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

A thesis submitted 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

2019

Contents

Copyright notice

Abstract

The Australian population has congregated in the capital cities and significant cities in each state. This pattern has resulted in very dense population centers and sparsely populated rural areas. The relationship between the Australian population and the geographic area they live on results in a heterogeneous distribution of the map space. The goal of many spatial visualisations is to gain a broad perspective of the values of statistics over the Australian population. However, the use of most mapping techniques can mislead, as the use of geographical areas unequally presents the spatial distribution of a dataset.

The algorithm presented in this thesis 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, 2019).

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

Keywords: maps; statistics; geospatial statistics; visual inference;

Declaration

(Thesis including published works 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 DProf. Kerrie Mengersen and Dr. Earl Duncan. It was also created under the supervision of the external supervisor Prof. Dianne Cook.

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

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 following is the suggested format for the required declaration provided at the start of any thesis chapter which includes a co-authored publication.

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:

Please state the publication title and date of publication or status:

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 _____

Acknowledgements

We would like to acknowledge Dr Earl Duncan (Research Associate at ARC Centre of Excellence for Mathematical & Statistical Frontiers, QUT). His time and effort given to edit and comment on this paper was invaluable.

We thank Dr Susanna Cramb (Spatial Modeller, Cancer Council Queensland) and Dr Peter Baade (Senior Research Fellow, Cancer Council Queensland) for providing the opportunity to cover this exploration of disease mapping methods in writing. This literature review would not be possible without the opportunity provided by Queensland University of Technology, and Cancer Council Queensland, and the roles of Professor Kerrie Mengersen (Professor of Statistics, Science and Engineering Faculty, QUT) and Dr Earl Duncan in contributing to the Australian Cancer Atlas and supervision of Stephanie's Research Masters.

It was in the development of this online cancer atlas that methods for disease map displays, and visual communication strategies were explored. We are thankful for the opportunity to write about these visualizations and the situations in which they are appropriate.

Several R (R Core Team, 2019) packages were used to produce this paper. For data analysis the tidyverse package (Wickham, 2017) provided many useful functions. ggplot2 package (Wickham, 2016) was used to create maps, with RColorBrewer package (Neuwirth, 2014) providing additional color palettes and ggthemes (Arnold, 2019) package providing themes. The png package (Urbanek, 2013) was used to access png images taken from online web atlases, cowplot package (Wilke, 2019) was used to arrange these plots into grouped images.

The following packages were used to create transformations from the sf (Pebesma, 2018) geographical shapes of the states of America (Bivand, Nowosad, and Lovelace, 2019) and the Australian Statistical Areas (Level 3): The cartogram (Jeworutzki, 2018) package creates contiguous, non-contiguous and Dorling cartograms. Hexagon tile map displays were created using the sugarbag package (Kobakian and Cook, 2019).

Preface

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

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

3. The details of the visual inference testing is documented in Chapter four has been submitted to the *IEEE Transactions of Visualisation and Computer Graphics* under the title "Which is Better: a Choropleth Map or Hexagon Tile Map? A comparison using visual inference."

4. The code for the algorithm documented in Chapter three has been submitted to *CRAN* as the package sugarbag.

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

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, another 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 Cancer Atlas of Australia, 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. An online, interactive web tool for exploring the impact of cancer on Australian communities. The prominent display used by the Australian Cancer Atlas is a visualisation of Incidence Rates or Excess Death Rates. The set of geographic units used is Australian Statisical Areas, at Level 2. There are almost 2,200 individual areas, ranging in size from

The choropleth map display used in the Australian Cancer Atlas 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 'accussed' 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 linup.

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. *Algorithm for creating hexagon tile maps of Australia:* The algorithm will take geospatial areas and create an alternative visualisation of the spatially 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, 2019) 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

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]. 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

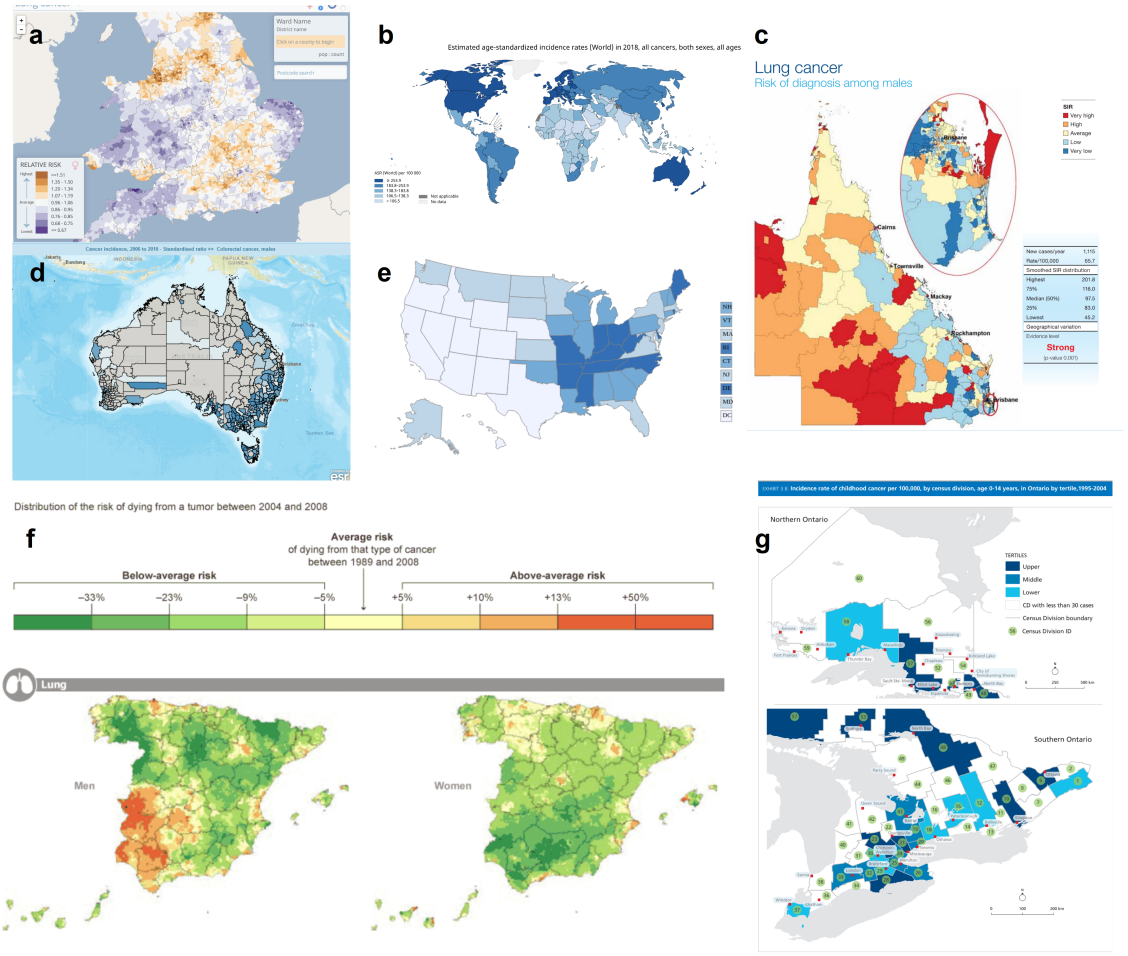


Figure 1: A selection of choropleth cancer maps from online atlases that are publicly available. Maps of various countries were chosen: United Kingdom, Australia, Spain, USA, Canada, and display several different colour palettes and legends. These atlases are described in Table 1.

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 modeling 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.

Map display	Details
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.
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 slghtly 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-

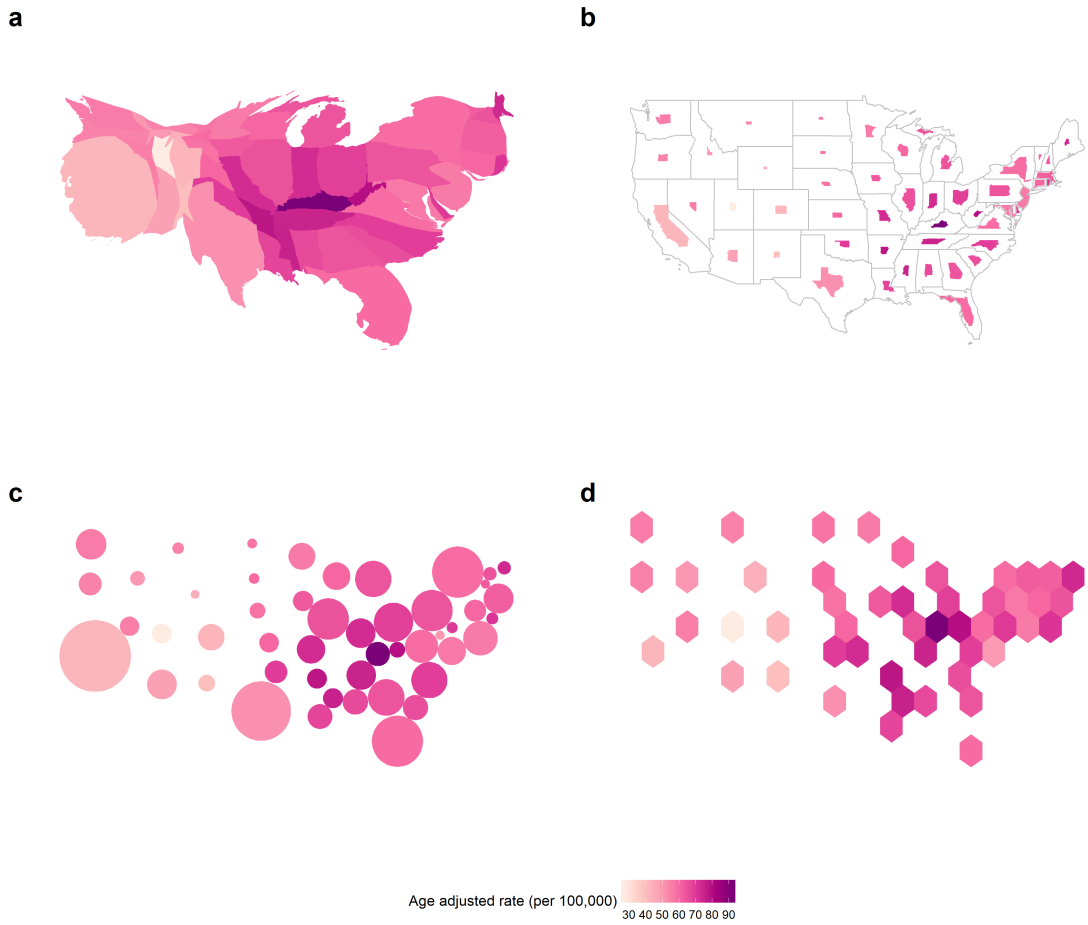


Figure 2: Common alternatives to maps, showing the same information for the United States of America: (a) contiguous cartogram, (b) non-contiguous, shape-preserved cartogram, (c) Dorling cartogram (non-contiguous), (d) hexagon tilemap (non-contiguous). Maps (a) - (c) are created by resizing and reshaping 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.

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, 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]. Tobler's Conformal mapping means to preserve angles locally so that the shapes of small areas on a traditional map and a cartogram would be similar. [33] presents 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.

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

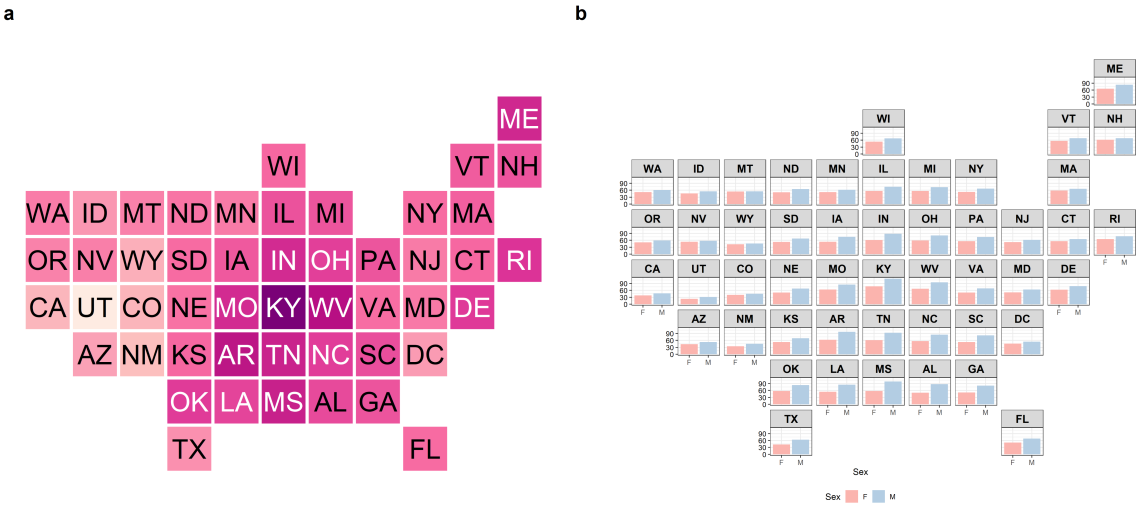


Figure 3: Two alternative displays, tile map (left) and geofaceted map (right), showing state age-adjusted rate of incidence for lung and bronchus in the USA. In the tile map, the layout approximates spatial location, with each state being an equal box filled with color representing cancer incidence. 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.

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 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.

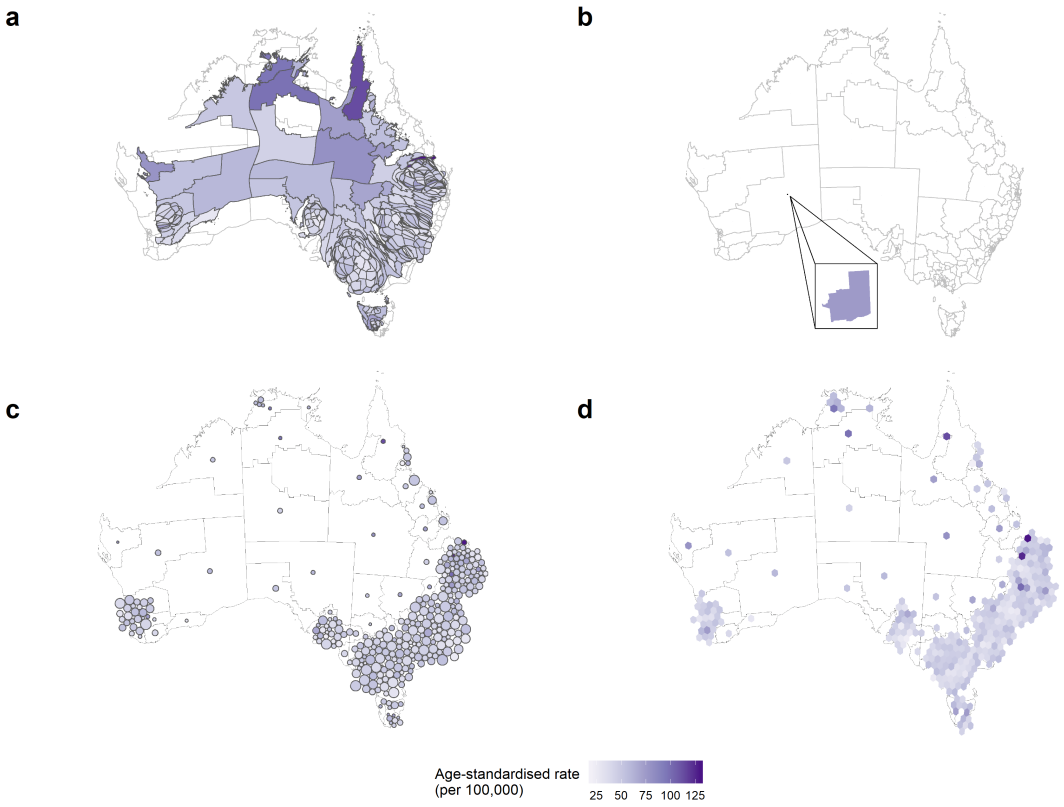


Figure 4: Cartograms showing melanoma incidence in Australia: (a) contiguous, partially population transformed, (b) non-contiguous shape preserved, (c) Dorling, (d) hexagon tile map. The contiguous cartogram has expanded the highly populated areas while preserving the full shapes of rural areas. If it 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.

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].

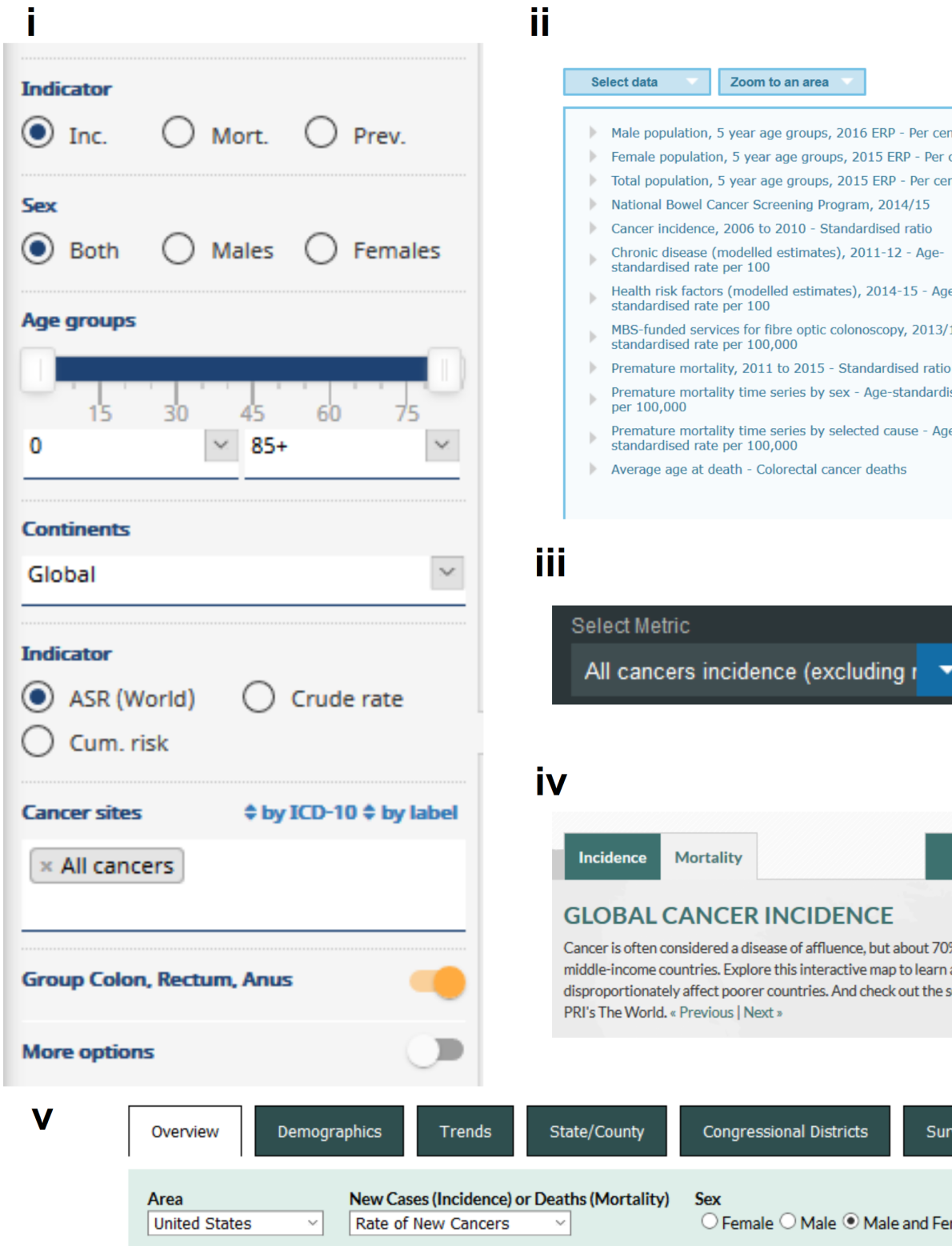
Animation, in contrast to interactivity, usually involves pre-computing views and showing these in a sequence. Lin Pedersen [60] provides an overview of animation for maps using the R package `gganimate` [61]. Animations are used to communicate a message by capturing and directing users' attention. It 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 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.



\begin{figure}

\caption{ 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: Cancer Today, (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.} \end{figure}

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 tradeoffs 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

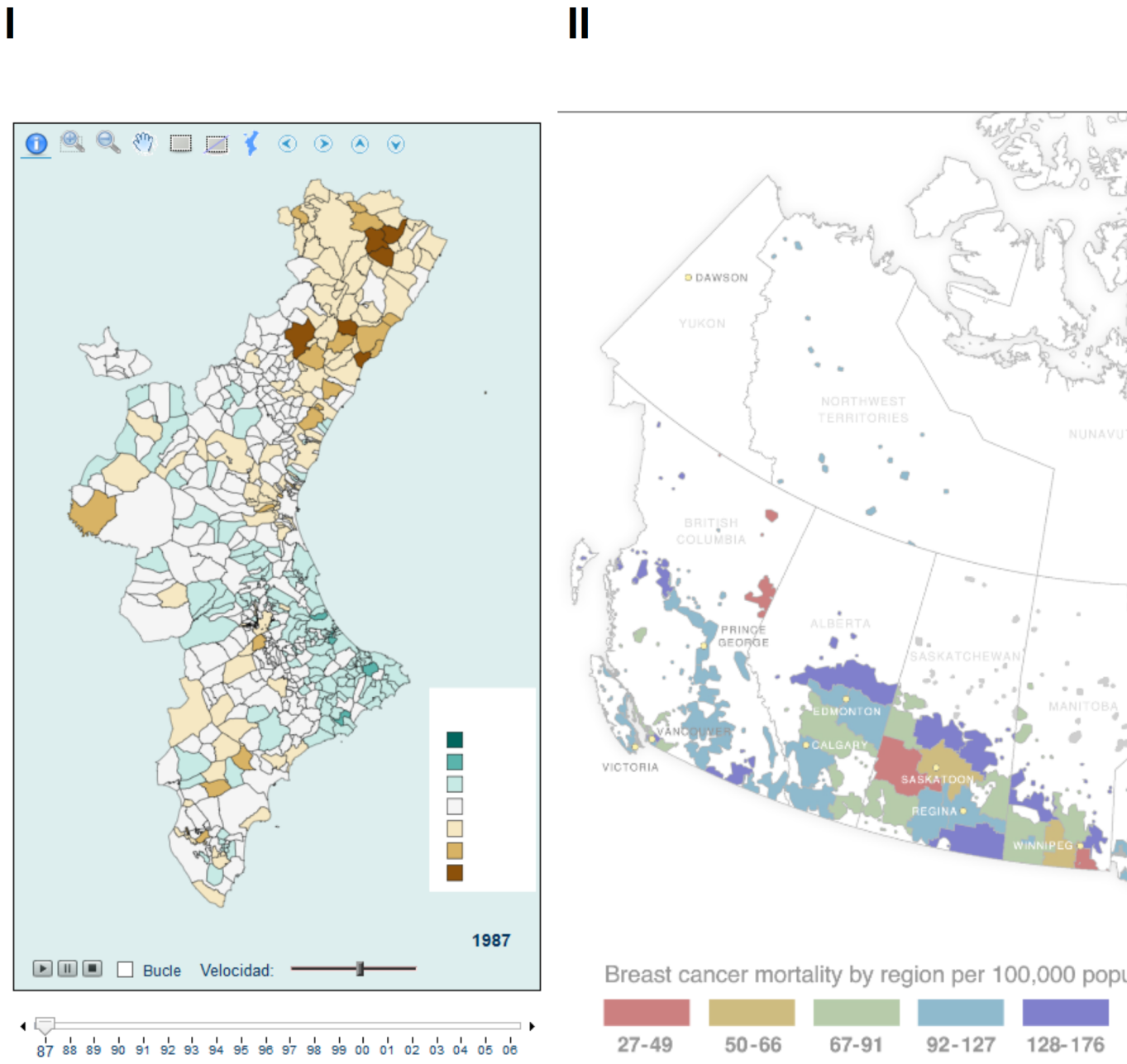


Figure 5: Two examples of advanced interactivity (and animation) in publicly available choropleth cancer maps: a. Linked maps and time-series line plots, with temporal animation in [Map of Cancer Mortality Rates in Spain](http://www.geeitema.org/AtlasET/atlas.jsp?causa=e27MEPOC), b. A highly responsive magnifying glass on a map of [Breast Cancer Mortality in Canada](http://www.ehatlas.ca/light-pollution/maps/breast-cancer-mortality).

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.

8 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 T (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 D, North S, 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 L, C.K. W (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

Algorithm

Journal of Statistical Software

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

Stephanie Kobakian
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 geogspatial 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, and the Australian Electoral Commission hold the responsibility for the division of geographic units. These boundaries are adjusted 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 amount of geographic area. 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 **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 must be provided to `create_hexmap`:

parameter	description
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. Tasmania

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:

parameter	description
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.

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

Step 1: Creating a tesselated grid

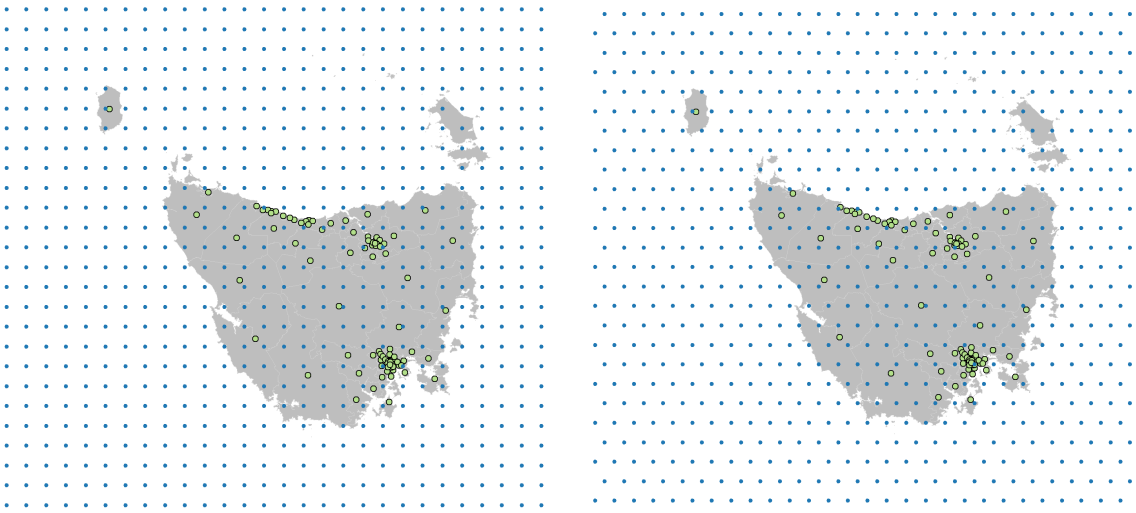


Figure 1: Grid points to create a tilegram.

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 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.

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

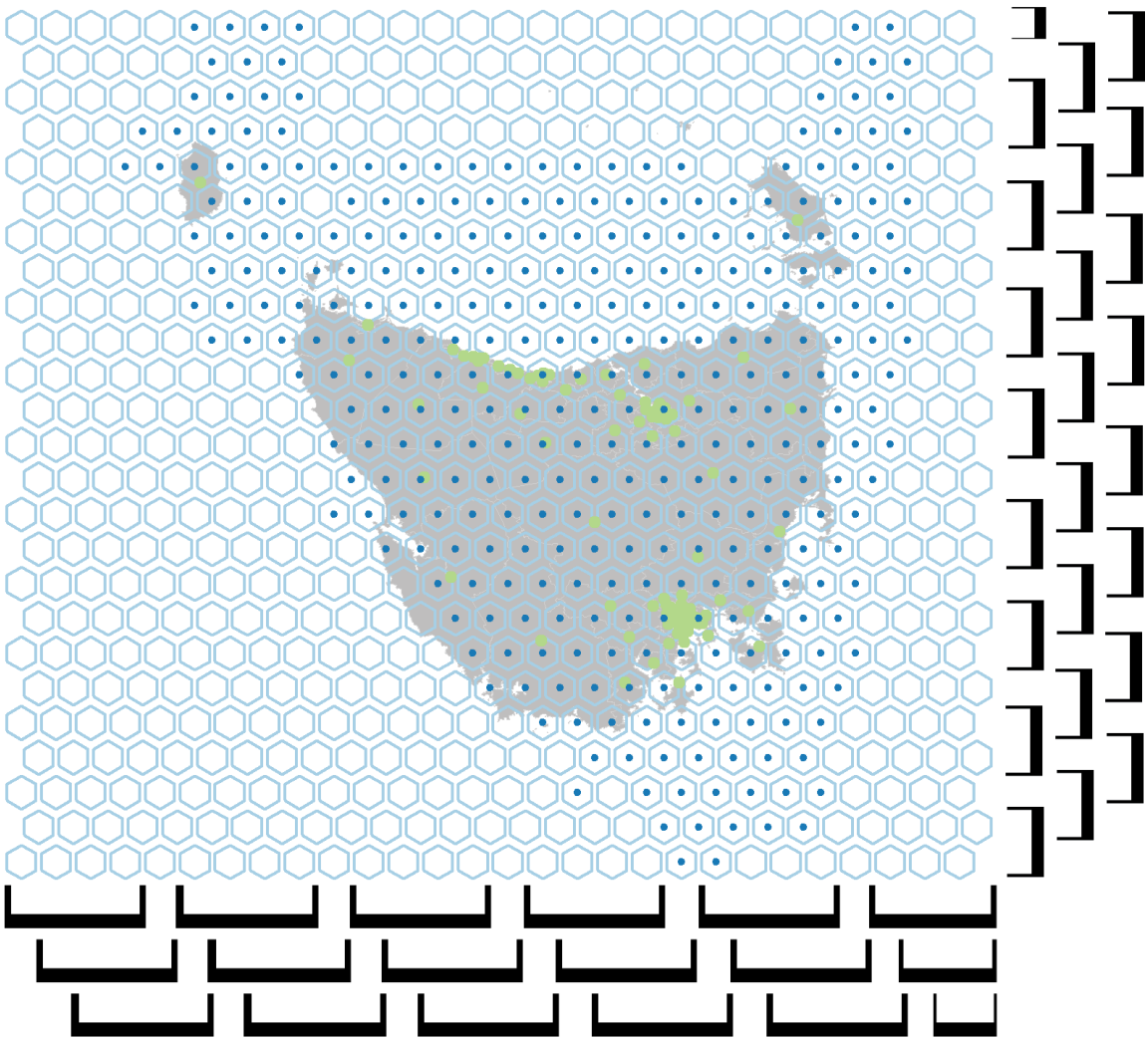


Figure 2: All possible hexagon locations from the initial grid are shown with blue outlines. The blue dots shown the grid points left as to choose from after the buffer step. The rolling windows to the right show the rows used to filter the hexagon locations.

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

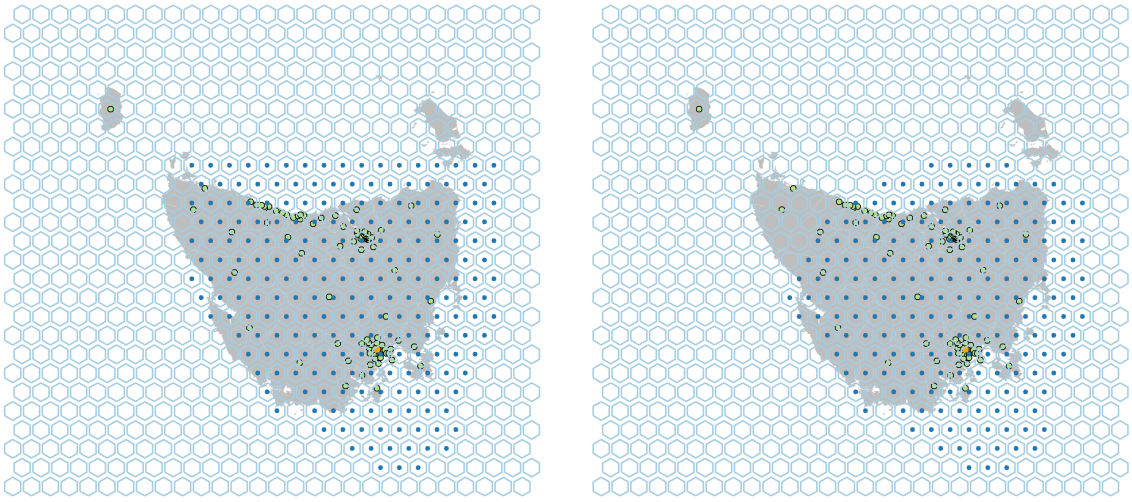
Keep only the available hexagon points, this will prevent multiple areas being allocated to the same hexagon.

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.

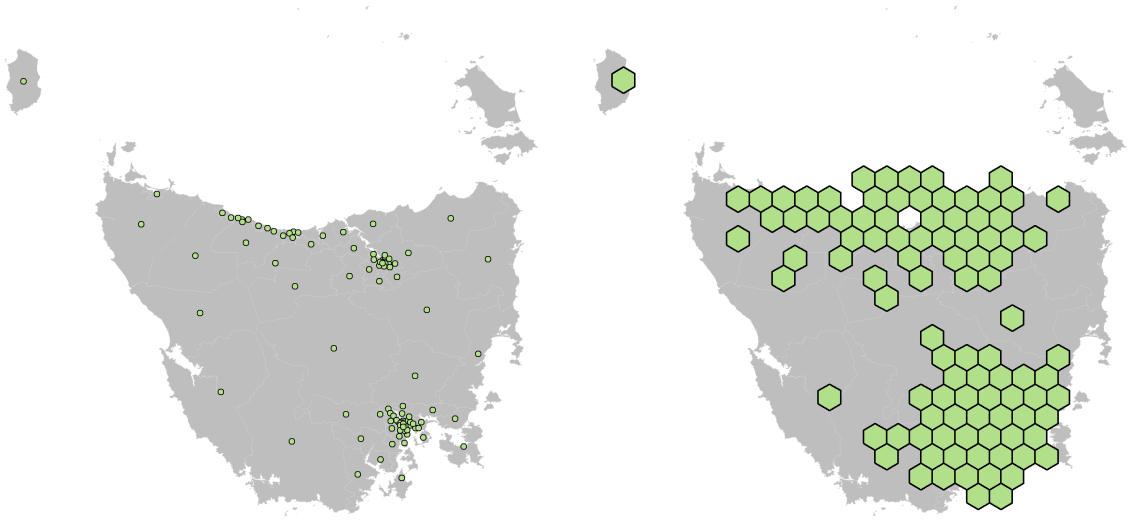
A box of possible hexagon locations around the centroid. The corners of the box may not look square as the buffer has already removed unnecessary points from over the ocean.

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 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.
The angle begins at 30 degrees by default, and will increase if no points can be found within the `hex_filter` distance. The allocation is returned and combined with the data relating to each polygon.



The following code creates a map for all the Statistical Areas at Level 2 in Tasmania:

3. Using sugarbag

```
R> # install.packages("sugarbag")
R> library(sugarbag)
R>
R> # Create centroids set
R> centroids <- create_centroids(tas_sa2, "SA2_NAME16")
R>
R> # Create hexagon location 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)
```

3.1. Neighbour relationships

It is possible to encourage the use of neighbourhood relationshipd, for stronger preservation of neighbour relations.

An additional step may 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 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.

4. Australian Cancer Atlas

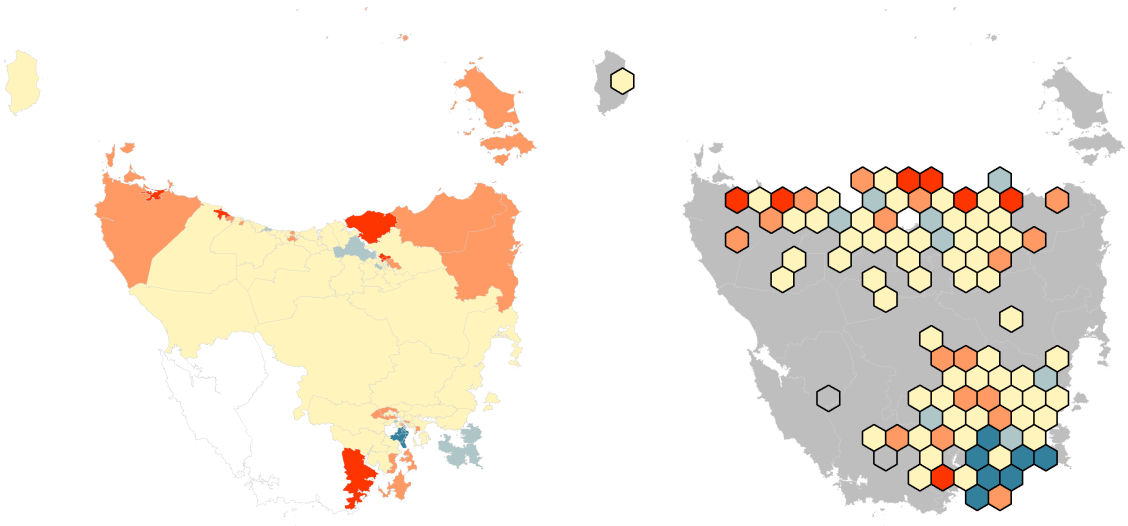


Figure 3: The Australian Cancer Atlas data has determined the colour of each Statistical Area of Australian at Level 2. A choropleth map (a) of Standardised Incidence Rates (SIRs) is paired with a hexagon tile map (b) to contrast the colours that are made obvious when every SA2 is equally represented.

The hexagon tile map visualisation method can be applied to the Australian Cancer Atlas data. A small example of Lung Cancer Standardised Incidence Rates (SIRs) allows two views of the same data produced by the Australian Cancer Atlas. This small example in Figure 3 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.

5. 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 Kobakian and Cook (2019) package written for 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.

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.

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 Mthematical Sciences, Science and Engineering Faculty, Brisbane, QLD, Australia
E-mail: stephanie.kobakian@qut.edu.au
URL: http://rstudio.com

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

Which is Better: a Choropleth Map or Hexagon Tile Map? A comparison using visual inference

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

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

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

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

Introduction

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

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

Motivation

Australian Cancer Atlas

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

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

Background

Methodology

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

Population focussed displays

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

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

Visual Inference

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

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

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

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

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

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

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

Study Design

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

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

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

Experimental design

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

TABLE I
THE EXPERIMENTAL DESIGN

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

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

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

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

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

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

The variables being manipulated and measured

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

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

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

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

The variables measured as a result of the changes were the probability of detection each display and the time taken to submit responses. To measure the accuracy of the detections, the plot chosen for each lineup evaluated was compared to the position of the real spatial trend plot in the lineup. A correct result occurs when the chosen plot matches the position of the

real plot, this was recorded in an additional binary variable; 1 = correct; 0 = incorrect. High efficiency occurs when a small amount of time is taken to evaluate each lineup. This will be measured as the numeric variable measuring the length of time taken to submit the answers to the evaluation of each line up.

Simulation process

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

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

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

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

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

Participants

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

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

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

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

Experiment procudure and data collection

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

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

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

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

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

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

The methods of data analysis used

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

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

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

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

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

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

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

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

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

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

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

Limitations of the data collection

This required internet connection for participants to access the survey

RESULTS

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

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

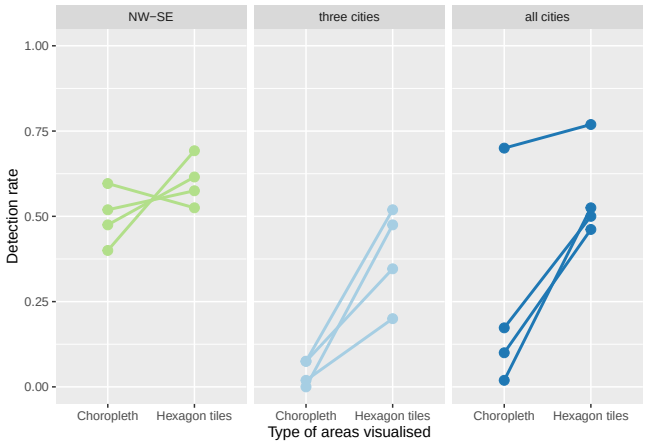


Fig. 1. Each point shows the probability of detection for the lineup display, separated by the trend model hidden in the lineup. The points for the same data set are linked to show the difference in the detection rate when the same data was seen in each display. 11 of the 12 real distribution plots were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

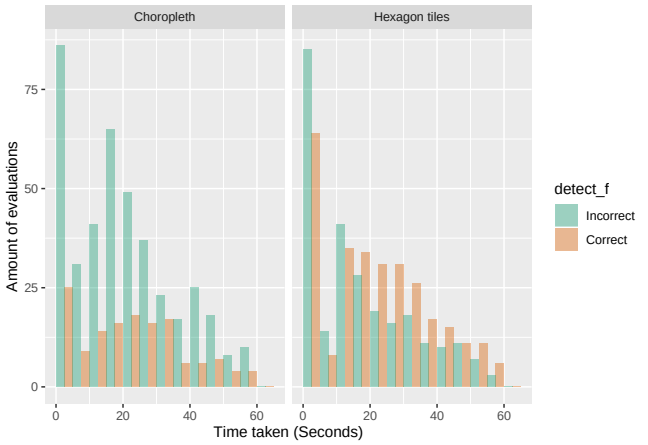


Fig. 2. The time taken to evaluate each display is broken into five second windows. The height of the histogram bars show how many evaulations were submitted within each time window. The distributions for the choropleth and hexagon tile maps are very similar. Both have a large peak at 0-5 seconds, and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

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

Demographics:

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

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

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

70 of the participants achieved a Bachelors or Masters degree.

Accuracy

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

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

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

Speed

Certainty

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

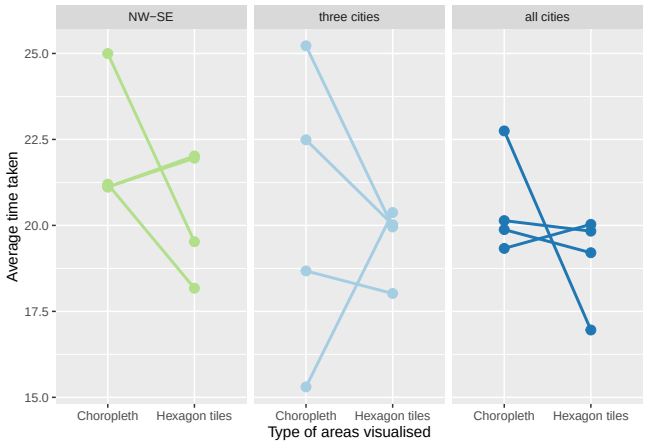


Fig. 3. Each point shows the average time taken for participants to evalute the lineup display, separated by the trend model hidden in the lineup. The points for the same data set are linked to show the difference in the avergae time taken when the same data was seen in each display. There is a lot of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

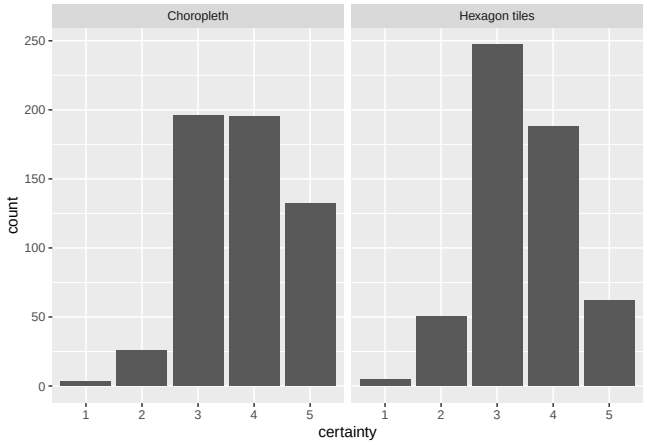
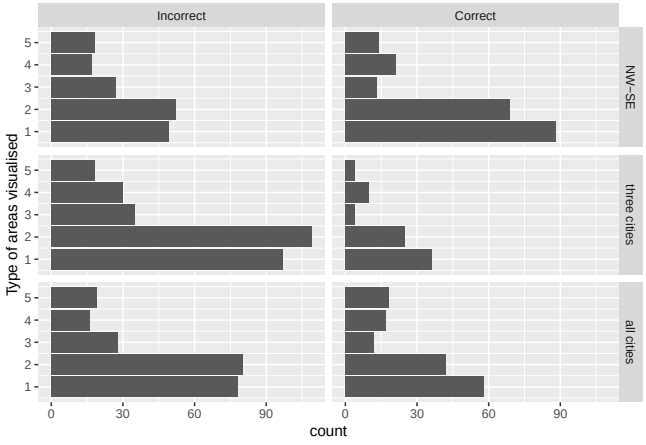


Fig. 4. The amount of times each level of certainty was chosen by participants when viewing hexagon tile map or choropleth displays. Participants were more likely to choose a high certainty when considering a Choropleth map. The default certainty of 3 was chosen most for the Hexagon tile map displays.

Reason

Contributors

Reason



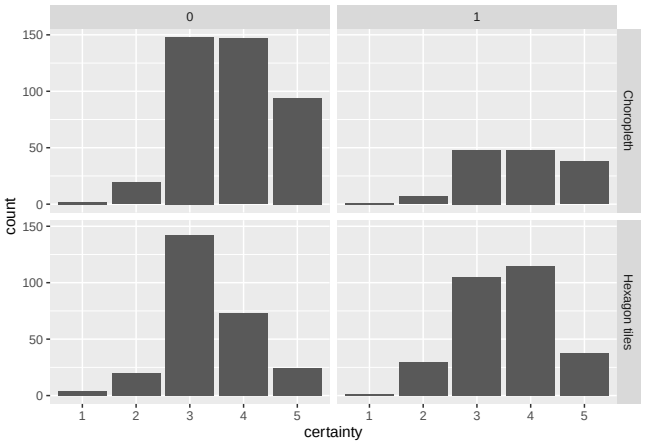


Fig. 5. The amount of times each level of certainty was chosen by participants. The columns shown whether a viewer correctly selected the real trend data plot when viewing hexagon tile map or choropleth displays.

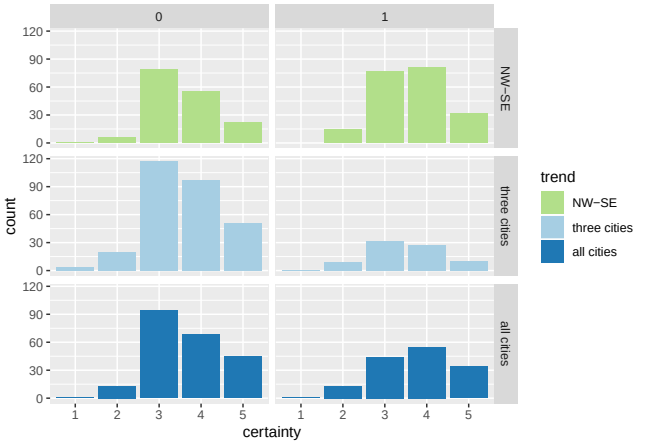


Fig. 6. The amount of times each level of certainty was chosen by participants. The columns shown whether a viewer correctly selected the real trend data plot when viewing hexagon tile map or choropleth displays. The rows show the type of the trend model added in the real data plot. The default certainty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

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

Anomolies

Modeling the difference

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

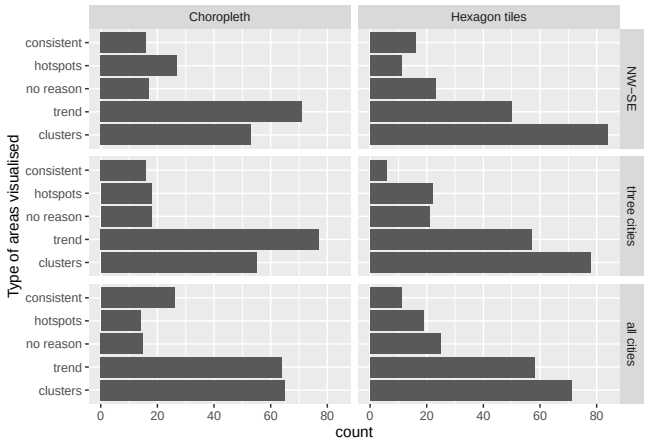


Fig. 7. The most common reason for choice of plot when looking at each trend model shown in Choropleth and Hexagon Tile maps. Clusters were the most common reason when viewing a Hexagon Tile map, trend was the most common choice for choropleth displays except for the all cities display.

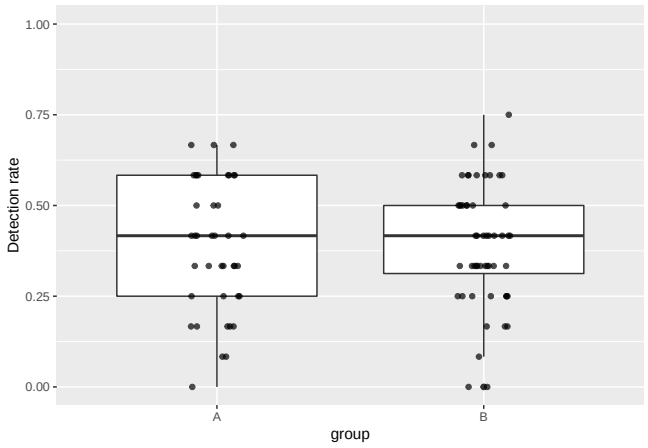


Fig. 8. The probablity of detection acheived by the contributors in each group is shown by the points. Group B has a larger range and a smaller inter-quartile range. Group A and both had 3 people who did not find any of the data maps in the displays.

Detection Rates:

```
## # A tibble: 6 x 5
##   term                                    estimate std.error statistic p.value
##   <chr>                                      <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)                                0.021     0.147     0.144   0.886
## 2 typeHexagon tiles                          0.420     0.121     3.47    0.001
## 3 trendthree cities                         -0.465     0.125    -3.72    0
## 4 trendall cities                           -0.241     0.120    -2.01    0.044
## 5 typeHexagon tiles:trendthree cities        0.465     0.188     2.47    0.013
## 6 typeHexagon tiles:trendall cities          0.312     0.171     1.82    0.069

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: detect_f
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr
## NULL                         1103     1477.0
## type         1   83.526      1102     1393.5 < 2
## trend        2  101.699      1100     1291.8 < 2
## type:trend   2   35.517      1098     1256.2 1.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
##
##              Incorrect Correct
##   Incorrect        431     121
##   Correct          242     310

## # A tibble: 1 x 4
##   sigma logLik   AIC   BIC
##   <dbl>  <dbl> <dbl> <dbl>
## 1 0.471 -657. 1321. 1357.

## # A tibble: 8 x 5
##   term                                    estimate std.error statistic p.value
##   <chr>                                      <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)                                0.505     0.147     3.44    0.001
## 2 typeHexagon tiles                          0.103     0.121     0.847   0.397
## 3 trendthree cities                         -0.467     0.125    -3.74    0
## 4 trendall cities                           -0.277     0.120    -2.31    0.021
```

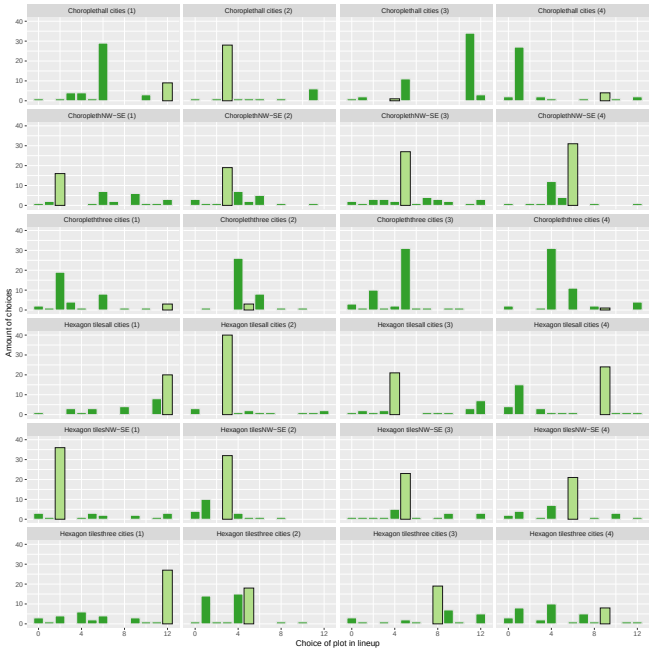


Fig. 9. Each facet is associated with one lineup, the height of the bars count the choices made by the participants considering each lineup. The bars coloured with black outlines show the map which contained a trend model, these are the correct choices. The numbers differentiate the replicates of each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

```
## 5 typeHexagon tiles:trendthree cities
## 6 typeHexagon tiles:trendall cities
## 7 sd_(Intercept).contributor
## 8 sd_Observation.Residual

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

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

Certainty:

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

```
0.250     0.0633   3.95 fixed
0.239     0.0633   3.77 fixed
0.118      NA       NA  contributor
0.430      NA       NA  Residual
```

```
                    typeHexagon tiles:trendall cities
                                               0.1035
                                              -0.2772
typeHexagon tiles:trendall cities                0.2391
```

Discussion

Conclusion

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

- For USA alternative methods can also be helpful

Supplementary Materials

Training

Survey application

Subject specific anomolies (0% detection)

Acknowledgment

The authors would like to thank…

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

Bibliography styles

References

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

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

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

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

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

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

Chapter 5

Discussion

Visualisation methods have improved in iterations over many years. Cartograms showed great promise and several algorithms were presented to create cartograms in the 1960s and 1970s. 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. Nusrat and Kobourov provide examples for techniques to evaulate the statistical, geographical, and topological accuracy of alternative visualisation displays.

Chapter 6

Conclusion

The goal 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, 2019) 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 alternative visualisation method solves the misprepresentation 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 several working example helps users to apply the functions to their own data sets.

The method outputs data sets to users that 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 evaulate 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.

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 diaplays, 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, Cook and Hofmann (Wickham et al., 2010), (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.

SURVEY APPLICATION DESIGN
– Survey –

Assessing the effectiveness of different visualisation methods for Australian spatial data.

QUT Ethics Approval Number 1900000991

This survey will be conducted using a web application hosted on a shiny server.
Participants will click the link from their Figure-Eight job after reading an introduction to the task.
It will be found at this web address:

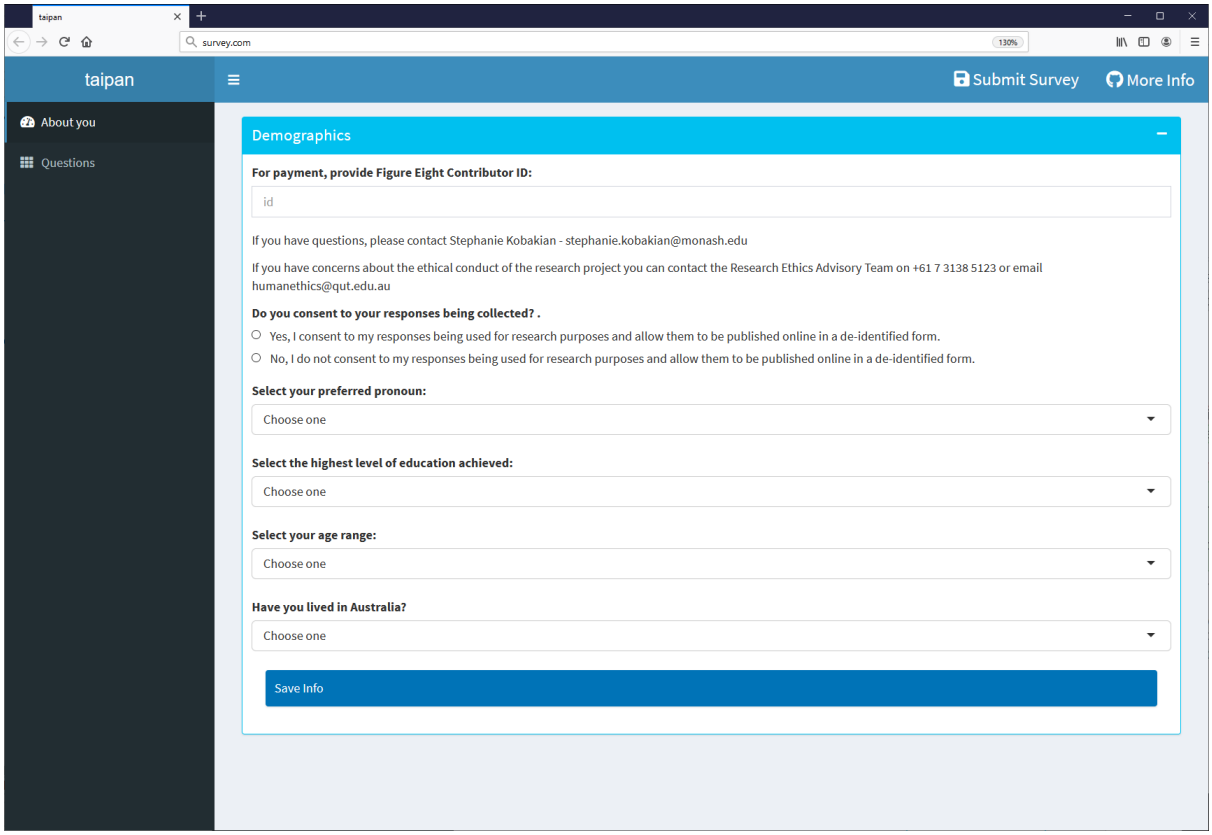


Figure 1: Web application survey, demographics questions page will allow users to provide information about themselves.

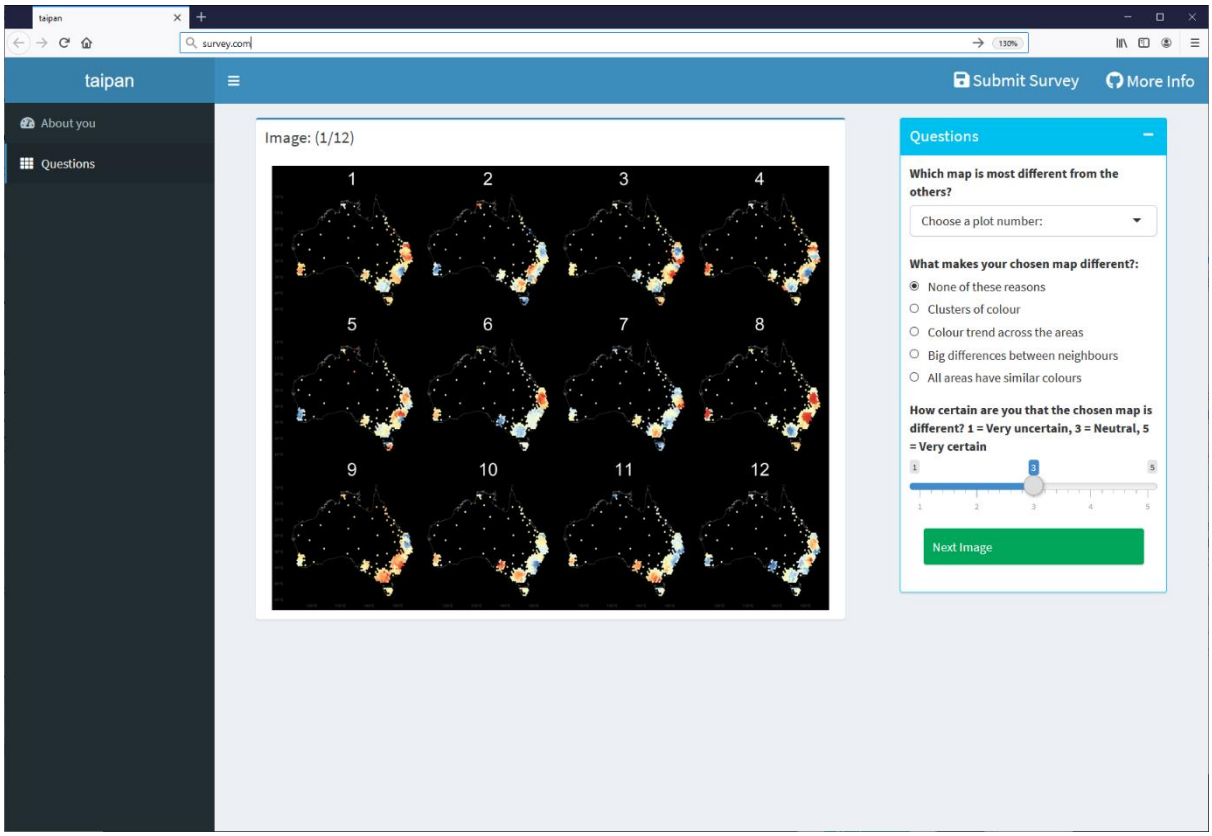


Figure 2: Web application page containing the survey questions that ask participants to choose a hexagon tile map from a lineup of maps

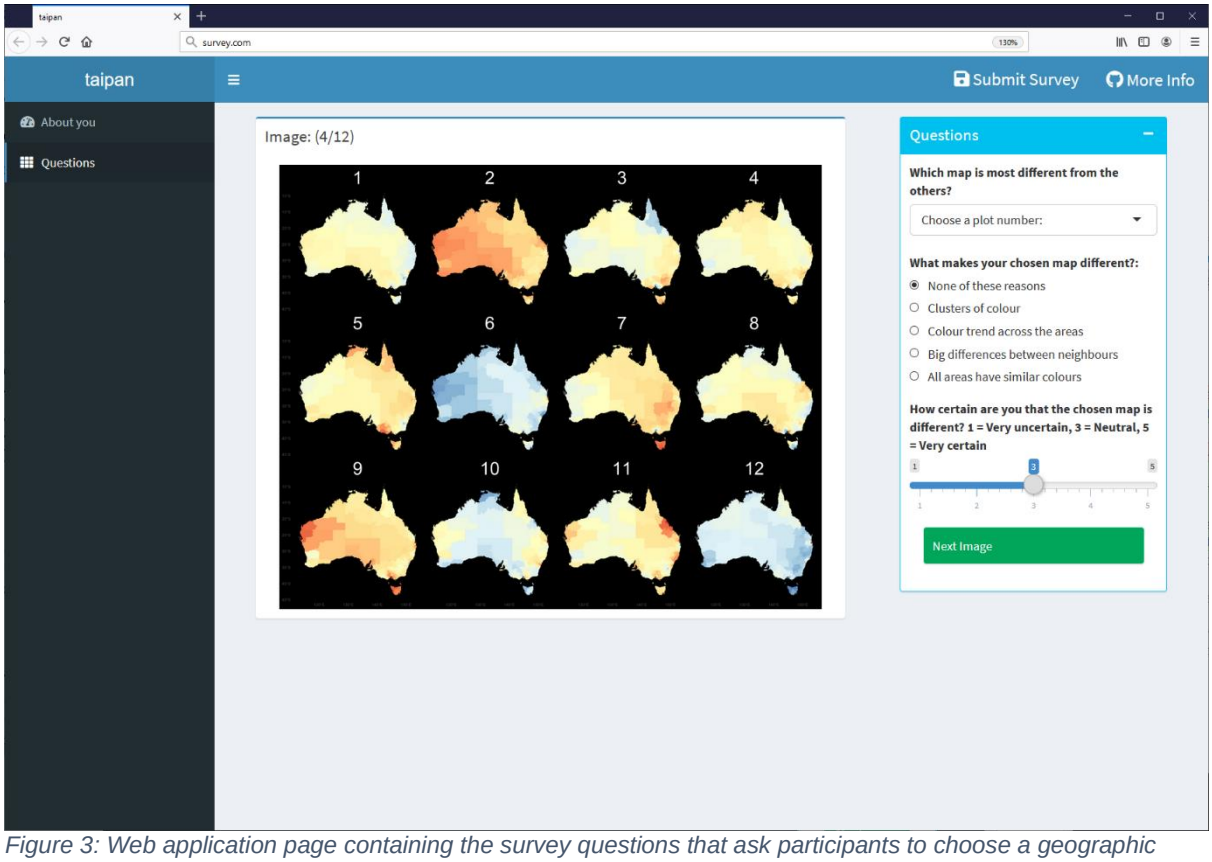


Figure 3: Web application page containing the survey questions that ask participants to choose a geographic map from a lineup of geographic maps

Appendix A

Ethics Approval

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:

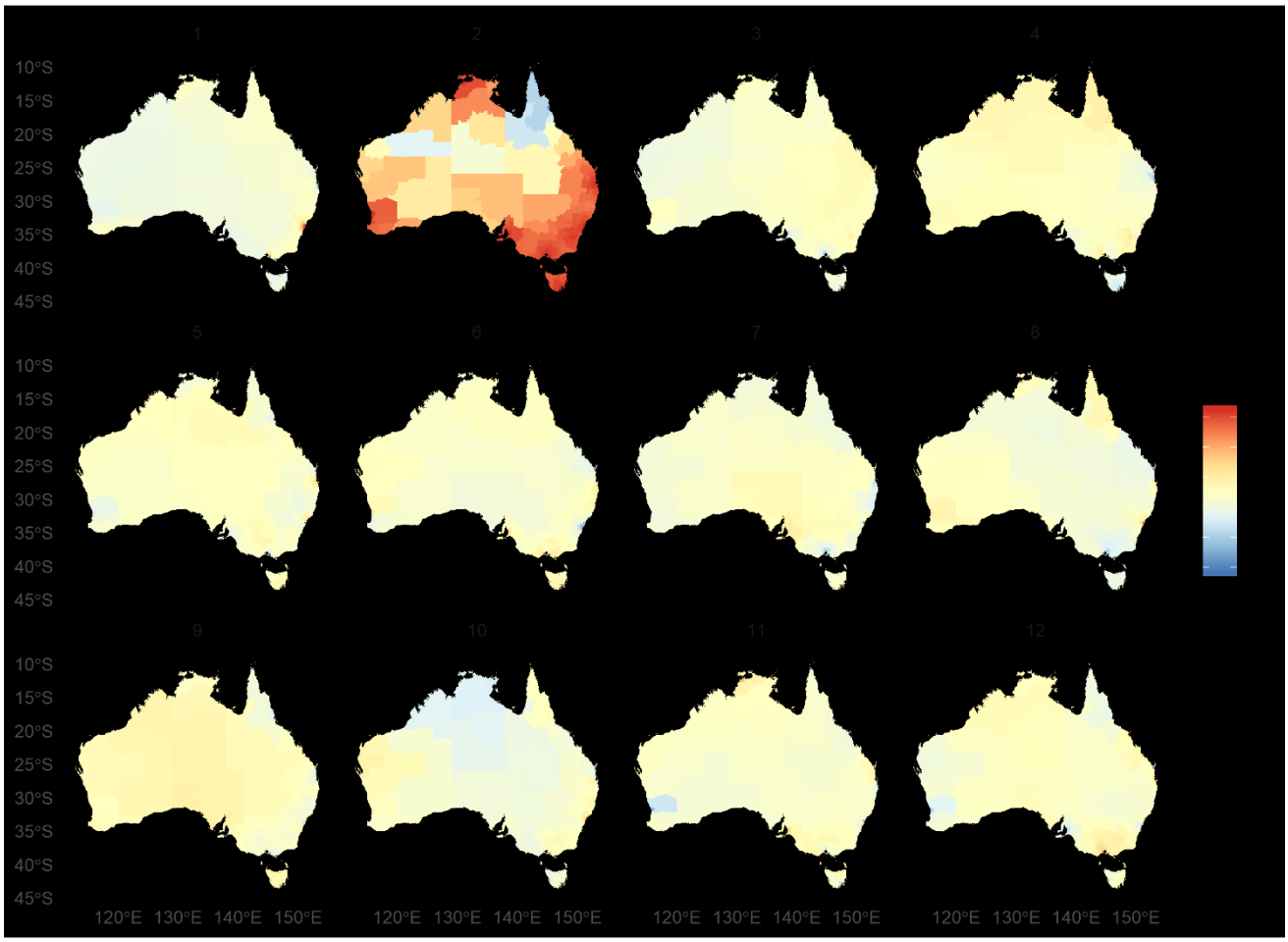


Figure 1. A lineup of geographic maps of Australia. Each sa3 has been coloured according to a simulated data set. Only one of these maps displays a spatial relationship, the rest are null plots, with colours shuffled between the areas.

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

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

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.

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.

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.

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

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

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

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.

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.

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.