Geography 187 Lab 1: Exploring autocorrelation using *GeoDa*

Introduction

This document will introduce you to the *exploratory spatial data analysis* (ESDA) software *GeoDa*, and how it can be used to measure and describe spatial autocorrelation (i.e. spatial dependence) in data aggregated to polygons (in this case, census data).

The *GeoDa* software is freely available for download and installation here:

http://geodacenter.github.io/download.html

Versions are available for Mac OSX, Windows and Linux so you can install a copy on your own computer and work on this lab in your own time – you don't need to be at the timetabled lab sections to complete the assignment, although you may find it helpful to attend to get assistance from the GSI.

Materials: data

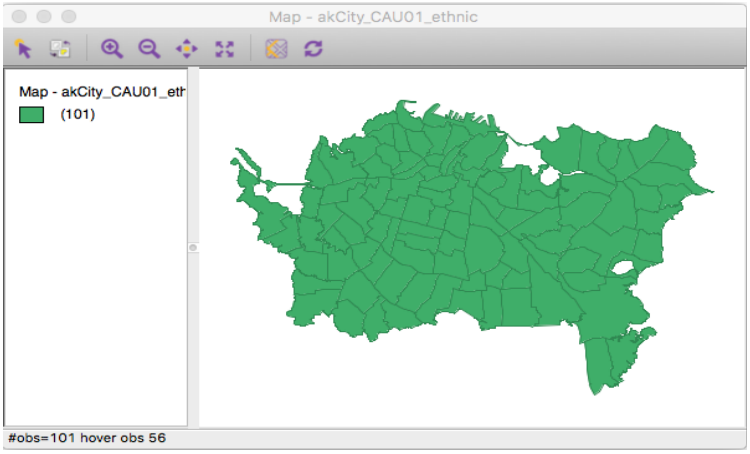
Two datasets are provided. One is from Auckland, New Zealand, and you can use it to check that you understand the instructions below (the illustrations in the instructions were created from this dataset). The other dataset is from three East Bay counties (from north to south these are Alameda, Contra Costa and Santa Clara), and focuses on the 2010 Census data at the tract level (areas of up to 10,000 population but more typically around 3-4,000). The data in this shapefile records percentages of the 2010 population in the major racial/ethnic categories used by the Census Bureau. You will use these data to get some idea of patterns of population in the East Bay.

The two datasets are ESRI shapefiles, available in the zipped archive **geog187-lab1.zip** on *bCourses*. You should download this zip archive and *unpack it* to the machine you are working on. It has to be unzipped. You will find that shapefiles are not single files, but actually consist of several files, although usually you only need to select the file with the **.shp** filename extension. All the other files must be present for things to work, but you only need to select the **.shp** file to access them.

A quick introduction to GeoDa

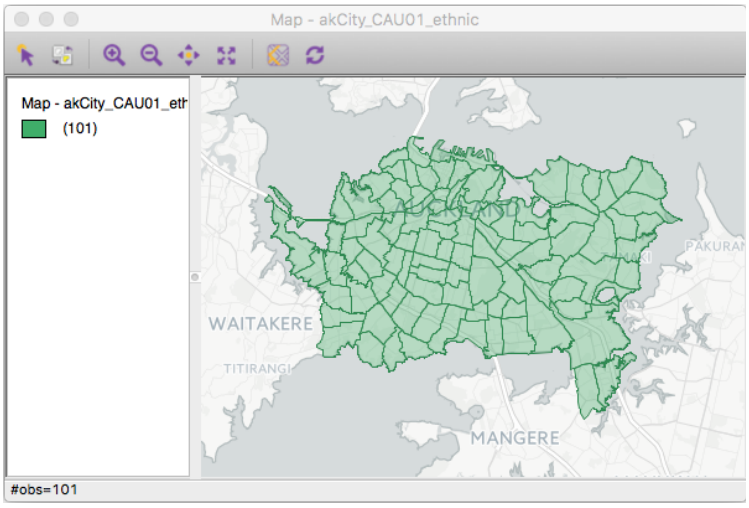
GeoDa is simple to use, because it doesn't try to do too many things. I have used the Auckland dataset to illustrate these instructions, and suggest you use that dataset to check you are getting things right.

First, you need to open a data file. To do this, go to the main menu **File – New Project From – ESRI Shapefile** and navigate to the shapefile you want. When it opens you should see a 'themeless' map as shown. So that you can tell where this is you can also add a web basemap to the map view by clicking the button on the map window tool bar second from right. This allows you to select a web-base map for the map view, which will give you some context.



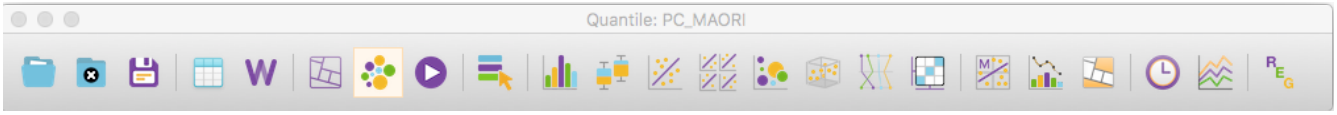
With 'Carto light' selected the map above looks as shown to thr right, and the census area units become semi-transparent to make it possible to make sense of the basemap. You can change the transparency by clicking again on the basemap button, and selecting the **Change map transparency**

option.



So far so good. You can also zoom in and out, and generally explore the map in the usual sorts of way (although panning is a little clunky). You can see a view of the data associated with the map by selecting the button that looks like a table from the *Geoda* button toolbar.

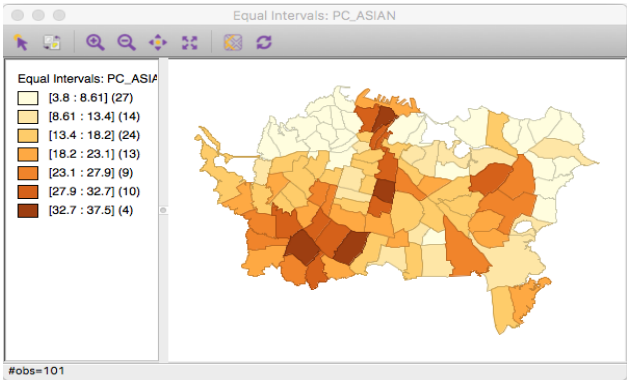
That's all very interesting, but we really want to see is geographical pattern in the data. You can do this either from **Map – <map type>** in the main menu (which will make a new map window), or by right-clicking in the map area and selecting **Change Current Map Type – <map type>** or from the main



menu by **Options – Change Current Map Type - <map type>**. Pick any of the map types and you will be asked for the variable to map. If you select (say) the **Equal Intervals** type with 7 intervals and PC_ASIAN (percentage Asian) as the variable, then you should get a map as shown below (without a basemap). The most relevant map types are as described below, along with the output you should see for this same variable:

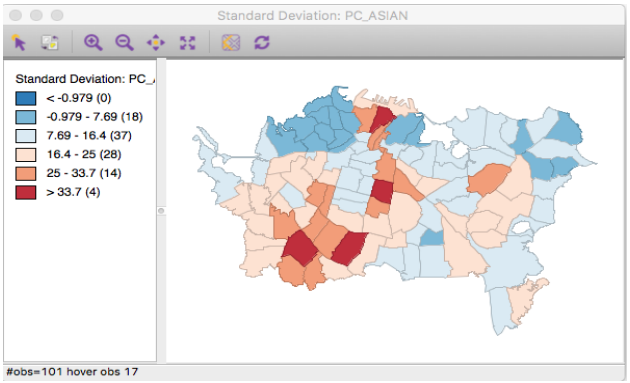
Equal Intervals Map

In this map the range of values in the data is divided into the requested number of equal ranges ('intervals') from the minimum to the maximum value, and areas are colored from light through to dark. The range associated with each color is shown in the legend (for example, the palest color shown is values from 3.8 to 8.61 percent), along with the number of areas in the interval in parentheses (27 for the palest color shown).



Standard Deviations Map

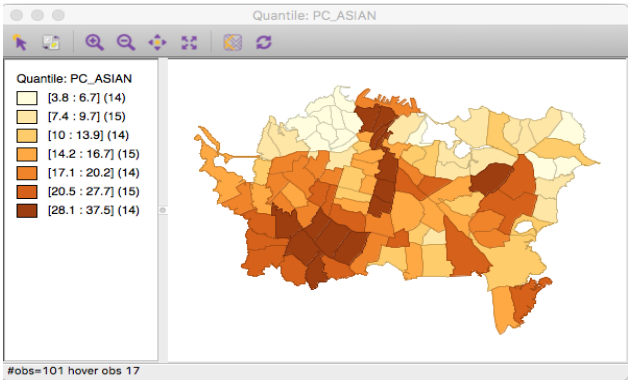
The standard deviation map is also an equal interval map, but the interval is determined based on the standard deviation of the values in the data. With standard deviation, colors are based on how much values stray from the mean. Values below the mean are in shades of blue, and values above the mean are colored in shades of red. This makes it easier to see both extreme low and high values (in dark colors) as well as the more 'typical' values (in pale colors).



Quantiles Map

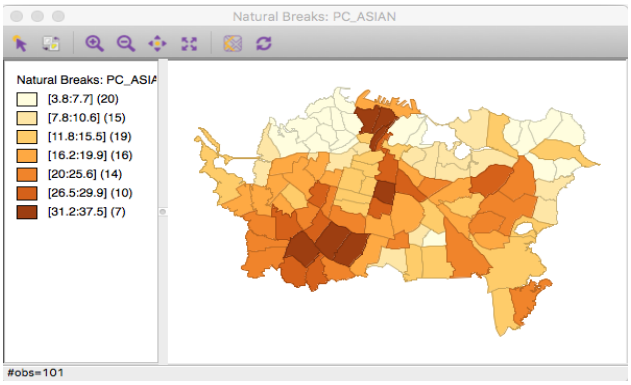
This map divides the areas into roughly equally sized

groups. Notice how each color is represented by 14 or 15 areas, so that all groups are as close to the same size as possible. To do this, the areas are arranged in order from lowest value to highest, and 'breaks' between categories are determined as needed. Many people like quantile maps, because they often have a good balance of colors (since there are equal numbers of areas in each color), but it is important to realize that the break values may not be very meaningful.



Natural Breaks Map

The natural breaks method chooses the breaks between categories to be in the largest 'gaps' in the data. These gaps may not be very obvious (they aren't very obvious in this map), but in theory, what this approach does is create categories whose members are similar to each other, and at least a little bit different from the other categories (because of the gaps).

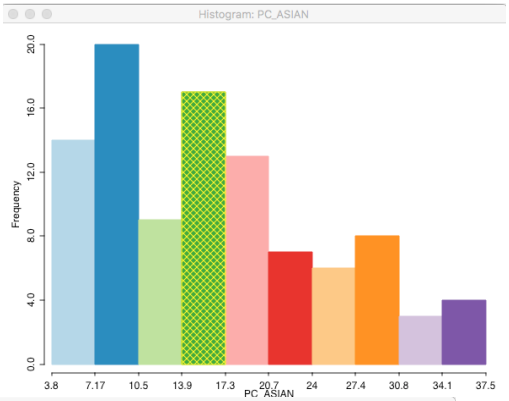


Data Exploration

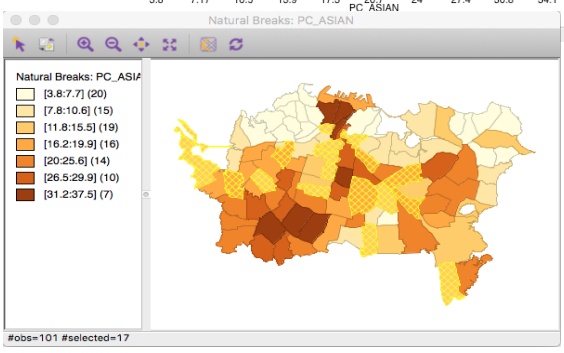
OK. The main point here is to get you to *explore* the data and get a feel for it. It is easy to switch between map types, or to make multiple map types quickly and compare them to one another. That's the idea of an exploratory analysis tool like this.

There are additional tools to help you with this exploration.

You can also see a statistical view of the data. To do this go to the **Explore –** menu and pick one of the options. For example, you might choose **Histogram** which will give you something like the display shown to the right. This is a traditional statistical chart. Where it gets interesting is that now, if you click anywhere, either in the map, or in the histogram, the corresponding areas in the chart and in the map will be highlighted. So if you select a histogram bar, all the map areas with values corresponding to that data range will be highlighted in the map. Such a linkage is shown in the two displays at right.



Equally, if you click-drag and select an area in the map, where the associated data values are in the statistical display will be highlighted. This remains true, even if the variable shown in the histogram or box plot (or whatever other chart you make) is displaying a *different* variable from that in the map. So, for example, you can use this to start to understand how different variables are related to each other. You can select, say, high percent Hispanic-Latino cases in a statistical chart, and at the same time see both where those areas are in



the map, and if they have high or low (say) Asian populations.

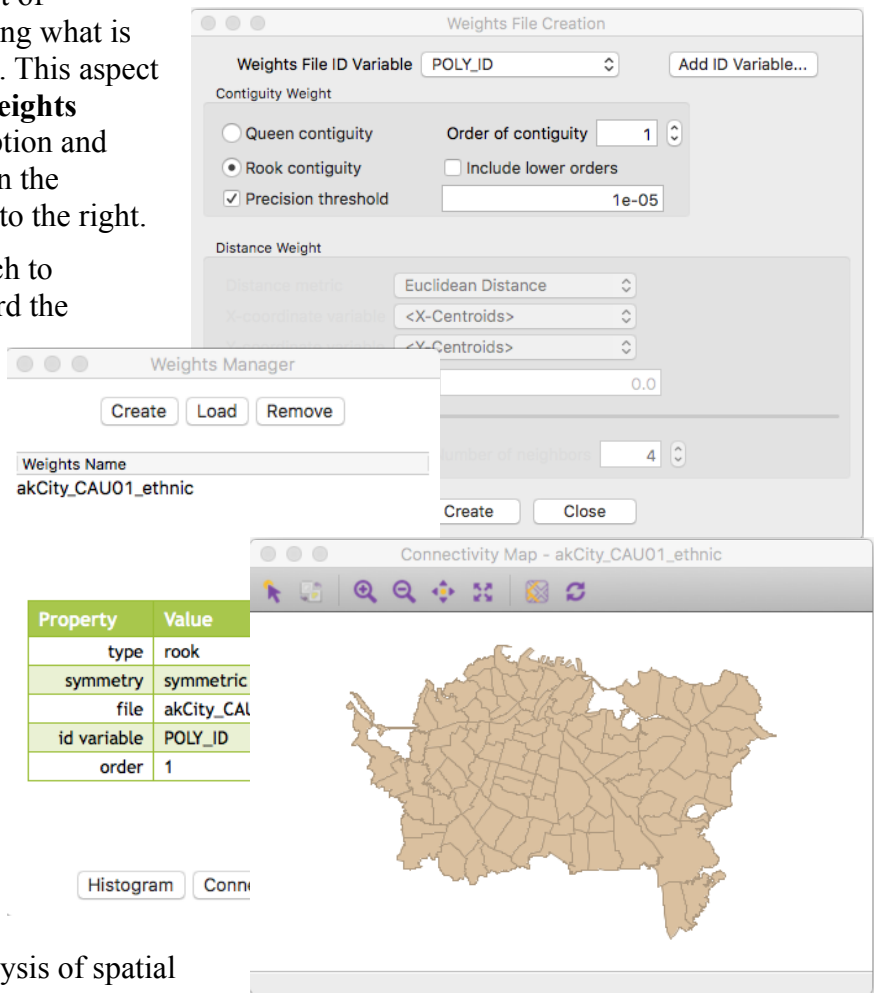
Another capability is *linked brushing*. This allows you to dynamically change the selection, while moving around either in the map or in the statistical chart, and to see the selection in other displays update at the same time. To perform linked brushing when you click-drag to select in a map or chart, also hold down the **<COMMAND>** key (on Mac) or **<CTRL>** key (on Windows). When you release the click, you will find that you can move the selection area around and that the selected elements in all windows will dynamically update. This can be a very powerful way to really get to know a dataset. To stop the dynamic updating just **<COMMAND>** or **<CTRL>** click again in any display area.

Measuring spatial autocorrelation

The main point of this assignment is to assess the spatial autocorrelation in the data. So, data exploration is an important first step to understanding the pattern, but we now have to run some analyses.

Building the Map Adjacency Structure

As we have seen in lectures, a key part of autocorrelation measurement is deciding what is 'near' for inclusion in the calculations. This aspect is handled by Geoda in the **Tools – Weights Manager**. To start select this menu option and click the **Create** button. This will open the **Weights File Creation** dialog shown to the right.



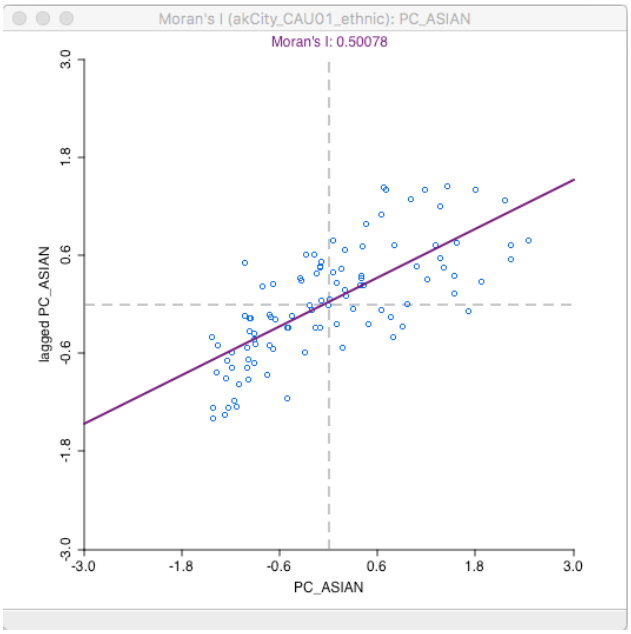
Here you can decide the basis on which to construct the spatial weights that record the adjacency structure you want to use for analysis. Select the 'Weights File ID' variable as 'POLY_ID' and 'Rook contuiguity' to make a basic weights file based on adjacency. Click **Create** and you will be asked where to save the file, and a name for it. Once you have done that you should then see in the **Weights Manager** dialog, the newly created file. To see the adjacency it is based on you can open the **Connectivity Map**. Note that you can use this tool to make multiple different weights files each based on different criteria and to manage their use in analysis. As has been discussed in lectures this is an important aspect to the analsysis of spatial autocorrelation.

With that done, there are two analysis tools you will use: **Univariate Moran's I**, and **Univariate Local Moran's I**, which you will find in the **Space** menu. Actually running these analyses is straightforward, and is described over.

Univariate Moran's I

Select the **Space – Univariate Moran's I** option, and select the variable you are interested in from the list that appears. Also select the Weights file you wish to apply.

The result of this analysis for the PC_ASIAN variable in the Auckland data is shown at right. This scatterplot is the *Moran scatterplot*, which will be explained in lectures. It is also discussed in some detail in sections 7.4 and 7.5 (pages 199-210) of O'Sullivan and Unwin (the chapter is available on *bCourses* in the **Files** folder).

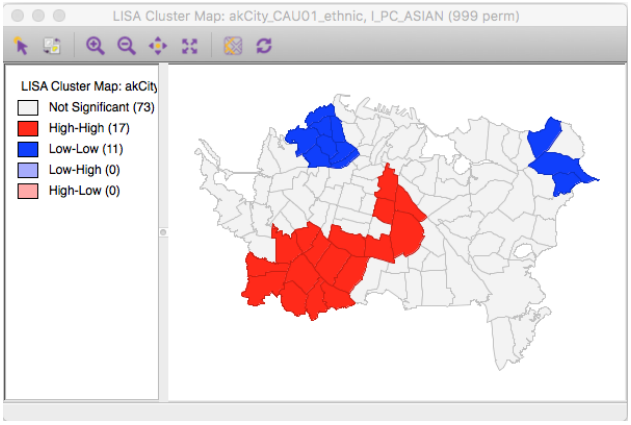


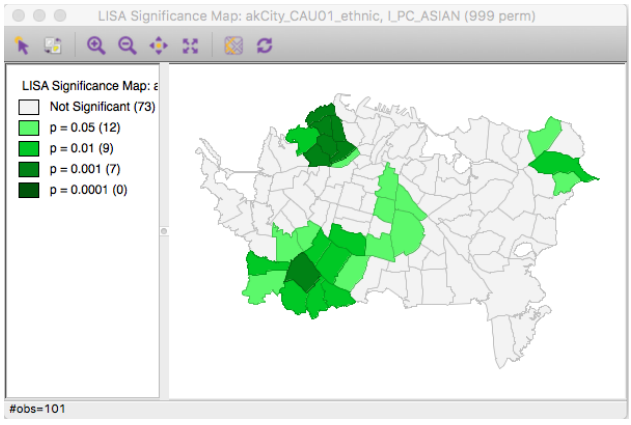
Apart from interpretation of the chart, there are two other things to note. First, the value of Moran's *I* is reported (in the example shown it is 0.50078, indicating strong positive spatial autocorrelation). Second, the Moran scatterplot is a statistical chart, just like any other, and so you can perform linked brushing between it and the map as you can with any other chart. This can be particularly useful for understanding the results you are looking at.

Univariate Local Moran's I

This really couldn't be easier. Select the appropriate menu option, request all the charts (Significance Map, LISA Cluster Map and Scatter Plot) and select the variable to analyze and hit OK. This will produce three outputs. One of them is the Moran scatterplot shown above; the other two are the Significance and LISA Cluster Maps, shown below (for the PC_ASIAN variable in the Auckland data).

Note that you might not get exactly the same maps as shown, due to the stochastic nature of the





permutation procedure in use.

Again, interpretation of these maps will be discussed in lectures in more detail. But briefly:

In the **Cluster Map**, the **RED** areas are ones with high values whose neighbors are also high values. **PINK** areas have high values but their neighbors have low values. **DARK BLUE** areas have low values and low value neighbors, and finally **PALE BLUE** areas have low values with high value neighbors. You can get a feel for what this means using linked brushing between the original thematic map and the scatter plot view.

The **Significance Map** shows how unusual the results are associated with any particular location and is what filters the areas that are colored in the cluster map. Only areas that are more unusual than some

chosen significance filter are shown in either map. You can change this filter setting by right-clicking in the Significance or Cluster map, and selecting **Significance Filter**. A lower setting will reduce the number of areas that are considered significant. If you want to set a value less than 0.01 (1 in 100) then you may need to increase the number of *permutations* that are used, which is another option available when you right-click in the map and select the **Randomization** option.

Getting maps and plots out of GeoDa

One downside of an exploratory tool is that it's really at its best *in use*. Most of what you learn from it happens *as you use the interactive features*. That makes it hard to capture what you learn in a static view, and also it means that *GeoDa* doesn't really put much emphasis on getting nice maps out of the program so that you can put them into documents and reports. Essentially, you are limited to using either:

Right-click in a window and choose either **Save Image As** or **Copy Image to Clipboard**. If you save to an image you can choose a format. Probably, you will want PNG (portable network graphic). If you copy to the clipboard you can then paste into a word document or other program. In theory, you can also copy the legend of a map in this way, but I have been unable to get this to work completely reliably. Try it and see if it works for you. If you have no luck, you will need the other option, , which is…

Use the operating system window and screen capture facilities. On Mac OSX the best bet is the **Preview** program, when you can use the **File – Take Screenshot – From Window…** or **– From Selection…** options. On Windows, you want the **Snipping** tool, which you can find by typing the first few letters of its name into the **Start** menu search bar. This provides similar options for making images from parts of the screen to those on Mac OSX.

OK… on to the assignment!

Assignment deliverables

Using the East Bay data provided, you need to do three things:

FIRST Determine the Moran's *I* statistic for the four 'major' census groupings, White, Black, Asian, and Hispanic-Latino. [***20%*** *for inclusion in your answer to* ***Q1*** *below*]

SECOND For **one** (you choose) of the four major groupings perform the Univariate Local Moran's analysis. You should produce a Moran scatter plot, Significance map and Cluster map for this analysis, and also a standard map (choose the map type you consider most informative) [***20%*** *for inclusion of outputs in your answer to* ***Q4*** *below*].

THIRD Write a short report addressing the following questions:

> **Q1** Compile a table of the Univariate Moran's *I* results for each of the four major census groups. Which is most 'aggregated' based on these results? Do you think this result is very meaningful? Explain your answer with respect only to the statistical results, *not* the geographical distributions (that's for the next question). [***15%***]
>
> **Q2** [The Sesame St. question] One of these group's geographic distributions is not like the others. Examining maps of each group, which group is different, and how? [***15%***]
>
> **Note, Q3 is difficult, but logically belongs here. You may want to go on to Q4 and come back to this one!**

Q3 What changes to the Moran's *I* approach might identify the difference among the groups more effectively? [**Hint:** This is not an easy question, and you are not expected to come up with a definitive answer. Think about (among other things): *scale*, the *census polygons* being used, the *total populations of each group*, and how we are considering 'near' when we use *polygon contiguity*] [***10%***]

Q4 For the Univariate Local Moran's analysis case that you carried out, present the various output displays, and comment on them. Where does this analysis suggest that communities of the group you have chosen are particularly concentrated? (The web map may be useful here). Do you feel this analysis provides any insight beyond a basic map of the census group in question? [***20%***]

You should prepare a short report addressing these questions, and including the various elements requested. Your report should not be longer than two pages of written answers (not including the figures) and almost certainly does not need to be as long as that!

Make sure your name is on your answer document and upload it using the *bCourses* submission box for this course. The submission deadline is February 13th at 11:59pm.

David O'Sullivan & Valerie Francella
January 27, 2016