Advanced Undergraduate Research Opportunity Programme in Science

Visualisation of Area-Specific Demographic Distributions on a Physical Map

Dion Kwan Zheng Kai

Supervisor:
A/Prof Yap Von Bing

Department of Statistics
National University of Singapore
AY19/20 Semester II

A report submitted in fulfillment of the requirements of NUS Advanced Undergraduate Research Opportunities Programme in Science (UROPS) in Statistics and Applied Probability

Contents

Abstract

The Singapore census yields summary statistics on multiple demographic variables for each planning area. These type of demographic variables include - Race, Religion, Population size, etc. The visual representation of categorical variables on a physical map can be a rich source of information, and may lead to new insights on the evolution of demographic variables in time. Very commonly seen in other National Statistical Boards are breakdowns of such multivariate variables into univariate/bivariate graphical outputs for ease of computer generation and audience understanding.

Commonly used thematic map types and techniques for visualising spatial data include - Choropleth maps, Heat maps, Proportional symbol maps and Dot density maps.

The objective of this project is to create tools in the programming language R for representing such demographic summaries on a map, via one of the methods as mentioned above, or a combination of them.

Introduction

Representing geospatial data has always been of interest all around the world. The oldest known maps are dated back to the Babylonian era of 2300BC. Maps and cartography have since evolved to take many different forms, representing different variables. Geographic information systems (GIS) emerged in the 1970-80s period and we are now able to generate infographic maps, and even use them for prediction.

We can view a collection of multivariate maps via this link.[1] Many of these multivariate maps indicate a 3-Dimensional projection, and the trivariate maps seem to be too overwhelming for the normal population to comprehend. However, an interesting easy-to-comprehend thematic scheme can be seen in the embedded maps.

We will use the US Official Census Data Mapper[2] as an example of cross-examining/referencing the use of multivariate mapping in Government Statistics Boards. While the Singapore Department of Statistics(DOS) does not provide flexible applets to chart out demographic information on its website, we can see that the US Census Mapper only allows for univariate charting according to the selected variable.

Thus, the main goals of our study include:

1. Create tools using R and GIS capabilites to produce a novel multivariate variable map to represent demographic summaries.

2. The new type of multivariate variable map must be easily comprehensive, without

significant loss of information.

3. To be able to extend such graphing functions to a larger scale, possibly including the generation of dashboards, and to apply it on more datasets.

Project Scope

Following the project main goals, We will hence define the scope for this project:

- We will only examine the 4 ethnic proportions(Chinese, Malay Indian, Others) of SG residing in each Planning Area(PA) defined by the Urban Redevelopment Authority.
  - This is due to the fact that Singapore is a small country - $721.5km^2$. It is impractical in practice to represent such summaries onto a country with 323 subzones(SZ) and sub-subzones(SSZ). Each small area will have to be blown up to significant proportions to be seen properly on the actual map of Singapore.
  - Example of SZ: Yio Chu Kang, Ang Mo Kio
  - Example of SSZ: Yio Chu Kang North, Ang Mo Kio Central
- There is insufficient data on the ethnic proportions for us to be able to graph each of the 323 areas in the Singapore Map, resulting in many areas having missing data.
- We define the Total Population of SG as N, and N_C, N_I, N_M, N_O for the Population of each ethnic group: Chinese, Indian, Malay, Others, respectively.
- Similarly, we use p to express ratio of ethnic groups in each PA e.g: $\sum_{i \in \{C,M,I,O\}} p_i = 1$
- For specific PA: the PA's abbrieviation code will be added to the subscript of the symbols e.g: For Ang Mo Kio, $N_{AMK} = N_{AMK,C} + N_{AMK,I} + N_{AMK,M} + N_{AMK,O}$ Likewise for the ethnic proportions: $p_{AMK} = p_{AMK,C} + p_{AMK,I} + p_{AMK,M} + p_{AMK,O}$
- For the use of further extensions, we will also touch on the gender proportions of each p_i, i.e the Chinese male proportion of Ang Mo Kio is denoted as $p_{AMK,C,M}$, and the Chinese Female proportion of Ang Mo Kio is denoted as $p_{AMK,C,F}$.
- Similarly, we assume that these gender proportions sum to 1, i.e $p_{AMK,C,M} + p_{AMK,C,F} = 1$

Data

Data Collection

For the course of this project, we will be using the summary statistics provided by the Singapore Department of Statistics(DOS) Census 2015 and the 2015 General Household Survey, which are the most recent as C2020 is still under work.

We will obtain data from Singstat Table Builder Website[3], by following the steps:

1. Subject: Population → Geographic Distribution
2. Topic: General Household Survey 2015
3. Title: Table 8 Resident Population by Planning Area/Subzone, Ethnic Group & Sex

We will also obtain the files containing the all the geographic information of the Singapore's Planning Area Boundary(including spatial coordinates) via the Singapore's Government open data website[4]. The shapefiles we are using are:

- Master Plan 2019 Planning Area Boundary(No Sea)[5] : GEOJSON file format
- Master Plan 2014 Planning Area Boundary(No Sea)[6] : SHP file format

We will use the 2014 SHP file unless the 2019 GEOJSON Boundary Data is needed. This is because SHP files are more malleable in R, compared to GEOJSON files, and it will be easier to perform geospatial techniques in R.

* The processed data for the use of this project will be uploaded onto Google Drive in case of data loss. This also allows the data files to be readily available when needed.

* The codes for use in this project will also be uploaded onto Github for future references and management of code.

Data Processing/Cleaning

A quick look at the data table reveals that there is data provided for all the zones that make up each subzone and Planning Area, and the total population of each ethnic group. We will hence combine the population residing in each zone into their respective Planning Areas. As such, we will have a total of 55 Planning Areas, as given in the table in the next page, each separated into regions:

Planning Area(PA) of Singapore by Regions

Central (A-N)	Central (N-Z)	East	North	North-East	West
Bishan	Novena	Bedok	Central Water Catchment	Ang Mo Kio	Boon Lay
Bukit Merah	Orchard	Changi	Lim Chu Kang	Hougang	Bukit Batok
Bukit Timah	Outram	Changi Bay	Mandai	North-Eastern Islands	Bukit Panjang
Downtown Core	Queenstown	Pasir Ris	Sembawang	Punggol	Choa Chu Kang
Geylang	River valley	Paya Lebar	Simpang	Seletar	Clementi
Kallang	Rochor	Tampines	Sungei Kadut	Sengkang	Jurong East
Marina East	Singapore River		Woodlands	Serangoon	Jurong West
Marina South	Southern Islands		Yishun		Pioneer
Marine Parade	Straits View				Tengah
Museum	Tanglin				Tuas
Newton	Toa Payoh				Western Islands
					Western Water Catchment

Table 1: Planning Area(PA) of Singapore by Regions

We reformat the data file in MS Excel for easier processing in R:

Planning Area/Subzone	Total (population)	Chinese	Malays	Indian	Others
		Total(%)	Total(%)	Total(%)	Total(%)
SG Total	...	...	...	...	...
Ang Mo Kio - Total	...	...	...	...	...
⋮	⋮	⋮	⋮	⋮	⋮
Yishun- Total	...	...	...	...	...

Table 2: Formatted Excel Sheet

We implement some rules to further clean the ethnic data as there are some abnormalities:

1. Total Population in the first column is not always equal to the row sums of the ethnic population of the 4 races. In particular:

$$\exists PA \in SG \mid N_{PA} \neq \sum_{i \in \{C,M,I,O\}} N_{PA,i} \tag{1}$$

   Also if the numbers differ, the max difference is 10 eg:

$$\exists PA \in SG \mid \max \left\{ \mid N_{PA} - \sum_{i \in \{C,M,I,O\}} N_{PA,i} \mid \right\} = 10 \tag{2}$$

   - We will standardise the total N_{PA} to be the sum of all $N_{PA,i}$, instead of using the row sum, as fixing the total population to the row sum is simpler and saves computation if we were to scale each of the $N_{PA,i}$ to sum up to N_{PA}.

2. Some small Planning Areas have a nonzero population but there is no information on the number of the ethnic breakdown of that PA e.g:

$$\exists PA \in SG \mid \{N_{PA} \neq 0 \cap (\forall i \in \{C, M, I, O\} \mid N_{PA,i} = 0)\} \tag{3}$$

- We will thus remove these PA's from our dataset, as they provide no information at all for us to use practically.

3. Some PA's have populations that are nonzero, but have some ethnic proportions that are zero e.g:

$$\exists PA \in SG \mid \{N_{PA} \neq 0 \cap (\exists i \in \{C, M, I, O\} \mid N_{PA,i} = 0)\} \tag{4}$$

- In this case, we force the ethnic group in that PA to have 0 population, thus $N_{PA,i} = 0$. Consequently, this will not change the results as the row sum is already predetermined by the rule as stated in point 1.

4. Some minor cleaning for the ethnic data include:

- Capitalising all the names of the PA, as well as removing the string "-Total" from each PA name, to make the data easier to work with in R, due to the nature of the read SHP and GEOJSON file in R.

- Sorting the data frame in R in alphabetical order, according to PA.

- Removed the first column relating to the total $N_{PA,i}$ as we are working with the ethnic proportions, instead of ethnic counts.

- Dropping the Z^{th} Coordinates of the 2019 GEOJSON file for better usage.

5. We also subset the ethnic data by Sex:

- Subset 1: Male Proportions of each Ethnic Group by PA

- Subset 2: Female Proportions of each Ethnic Group by PA

- Minor Cleaning for these subsets include:

  (a) Some PAs have missing data for the number of Males and Females for some Ethnic Groups, we define these proportions to be 0.

  (b) The counts for these proportion data is kept separately as another data frame for future use.

Data Exploration

Ethnic Data

We will now take a look at the cleaned up data. The new data frame consists of 41 Planning Areas, as the other 14 Planning Areas have been removed due to the imposed data cleaning constraints. The other 41 rows are kept and represent the ethnic proportions $p_{PA,i}$:

Planning Area(PA)	Chinese(%)	Malays(%)	Indian(%)	Others(%)
TOTAL	...	...	...	...
ANG MO KIO	...	...	...	...
⋮	⋮	⋮	⋮	⋮
YISHUN	...	...	...	...

Table 3: R data frame: Ethnic Proportions by Planning Area

From the new reformatted data, the remaining PA's are shown below,

Remaining PA's

Central (A-N)	**Central (O-Z)**	**East**	**North**	**North-East**	**West**
Bishan	Orchard	Bedok	Mandai	Ang Mo Kio	Bukit Batok
Bukit Merah	Outram	Changi	Sembawang	Hougang	Bukit Panjang
Bukit Timah	Queenstown	Pasir Ris	Sungei Kadut	Punggol	Choa Chu Kang
Downtown Core	River Valley	Tampines	Woodlands	Seletar	Clementi
Geylang	Rochor		Yishun	Sengkang	Jurong East
Kallang	Singapore River			Serangoon	Jurong West
Marine Parade	Southern Islands				Western Water Catchment
Museum	Tanglin				
Newton	Toa Payoh				
Novena					

Table 4: Remaining PA's

The removed PA's are:

- Central: Marina East, Marina South, Straits view
- East: Changi Bay, Paya Lebar
- North: Central Water Catchment, Lim Chu Kang, Simpang
- North-East: North-Eastern Islands
- West: Boon Lay, Pioneer, Tengah, Tuas, Western Islands

A simple summary of the ethnic proportions of the remaining 41 PA's is shown in the table below:

	Chinese	Malay	Indian	Others
Min	0.4743	0	0.04901	0.02218
1st Qu	0.7009	0.01153	0.08182	0.02744
Median	0.7442	0.07939	0.09245	0.03408
Mean	0.7371	0.08995	0.09980	0.07315
3rd Qu	0.8017	0.13909	0.10800	0.07373
Max	0.8614	0.28458	0.20274	0.33784
Baseline Population(Total)	0.7431	0.13348	0.09095	0.03249

Table 5: Summary of Ethnic Proportions in the 41PA's

From Table 5, we can see that the Chinese and Malay Populations have relatively similar IQR values ≈ 0.1, whereas the IQR for the Indian Population and that of the Others are significantly lower (< 0.05), around half the IQR of the Chinese and Malay Population.

A boxplot of the ethnic proportions is shown below:

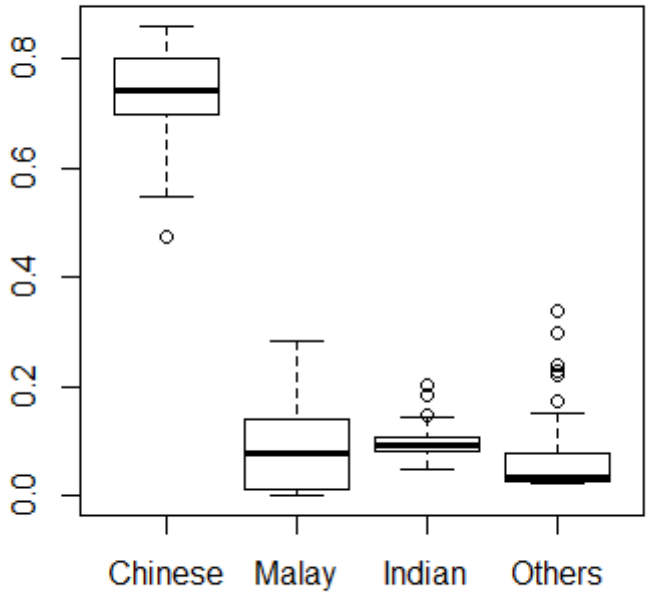

Figure 1: Boxplot of Ethnic Proportions

From Figure 1, we see that the the Indian population proportion distribution is fairly uniformly distributed, the the exception of a few outliers that lie outside of $(\mu \pm 1.5IQR)$.

For the rest of the ethnic groups, there is some skewness of the proportion distribution. The ethnic proportion distribution of the Chinese is very left skewed, whereas the ethnic

proportion distribution of that of Others and Malay are more right skewed.

There are more outlier values lying outside ($\mu \pm 1.5IQR$) for the Other ethnic groups across all PA's. Whereas there are fewer outliers for the other ethnic groups.

Also, we can see that the spread for Chinese and Malay Ethnic Proportions tend to be larger compared to that of Indian and Others. This information will be important for later quantile based computations and splits.

Shape Files

The data frame of the 2019 GEOJSON file read in R contains a list of 5 different items:

1. 'data': Consisting of 12 variables:

    - Object ID, PLN_AREA_N, PLN_AREA_C, CA_IND, REGION_N, REGION_C, INC_CRC, FMEL_UPD_D, X_ADDR, Y_ADDR, SHAPE_Leng, SHAPE_Area

2. 'polygons': Consisting a list of 55 Polygon Shape class, with some variables indicating the coordinates for each PA.

3. 'plotOrder': Indicative of the order to plot which PA's first in R.

4. 'bbox': Data containing the maximum and minimum coordinate ranges.

5. 'proj4string': Data of the class "CRS" (Coordinate Reference System Arguments)

The data frame of the 2014 SHP file read in R contains a data frame consisting of the following:

1. 'data': Consisting of 13 variables:

    - Object ID, PLN_AREA_N, PLN_AREA_C, CA_IND, REGION_N, REGION_C, INC_CRC, FMEL_UPD_D, X_ADDR, Y_ADDR, SHAPE_Leng, SHAPE_Area, Geometry

The largest difference between these 2 shapefiles is that the 2014 SHP file contains less information, as 'polygons' in the 2019 GEOJSON file is now in the 'geometry' column of the SHP file. The rest of the dropped variables are not relevant to the scope of this project.

A simple plot in R is generated below using the 2019 GEOJSON shapefile, depicting Singapore, with a colour palette of 16 distinct colours being used to shade each indicative polygon of all 55 PA.

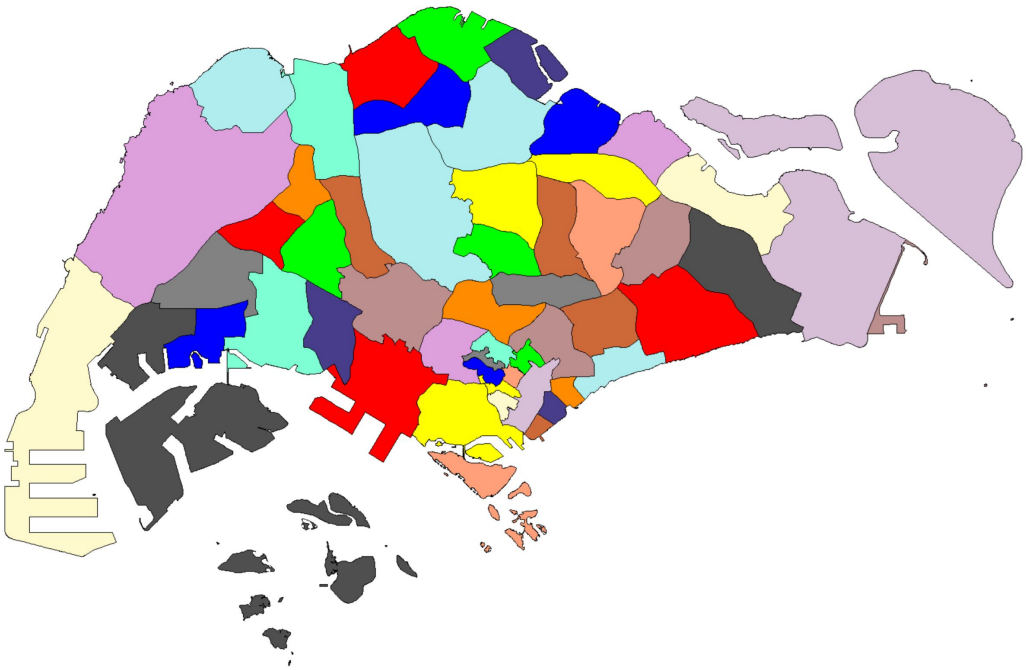

Figure 2: SG Map in Singapore with 16 distinct Shades For all 55 PA (GEOJSON)

We do the same for the 2014 SHP file:

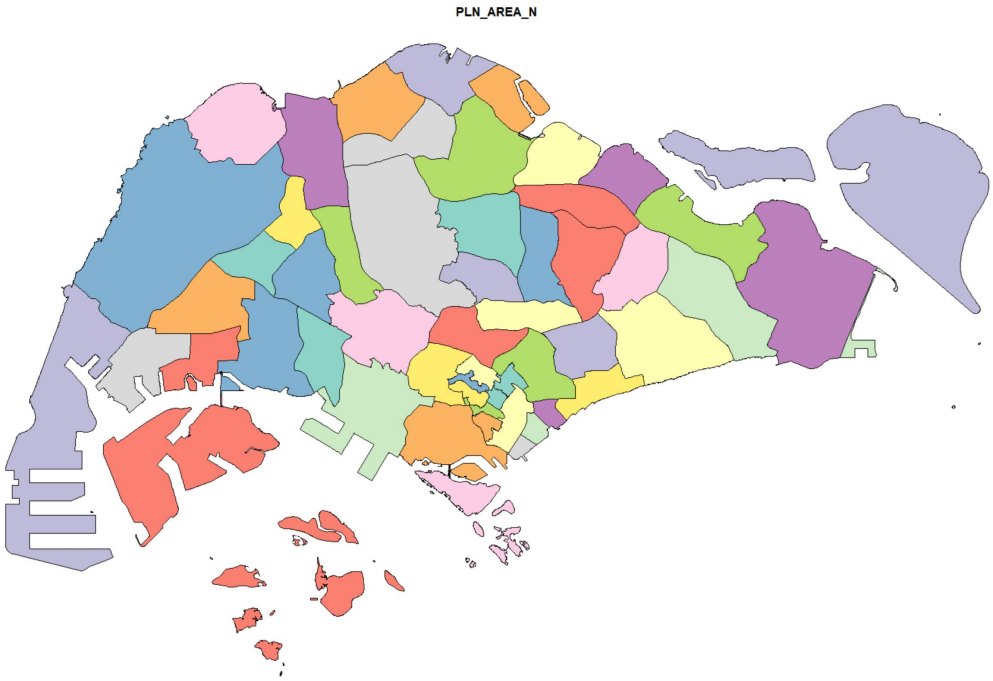

Figure 3: SG Map in Singapore with 16 distinct Shades For all 55 PA (SHP)

A quick look at both Figure 2 and Figure 3 shows that there is no stark difference in using the 2 shapefiles to plot the map of Singapore.

Some observations:

- The outlines for Figure 2 seems to be less sharp than that of Figure 3
- Some PA's that take up a very small area, which can be seen in the Southern part of Singapore.
- Some places seem to consist of more smaller islands, such as Western Islands shaded in salmon colour at the bottom of Figure 3, this in turn, might hinder our ability to understand the plot.

This brings forth some questions:

1. If each PA is so small, how to we label them to make them visible for the layman reader?
2. How do we represent the 4 Ethnic Groups of Singapore on a 2 Dimensional Map?
3. If we are able to do so, how do we do it in the way in which the least information is lost?
4. Can we combine statistical and graphical techniques to produce a plot that is simple enough for any reader to understand?
5. Are we able to extend this idea to jointly distributed variables? (In this case whether the proportion of ethnic groups in each PA is jointly distributed with the proportion of Males/Females in each PA)
6. How do we represent missing information on the map?
7. Can we find the most simple yet novel method that combines all the answers to the above questions?

To answer these questions, we first look at the statistical techniques that we can perform on the ethnic datasets.

Statistical Techniques

K-Means Clustering

K-Means Clustering is a popular choice of clustering.

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional vector, K-Means clustering partitions the n observations into $k \leq n$ sets. This is done by minimizing the within-cluster sum of squares, in other words, the objective for minimisation is:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} \mid S_i \mid Var\, S_i \tag{5}$$

In our case, referring back to Table 3, we have a 4-dimensional vector for each of the 55 observations. However, due to the sum constraint mentioned in the Project Scope, we will consider the first 3 dimensions instead.

Also, because we have 14 PA's with missing data, it will not be wise to use those values, as those values are set to 0, and we would have 14 PAs in 1 cluster, which is not wise as they would waste up one of the k clusters.

We first do a scree plot to decide the value of k to use:

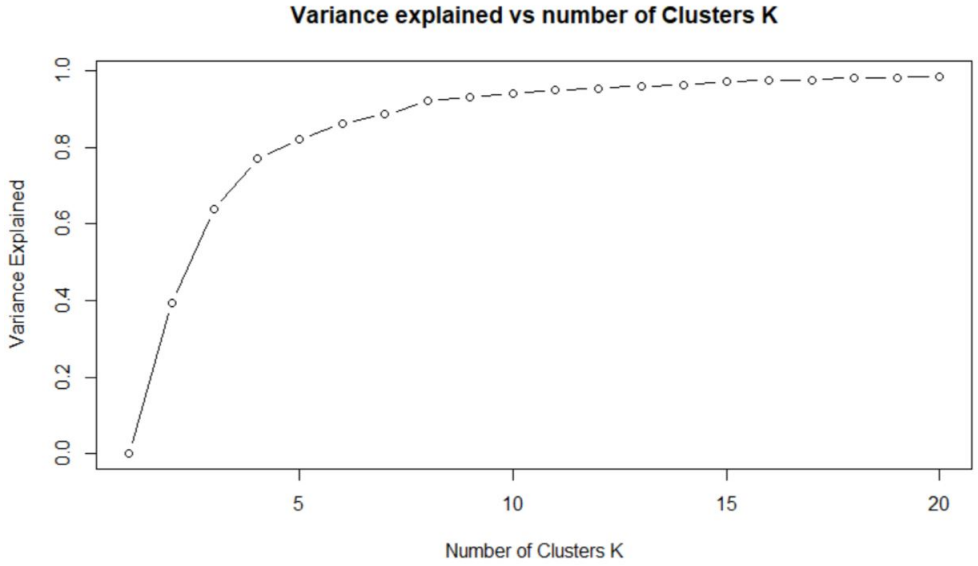

Figure 4: Variance Explained by K-Means vs Number of Clusters K

We thus choose the number of elbows to be 5, as increasing k will not significantly reduce the variance of K-Means.

Fitting the K-Means model on the ethnic proportions of Singapore results in the following plot:

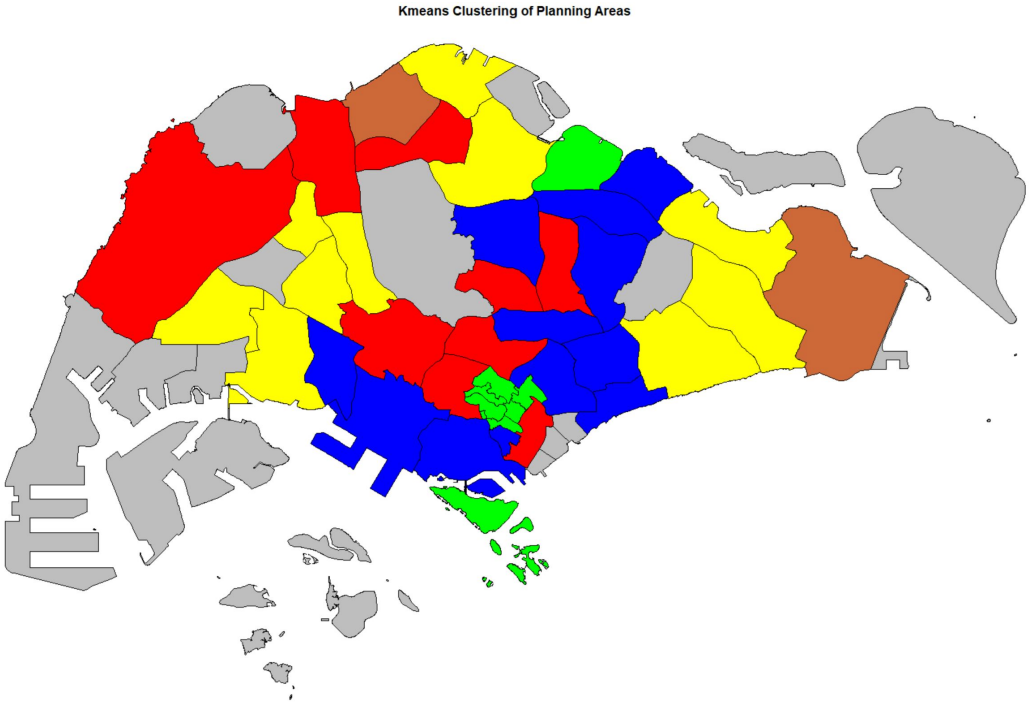

Figure 5: Plot of SG by Ethnic Proportion Clusters (K-Means)

From Figure 5 above, we can see that K-Means is able to break down the PA's into clusters, with the omitted 14 PA's shaded as grey above.

However, by performing this technique, we only answer Questions 2,4 and 6 from Page 13. This is not good enough.

By performing K-Means on the data, we have essentially lost a significant amount of information i.e

- We only know that PA's in the same cluster are similar.
- However, we do not know how different are the PA's from each other.
- Also, we lose information on the ethnic proportions in each PA, as now the information is condensed into a distance matrix, which is not optimal.

Also, these PA's are unlabelled, thus no one would know which PA's are which, unless they are well versed in the geography of Singapore. Any attempts to label these PA's will not be useful as some PA's are too small, and it would be messy and each PA would then be unidentifiable, especially in the Southern part of Singapore.

K-Medians Clustering

K-Medians is also a commonly used clustering method, it works in the same way as K-Means, however this method is slight more robust to outliers as it minimises over the ℓ_1 norm instead of the ℓ_2 norm in K-Means.

Similarly, we choose the number of $k = 5$, and we obtain the following plot:

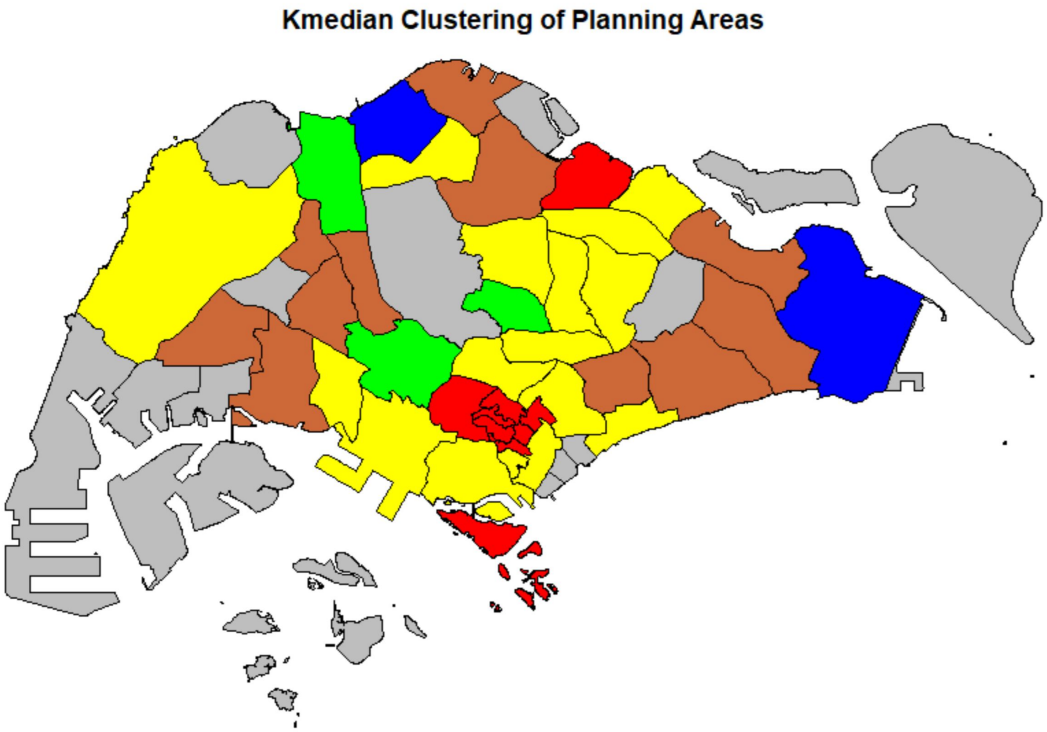

Figure 6: Plot of SG by Ethnic Proportion Clusters (K-Medians)

Similarly, the grey shaded areas are the 14 PA's that we do not have information on.

We can also conclude that the limitations of this method is same as the ones as described in Page 14.

One strong advantage of using K-Medians over K-Means, is that in K-Means, if the observations tend to be very similar in number, K-Means Clustering performs poorer than K-Medians, as the random initialisation of K-Means Clustering might results in significantly different clusters when run many times. K-Medians clustering in this case is more stable compared to K-Means.

Hierarchical Clustering

Hierarchical Clustering is a cluster analysis method that seeks to build a hierarchy of clusters. This is usually done in 2 different approaches:

1. Agglomerative approach: Bottom Up approach - each observation starts in their own cluster, then pairs of clusters are merged as they move up the hierarchy.

2. Divisive Approach: Top Down approach - all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Also, in order to decide which clusters to merge/split, a measure of dissimilarity between sets of observations is required. The measures of dissimilarity are usually functions of the pairwise distances between the sets. These functions are known as linkages, and the most common linkages are as summarised below:

- Complete Linkage - The maximum distance between sets of observations i.e:

$$max\left\{d(a,b) : a \in A, b \in B\right\} \tag{6}$$

- Single Linkage - The minimum distance between sets of observations i.e:

$$min\left\{d(a,b) : a \in A, b \in B\right\} \tag{7}$$

- Average Linkage(unweighted) - The unweighted average distance between sets of observations i.e:

$$\frac{1}{|A|\,|B|}\sum_{a\in A}\sum_{b\in B} d(a,b) \tag{8}$$

- Centroid Linkage - The distance between centroids of clusters i.e:

$$\|c_s - c_t\| \text{ where } c_s \text{ and } c_t \text{ are the centroids of clusters } s \text{ and } t \text{ respectively} \tag{9}$$

- Median Linkage - Related to Centroid Linkage, but is weighted

For the purpose of this project, we will work with the Agglomerative approach, and we plot the maps for the above-mentioned approaches. We choose the number of clusters = 5, as there are the PA's are split into 5 regions, and we also want to check partially if the ethnic proportions are related to the 5 regions. Also, $k = 5$ is set at a fixed number because we can cut the dendrogram at any height, and it will be inconsistent if we were to .

Complete Linkage

The hierarchical model using complete linkage is as shown below:

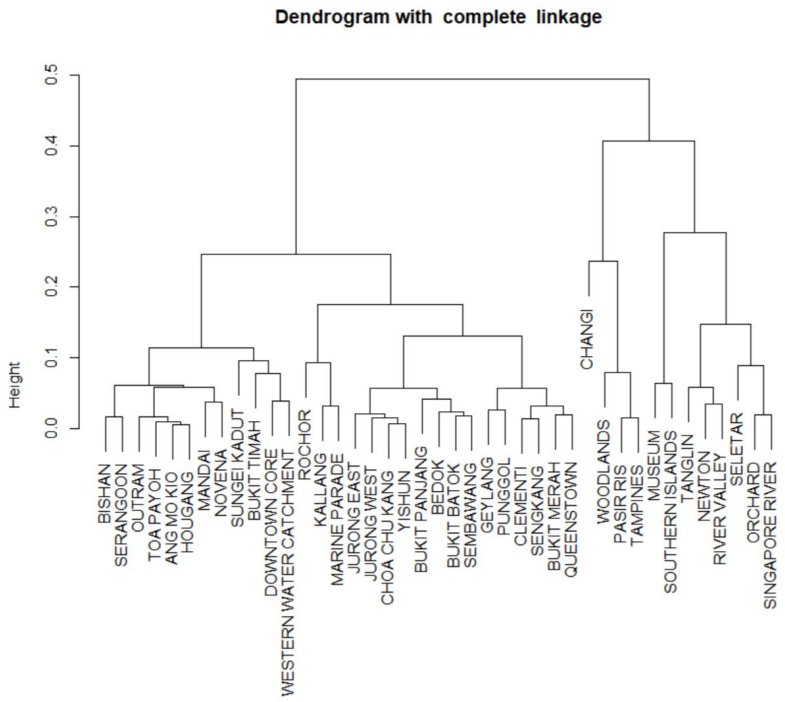

Figure 7: Dendrogram of Complete Linkage

The following plot of SG map given $k = 5$ clusters is shown below:

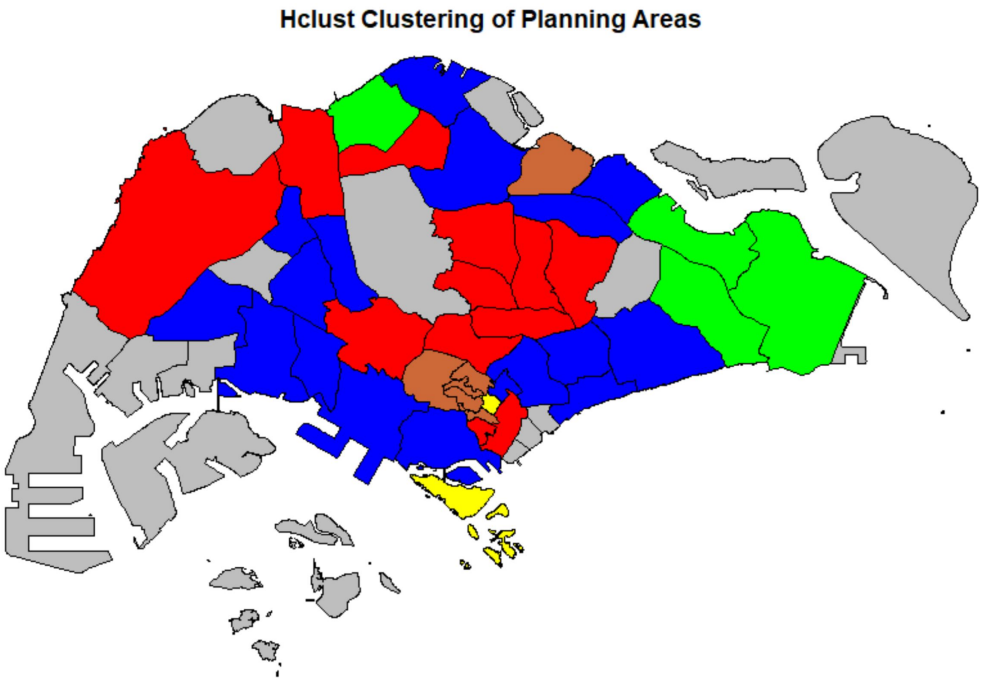

Figure 8: Plot of SG by Ethnic Proportion Clusters (Hclust - Complete Linkage)

Single Linkage

The hierarchical model using complete linkage is as shown below:

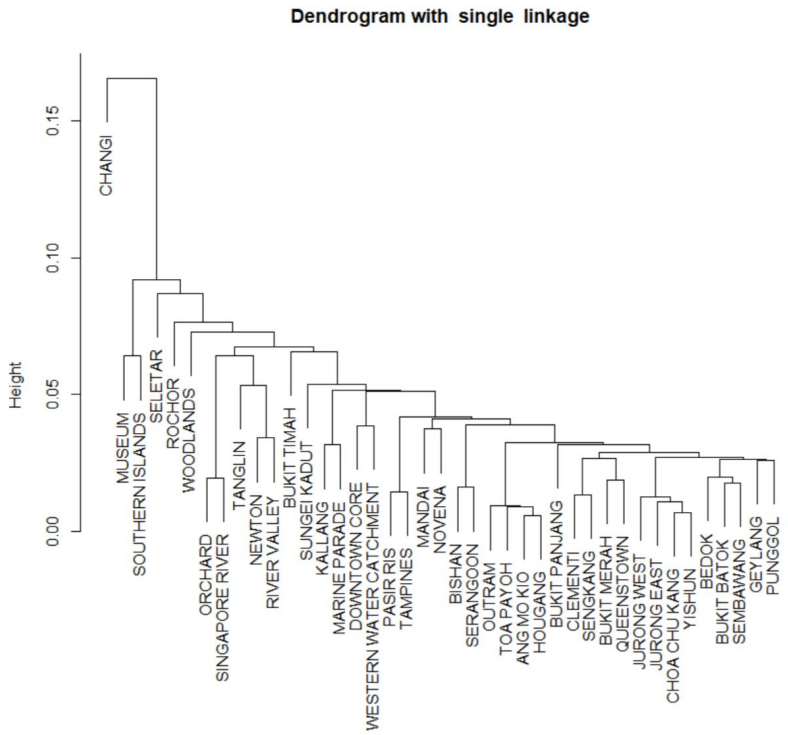

Figure 9: Dendrogram of Single Linkage

The following plot of SG map given $k = 5$ clusters is shown below:

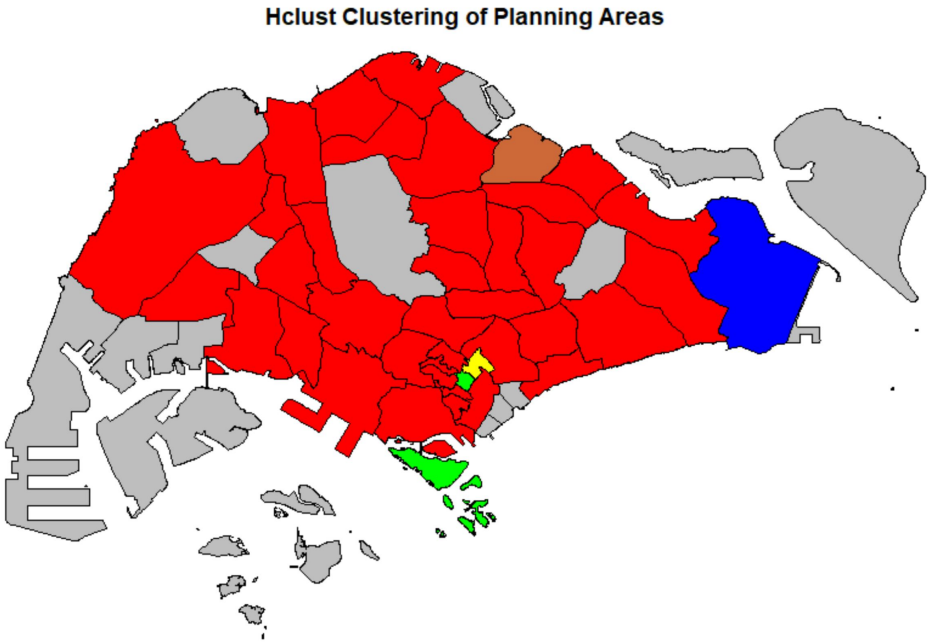

Figure 10: Plot of SG by Ethnic Proportion Clusters (Hclust - Single Linkage)

Average Linkage

The hierarchical model using average linkage is as shown below:

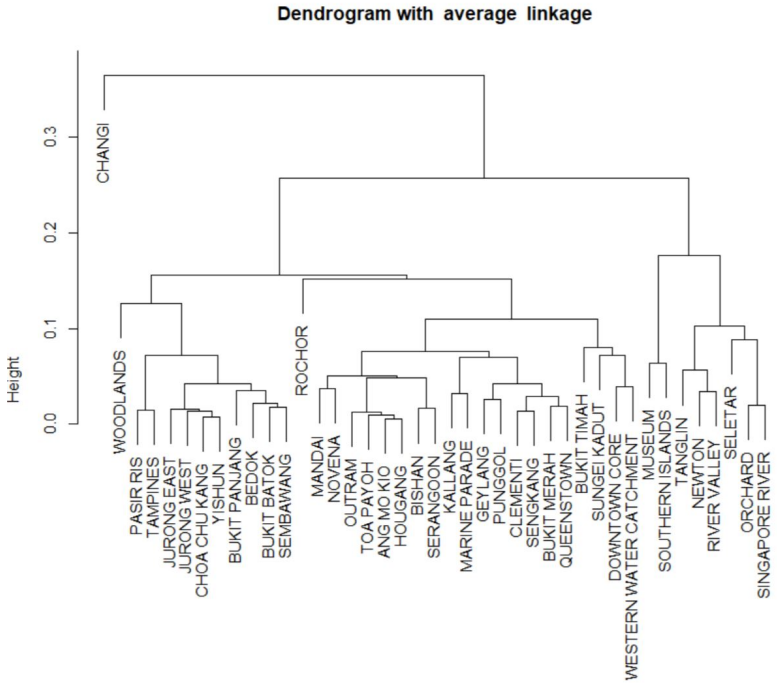

Figure 11: Dendrogram of Average Linkage

The following plot of SG map given $k = 5$ clusters is shown below:

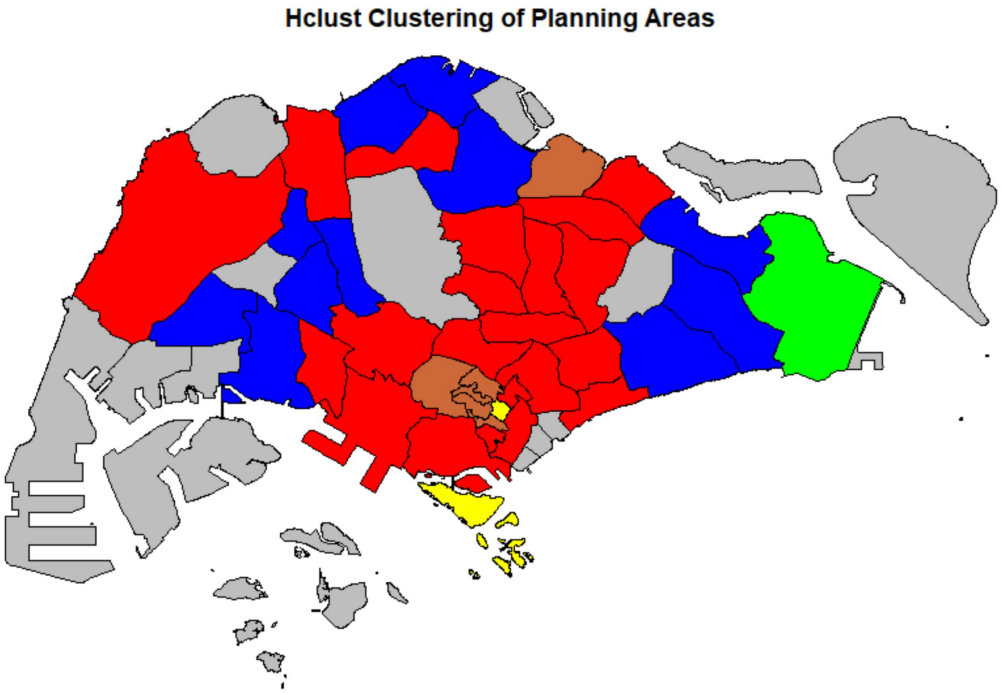

Figure 12: Plot of SG by Ethnic Proportion Clusters (Hclust - Average Linkage)

Median Linkage

The hierarchical model using Median linkage is as shown below:

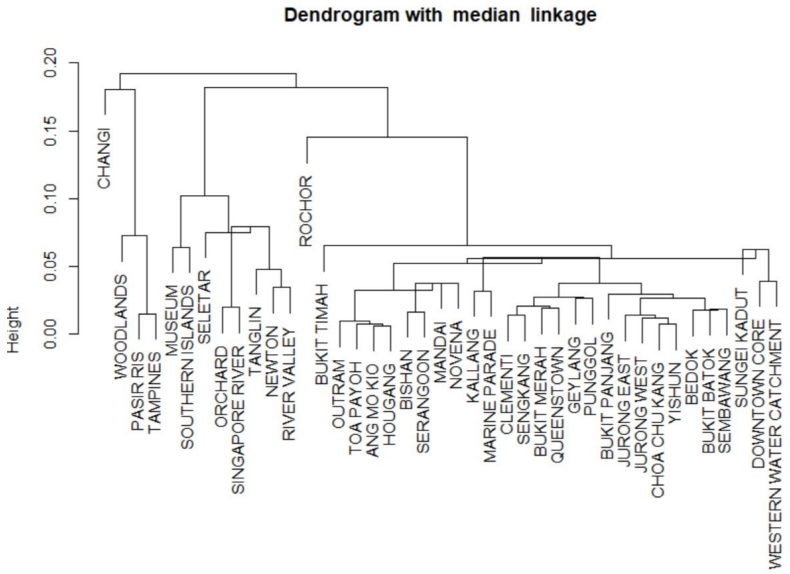

Figure 13: Dendrogram of Median Linkage

The following plot of SG map given $k = 5$ clusters is shown below:

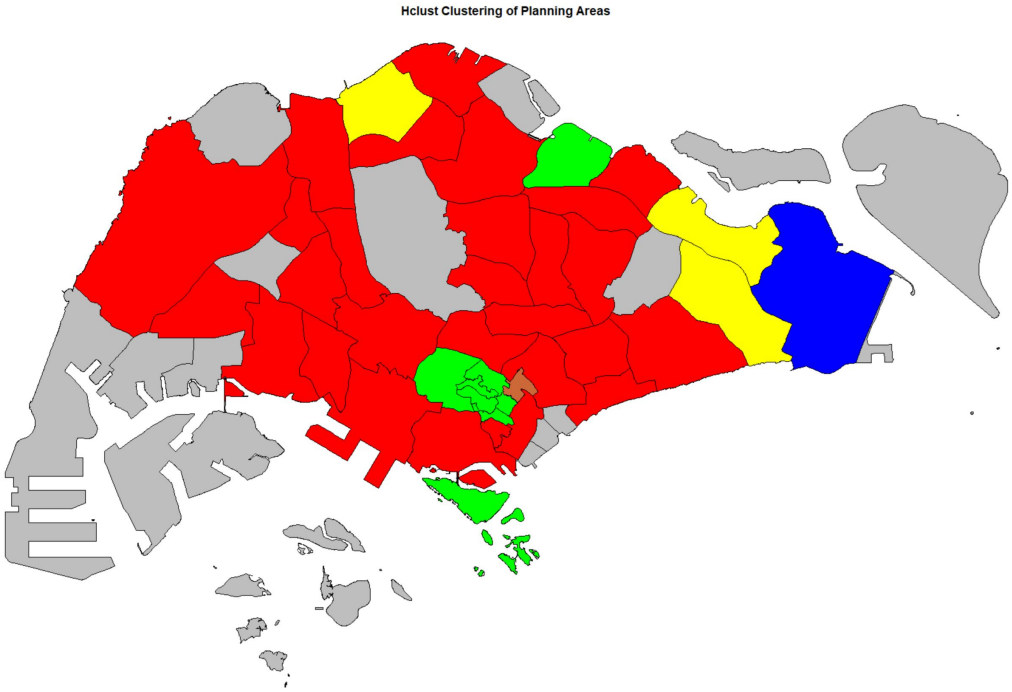

Figure 14: Plot of SG by Ethnic Proportion Clusters (Hclust - Median Linkage)

Centroid Linkage

The hierarchical model using Centroid linkage is as shown below:

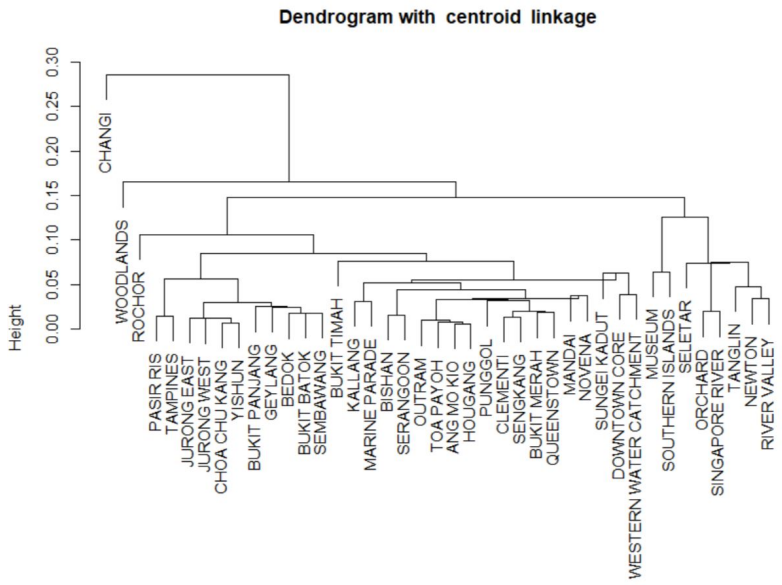

Figure 15: Dendrogram of Centroid Linkage

The following plot of SG map given $k = 5$ clusters is shown below:

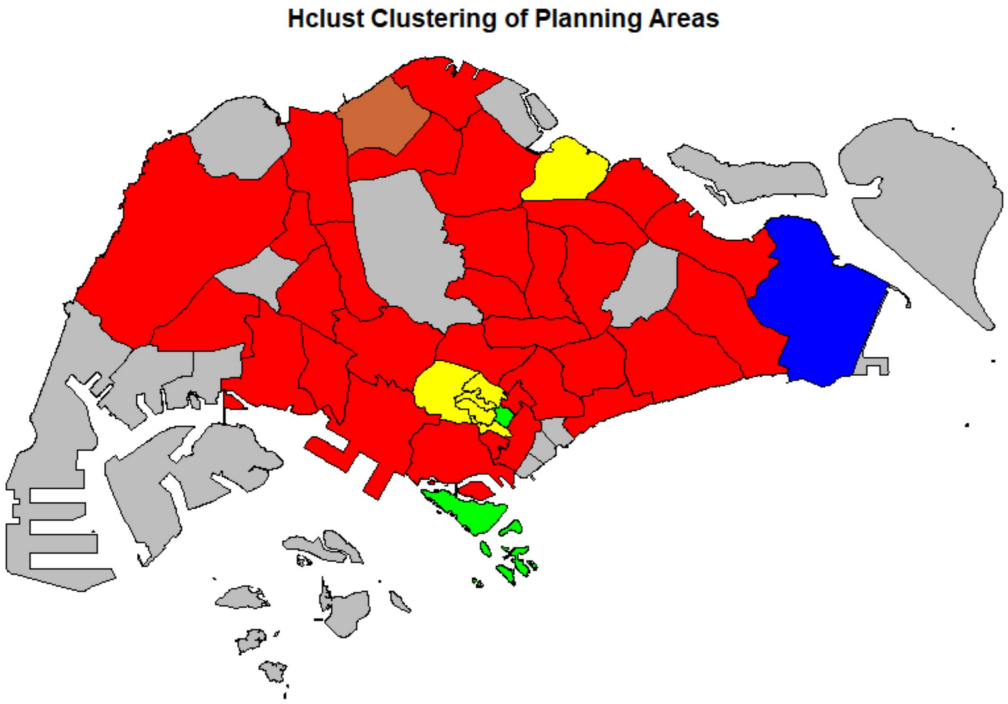

Figure 16: Plot of SG by Ethnic Proportion Clusters (Hclust - Centroid Linkage)

Summary of Hierarchical Clustering

In summary, by looking at the dendrograms and the repective plots of Singapore, we notice that Single Linkage performs the worst (Refer to Figure 9 and Figure 10). This is because by using Single Linkage, clusters are merged with another quickly, forming large clusters even at a large height in the dendrogram. This causes the Singapore map to have a relatively uniform colour, as cutting the dendrogram into 5 clusters results in a cluster of 37 PA's and 4 clusters of 1 PA. Median Linkage and Centroid Linkage shows the same trend as Single Linkage, but is considerably better as they do not form large clusters so quickly. (Refer to Figure 13, Figure 14, Figure 15 , Figure 16)

Also, Complete Linkage and Average Linkage performs better, as they do not results in long chains of clusters, or large clusters, and these clusters are evenly spread out.

Comparing Figure 8 and Figure 12 we can see that several places are quite similar, as suggested by the clusters highlighted. There seems to be one cluster generally in the Central Region of Singapore, and cluster linking the ring outside the Central Region, as well as one more cluster generally in the Changi Area. This suggests that the Ethnic Proportion makeup of these PA's are generally very similar in these clusters.

Statistical Techniques - Conclusion

In conclusion, whilst these techniques prove to be statistically useful in clustering the areas, we note that there are certain drawbacks.

Since this hierarchical clustering and K-Medians are just another few clustering methods, the method of plotting these clusters on the Singapore map is the same. Referring back to Page 13, we leave some questions unanswered.

- We have not yet found a way to make the smaller PAs visible and labelled.
- Also, we have not been able to accurately represent the 4 Ethnic Populations onto the map of Singapore without losing information.
- We also have not extended this to jointly distributed variables.

.

This concludes the section on Statistical Techniques we can use, and we move on to possible Graphical techniques we can employ, to further refine our current methods.

Graphical Techniques

Possible Types of Maps

Cartograms

Cartograms are a type of map which is distorted based on some information variable or characteristic possessed by the certain plotted area.

For this we employ the use of an online cartogram widget from a paper (Gastner, V. , 2018)[7], which allows us to have a look at the Map of Singapore plotted before and after the Cartogram method is applied:

Before Cartogramming:

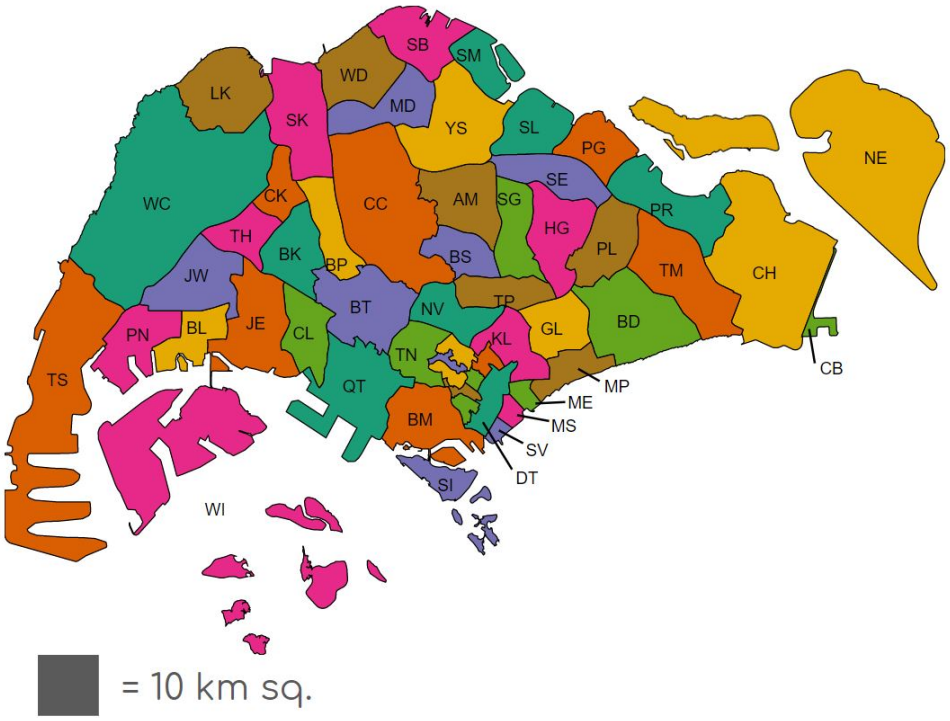

Figure 17: Singapore Map before Cartogramming

We can see that before the Cartogramming, the map of Singapore is exactly the same as the ones generated above, with labels.

However, we can see that some of the labels lie in awkward places (e.g 'WI' lies in the middle of the sea even though it is in the centroid of the defined polygon for the Western Islands, and 'BP' lies on the boundary of the polygon which does not look too nice.)

After Cartogramming:

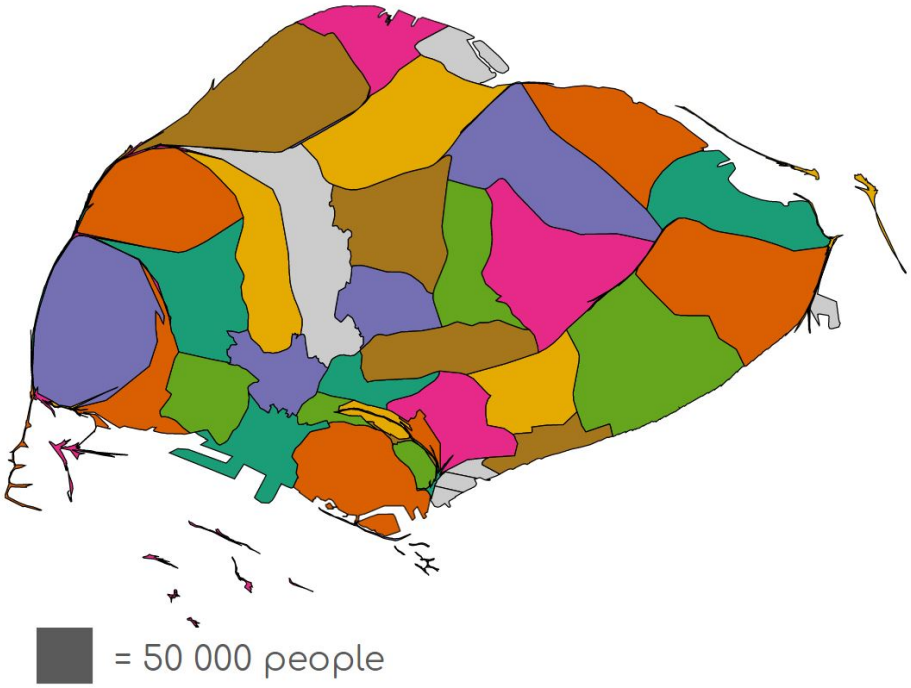

Figure 18: Singapore Map after cartogramming

After Cartogramming, we can see that certain areas are blown up to the relative size of the PA's population. This causes some distortion in areas with significantly larger population than the other areas around it. Because of this, the smaller PA's have become significantly smaller, and cannot be seen at all. If we use the inverse weights, the opposite will happen, where the significantly smaller PA's will be too large.

Drawbacks of this method include the inability to comprehend the plot easily, and we lose readability in the small PA's. Labelling of PA's become significantly harder, and we are still not able to accurately project the Ethnic Proportion of each PA without loss of information.

Cartograms are also harder for the general audience to comprehend, and required a legend and more visual aids. Thus, we will not consider cartograms as a graphical method.

Chloropleth/Heat Maps

Another type of map commonly used is the Chloropleth Map, an example of the Singapore Chloropleth map is shown below:

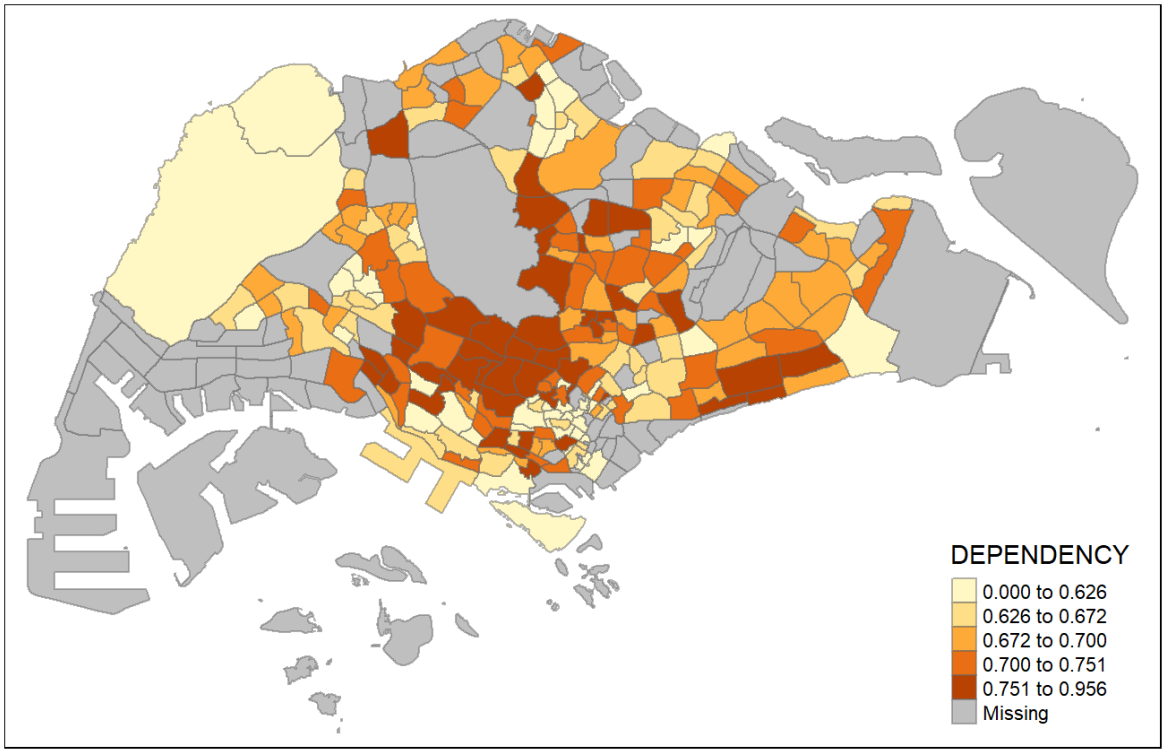

Figure 19: Example of Singapore Chloropleth Map

We can see that in general, this map does not differ too much to the original maps plotted by the clustering methods in the previous section, the only difference is that whilst those use clustering approaches, the colour used here is for variable with continuous scale.

However, this might not be so useful in our context, as we have categorical variables, with 4 factors - Chinese, Malay, Indian, Others.

Furthermore, addition of colours make it less intuitive for the reader unless a legend is specified, also like with the previous issues of the above methods, information is lost, and the areas are still too small to be labelled and interpreted easily.

Therefore, we are looking for a method which can standardise the size of each PA, and not be able to lose information in the method that is used for plotting.

Radar Maps

We can further draw some inspiration from radar maps, from a paper(A.M Guerry, 2007)[8] published on Multivariate Spatial Analysis.

These radar maps aim to improve the graphical techniques for multivariate spatial analysis, and one such example from the paper is shown below:

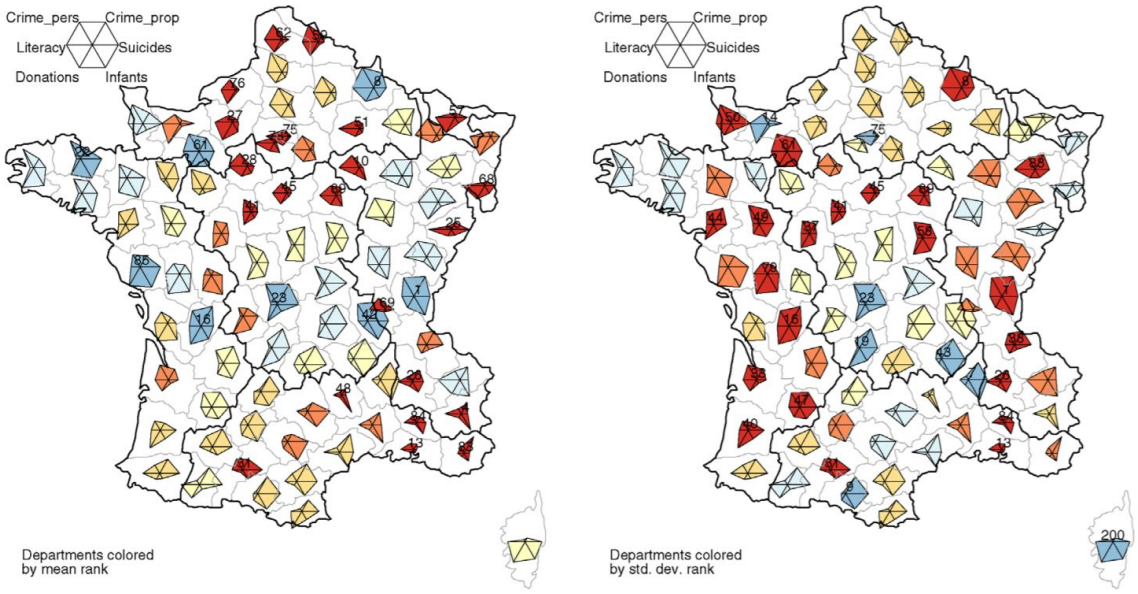

Figure 20: Radar Map (A.M Guerry, 2007)

From Figure 20, we can see that more variables can be represented in each area, which is exactly what we are looking for! By combining the 'radar' essence of the radar map, we can add the Ethnic features into a map of our own. However, instead of the irregular hexagons as seen in Figure 20, we will use arrows to represent the Chinese, Malay, Indian, and Others proportion in our map.

In Figure 20, as the distance from the centroid of the irregular hexagons increases. the corresponding variable value also increases, this can be applied to our case of the Ethnic Proportions.

However, some areas are still not magnified enough, we still need to find a way to standardize the size of each PA.

Hexmaps/Tile Grid Maps

One of the ways to standardize the size of each area, is to standardize the shape of each PA. One way of doing this is to find polygons which are able to tessellate themselves.

There are only 3 regular polygons that tessellate: triangles, squares and hexagons.

We thus look at 2 main shapes: hexagons and squares, which have enough area to add labels in their geometric centroids, and also enough space to add lines/arrows within them.

A less common type of mapping is called the hexmap, an online example can be found here[9]. An illustration of a hexmap is shown below:

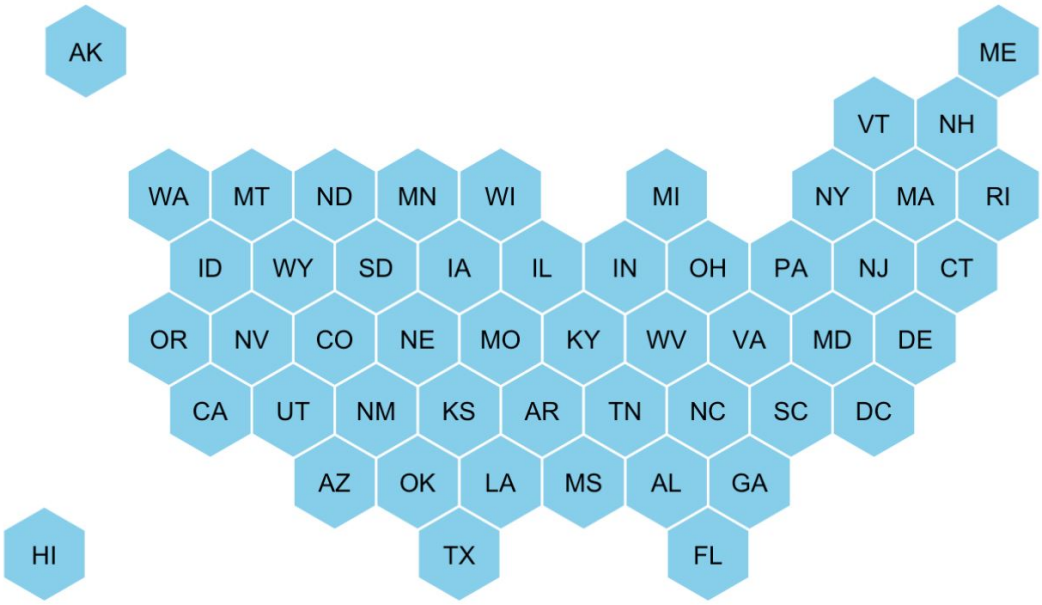

Figure 21: Example of a Hexmap (States of the USA)

Figure 21 above shows the plotted hexmap of the stats of the USA. From this map above, we can see that the size of each area is standardized, and labels fit perfectly into the geometric centroid of each hexagon. This comes at the expense of losing the original shapes of the areas, but still preserves the map shape as much as possible.

Another type of map that uses squares instead of hexagons, is known as the Tile Grid Maps, albeit not commonly used, as it distorts the original shape of the map more than a hexmap does. An example of a Tile Grid Map is found here[10] and is shown in the next page.

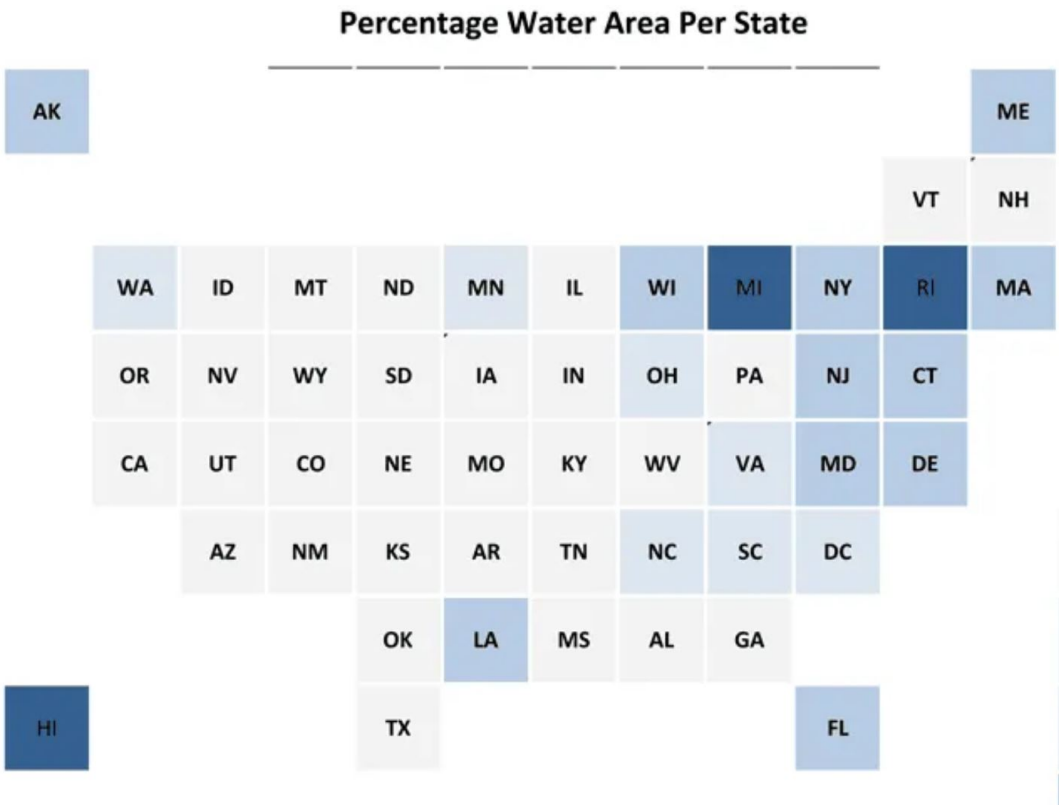

Figure 22: Example of a Tile Grid Map (States of the USA)

We can see that the Tile Grid Map shown above in Figure 22 also manages to preserve the shape of USA. This also allows for annotations and colours inside of each area.

Thus, these types of maps meet some criteria that we are looking for:

- Each area has a standardized size and is large enough to support colours, labels and extra annotations.
- The shape of the Hexmaps and Tile Grid Maps are also relatively preserved, thus we can still make out the general regions of the map.
- As these maps support labels, this implies that these maps can support multivariable extensions of Ethnic Distributions.
- This all in turn also leads to lesser loss of information which answers most of the unanswered questions found in Page 13.

Since we are interested in representing the Ethnic Proportions of Singapore on a physical map, it is perfectly fine to lose some accuracy in the shape of the maps whilst trying our best to preserve the original shape, as this issue does not lie in the original project scope.

Graphical Techniques - Conclusion

Out of the possible surveyed graphical techniques, we find that a combination of 2 of them care able to answer the questions found in Page 13.

We can use the statistical methods(including plotting) to:

1. Represent the 4 Ethnic Groups of Singapore on a 2D map.

2. Represent missing information which is easy, as each PA just has to have either no fill, or a shade colour representing missing data.

The graphical methods enable us to:

1. Transform each PA to the same size, enabling better readability for the layman,

2. Can reduce drastically the amount of information lost in plotting the Ethnic Distributions of Singapore,

3. Able to extend this idea to jointly distributed variables.

With a combination of Radar maps, as well as Hexmaps/Tile Grid Maps, as well as some of the above mentioned statistical techniques, we propose a new type of map that is reader intuitive, and easy to understand, answering most, if not all of the questions in Page 13.

Methodology

Shapefile Processing

To transform the shapefiles into the hexmap that we want, we need a list of packages in R - `sf , geojsonio, geogrid , dplyr`. The library 'geogrid' allows us to convert the shapefiles into hexagonal or tile grids, by simulation.

This simulation is done by the Hungarian matching algorithm(Kuhn, 1955)[11], of the geometric centroids of each PA. However, different initial seeds and different learning rates, thus it will be best to first find one good pair of values to give us the best simulated grid as possible, in our case, trying values between $seed \in [1, 200]$, $learning \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. The best pair of values obtained were $seed = 14, learn = 0.5$, we will use this initialisation to plot the simulated map for all future references.

The plotted Hexmap with unique abbreviated labels for each PA is shown below:

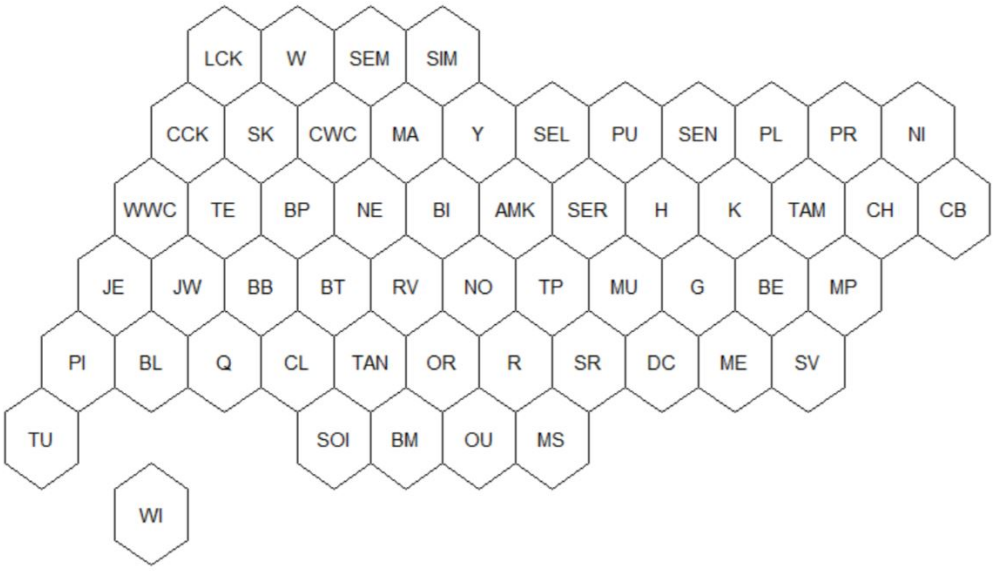

Figure 23: Plotted Raw Hexmap of the 55 PA's in Singapore

From Figure 23 above, we can see that the hexagons are neatly plotted, and each PA is identifiable due to their uniqueness.

Cross referencing back to the original plot of Singapore (Figure 2), we can see that the overall shape of Singapore is preserved. However, due to Southern Singapore having small PA's, the new regions around these small PA's are pushed out, possibly distorting the location and label of PA's

We also try plotting the simulated Tile Grid Map of Singapore, as shown below:

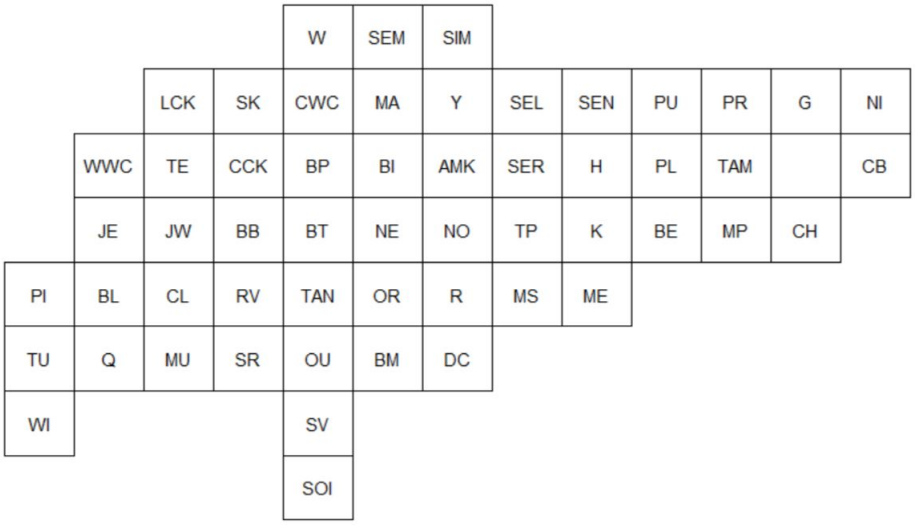

Figure 24: Plotted Raw Tile Grid Map of the 55 PA's in Singapore

Cross referencing back to the original plot of Singapore (Figure 2), as well as the new simulated Figure 23, we can see that the overall shape of Singapore is preserved and are relatively similar. However, there seems to be a small unlabelled gap in the Changi Bay area (labelled 'CB') as the tessellation algorithm forces the grids to go around the empty square instead of merging them together.

The plotted area of the squares in Figure 24 are also smaller in size compared to the hexagons, thus for these reasons, we will opt to use the hexagonal grid for further plotting.

Annotations of Plot

We then add the 4 arrows each pointing in $\{0°, 90°, 180°, 270°\}$ to signify the Chinese, Malay, Indian, and Others respectively, with $0°$ representing Geographical North, and moving in a clockwise fashion.

The lengths of these arrows, are determined by a quantile cut, into 4 not necessarily equal sized quantiles, meaning for each Ethnic Proportion vector, there will be 4 levels specifying different lengths of the arrows.

We also use 4 different colours to specify different Ethnic Groups, and add a legend to the plot which is easy for the reader to understand.

The new plot is as shown in the next page.

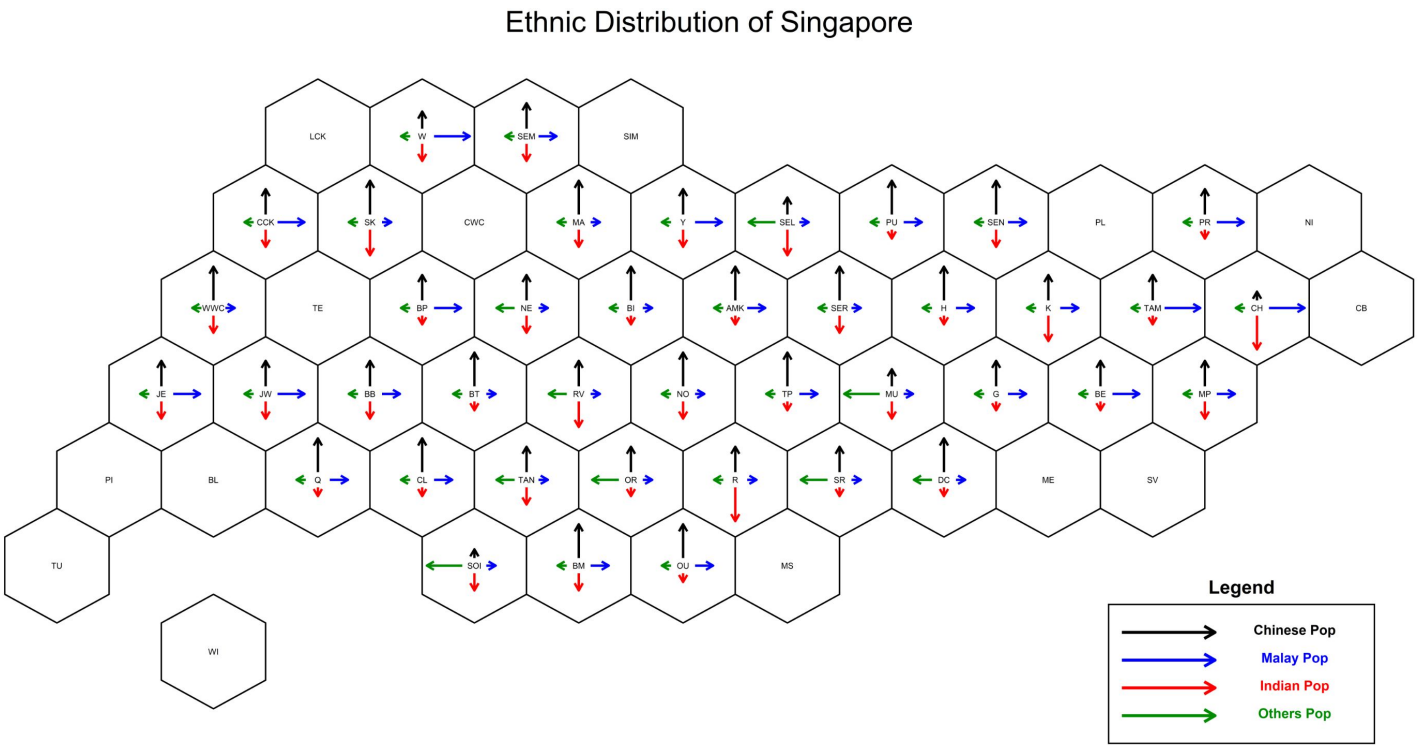

Figure 25: Plot of Ethnic Distribution in Singapore

From Figure 25 above, it is easy to tell which Ethnic Proportions are belonging to which Ethnic Group. In closer comparison, we can tell how close are the PA's are by pairwise comparison, i.e In the Northern Part of Singapore, we can tell that Woodlands (denoted by 'W') has a lower perentage of Chinese compared to Sembawang (denoted by 'SEM'), also, Woodlands also has a higher proportion of Malays, compared to Sembawang.

Also, Figure 25 has a legend, and plot title equipped for easier readability and customisation. Furthermore, PA's with missing data are also preserved in their places, with no arrows attached, so it is easy for us to tell which PA's have missing data.

However, this is not enough as we have not included the statistical methods for this plot, we will highlight the name labels of the PA's if they are determined to be in the same cluster.

We will allow all the statistical clustering methods mentioned in the previous section to be used, and we can specify the which method we want to use, as well as if we want to represent them on the map. If we do not want the clusters to be represented, the labels will remain as black, however, if we want the clusters to be represented, each label will be labelled with a colour that belongs to a certain cluster.

The resulting plot is shown in the next page.

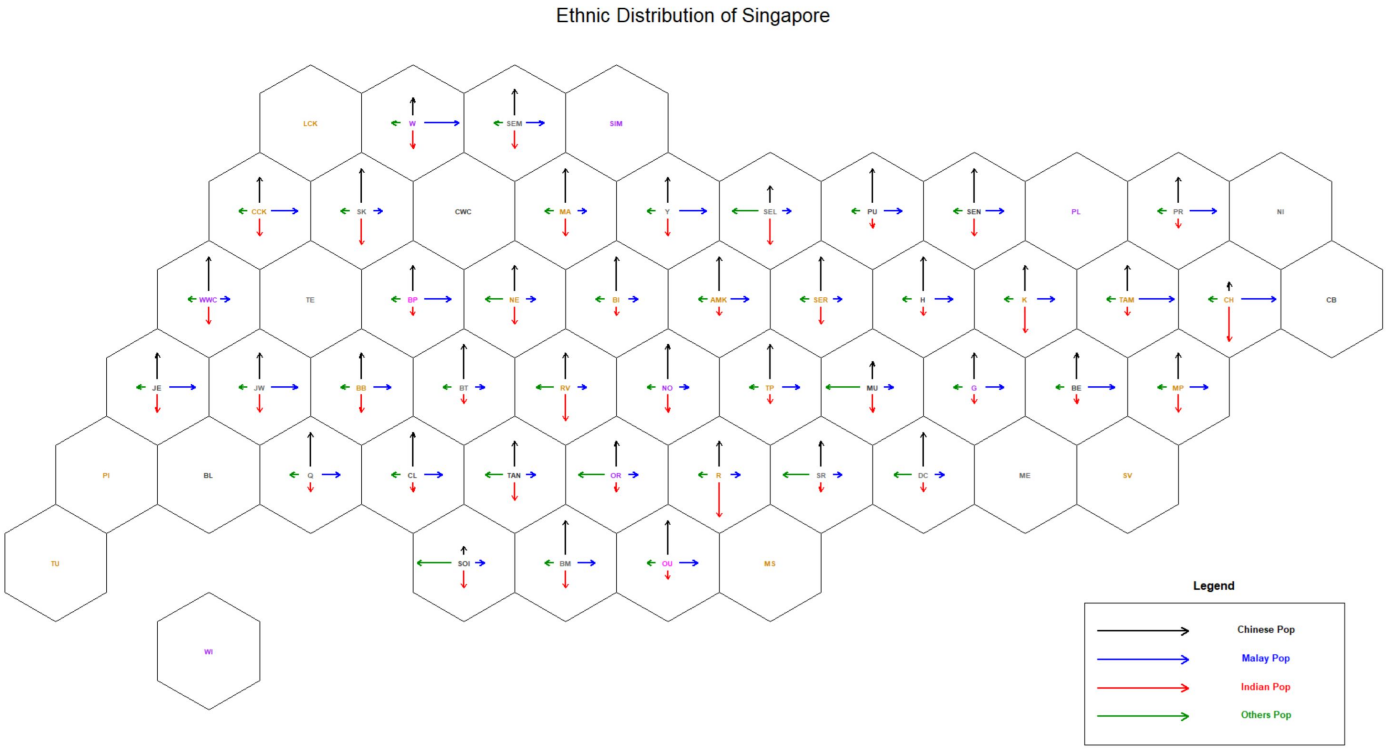

Figure 26: Plot of Ethnic Distributions with K-Means Clustering

From Figure 26 above, the resultant labels are labelled according to which cluster they belong in using K-Means Clustering. The user can then input whichever clustering method they want to use, and choose whether they want the labels to be labelled according to the clusters. Users can also choose whether they want the plot to be saved.

Multivariate Extensions

We also extend this plotting function in R to be able to handle data that are jointly distributed, in this case, we use the gender proportions of each Ethnic Group in each PA as the other variable.

For this, we use the thickness of the arrows to depict the male gender proportions in each PA for each Ethnic Group, the thicker the arrow, the higher percentage males there are.

Similarly, clustering methods can be applied to the data, but users have to specify which data of the 2 they want to show on the plot labels.

One such plot is shown in the next page.

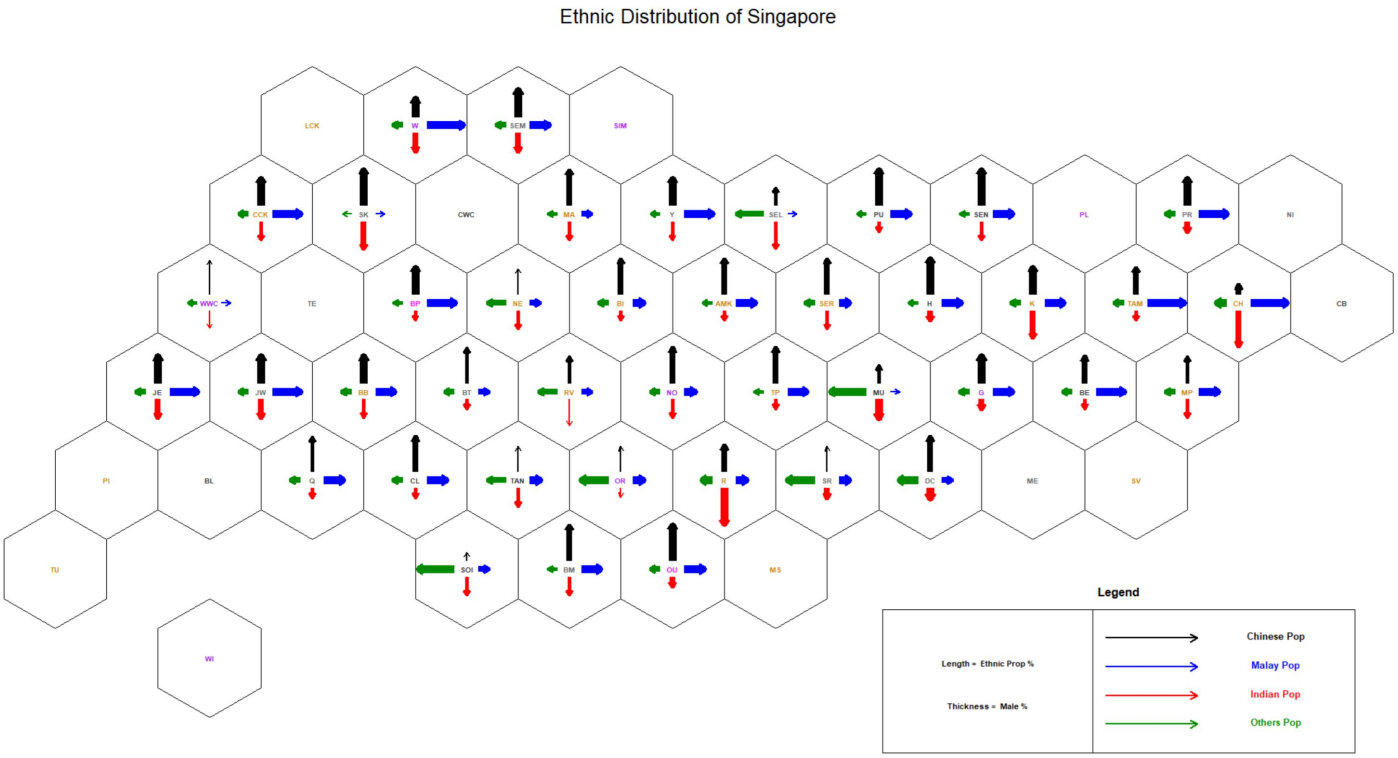

Figure 27: Plot of Ethnic Distributions + Gender with K-Means Clustering

We can see that in Figure 27, The legend helps readers to understand the plot better as the Legend labels for the length and thickness of arrows are specified, along with the colours for the arrows denoting which Ethnic Group they refer to.

Also, in Figure 27, the labels are chosen to be the clusters obtained from K-Means of the Gender proportions. We can easily see the difference between each PA, and this plotting function comes with heavy user customisability.

Thus, we are also able to solve the problem of plotting many variables on a single map.

Some Limitations

- The PA's not being represented at their exact geographical location - Jurong East (labelled 'JE') is on the west of Jurong West (labelled 'JW').
- We are not able to tell the exact values of each data, but we can tell the relative difference between each Planning Area in the most succinct manner.
- The areas of the hexagons are fixed, thus there is no way to change the scale of the map to allow for many more variables.

Conclusion

In conclusion, we are able to answer all the questions in Page 13. The plots shown as part of the methodology are:

1. Very easy for the reader to understand. missing data is plotted as grey areas, or areas that have no fill, whilst the labels for the PAs which contain missing data are kept.

2. Able to show as much information as possible, while combining statistical techniques and graphical techniques by adding the results of the statistical techniques into the plot (without loss of information).

3. Able to plot joint distributions.

4. Able to standardise the size of each PA in the map, and labels are easily added.

A quick check on Google Scholars shows us that these kind of maps have not been seen before, and these maps are easy to understand by the layman reader.

Thus, looking back at the Project Scope in Page 5, we find that through this project, we have been able to meet the main goals of our study as we have successfully:

- Created tools in R, with GIS capabilities to produce a novel univariate/bivariate variable map to represent demographic summaries.

- New type of map is easy to comprehend, and minimises loss of information

- Able to be applied to more datasets (only applies or is limited to datasets with row names matching the 55 PA's as well as has 4 categories)

References, and and a user's guide to the implementation of the R code can be found in the Appendix.

Data files are uploaded to this Google Drive link, and the R codes are uploaded to this Gihub Page.

Appendix A

References

(1) Jim Vallaningham (2017) Multivariate Map Collection: Retrived from:

https://vallandingham.me/multivariate$_m$*aps.html*

(2) Census Data Mapper: Retrieved from United States Census Bureau:

https://datamapper.geo.census.gov/map.html

(3) Singstat Table Builder: Retrieved from Department of Statistics Singapore:

https://www.tablebuilder.singstat.gov.sg/publicfacing/mainMenu.action

(4) Singapore Public Data: Retrieved from Govtech Public Data Source:

https://data.gov.sg/

(5) Master Plan 2019 Planning Area Boundary (No Sea): Retrieved from Govtech Public Data Source:

https://data.gov.sg/dataset/master-plan-2019-planning-area-boundary-no-sea

(6) Master Plan 2014 Planning Area Boundary (No Sea): Retrieved from Govtech Public Data Source:

https://data.gov.sg/dataset/master-plan-2014-planning-area-boundary-no-sea

(7) Gastner, M.T. , Seguy, V. , More, P.(2018) Fast flow-based algorithm for creating density-equalizing map projections: retrieved from PNAS:

https://www.pnas.org/content/115/10/E2156

(8) Friendly, Michael. A.-M. Guerry's Moral Statistics of France : Challenges for Multivariable Spatial Analysis. Statist. Sci. 22 (2007), no. 3, 368–399. doi:10.1214/07-STS241. Extracted from:

https://projecteuclid.org/euclid.ss/1199285037

(9) Hexbin map in R: an example with US states: Retrieved from:

https://www.r-graph-gallery.com/328-hexbin-map-of-the-usa.html

(10) Dempsey, C. (2015) How to Make a Tile Grid Map Using Excel: Extracted from:

https://www.gislounge.com/how-to-make-a-tile-grid-map-using-excel/

Appendix B

R Implementation

The R code is found in 2 R files:

- Lib_Data_Source.R - Contains R code for the following purposes:
  - Extract current working directory and set working directory.
  - Check for required packages and install them if they are missing in the package list.
  - Ask user if they want to save the current workspace.
  - Check if Internet Connection is available.
  - Check if the required files exist, if not, download all the required files from the Google Drive link
  - Process data files that are required for this project.
- Hexagonal_Simulation_Combined_Var.R -
  - Defined function that takes in 16 parameters, some of which are optional and will be explained in greater detail.

A third R file should be opened to produce the plots, sourcing these 2 files before any code is run in this order: (1) Lib_Data_Source.R ; (2) Hexagonal_Simulation_Combined_Var.R

Appendix C

Lib_Data_Source_R - User Guide

This R file uses the following packages:

- 'tcltk','svDialogs': For dialog/message popup box displays.
- 'downloader','readxl', 'plyr', 'dplyr', 'broom': To download and tidy data.
- 'ggplot2','grid': For plotting purposes.
- 'colorspace', 'RColorBrewer', 'viridis' : For colours in R.
- 'sf', 'geojsonio', 'geogrid': To manipulate Geospatial Data in R.
- 'devtools', 'Gmedian,' , 'XML': Miscellaneous Purposes

When `source{Lib_Data_Source.R}` is called for the first time, the user should expect to see a couple of things:

1. A popup will appear asking if the user wants to save the data:

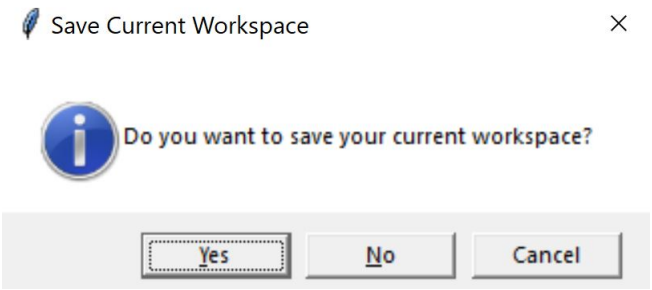

Figure C.1: Popup Message 1

2. If 'No' is selected, the code stops. If 'Yes' is selected, another popup message appears:

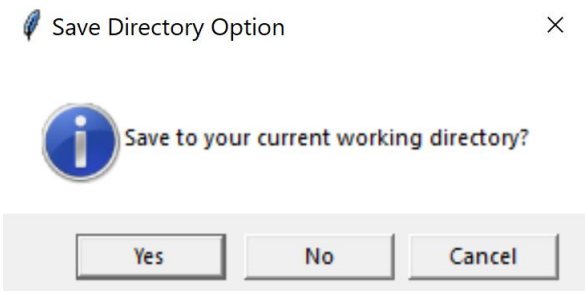

Figure C.2: Message 2

3. If 'Yes' is selected, the current workspace is saved to the current directory, but if 'No' is selected, another popup appears:

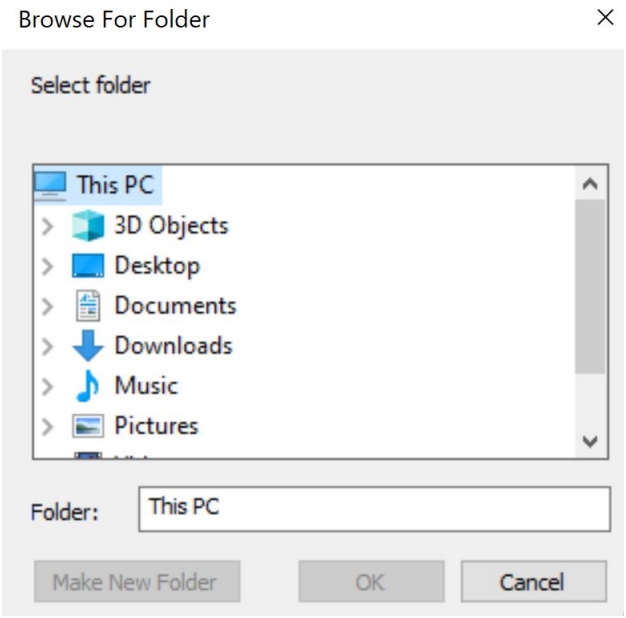

Figure C.3: Message 3

4. Once the directory is selected, the workspace is saved into the chosen folder.

5. The code will then attempt to install missing packages from the ones specified in Page 40, as shown in the next page.

6. In this figure in the next page, note that 'downloader' is to be installed, thus the code runs as so.

```
Checking for packages to be installed
downloader  to be installed!
Installing package into 'C:/Users/dkzk/OneDrive/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/downloader_0.4.zip'
Content type 'application/zip' length 24920 bytes (24 KB)
downloaded 24 KB

package 'downloader' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\dkzk\AppData\Local\Temp\Rtmp4afBqX\downloaded_packages
downloader  installed!
Packages installed and in use!
```

Figure C.4: Installing Missing Packages

7. It then checks whether the required files are installed, if not it will be downloaded from Google drive into the current working directory.

```
Checking if required files are available...
SHP_14 not found!
GJSON_19 not found!
SG_ethnic_composition not found!
SG_ethnic_composition_counts not found!
Ethnic_Breakdown_by_PA not found!
Download and unzip of 2014 SHP data done!
Download and unzip of 2019 GJSON data done!
Download of SG ethnic composition data done!
Download of SG ethnic composition counts data done!
Download of SG ethnic composition by gender data done!
All files downloaded!
```

Figure C.5: Checking and Downloading Missing Data

8. It then prepares the files to be ready for processing, and once the script ends, these messages will appear:

```
Dropping unused Z coordinates into 2D file

Success! File is at 2D_new19PA.geojson
Files ready for processing!
```

Figure C.6: End of Script

Appendix D

Hexagonal_Simulation_Combined_Var.R - User Guide

This R Script only consists one chunk of code, defining the plotting function for Figure 27 above. The function is called `plot_sg_joint_var()` and it takes in 16 parameters, some of which are optional:

1. `shp_file`: The Shape file to plot (default is PA_14 Singapore Shapefile)
2. `var1_matrix`: Full data matrix of the first variable to plot
3. `var2_matrix`: Full data matrix of the second variable to plot (default = NULL)
4. `areas_to_plot`: Names of Planning Areas to Plot/Perform Clustering on
5. `area_names_var`: Variable Name for the Row Names of Data Frame
6. `learn`: Learning rate of the simulation algorithm (default = 0.5)
7. `grid_seed`: Seed initialisation of the simulation algorithm (default = 14)
8. `grid`: 'Hexagonal' for Hexmap(default) and 'Regular' for Tile Grid Map
9. `category_names`: Names of the 4 categories for the variables
10. `variable_names`: Variable name sindicating the length and thickness of arrows (default = NULL)
11. `plot_main`: Main title of Plot Output

12. `save_plot`: Whether to save plot in current working directory

    For this, the saved plots are saved as TIFF images into the current working directory, with a width of 15 inches against a height of 7.5 inches, with a resolution of 1000ppi.

13. `clust_method`: Choice of clustering method to perform: 'Kmeans', 'Hclust', 'KMedian'

14. `clust_num`: Number of clusters

15. `which_cols`: Which columns of `var1_matrix` or `var2_matrix` to perform clustering methods on

16. `include_cluster`: Include clustering labels in plot (default = TRUE)

Some Examples

We first define some variables to help understand the code better:

```
#Plotted areas are the 41 PA's with NAs removed
areas_to_plot = rownames(ethnic_prop[-1,])

#The names of the 4 Ethnic Groups
category_names = c('Chinese Pop'  , 'Malay Pop' , 'Indian Pop' , 'Others Pop')
```

Example 1

We want to plot the Ethnic Groups on the map, using K-Means Clustering, with 5 clusters, and use the first 3 columns of the data matrix for clustering, and save the plots.

We also want to include the clustering in the labels, and use a hexmap.

```
plot_sg_jointvar(shp_file = PA_14_shp , var1_matrix = ethnic_0,
                 area_names_var = 'PLN_AREA_N' , areas_to_plot = areas_to_plot,
                 clust_method = 'Kmeans', clust_num = 5, which_cols = 1:3,
                 category_names = category_names , save_plot = TRUE,
                 include_cluster = TRUE,
                 plot_main = 'Ethnic Distribution of Singapore',
                  grid = 'hexagonal')
```

A popup message will appear as such:

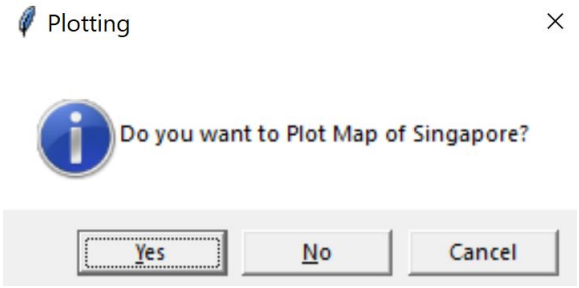

Figure D.1: Popup Message - Plotting Singapore Map for K-Means

If 'Yes' was selected, an additional plot will be generated, plotting the original clustered map without the simulated map. Otherwise, the plot is shown below:

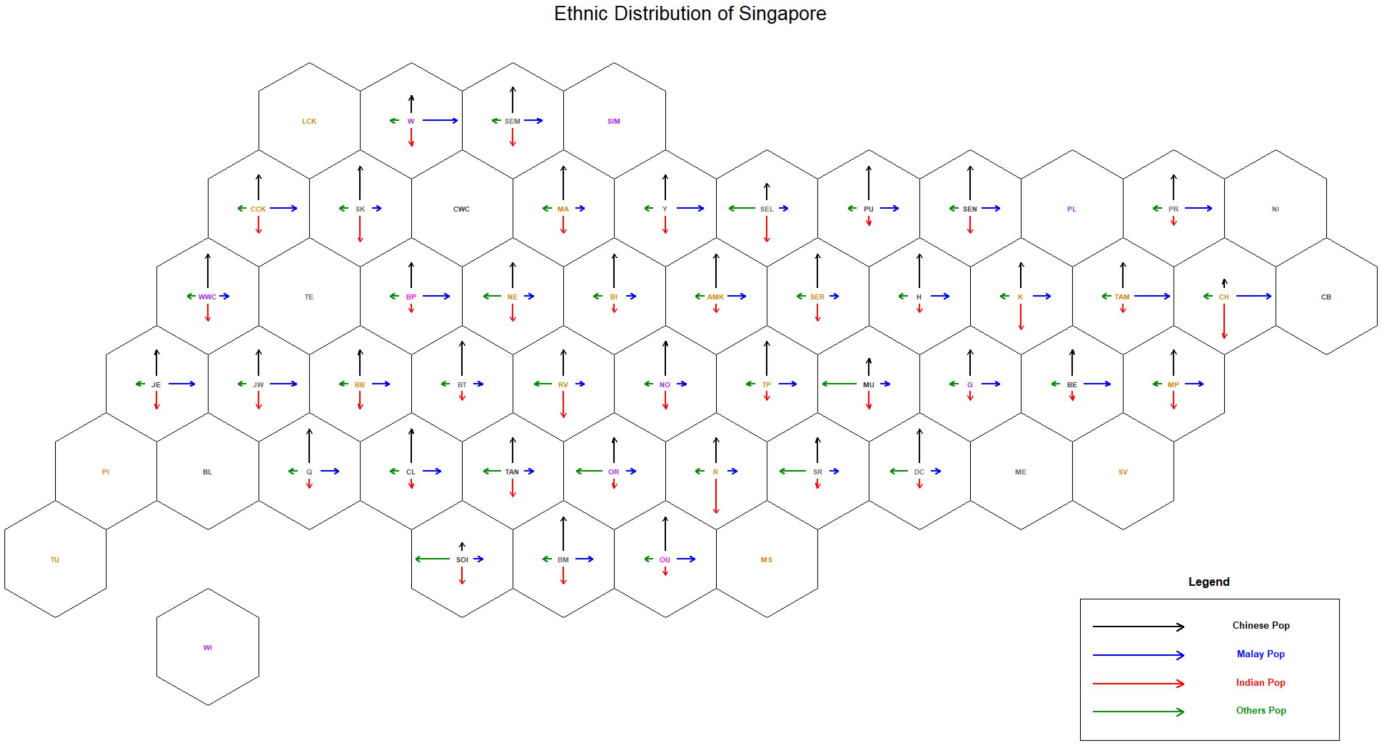

Figure D.2: Resulting Plot

Example 2

We want to plot the Ethnic Groups on the map, using Hierarchical Clustering, with 5 clusters, and use the first 3 columns of the data matrix for clustering, and save the plots.

However, We do not want to include the clustering in the labels, and use a hexmap.

```
plot_sg_jointvar(shp_file = PA_14_shp , var1_matrix = ethnic_0,
                 area_names_var = 'PLN_AREA_N' , areas_to_plot = areas_to_plot,
                 clust_method = 'Hclust', clust_num = 5, which_cols = 1:3,
                 category_names = category_names , save_plot = TRUE,
                 include_cluster = FALSE,
                 plot_main = 'Ethnic Distribution of Singapore',
                  grid = 'hexagonal')
```

In this case, the same popup as Figure D.1 will popup, which performs exactly the same function as mentioned above. However, not the user is required to input the linkage method to be used for the Hierarchical Clustering:

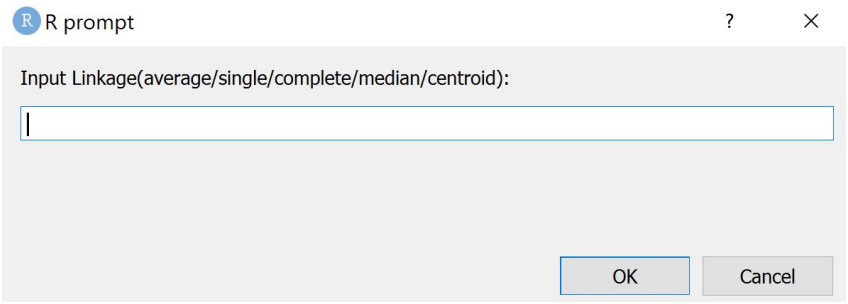

Figure D.3: Choosing Linkage for HClust Method

After keying in the linkage method, another popup will show:

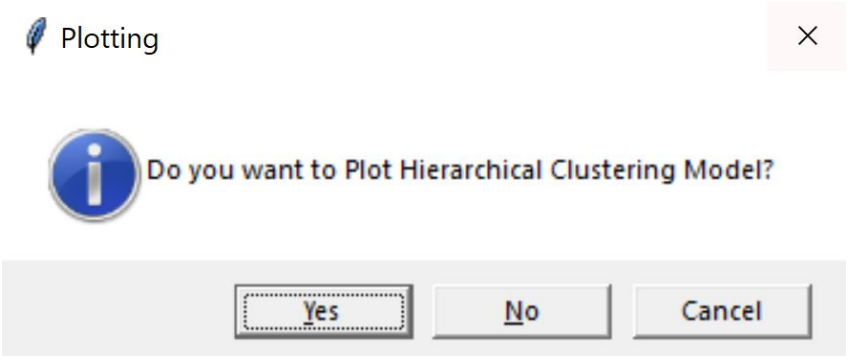

Figure D.4: To plot the Dendrogram for the Hclust Method

In this case, the corresponding output is as shown below:

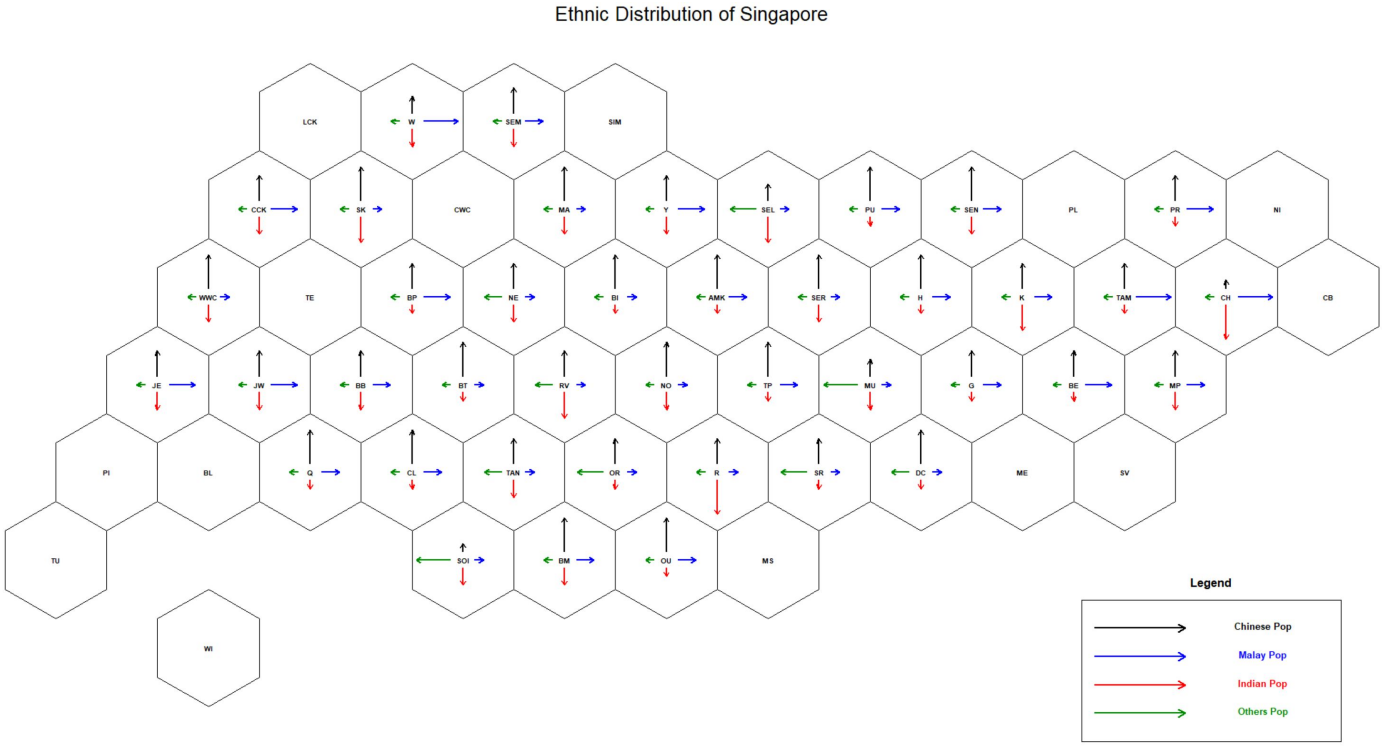

Figure D.5: Resulting Plot

Example 3

We now extend the plot to a multivariable map:

We want to plot the Ethnic Groups and gender data on the map, using K-Median Clustering, with 5 clusters, and use the first 3 columns of the data matrices for clustering, and save the plots.

We do not want to include the clustering in the labels, and use a hexmap.

```
plot_sg_jointvar(shp_file = PA_14_shp , var1_matrix = ethnic_0,
                 area_names_var = 'PLN_AREA_N' , var2_matrix = gender_male_0,
                 areas_to_plot = areas_to_plot, clust_method = 'Kmedian',
                 clust_num = 5, which_cols = 1:3,
                 category_names = category_names ,
                 save_plot = TRUE, include_cluster = TRUE,
                 plot_main = 'Ethnic Distribution of Singapore',
                 grid = 'hexagonal')
```

Since `variable_names` is not specified, the function will have a popup message asking the user to specify the variables indicating the length and thickness of the arrows.

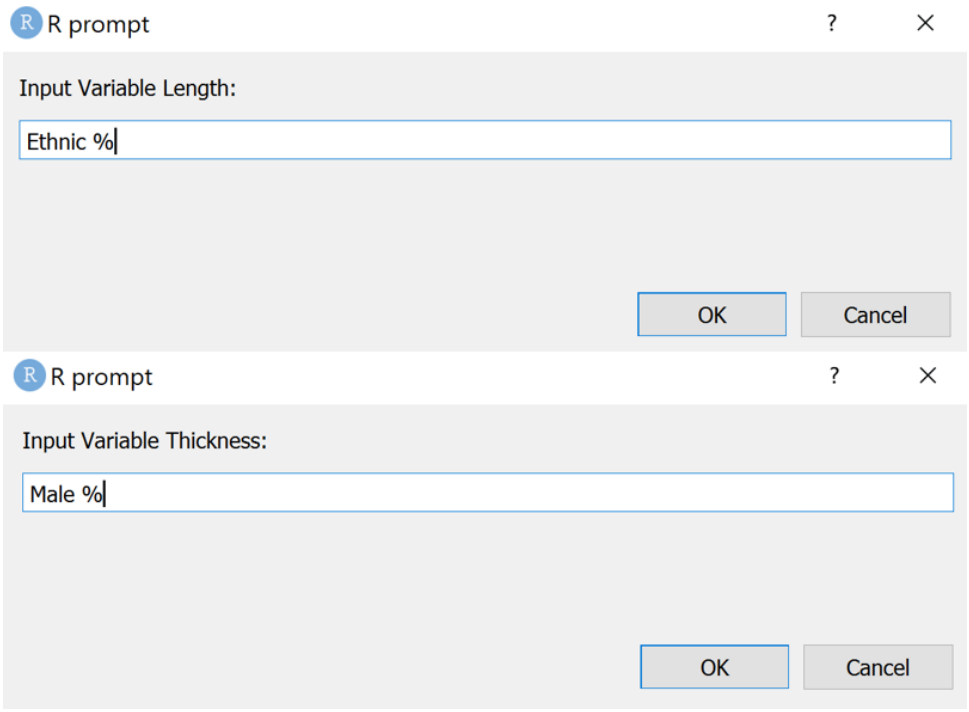

Figure D.6: Specifying Labels

After this is done, another popup message appears, asking the user to choose which matrix the clustering is to be applied on:

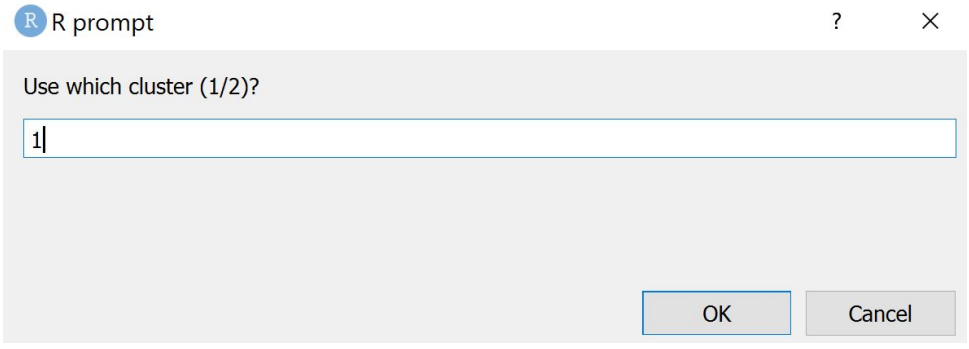

Figure D.7: Choosing the Data Matrix for Clustering

The final output is thus:

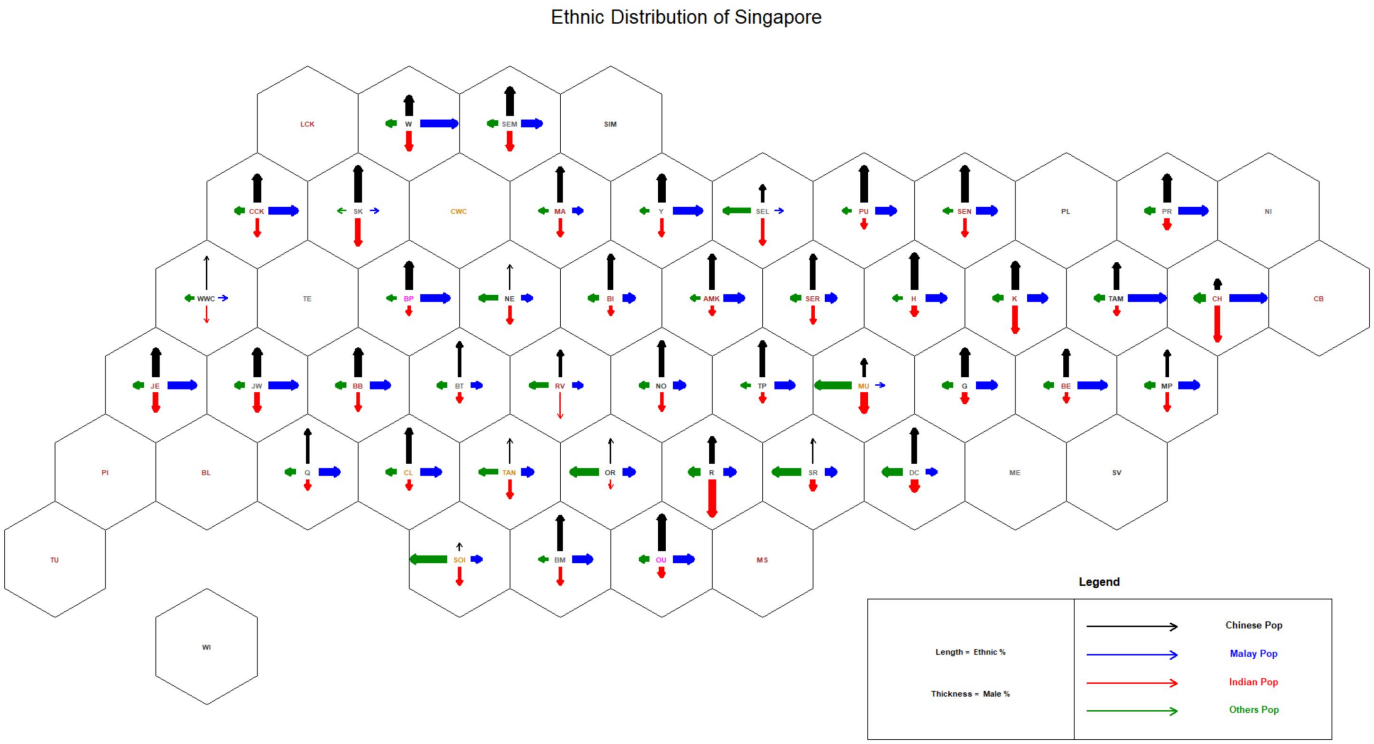

Figure D.8: Resulting Plot

Example 4

We try plotting the multivariable Tile Grid Map, using K-Means Clustering, with 5 clusters, and use the first 3 columns of the data matrices for clustering, and save the plots, and do not label the clusters.

```
plot_sg_jointvar(shp_file = PA_14_shp , var1_matrix = ethnic_0,
                 area_names_var = 'PLN_AREA_N' , var2_matrix = gender_male_0,
                 areas_to_plot = areas_to_plot, clust_method = 'Kmeans',
                 clust_num = 5, which_cols = 1:3,
                 category_names = category_names ,
                 save_plot = TRUE, include_cluster = FALSE,
                 plot_main = 'Ethnic Distribution of Singapore',
                 grid = 'regular')
```

In this case we get popups similar to Figure D.6 and Figure D.7, in which we input the same values as seen in the mentioned figures.

We thus obtain the following figure:

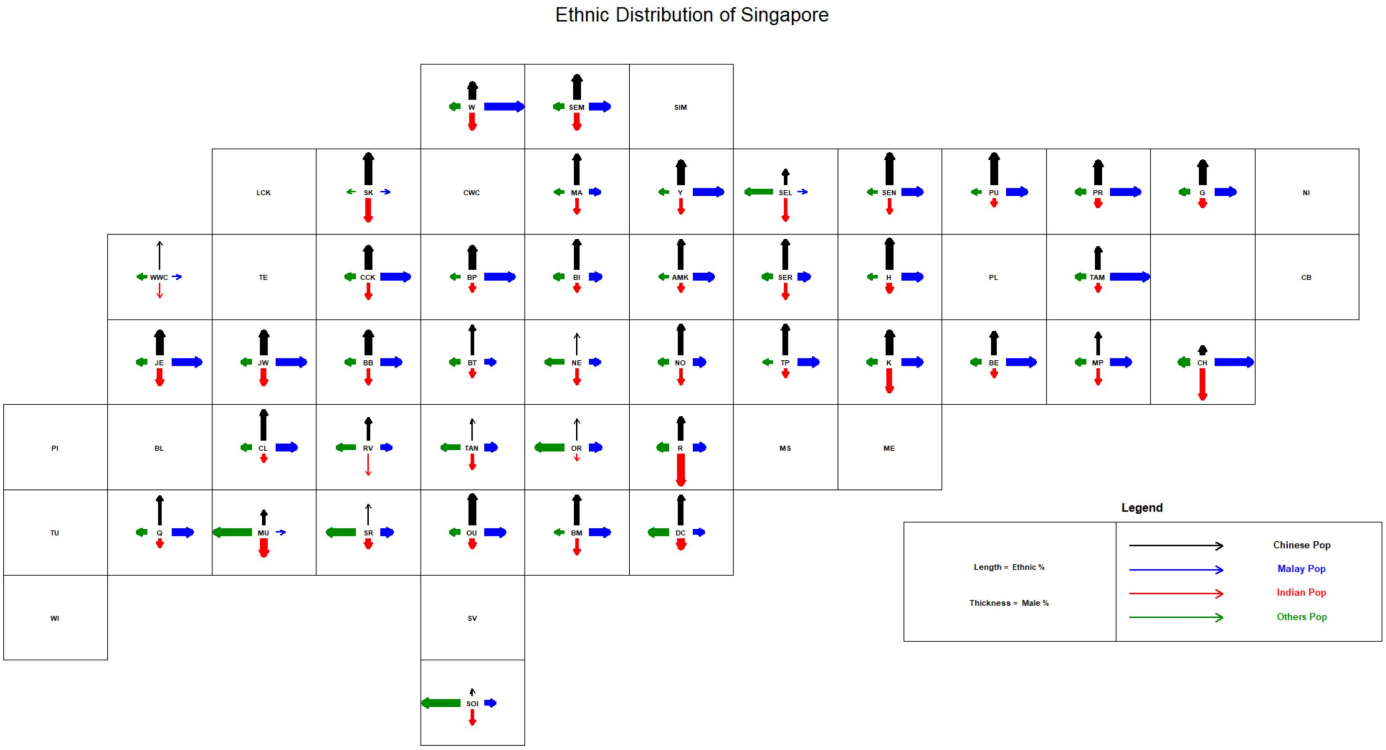

Figure D.9: Resulting Plot