An open-source framework for non-Spatial and Spatial segregation measures: the PySAL `segregation` module

Renan Xavier Cortes[1*]
Sergio Rey[1]
Elijah Knaap[1]
and
Levi John Wolf[2]
[1]*Center for Geospatial Sciences, University of California, Riverside.*
[2]*School of Geographical Sciences, University of Bristol*

July 9, 2019

Abstract

In urban geography and social sciences, segregation, usually consider five dimensions in a given society such as evenness, isolation, clustering, concentration and centralization. All of these measure can either ignore spatial context or take it into consideration. Currently, several segregation measures are available in the literature, but they lack of wide spread use, in part, due to their complex calculations. In addition, there are only a few works that address the problem of inference in segregation measures for either single measure or for comparison between multiple measures. This work tries to fill this gap by constructing an open-source segregation module in the Python Spatial Analysis Library (PySAL). This new module tackles the problem of segregation point estimation for some well-known non-spatial segregation indexes such as Dissimilarity (and its related), Gini, Entropy, Isolation, Concentration Profile, Correlation Ratio, and spatial indexes such as Spatial Proximity, Relative Clustering, Relative Concentration, Relative Centralization. Furthermore, it also presents a novel feature that performs inference for segregation and for comparative segregation, relying on simulations under the null hypothesis. We illustrate the use of this new library using tract level census data in American counties of non-Hispanic black population.

Keywords: open-source; segregation; PySAL; spatial analysis.

*We are grateful for the support of National Science Foundation (NSF) (Award 1831615) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) foundation (Process 88881.170553/2018-01).

1 Introduction

Segregation measures date back to the pioneer work of Park (1926). However, it was the work of Duncan and Duncan (1955) that leveraged the study by performing a deep analysis in segregation measures through the "segregation curve" of the most used indexes by that time.

Later on, Massey and Denton (1988) formalized the concept of segregation as a multidimensional phenomenon assuming that its extent depends on several factors for a given group in a given society. The latter called "hypersegregation" approach (Massey and Denton, 1989) assumes that segregation has five broad dimensions: Evenness, Isolation, Clustering, Concentration and Centralization.[1] Their work discuss some of the well-known indexes such as the Dissimilarity (D), Gini (G), Entropy (H), Isolation (xPx), Relative Concentration (RCO), Relative Centralization (RCE) and the Relative Clustering (RCL).

These indexes have an extensive literature in terms of methodological aspects.[2] Carrington and Troske (1997) proposed a modification on D and G indexes. Their approach relies on the fact that these measures could overestimate segregation, specially when small units are present, because most indices are functions of proportions and they can suffer from small sample problem due to large sampling variance of the denominators. Also, they argue that these indexes assess the distance from *evenness* rather than *randomness*. Rathelot (2012) also address this upward behavior of classical segregation indexes by building a parametric approach assuming that the frequency of population under study is draw from a probability following a beta mixture.[3] Also Allen et al. (2015) propose a bias-corrected approach and a density-corrected approach for D. In terms of spatial indexes, Morrill (1991) and Wong (1993) propose spatial corrections for the same classical index. Furthermore, Hong and Sadahiro (2014) developed two indexes, the concentration profile and the spatial proximity profile, which also tries to overcome some limitations in previous versions of spatial and non-spatial segregation measures.

However, the complexity of the calculations can represent a deterrent for a broader use of these measures. There are some open-source options to perform segregation analysis that enables the

[1] For a literature review on segregation, we refer to Royuela et al. (2010).

[2] For application examples, see Massey and Denton (1993), Carrington and Troske (1998), Hellerstein and Neumark (2008), Söderström and Uusitalo (2010) and Massey and Tannen (2015).

[3] More recently, d'Haultfoeuille and Rathelot (2017) addressed this problem assuming a nonparametric binomial mixture of the frequencies.

user to compute several measures such as the `seg` package of Hong et al. (2014) for the `R` language (R Development Core Team, 2008) and the Geo-Segregation Analyzer (GSA) (Apparicio et al., 2014).[4] The former, comprises 12 measures such as the D, three version of modified D, spatial proximity, concentration profile, spatial exposure, spatial isolation, spatial information theory, spatial relative diversity, spatial dissimilarity (surface based) and the decomposable measure of segregation. All these measures are wrapped in generic functions that produce outputs unique to each type of index. The latter has a vast range of 41 indexes[5] for either one group, two groups, multi-group or local indices. Although GSA represents a feasible way to estimate these indices, it may not be as convenient as a module under a broader spatial analysis tool, since the user has to download and install it independently only to perform segregation analysis. In addition, this option relies only on the use of *shapefiles* which, despite being one of the most popular Geographic Information Systems (GIS) extensions to handle spatial data, were developed and regularized by the Environmental Systems Research Institute (ESRI).

More recently, an important open-source contribution was made by Tivadar (2019) with the `OasisR` package. In this tool, a set of 50 indices are available comprising non-spatial and spatial measures, multi-group segregation measures and an inference framework for single values of segregation. Tivadar (2019) also discusses in detail several inconsistencies in classical segregation formulas.[6] Due to the vast number of studies and indexes that are present in the literature, the `OasisR` package poses as one of the most complete options for `R` users. So far, this used to be the only software option that provided some statistical inference framework for single values of segregation.

The free and open-source software, which allows the user to have full access to the algorithmic implementations, is an excellent option for researchers. The advantage of a full transparent and community-led development that open-source has can lead to more transparency, reliability; also, it allows virtually anyone to get involved in the development process. Therefore, our current

[4]Table 2 of (Apparicio et al., 2014) cites other options of software that also put effort to calculate these indices such as Reardon (2002) and Wong (2003), but not as open-source.

[5]In the original paper, they consider 43 different indexes, due to three Atkinson indexes versions. However, these indexes only differ in terms of the value of the parameter b, therefore we consider this index only once.

[6]One of the most prominent is the indexes issues presented in Wong (1993) discussed in the bottom of page 6 of Tivadar (2019). During the construction of the present module, the same problems were identified and the default approach of these indexes follows actually the latter study for this Python package.

approach has more power to broaden the use of segregation analysis in regional science since it relies in a fully open-source approach and can handle multiple types of spatial data input. The Python Spatial Analysis Library (PySAL) (Rey and Anselin, 2010) is a well-established library of the Python programming language (Rossum, 1995) for spatial analysis. Currently, PySAL has several features and modules comprising exploratory spatial data analysis, geospatial distribution dynamics, spatial econometrics, spatial graphs, geoprocessing, spatial graphs, data visualization of spatial data and models. Since PySAL has a broad scope of use and an active community of users and developers, it could be considered an ecosystem itself to perform geospatial data science. In this sense, the `segregation` module of PySAL intends to fill the gap of segregation analysis in this current library and Python scientific ecosystem.

Besides allowing the user to estimate the main spatial and non-spatial measures, this work also covers a functionality that is not usually object of concern in the segregation literature which is inference. In terms of previous work, Boisso et al. (1994) works with simulations to perform inference in a multidimensional version of the clasical gini index. Also, Ransom (2000) develops a sampling exercise of a multinomial distribution for the dissimilarity index and gini index in order to build asymptotic distribution of the estimators. Allen et al. (2015) builds an inference framework developed a likelihood ratio test for the presence of any systematic segregation for a bias-modified D. In addition, like Ransom (2000), they develop tests for this measure relying on the asymptotic distributions. More recently, Rathelot (2012) and d'Haultfoeuille and Rathelot (2017) tackles the issue of inference on segregation. Rathelot (2012) developed a beta mixture approach for the dissimilarity, Gini and entropy indices trying to overcome the small unit problem and a bootstrap and the delta method was proposed to provide inference. The more sophisticated approach of d'Haultfoeuille and Rathelot (2017) assumes a mixture of binomial distributions and build testable assumption, bootstrap confidence intervals for the bottom and upper limits of the probability parameters of the distributions. Also more recently, Napierala and Denton (2017) discuss the behavior of the dissimilarity index under uncertainty of American Community Survey data under simulations studies.

This current work tackles the inference framework for segregation making use of distributions for these measures under the null hypothesis where segregation does not hold. To perform inference for a single measure, we follow an extension of the procedure described in Allen et al. (2015)

where we generate the distribution of each measure under the null of no systematic segregation by creating multiple samples generated using restricted conditional probabilities (absence of *systematic segregation*). Also, in order to generalize the use of our inference approach for single measures, the PySAL `segregation` module comprises different approaches to the null hypothesis of absence of segregation assuming *evenness*, *spatial permutations*, *absence of systematic segregation with permutation* and *evenness with permutation*, which will be covered later.

The major contribution of our framework is the ability to perform inference to compare more than one segregation measure.[7] We rely in an extended version of Rey and Sastré-Gutiérrez (2010) where is provided an inferential basis for comparisons of regional statistics. Their approach relies on random labelling where, in each permutation, each data are randomly assigned to a point in time. However, our approach for comparative segregation comprises two situations: firstly, a single region evolution between two points in time and, secondly, two regions comparison in the same point in time. The former is a straightforward case of Rey and Sastré-Gutiérrez (2010), but the latter is more challenging due to the possibility of totally different spatial contexts of each city which may directly affect the segregation measure. To try to provide alternative ways to assess the absence of segregation, our framework comprises not only random data labelling (*"random label"* approach), but also a randomization labelling process accordingly to cumulative distribution function of the percentage of the interest group in each unit (*"counterfactual composition"* approach).

2 The PySAL `segregation` module

The PySAL `segregation` module (hereafter referred as SM)[8] can be divided into two frameworks: point estimation and inference wrappers. The first framework can be, in turn, subdivided into non-spatial indexes and spatial indexes. The inference wrappers present functions to perform inference through simulations over the null hypothesis for a single value or for comparison between two values.

Each framework is explained separately below.

[7] In terms of software, so far, we are unaware of any that performs inference for comparison between them.

[8] Available at https://github.com/pysal/segregation.

2.1 Point Estimation

Originally, SM had 25 segregation indexes ranging from non-spatial indexes and spatial indexes that can be summarized in Table 1.[9][10] This table presents the main information of each function given the appropriate name of each measure, the class/function name, whether it is spatial or not and what are the input parameters. A detailed description of each index and respective literature, presented as a table, can be found in the Appendix A.

Table 1: Segregation Measures available in the PySAL `segregation` module

Measure	**Class/Function**	**Spatial?**	**Function Inputs**
Dissimilarity (D)	Dissim	No	-
Gini (G)	GiniSeg	No	-
Entropy (H)	Entropy	No	-
Isolation (xPx)	Isolation	No	-
Exposure (xPy)	Exposure	No	-
Atkinson (A)	Atkinson	No	b
Correlation Ratio (V)	CorrelationR	No	-
Concentration Profile (R)	ConProf	No	m
Modified Dissimilarity (Dct)	ModifiedDissim	No	iterations
Modified Gini (Gct)	ModifiedGiniSeg	No	iterations
Bias-Corrected Dissimilarity (Dbc)	BiasCorrectedDissim	No	B
Density-Corrected Dissimilarity (Ddc)	DensityCorrectedDissim	No	xtol
Spatial Proximity Profile (SPP)	SpatialProxProf	Yes	m
Spatial Dissimilarity (SD)	SpatialDissim	Yes	w, standardize
Boundary Spatial Dissimilarity (BSD)	BoundarySpatialDissim	Yes	standardize
Perimeter Area Ratio Spatial Dissimilarity (PARD)	PerimeterAreaRatioSpatialDissim	Yes	standardize
Distance Decay Isolation (DDxPx)	DistanceDecayIsolation	Yes	alpha, beta
Distance Decay Exposure (DDxPy)	DistanceDecayExposure	Yes	alpha, beta
Spatial Proximity (SP)	SpatialProximity	Yes	alpha, beta
Relative Clustering (RCL)	RelativeClustering	Yes	alpha, beta
Delta (DEL)	Delta	Yes	-
Absolute Concentration (ACO)	AbsoluteConcentration	Yes	-
Relative Concentration (RCO)	RelativeConcentration	Yes	-
Absolute Centralization (ACE)	AbsoluteCentralization	Yes	-
Relative Centralization (RCE)	RelativeCentralization	Yes	-

All input data for SM rely on `pandas` DataFrames (McKinney, 2011) for the non-spatial mea-

[9]More recently, some other measures were added to SM, but we conducted the current work with the original 25.

[10]In addition, the module has a function/class named `Compute_All_Segregation` that performs point estimation of several segregation measures at once.

sures and `geopandas` DataFrames (Jordahl, 2014)[11] for spatial ones. Loosely speaking, the user needs to pass the `pandas` DataFrame as its first argument and then two strings that represent the variable name of population frequency of the group of interest (variable `group_pop_var`) and the total population of the unit (variable `total_pop_var`). So, for example, if a user would want to fit a dissimilarity index (D) to a DataFrame called `df` to a specific group with frequency `freq` with each total population `population`, a usual SM call would be something like this:

```
index = Dissim(df, "freq", "population")
```

In addition, every class of SM has a `statistic` and a `core_data` attributes. The first is a direct access to the point estimation of the specific segregation measure and the second attribute gives access to the main data that SM uses internally to perform the estimates. To see the estimated D in the generic example above, the user would have just to type `index.statistic` to see the fitted value.

2.2 Inference Wrappers

Once the segregation classes described in Section 2.1 are fitted, the user can perform inference to shed light for statistical significance in regional analysis. Currently, it is possible to make inference for a single measure or for two values of the same measure. The summary of the inference wrappers is presented in Table 2.

Table 2: Inference Wrappers available in PySAL `segregation` module

Inference Type	Class/Function	Function main Inputs	Function Outputs
Single Value	InferSegregation	seg_class, iterations_under_null, null_approach, two_tailed	p_value, est_sim, statistic
Two Values	CompareSegregation	seg_class_1, seg_class_2, iterations_under_null, null_approach	p_value, est_sim, est_point_diff

2.2.1 A single value

The function `InferSegregation` of SM perform inference through simulations for a single value of a specific index. The user needs to specific inputs that rely on previous SM estimations with the

[11] It is worth to mention, that using a `geopandas` for the non-spatial indexes is also valid since it "behaves" as a usual `pandas` dataframe.

seg_class parameter, number of iterations under the null hypothesis with the iterations_under_null parameter, specify which type of null hypothesis the inference will iterate with the null_approach parameter, set if the p-value estimated will be two-tailed estimated with the two_tailed parameter and could pass additional parameters for the segregation estimation. Therefore, a usual call for this function would be:

```
inference_result = InferSegregation(index, iterations_under_null = 10000,
null_approach = "systematic", two_tailed = True)
```

The null_approach parameter in this single measure framework present several options. The default "systematic" draws multinomial simulations assuming that every group has the same probability with restricted conditional probabilities given by the share unit of the the total population (Allen et al., 2015)[12], "evenness" draws independent binomial distributions assuming that each unit has the same global probability of the group under study, "permutation" randomly allocates the units over space keeping the original values as proposed by Rey (2004) for regional measures, the "systematic_permutation" is a combination of "systematic" and "permutation" assuming absence of systematic segregation and randomly allocates the units over space and, lastly, "even_permutation" is a combination of "evenness" and "permutation" assuming that each measure have same global binomial probability and randomly allocates the units over space.[13]

The user can access the results of the function with the p_value and est_sim.[14] The first is the pseudo p-value estimated from the simulations and the second are the estimates of the segregation measure under the null hypothesis previously established.

2.2.2 Comparative Inference

To compare two different values, the user can rely on the CompareSegregation function. Similar to the previous function, the user needs to pass two segregation SM classes (seg_class_1

[12]Assuming that n_{ij} is the population of unit i of group j, this approach assumes that the distribution of people from each j group is a multinomial distribution with probabilities given by $\frac{\sum_j n_{ij}}{\sum_i \sum_j n_{ij}} = \frac{n_{i.}}{n_{..}}$.

[13]We are aware that for some measures some approaches would not be appropriate, but we chose to let this to let this framework as generic as possible. For example, the modified Dissimilarity (Dct) and Gini (Gct), rely exactly on the distance between evenness through sampling which, therefore, the "evenness" value for null_approach would not be the most appropriate for these indexes.

[14]There is also a statistic attribute to access the original point estimation of the measure.

and `seg_class_2`) to be compared, establish the number of iterations under null hypothesis with `iterations_under_null`, specify which type of null hypothesis the inference will iterate with `null_approach` argument and, also, can pass additional parameters for each segregation estimation.[15] Therefore, after fitting two measures, a usual call for this function would be:

```
index_1 = Dissim(df1, "freq", "population")
index_2 = Dissim(df2, "freq", "population")
compare_result = CompareSegregation(index_1, index_2,
iterations_under_null = 10000, null_approach = "random_label")
```

Assuming that 1 and 2 are the subindexes for two measures, the null hypothesis to compare them is

$$H_0 : Segregation\ Measure_1 - Segregation\ Measure_2 = 0$$

and, therefore, the `null_approach` plays an important role, once again, in the inference framework. The default `"random_label"` approach follows directly the approach of Rey and Sastré-Gutiérrez (2010) where SM random labels the data in each iteration. In this approach, the data swap between the two groups allowing them to be either two points in time for the same region, in order to compare its evolution, or two different regions in the same point in time, thus comparing different spatial contexts. The `"counterfactual_composition"` approach introduced in Section 1 tackles the null hypothesis in a different way. In this framework, the population of the group of interest in each unit is randomized with a constraint that depends on both cumulative density functions (cdf) of the group of interest composition[16] distribution. In each unit of each iteration, there is a probability of 50% of keeping its original value or swapping to its corresponding value according of the other composition distribution cdf that it is been compared against. Thus, we build artificial values that can represent what would be the frequency of a specific group if it would have presented another cdf for the composition. This latter approach can be considered as a special case of a inverse re-sampling (Devroye, 1986) where you sub-sample 50%, on average, the existing empirical distribution with the data of another distribution according to its cdf.

[15] Note that in this case, each measure has to be the same SM class as it would not make much sense to compare, for example, a Gini index with a Delta index.

[16] We refer the word *composition* to the group of interest frequency of each unit. For example, if a unit has total population of 50 and 5 people belonging to group A, the group A composition of this unit is 10%.

Lastly, this function also return a `p_value` and `est_sim` attributes. The first is the two-tailed p-value generated from the simulations and the second is the estimation differences under the null hypothesis that need to be compared to zero in the absence segregation difference. In addition, the user can access the `est_point_diff` attribute which is the point estimation of the difference between the two values.

2.2.3 The `plot` method

The `plot` method of the SM inference framework is a visual representation of the segregation under the null hypothesis confronted with the value under study. It relies on `matplotlib` (Hunter, 2007) and `seaborn` (Waskom et al., 2017) functions.

For single measures, the distribution is the point estimation along all iterations, while a vertical red line represents the actual value. On the other hand, for inference comparison, the distribution represents the differences between the measures in each iteration while a vertical red line represent the estimated difference using the original data. In the latter visual representation, values closer to zero indicates an absence of segregation difference. The user can visually inspect the results with `inference_result.plot()` or `compare_result.plot()`.

3 Performance Comparison Study

A very important aspect to investigate in the module is the time necessary for its estimations. Since the nature of each index can vary in terms of the mathematical operations involved, either due to the dimension of segregation assessed or due to internal simulations/optimizations, the difference in time between the indexes can change drastically.[17]

Figure 1 depicts a time comparison for a single estimation of each index of Table 1 in seconds for a 10 x 10 regular lattice with simulated data[18].[19] From this figure, it is possible to see that the

[17]We also noticed that for most of the indexes, specially the spatial ones, SM was much faster to estimate than the implementation of Tivadar (2019).

[18]We used the total population of 100,000 and generated a random composition for each unit given from a Uniform distribution between 0 and 1.

[19]The indexes were fitted used the default values for input. Although this can be a source for difference in the values, we highlight that these default values are roughly comparable since all indexes that rely on simulations (Dct, Gct, and Dbc) have the same value of 500 for the iterations and indexes that rely on integration (R and SPP)

Modified Gini (Gct) poses as the most time-consuming index among all the set of indexes. This is due to the fact its construction relies on a bootstrap simulation of multiple binomial distributions for each unit and also because its calculation, given by Equation 5 in Appendix A, rely on an outer product of vectors which can be expensive depending on the size of the data. The second most time expensive index is the Density-Corrected Dissimilarity that relies on numerical optimizations to estimate a θ_j component in its formula. The following positions are filled by simulations based indexes such as the Modified Dissimilarity (Dct) and Bias-Corrected Dissimilarity (Dbc). At last, the Boundary Spatial Dissimilarity (BSD) presented a significant value among all the set of indexes.

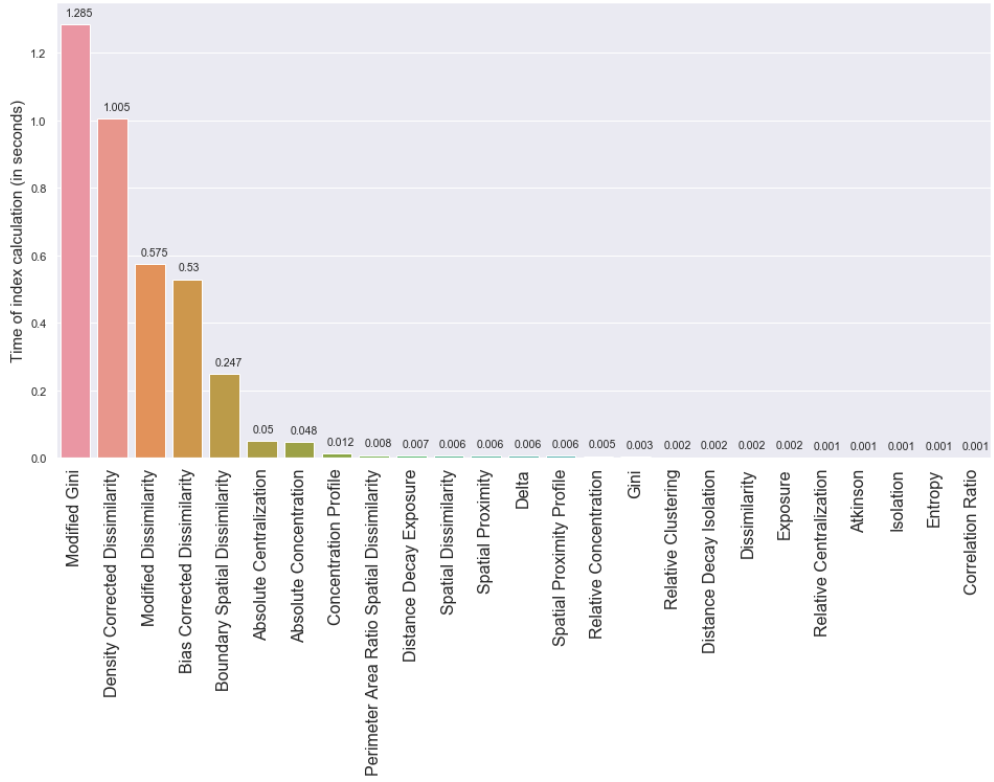


Figure 1: Time comparison estimation between all indexes of SM for a 10 x 10 regular lattice

have the same number of thresholds for integral approximation of 1000. The index Ddc has a degree of tolerance in the optimization of 10^{-5}.

4 non-Hispanic Black population in Los Angeles and New York: `segregation` application

Segregation in US counties, and metropolitan areas in general, has been object study for a vast literature. Allen and Turner (2012) used the D index to many US counties to assess Black-White and Hispanic-White segregation. Also Massey and Tannen (2015) made a vast metropolitan study for a 40-year period on hypersegregation of black population. More recently, Clark and Östh (2018) studied ethnic residential segregation of metropolitan regions of California using a different type of spatial isolation.

In this section we rely on SM to perform several segregation measures for Los Angeles County, CA, and New York City[20], NY, census data tract level for non-Hispanic black population (`nhblk`).[21] It is of interest to inspect how segregated Los Angeles along all five dimensions (evenness, isolation, clustering, concentration and centralization) using all indexes available to making point estimation and inference for 2010. For comparisons, this section studies the evolution of these estimates for Los Angeles county between 2000 and 2010 (two cross-sections in two times) and, in addition, make the New York comparison for the year of 2010 (one cross-section for two spatial contexts).[22]

In Figure 2 we can see the spatial distribution pattern of `nhblk` between the tracts of Los Angeles where the color gradient represent the relative percentage of `nhblk` within each tract (`nhblk` divided by total tract population), i. e., the composition. There is a clear pattern of spatial concentration and unevenness in terms of frequency and, therefore, a segregation regional analysis is reasonable to perform. It is worth to highlight the unusual spatial distribution of census tract along Los Angeles County where it is heavily affected by an asymmetry of tracts areas. This might affect the spatial estimates as well as the inference for spatial measures.

Figure 3 presents the simulations for each measure under different null hypothesis. These

[20]Composed by five counties: New York County, Bronx County, Kings County, Queens County and Richmond County.

[21]Both regions are close in terms of number of spatial units, as Los Angeles County has 2346 census tracts in 2010 and New York City has 2168.

[22]Once again, all simulation were run using the default values of the input parameters and 500 iterations in parallel with 6 cores in a Jupyter Notebook (Kluyver et al., 2016) using an Intel (R) Core (TM) i7-8750H CPU with 2.21GHz and 16GB of RAM. It was necessary approximately 34.7 hours to run all application results here presented.

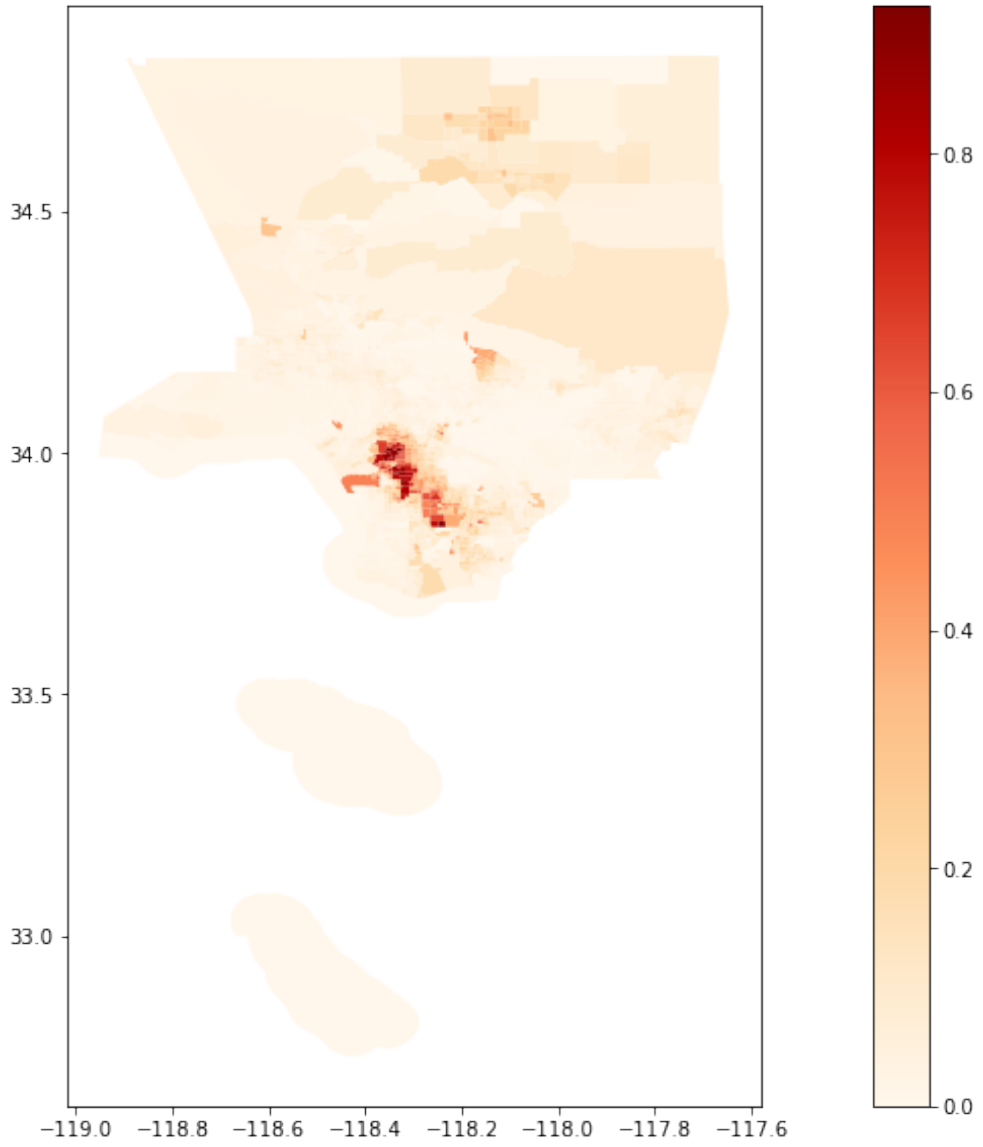


Figure 2: non-Hispanic Black population (`nhblk`) in Los Angeles county composition in 2010

graphs have the distribution under the null hypothesis as a blue density curve and a vertical red line dot that represents the point estimation of the measure. In addition, the value of each segregation measure is highlighted in each title.

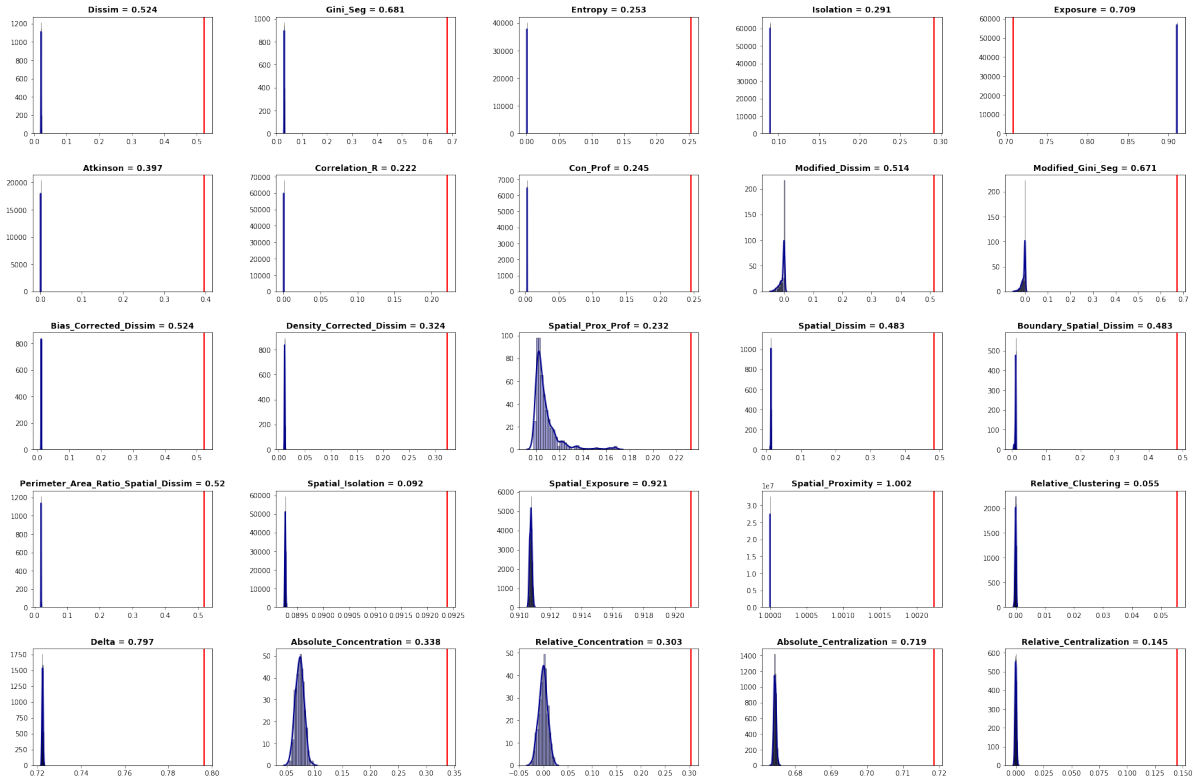
In Subfigure 3a the simulations were draw assuming a multinomial distribution with no systematic segregation. It is clear the unusual behavior of the distributions when comparing to the actual value estimated from the data. Basically, all 25 measures are highly positive significantly, with the exception of the Exposure index. The majority of the distributions present values close to zero which is in accordance to the mathematical property of some measures that assumes zero when there is no segregation in the data. Subfigure 3b present the current 13 spatial segregation measures under the spatial permutation approach.[23] In this case, the statistical significance of each measure are not as highlighted as it was previously. It is possible to notice that the Spatial Proximity Profile (p-value ≈ 0.068), the Absolute Concentration (p-value ≈ 0.272) and the Relative Concentration (p-value ≈ 0.184) present values that may not be significant in a statistical perspective. However, it is possible to see that even the distributions are closer to the original values represented in the red line, all measures, except those three previous mentioned, are highly statistically significant (p-values < 0.001), reinforcing that Los Angeles is, indeed, segregated in terms of non-Hispanic black population.

One of the major contributions of SM is the ability to easy assess segregation difference between two distinct measures. If Los Angeles county was statistically segregated in 2010, a question that may rise is "Is Los Angeles County more or less segregated in 2010 than in 2000?"[24]. Figure 4 depicts the composition spatial distribution of this county using census data of 2000. Despite the very similarities, it is possible to notice that this is slightly different from the one presented in Figure 2 of 2010. The `nhblk` composition did not change in the most concentrated part of the map, but the outskirts of this highlighted region presented changes.

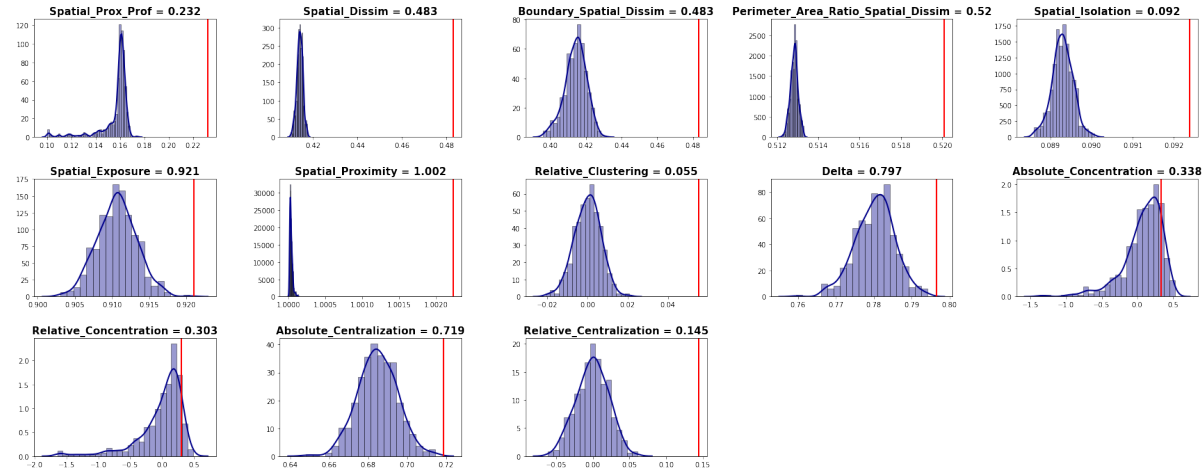
To assess the statistical significance of the Los Angeles county evolution over this decade, we rely on the `CompareSegregation` function of SM with the `random_label` approach. Figure 5 represent the results for the difference between 2000 and 2010 for each of the measures. In general, it is clear from the graph that the year of 2000 was, actually, more segregated than 2010 since the

[23] This approach does not apply to measures that do not take spatial context into consideration since each value for the simulations would be the same along the permutations.

[24] $H_0 : LosAngeles\ Segregation_{2010} - LosAngeles\ Segregation_{2000} = 0$



(a) `systematic` null approach



(b) `permutation` null approach

Figure 3: Simulations using SM for non-Hispanic Black population (`nhblk`) in Los Angeles in 2010. The point estimation of each segregation measure in presented in each title. Here, Distance Decay Isolation/Exposure are named Spatial Isolation/Exposure.

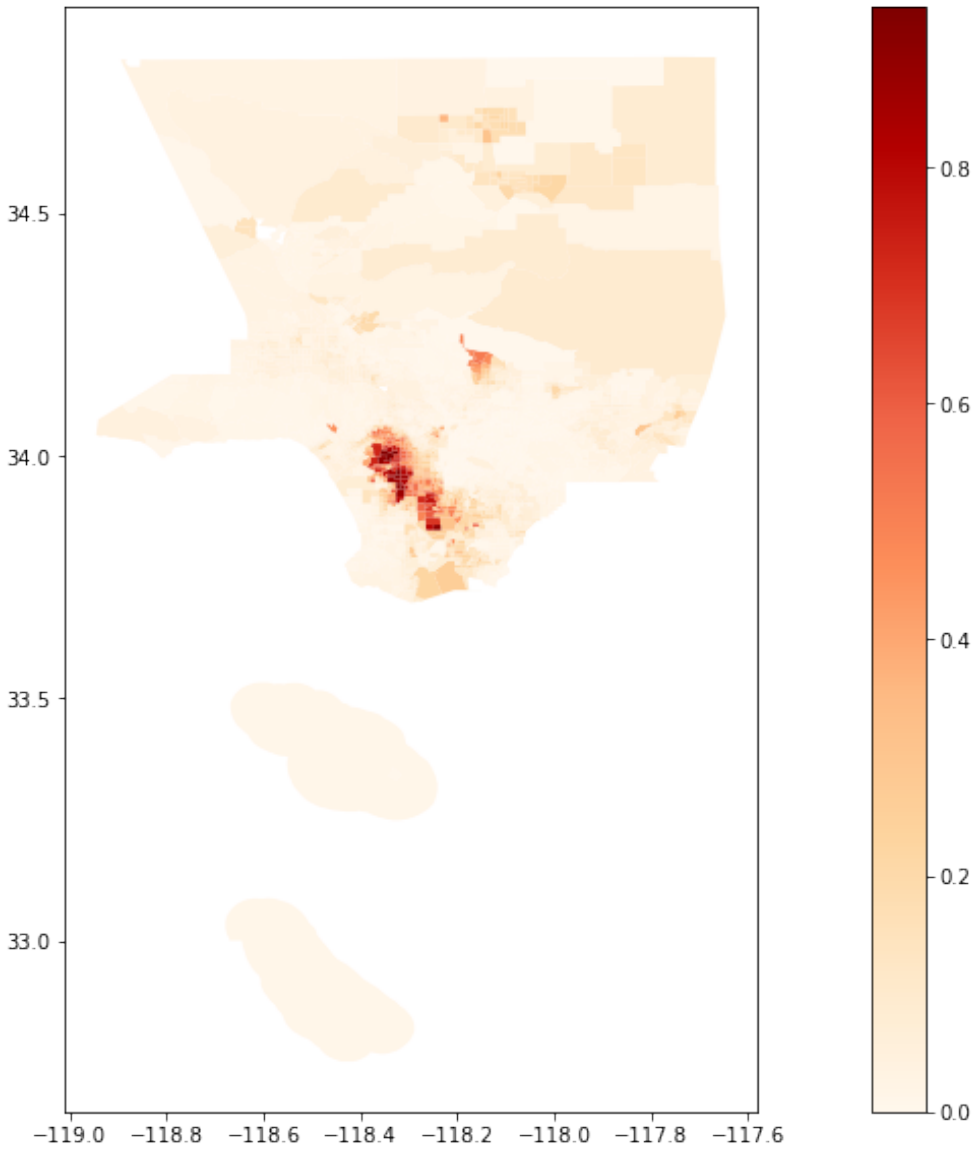


Figure 4: non-Hispanic Black population (`nhblk`) in Los Angeles county composition in 2000

majority of vertical red lines are located on negative values. Moreover, for almost all segregation measures available these difference values seem to be statistically significant since they are on the far left tail of each distribution.[25]

However, some particularities shall be pointed. For two of the concentration dimensions (ACO and RCO) Los Angeles showed to be more not statistically significant.[26] Also, the same non-significant difference was indicated by RCE (p-value ≈ 0.136) and, in part, by ACE (p-value ≈ 0.022). These results make sense with what was discussed when comparing the composition spatial distribution of both maps. There was no visual difference in terms of concentration and centralization of `nhblk` as both maps presented the same hotspot in 2000 and in 2010. Also, under the same argument, it is worth to mention the non-significance of the Spatial Proximity Profile (p-value ≈ 0.096), related to the clustering dimension of segregation.

The ability to make comparisons between regions is also possible with SM. Regarding this, since the `CompareSegregation` function can generic handle two classes previous fitted, a user can, therefore, pass two segregation measures from two different spatial contexts. Figure 6 present the New York City which is, unlike Los Angeles, located at the east coast of US.

The composition of New York has a unique pattern that contrasts with Los Angeles. The former presents multiple hotspots of `nhblk` people mostly concentrated in the Kings County (center of the map), in part of the Queens County (east side of the map) and, with less intensity, in the Bronx County (north of the map).

A feasible question of research in social science would arrive to check statistically significance difference between these two regions in terms of segregation. To shed light on this question, Figure 7 depicts the comparison for both cities[27] for 2010 census tract data for all measures using the `random_label` approach.

From this graph, we can see that all indexes (with the exception of ACO, RCO and ACE) resulted in significant values. For an expressive number of measures (D, G, H, xPx, A, V, R, Dct, Gct, Dbc, SPP, SD, BSD, DDxPx and DEL) New York presented higher values of segregation.[28] This indicates that New York is, in general, more segregated than Los Angeles in terms of the

[25] With the caveat that the Exposure is inversely proportional of the segregation and, thus, it's located on the right-tail of the distribution under null hypothesis.

[26] The p-value of ACO was ≈ 0.74 and of RCO was ≈ 0.816.

[27] $H_0 : Los\ Angeles\ Segregation - New\ York\ Segregation = 0$

[28] For the xPy and DDxPy it presented lower values, but the interpretation is the same.

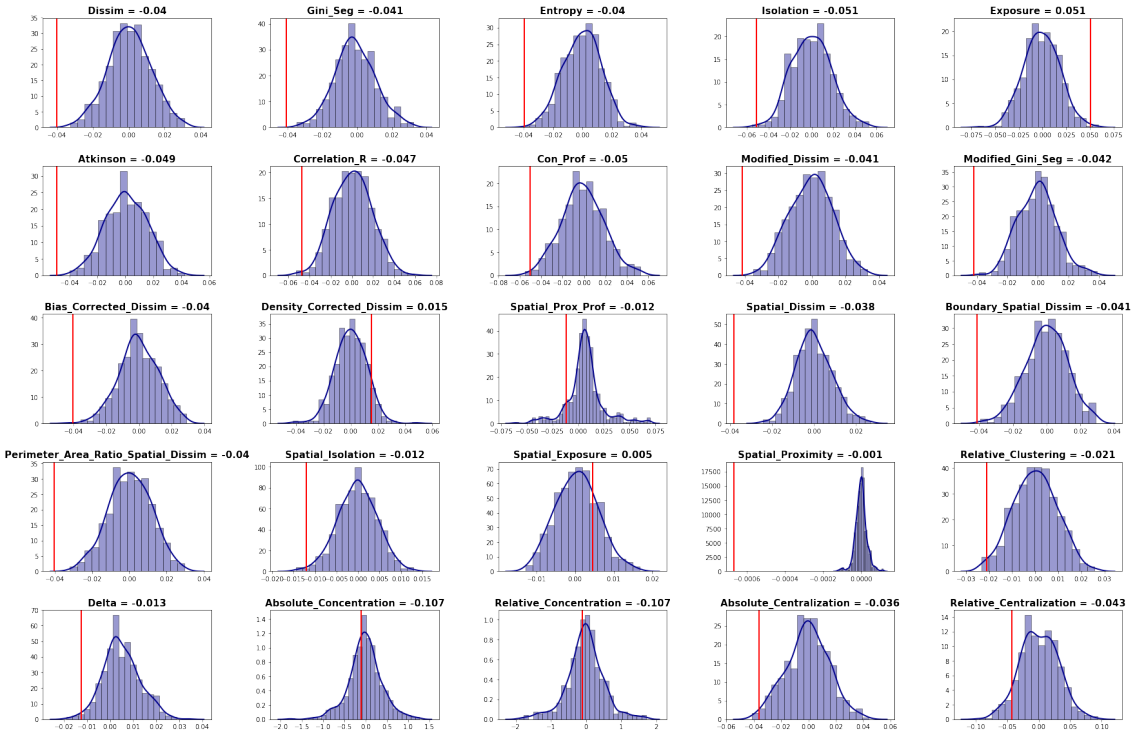


Figure 5: Simulations using SM for Los Angeles comparison between 2000 and 2010 using the `random_label` null approach. The point estimation of the **difference** of each segregation measure in presented in each title. Here, Distance Decay Isolation/Exposure are named Spatial Isolation/Exposure.

`nhblk` people.[29] On the other hand, some interesting results can also be pointed in terms of the clustering and centralization dimensions for some measures. Is was possible to see that Los Angeles is more clustered (in terms of SP and RCL) and more centralized (in terms of ACE, which resulted in a p-value ≈ 0.06, and RCE) which is in consent with the map comparison discussion that it is more concentrated in a single `nhblk` hotspot, unlike New York which presents multiple hotspots in its composition.

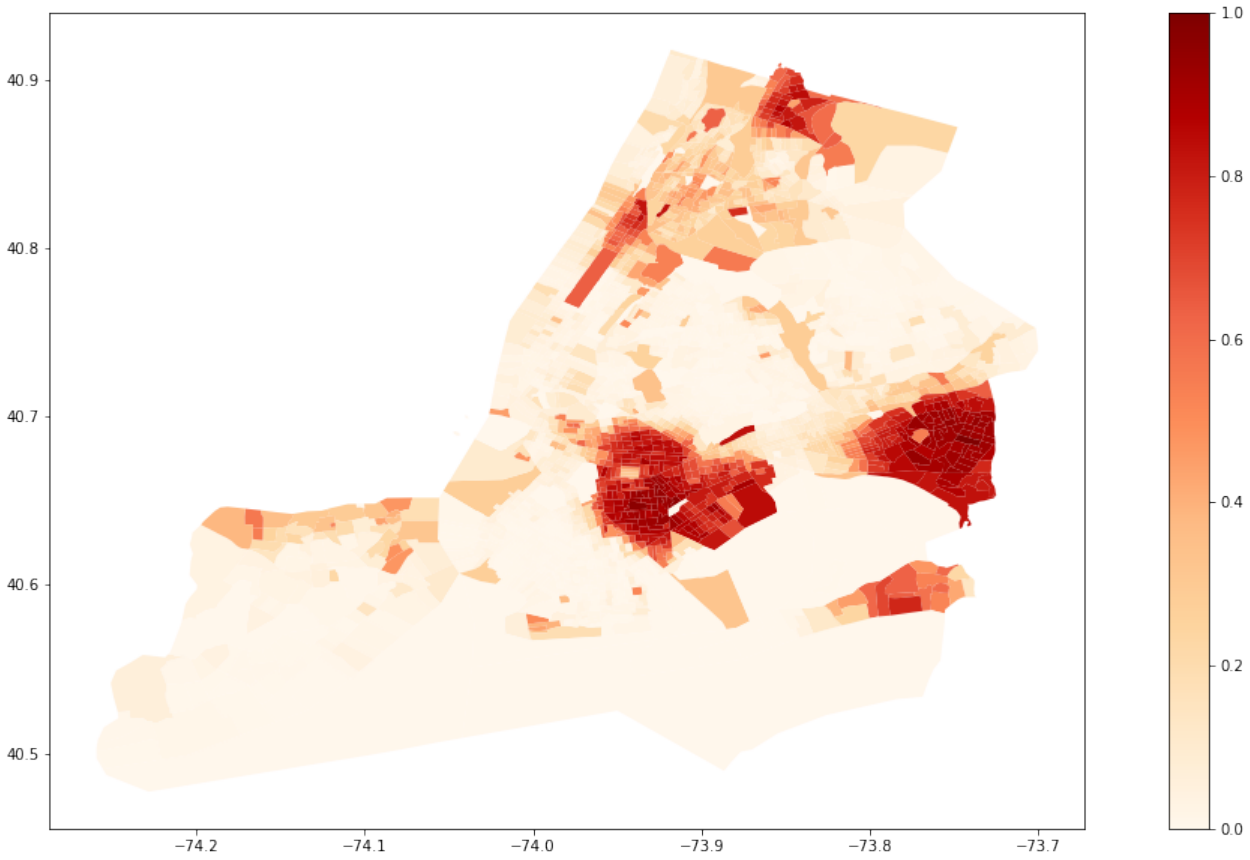


Figure 6: non-Hispanic Black population (`nhblk`) in New York composition in 2010

This unexpected result highlights the importance of comparative inference for segregation measures to be dependent of what dimensions it is under evaluation. One might argue that a place is highly segregated than another, but this might be not true if you change the dimension perspective of segregation. The same behavior can arise when comparing the same city for two

[29]However, an unexpected result arose from the fact that for the Ddc index Los Angeles was, significantly, more segregated.

distinct periods as what happened with ACO and RCO, for example, for Los Angeles County in 2010 versus itself in 2000.

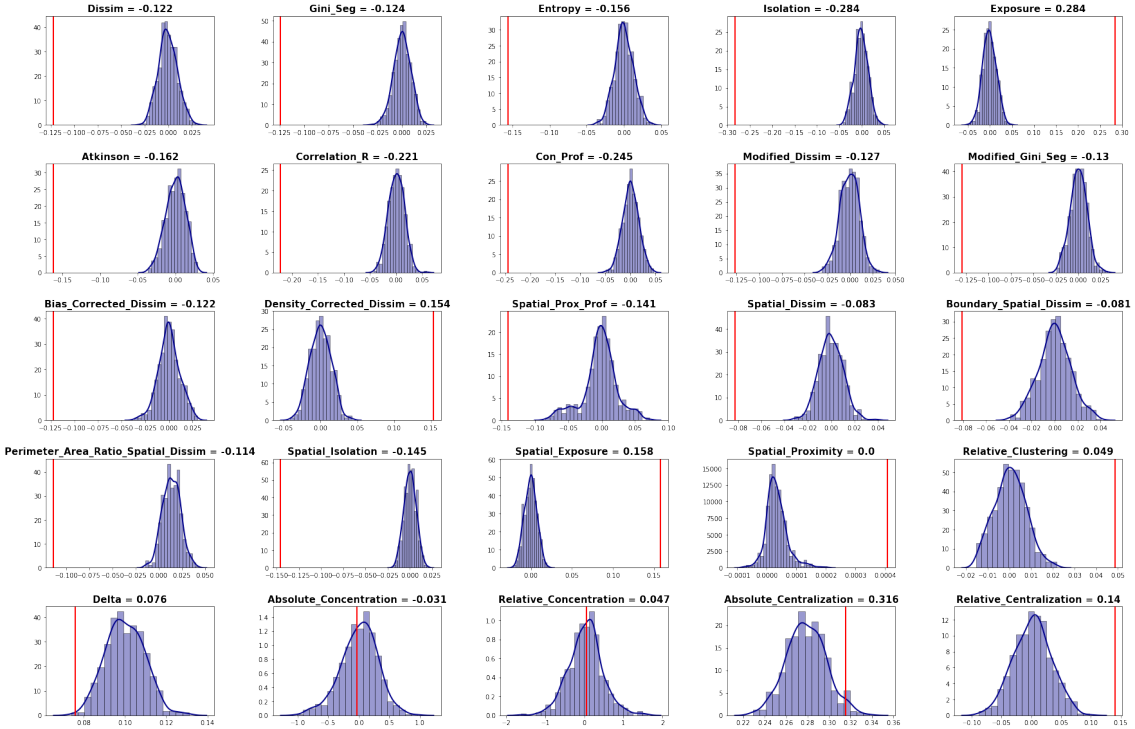


Figure 7: Simulations using SM for Los Angeles and New York comparison in 2010 using the `random_label` null approach. The point estimation of the **difference** of each segregation measure in presented in each title. Here, Distance Decay Isolation/Exposure are named Spatial Isolation/Exposure.

5 Conclusion

Segregation measurements have a vast literature and an extensive use since the first half of the 20th century. This field is constantly under progress with increasingly works discussing the properties of the indexes, better ways to overcome limitations, illustrate applications, etc. This work is an attempt to make an advance in the use of segregation measure through an open-source framework under the `PySAL` ecosystem - the, so called `PySAL segregation` module (SM). Moreover, our contribution is not restricted to ease the assess of several well-known non-spatial and spatial segregation measures, but also to build a consistent inference framework software for them never

considered before.

We provide a flexible way to estimate non-spatial and spatial segregation, perform inference for testing the significance of a single value or for comparative values. Each measure of SM has its own function that depends on the nature of the index for the data type and parameters inputs. Also, two main functions depict the inference for testing framework: the `InferSegregation` and `CompareSegregation`. Each one of these represent a wrapper function for the segregation classes fitted previously, where the first is used to perform inference for a single measure, while the second allow comparison between two measures. Both of them relying on simulations under the null hypothesis chosen.

As an illustration, Los Angeles County and New York City were used to perform regional segregation analysis using census tract data. We studied the degree of how non-Hispanic black population was in 2010 by inspecting the significance of each of the measures and concluding that it was, indeed, statistically significant for all measures, even assuming different approaches for the null hypothesis. In order to illustrate the `CompareSegregation`, two types of comparisons were made: same space between two periods and two spaces for the same period. The former assess the evolution of Los Angeles between 2000 and 2010 concluding that it was statistically more segregated in the past and the latter compared Los Angeles and New York and concluded that, in general, the latter city is statistically more segregated than the former, although some differences might be considered for specific dimensions of segregation.

This PySAL module is actively under development and some new features and functionalities were developed recently. To cite some of the topics not covered here, SM currently has a set of multigroup segregation measures, a set of local segregation measures, new approaches for the null hypothesis of the inference wrappers, a decomposition framework and an innovative street network based segregation measures. The first feature is based mostly in Reardon and Firebaugh (2002), the second draw inspiration from Tivadar (2019), the new inference approaches include the *bootstrap* for single value measures and different way to generate the counterfactual distributions for comparative segregation, the decomposition framework is based on Shapley (1953) and, finally, the street network based measures draw inspiration from Roberto (2018) and uses a handful of libraries from the Urban Data Science Toolkit[30]. The combination of all functionalities present in

[30]https://github.com/UDST

this paper with all these other features mentioned, we are confident that the current module is one of the most complete tools to deal with segregation currently available.

Additionally, several aspects are still to be explored. Possible extensions comprise more measures that can be added such as the Proportion of Central City number (PCC) (Massey and Denton, 1988), other indices present in Tivadar (2019) and the parametric and nonparametric approach of the class of indexes of, respectively, Rathelot (2012) and d'Haultfoeuille and Rathelot (2017). Another landscape of opportunity is not only "zone-based" measures, but also "surface-based" methods as quoted in Hong et al. (2014). In this regard, spatial counterfactual approaches (Carrillo and Rothbaum, 2016) can be considered to develop alternatives for the inference framework that could rely on the counterfactual distribution between two measures. Currently, the street network based measures already deal with this kind of data.

To conclude, we believe that this current work will represent a useful option for social scientist researchers. By allowing them to use a simple and friendly framework to assess segregation in many different contexts of a given group, we empower the society by broaden the use of these measures and make an important advance by combining open-source tools, urban planning, statistics and segregation.

References

Allen, J. P., Turner, E., 2012. Black–white and hispanic–white segregation in US counties. The Professional Geographer 64 (4), 503–520.

Allen, R., Burgess, S., Davidson, R., Windmeijer, F., 2015. More reliable inference for the dissimilarity index of segregation. The econometrics journal 18 (1), 40–66.

Apparicio, P., Martori, J. C., Pearson, A. L., Fournier, É., Apparicio, D., 2014. An open-source software for calculating indices of urban residential segregation. Social Science Computer Review 32 (1), 117–128.

Boisso, D., Hayes, K., Hirschberg, J., Silber, J., 1994. Occupational segregation in the multidimensional case: decomposition and tests of significance. Journal of Econometrics 61 (1), 161–171.

Carrillo, P. E., Rothbaum, J. L., 2016. Counterfactual spatial distributions. Journal of Regional Science 56 (5), 868–894.

Carrington, W. J., Troske, K. R., 1997. On measuring segregation in samples with small units. Journal of Business & Economic Statistics 15 (4), 402–409.

Carrington, W. J., Troske, K. R., 1998. Interfirm segregation and the black/white wage gap. Journal of Labor Economics 16 (2), 231–260.

Clark, W. A., Östh, J., 2018. Measuring isolation across space and over time with new tools: Evidence from californian metropolitan regions. Environment and Planning B: Urban Analytics and City Science, 2399808318756642.

Devroye, L., 1986. Sample-based non-uniform random variate generation. En: Proceedings of the 18th conference on Winter simulation. ACM, pp. 260–265.

d'Haultfoeuille, X., Rathelot, R., 2017. Measuring segregation on small units: A partial identification analysis. Quantitative Economics 8 (1), 39–73.

Duncan, O. D., Duncan, B., 1955. A methodological analysis of segregation indexes. American sociological review 20 (2), 210–217.

Hellerstein, J. K., Neumark, D., 2008. Workplace segregation in the united states: Race, ethnicity, and skill. The review of economics and statistics 90 (3), 459–477.

Hong, S.-Y., O'Sullivan, D., Sadahiro, Y., 2014. Implementing spatial segregation measures in r. PloS one 9 (11), e113767.

Hong, S.-Y., Sadahiro, Y., 2014. Measuring geographic segregation: a graph-based approach. Journal of Geographical Systems 16 (2), 211–231.

Hunter, J. D., 2007. Matplotlib: A 2d graphics environment. Computing In Science & Engineering 9 (3), 90–95.
DOI: 10.1109/MCSE.2007.55

Jordahl, K., 2014. Geopandas: Python tools for geographic data. URL: https://github.com/geopandas/geopandas.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., et al., 2016. Jupyter notebooks-a publishing format for reproducible computational workflows. En: ELPUB. pp. 87–90.

Massey, D. S., Denton, N. A., 1988. The dimensions of residential segregation. Social forces 67 (2), 281–315.

Massey, D. S., Denton, N. A., 1989. Hypersegregation in us metropolitan areas: Black and hispanic segregation along five dimensions. Demography 26 (3), 373–391.

Massey, D. S., Denton, N. A., 1993. American apartheid: Segregation and the making of the underclass. Harvard University Press.

Massey, D. S., Tannen, J., 2015. A research note on trends in black hypersegregation. Demography 52 (3), 1025–1034.

McKinney, W., 2011. pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing, 1–9.

Morgan, B. S., 1983. A distance-decay based interaction index to measure residential segregation. Area, 211–217.

Morrill, R. L., 1991. On the measure of geographic segregation. En: Geography research forum. Vol. 11. pp. 25–36.

Napierala, J., Denton, N., 2017. Measuring residential segregation with the acs: How the margin of error affects the dissimilarity index. Demography 54 (1), 285–309.

Park, R. E., 1926. The urban community as a spatial pattern and a moral order. Urban social segregation, 21–31.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL: `http://www.R-project.org`

Ransom, M. R., 2000. Sampling distributions of segregation indexes. Sociological Methods & Research 28 (4), 454–475.

Rathelot, R., 2012. Measuring segregation when units are small: a parametric approach. Journal of Business & Economic Statistics 30 (4), 546–553.

Reardon, S. F., 2002. Seg: Stata module to compute multiple-group diversity and segregation indices.

Reardon, S. F., Firebaugh, G., 2002. Measures of multigroup segregation. Sociological methodology 32 (1), 33–67.

Rey, S. J., 2004. Spatial analysis of regional income inequality. Spatially integrated social science 1, 280–299.

Rey, S. J., Anselin, L., 2010. PySAL: A Python library of spatial analytical methods. En: Handbook of applied spatial analysis. Springer, pp. 175–193.

Rey, S. J., Sastré-Gutiérrez, M. L., 2010. Interregional inequality dynamics in mexico. Spatial Economic Analysis 5 (3), 277–298.

Roberto, E., 2018. The spatial proximity and connectivity method for measuring and analyzing residential segregation. Sociological Methodology 48 (1), 182–224.

Rossum, G., 1995. Python reference manual. Tech. rep., Amsterdam, The Netherlands, The Netherlands.

Royuela, V., Vargas, M., et al., 2010. Residential segregation: A literature review. Tech. rep.

Shapley, L. S., 1953. A value for n-person games. Contributions to the Theory of Games 2 (28), 307–317.

Söderström, M., Uusitalo, R., 2010. School choice and segregation: evidence from an admission reform. Scandinavian Journal of Economics 112 (1), 55–76.

Tivadar, M., 2019. Oasisr: An R package to bring some order to the world of segregation measurement. Journal of Statistical Software 89 (1), 1–39.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer,

S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., Qalieh, A., Sep. 2017. mwaskom/seaborn: v0.8.1 (september 2017).
URL: https://doi.org/10.5281/zenodo.883859
DOI: 10.5281/zenodo.883859

Wong, D. W., 1993. Spatial indices of segregation. Urban studies 30 (3), 559–572.

Wong, D. W., 2003. Implementing spatial segregation measures in gis. Computers, Environment and Urban Systems 27 (1), 53–70.

A Point Estimation details

Here, we present and explain each formula for the segregation measures presented in Table 1 of Section 2.1. The respective literature used for each measure can be found in Table 3[31,32] in addition with the respective dimension.

For consistency of notation, we assume that n_{ij} is the population of unit $i \in \{1, ..., I\}$ of group $j \in \{x, y\}$, also $\sum_j n_{ij} = n_{i.}$, $\sum_i n_{ij} = n_{.j}$, $\sum_i \sum_j n_{ij} = n_{..}$, $\tilde{s}_{ij} = \frac{n_{ij}}{n_{i.}}$, $\hat{s}_{ij} = \frac{n_{ij}}{n_{.j}}$. The segregation indexes can be build for any group j of the data.

The Dissimilarity Index (D) is given by:

$$D = \sum_{i=1}^{I} \frac{n_{i.} \mid \tilde{s}_{ij} - \frac{n_{.j}}{n_{..}} \mid}{2n_{..} \frac{n_{.j}}{n_{..}} \left(1 - \frac{n_{.j}}{n_{..}}\right)}. \tag{1}$$

The spatial D (SD) is given by:

$$SD = D - \frac{\sum_{i_1=1}^{I} \sum_{i_2=1}^{I} \left|\tilde{s}_{ij}^{i_1} - \tilde{s}_{ij}^{i_2}\right| c_{i_1 i_2}}{\sum_{i_1=1}^{I} \sum_{i_2=1}^{I} c_{i_1 i_2}} \tag{2}$$

[31] This table doesn't reflect necessarily the original/pioneer paper of each measure, but rather the related literature of the formulas presented in this Appendix.

[32] We considered to include the mixture of betas approach of Rathelot (2012) for the D, G and H indexes, as the author kindly shared the original code. However, due to convergence problems we chose not to include it in the current version of SM.

Table 3: Segregation Measures related literature for PySAL `segregation` module point estimations

Measure	**Related literature of the function**	**Dimension**
Dissimilarity (D)	Massey and Denton (1988)	Evenness
Gini (G)	Massey and Denton (1988)	Evenness
Entropy (H)	Massey and Denton (1988)	Evenness
Isolation (xPx)	Massey and Denton (1988)	Isolation
Exposure (xPy)	Massey and Denton (1988)	Isolation
Atkinson (A)	Massey and Denton (1988)	Evenness
Correlation Ratio (V)	Massey and Denton (1988)	Isolation
Concentration Profile (R)	Hong and Sadahiro (2014)	Evenness
Modified Dissimilarity (Dct)	Carrington and Troske (1997)	Evenness
Modified Gini (Gct)	Carrington and Troske (1997)	Evenness
Bias-Corrected Dissimilarity (Dbc)	Allen et al. (2015)	Evenness
Density-Corrected Dissimilarity (Ddc)	Allen et al. (2015)	Evenness
Spatial Proximity Profile (SPP)	Hong and Sadahiro (2014)	Clustering
Spatial Dissimilarity (SD)	Morrill (1991)	Evenness
Boundary Spatial Dissimilarity (BSD)	Hong et al. (2014)	Evenness
Perimeter Area Ratio Spatial Dissimilarity (PARD)	Wong (1993)	Evenness
Distance Decay Isolation (DDxPx)	Morgan (1983)	Isolation
Distance Decay Exposure (DDxPy)	Morgan (1983)	Isolation
Spatial Proximity (SP)	Massey and Denton (1988)	Clustering
Relative Clustering (RCL)	Massey and Denton (1988)	Clustering
Delta (DEL)	Massey and Denton (1988)	Concentration
Absolute Concentration (ACO)	Massey and Denton (1988)	Concentration
Relative Concentration (RCO)	Massey and Denton (1988)	Concentration
Absolute Centralization (ACE)	Massey and Denton (1988)	Centralization
Relative Centralization (RCE)	Massey and Denton (1988)	Centralization

where $\tilde{s}_{ij}^{i_1}$ and $\tilde{s}_{ij}^{i_2}$ are the proportions of the minority population in the units i_1 and i_2, respectively and where $c_{i_1 i_2}$ denotes an element at (i_1, i_2) in a matrix C, which becomes one only if i_1 and i_2 are considered neighbors.

The boundary spatial D (BSD) is given by:

$$BSD = D - \frac{1}{2}\sum_{i_1=1}^{I}\sum_{i_2=1}^{I} w_{i_1 i_2} \left|\tilde{s}_{ij}^{i_1} - \tilde{s}_{ij}^{i_2}\right| \tag{3}$$

where

$$w_{i_1 i_2} = \frac{cb_{i_1 i_2}}{\sum_{i_2=1}^{I} d_{i_1 i_2}}$$

where $\tilde{s}_{ij}^{i_1}$ and $\tilde{s}_{ij}^{i_2}$ are the proportions of the minority population in the units i_1 and i_2, respectively, and $cb_{i_1 i_2}$ is the length of the common boundary of areal units i_1 and i_2.

The perimeter/area ratio Spatial D (PARD) is a spatial dissimilarity index that takes into consideration the perimeter and the area of each unit by adding a specific multiplicative term in the second term of BSD (the spatial effect):

$$\frac{\frac{1}{2}\left[\left(\frac{P_i}{A_i}\right) + \left(\frac{P_j}{A_j}\right)\right]}{MAX\left(\frac{P}{A}\right)} \tag{4}$$

where P_i and A_i are the perimeter and area of unit i, respectively and $MAX(P/A)$ is the maximum perimeter-area ratio or the minimum compactness of an areal unit found in the study region.

The Gini coefficient (G) is given by:

$$G = \sum_{i_1=1}^{I}\sum_{i_2=1}^{I} \frac{n_{i_1 .} n_{i_2 .} \mid \tilde{s}_{ij}^{i_1} - \tilde{s}_{ij}^{i_2} \mid}{2n_{..}^2 \frac{n_{.j}}{n_{..}}\left(1 - \frac{n_{.j}}{n_{..}}\right)} \tag{5}$$

The global entropy (E) is given by:

$$E = \frac{n_{.j}}{n_{..}}\, log\left(\frac{1}{\frac{n_{.j}}{n_{..}}}\right) + \left(1 - \frac{n_{.j}}{n_{..}}\right) log\left(\frac{1}{1 - \frac{n_{.j}}{n_{..}}}\right) \tag{6}$$

while the unit's entropy is analogously:

$$E_i = \tilde{s}_{ij}\, log\left(\frac{1}{\tilde{s}_{ij}}\right) + (1 - \tilde{s}_{ij})\, log\left(\frac{1}{1 - \tilde{s}_{ij}}\right). \tag{7}$$

Therefore, the entropy index (H) is given by:

$$H = \sum_{i=1}^{I} \frac{n_{i.}\left(E - E_i\right)}{E n_{..}} \tag{8}$$

The Atkinson index (A) is given by:

$$A = 1 - \frac{\frac{n_{.j}}{n_{..}}}{1 - \frac{n_{.j}}{n_{..}}} \left| \sum_{i=1}^{I} \left[\frac{\left(1 - \tilde{s}_{ij}\right)^{1-b} \tilde{s}_{ij}^{b} t_i}{\frac{n_{.j}}{n_{..}} n_{..}} \right] \right|^{\frac{1}{1-b}} \tag{9}$$

where b is a shape parameter that determines how to weight the increments to segregation contributed by different portions of the Lorenz curve.

The Concentration Profile (R) measure is discussed in Hong and Sadahiro (2014) and tries to inspect the evenness aspect of segregation. The threshold proportion t is given by:

$$\upsilon_t = \frac{\sum_{i=1}^{I} n_{ij} g(t,i)}{\sum_{i=1}^{I} n_{ij}}. \tag{10}$$

In the equation, $g(t,i)$ is a logical function that is defined as:

$$g(t,i) = \begin{cases} 1 & \text{if } \frac{n_{ij}}{n_{i.}} \geqslant t \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The Concentration Profile (R) is given by:

$$R = \frac{\frac{n_{.j}}{n_{..}} - \left(\int_{t=0}^{\frac{n_{.j}}{n_{..}}} \upsilon_t dt - \int_{t=\frac{n_{.j}}{n_{..}}}^{1} \upsilon_t dt \right)}{1 - \frac{n_{.j}}{n_{..}}}. \tag{12}$$

The spatial proximity profile (SPP) is similar to the Concentration Profile, but with the addition of the spatial component in the connecting function.

$$\eta_t = \frac{k^2 - k}{\sum_{i_1} \sum_{1_2} \delta_{i_1 i_2}} \tag{13}$$

where k refers to the sum of $g(t,i)$ for a given t and δ_{ij} is the distance between i_1 and i_2. One way of determining $\delta_{i_1 i_2}$ would be to use a spatial structure matrix, W. The matrix W present ones if i_1 and i_2 are contiguous and zero, otherwise. The distance $\delta_{i_1 i_2}$ between i_1 and i_2 is given by is the order of how neighbors is needed to reach from i_1 to i_2. For example, two census tracts, x_1 and x_2, that do not have a common boundary but both are adjacent to the same unit, x_3,

are second-order neighbors, so δ_{12} becomes 2. Like the Concentration Profile, if the number of thresholds used is large enough, a smooth curve, or a *spatial proximity profile*, can be constructed by plotting and connecting η_t.

Isolation (xPx) assess how much a minority group is only exposed to the same group. In other words, how much they only interact the members of the group that they belong. Assuming $j = x$ as the minority group, the isolation of x is giving by:

$$xPx = \sum_{i=1}^{I} \left(\hat{s}_{ix}\right)\left(\tilde{s}_{ix}\right). \tag{14}$$

The Exposure (xPy) of x is giving by

$$xPy = \sum_{i=1}^{I} \left(\hat{s}_{iy}\right)\left(\tilde{s}_{iy}\right). \tag{15}$$

The correlation ratio (V or Eta^2) is given by

$$V = Eta^2 = \frac{xPx - \frac{n_{.x}}{n_{..}}}{1 - \frac{n_{.x}}{n_{..}}}. \tag{16}$$

The Spatial Proximity Index (SP) is given by:

$$SP = \frac{XP_{xx} + YP_{yy}}{TP_{tt}} \tag{17}$$

where

$$P_{xx} = \sum_{i_1=1}^{I}\sum_{i_2=1}^{I} \frac{n_{i_1x}n_{i_2x}\zeta_{i_1i_2}}{n_{.x}^2}$$

$$P_{yy} = \sum_{i_1=1}^{I}\sum_{i_2=1}^{I} \frac{n_{i_1y}n_{i_2y}\zeta_{i_1i_2}}{n_{.y}^2}$$

$$P_{tt} = \sum_{i_1=1}^{I}\sum_{i_2=1}^{I} \frac{n_{i_1.}n_{i_2.}\zeta_{i_1i_2}}{n_{..}^2}$$

$$\zeta_{i_1i_2} = exp(-d_{i_1i_2})$$

$d_{i_1i_2}$ is a pairwise distance measure between area i_1 and i_2 and d_{ii} is estimated as $d_{ii} = (\alpha a_i)^{\beta}$ where a_i is the area of unit i. The default is $\alpha = 0.6$ and $\beta = 0.5$ and for the distance measure, we first extracts the centroid of each unit and calculate the euclidean distance.

The Relative Clustering Measure (RCL) is given by:

$$RCL = \frac{P_{xx}}{P_{yy}} - 1 \tag{18}$$

The Distance Decay Isolation (DDxPx) is given by:

$$DDxPx = \sum_{i_1=1}^{I} (\hat{s}_{i_1 x}) \left(\sum_{i_2=1}^{I} P_{i_1 i_2} (\tilde{s}_{i_1 x}) \right) \tag{19}$$

where

$$P_{i_1 i_2} = \frac{\zeta_{i_1 i_2} n_{i_2 .}}{\sum_{i_2=1}^{I} \zeta_{i_1 i_2} n_{i_2 .}}$$

such that

$$\sum_{i_2=1}^{I} P_{i_1 i_2} = 1.$$

where $\zeta_{i_1 i_2}$ is defined as before. This also could be seen as the *probability of contact* of members of group x to each other weighted by the inverse of distance.

The Distance Decay Exposure (DDxPy) is given by:

$$DDxPy = \sum_{i_1=1}^{I} (\hat{s}_{i_1 x}) \left(\sum_{i_2=1}^{I} P_{i_1 i_2} (\tilde{s}_{i_1 y}) \right) \tag{20}$$

where $P_{i_1 i_2}$ is defined as before.

The Delta (DEL) measure is given by the following equation:

$$DEL = \frac{1}{2} \sum_{i=1}^{I} \left| \hat{s}_{ij} - \frac{a_i}{A} \right| \tag{21}$$

where a_i is the area of unit i and A is the total area of the given region $A = \sum_{i=1}^{I} a_i$.

The Absolute Concentration Index (ACO) is given by:

$$ACO = 1 - \frac{\sum_{i=1}^{I} \left(\frac{n_{ij} a_i}{n_{.j}} \right) - \sum_{i=1}^{n_1} \left(\frac{n_{i.} a_i}{T_1} \right)}{\sum_{i=n_2}^{I} \left(\frac{n_{i.} a_i}{T_2} \right) - \sum_{i=1}^{n_1} \left(\frac{n_{i.} a_i}{T_1} \right)} \tag{22}$$

where the units are **ordered** from **smallest to largest** in areal size. In this formula, n_1 is the rank of the unit where the cumulative total population equal the total minority population, n_2 is

the rank of the unit where cumulative total population equal equal the total minority population **from the largest unit down**. In addition,

$$T_1 = \sum_{i=1}^{n_1} n_{i.}$$

and

$$T_2 = \sum_{i=n_2}^{n} n_{i.}$$

Another measure of concentration is the Relative Concentration Index (RCO).

$$RCO = \frac{\frac{\sum_{i=1}^{I}\left(\frac{n_{ix}a_i}{n_{.x}}\right)}{\sum_{i=1}^{I}\left(\frac{n_{iy}a_i}{n_{.y}}\right)} - 1}{\frac{\sum_{i=1}^{n_1}\left(\frac{n_{i.}a_i}{T_1}\right)}{\sum_{i=n_2}^{I}\left(\frac{n_{i.}a_i}{T_2}\right)} - 1} \tag{23}$$

where n_1, n_2, T_1 and T_2 are defined as before.

The degree of centralization can be evaluated through the Absolute Centralization Index (ACE) or through the Relative Centralization Index (RCE):

$$ACE = \left(\sum_{i=2}^{I} X_{i-1}A_i\right) - \left(\sum_{i=2}^{I} X_i A_{i-1}\right) \tag{24}$$

$$RCE = \left(\sum_{i=2}^{I} X_{i-1}Y_i\right) - \left(\sum_{i=2}^{I} X_i Y_{i-1}\right) \tag{25}$$

where A_i is the cumulative area proportion through unit i, X_i is the cumulative frequency proportion through unit i of group x and Y_i is the analogous for group y. In this measure, the area units are ordered by increasing distances from the central business district, which we assume being located in the average latitude and average longitude among all centroid.

The Modified Dissimilarity Index (Dct) based on Carrington and Troske (1997) evaluates the deviation from simulated evenness. This measure is estimated by taking the mean of the classical D under several simulations under evenness from the global minority proportion.

Let D^* be the average of the classical D under simulations draw assuming evenness from the global minority proportion. The value of Dct can be evaluated with the following equation:

$$Dct = \begin{cases} \frac{D-D^*}{1-D^*} & \text{if } D \geqslant D^* \\ \frac{D-D^*}{D^*} & \text{if } D < D^* \end{cases} \tag{26}$$

Similarly, the Modified Gini (Gct) based also on Carrington and Troske (1997) evaluates the deviation from simulated evenness. This measure is estimated by taking the mean of the classical G under several simulations under evenness from the global minority proportion.

Let G^* be the average of G under simulations draw assuming evenness from the global minority proportion. The value of Gct can be evaluated with the following equation:

$$Gct = \begin{cases} \frac{G-G^*}{1-G^*} & \text{if } G \geqslant G^* \\ \frac{G-G^*}{G^*} & \text{if } G < G^* \end{cases} \tag{27}$$

Lastly, the Bias-Corrected (Dbc) and Density-Corrected (Ddc) Dissimilarities indexes are presented in Allen et al. (2015). The Dbc is given by:

$$D_{bc} = 2D - \bar{D}_b \tag{28}$$

where $\bar{D}_b$ is the average of B resampling using the observed conditional probabilities for a multinomial distribution for each group independently.

The Ddc measure is given by:

$$D_{dc} = \frac{1}{2} \sum_{i=1}^{I} \hat{\sigma}_i n\left(\hat{\theta}_i\right) \tag{29}$$

where

$$\hat{\sigma}_i^2 = \frac{\hat{s}_{ix}(1-\hat{s}_{ix})}{n_{.x}} + \frac{\hat{s}_{iy}(1-\hat{s}_{iy})}{n_{.y}}$$

and $n\left(\hat{\theta}_i\right)$ is the θ_i that maximizes the folded normal distribution $\phi(\hat{\theta}_i - \theta_i) + \phi(\hat{\theta}_i + \theta_i)$ where

$$\hat{\theta}_i = \frac{|\hat{s}_{ix} - \hat{s}_{iy}|}{\hat{\sigma}_i}.$$

and ϕ is the standard normal density.