* Towards an app for exploring segregation measures
* A toolkit
* Background.
* Massey and Denton’s 1988 paper on segregation measures aimed to move beyond a ‘Pax Duncana’, in which social researchers interested in segregation focused too much on the measure of dissimilarity proposed in Duncan 1955. M&D explored additional measures of segregation, and argued that there are five separate dimensions of segregation, and a number of indices which attempt to measure each of them. They catalogued and defined around 20 measures spread among these five dimensions.
* Despite this the Duncan index D is still predominant in social research on segregation. It is relatively easy to compute but there are a number of theoretical and conceptual problems. Firstly, it is aspatial: the order in which areal units are fed into it, and the spatial proximity between units, are not considered in the model. Secondly, it is a simple point estimate, and does not incorporate anything on uncertainty around this point estimate. Thirdly, it measures only one of the potentially distinct dimensions of segregation which social researchers should be interested in.
* The aim of this paper is to discuss the development of an app/toolkit, which can be used to help researchers see how different measures of segregation would summarise the same dataset. It will be a pedagogic and exploratory tool.
* It will explore the modifiable areal unit problem (MAUP), by showing how different segregation index values vary as the size of areal unit varies.
* It will introduce the dimensions of segregation highlighted by Massey & Denton, and explore whether correlation between indices within a dimension are greater than correlations between indices in different dimensions. So, it will explore whether the dimensions identified by Massey & Denton are statistically consistent for a range of datasets.
* It will look at the issue of uncertainty, and how not providing confidence intervals around the point estimate can lead researchers into falsely inferring a signal when what they are seeing is noise.
* It will look at how randomly generated data are categorised by the variety of segregation indices covered.
* It will look at how these different measures change as segregation increases. It will do this with synthetic data – using Schelling processes.
* It will include Moran’s I alongside the conventional segregation index, in order to see how this statistical measure of spatial autocorrelation could work as a segregation index.
* It will look at the relative contribution of the scaling effect and the zoning effect of the MAUP. The scaling effect is about how the segregation index value varies as the size of the areal unit changes. The zoning effect is about how changing the shape of an areal unit, without changing the size, leads to changes in the reported value.
* It will allow users to input their own spatial datasets.
* Calculating some of the indices will require that users specify the location of the ‘Centre’. Alternatively, clustering approaches could be used but that’s a task in itself.
* The aim of the research should be manifold. It should not just be able creating a tool, but also about produing a number of articles. These two functions should work together effectively: the creation of a p – publishing a paper should make it easier - help to promote the tool developed. A well developed toolkit should make developing the w research easier.
* The project needs to be clearly specified from the outset, such that there is a minimal version of it which will satisy funder requirements. Depending on the time this minimal product takes tp produce, modular increments and developments of the toolkit could then be produced.
* So, it is important to disgintuish between ‘what we need’, and ‘what we would like’. We should not bite too much than we can chew from the outset.
* The design architecture of the paper is as follows:
* The \
* The main bits of the research are:
* Building the code in R/Rcpp
* Deciding on the aesthetic of the toolkit so that it meets everyone’s expectations.
* Shiny will be used. Initially this will just run locally rather than be web-based.
* Coming up with a list of research questions we want to be able to answer using the toolkit, and/or using the code developed in writing the toolkit.
* Working out how to effectively iteratively develop the toolkit, so that it incorporates feedback from users/funders.
* A fairly typical user interface design within Shiny will be for the user inputs to be on a pane on the left side of the screen, and for outputs based on these user inputs to be on the right hand of the screen. There can also be tabs on either the right or left hand, which allow users to select from a range of grouped outputs.
* I think a good starting point would be:
* 1) have a user input that allows the user to select the data : initially this will be synthetic data only. Scheling type models. The user will specify the size of array, and the minority proportion. Initially there will only be two groups. From a drop down menu, a number of pre-built shapes will also be selectable, including checkerboard patterns, concentric circles, and so on.
* 2) the user specifies the size of the areal unit: how many squares are included. This will help address the scaling component of the MAUP. The user also specifies the aspect ratio of the areal unit: i.e. the ratio between height and width. This will help address the zoning component of the MAUP.
* 3) The user will be able to select from a range of metrics which should be calculated. These will be executed when the user clicks apply. Where indices require different parameters, additional menu buttons will appear, showing the default values used. The user can then change these if they want to.
* 4) On the right hand side of the pane: the user will see a summary of all of the metrics. This summary will include means, medians, SDs, 95% confidence intervals (if appropriate).
* 5) there will be a number of tabs on the right hand side of the app: these will allow exploration of individual metrics in more detail. The additional detail will include: density plots/histograms of the estimated values; computing time
* 6) There will also be a tab which shows correlation between indices. Selecting this tab on the right hand side will lead to a new option appearing on the left hand side of the pane: perform cluster analysis: on selecting this option the software will try to cluster the indices by similarity/dissimilarity. It will report concordance/discordance with the five dimensions specified by Massey & Denton.