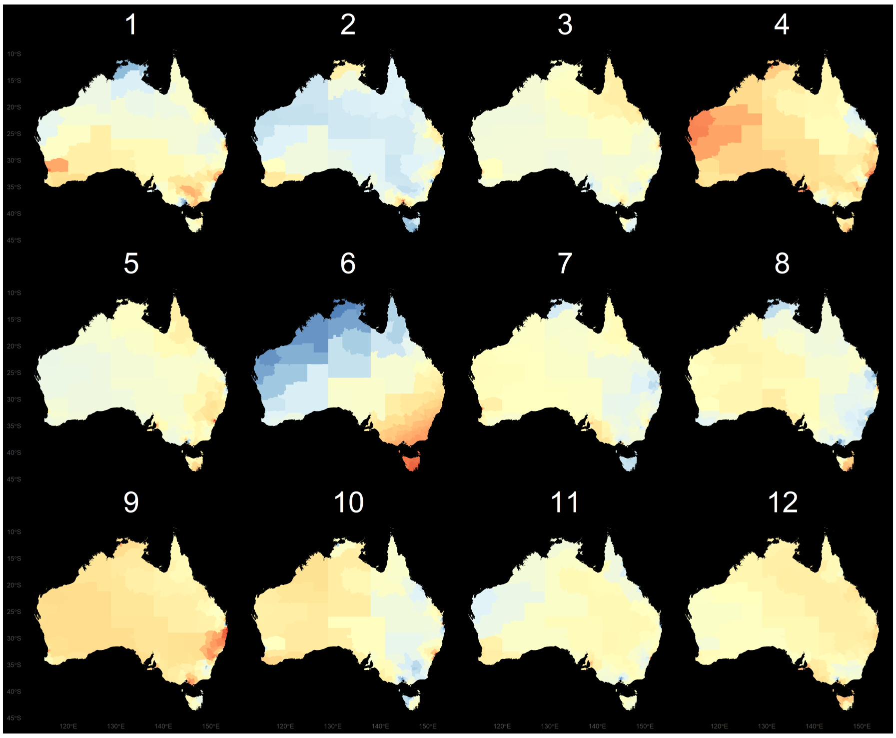


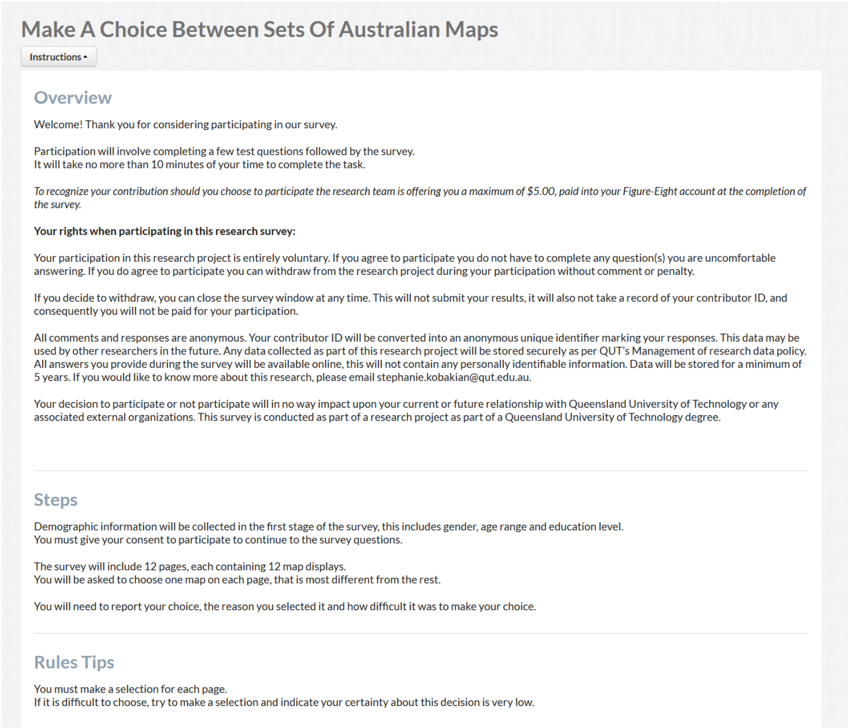
Figure A.28: *The lineup of hexagon tile map displays, location 9 contains a distribution that affects all capital nwse.*

Appendix B

Visual inference study survey procedure

Participants were recruited via advertising on the Figure-Eight crowdsource platform. Choosing the task from the list directed all potential participants to the page of instructions. This page contained written instructions and is shown in Fig. B.1.

```
ggdraw() +
  draw_plot(grid::rasterGrob(readPNG("figures/fig8_instruct.png")))
```



Make A Choice Between Sets Of Australian Maps

Instructions

Overview

Welcome! Thank you for considering participating in our survey.

Participation will involve completing a few test questions followed by the survey.
It will take no more than 10 minutes of your time to complete the task.

To recognize your contribution should you choose to participate the research team is offering you a maximum of $5.00, paid into your Figure-Eight account at the completion of the survey.

Your rights when participating in this research survey:

Your participation in this research project is entirely voluntary. If you agree to participate you do not have to complete any question(s) you are uncomfortable answering. If you do agree to participate you can withdraw from the research project during your participation without comment or penalty.

If you decide to withdraw, you can close the survey window at any time. This will not submit your results, it will also not take a record of your contributor ID, and consequently you will not be paid for your participation.

All comments and responses are anonymous. Your contributor ID will be converted into an anonymous unique identifier marking your responses. This data may be used by other researchers in the future. Any data collected as part of this research project will be stored securely as per QUT's Management of research data policy. All answers you provide during the survey will be available online, this will not contain any personally identifiable information. Data will be stored for a minimum of 5 years. If you would like to know more about this research, please email stephanie.kobakian@qut.edu.au.

Your decision to participate or not participate will in no way impact upon your current or future relationship with Queensland University of Technology or any associated external organizations. This survey is conducted as part of a research project as part of a Queensland University of Technology degree.

Steps

Demographic information will be collected in the first stage of the survey, this includes gender, age range and education level.
You must give your consent to participate to continue to the survey questions.

The survey will include 12 pages, each containing 12 map displays.
You will be asked to choose one map on each page, that is most different from the rest.

You will need to report your choice, the reason you selected it and how difficult it was to make your choice.

Rules Tips

You must make a selection for each page.
If it is difficult to choose, try to make a selection and indicate your certainty about this decision is very low.

Figure B.1: *The training lineups of choropleth maps.*

B.1 Training

The participants were trained using three displays. There were relatively simple lineups, they are displayed in Fig. B.2 and Fig. B.3.

```r
ggdraw() +
  draw_plot(grid::rasterGrob(readPNG("figures/fig8_training1.png")))
```

```r
ggdraw() +
  draw_plot(grid::rasterGrob(readPNG("figures/fig8_training2.png")))
```

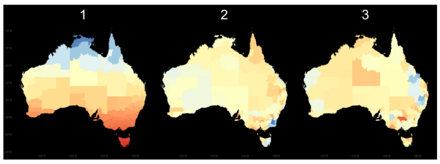
Simple Examples

The sets of images below show three maps of Australia, either a choropleth map or a hexagon tile map. You will see similar maps in the survey. In a choropleth map, geographic regions are colored according to a numerical value. In a hexagon tile map, each hexagon is colored according to a numerical value.

Please look at these images and read the explanations prior to taking the survey.

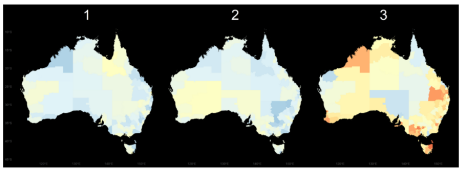
Set 1:

In this set of three maps of Australia: Which map is most different?



Answer: **MAP 1.**
This map has many blue areas at the top, and many red areas at the bottom. Tasmania, the island on the bottom right has much more red than the rest of Australia. This is very different to the other maps, which do not show the same color trend from North to South.

Set 2:
In this set of three maps of Australia: Which map is most different?

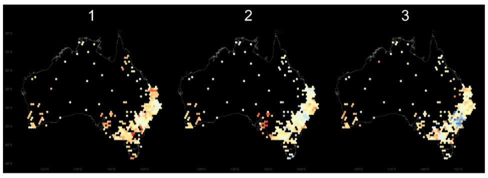


Answer: **MAP 3.**
This map has more red and orange areas around the coast.

Figure B.2: *The training lineups of choropleth maps.*

Set 3:

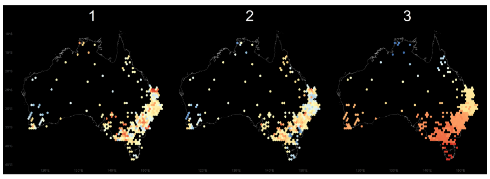
In this set of three maps of Australia: Which map is most different?



Answer: **MAP 2.**
This map has a group of red areas near each other. The other two maps are have scattered colors.

Set 4:

In this set of three maps of Australia: Which map is most different?



Answer: **MAP 3.**

This map has some blue areas at the top, and many red areas at the bottom.

Figure B.3: *The training lineups of hexagon tile maps.*

B.2 Survey application

The survey application was a `shinydashboard` we application, hosted on a website external to the Figure-Eight platform. The link to the survey was located at the bottom of the instructions and training page. Only participants who had read all of the instructions and seen the example image sets continued to the survey via the link. This page also contained a question that asked participants for a validation code. The participants unique validation code was generated upon them opening the web application. This code was released to participants when they had considered all twelve lineups and submitted their responses to the `googlesheets` data set. Their validation codes were contained in the data set and associated with each of their responses.

The demographic and consent page of the `shinydashboard` web application are displayed in Fig. B.4. Two example lineups are shown, one choropleth map lineup in Fig. B.5 and one heaxgon tile map lineup in Fig. B.6.

```
ggdraw() +
  draw_plot(grid::rasterGrob(readPNG("figures/survey_demogs.png")))
```

```
ggdraw() +
  draw_plot(grid::rasterGrob(readPNG("figures/survey_choro.png")))
```

```
ggdraw() +
  draw_plot(grid::rasterGrob(readPNG("figures/survey_hex.png")))
```

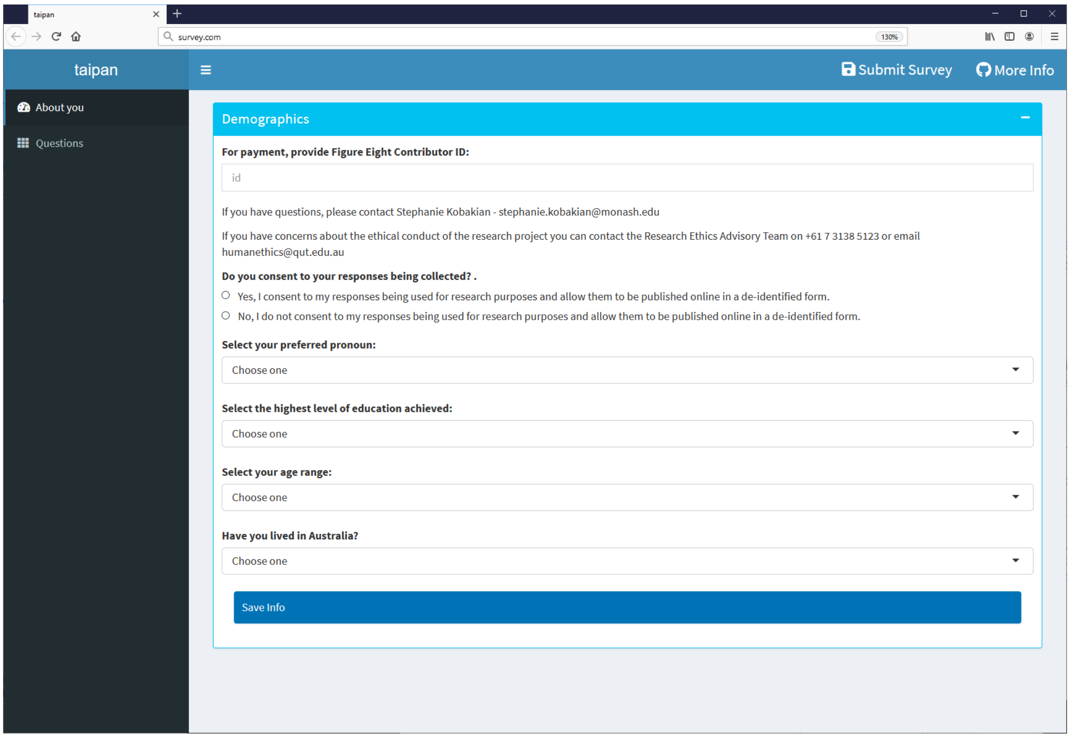


Figure B.4: *The demographics questions tab of the shinydashboard survey application.*

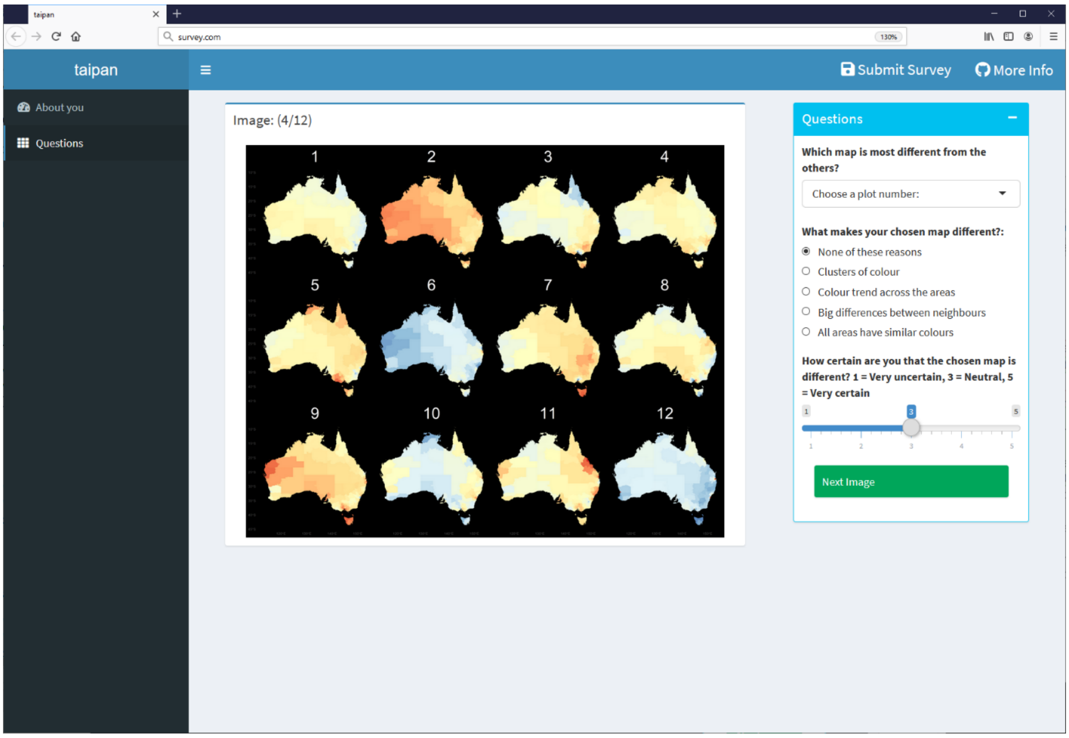


Figure B.5: *An example of the choropleth map lineup shown in the survey tab of the shinydashboard app.*

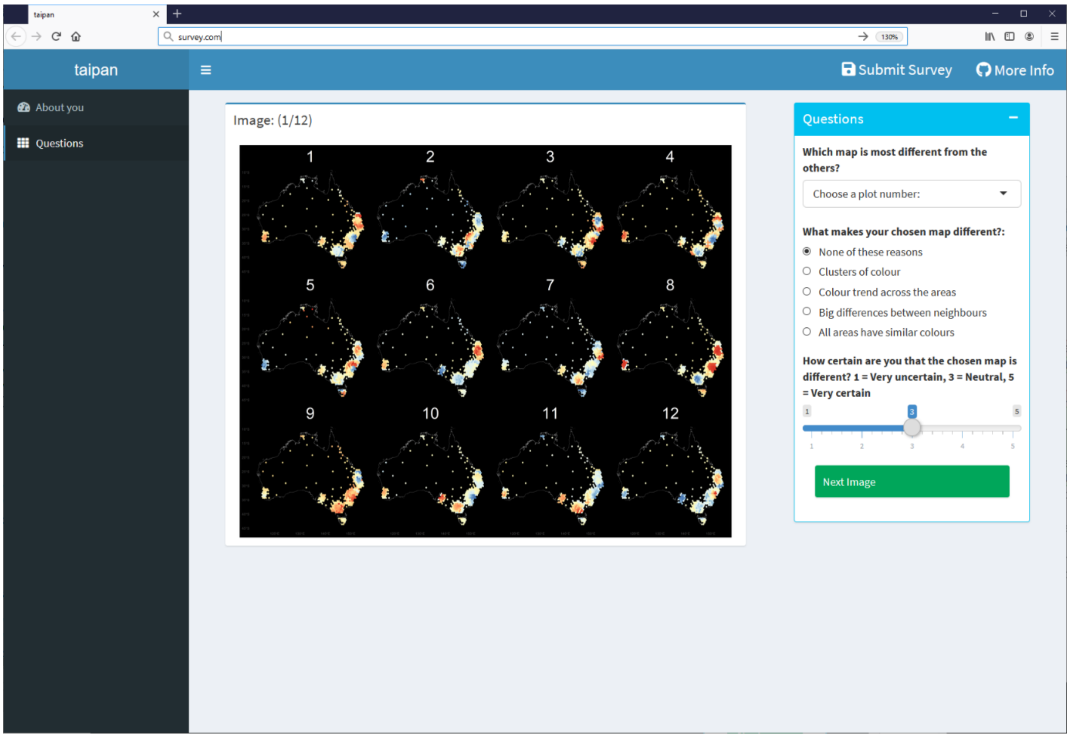


Figure B.6: *An example of the hexagon tile map lineup shown in the survey tab of the shinydashboard app.*

Appendix C

Ethics Approval

QUT	**PARTICIPANT INFORMATION FOR QUT RESEARCH PROJECT** **– Survey –**
	Assessing the effectiveness of different visualisation methods for Australian spatial data. **QUT Ethics Approval Number** 1900000991

Research team

Principal Researcher:	Stephanie Kobakian	Masters Student
Associate Researchers:	Kerrie Mengersen	Principal Supervisor
	Earl Duncan	Associate Supervisor
	Dianne Cook	External Supervisor
	Faculty of Science and Engineering	
	Queensland University of Technology (QUT)	

Why is the study being conducted?

The purpose of this research project is to test the effectiveness of two types of spatial displays: a choropleth map, and a hexagon map, where each geographic region is represented by a hexagon. This will examine the use of different map styles in communicating a relationship between geographic areas. The purpose of these displays is to convey the spatial distribution of the disease occurrence, or incidence. This can mean detecting hot spots corresponding to outbreaks, spatial trends, for example, indicating occurrence is related to latitude or even rural vs urban differences. Effectiveness of the display will be measured by accurate and efficient perception of these patterns.

This research project is being undertaken as part of a Masters study for Stephanie Kobakian, a student at Queensland University of Technology. You are invited to participate in this research project because you have had experience answering surveys and participating in crowdsource activities.

What does participation involve?

Participation will involve completing a few test questions followed by a survey.
Each survey item will contain a grid of maps, you will be asked the following question:

Which map is most different from the others?

Report your choice and the reason you selected it, and how difficult your decision was to make. It will take no more than 10 minutes of your time to complete the task.

Questions will include images similar to Figure 1 below:

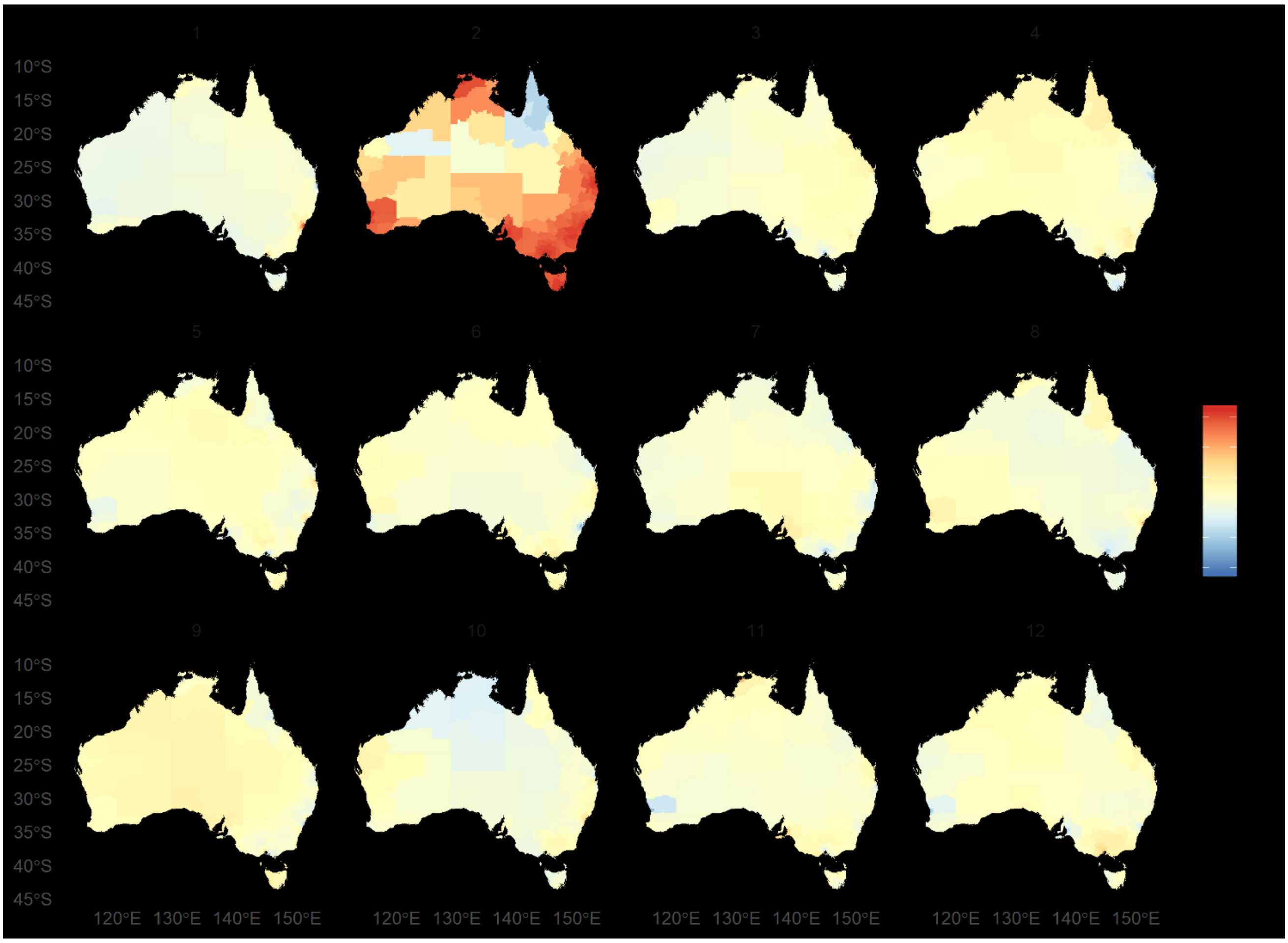


Figure 1. A lineup of geographic maps of Australia. Each sa3 has been coloured according to a simulated data set. Only one of these maps displays a spatial relationship, the rest are null plots, with colours shuffled between the areas.

Your participation in this research project is entirely voluntary. If you agree to participate you do not have to complete any question(s) you are uncomfortable answering. Your decision to participate or not participate will in no way impact upon your current or future relationship with QUT (for example your grades) or associated external organisation. If you do agree to participate you can withdraw from the research project during your participation without comment or penalty. However, as the survey does not request any personal identifying information, once it has been submitted it will not be possible to withdraw.

What are the possible benefits for me if I take part?

It is expected that this research project will directly benefit you as a paid member of the Figure-Eight platform. The outcomes of the research may also benefit researchers who have the options to consider the maps they use to communicate spatial information to the general public.

To recognise your contribution should you choose to participate the research team is offering you $5.00 paid into your Figure-Eight account at the completion of the survey.

What are the possible risks for me if I take part?

Risks:

Monetary risk: if participants do not accurately provide their Contributor ID in our external survey we will not be able to confirm they participated, and provide the payment to their account. As the figure Eight platform encourages paying participants after the survey collection period has finished.

Privacy risk: data on contributors, such as location, channel, time, and IP address will be provided to researchers by the platform. This information will be held on the researcher's personal laptops.

Psychological risks of taking part in this survey involves a possible negative affective state such as anxiety as participants will be asked to evaluate maps that are very unfamiliar. There is also the potential risk of anxiety resulting from the colouring of red areas, this colour scheme is best for all colour blindness types except greyscale.

What about privacy and confidentiality?

All comments and responses are anonymous. It will only be possible to identify due to your contributor ID provided in the research, personal identifying information is not sought in any of the responses.
It will not be possible to re-identify you using your contributor ID, but it will be removed and not stored in a public space after is received by the researchers. This data may be used by other researchers in the future.

Any data collected as part of this research project will be stored securely as per QUT's Management of research data policy. All answers you provide during the survey will be available online, this will not contain any personally identifiable information. Data will be stored for a minimum of 5 years. It will be available publicly at the web address: https://github.com/srkobakian/experiment.
The research project is funded by ACEMS and they will have access to the data obtained during the project as it will be publicly available.

How do I give my consent to participate?

The survey will ask if each participant gives their consent for their responses to be used.
The selection of the "yes" checkbox will allow continuation to the survey questions.
Submission of the completed survey is accepted as an indication of your consent to participate in this research project, you may withdraw by completing less than 50% of the questions, after checking the "yes" checkbox.

What if I have questions about the research project?

If you have any questions or require further information please contact one of the listed researchers:

Stephanie Kobakian	stephanie.kobakian@hdr.qut.edu.au	+61 433699797
Kerrie Mengersen	k.mengersen@qut.edu.au	+61 731382063
Earl Duncan	earl.duncan@qut.edu.au	+61 410874218
Dianne Cook	dicook@monash.edu	+61 399052608

What if I have a concern or complaint regarding the conduct of the research project?

QUT is committed to research integrity and the ethical conduct of research projects. If you wish to discuss the study with someone not directly involved, particularly in relation to matters concerning policies, information or complaints about the conduct of the study or your rights as a participant, you may contact the QUT Research Ethics Advisory Team on +61 7 3138 5123 or email humanethics@qut.edu.au.

Thank you for helping with this research project.
Please print this sheet for your information.

Bibliography

Allaire, J, Y Xie, J McPherson, J Luraschi, K Ushey, A Atkins, H Wickham, J Cheng, W Chang, and R Iannone (2019a). *rmarkdown: Dynamic Documents for R*. R package version 1.15. https://github.com/rstudio/rmarkdown.

Allaire, J, 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, MA Desautels, D Leutnant, MDPI, Taylor and Francis, O Ögreden, D Hance, D Nüst, P Uvesten, E Campitelli, J Muschelli, ZN Kamvar, N Ross, and R Cannoodt (2019b). *rticles: Article Formats for R Markdown*. R package version 0.13. https://CRAN.R-project.org/package=rticles.

Arnold, JB (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes.

Bates, D, M Maechler, B Bolker, and S Walker (2019). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. R package version 1.1-21. https://CRAN.R-project.org/package=lme4.

Bell, BS, RE Hoskins, LW Pickle, and D Wartenberg (2006). Current Practices in Spatial Analysis of Cancer Data: Mapping Health Statistics to Inform Policymakers and the Public. *International Journal of Health Geographics* **5**, 49.

Bivand, R, J Nowosad, and R Lovelace (2019). *spData: Datasets for Spatial Analysis*. R package version 0.3.0. https://CRAN.R-project.org/package=spData.

Cook, D, A Ebert, J Forbes, H Hofmann, R Hyndman, T Lumley, B Marwick, C Sievert, M Sun, D Talagala, N Tierney, N Tomasetti, E Wang, and F Zhou (2019). *eechidna:*

Exploring Election and Census Highly Informative Data Nationally for Australia. R package version 1.4.0. https://CRAN.R-project.org/package=eechidna.

d'Onofrio, A, C Mazzetta, C Robertson, M Smans, P Boyle, and M Boniol (2016). Maps and Atlases of Cancer Mortality: A Review of a Useful Tool to Trigger New Questions. *Ecancermedicalscience* **10**, 670–670.

Dorling, D (2011). "Area Cartograms: Their Use and Creation". In: *Concepts and Techniques in Modern Geography (CATMOG)*. Vol. 59, pp. 252–260.

Dorling, D (2012). *The Visualisation of Spatial Social Structure*. John Wiley and Sons Ltd.

Dougenik, JA, NR Chrisman, and DR Niemeyer (1985). An Algorithm to Construct Continuous Area Cartograms. *The Professional Geographer* **37**(1), 75–81. eprint: https://doi.org/10.1111/j.0033-0124.1985.00075.x.

Griffin, T (1980). Cartographic Transformation of the Thematic Map Base. *Cartography* **11**(3), 163–174. eprint: https://doi.org/10.1080/00690805.1980.10438102.

Hafen, R (2019). *geofacet: 'ggplot2' Faceting Utilities for Geographical Data*. R package version 0.1.10. https://CRAN.R-project.org/package=geofacet.

Hofmann, H, L Follett, M Majumder, and D Cook (2012). Graphical Tests for Power Comparison of Competing Designs. *IEEE Transactions on Visualization and Computer Graphics* **18**(12), 2441–2448.

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

Kobakian, S and D Cook (2019). *sugarbag: Create Tessellated Hexagon Maps*. R package version 0.1.0. https://CRAN.R-project.org/package=sugarbag.

Kobakian, S and M O'Hara-Wild (2018). *taipan: Tool for Annotating Images in Preparation for Analysis*. R package version 0.1.2. https://CRAN.R-project.org/package=taipan.

Lucchesi, L and W C.K. (2017). Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation and Glyph Rotation. *Stat*.

Mackey, W. F. (2019). *absmapsdata: A catalogue of ready-to-use ASGS mapping data*. R package version 0.1.1.

Min Ouyang and P Revesz (2000). Algorithms for Cartogram Animation. In: *Proceedings 2000 International Database Engineering and Applications Symposium (Cat. No.PR00789)*, pp.231–235.

Neuwirth, E (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer.

Olson, JM (1976). Noncontiguous Area Cartograms. *The Professional Geographer* **28**(4), 371–380. eprint: https://doi.org/10.1111/j.0033-0124.1976.00371.x.

Pebesma, E (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**(1), 439–446.

Pebesma, E and B Graeler (2019). *gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation*. R package version 2.0-3. https://CRAN.R-project.org/package=gstat.

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

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

Raisz, E (1963). Rectangular Statistical Cartograms of the World. *Journal of Geography* **35** (1), 8–10.

Roy Chowdhury, N (2014). “"Explorations of the lineup protocol for visual inference: application to high dimension, low sample size problems and metrics to assess the quality"”. https://lib.dr.iastate.edu/etd/13988.

Tufte, ER (1990). *Envisioning Information*. Graphics Press.

Urbanek, S (2013). *png: Read and write PNG images*. R package version 0.1-7. https://CRAN.R-project.org/package=png.

Wickham, H, D Cook, H Hofmann, and A Buja (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* **16**(6), 973–979.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H (2017). *tidyverse: R packages for data science*. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Wickham, H, NR Chowdhury, D Cook, and H Hofmann (2018). *nullabor: Tools for Graphical Inference*. R package version 0.3.5. https://CRAN.R-project.org/package=nullabor.

Wilke, CO (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot.

Xie, Y (2014). "knitr: A Comprehensive Tool for Reproducible Research in R". In: *Implementing Reproducible Computational Research*. Ed. by V Stodden, F Leisch, and RD Peng. ISBN 978-1466561595. Chapman and Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, H (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra.