Journal of Computational Social Science
https://doi.org/10.1007/s42001-019-00059-3

RESEARCH ARTICLE

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] · **Levi John Wolf**[2]

Received: 29 June 2019 / Accepted: 30 October 2019

Abstract

In human geography and the urban social sciences, the segregation literature typically engages with five conceptual dimensions along which a given society may be considered segregated: evenness, isolation, clustering, concentration and centralization (all of which can incorporate or omit spatial context). Over the last several decades, dozens of segregation indices have been proposed and studied in the literature, each of which is designed to focus on the nuances of a particular dimension, or correct an oversight in earlier work. Despite their increasing proliferation, however, few of these indices remain used in practice beyond their original conception, due in part to complex formulae and data requirements, particularly for indices that incorporate spatial context. Furthermore, existing segregation software typically fails to provide inferential frameworks for either single-value or comparative hypothesis testing. To fill this gap, we develop an open-source Python package designed as a submodule for the Python Spatial Analysis Library, PySAL. This new module tackles the problem of segregation point estimation for a wide variety of spatial and aspatial segregation indices, while providing a computationally based hypothesis testing framework that relies on simulations under the null hypothesis. We illustrate the use of this new library using tract-level census data in two American cities.

Keywords Open-source · Segregation · PySAL · Spatial analysis

✉ Renan Xavier Cortes
renanxcortes@gmail.com

Sergio Rey
sergio.rey@ucr.edu

Elijah Knaap
elijah.knaap@ucr.edu

Levi John Wolf
levi.john.wolf@bristol.ac.uk

[1] Center for Geospatial Sciences, University of California, Riverside, USA

[2] School of Geographical Sciences, University of Bristol, Bristol, UK

Published online: 23 November 2019

Introduction

Segregation literature is voluminous, decade spanning, and often traces its lineage to the pioneering work of Ref. [32]. The traditional antecedant to a discussion of modern segregation indices is, however, [13], thanks to its introduction of the "segregation curve", the quantitative approach that came to dominate segregation measurement methods at the time. Despite the importance of these early contributions, the vast majority of the segregation literature in recent decades begins with a discussion of Ref. [24], who formalized the concept of segregation as a multidimensional phenomenon, articulating that because the mechanisms that divide people into disparate locations of a city can take several forms (namely evenness, isolation, clustering, concentration and centralization), so too can segregation indices vary in their ability to uncover these different dimensions.[1] Over the years, each of the dimensions in Massey's taxonomy has developed something of a "champion" index, which is used predominantly in the study of that particular dimension, including well-known indices such as the Dissimilarity (D), Gini (G), Entropy (H), Isolation (xPx), Relative Concentration (RCO), Relative Centralization (RCE) and the Relative Clustering (RCL).

More recently, scholars have questioned the validity of the five dimensional classification, arguing there may be only two dimensions of segregation in reality, since concentration evenness and clustering exposure can each be viewed as a single continuum with two poles. Meanwhile, these scholars contest the validity of the centralization dimension, which relies on a subjective definition of the city center [5, 19, 38]. While this discussion is lively in the contemporary literature, for the remainder of the paper, we adopt the classic conception from Ref. [25].

Literature focused on the methodological aspects of segregation indices and their properties, specifically, is extensive.[2] Apart from proposing new indices with a variety of desirable properties, the literature is rife with discussions both about corrections and estimation issues inherent in classical indices, and their proper classification in the Massey taxonomy. [7], for instance, propose a modification to D and G indices designed to overcome overestimation issues that arise specially when enumeration units are small. Because most indices are functions of proportions, they can suffer bias arising from small samples, thus large sampling variance of the denominators. Reference [7] further argue that the G and D indices assess the distance from evenness rather than randomness. [35] also addresses this upward behavior of classical segregation indices by building a parametric approach, assuming that the frequency of a population under study is drawn from a probability distribution following a beta mixture.[3] Reference [2] also propose a bias-correction approach and a density-correction approach for D. In terms of spatial indices, References [30]

[1] For a literature review on segregation, we refer to Ref. [44]. We also refer to Refs. [10] and [18] as important literature in segregation.

[2] For application examples, see [8, 14, 26, 27, 46].

[3] More recently, Ref. [12] addressed this problem assuming a nonparametric binomial mixture of the frequencies.

and [49] propose spatial adaptations for the same classical *D* index. Recently, Ref. [16] also developed two indices, the Concentration Profile and the Spatial Proximity Profile (SPP), which similarly attempt to overcome limitations in previous versions of spatial and non-spatial segregation measures.

These discussions in the literature make clear the scholarly value inherent in each of the various indices and helps elucidate the context for which each is best suited, given a set of study parameters. The importance of their contribution to the literature notwithstanding, however, the formulaic complexity in dozens of segregation indices continues to be a major deterrent to their broader adoption in applied settings. Currently, there are a small handful of open-source platforms designed for segregation analysis, but they remain limited in both the variety of indices they can calculate and the inferential frameworks they provide (if any).

Current examples include the `seg` package of Ref. [15] for the `R` language [33] and the Geo-Segregation Analyzer (GSA) [3].[4] The former, comprises 12 measures such as the *D*, three version of modified *D*, spatial proximity (SP), 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 indices[5] for either one group, two groups, multi-group or local indices. Although GSA represents a feasible way to estimate these indices, it is less convenient for modern data science workflows or the broader ecosystem of spatial analysis, since it is isolated from other scientific computing environments and must be downloaded and installed independently for the sole purpose of segregation analysis. In addition, this option relies exclusively on the use of shapefiles which, despite being one of the most popular geographic information systems (GIS) formats for storing spatial data, is a proprietary format belonging to the Environmental Systems Research Institute (ESRI), and suffers from several well-known drawbacks.[6] Shapefiles are being phased out rapidly as the format-of-choice for spatial analysts; so reliance on shapefiles is becoming a dated and limiting factor quickly.

More recently, an important open-source contribution was made by Ref. [47] with the `OasisR` package. In this tool, a set of 50 indices is available comprising non-spatial and spatial measures, multi-group segregation measures and an inference framework for single values of segregation. Reference [47] also discusses in detail several inconsistencies in classical segregation formulas.[7] Due to the vast

[4] Table 2 of [3] cites other options of software that also put effort to calculate these indices such as Refs. [36] and [50], but not as open-source.

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

[6] Most notably shapefiles are limited to ten character column names and they are difficult to transport across computing environments because the specification is actually a minimum of four files, not a single file as the name would suggest.

[7] One of the most prominent is the indices issues presented in Ref. [49] discussed in the bottom of page 6 of Ref. [47]. During the construction of the present module, the same problems were identified and the default approach of these indices follows actually the latter study for this Python package.

number of studies and indices that are present in the literature, the `OasisR` package has emerged recently as one of the most complete options for `R` users. To our knowledge, this is the only software currently available that provides any form of statistical inference framework for single values of segregation.

As data science has risen in prominence over the last decade, the benefits of free and open-source software (FOSS) in the academic realm have become clear. This is particularly true in collaborative research environments where open-source platforms allow users full access to underlying algorithmic implementations, a critical advantage for transparency, reliability, and reproducibility; FOSS platforms also promote inclusivity by allowing virtually anyone to get involved in the development process. For these reasons, we argue there is a clear need for FOSS platforms designed explicitly for the analysis of urban segregation, particularly those that facilitate the generation of a wide variety of segregation statistics, hypothesis testing, and comparative analysis.

Toward that end, we introduce the `segregation` module for the Python Spatial Analysis Library (PySAL) that addresses each of the limitations identified above. We argue that our current approach has considerable 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. PySAL [40] is a well-established library of the Python programming language [43] for spatial analysis. Currently, PySAL has several features and modules comprising exploratory spatial data analysis, geospatial distribution dynamics, spatial econometrics, spatial network and graph analysis, geoprocessing, and spatial data visualization. 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, this manuscript intends to fill the of segregation analysis in this current library and Python scientific ecosystem.

Apart from allowing users to estimate spatial and non-spatial segregation statistics, the `segregation` package also includes functionality that is conspicuously absent in the segregation literature: statistical inference. In terms of previous work, Ref. [4] works with simulations to perform inference in a multidimensional version of the classic gini index. Also, Ref. [34] develops a sampling exercise of a multinomial distribution for the Dissimilarity Index and Gini Index to build asymptotic distribution of the estimators. Reference [2] builds an inference framework developing a likelihood ratio test for the presence of any systematic segregation for a bias-modified D. In addition, like [34], they develop tests for this measure relying on the asymptotic distributions. Reference [23] presents a Bayesian inference approach for the Dissimilarity Index and Ref. [20] develops a multilevel inference framework for residential segregation. More recently, Refs. [12] and [35] tackle the issue of inference on segregation. Reference [35] 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 Ref. [12] 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, Ref. [31] discuss the behavior

of the Dissimilarity Index under uncertainty of American Community Survey data under simulations studies.

The `segregation` module provides an 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 Ref. [2] where we generate the distribution of each measure under the null hypothesis of no systematic segregation by creating multiple samples generated using restricted conditional probabilities (absence of systematic segregation). Also, to generalize the use of our inference approach for single measures, the PySAL `segregation` module comprises different approaches to the null hypothesis assuming evenness, spatial permutations, absence of systematic segregation with permutation and evenness with permutation, which we discuss in detail later.

The major contribution of our framework is the ability to perform inference to compare more than one segregation measure.[8] To do so, we extend [41], who provide an inferential basis for comparisons of regional statistics. Their approach relies on a random labeling approach, where in each permutation, each unit in the dataset is assigned randomly to a point in time. However, our approach for comparative segregation stands as more generic and may be applied in any situation where two spatial contexts are compared. For example, a user can compare the evolution of a single region between two points in time, two regions in the same point in time, and, also, two regions between two points in time.[9] The first case is a straightforward adaptation of [41], but the second differs, given the possibility that each region may have entirely different spatial contexts. To try to provide alternative ways to assess the absence of segregation difference, our framework comprises not only a random data labeling ("random label" approach), but also a labeling process that randomizes observations according to the cumulative distribution function representing the population share for the group of interest in each unit ("counterfactual composition" approach).

The PySAL `segregation` module

The PySAL `segregation` module (hereafter referred as SM)[10] can be divided into two frameworks: point estimation and inference wrappers. The first framework can be, in turn, subdivided into non-spatial indices and spatial indices. 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.

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

[9] This last case is unusual, but our framework permits any of these combinations, as presented in Sect. ??.

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

Point estimation

Originally, SM had 25 segregation indices ranging from non-spatial indices and spatial indices that can be summarized in Table 1.[11,12] This table presents the main information of each function including its nomenclature in the literature, its class/function name in the `segregation` package, its input parameters, and whether it considers spatial context. A detailed description of each index and respective literature, presented as a table, can be found in the Appendix A.

All input data for SM rely on `pandas` DataFrames [28] for the non-spatial measures and `geopandas` DataFrames [21][13] 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` attribute. The first provides direct access to the point estimate of the segregation measure and the second gives access to the input data that SM uses internally to perform the estimates. To see the estimated D in the generic example above, the user would call `index.statistic` to see the fitted value.

Inference wrappers

Once the segregation classes described in "Point estimation" are fitted, a user can proceed with hypothesis testing to shed light on the statistical significance of her findings. Currently, the module facilitates hypothesis testing using either a single measure, or two values of the same measure. The summary of the inference wrappers is presented in Table 2.

A single value

The function `SingleValueTest` of SM performs inference through simulations for a single value of a given segregation index. To do so, a user must provide two fitted segregation statistics to the `seg_class` argument, the number of iterations to simulate under the null hypothesis to the `iterations_under_null` argument,

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

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

[13] It is worth mentioning, that using a `geopandas` GeoDataFrame for the non-spatial indices is also valid since it "behaves" as a usual `pandas` dataframe.

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 (xP*x*)	Isolation	No	–
Exposure (xP*y*)	Exposure	No	–
Late 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 (DDxP*x*)	DistanceDecayIsolation	Yes	Alpha, beta
Distance Decay Exposure (DDxP*y*)	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	–

which type of null hypothesis the inference will iterate with the `null_approach` argument, and whether the estimated *p* value will be single-tailed or two-tailed with the `two_tailed` argument. Certain calls can also include additional arguments to parameterize the estimate. A typical call for this function might be

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

The `null_approach` argument in this single measure framework includes several options. The default "`systematic`" draws multinomial simulations assuming that every group has the same probability with restricted conditional probabilities

Table 2 Inference wrappers available in PySAL `segregation` module

Type	Class/function	Function main inputs	Function outputs
Single value	Single value test	seg_class, iterations_under_null, null_approach, two_tailed	p_value, est_sim, statistic
Two values	Two value test	seg_class_1, seg_class_2, iterations_under_null, null_approach	p_value, est_sim, est_point_diff

given by the share unit of the the total population [2],[14] `"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 Ref. [39] 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.

Beyond simply providing flexibility for end-users, this choice has a critical impact on how a user may interpret her results, since the different approaches for null hypotheses affect directly the results of the inference test, depending on the combination of the index type of `seg_class` and the `null_approach` chosen. Therefore, the user must be aware of how these approaches affect the data generation process within the simulations if she means to draw meaningful conclusions within the scope of the analysis. More specifically, it is not true that in all cases, the null hypothesis represents the absence of segregation.[15]

Since little in the literature has compared different approaches for statistical inference in the segregation context, and this is among the primary motivations for for our work, it is important to discuss here some details of the inference frameworks provided in SM. Usually, in measuring segregation, the variables of concern are population counts or compositional ratios with statistical properties different from typical variables. Therefore, clarifying how we treat the population in each approach is a relevant matter.[16]

In SM's single-value inference framework, for the `"systematic"` approach, two multinomial distributions with the same probability parameters are generated

[14] 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 $\frac{\sum_j n_{ij}}{\sum_i \sum_j n_{ij}} = \frac{n_{i.}}{n_{..}}$.

[15] We are aware that for some measures, some approaches would not be appropriate, but we chose to allow these combinations, allowing our framework to remain 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 indices.

[16] We thank a reviewer for drawing attention to this point in the manuscript.

for the minority group and complementary group (i.e., total population of unit i = minority group of unit i + complementary group of unit i) and the total is given by their sum. Therefore, the total population of each simulation of this approach may differ from the original data. However, this is necessary to prevent unrealistic scenarios where the minority population would be greater than the total population in some units. In such a case, the total population of each unit cannot be fixed, although the total population of the entire spatial extent is fixed, since each size of the multinomial distribution is the original size of the data. The `"evenness"` approach draws from a binomial distribution in each unit with the same probability value given by the global proportion of the minority group. In this approach, the total population of each unit is fixed, but relaxes the total minority population in the units and also in the spatial extent under study. On the other hand, the `"permutation"` approach fixes the total population and the total minority population of the whole spatial extent while allowing spatial randomization and, therefore, letting each population of the units vary.

In terms of the software, the user can access the results of the function with the `p_value` and `est_sim`.[17] 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.

Comparative inference

To compare two different values, the user can rely on the `TwoValueTest` function. Similar to the previous function, the user needs to pass two segregation SM classes (`seg_class_1` 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. Optionally, the user may also pass additional parameters for each segregation estimation.[18] 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 = TwoValueTest(
seg_class_1 = index_1,
seg_class_2 = index_2,
iterations_under_null = 10000,
null_approach = "random_label"
```

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

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

[18] 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 (DEL) Index.

$$H_0 : \text{segregation measure}_1 - \text{segregation measure}_2 = 0, \quad (1)$$

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 Ref. [41] where SM uses random labeling applied to the data in each iteration.

Assuming a scenario with two different maps (regardless of being from the same city or different cities), each map has a set of polygons with a pair of values (`freq` and `population`) associated with each polygon. The concept underlying the `"random_label"` approach is to gather all pairs of values, regardless of the pair's polygon of origin, and randomly allocate each pair to a polygon in both maps, assuming a uniform probability among all polygons. Once all value pairs are allocated, the segregation measure is recalculated for each map and the difference between each map's segregation index is recorded. This process is repeated a sufficient number of times to build an artificial distribution of the differences of the null hypothesis.

The `"counterfactual_composition"` approach introduced in "Introduction" 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[19] 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 against which it is being compared[20]. 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 [11] where an analyst would sub-sample 50%, on average, the existing empirical distribution with the data of another distribution according to its CDF.

Lastly, this function also returns a `p_value` and `est_sim` attributes. The first is the two-tailed p value generated from the simulations and the second is the estimated difference under the null hypothesis (i.e., the divergence from zero in the absence of difference between segregation levels). In addition, the user can access the `est_point_diff` attribute which is the point estimate of the difference between the two values.

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` [17] and `seaborn` [48] functions.

[19] 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%.

[20] The details of the construction of these counterfactual values are presented in Appendix B.

For single measures, the distribution is generated from the point estimates among 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 indicate an absence of segregation difference. The user can visually inspect the results with `inference_result.plot()` or `compare_result.plot()`.

Performance comparison and reliability 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 indices can change drastically.[21]

Figure 1 depicts a time comparison for a single estimation of each index of Table 1 in seconds for a 10×10 regular lattice with simulated data.[22],[23] From this figure, it is clear that the Modified Gini (Gct) is the most time-consuming index to compute among the set of indices. This is due to the fact that its construction relies on a bootstrap simulation of multiple binomial distributions for each unit and also because its calculation, given by Eq. (6) in Appendix A, relies on an outer product of vectors which can be computationally 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 indices 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 indices.

In Table 3, we present the results of a benchmark test that verifies the correctness of the point estimations of SM. Since `OasisR` has done an extensive comparison with different tools showing virtually the same results for all of its indices, this package will be mainly our benchmark for the comparisons [47]. We cannot identify any existing software package that calculates the SPP Index from Ref. [16] and, therefore, it is not present in this table due to the lack of a benchmark. Also, since SM is open source, the Python code used to calculate the indices is available to the public to check in its entirety.

[21] We also noticed that for most of the indices, specially the spatial ones, SM was much faster to estimate than the implementation of Ref. [47].

[22] 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.

[23] The indices 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 indices that rely on simulations (Dct, Gct, and Dbc) have the same value of 500 for the iterations and indices that rely on integration (R and SPP) 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}.

This table makes clear that each of the implementations in SM generate reliable values, as they match their expected values from the benchmark. Specifically, the *R* Index was tested with the `seg` package; whereas, the Dct, Gct, Dbc, and Ddc were checked with the values provided in their respective literature. All of these indices, except Ddc, rely on simulations and, therefore, while there is some variance between our estimates and their benchmark comparisons, such variance is expected and disappears given certain numerical thresholds. One thing to notice is that DDxP*x*, DDxP*y* and SP resulted in slightly different values since the specification of the distance of spatial unit *i* with itself is calculated in SM following exactly [24], unlike `OasisR`.[24]

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

Racial segregation in the United States has been a topical focus for a vast literature. Recently Ref. [1] used the *D* Index to study Black–White and Hispanic–White segregation in counties across the US. In another recent contribution, Ref. [27] made a vast metropolitan study for a 40-year period on hypersegregation of black population. Even more recently, Ref. [9] studied ethnic residential segregation of metropolitan regions of California using a different type of spatial isolation. Using this literature as a backdrop, we use the following sections to present an example study of racial segregation in the US to demonstrate the unique functionality now available to researchers using SM.

In this section, we rely on SM to calculate several segregation measures for Los Angeles County, CA, and New York City,[25] NY, census data tract level for two groups: non-Hispanic black population (`nhblk`) and others.[26] In this example, we examine the total measured level of segregation in Los Angeles along all five dimensions (evenness, isolation, clustering, concentration and centralization) using all indices available to making point estimates 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). We also use these esimates as points of comparison between Los Angeles and New York in 2010 (one cross section for two spatial contexts).[27]

[24] The values marked with * are virtually the same although `OasisR` has a mispecification in d_{ii} that does not follow [24]. This difference can be checked in https://github.com/cran/OasisR/pull/1/commits/cc3681dae96188663230cf140d0cf41fd90e45cd.

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

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

[27] 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 [22] using an Intel (R) Core (TM) i7-8750H CPU with 2.21 GHz and 16 GB of RAM. It was necessary approximately 34.7 h to run all application results here presented.

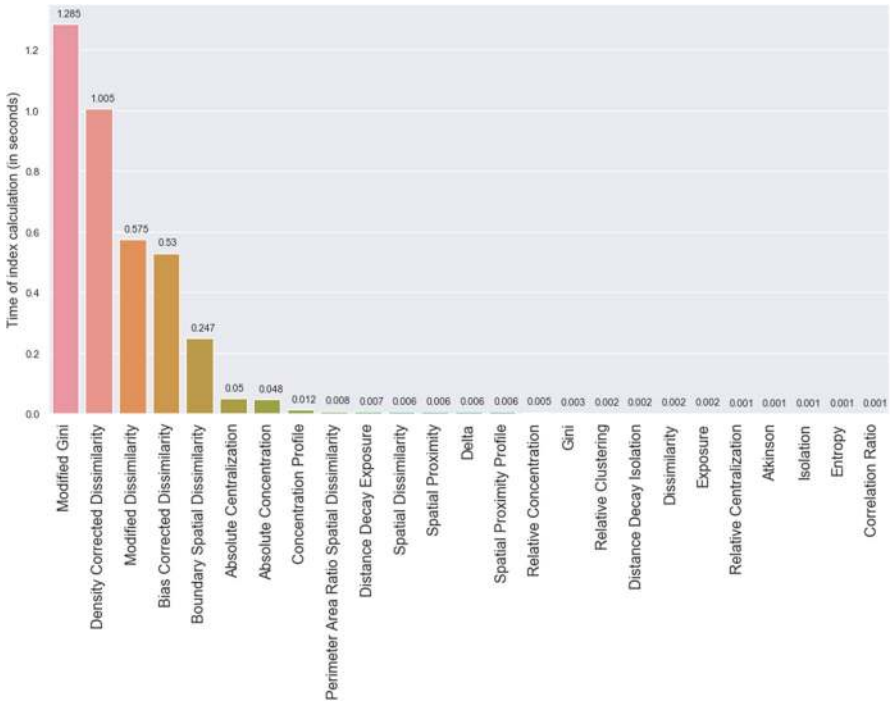

Fig. 1 Time comparison estimation between all indices of SM for a 10×10 regular lattice

Figure 2 displays the clear spatial patterning of `nhblk` in the Los Angeles metropolitan region where the color gradient represents the relative share of non-Hispanic black residents living in each tract (`nhblk` divided by total tract population), i. e., the composition. The maps show an obvious pattern of spatial concentration and unevenness in terms of frequency of the non-Hispanic Black population and, therefore, it is reasonable to perform a regional segregation analysis. We also note the unusual spatial distribution of census tracts within Los Angeles County, where topographical features lead to considerable asymmetry of tracts areas. Such a condition could affect the spatial estimates as well as the inference for spatial measures.

Figures 3 and 4 present the simulations for each measure under different null hypotheses. These graphs display the distribution under the null hypothesis as a blue density curve and a vertical red line that represents the point estimate for the measure. In addition, the value of each segregation measure is highlighted in each title.

In Fig. 3, the simulations were drawn assuming a multinomial distribution with no systematic segregation. when comparing the actual value with that estimated from the data, the unusual behavior of the distributions becomes clear: all 25 measures are highly significant, with the exception of the Exposure Index. The majority of the distributions present values close to zero, which is in accordance with the mathematical property of some measures that assumes zero when there is no systematic segregation in the data. Figure 4 shows the current 13 spatial segregation

Table 3 Benchmark testing for PySAL `segregation` module point estimations

Measure	Benchmark	Result
Dissimilarity (*D*)	`OasisR`	Same value
Gini (*G*)	`OasisR`	Same value
Entropy (*H*)	`OasisR`	Same value
Isolation (xP*x*)	`OasisR`	Same value
Exposure (xP*y*)	`OasisR`	Same value
Atkinson (*A*)	`OasisR`	Same value
Correlation ratio (*V*)	`OasisR`	Same value
Concentration Profile (*R*)	`seg`	Same value
Modified Dissimilarity (Dct)	Table 1 of [7]	Same value with 2 digits precision
Modified Gini (Gct)	Table 1 of [7]	Same value with 2 digits precision
Bias-Corrected Dissimilarity (Dbc)	Table 1 (a) of [2]	Same value with 2 digits precision
Density-Corrected Dissimilarity (Ddc)	Table 1 (a) of [2]	Same value with 2 digits precision
Spatial Dissimilarity (SD)	`OasisR`	Same value
Boundary Spatial Dissimilarity (BSD)	`OasisR`	Same value
Perimeter Area Ratio Spatial Dissimilarity (PARD)	`OasisR`	Same value
Distance Decay Isolation (DDxP*x*)	`OasisR`	Same value*
Distance Decay Exposure (DDxP*y*)	`OasisR`	Same value*
Spatial Proximity (SP)	`OasisR`	Same value*
Relative Clustering (RCL)	`OasisR`	Same value
Delta (DEL)	`OasisR`	Same value
Absolute Concentration (ACO)	`OasisR`	Same value
Relative Concentration (RCO)	`OasisR`	Same value
Absolute Centralization (ACE)	`OasisR`	Same value
Relative Centralization (RCE)	`OasisR`	Same value

measures under the spatial permutation approach.[28] In this case, the statistical significance of each measure is not as highlighted as the prior results. Here, the SPP (p value ≈ 0.068), the Absolute Concentration (ACO) (p value ≈ 0.272) and the Relative Concentration (RCO) (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).

One of the major contributions of SM is the ability to assess differences in segregation levels between two distinct measures easily. If Los Angeles county was statistically segregated in 2010, a natural question that may arise is "Is Los Angeles

[28] 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.

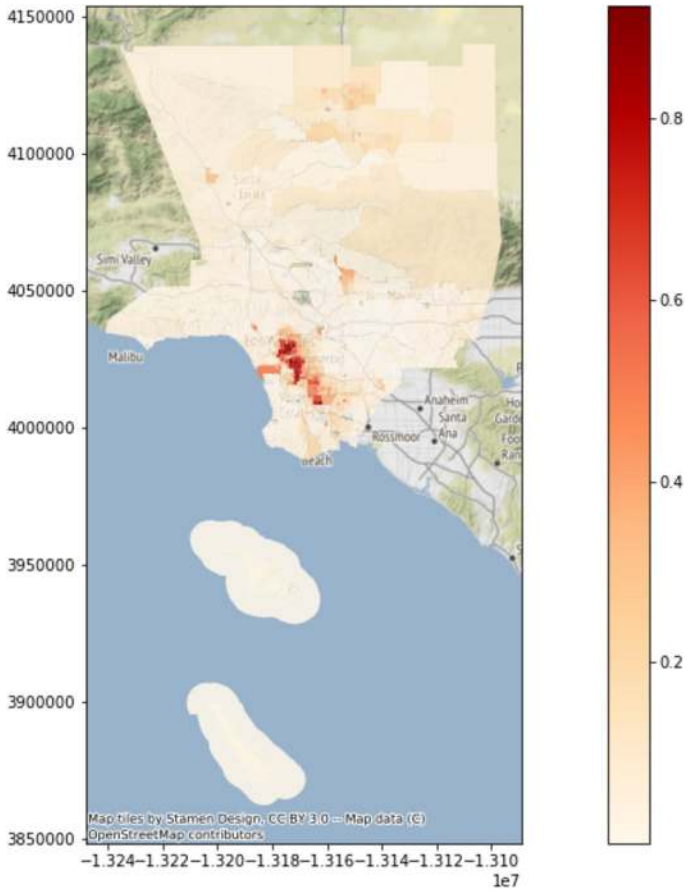

Fig. 2 Non-Hispanic Black population (`nhblk`) in Los Angeles county composition in 2010

County more or less segregated in 2010 than in 2000?".[29] Figure 5 depicts the composition spatial distribution of this county using census data from 2000. Despite the similarities, the graph shows a slightly different conclusion from the one presented in Fig. 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 evolution of Los Angeles county over this decade, we rely on the `TwoValueTest` function of SM with the `random_label` approach. Figure 6 displays the results for the difference between 2000 and 2010 for each of the measures. In general, it is clear from the graph that 2000 was more segregated than 2010, since the majority of vertical red lines are located on negative values. Moreover, for almost all segregation measures available, these

[29] H_0 : Los Angeles segregation$_{2010}$ − Los Angeles segregation$_{2000}$ = 0.

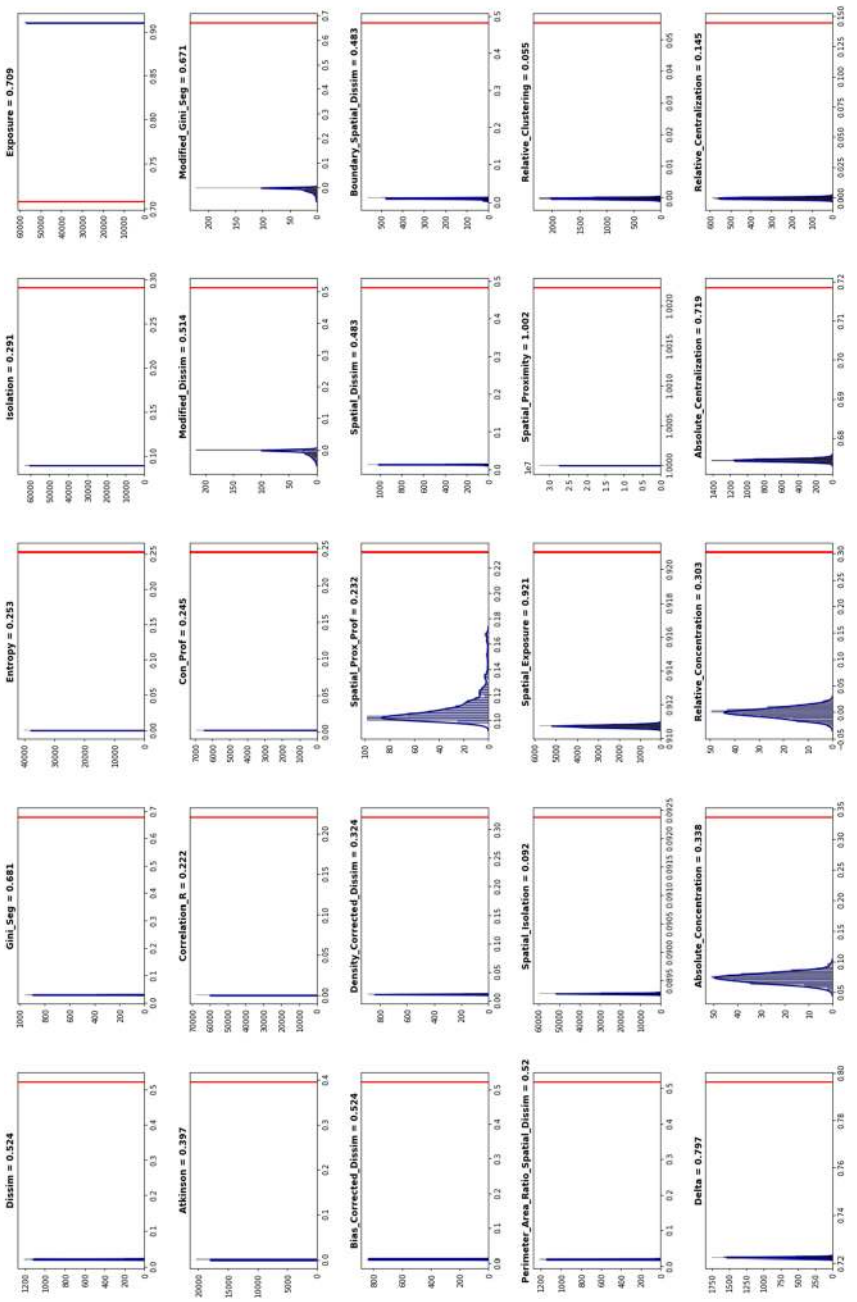

Fig. 3 Simulations using SM for non-Hispanic Black population (`nhblk`) in Los Angeles in 2010: `systematic` null approach. The point estimation of each segregation measure in presented in each title. Here, Distance Decay Isolation/Exposure are named Spatial Isolation/Exposure

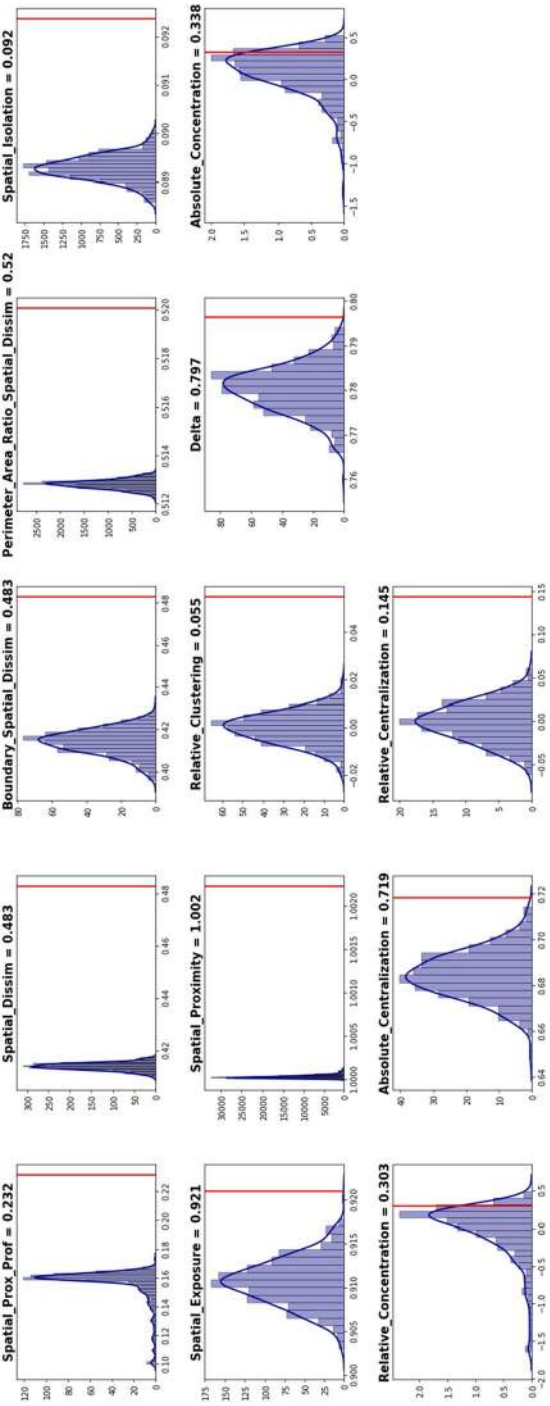

Fig. 4 Simulations using SM for non-Hispanic Black population (`nhblk`) in Los Angeles in 2010: `permutation` null approach. The point estimation of each segregation measure in presented in each title. Here, Distance Decay Isolation/Exposure are named Spatial Isolation/Exposure

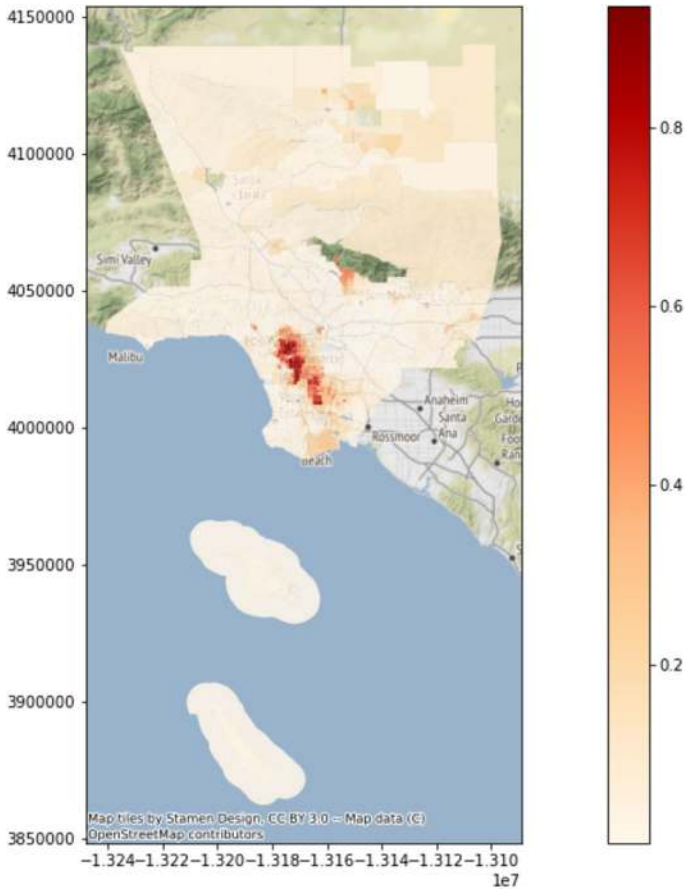

Fig. 5 Non-Hispanic Black population (`nhblk`) in Los Angeles county composition in 2000

difference values seem to be statistically significant since they are on the far left tail of each distribution.[30]

However, some particularities emerge. For two of the concentration dimensions in Los Angeles (ACO and RCO), the measures are not statistically significant.[31] 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 are sensible, given the earlier discussion 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 mentioning the lack of statistical significance for the SPP (p value ≈ 0.096), related to the clustering dimension of segregation.

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

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

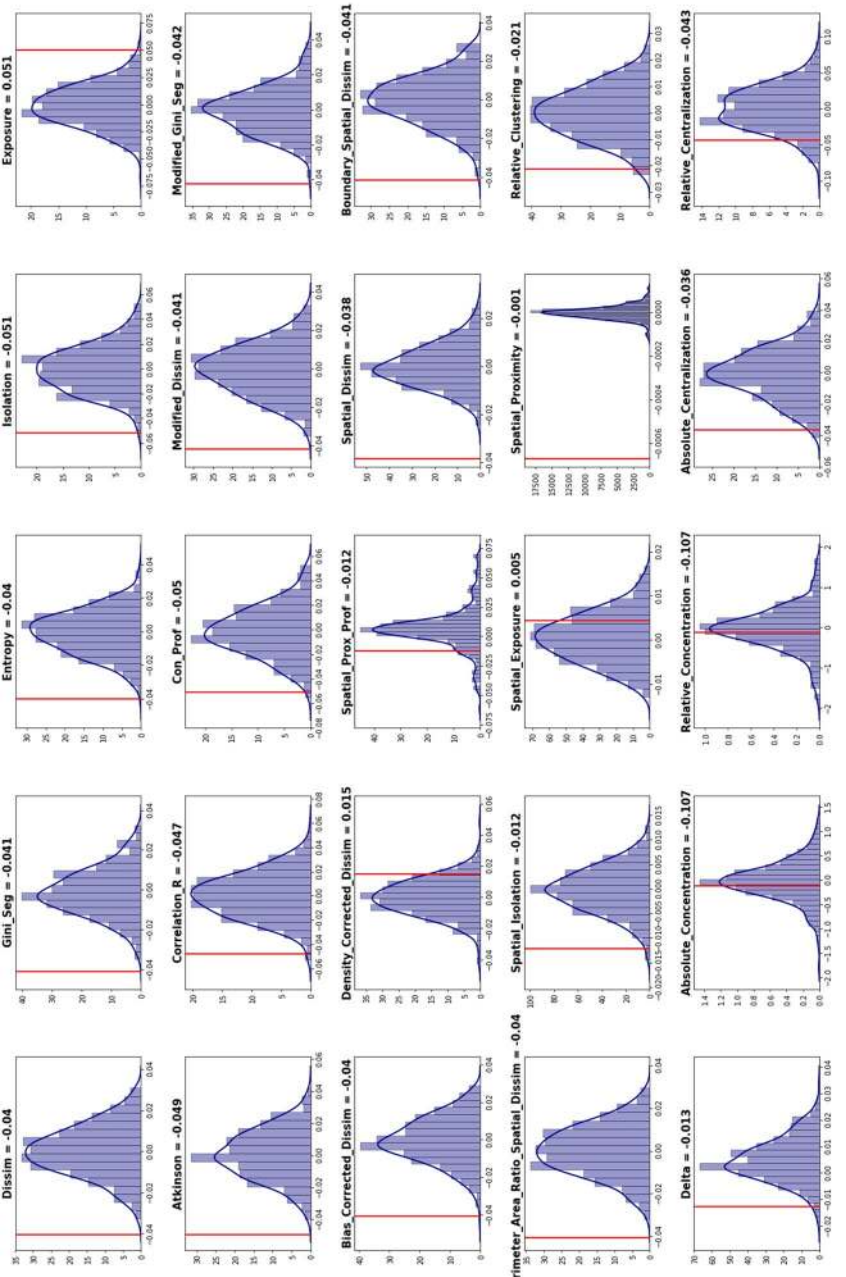

Fig. 6 Simulations using SM for Los Angeles comparison between 2000 and 2010 using the `random_label` null approach. The point estimation of the difference of each segregation measure is presented in each title. Here, Distance Decay Isolation/Exposure is named Spatial Isolation/Exposure

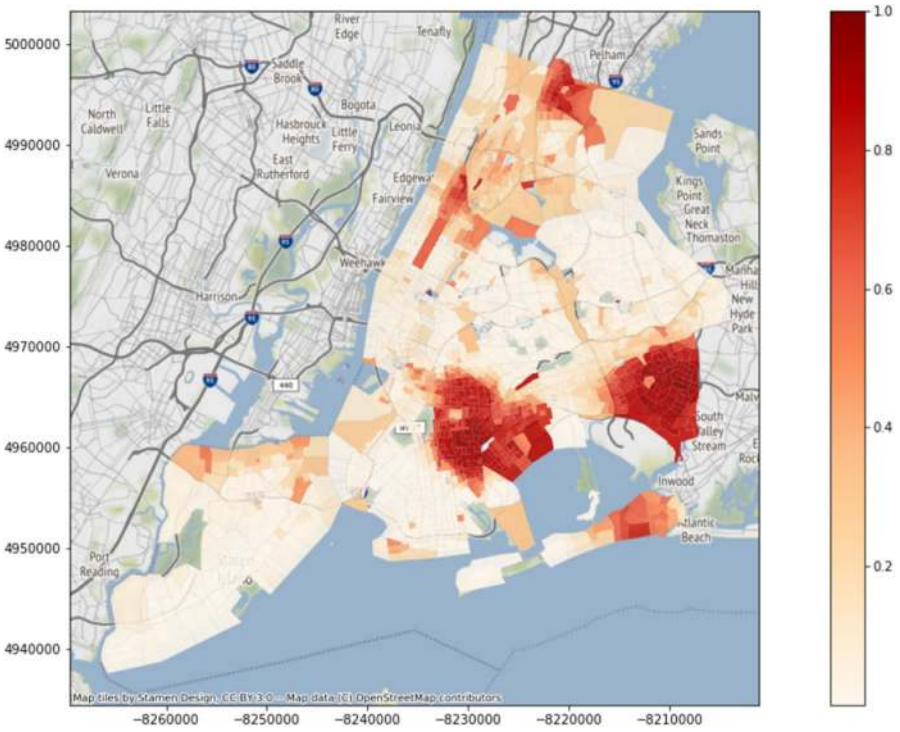

Fig. 7 Non-Hispanic Black population (`nhblk`) in New York composition in 2010

The ability to make comparisons between regions is also possible with SM. Since the `TwoValueTest` function can handle two classes fitted previously in a generic framework, a user can pass two segregation measures from two different spatial contexts. Figure 7 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). With these two maps in hand, a natural question in the social sciences might be to assess the statistical significance of the difference in measured segregation levels between the two metropolitan areas. To shed light on this question, Fig. 8 depicts the comparison for both cities[32] for 2010 census tract data for all measures using the `random_label` approach.

From this graph, it is clear that all indices (with the exception of ACO, RCO and ACE) resulted in significant values. For an expressive number of measures (*D*, *G*, *H*, xP*x*, *A*, *V*, *R*, Dct, Gct, Dbc, SPP, SD, BSD, DDxP*x* and DEL) New York shows

[32] H_0 :Los Angeles segregation − New York segregation = 0.

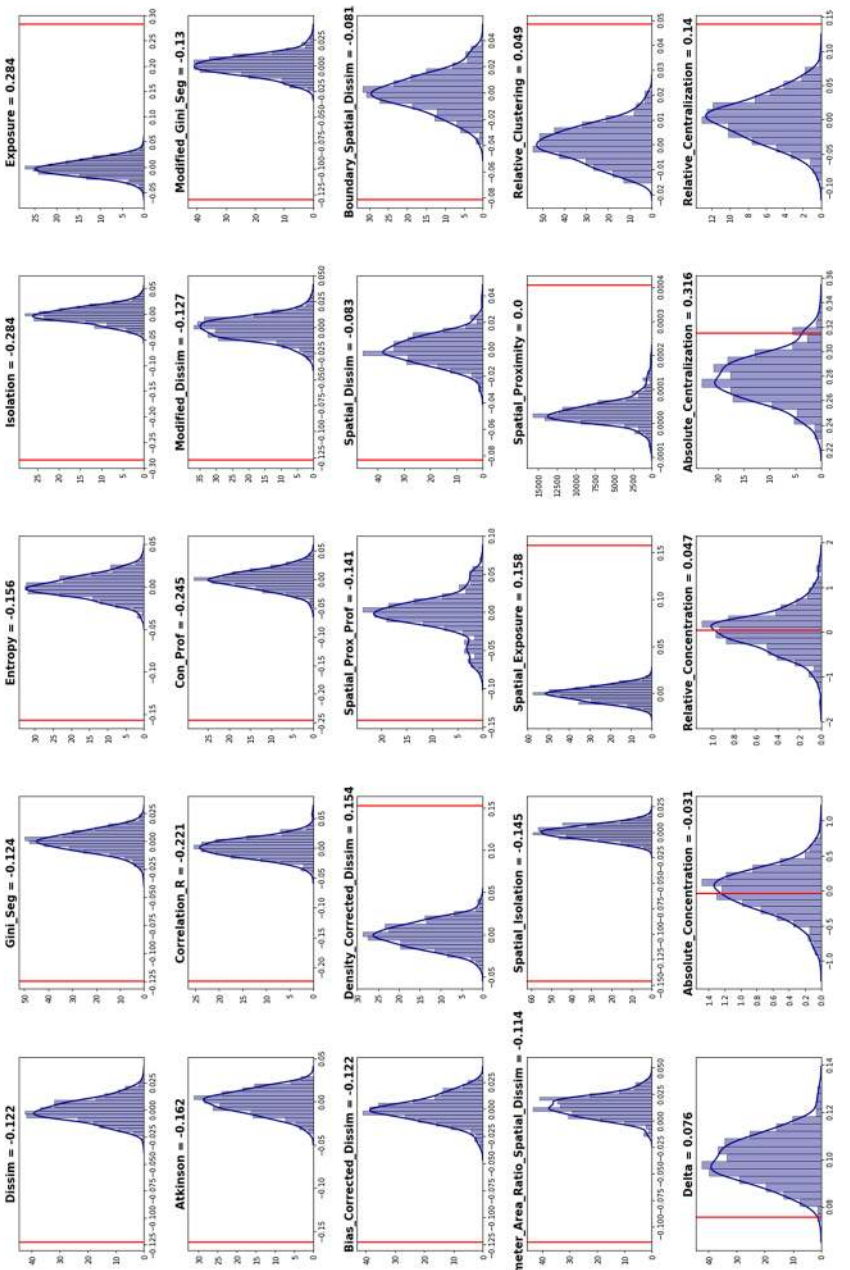

Fig. 8 Simulations using SM for Los Angeles and New York comparison in 2010 using the `random_label` null approach. The point estimation of the difference of each segregation measure in presented in each title. Here, Distance Decay Isolation/Exposure is named Spatial Isolation/Exposure

higher levels of segregation.[33] This indicates that, in general, non-Hispanic blacks are more segregated in New York than Los Angeles.[34] On the other hand, some interesting results also emerge from the clustering and centralization dimensions for some measures. The results show 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 consistent with the discussion comparing maps from each city which shows that that the non-Hispanic black population in Los Angeles is more concentrated in a single `nhblk` hotspot, unlike New York has multiple hotspots.

This unexpected result highlights the importance of defining the appropriate dimension of segregation an analyst wishes to study using comparative inference. One might argue that a given city is considerably more segregated than another, but this may not be true from the perspective of a different dimension of segregation. The same behavior can arise when comparing the same city for two distinct periods as what happened with ACO and RCO, for example, for Los Angeles County in 2010 versus itself in 2000.

Conclusion

Segregation measurements have a vast literature and an extensive use since the first half of the twentieth century. This field is constantly under progress with increasingly works discussing the properties of different segregation indices, better ways to overcome limitations, illustrate applications, etc. This work is an attempt to advance the use of segregation measure through an open-source framework within the `PySAL` ecosystem—the `PySAL segregation` module (SM). Moreover, our contribution is not simply to provide an easy method to estimate a wide variety of well-known non-spatial and spatial segregation measures, but also to build a consistent software framework for conducting statistical inference that has not been considered before.

In so doing, 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 `SingleValueTest` and `TwoValueTest`. Each one of these represents a wrapper function for the segregation classes fitted previously, where the first is used to perform inference for a single measure, while the second allows comparison between two measures. Both functions depend on techniques for simulating distributions for the null hypothesis chosen.

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

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

As an illustration, we used Los Angeles County and New York City to perform regional segregation analysis using census tract data. We studied the degree to which the non-Hispanic black population in these cities was segregated 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. To illustrate the `TwoValueTest`, two types of comparisons were made: same space between two time periods and two spaces for the same time period. The former assesses the evolution of Los Angeles between 2000 and 2010 concluding that it was statistically more segregated in the past; 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. These illustrations make clear that SM can be a powerful tool for further research into the validity of the five dimensions taxonomy of Ref. [24].

This PySAL module is under active 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 Ref. [37], the second draws inspiration from Ref. [47], 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 [45] and, finally, the street network-based measures draw inspiration from [42] and use a handful of libraries from the Urban Data Science Toolkit.[35] Given all functionalities present in this paper and all these other features mentioned, we are confident that the current module is one of the most complete tools currently available for analyzing urban segregation.

Additionally, several aspects remain to be explored. Possible extensions comprise more measures that can be added such as the Proportion of Central City number (PCC) [24], other indices present in Ref. [47] and the parametric and nonparametric approach of the class of indices of, respectively, Refs. [12] and [35]. Another landscape of opportunity is not only "zone-based" measures, but also "surface-based" methods as quoted in Ref. [15]. In this regard, spatial counterfactual approaches [6] 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 these kinds of data.

Acknowledgements 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).

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

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 4[36,37] in addition with the respective dimension.

For consistency of notation, we assume that n_{ij} is the population of unit $i \in \{1, \ldots, I\}$ 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 indices can be build for any group j of the data.

The Dissimilarity Index (D) is given by:

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

The spatial D (SD) is given by:

$$SD = D - \frac{\sum_{i_1=1}^{I} \sum_{i_2=1}^{I} \left| \tilde{s}_{ij}^{i_1} - \tilde{s}_{ij}^{i_2} \right| c_{i_1 i_2}}{\sum_{i_1=1}^{I} \sum_{i_2=1}^{I} c_{i_1 i_2}}, \tag{3}$$

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:

$$BSD = D - \frac{1}{2} \sum_{i_1=1}^{I} \sum_{i_2=1}^{I} w_{i_1 i_2} \left| \tilde{s}_{ij}^{i_1} - \tilde{s}_{ij}^{i_2} \right|, \tag{4}$$

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):

[36] This table does not reflect necessarily the original/pioneer paper of each measure, but rather the related literature of the formulas presented in this Appendix.

[37] We considered to include the mixture of betas approach of Ref. [35] for the D, G and H indices, 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 4 Segregation measures-related literature for PySAL `segregation` module point estimations

Measure	Related literature	Dimension
Dissimilarity (D)	[24]	Evenness
Gini (G)	[24]	Evenness
Entropy (H)	[24]	Evenness
Isolation (xPx)	[24]	Isolation
Exposure (xPy)	[24]	Isolation
Atkinson (A)	[24]	Evenness
Correlation ratio (V)	[24]	Isolation
Concentration Profile (R)	[16]	Evenness
Modified Dissimilarity (Dct)	[7]	Evenness
Modified Gini (Gct)	[7]	Evenness
Bias-Corrected Dissimilarity (Dbc)	[2]	Evenness
Density-Corrected Dissimilarity (Ddc)	[2]	Evenness
Spatial Proximity Profile (SPP)	[16]	Clustering
Spatial Dissimilarity (SD)	[30]	Evenness
Boundary Spatial Dissimilarity (BSD)	[15]	Evenness
Perimeter Area Ratio Spatial Dissimilarity (PARD)	[49]	Evenness
Distance Decay Isolation (DDxPx)	[29]	Isolation
Distance Decay Exposure (DDxPy)	[29]	Isolation
Spatial Proximity (SP)	[24]	Clustering
Relative Clustering (RCL)	[24]	Clustering
Delta (DEL)	[24]	Concentration
Absolute Concentration (ACO)	[24]	Concentration
Relative Concentration (RCO)	[24]	Concentration
Absolute Centralization (ACE)	[24]	Centralization
Relative Centralization (RCE)	[24]	Centralization

$$\frac{\frac{1}{2}\left[\left(\frac{P_i}{A_i}\right)+\left(\frac{P_j}{A_j}\right)\right]}{\mathrm{MAX}\left(\frac{P}{A}\right)}, \tag{5}$$

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:

$$G=\sum_{i_1=1}^{I}\sum_{i_2=1}^{I}\frac{n_{i_1.}n_{i_2.}\mid \tilde{s}_{ij}^{i_1}-\tilde{s}_{ij}^{i_2}\mid}{2n_{..}^2\frac{n_{.j}}{n_{..}}\left(1-\frac{n_{.j}}{n_{..}}\right)}. \tag{6}$$

The global entropy (E) is given by:

$$E = \frac{n_{.j}}{n_{..}} \log\left(\frac{1}{\frac{n_{.j}}{n_{..}}}\right) + \left(1 - \frac{n_{.j}}{n_{..}}\right)\log\left(\frac{1}{1 - \frac{n_{.j}}{n_{..}}}\right), \tag{7}$$

while the unit's entropy is analogously:

$$E_i = \tilde{s}_{ij} \log\left(\frac{1}{\tilde{s}_{ij}}\right) + \left(1 - \tilde{s}_{ij}\right) \log\left(\frac{1}{1 - \tilde{s}_{ij}}\right). \tag{8}$$

Therefore, the Entropy Index (H) is given by:

$$H = \sum_{i=1}^{I} \frac{n_{i.}\left(E - E_i\right)}{En_{..}} \tag{9}$$

The Atkinson Index (A) is given by:

$$A = 1 - \frac{\frac{n_{.j}}{n_{..}}}{1 - \frac{n_{.j}}{n_{..}}} \left| \sum_{i=1}^{I} \left[\frac{\left(1 - \tilde{s}_{ij}\right)^{1-b} \tilde{s}_{ij}^{b} t_i}{\frac{n_{.j}}{n_{..}} n_{..}} \right] \right|^{\frac{1}{1-b}}, \tag{10}$$

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 Ref. [16] and tries to inspect the evenness aspect of segregation. The threshold proportion t is given by:

$$\upsilon_t = \frac{\sum_{i=1}^{I} n_{ij} g(t, i)}{\sum_{i=1}^{I} n_{ij}}. \tag{11}$$

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

$$g(t, i) = \begin{cases} 1 \text{ if } \frac{n_{ij}}{n_{i.}} \geq t \\ 0 \text{ otherwise} \end{cases}. \tag{12}$$

The Concentration Profile (R) is given by:

$$R = \frac{\frac{n_{.j}}{n_{..}} - \left(\int_{t=0}^{\frac{n_{.j}}{n_{..}}} \upsilon_t \, \mathrm{d}\, t - \int_{t=\frac{n_{.j}}{n_{..}}}^{1} \upsilon_t \, \mathrm{d}\, t \right)}{1 - \frac{n_{.j}}{n_{..}}}. \tag{13}$$

The SPP is similar to the Concentration Profile, but with the addition of the spatial component in the connecting function:

$$\eta_t = \frac{k^2 - k}{\sum_{i_1} \sum_{1_2} \delta_{i_1 i_2}}, \tag{14}$$

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 SPP, 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:

$$\mathrm{xP}x = \sum_{i=1}^{I} \left(\hat{s}_{ix}\right)\left(\tilde{s}_{ix}\right). \tag{15}$$

The Exposure (xPy) of x is giving by

$$\mathrm{xP}y = \sum_{i=1}^{I} \left(\hat{s}_{iy}\right)\left(\tilde{s}_{iy}\right). \tag{16}$$

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

$$V = \mathrm{Eta}^2 = \frac{\mathrm{xP}x - \frac{n_{.x}}{n_{..}}}{1 - \frac{n_{.x}}{n_{..}}}. \tag{17}$$

The SP Index is given by:

$$\mathrm{SP} = \frac{XP_{xx} + YP_{yy}}{TP_{tt}}, \tag{18}$$

where

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

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

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

$$\zeta_{i_1 i_2} = \exp(-d_{i_1 i_2}),$$

$d_{i_1 i_2}$ 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 extract the centroid of each unit and calculate the euclidean distance.

The RCL measure is given by:

$$\mathrm{RCL} = \frac{P_{xx}}{P_{yy}} - 1. \tag{19}$$

The Distance Decay Isolation (DDxPx) is given by:

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

where

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

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:

$$\mathrm{DDxP}y = \sum_{i_1=1}^{I} \left(\hat{s}_{i_1 x}\right)\left(\sum_{i_2=1}^{I} P_{i_1 i_2}\left(\tilde{s}_{i_1 y}\right)\right) \tag{21}$$

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

The DEL measure is given by the following equation:

$$\mathrm{DEL} = \frac{1}{2} \sum_{i=1}^{I} \left| \hat{s}_{ij} - \frac{a_i}{A} \right|, \tag{22}$$

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 ACO Index is given by:

$$\mathrm{ACO} = 1 - \frac{\sum_{i=1}^{I}\left(\frac{n_{ij} a_i}{n_j}\right) - \sum_{i=1}^{n_1}\left(\frac{n_i a_i}{T_1}\right)}{\sum_{i=n_2}^{I}\left(\frac{n_i a_i}{T_2}\right) - \sum_{i=1}^{n_1}\left(\frac{n_i a_i}{T_1}\right)}, \tag{23}$$

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 RCO Index:

$$\mathrm{RCO} = \frac{\frac{\sum_{i=1}^{I}\left(\frac{n_{ix}a_i}{n_{.x}}\right)}{\sum_{i=1}^{I}\left(\frac{n_{iy}a_i}{n_{.y}}\right)} - 1}{\frac{\sum_{i=1}^{n_1}\left(\frac{n_{i.}a_i}{T_1}\right)}{\sum_{i=n_2}^{I}\left(\frac{n_{i.}a_i}{T_2}\right)} - 1}, \tag{24}$$

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 RCE:

$$\mathrm{ACE} = \left(\sum_{i=2}^{I} X_{i-1}A_i\right) - \left(\sum_{i=2}^{I} X_i A_{i-1}\right), \tag{25}$$

$$\mathrm{RCE} = \left(\sum_{i=2}^{I} X_{i-1}Y_i\right) - \left(\sum_{i=2}^{I} X_i Y_{i-1}\right), \tag{26}$$

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 Dct Index based on [7] 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:

$$\mathrm{Dct} = \begin{cases} \frac{D-D^*}{1-D^*} & \text{if } D \geq D^* \\ \frac{D-D^*}{D^*} & \text{if } D < D^* \end{cases}. \tag{27}$$

Similarly, the Gct based also on Ref. [7] 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:

$$\text{Gct} = \begin{cases} \frac{G-G^*}{1-G^*} & \text{if } G \geq G^* \\ \frac{G-G^*}{G^*} & \text{if } G < G^* \end{cases}. \tag{28}$$

Lastly, the Bias-Corrected (Dbc) and Density-Corrected (Ddc) Dissimilarities indices are presented in Ref. [2]. The Dbc is given by:

$$D_{\text{bc}} = 2D - \bar{D}_{\text{b}}, \tag{29}$$

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:

$$D_{\text{dc}} = \frac{1}{2}\sum_{i=1}^{I} \hat{\sigma}_i n\left(\hat{\theta}_i\right), \tag{30}$$

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

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

and ϕ is the standard normal density.

B: Counterfactual composition details

Following the same notation of A and assuming building counterfactual values fro two different cities, we form the cumulative distribution functions (CDF) for these values taken over all the tracts in City 1: $F^{(1)}(\tilde{s}_{i,j}^{1,t})$, and City 2: $F^{(2)}(\tilde{s}_{i,j}^{2,t})$. To create a counterfactual distribution that imposes the attribute distribution of City 2 on the spatial structure of City 1 we take $p_{i,j}^{1,t} = F^{(1)}(\tilde{s}_{i,j}^{1,t})$ and then generate $n_{i,j}^{1,t}|_{attr=2} = {F^{(2)}}^{-1}(p_{i,j}^{1,t})n_{i,.}^{1,t}$, where $attr = 2$ means that this population is calculated given the attributes of City 2. This entire process is done for all tracts of a group in City 1 and the majority group population is given by the difference $n_{i,.}^{1,t} - n_{i,j}^{1,t}|_{attr=2}$. The populations for City 2 are generated analogously.

References

1. Allen, J. P., & Turner, E. (2012). Black-White and Hispanic-White segregation in US counties. *The Professional Geographer*, *64*(4), 503–520.
2. 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.
3. 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.
4. 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.
5. Brown, L. A., & Chung, S. Y. (2006). Spatial segregation, segregation indices and the geographical perspective. *Population, Space and Place*, *12*(2), 125–143.
6. Carrillo, P. E., & Rothbaum, J. L. (2016). Counterfactual spatial distributions. *Journal of Regional Science*, *56*(5), 868–894.
7. Carrington, W. J., & Troske, K. R. (1997). On measuring segregation in samples with small units. *Journal of Business & Economic Statistics*, *15*(4), 402–409.
8. Carrington, W. J., & Troske, K. R. (1998). Interfirm segregation and the Black/White wage gap. *Journal of Labor Economics*, *16*(2), 231–260.
9. 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*, *45*(6), 1038–1054.
10. Cowgill, D. O., & Cowgill, M. S. (1951). An index of segregation based on block statistics. *American Sociological Review*, *16*(6), 825–831.
11. Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on winter simulationACM* (pp. 260–265).
12. d'Haultfoeuille, X., & Rathelot, R. (2017). Measuring segregation on small units: A partial identification analysis. *Quantitative Economics*, *8*(1), 39–73.
13. Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210–217.
14. 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.
15. Hong, S. Y., O'Sullivan, D., & Sadahiro, Y. (2014). Implementing spatial segregation measures in R. *PloS One*, *9*(11), e113767.
16. Hong, S. Y., & Sadahiro, Y. (2014). Measuring geographic segregation: A graph-based approach. *Journal of Geographical Systems*, *16*(2), 211–231.
17. Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
18. James, D. R., & Taeuber, K. E. (1985). Measures of segregation. *Sociological Methodology*, *15*, 1–32.
19. Johnston, R., Poulsen, M., & Forrest, J. (2007). Ethnic and racial segregation in us metropolitan areas, 1980–2000: The dimensions of segregation revisited. *Urban Affairs Review*, *42*(4), 479–504.
20. Jones, K., Johnston, R., Manley, D., Owen, D., & Charlton, C. (2015). Ethnic residential segregation: A multilevel, multigroup, multiscale approach exemplified by London in 2011. *Demography*, *52*(6), 1995–2019.
21. Jordahl, K. (2014). Geopandas: Python tools for geographic data. https://github.com/geopandas/geopandas. Accessed 3 Apr 2019.
22. 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. In *ELPUB* (pp. 87–90). https://jupyter.org/. Accessed 3 Apr 2019.
23. Lee, D., Minton, J., & Pryce, G. (2015). Bayesian inference for the dissimilarity index in the presence of spatial autocorrelation. *Spatial Statistics*, *11*, 81–95.
24. Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces*, *67*(2), 281–315.
25. Massey, D. S., & Denton, N. A. (1989). Hypersegregation in us metropolitan areas: Black and hispanic segregation along five dimensions. *Demography*, *26*(3), 373–391.

26. Massey, D. S., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Cambridge: Harvard University Press.
27. Massey, D. S., & Tannen, J. (2015). A research note on trends in black hypersegregation. *Demography*, *52*(3), 1025–1034.
28. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt, J. Millman (Ed.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).
29. Morgan, B. S. (1983). A distance-decay based interaction index to measure residential segregation. *Area*, *15*(3), 211–217.
30. Morrill, R. L. (1991). On the measure of geographic segregation. *Geography Research Forum*, *11*, 25–36.
31. 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.
32. Park, R. E. (1926). The urban community as a spatial pattern and a moral order. In *Urban social segregation* (pp. 21–31).
33. R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. http://www.R-project.org. Accessed 3 Apr 2019.
34. Ransom, M. R. (2000). Sampling distributions of segregation indexes. *Sociological Methods & Research*, *28*(4), 454–475.
35. Rathelot, R. (2012). Measuring segregation when units are small: A parametric approach. *Journal of Business & Economic Statistics*, *30*(4), 546–553.
36. Reardon, S. F., & Townsend, J. B. (1999). SEG: Stata module to compute multiple-group diversity and segregation indices. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s375001.html. Accessed 3 Apr 2019.
37. Reardon, S. F., & Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological Methodology*, *32*(1), 33–67.
38. Reardon, S. F., & O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology*, *34*(1), 121–162.
39. Rey, S. J. (2004). Spatial analysis of regional income inequality. *Spatially Integrated Social Science*, *1*, 280–299.
40. Rey, S. J., & Anselin, L. (2010). PySAL: A Python library of spatial analytical methods. In *Handbook of applied spatial analysis* (pp. 175–193). Springer.
41. Rey, S. J., & Sastré-Gutiérrez, M. L. (2010). Interregional inequality dynamics in Mexico. *Spatial Economic Analysis*, *5*(3), 277–298.
42. Roberto, E. (2018). The spatial proximity and connectivity method for measuring and analyzing residential segregation. *Sociological Methodology*, *48*(1), 182–224.
43. Rossum, G. (1995). Python reference manual. Technical report. The Netherlands: Amsterdam.
44. Royuela, V., & Vargas, M., et al. (2010). Residential segregation: A literature review. Technical report.
45. Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.
46. Söderström, M., & Uusitalo, R. (2010). School choice and segregation: Evidence from an admission reform. *Scandinavian Journal of Economics*, *112*(1), 55–76.
47. 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.
48. 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. (2017). mwaskom/seaborn: v0.8.1 (september 2017). https://doi.org/10.5281/zenodo.883859.
49. Wong, D. W. (1993). Spatial indices of segregation. *Urban Studies*, *30*(3), 559–572.
50. Wong, D. W. (2003). Implementing spatial segregation measures in GIS. *Computers, Environment and Urban Systems*, *27*(1), 53–70.