AQMeN segregation index notes

Duncan Lee

May 14, 2013

1 Massey and Denton (1988)

Consider a population living in a city which has been partitioned into n areal units. Then consider two sub-populations, the majority population and the minority population. The aim is to describe how segregated the minority population are in the city. Massey and Denton identify five different dimensions to measuring the degree of segregation, and highlight the measures used in each dimension.

1.1 Evenness

The degree to which the minority population are evenly spread across the n areal units. Let P denote the proportion of the minority population across the entire study region, and let p_i denote the proportion in areal unit i. Then let T and t_i respectively denote the total population sizes from the whole study region and areal unit i. Then the following indexes of evenness are commonly used.

1. **Dissimilarity index** - measures the population weighted absolute difference between each areas minority proportion and the study region minority proportion.

$$D = \sum_{i=1}^{n} \frac{t_i|p_i - P|}{2TP(1-P)}$$

   Its value lies in $[0, 1]$, where 0 represents complete evenness ($p_i = P \ \ \forall i$) and 1 is complete segregation. This seems a reasonable measure of evenness.

2. **Gini index** - measures the weighted average of all pairwise differences in the minority proportions.

$$G = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{t_it_j|p_i - p_j|}{2T^2P(1-P)}$$

   Its value lies in $[0, 1]$, where 0 represents complete evenness ($p_i = p_j \ \ \forall i, j$) and 1 is complete segregation. This seems a reasonable measure of evenness.

3. **Entropy index** - measures the weighted average of the difference between the entropy of each area to the entropy for the entire study region.

$$H = \sum_{i=1}^{n} \frac{t_i(E - E_i)}{ET}$$

   where $E = P\log(1/P) + (1-P)\log(1/(1-P))$ is the study wide entropy measure and $E_i = p_i\log(1/p_i) + (1-p_i)\log(1/(1-p_i))$ is the entropy measure for areal unit i. For the latter E_i has its peak when $p_i = 0.5$ and decays to zero as p_i approaches either zero or one. The entropy index lies in $[0, 1]$, where 0 represents complete evenness ($E_i = E \ \ \forall i$) and 1 is complete segregation ($E_i = 0, 1 \ \ \forall i$). The issue with this (compared with the others) is that the difference $E - E_i$ does not have an absolute value around it, thus positive and negative values cancel each other. Thus one could get a situation where large differences in E_i across areal units (segregated study region) gives a very small Entropy index (even study region). This doesn't seem a good measure.

4. **Atkinson index** - is given by

$$A = 1 - \frac{P}{1-P}\left|\sum_{i=1}^{n}\frac{(1-p_i)^{1-b}p_i^b t_i}{PT}\right|^{1/(1-b)}$$

   where $b \in (0,1)$ is a shape parameter to be chosen by the researcher that determines whether large or small p_i values are treated as more important. I'm not sure what this measures, as p_i is not explicitly compared to anything, i.e. either P or p_j. Also, b involves choice so how should it be chosen? This does not seem a sensible measure.

1.2 Exposure

The second dimension is exposure, i.e. the extent to which people from the minority proportion are in contact with (live in the same areal unit) people from the majority proportion. Let x_i and y_i respectively denote the numbers of people from the minority (x_i) and majority (y_i) populations who live in areal unit i, with t_i denoting total population size in areal unit i as before. Finally, let X denote the total number of people from the minority population who live in the study region. Then the following indexes have been proposed.

1. **Index of exposure** - The index of exposure is given by

$$xP^*y = \sum_{i=1}^{n}\left(\frac{x_i}{X}\right)\left(\frac{y_i}{t_i}\right)$$

   and again ranges between $[0,1]$. It can be interpreted as the probability that a randomly drawn member of the minority population lives in the same areal unit as a member from the majority population. Thus when the minority population do not share any areal units with the majority (unexposed) then the above index equals 0. This index seems reasonable.

2. **Index of isolation** - The index of isolation is given by

$$xP^*x = \sum_{i=1}^{n}\left(\frac{x_i}{X}\right)\left(\frac{x_i}{t_i}\right)$$

   and again ranges between $[0,1]$. It can be interpreted as the probability that a randomly drawn member of the minority population lives in the same areal unit as a member from the same minority population. Thus when the minority population do not share any areal units with the majority (unexposed) then the above index equals 1. This index seems reasonable.

3. **Eta^2 index** - The Eta^2 index is given by

$$\text{Eta}^2 = \frac{xP^*x - P}{1-P}$$

   This is a modification of the isolation index that controls for population composition (via P). This seems slightly unnecessary, as the isolation index range over a specified interval and has a meaning.

1.3 Concentration

The third dimension is concentration, which quantifies the amount of physical space taken up by the minority population relative to the majority population. Concentrated minority populations live in a small numbers of areas which themselves have a small geographical area, and are said to be more segregated. In contrast, if the minority populations live in a large number of areas of large size they are less concentrated and hence less segregated. Let (x_i, X) be as before (number of minority population living in areal unit i and the total minority population), while (a_i, A) are the land area of areal unit i (a_i) and the study region (A).

1. **Delta** - The sum over all areal units of the absolute difference between the proportions of minority population and geographical size of each unit.

$$\text{DEL} = 0.5\sum_{i=1}^{n}\left|\frac{x_i}{X} - \frac{a_i}{A}\right|$$

   This ranges between $[0, 1)$, where 0 means the minority population are not concentrated as they are spread out in proportion to the land area. In contrast, if all the minority population live in a single small area (highly concentrated) then the metric is nearly equal to one $(1 - a_1/A)$. This measure seems reasonable.

2. **ACO** - Measures the total area inhabited by the minority population relative to the maximum and minimum areas that could be inhabited by the minority population. Assume the areal units are ordered by geographical size, so that $a_1 \leq a_2 \leq \ldots \leq a_n$. Now let n_1 be the rank of the areal unit (i.e. which one once they are in size order) where the cumulative total population of the areal units $1, \ldots, n_1$ equals the cumulative minority population. Conversely, n_2 is the rank of the areal unit where he cumulative total population of the areal units $n_2, \ldots, n$ equals the cumulative minority population but starting from the largest unit not the smallest. Finally, T_1 is the total population from areal units 1 to n_1 and T_2 is the total population from areal units n_2 to n. Then the measure is given by:

$$\text{ACO} = 1 - \frac{\sum_{i=1}^{n} x_i a_i/X - \sum_{i=1}^{n_1} t_i a_i/T_1}{\sum_{i=n_2}^{n} t_i a_i/T_2 - \sum_{i=1}^{n_1} t_i a_i/T_1}$$

   This statistic ranges between $[0, 1]$, with a value of 1 meaning the minority population is most concentrated (segregated). Here $\sum_{i=1}^{n}(x_i a_i/X)$ captures the amount of space taken up by the minority population, so the bigger it is the less concentrated the minority population are. This measure seems reasonable.

3. **RCO** - The ACO above only looks at the spatial distribution of the minority group, and not at the relative concentrations of the minority and majority groups. The RCO rectifies this.

$$\text{RCO} = \frac{\sum_{i=1}^{n}(x_i a_i/X)/(\sum_{i=1}^{n}(y_i a_i/Y)) - 1}{(\sum_{i=1}^{n_1} t_i a_i/T_1)/(\sum_{i=n_2}^{n} t_i a_i/T_2) - 1}$$

   This index goes between $[-1, 1]$, with a score of zero meaning the two populations are equally segregated. A value of 1 means that the minority population are much more concentrated than the majority population. This measure seems reasonable.

1.4 Centralisation

Centralisation measures the extent to which the minority population are located in the city centre.

1. **PCC** - The proportion of the minority population who live within the boundary of the central city.

$$\text{PCC} = X_{CC}/X$$

   where X_{CC} is the number of members of the minority population who live within the city centre boundary and X is the total minority population size. The problem here is how do you define the 'city centre'. In addition, no comparison is made to the level of centrality of the majority population. This measure does not seem reasonable.

2. **Relative centralisation index** - This measures relative to the majority population how close to the centre of the city do the minority population live. Order the areal units by their distance to the central business district, where areal unit 1 is the closest. Then let X_i and Y_i be the cumulative proportions of the minority (X) and majority (Y) populations in areal unit i. It is defined as

$$\text{RCE} \; = \; \sum_{i=1}^{n} X_{i-1}Y_i - \sum_{i=1}^{n} X_iY_{i-1}$$

   This index varies between $[-1, 1]$ with a value of zero meaning that the two populations have the same level of centrality. A value of one means the minority population are more central than the majority population. However, this index is rather ambiguously defined. Firstly, is X_i the proportion of the minority population in areal unit i, or a cumulative proportion based on areal units 1 to i? Secondly, ordering the areal units in distance from the business district is problematic, firstly it assumes the business district is the centre and secondly space is two-dimensional so assumes isotropy. Also, mathematically it uses X_0, what is this? Is it zero? This does not seem reasonable.

3. **Absolute centralisation index** - This measures the level of centrality of the minority population relative to land area and not the majority population.

$$\text{ACE} \; = \; \sum_{i=1}^{n} X_{i-1}A_i - \sum_{i=1}^{n} X_iA_{i-1}$$

   where A_i is the cumulative proportion of land area through unit i. This has the same $[-1, 1]$ range as the RCE index, with positive values meaning the minority population are likely to reside near the city centre. This does not seem reasonable for the reasons outlined above for the RCE.

1.5 Clustering

Clustering measures the degree to which areas containing the minority population are close to each other forming a large cluster or enclave.

1. **ACL** - Let C be an $n \times n$ spatial contiguity matrix. Then

$$\text{ACL} \; = \; \frac{[\sum_{i=1}^{n}(x_i/X)\sum_{j=1}^{n} c_{ij}x_j] - [(X/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}]}{[\sum_{i=1}^{n}(x_i/X)\sum_{j=1}^{n} c_{ij}t_j] - [(X/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}]}$$

   This index ranges from $[0, 1)$, with larger values suggesting stronger clustering. They also suggest a non binary C matrix given by $c_{ij} = \exp(-d_{ij})$ where d_{ij} is the distance between areal units i and j. This is clearly a bad idea for areal units of differing size. Aside from that if they used a binary contiguity matrix based on sharing a common border then this seems reasonable.

2. **SP** - The average proximity (degree of clustering) between minority group (X) members and other minority group members can be quantified as:

$$P_{xx} \ = \ \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{c_{ij}x_ix_j}{X}$$

   where large values represent more clustering. Similar descriptions for P_{yy} and P_{tt} - the degrees of clustering for the majority group (Y) and the total population (T). Then the measure of spatial clustering is:

$$\text{SP} \ = \ \frac{X \times P_{xx} + Y \times P_{yy}}{T \ \times P_{tt}}$$

   This index equals one if there is no difference between within and between group clustering, i.e. the minority population and the majority populations are not clustered together. Values above one suggest both groups X and Y tend to cluster with their own kind and away from the other. This metric seems reasonable

3. **RCL** - A simple alternative to the above is

$$\text{RCL} \ = \ P_{xx}/P_{yy} - 1$$

   which equals one if both groups have the same level of spatial clustering. It is greater than one if the minority group exhibit greater clustering. This metric does not have a maximum or minimum though. This metric seems reasonable.

4. **Distance weighted exposure and isolation** - These are essentially distance weighted versions of the exposure and isolation indices in 1.2. Consider the distance weight function

$$K_{ij} \ = \ \frac{\exp(-d_{ij})t_j}{\sum_{i=1}^{n}\exp(-d_{ij})t_j}$$

   Then exposure is quantified as

$$\text{DP}^*_{xy} \ = \ \sum_{i=1}^{n}(x_i/X)\sum_{j=1}^{n}(K_{ij}y_j/t_j)$$

   while isolation is quantified as

$$\text{DP}^*_{xx} \ = \ \sum_{i=1}^{n}(x_i/X)\sum_{j=1}^{n}(K_{ij}x_j/t_j)$$

   Massey and Denton argue that these measures add little to the non-spatial exposure measures.

1.6 Improvements

From a statistical perspective, this paper could be improved by considering the following.

1. The data analysis looked at five factors, because that is what they wanted to look at (i.e. they wanted to show five dimensions of segregation). Were there really five important factors, or could there be four or six? They discuss this briefly but do not quantify it statistically as to why they chose five.

2. None of these measures of segregation have any measure of uncertainty in them. Thus one cannot say is city A significantly more segregated than city B.
3. In the segregation measures they consider two groups, whites and the majority. Clearly this is not realistic, and it would be better to jointly consider the presence of more groups. Examples could be the four racial groups (they had white, black, asian and hispanic) or it could be split by religion.
4. Some of the indices are based on absolute differences, whereas in a statistical world squared differences are more commonly seen for some metrics. I'm not saying this is better but it would be interesting to see if it makes any difference.
5. How do they know the answers they get are correct? The true level of segregation in the population of real data is unknown. They should simulate data with a given segregation pattern and see which measures actually capture it best.

2 Massey et al. (1996)

The paper by Massey *et al.* (1996) updates the data analysis in the original Massey and Denton (1988) paper using data from 1990 rather than 1980. Essentially it answers the following questions.

1. It compares the results obtained by Massey and Denton in 1980 to those from 1990 using the same methodology and data.
2. It extends the analysis of Massey and Denton to all urban areas in the USA, rather than just the 50 largest (+ 10 with large hispanic populations) metropolitan areas.
3. It compares the results across different racial groups (black, asian, hispanic).

3 Reardon and Firebaugh (2002)

Reardon and Firebaugh (2002) extend the ideas in Massey and Denton by considering the realistic situation of having more than two groups (e.g. races, religions, etc) within a population. They define different approaches for creating measures of segregation, and show that some of these lead to the same segregation indices. Consider a population having M groups, where π_m are the proportions of people from each group over the entire study region. Then let π_{jm} denote the proportion of people from group m in areal unit j for $j = 1, \ldots, n$. Also, let t_j denote the total population for areal unit j, and T be the total population across the study region.

They describe two intermediate functions relating to the entire study population which go into the final indexes of segregation. The functions are given by

1. $I = \sum_{m=1}^{M} \pi_m(1 - \pi_m)$ - the interaction index.
2. $E = \sum_{m=1}^{M} \pi_m \ln(1/\pi_m)$ - the entropy index.

and equal 0 in the case of maximum segregation where all the population come from one group ($\pi_n = 1$ and the rest equal zero). At the other extreme, both functions are maximised (have maximum diversity and thus minimum segregation) when $\pi_m = 1/M \;\; \forall m$. Then they describe the following indexes of segregation.

1. **Dissimilarity index** extended from that presented in Massey and Denton

$$D \;=\; \sum_{m=1}^{M}\sum_{j=1}^{n}\frac{t_j}{2TI}|\pi_{jm} - \pi_m|$$

   This equals zero under no segregation.

2. **Gini index** extended from that presented in Massey and Denton

$$G = \sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{t_j t_i}{2T^2 I}|\pi_{jm} - \pi_{im}|$$

   This equals zero under no segregation.

3. **Information index** apparently equivalent (in the two group case) to the Entropy index H discussed by Massey and Denton

$$H = \sum_{m=1}^{M}\sum_{j=1}^{n}\pi_{jm}\frac{t_j}{TE}\ln\left(\frac{\pi_{jm}}{\pi_m}\right)$$

   This equals zero under no segregation.

4. **Squared coefficient of variation** similar to the dissimilarity index except squared rather than absolute difference is used.

$$C = \sum_{m=1}^{M}\sum_{j=1}^{n}\frac{t_j}{T}\frac{(\pi_{jm} - \pi_m)^2}{(M-1)\pi_m}$$

   This equals zero under no segregation.

5. **Relative diversity index** similar to the squared coefficient of variation.

$$R = \sum_{m=1}^{M}\sum_{j=1}^{n}\frac{t_j}{TI}(\pi_{jm} - \pi_m)^2$$

   This equals zero under no segregation.

6. **Normalised exposure** similar to the squared coefficient of variation.

$$P = \sum_{m=1}^{M}\sum_{j=1}^{n}\frac{t_j}{T}\frac{(\pi_{jm} - \pi_m)^2}{1 - \pi_m}$$

   This equals zero under no segregation.

The problem with this paper is that it is completely theoretical, they never actually apply any of their indices to data to see what they measure.

4 Reardon and O'Sullivan (2004)

This paper basically extends the Reardon and Firebaugh (2002) paper by considering measures of segregation that are inherently spatial, that is take account of the proximity between different minority groups in the population. One of the main points of the paper is that the use of census tracts or other areal units leaves the results open to the modifiable areal unit problem (MAUP). Therefore, they propose measures assuming point level data are available. Let (τ_p, τ_{pm}) be the population densities of the entire population and that of group m at point p in the study region R. Further let $(\tilde{\tau}_p, \tilde{\tau}_{pm})$ be the population densities of the entire population and of group m in the local environment of point p (defined as a spatial weighted average of population densities in the paper). Then

$$\tilde{\pi}_{pm} = \frac{\tilde{\tau}_{pm}}{\tilde{\tau}_p}$$

be the proportion in group m in the local environment of point p. Then the paper proposes the following measures.

4.1 Spatial exposure measures

They define exposure and isolation measures as the spatial analogue of the measures described in Massey and Denton. These are:

1. **Exposure** - Here we have

$$m\tilde{P}^*_n = \int_{q \in R} \frac{\tau_{qm}}{T_m} \tilde{\pi}_{qn} \mathbf{d}q.$$

   where T_m is the total population from group m across the study region.

2. **Isolation** - Here we have

$$m\tilde{P}^*_m = \int_{q \in R} \frac{\tau_{qm}}{T_m} \tilde{\pi}_{qm} \mathbf{d}q.$$

   where again T_m is the total population from group m across the study region.

4.2 Spatial evenness meaures

Three measures of spatial evenness are computed.

1. **Entropy**- The spatial equivalent of the entropy H index given by

$$\tilde{H} = 1 - \frac{1}{TE} \int_{p \in R} \tau_p \tilde{E}_p \mathbf{d}p$$

   where T is the total population and E and $\tilde{E}_p$ are measures of entropy of the population and at point p, where the latter is given by

$$\tilde{E}_p = -\sum_{m=1}^{M} \tilde{\pi}_{pm} \log_M(\tilde{\pi}_{pm})$$

   while E is defined similarly. The entropy measure will equal one under maximum segregation and zero under complete integration.

2. **Spatial relative diversity**- This is given by

$$\tilde{R} = 1 - \frac{1}{TE} \int_{p \in R} \frac{\tau_p \tilde{I}_p}{TI} \mathbf{d}p$$

   where $\tilde{I}_p$ is given by

$$\tilde{I}_p = \sum_{m=1}^{M} \tilde{\pi}_{pm}(1 - \tilde{\pi}_{pm})$$

   and I is defined similarly. I'm not sure about this because I equals zero under complete population wide segregation, in which case the integrand is undefined.

3. **Spatial dissimilarity index**- This is given by

$$\tilde{D} \;=\; \sum_{m=1}^{M} \int_{p \in R} \frac{\tau_p}{2TI} |\tilde{\pi}_{pm} - \pi_m| \mathbf{d}p$$

The paper has a number of drawbacks.

1. First, no data are used and no computation is done to see how these measures perform in practice.
2. This also allows them to overlook the practical difficulties in estimating the above measures. They say that one would not have data quantifying (τ_p, τ_{pm}), so instead they use areal level data. They suggest one could compute the population density in each areal unit, and assign all points in that unit that population density (constant population density across areal units). However, this is affected by MAUP, because if the areal units change so does these estimated population densities. Thus this paper is untruthful in claiming that it has overcome MAUP.
3. Also, in practical terms one would have to split the study region up into a finite set of points to allow computation, so again we are back to problems of how to choose this finite set of points.

5 Other metrics not considered

The two commonly used statistical measures of spatial clustering are Moran's I and Geary's C. Let the contiguity matrix be defined by C as before and consider the two group case, where p_i is the minority proportion in areal unit i. Note, Moran's I and Geary's C are not to my knowledge used in this segregation context, and hence it is only applied to one data set (equivalent to the two group case where the data are p_i). Although of course we could propose extensions of them. They are essentially global (i.e. across the entire study region) measures of spatial clustering, and are given by

$$I \;=\; \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}(p_i - \bar{P})(p_j - \bar{P})}{\left(\sum_{i \neq j} c_{ij}\right) \sum_{i=1}^{n} (p_i - \bar{P})^2}$$

and

$$C \;=\; \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}(p_i - p_j)^2}{\left(\sum_{i \neq j} c_{ij}\right) \sum_{i=1}^{n} (p_i - \bar{P})^2}$$

where $\bar{P}$ is the mean minority proportion over the study region. Typically c_{ij} is binary and based on whether or not two areas share a border. For Moran's I values under no spatial clustering are close to zero (the asymptotic mean is $-1/(n-1)$) and positive values correspond to spatial clustering of minority populations. Conversely, negative value indicate negative clustering, that minority populations are more likely to be closer to the majority population. In contrast, for Geary's C no clustering corresponds to a value of one. Values below one (C is never negative) correspond to clustering and values above one correspond to hyper integration.

There are other measures of *local spatial clustering/correlation*, that is the extent to which an area is similar to all of its neighbours. These include local Moran's I and local Geary's C and have been called Local Indicators of Spatial Association (LISA). They are essentially jus the numerator in the above with only one summation. That is the association for area i is computed based on the quantity (via local Moran's I)

$$I \;=\; p_i \sum_{j=1}^{n} c_{ij} p_j$$

which is then appropriately scaled. However, these measures capture average similarity between area i and all of its neighbours, which does not capture the variability in the relationship between area i and each of its neighbours. That is area i may be related to some but not all of its neighbours. The obvious way to measure similarity between areas (i, j) are through statistics of the form

1. $d_{ij} \;=\; (p_i - p_j)^2.$
2. $d_{ij} \;=\; |p_i - p_j|.$
3. $d_{ij} \;=\; (p_i - \bar{P})(p_j - \bar{P}).$

The problem here is that you have only one observation, so you would need to have replication of some sort. This could potentially be over M different groups.