

# MPhil Research Proposal

*Stephanie Kobakian*

*2018-11-17*

## Contents

<b>The Proposed Title</b>	<b>1</b>
<b>The Proposed Supervisory Team</b>	<b>1</b>
<b>Background and Literature Review(Maximum 1,400 words, 760ish now)</b>	<b>2</b>
Introductory Statement . . . . .	2
Literature Review . . . . .	2
Research Problem (e.g. aims, questions and/or hypotheses) . . . . .	4
<b>Program and Design of the Research Investigation (Maximum 2,000 words)</b>	<b>5</b>
Objectives, Methodology and Research Plan . . . . .	5
Resources and Funding Required . . . . .	6
Timeline for Completion of the Program . . . . .	6
<b>Reference List (Word count not included for proposal)</b>	<b>7</b>
Appendix (Word count not included for proposal) . . . . .	8
Coursework . . . . .	8

## The Proposed Title

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

## The Proposed Supervisory Team

*Principal Supervisor:* Professor Kerrie Mengersen; Science and Engineering Faculty, School of Mathematical Sciences

*Associate Supervisor:* Dr Earl Duncan; Science and Engineering Faculty, School of Mathematical Sciences

*Non QUT Associate Supervisor:* Professor Dianne Cook; Econometrics & Business Statistics Faculty, Monash University.

# Background and Literature Review(Maximum 1,400 words, 760ish now)

## Introductory Statement

Data visualisations have enhanced the understanding of problems in many scientific fields. A classic example in disease data is the map of London, pinpointing water pump locations among cholera occurrences, made by Dr John Snow, that made the case for the cause being contaminated water. A nice discussion of this can be found in (Tufte et al. 1998).

Many disease data sets are distributed in aggregate on political areas rather than occurrence location. This may be for various reasons, such as privacy, and also ability to respond often depends on the political entity responsible. This data has measurements associated with small spatial areas, and the typical visualisation is a choropleth map, where areas are coloured by the numerical value. A purpose of making the map is to understand the spatial distribution of disease occurrence, and also locate disease clusters where remedial action may be needed. The problem is, especially for Australia, is that the spatial distribution is obscured by the prevalence of vastly different sizes of areas. This research explores more effective visualisation to better represent the spatial distribution, with a focus on Australian disease mapping. The work is motivated by the Cancer Atlas of Australia, which presents the spatial patterns of Incidences of many cancers in Australia.

## Literature Review

### Statistical maps of geographic areas

Waldo Tobler (2004) explores the development process from choropleth maps to cartograms and considers the situations which favour either display. The visualisations commonly used for mapping geographical data are choropleth maps and cartograms. Choropleth methods produce geographically accurate maps, and represent data using colour, transparency and other methods. The geographical areas are presented according to their geographic shape, size and location. This keeps spatial relationships intact, and the data of the surrounding neighbours can reasonably be considered related. Alternatively, cartograms are used to focus on the spatial distribution within the geographically related data. The geographical features are foregone, in order to visually analysis the spatial distribution and emphasise the statistics.

Cartograms have been used as a visual display since 1851 (Tobler 2004). Sabrina Nusrat and Stephen G. Kobourov (2018) recognise four types of cartograms: Contiguous, Non-Contiguous, Dorling and Rectangle. All of these methods change the size of the areas to represent a statistic. Rectangular (Kreveld and Speckmann 2007) cartograms have consistent shapes to represent regions and sizes differ to represent a statistic, Dorling cartograms use circles (Dorling 2012). The term ‘mosaic cartograms’ was introduced (K. Buchin & T. Castermans & A. Pieterse & W. Sonke & B. Speckmann 2015) to describe an alternative cartogram display. Carr, Olsen and White’s (Carr, Olsen, and White 1992) ‘hexagon mosaic map’ discussion cites Carr’s previous work that finds hexagons provide visual appeal and ‘representational accuracy’. The authors credit nature for the concept of tessellated hexagons tiles, which their design employs. These maps also provide as little disruption as possible to the area adjacencies, especially at lower resolutions where smaller hexagons capture more geographical details.

Daniel Dorling’s cartograms allow one circle for each area (Dorling 2012). The size of these circles differs based on the value of the statistic. It maintains the spatial distribution for each area by ‘touching as many others that it originally neighboured as possible while touching as few as possible that it did not.’ This concept is key to the visual analysis of spatial distributions, especially for the visual analysis of small geographic areas.

“On any traditional map of an urbanised country, the majority of political constituencies are literally not visible to the naked eye (Figures 3.8 and 3.9). The problem is particularly acute in countries such as Canada and Australia, but still fundamental in all other regions of the world. The argument is not that the conventional map distorts the message; it is that it cannot correctly convey even a very small fraction of it.” (Dorling 2012)

### **Evaluating alternative representations**

The cartogram methods differ based on the preservation of shape, topography and the statistical accuracy. The effectiveness of the methods have been contrasted (Nusrat, Alam, and Kobourov 2018) and the algorithms have been evaluated in terms of time and error, and the visualisation subject preferences. These measures may now be supplemented with a new approach.

Graphical Inference is introduced by Wickham, Cook, Hofmann and Buja to determine if patterns that appear in visualisations are ‘really there’(2010). Their approach has been applied to common graphics, and presents the opportunity to measure effectiveness of a new graphic in comparison to established visualisation methods.

The question: “Is there a spatial trend?” is historically associated with choropleth maps (Wickham et al. 2010). An example has been displayed of one true data set of spatial data, presented in a line-up of null plots. These tests acknowledge a trend, pattern or relationship in the true data, if it appears significantly different from the null plots in the line-up.

Adapting this approach could be used to test the effectiveness of the choropleth map, against an alternative graphic. By contrasting the results of the choropleth and alternative line-ups that present the same true data in line-ups of null plots. This contrast could be used to test statistically the effectiveness of the visual display in communicating spatial data. The R package, ggplot2 (Wickham 2016), is able to create plots and line-ups of plots that can be used in this testing.

### **Enhancing applications using animations**

They are regularly used in the United States, most recently to present results and race calls from The Associated Press for the mid-term elections (Almukhtar et al. 2018) along with the 2017 primary elections (Kimelman, Chaumont, and Swartz 2017).

Applications of cartograms (Nusrat and Kobourov 2016) have been seen in social, political and epidemiological contexts. The Christopher Kocmoud, and Donald House, are cited stating ‘cartogram can better show the distribution’ (Kocmoud and House 1998) when considering the spread of a disease. The move toward alternative mapping techniques has progressed with the accessibility of animations. Especially in the statistical computing community which has recently been improved with the development of ganimate (Pedersen and Robinson 2017).

Projecting data into the recognisable form of the geography, and then transforming to an alternative representation allows deeper understanding of the data. Thomas Lin Pedersen (Pedersen and

Robinson 2018) suggests animating visualisations as it ‘demands attention’. This is especially convenient to direct attention, as animations do not assume captive audiences, instead inform by narrating the intended message to be communicated by the visualisation.

### **Research Problem (e.g. aims, questions and/or hypotheses)**

One feature of Australian maps is that no group of areas are homogeneous. The size of the states vary greatly, and this is feature persists for all ABS Structures and Non ABS Structures. Instead, the size of these areas is usually driven by population density. The inverse relationships means less significant areas can take up substantial map space. Using most mapping techniques to get a broad perspective of Australia can be misleading, the use of geographical areas misrepresents the spatial distribution of a dataset.

We will represent each area equally on the map space. The use of colour, transparency and symbols will still be available for variables.

To test the effectiveness of these visualisations we will examine the accuracy of reading spatial distributions. Applying visual inference tests will measure the increase in accurate understanding of spatial distributions. These experiments will be based on Hofmann’s (Hofmann et al. 2012) work in visual inference.

### **Aims**

1. *Algorithm for hexmapping Australia:* The algorithm will take geospatial areas in the form of polygons, and create an alternative graphical display of the variable that is believed to be spatially distributed.
2. *Test the effectiveness of the hexmap relative to the choropleth map for providing a more accurate reading of spatial distribution for Australia:* The display produced by the algorithm will be contrasted with the traditional choropleth map, applying the same colour and transparency methods to represent the data values. The maps will be presented in the form of experiments, to test the effectiveness by asking for interpretation of spatial distributions.
3. *Communicating the relationship between the hexmap and choropleth map through animation algorithm:* Finally, we recognise the value of presenting users with standard maps that are familiar, if we prove the belief that the alternative method enhances interpretation of the spatial distribution we will aim to maximise the benefits of both. The use of 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 has been used for many years, and is gaining popularity as access to computing power is increasing the amount of applications.

# Program and Design of the Research Investigation (Maximum 2,000 words)

## Objectives, Methodology and Research Plan

*should clearly identify the tasks to be undertaken and how these address your research problem; may be organised in relation to each of the individual aims or questions, and identify specific methods of experimentation for those conducting laboratory based work;*

*And should include a clear if preliminary statement of the theoretical/experimental framework underpinning how you are going to carry out the design / plan.*

## Chapter 1: Algorithm for hexmapping Australia

Produce alternative mapping strategies (cartograms informed by Xiaoyue Cheng). While there are visualisations available for spatial data, we can benefit by exploring and creating new displays. During the past few months work has progressed that contributes to this first aim. We have been able to achieve a working algorithm for a hexmap of Australia [link Appendix with code]. This algorithm takes a set of polygons and creates a hexmap, representing each geographical area with a single hexagon. They are arranged so that hexagons are placed to replicate spatial relationships of areas in each city.

To create a display, the user needs to provide the desired size of the hexagons for each individual area. Using this, we create a grid of all possible locations for relocated, tessellated polygon centroids. This grid spans the bounding box of the centroids for the geographical areas.

The order of allocation depends on the ordered dataset of polygon centroids passed to the function. For this set of polygons, the distance has been calculated from each polygon centroid to the closest capital city. This approach allows the polygons in each city to progressively expand out. It is also a valid approach for rural areas, as they have little competition for the grid point closest to their polygon centroid.

To complete Chapter 1, we plan to incorporate repelling and attraction [link Xiaoyue Cheng] mechanisms inspired by Xiaoyue Cheng. This could allow rural areas to bunch together and allow easier comparison of colours or symbols.

## Chapter 2: Testing the effectiveness of the hexmap relative to the choropleth map

To test the effectiveness of the alternative mapping techniques, we will produce an experiment hosted on Amazon's (2008) Mechanical Turk. The 'concept of a 'null distribution of plots' as the analogue of the null distribution of tests statistics' [stat inf EDA Buja] using 'the line-up' protocol Buja suggests.

We have the opportunity to treat a visualisation as a statistic, visualisation may be compared just as numerical features of models or data summaries can. This is worth doing if methods show statistical distributions are present and not noticed using geographical maps. Results for aggregated areas in disease data are often standardised to be comparable to other areas, and area boundaries are driven by population.

Therefore we propose the following experiment for 30 people:

Factors: 1. Plot Type: - Hex map and Choropleth map 2. Distribution of estimates: - Normal - Skewed: right; similar to the Cancer Atlas data - Clustered 3. Spatial dependence - Random: values of spatial areas are randomly pulled from relevant distribution - Spatial Dependence: values of spatial areas are randomly pulled from a range surrounding their neighbours, within the relevant distribution

Replications: 4 of each 6 control combinations, where the four replications are randomised draws of data

Randomisation: Participants will be shown 24 of the line-ups produced. Two reps of each combination of 2 and 3.

The entire experiment will be replicated with new simulated data sets and a different set of 30 people.

The experiment will form a survey of line-up plots of Australian maps. Each participant will see a random selection of the 6 (or 12) control combinations, they will see both the choropleth map and the hex map.

To distinguish the effectiveness of the maps, we will contrast the participant's recognition of the spatial distributions when displayed in each map type. A statistically significant difference in the group means will be computed in R, a software for statistical analysis.

### **Chapter 3: Communicating the relationship between the hexmap and choropleth map: animation algorithm**

Animations will be created using R statistical software.

### **Resources and Funding Required**

A request for funding will be submitted to undertake the line-up experiment on Amazon Mechanical Turk. We suggest 90 respondents at \$5 each,

We will utilise the currently available equipment. A self provided laptop with access to the open source R software. We acknowledge the funding of the Research Training Program (RTP), ACEMS for supporting the author.

### **Timeline for Completion of the Program**

1. *Algorithm for hexmapping Australia*
2. *Testing the effectiveness of the hexmap relative to the choropleth map*
3. *Communicating the relationship between the hexmap and choropleth map: animation algorithm*
  - A. Addition activities: coursework units, deadlines, thesis preparation

(Expected) Months	(Expected) Time	Task	Status
<i>Semester 2 2018</i>		- (A) FIT9133: Programming fundamentals in Python: Semester 2, Monash University	

(Expected) Months	(Expected) Time	Task	Status
June - August	<b>3 months 1 week</b>	- (1) Allocation of polygon area centroids to hex map grid centroids - (A) useR!2018 Conference - (1) Arrangement of centroids to allocate in order - (1) Incorporate capital city relationships and directional centroid allocations	Completed
September	<b>1 month</b>	- (1) Tidy to allow provision of polygons	Planned
October - November	<b>2 months</b>	- (1) Automate tuning for size of hexagons	
December	<b>1 month</b>	- (A) Prepare thesis chapter for submission	Planned
January - February	<b>2 month</b>	- (2) Preparation of survey questions	
<i>Semester 1 2019</i>		- (A) ETC4541: Bayesian time series econometrics: Semester 1, Monash University	
March	<b>1 month</b>	- (2) Execution of experiment	
April	<b>1 month</b>	- (A) Annual Progress Report	
May - June	<b>2 months</b>	- (2) Data Analysis	
July	<b>1 month</b>	- (A) Prepare thesis chapter for submission	
<i>Semester 2 2019</i>			
(Expected) Months	(Expected) Time	Task	Status
August - September	<b>2 months</b>	- (3) Apply transformation algorithm, polygons to hexagons	Planned
October	<b>1 month</b>	- (3) Prepare thesis chapter for submission	
November	<b>1 month</b>	- (A) Complete masters thesis	

## Reference List (Word count not included for proposal)

- Almukhtar, Sarah, Mike Andre, Wilson Andrews, Matthew Bloch, and Jeremy Bowers. 2018. "House Election Results: Democrats Take Control." *The New York Times*, November. A. G. Sulzberger. <https://www.nytimes.com/interactive/2018/11/06/us/elections/results-house-elections.html>.
- Carr, D. B, A. R. Olsen, and D. White. 1992. "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographi- Cal Data." *Cartography and Geographic Information Systems* 19 (4): 228–36.
- Dorling, Daniel. 2012. *The Visualisation of Spatial Social Structure*. John Wiley & 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): 2441–8. <https://doi.org/10.1109/TVCG.2012.230>.

K. Buchin & T. Castermans & A. Pieterse & W. Sonke & B. Speckmann, R. G. Cano &. 2015. “Mosaic Drawings and Cartograms.” *Computer Graphics Forum* 34 (3): 361–70.

Kimelman, Jeremia, Kriss Chaumont, and Samantha Swartz. 2017. “Mapping the Race to Control the House.” *Nbcnews*, March. NBCUniversal. <https://www.nbcnews.com/politics/elections/mapping-race-control-house-n906076>.

Kocmoud, Christopher, and Donald House. 1998. “A Constraint-Based Approach to Constructing Continuous Cartograms,” January.

Kreveld, M. van, and B. Speckmann. 2007. “On Rectangular Cartograms.” *Computational Geometry: Theory and Applications* 37 (3): 175–87. <https://doi.org/10.1016/j.comgeo.2006.06.002>.

Nusrat, Sabrina, and Stephen G. Kobourov. 2016. “The State of the Art in Cartograms.” *Comput. Graph. Forum* 35: 619–42.

Nusrat, S., M. J. Alam, and S. Kobourov. 2018. “Evaluating Cartogram Effectiveness.” *IEEE Transactions on Visualization and Computer Graphics* 24 (2): 1077–90. <https://doi.org/10.1109/TVCG.2016.2642109>.

Pedersen, Thomas Lin, and David Robinson. 2017. *Gganimate: A Grammar of Animated Graphics*. <http://github.com/thomasp85/gganimate>.

———. 2018. “The Grammar of Animation.” 2018. <https://youtu.be/21ZWDrTukEs>.

Tobler, Waldo. 2004. “Thirty Five Years of Computer Cartograms.” *Annals of the Association of American Geographers* 94 (1). Routledge: 58–73. <https://doi.org/10.1111/j.1467-8306.2004.09401004.x>.

Tufte, Edward R, Susan R McKay, Wolfgang Christian, and James R Matey. 1998. *Visual Explanations: Images and Quantities, Evidence and Narrative*. AIP.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <http://ggplot2.org>.

Wickham, H., D. Cook, H. Hofmann, and A. Buja. 2010. “Graphical Inference for Infovis.” *IEEE Transactions on Visualization and Computer Graphics* 16 (6): 973–79. <https://doi.org/10.1109/TVCG.2010.161>.

## Appendix (Word count not included for proposal)

### Coursework

In all cases, your required coursework needs to be based upon a research degree skills audit and a written plan briefly setting out the educational outcomes expected from the coursework. This coursework is planned together with your Supervisors to contribute and provide structure to your overall program of research.

Courses undertaken at QUT: IFN001 Advanced Information Research Skills (AIRS)



Courses undertaken via cross institutional study at Monash University: FIT 9133 Programming fundamentals in Python ETC 4541 Bayesian time series econometrics