Proposal

Stephanie Kobakian

21 August 2018

Contents

The Proposed Title **1**

The Proposed Supervisory Team **1**

Background and Literature Review(Maximum 1,400 words, 760ish now) **2**
Introductory Statement 2
Literature Review (Creat bib references file) 2
Research Problem (e.g. aims, questions and/or hypotheses) 3

Program and Design of the Research Investigation (Maximum 2,000 words) **3**
Objectives, Methodology and Research Plan 4
Resources and Funding Required 5
Timeline for Completion of the Program 5

Reference List (Word count not included for proposal) **7**
Appendix (Word count not included for proposal) 7
Coursework 7

The Proposed Title

What are good practices in visualising geo-spatial and temporal disease data, and estimates?

Comparison of methods for visualising 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

Visualisation of data has contributed to understanding of problems in many fields. Disease data has been visualised for centuries, Edward Tufte and his co-authors (1998) recognised that Dr. John Snow's dot plot map of cholera occurences provided a 'direct and poweful testimony about a possible cause-effect relationship' as it drew attention to a contaminated water pump. The connection between disease occurence and their causes often lies in the physical locations of the people affected. Dr. Snow's analysis highlighted this, and the benefits to using a map to communicate disease data.

A typical approach to plot this data related to small areas is to make a map with colour representing the numerical value associated with the spatial area, which is called a choropleth map. The purpose of making the visualisation is to understand the spatial distribution of disease occurrence. When some areas are small and others are large, maps can hide features of a spatial distribution. This research explores ways to better represent the spatial distribution, with a focus on Australian disease mapping. The work is motivated by the Cancer Atlas of Australia.

Disease maps may seem common now, and the visual analysis of spatial distributions has progressed to exploring aggregated areas, incorporating colours and transparency as well as symbols. Disease data in Australia is usually distributed by medical registries and government organisations as aggregated values for small spatial areas. Australia utilises ABS Structures, 'areas that the ABS designs specifically for outputting statistics' and Non ABS Structures which are politically determined 'administrative areas for which the ABS is committed to providing a range of statistics'. We will aim to contribute an alternative map to help communicate spatial distributions.

Literature Review (Creat bib references file)

Two visualisations commonly used for mapping geographical disease data are choropleth maps, and cartograms. Choropleth methods use geographical map bases, and represent data using colour, transparency and other methods. The geographical areas represented are kept constant, this keeps spatial relationships intact, the data of the surrounding neighbours can reasonably be considered related.

Waldo Tobler (2004) explores the development from choropleth maps to cartograms and differentiates their use by stating 'in contrast to the usual geographic map, the most common use of cartograms is solely for the display and emphasis of a geographic distribution'. This allows an acknowledgement that choropleths are useful for looking at specific geographical areas, and their associated value, but to consider the spatial distribution of a group of areas, cartograms should be favoured. Cartograms have been recognised as a viual display since 1851 (Tobler 2004) however their use in Australia has been limited, namely due to the way Australian area boundaries have been influenced by the population. There is need for alternative methods to maintain spatial relationships when visualising spatial distributions.

One feature of Australian mapping is that no group of areas are homogenous. The size of the states vary greatly, and this is true for all ABS Structures and Non ABS Structures. However the size of these areas is usually driven by population density. The inverse relationships mean less siginificant

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 misprepresents the spatial distribution of a dataset. Rather than cartogramming, which manipulates the map space of a geogrpahical area according to the value, we will consider representing each area equally on the map space. The use of colour, transparency and symbols is still available.

This project proposes two key stages, algorithm development and tests of effectiveness. Proposing ideas is valuable, and the creation of an algorithm to create the maps useful too. However they must be able to communicate spatial distributions effectively. The implementation of the maps will only be helpful if they can allow the people using them to draw conclusions or raise qustions that were not immediately obvious using a geographical map.

To test the effectiveness of types of plots, and the accuracy of reading spatial distributions we will apply visual inference tests. These experiments will be based on Hofmann's [] work in visual inference.

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

Three key problems this work will address.

1. Algorithm for hexmapping Australia
2. Test the effectiveness of the hexmap relative to the choropleth map for providing a more accurate reading of spatial distribution for Australia.
3. Communicating the relationship between the hexmap and choropleth map: animation algorithm

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.

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.

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 tranform a reconisable 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)

We intend to address the program in three chapters:

1. Produce an 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

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 algoirthm 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, tessalated 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 aprroach 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 lineup' 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: - Hexmap 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 lineups 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 lineup 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 lineup 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

Your schedule for completing the various aspects of your program needs to be illustrated by a timeline so that you, your supervisors, your faculty and the Research Degrees Committee can see your expected rate of progress. The timeline includes each of the tasks identified in Section 4.1 and should normally include monthly targets. As well as your significant milestones (Annual Progress Report date/s, planned Research Seminar and Expected Lodgement Due Date), it could include your coursework requirements, transferable skills and other module training as well as proposed conference attendance and field trip travel. Research Degrees Committee will expect to see a proposed eighteen month* (from admission) full time equivalent completion timeline. Your proposed timeline is to support you in maintaining progress as well as ensuring that your project is realistically scoped within the maximum course completion time frame. Approaches to creating a

timeline are many and varied but this is an important section of your application. For a simple and generic example that includes details that you should be incorporating into your own timeline please contact your faculty.

Proposed Timeline:

Beginning June 2018, approximate conclusion December 2019

1. *Algorithm for hexmapping Australia*

Status	Task	(Expected) Time	(Expected) Months
Completed	- Allocation of polygon area centroids to hex map grid centroids	**3 months**	June - August
	- useR!2018 Conference	**1 week**	
	- Arrangement of centroids to allocate in order		
	- FIT9133: Programming fundamentals in Python: Semester 2, Monash University		
	- Incorporate capital city relationships and directional centroid allocations		
Planned	- Tidy to allow provision of polygons	**1 month**	September
	- Automate tuning for size of hexagons	**2 months**	October - November
	- Prepare thesis chapter for submission	**1 month**	December

2. *Testing the effectiveness of the hexmap relative to the choropleth map*

Status	Task	(Expected) Time	(Expected) Months
Completed	- Factorial Design		
Planned	- Preparation of survey questions	**2 month**	January - February
	- ETC4541: Bayesian time series econometrics: Semester 1, Monash University		
	- Execution of experiment	**1 month**	March
	- Annual Progress Report	**1 month**	April
	- Data Analysis	**2 months**	May - June
	- Prepare thesis chapter for submission	**1 month**	July

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

Status	Task	(Expected) Time	(Expected) Months
Planned	- Apply transformation algorithm, polygons to hexagons	**2 months**	August - September
	- Prepare thesis chapter for submission	**1 month**	October

Status	Task	(Expected) Time	(Expected) Months
	- Complete masters thesis	**1 month**	November

Reference List (Word count not included for proposal)

Tobler, Waldo. 2004. "Thirty Five Years of Computer Cartograms." *Annals of the Association of American Geographers* 94 (1). Routledge: 58–73. doi: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.

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