*Journal of Statistical Software*

*May 2019, Volume 89, Issue 7.* *doi: 10.18637/jss.v089.i07*

# OasisR: An R Package to Bring Some Order to the World of Segregation Measurement

**Mihai Tivadar**
Université Grenoble Alpes

**Abstract**

Interest in social segregation measurement has increased strongly over the years and the number of segregation indices proposed in the literature have become more complex. However there are only a few software applications that can be employed to analyze social segregation, and these are usually available as a plug-in/package in geographic information system (GIS) software or as limited stand-alone application. Thus, the development of a package which exploits the power and versatility of the R environment for statistical computing and graphics would be desirable. Also, analysis of the segregation indices shows that there are ambiguities and errors in the literature, and consequently in the available software applications. This is an even more important reason why we need to develop a new tool to bring some order to the world of segregation measurement. This paper contributes also by proposing an automatic statistical testing methodology for these indices, using several resampling techniques: randomization tests, bootstrap and jackknife.

*Keywords*: segregation indices, spatial segregation, social segregation, resampling tests, R.

## 1. Introduction

Segregation refers to the organizational (school, occupation, health, etc.) or spatial (residential) separation of social groups. Social segregation is an important issue in modern society because of its consequences for economic efficiency, social cohesion and equity. Over the past few decades, the political agendas in several countries have set objectives and introduced measures to promote socio-spatial diversity, and segregation analyses are being published in official reports and statistics (Iceland, Weinberg, and Steinmetz 2002; Maurin and Schneider 2015).

Despite the growing use of segregation indices and their increased complexity, few software

tools are available. Thus, the first objective of this work is to provide a relatively comprehensive segregation application which exploits the power and versatility of R, an open-source and free statistical software package (R Core Team 2019). In parallel, we conduct a thorough analysis of segregation indices which show that there are contradictions, ambiguities and errors, reflected in existing software tools as well. This is perhaps an even more important reason to develop a new tool that will bring some order to the world of segregation measurement. The third major contribution of our paper is the development of an automatic statistical testing methodology for these indices, which uses several resampling techniques, such as randomization tests, bootstrap and jackknife.

The segregation measurement debate started after World War II, with the work of several sociologists (Jahn, Schmid, and Schrag 1947; Hornseth 1947; Williams 1948; Shevky and Williams 1949; Cowgill and Cowgill 1951). Then the work of Duncans (Duncan and Duncan 1955a,b) heralded an era of "peace", disrupted in the 70's, by several critical articles (Cortese, Falk, and Cohen 1976; Winship 1977; Falk, Cortese, and Cohen 1978). A new period of peace followed publication of Massey and Denton (1988), who empirically and conceptually established a typology of segregation (the segregation dimensions). The development of computation power heralded a new era in which spatial interactions were incorporated explicitly into segregation measurement (Morgan 1983b; Morrill 1991; Wong 1993). A synthesis of this evolution from aspatial to spatial and from global to local measures is provided in Wong (2016). In the new millennium, in addition to the development of spatial and local indices, developments have been made in the direction of multi-group indices (Reardon and Firebaugh 2002; Reardon and O'Sullivan 2004) and of specific measures for ordered groups (Reardon 2009), income segregation (Reardon 2011; Reardon and Bischoff 2011) and activity space (Wong and Shaw 2011; Farber, Páez, and Morency 2012; Farber, O'Kelly, Miller, and Neutens 2015). Also, efforts have been made to respond to the lack of statistical inference methods for segregation measures, such as bootstrap tests (Boisso, Hayes, Hirschberg, and Silber 1994; Lee, Minton, and Pryce 2015), randomization tests/Monte Carlo simulations (Feitosa, Câmara, Monteiro, Koschitzki, and Silva 2007; Tivadar, Schaeffer, Torre, and Bray 2014) and Bayesian inference (Lee *et al.* 2015). For a review of these recent topics in segregation measurement, see Yao, Wong, Bailey, and Minton (2019).

Since "modern" indices are based on spatial information, the first applications were integrated into geographic information system (GIS) software, with small numbers of indices: **ArcInfo** 7 (Wong and Chong 1998), **ArcView** 3.2 (Wong 1996, 2003). Apparicio (2000) developed a **MapInfo** application that computes a large number of indices. The main disadvantage of these types of applications is that they are attached to commercial GIS software and to use the segregation results in regressions, simulations and further analysis, requires an exportation procedure. Also, automatization is not possible.

Another family of tools is comprised of stand-alone applications, first proposed by Konstantinidis and Townshend (1999). However, it was Apparicio and colleagues who proposed two stand-alone applications (Apparicio, Petkevitch, and Charron 2008; Apparicio, Martori, Pearson, Fournier, and Apparicio 2014) that allow computation of large numbers of indices, including complex indices based on spatial information and multi-group measures. There are several important advantages to stand-alone software solutions: Their use is noncommercial and they are user-friendly. The main disadvantage is that they can be used only in a predefined context, and do not allow data manipulation, integration with other software, or automatization.

| Name (Language) | Integration | Indices | Authors |
|---|---|---|---|
| (**AML-Splus**) | **ArcInfo** 7 | 4 | Wong and Chong (1998) |
| (**Avenue**) | **ArcView** 3.2 | 7 | Wong (1996, 2003) |
| (**MapBasic**) | **MapInfo** 4.5 | 24 | Apparicio (2000) |
| **SEGCALC** | standalone | 18 | Konstantinidis and Townshend (1999) |
| **Segregation Analyzer** (C#) | standalone | 42 | Apparicio *et al.* (2008) |
| **GSA** (Java) | standalone | 43 | Apparicio *et al.* (2014) |
| Oasis (R, **PostgreSQL/PostGIS**, **pl-R**, **MapServer**) | web platform | 33 | Tivadar *et al.* (2014) |
| **-seg-** (Stata) | Stata module | 9 | Reardon and Firebaugh (2002) |
| **seg** (R) | R package | 11 | Hong, O'Sullivan, and Sadahiro (2014) |
| **OasisR** (R) | R package | 50 | Tivadar (2019) |

Table 1: Available software tools for segregation analysis.

The Oasis web platform developed by Tivadar *et al.* (2014) is novel in not requiring any software installation, requests are made via the web navigator, and computation is conducted on distant servers. The user has the possibility of wide segregation analysis on a territory (including auto-correlation indices, descriptive statistics and web mapping) using either a historical French data base or their own data. Another original feature of Oasis is that it allows Monte Carlo simulations (permutation tests) to test the statistical significance of the indices. However, this tool basically has the same advantages and disadvantages of standalone applications.

Other applications have been developed as statistical software packages, but have a small number of indices: e.g., Reardon and Firebaugh (2002)'s Stata (StataCorp 2017) module, and Hong and O'Sullivan (2018)'s R package **seg**. The R package **seg** differs in that it is able to compute more recent, surface-based measures, developed in response to the so-called modifiable areal unit problem. Similar to package **seg**, package **OasisR** (Tivadar 2019) is a package implemented in R, an open source software environment for statistical computing and graphics. The main advantage of these tools is that they are flexible, and allow control over many input parameters. They benefit also from the advantage of integration in statistical software which provides the possibility of further analysis without exporting results, integration into other software, automatization, etc. Their main disadvantage is that they require basic knowledge of programming in statistical software.

The article is structured in two main sections. In Section 2, we provide a brief summary of segregation measurement, focusing on aspects that are relevant to the present work, i.e., definition and computation. Section 3 uses some practical examples to show how to use **OasisR**. The article ends with conclusions and further developments.

## 2. Segregation indices

The objective of this paper is not to provide a comprehensive analysis of segregation indices, but rather to unravel several errors and ambiguities, and to provide clearer definitions. The

mathematical formulas are presented in Appendix A.

In line with most of the existing literature, we present the indices developed in **OasisR** following the five dimensions of segregation defined by Massey and Denton (1988). These dimensions are: evenness (population distribution across units), exposure (potential contact between individuals), concentration (space occupied by social groups), clustering (population concentration in contiguous spatial units) and centralization (spatial distribution around the area's center). There are many critics of this classification. According to some, the relevance of centralization has diminished due to the contemporary polycentric form of cities (Brown and Chung 2006), while others believe that the five dimensions could be reduced to a two-dimensional continuum: evenness-clustering and isolation-exposure (Reardon and O'Sullivan 2004), or evenness-concentration and clustering-exposure (Brown and Chung 2006), or separation-location (Johnston, Poulsen, and Forrest 2007). A distinction is made between one-group indices (segregation of a group compared to the rest of the population), between group indices (which measure the segregation between pairs of groups), multi-group indices (which analyze the distribution of several population groups simultaneously) and social diversity indices (which can be understood as zonal multi-group or local indices).

We compare the results of **OasisR** with those from two other software implementations. The geo-segregation analyzer (**GSA** version 1.1) developed by Apparicio *et al.* (2014), is the most complete automatic tool available so far. The application needs an input shape file (data and maps), and produces 43 indices. The second tool is the R package **seg**, version 0.5-1, developed by Hong *et al.* (2014). It computes 11 indices, most available only for a population composed of two social groups. For comparison, we used hypothetical segregation patterns with two groups (Morrill 1991; Wong 1993; Hong *et al.* 2014). Theoretical examples are available in **OasisR** as a data object, adapted from Hong and O'Sullivan (2018). The space is represented by a 10×10 checkboard, with different distributions of the two social groups in the area.

## 2.1. Evenness

Evenness refers to the distribution of different groups across spatial units and can be interpreted as a form of spatial inequality: The more uneven the group distribution compared to other social groups, the more segregated is that group. This is the reason why several evenness indices are based on a spatial form of the Lorenz inequality curve, also called segregation curve.

### *Standard evenness indices*

Indices were introduced initially by Jahn *et al.* (1947) to measure "ecological" segregation between black and white populations. Duncan and Duncan (1955a) demonstrated the mathematical relationships between these indices, and provided a graphical interpretation. They showed that the information provided by previous indices could be derived from the dissimilarity index $D^{k_1 k_2}$ and the social group proportions. The dissimilarity index can be interpreted as the share of a group $k_1$ that would have to move to achieve an even distribution compared to group $k_2$. Similar to many other indices, the index was defined in the context of a two-group population (minority vs. majority).

Duncan and Duncan (1955b) adapted the dissimilarity index to a one-group form, known as Duncan's segregation index $IS^k$. It measures the dissimilarity between a group $k$ and the rest of the population. In the case of two group populations, the segregation and dissim-

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| Duncan segregation | × | × | – |
| Duncan dissimilarity | × | × | only for 2 groups |
| Gini | × | × | – |
| Gini2 | × | – | – |
| Atkinson | × | limited to 3 values of $\delta$ | – |
| Gorard | × | – | – |
| Entropy | × | errors for 2 groups | – |

Table 2: Standard evenness indices comparison.

ilarity indices are identical. The Gorard segregation index $GS^k$ (Gorard and Taylor 2002) is a slightly different form which computes the dissimilarity between a group and the total population. Gorard's index has some disadvantages (the upper boundary is less than 1, the index is asymmetric) but has the advantage that unlike dissimilarity based indices it is a strong composition invariant index.

Another standard one-group index based on the segregation curve is $G^k$, the spatial version of the Gini index (Gini 1921) adapted by Duncan and Duncan (1955a). We developed a between group form of Gini, $G2^{k_1k_2}$, by computing the index for a sub-population formed by two groups.

The Atkinson index $A^k$ (Atkinson 1970) was adapted to a segregation context by James and Taeuber (1985), with the mathematical formula corrected by Massey and Denton (1988). Compared to other indices based on segregation curves, the Atkinson index allows the researcher to decide the weights of the spatial units in different zones of the segregation curve by introducing an inequality aversion parameter $\delta$ with values between 0 and 1. If $\delta < 0.5$, the spatial units where the minority is underrepresented compared to the average, contribute more to the segregation. The reverse is valid for $\delta > 0.5$.

The entropy index (or the information index) was proposed by Theil (Theil and Finizza 1971; Theil 1972) as an index of school segregation for a two-group population. It measures the departure from evenness as the population weighted average deviation of each spatial unit from the area's entropy (or social diversity). In the case of a population with more than two groups, the local and area entropy need to be calculated for each group as the "minority" and the rest of the population as the "majority".

The results obtained using **OasisR** and **GSA** are identical apart from the entropy index. According to its definition, the entropy index for a two group population should have the same value for both groups. If we take the example of complete segregation from the theoretical distributions, the index should be equal to 1 (as in **OasisR**), while **GSA** provides $H^1 = 0.75$ and $H^2 = 0.25$. Using empirical data with several social groups produces identical results. The Gorard index is computed only in **OasisR**, and in the case of the Atkinson index, the user is limited in **GSA** to three standard values of the inequality aversion parameter $\delta$ (0.1, 0.5 and 0.9). The **seg** package computes only the dissimilarity index for individual pairs of groups.

*Spatial evenness indices*

Spatial evenness indices were developed by geographers in response to a major criticism of standard segregation indices: Although satisfactory for organizational segregation studies,

they seem less appropriate in a geographical context where segregation is a "separation created by spatial structure" (Wong 1993, p. 559). For instance, if we make random permutations of the populations between spatial units, standard evenness indices do not change, but the social structure of the area is obviously different. Spatial evenness indices are based on the dissimilarity index, and were defined in the context of a two-group population. The first developments were made by Jakubs (1981) and Morgan (1983a) but they require complex linear programming methods.

Morrill (1991) developed a contiguity modified dissimilarity index $D^{k_1k_2}(adj)$, where the probability of contact between groups is modeled via the contiguity matrix: Interactions between groups emerge if two spatial units are adjacent. In the case of $D^{k_1k_2}(adj)$ with a population formed by more than two groups, an ambiguity arises from the spatial interaction term: It is not sufficiently clear how the population proportions across spatial units should be computed. In the original paper, the author uses total populations $t_i$ while Wong and Chong (1998) present these totals specifically as the sum of two groups $t_i^{k_1k_2}$ . This difference has consequences for the index if the population is composed of more than two groups. The only detailed generalized formula is that proposed by Apparicio *et al.* (2008) but it seems incorrect since the proportions are determined using the entire population. Instead, the partial total population should be used because Morrill's index is based on the dissimilarity index which compares the distributions between each pair of groups independently of the others, and it seems logical to assume that spatial potential interactions should also take account only of each pair of groups (see Appendix A).

For the theoretical two group distributions, **OasisR**, **GSA** and **seg** provide identical results. In the case of more than two groups, Morrill's index is computed only in **OasisR** and **GSA**. The results in **GSA** are incorrect since the spatial interaction term is based on population totals rather than the groups involved in the comparison. An empirical confirmation is provided by the fact that the result matrix is not symmetrical as it should be (the dissimilarity between two groups is by construction symmetrical).

Apparicio *et al.* (2008) adapted the original index to construct d Morrill's segregation index $IS^k(adj)$ (one-group version of the index). Similar to Duncan's dissimilarity and segregation indices, $IS^k(adj)$ can be interpreted as the dissimilarity between group $k$ and the rest of the population, and the use of a group's proportion within the total population of each unit in the spatial interaction is correct (see Appendix A). Computation of the index gives the same results in **OasisR** and **GSA** and is not provided in **seg**.

One limitation of Morrill's indices is that they take account only of direct interactions between adjacent spatial units, and it would be interesting to expand these interactions further in space. One solution would be to go beyond the first order contiguity by generalizing Morrill's indices to the $k$th order contiguity $D^{k_1,k_2}(Kadj)$. The contiguity order is considered to have a negative effect on spatial interactions. Generalized Morrill's indices are computed only in **OasisR**, and we propose two forms for the spatial function: negative exponential and reciprocal.

Wong (1993) developed two indices for a population with two groups: $D^{k_1k_2}(w)$, where spatial interactions between contiguous spatial units are proportional to the length of the shared boundary and $D^{k_1k_2}(s)$, which also includes the perimeter/area ratio. In Wong's original paper both indices have errors in their mathematical definition. The first error is the division by 2 of the spatial interaction term and the second is the row standardization of the spatial

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| Morrill's segregation | × | × | – |
| Morrill's dissimilarity | × | errors for more than 2 groups | only for 2 groups |
| Generalized Morrill's | × | – | – |
| Wong's segregation | × | errors | – |
| Wong's dissimilarity | × | errors | only for 2 groups |
| User's spatial matrix definition | × | – | only for 2 groups |

Table 3: Comparison of spatial evenness indices.

matrix instead of its overall standardization. Despite these problems, the results provided by Wong (1993) using the theoretical distributions are correct. If we re-compute the index according to Wong's formal definition, the results are not coherent, but if we adapt Morrill's index by replacing the contiguity matrix with the shared boundary matrix (as described in Wong's article) we obtain identical results.

Hong *et al.* (2014) obtained the same results for the theoretical examples but the authors do not provide the mathematical definition of Wong's indices. On his personal page, Hong (2014) developed scripts to present the R package **seg** and similar to **OasisR**, he defines the spatial matrix using a global standardization. Wong and Chong (1998) presented improved versions of the indices formulae, where the proportions in the spatial interaction effect are clearly defined, but the definition of the spatial interaction matrix seems incorrect since Wong and Chong (1998) use double standardization of the spatial matrix (row standardization followed by overall standardization).

Furthermore, there are ambiguities concerning the definition of each spatial unit's perimeter, necessary for the computation of $D^{k_1 k_2}(s)$. To obtain the same results as in Wong (1993) and Hong *et al.* (2014), we need to use only the "internal" perimeter of each spatial unit, defined as the sum of the boundaries shared with other spatial units, and ignore the area's external borders. To overcome this issue, in **OasisR** the user can choose the perimeter definition.

For the theoretical two-group distribution, using the corrected mathematical formula (see Appendix A) and the "internal" definition of the perimeter, **OasisR** provides the same results as the **seg** package and Wong's original paper. The results of **GSA** are incorrect because the software uses the original wrong mathematical definitions. We also generalized Wong's indices to a case with more than two groups, and its one-group form. Apparicio *et al.* (2008) define these indices mathematically but these definitions have similar problems to Morrill's index generalization, and carry the errors from the original definition. Finally, **seg** and **OasisR** allow the user to compute a modified version of the index using their own definition of the spatial interaction matrix.

## 2.2. Exposure

Exposure measures the potential contact between members of the same group (isolation) or between members of different groups (interaction) as the probability that they live in the same spatial unit.

*Standard exposure indices*

The first exposure indices were developed by Shevky and Williams (1949), and were normal-

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| Isolation | × | × | × |
| Interaction | × | × | × |
| Eta2 | × | × | – |
| Spatial isolation/interaction | × | – | × |
| Distance-decay isolation/interaction | × | errors | – |

Table 4: Exposure indices comparison.

ized and explained as a probabilistic model by Bell (1954). The notations of these indices were introduced by Lieberson (1981).

The isolation index $xPx^k$ is defined as the probability that a group member shares the same spatial unit with another member of the same group. Without the computational power of a computer, it was difficult to calculate the index according to its definition and Bell (1954) provided an approximate version of the index (see Appendix A for details). Presently, there are no particular reasons not to compute its exact value $xPx^{k^*}$, and in **OasisR**, the user has the possibility to choose between the two versions.

The isolation index can be adjusted to control for the effect of population composition, which has a strong effect on the index value. Bell (1954) also developed the normalized isolation index (an approximate version) which is equivalent to the correlation ratio $Eta2^k$ (White 1986) and to the mean square contingency or phi square[1] for a dichotomous population (Duncan and Duncan 1955a). Since the index can be computed in different ways (Bell 1954; Coleman 1966; Zoloth 1976), debate emerged over its dimension and interpretation (James and Taeuber 1985; Massey and Denton 1988; Stearns and Logan 1986).

The interaction index $xPy^{k_1k_2}$ (Lieberson 1981) is a between group segregation measure which computes the probability that a member of a group $k_1$ shares the same spatial unit with a member of group $k_2$. Similar to the isolation index, we can compute its exact or approximate value. The results of all the approximate standard exposure indices are the same in **OasisR**, **seg** and **GSA**. The exact versions can be computed only in **OasisR**.

*Spatial exposure indices*

Morgan (1983b) developed two exposure indices that take explicit account of the distance between spatial units, which influences the potential contact between members of the same social group (distance-decay isolation index $DPxx^k$ ) or different groups (distance-decay interaction index $DPxy^{k_1k_2}$). The hypothesis is that people also come into contact outside of their own spatial units, and the number of potential contacts increases with distance, but their intensity decreases.

Similar to the other indices based on distance, the use of a gravity exponential function makes the result sensitive to the distance measure. There are some ambiguities about the definition of distance within a spatial unit since it could be null or a function of the spatial unit shape (area, perimeter). In **OasisR**, the user can choose between different spatial matrix diagonal definitions: null, 0.6 of the area's square root, as proposed by White (1983), and a user matrix.

In **GSA**, the distance within a spatial unit is considered null. Results for the linear definition

[1] Williams (1948) defined the mean square contingency or phi square as a conversion of chi square for a population with two groups into an index (from 0 to 1) by dividing it by the total population.

of the distance are identical in **GSA** and **OasisR**. For the gravity version of the index, **GSA** results are incorrect: The diagonal of the distance matrix remains null after transforming to exponential form, and should be at its maximum level as $\exp(0) = 1$ (highest spatial interaction). If we compute these indices for theoretical two group patterns using the wrong null exponential diagonal, we obtain the same results as **GSA**. For more than two groups, the results provided by the two applications are different. **OasisR** and **GSA** provide metric conversion options (measure in and measure out), necessary for comparison between studies, and to avoid situations where indices cannot be computed because of the digital approximations that rapidly approach zero in the negative exponential distance function. These indices are not computed in **seg**.

Reardon and O'Sullivan (2004) develop several spatial indices, including a spatial version of the exposure/isolation index. The spatial exposure index is computed as the average percentage of a group within the local environment of each member of another group. The spatial isolation of a group is simply the spatial exposure of a group to itself. In **OasisR** we used only the functions developed by Hong and O'Sullivan (2018) in the **seg** package, formatting the output as the other **OasisR** functions.

## 2.3. Clustering

In clustering, the more contiguous spatial units occupied by a group (forming an enclave in the area) the more segregated that group. There are arguments in the literature about the need for a separate dimension since modern evenness indices take explicit account of the phenomena of space and clustering (Reardon and O'Sullivan 2004). The distinction between evenness and spatial clustering might be just an artifact of the reliance on spatial subareas at some chosen geographical scale of aggregation (evenness at one level of aggregation is strongly related to clustering at a lower level of aggregation). For Brown and Chung (2006), clustering and exposure constitute a single dimension since high clustering is a manifestation of low exposure, and vice versa: If the members of a group are located close to each other, especially in a large cluster, their exposure to other groups will be reduced.

*Proximity measures*

The first proximity indices were introduced by White (1983) for two groups, and later generalized to apply to more than two groups by White (1986): the mean proximity between the members of the same group $Pxx^k$ (one-group index), and between two different groups $Pxy^{k_1k_2}$ (between-group index), and the mean proximity between persons in the area without regard to the group $Poo$ (multi-group index). In the original papers, White proposed considering the distance within a spatial unit as non-null, and advised a function of the area ($0.6\sqrt{A}$) but a null diagonal distance matrix is most commonly used in the literature and computed using software packages (Apparicio *et al.* 2014; Tivadar *et al.* 2014). In **OasisR,** the user can choose among these options or exploit a user value. These measures can be determined using a linear function of the distance. The result represents the average distance between individuals (from the same or different groups). With a gravity form such as the exponential of the negative distance, the measure becomes an index. As for the other distance based measures, the exponential function makes the result sensitive to spatial measure units. Therefore, a metric converter is provided in **OasisR**.

By using a null distance within spatial units and a linear distance matrix in **OasisR**, the

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| One-group mean proximity | × | only null diagonal, errors | – |
| Between group mean proximity | × | only null diagonal, errors | – |
| Multi-group mean proximity | × | – | – |
| Multi-group mean proximity (between group) | × | – | – |
| Spatial proximity index (multi-group) | × | – | errors |
| Spatial proximity index (one-group) | × | – | – |
| Spatial proximity index (between group) | × | errors | – |
| Absolute clustering | × | only contiguity, errors | – |
| Relative clustering | × | only null diagonal | – |
| Relative clustering (linear) | × | – | – |

Table 5: Proximity measures and clustering indices comparison.

results for $Pxx^k$ and $Pxy^{k_1k_2}$ are the same as for **GSA**[2]. In relation to the other distance-based indices, their gravity form is incorrect in **GSA**: The diagonal for the negative exponential distance matrix is null but should be equal to 1. If we use this incorrect spatial definition, we obtain the same results for $Pxx^k$ but different ones for $Pxy^{k_1k_2}$. There is probably an additional error in the **GSA** computation, as the result matrix is not symmetric as it should be (the spatial proximity is identical if we permute the groups). Moreover, we tested the mathematical properties that these indices should respect for two group populations (White 1983); they hold only for **OasisR**. The **seg** package does not compute these measures.

Based on proximity measures, White (1983) computes a segregation statistic called spatial proximity which is simply the average of one-group proximities, weighted by the fraction of each group in the population. The initial index was defined for the case of two groups which created certain ambiguities related to its nature (between group or multi-group index) since the result is the same $SP = SP^{1,2}$ . White (1986) generalized the index to more than two groups by using the multi-group form of the index but with an error in the mathematical definition since the populations are squared. In contrast, Apparicio *et al.* (2008) keep the between group definition if the population includes more than two groups, and compute the index ignoring the rest of the population. This means also that the mean proximity between persons regardless of group should have a between group form $Poo^{k_1k_2}$, as used to compute $SP^{k_1k_2}$ . The spatial proximity index is computed using only the gravity form, but can also be used with linear distance in the opposite interpretation. If we compute the gravity form of the spatial proximity index using the wrong null diagonal, we obtain the same result as with the **seg** package (which computes only the multi-group form) and **GSA** (which provides only the between group version). The index can also be computed as the one-group version $SP^k$, which compares the proximity among the members of a group $Pxx^k$ and the average proximity of the population $Poo$.

### *Clustering indices*

Massey and Denton (1988) propose two clustering measures. The absolute clustering index $ACL^k$ expresses the average number of members of groups in nearby spatial units as a proportion of the total population in those proximate spatial units. Spatial interactions can

[2] **GSA** does not provide results for $Poo$.

be computed using the contiguity matrix, although the $ACL^k$ index could produce negative values despite Massey and Denton (1988)'s claim that its values always range between 0 and 1. The problem arises from missing information about the particular form of the contiguity matrix: Contiguity between a spatial unit and itself should be equal to 1 (Konstantinidis and Townshend 1999). As in White (1983), the index also has a gravity form (exponential of the negative distance), and it is recommended to use a non-null diagonal of the spatial interaction matrix as a function of the area. Using the gravity form, the index is subject to the same issue of sensitivity to the distance measure.

**GSA** provides only the contiguity form of the index, and the results appear incorrect, independent of the contiguity matrix diagonal definition (0 or 1). We used a very simple case of a theoretical 2×2 grid, where the first cell is inhabited exclusively by the minority, and all other cells include the majority. With a null contiguity matrix diagonal, the index has aberrant values (negative or superior to 1), and with a diagonal equal to 1, the index should be 0 for both groups which is not the case for **GSA**.

The relative clustering index $RCL^{k_1k_2}$ is a between-group index based on White's proximity measures which compares the average distance between the members of one group to the average distance between the members of another group. The index is computed using only the gravity form but we can easily adapt the index to linear distance. If we use the wrong null diagonal in the distance matrix for the exponential form of the index, the results in **OasisR** are similar to **GSA**.

## 2.4. Concentration

According to Massey and Denton (1988, p. 289) "the concentration refers to relative amount of physical space occupied by a group". The first index to measure spatial concentration is the Delta index $\Delta^k$, proposed by Hoover (1941) and adapted by Duncan, Cuzzort, and Duncan (1961). This is a dissimilarity index between the distribution of a group and the distribution of available space. Massey and Denton (1988) developed an absolute concentration index $ACO^k$, by comparing the average area inhabited by a group to the average land area they would inhabit under maximum spatial concentration (if they were all located in the smallest areal units). The relative concentration index $RCO^{k_1k_2}$ (Massey and Denton 1988) takes the ratio of one group concentration to another group concentration, and compares it to the maximum possible ratio that would be obtained if the first group was maximally concentrated and the second minimally concentrated. The index is standardized to obtain values between $-1$ and 1, but in contrast to what Massey and Denton (1988) claim, the index can be smaller than $-1$. Moreover, Egan, Anderton, and Weber (1998) identify several mathematical and conceptual problems with that index which is why $RCO^{k_1k_2}$ is no longer used in Census Bureau analyses (Iceland *et al.* 2002). In its mathematical formula, intermediary sums do not have the indices required to identify the maximum/minimum concentrations for each group which can lead to ambiguities. In the original paper, these parameters are presented as "defined as before", but they should differ from one group to another (see Appendix A).

It is impossible to compute concentration indices for the theoretical distributions (the denominator is null since spatial units have the same size). Thus, we use empirical examples for comparisons between **OasisR** and **GSA**. For one-group indices ($\Delta^k$ and $ACO^k$) the results are identical while for relative concentration only some of the results are the same. There is an error in **GSA** since the matrix should not be symmetric.

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| Delta | × | × | – |
| Absolute concentration | × | × | – |
| Relative concentration | × | errors | – |

Table 6: Concentration indices comparison.

## 2.5. Centralization

According to Massey and Denton (1988), centralization is the degree to which a group is spatially located near the center of an area. The first true centralization index was developed by Duncan and Duncan (1955b) to introduce some spatiality into segregation measuring. The index was presented as a one-group index, so we describe it as Duncan's absolute centralization index. The literature uses the relative centralization index $RCE^{k_1 k_2}$, adapted and proposed by Massey and Denton (1988)[3] which measures the extent of one group's centralization relative to another. These centralization indices are particular forms of the Gini index, and measure the localization unevenness of two groups around a specific point (the center) by ordering spatial units according to their distance from the center. Massey and Denton (1988) introduced an absolute centralization index $ACE^k$ which compares the spatial distribution of a group to the distribution of available land around the area's center. Since $ACE^k$ computation needs information on area, the results can sometimes contradict $RCE^{k_1 k_2}$. For this reason, in **OasisR**, we also compute the mathematical adaptation of $RCE^{k_1 k_2}$, to correspond to Duncan and Duncan (1955b)'s original description ($DCE^k$). With the exception of Duncan's centralization index, provided only by **OasisR**, the results of the other indices are similar to **GSA**.

One of the reasons why centralization lost popularity in the literature was that this dimension has little meaning in relation to increasingly polycentric and sprawling modern cities. To resolve this issue, we adapt the centralization indices to a polycentric spatial configuration. The option retained is to compute the distance between spatial units and each center, and to take account only of the distance to the closest point. This method is implemented in **OasisR** by the `RCEPoly`, `ACEPoly`, and `ACEDuncanPoly` functions. According to Folch and Rey (2016), we can spatially limit the effect of centrality. We consider two options: defining a parameter $k$ as the number of nearest neighbors affected by each center, or choosing the distance of influence $k_{\text{dist}}$. The constrained version of the index can be computed only for the indices developed by Duncan and Duncan (1955b) (`RCEPolyK` and `ACEDuncanPolyK`).

## 2.6. Other measures: Diversity, multi-group and local indices

*Social diversity indices*

Diversity indices measure the level of social diversity in an area, without taking account of the spatial distribution of different groups. Shannon-Wiener index $H_{SW}$ (Shannon 1948) is based on the entropy concept and measures the heterogeneity of a population from perfect homogeneity (0) to maximum heterogeneity (natural logarithm of the number of groups). The normalized version $\overline{H}_{SW}$ is obtained by dividing it by its maximum. Simpson's interaction

[3]There is a minor error in Massey's mathematical formula since the sums should start from 2 (Massey and Denton 1988, p. 292).

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| Relative centralization | × | × | – |
| Polycentric relative centralization | × | – | – |
| Constrained/local relative centralization | × | – | – |
| Duncan's absolute centralization | × | – | – |
| Duncan's polycentric absolute centralization | × | – | – |
| Duncan's constrained/local absolute centralization | × | – | – |
| Massey's absolute centralization | × | × | – |
| Massey's polycentric absolute centralization | × | – | – |

Table 7: Centralization indices comparison.

| Index | **OasisR** | **GSA** | **seg** |
|---|---|---|---|
| Diversity index | × | – | – |
| Normalized diversity index | × | – | – |
| Simpson's index | × | – | – |
| Multi-group dissimilarity | × | × | × |
| Multi-group Gini | × | × | – |
| Multi-group normalized exposure | × | × | × |
| Multi-group information theory index | × | × | × |
| Multi-group relative diversity | × | × | × |
| Multi-group squared coefficient of variation | × | × | – |
| Deviational ellipse index | – | × | – |
| Spatial multi-group dissimilarity | × | × | × |
| Spatial multi-group relative diversity index | × | – | × |
| Spatial multi-group information theory index | × | – | × |
| Ordinal information theory index | × | – | – |
| Ordinal variation ratio index | × | – | – |
| Ordinal square root index | × | – | – |
| Ordinal absolute difference index | × | – | – |
| Rank-order information theory index | × | – | – |
| Rank-order variation ratio index | × | – | – |
| Rank-order square root index | × | – | – |
| Poulsen typologies | – | × | × |
| Location quotient | × | × | – |
| Local diversity | × | – | – |
| Local entropy | × | × | – |
| Local Simpson | × | – | – |

Table 8: Diversity, multi-group and local indices comparison.

index $I_S$ measures the probability that individuals selected randomly from the area (regardless of their location), do not belong to the same social group (Simpson 1949). These indices are available only in **OasisR**.

*Multi-group indices*

We treat multi-group segregation indices separately since their appearance in the literature is later, and only some can be attributed to standard dimensions of segregation. These indices analyze the distribution of several population groups simultaneously. Reardon and O'Sullivan (2004) define a general approach to measuring spatial multi-group segregation for several indices: multi-group normalized exposure index $P^*$ (James and Taeuber 1985; Reardon and Firebaugh 2002) and a set of general multi-group spatial/clustering indices: multi-group information theory index $H^*$ (Theil 1972; Reardon and Firebaugh 2002)[4], multi-group relative diversity index $RD^*$ (Carlson 1992; Reardon 1998), and multi-group dissimilarity index $D^*$ (Morgan 1975; Sakoda 1981). Other multi-group indices provided in **OasisR** are multi-group Gini index $G^*$ (Reardon 1998) and the squared coefficient of variation $C^*$ (Reardon and Firebaugh 2002). Spatial versions of certain multi-group indices (dissimilarity, information theory and relative divesirty) were developed by Reardon and O'Sullivan (2004). For these spatial versions of multi-group indices, we formatted only the output of an existing function in the **seg** package. The results obtained in **OasisR**, **GSA** and **seg** are identical if the indices are available.

In addition to the previous measures, we developed functions in order to compute two specific types of multi-group segregation indices. Using the variation ratio approach described in Reardon and Firebaugh (2002); Reardon (2009) proposes four indices adapted to the particular case of groups defined by ordered categories: ordinal information theory index, ordinal variation ratio index, ordinal square root index, and ordinal absolute difference index. Reardon, Firebaugh, O'Sullivan, and Matthews (2006) and then Reardon (2011) and Reardon and Bischoff (2011), developed rank-ordered indices (rank-order information theory index, rank-order variation ratio index, and rank-order square root index), adapted from the ordinal methodology, to analyze segregation using a continuous variable (but not necessarily one that is interval-scaled) such as income. In practice, data on income distribution is available in classes ordered by income thresholds. Empirically, the methodology includes the following steps: First, for each threshold, we compute the corresponding ordinal segregation indices (ordered information theory, variation ratio, and square root index) between those above and below the income threshold; second, we fit a polynomial regression model to approximate the information theory/variation ratio/square root function; third we use the model's estimated coefficients to compute an estimate of the rank-order indices. All these outputs are included in **OasisR**.

*Local indices*

Local indices can be mapped which allow us to identify spatial patterns in the area. First, the location quotient $LQ_i^k$ (Isard 1960) identifies spatial units where a group is over-represented or under-represented. Moreover, the social diversity indices can be computed at the local level. The local entropy index $H2^i$ which is equivalent to $\overline{H}_{SW}$ (Theil 1972; Theil and Finizza 1971), measures social diversity within each spatial unit ($H2^i = 0$ for a homogeneous population and $H2^i = 1$ for maximal diversity, when all groups are equal in size). The results are identical in **OasisR** and **GSA**. We can adapt the local level Shannon's diversity index and Simpson's interaction index which are available only in **OasisR**. **GSA** provides Poulsen's typology (Poulsen, Johnston, and Forrest 2010, 2011) which is not developed in **OasisR**.

[4]Similar to the entropy index, computation of the multi-group version ignores the local entropy if a group is missing.

## 2.7. Resampling tests

In contrast to other tools, **OasisR** offers functions that allow the statistical testing of indices using resampling techniques (randomization tests, bootstrap and jackknife) based on individual or unit sampling.

*Random population distribution as comparison*

The idea that segregation should be analyzed as the deviation from a random pattern rather than from a complete theoretical desegregation was introduced at the end of the 1970s by Cortese *et al.* (1976) and Winship (1977). Winship (1977) considers the binomial distribution as a natural model for random segregation, while Cortese *et al.* (1976) and Falk *et al.* (1978) suggest using a hypergeometric distribution. Assuming these statistical hypotheses, it is possible to parameterize Duncan's dissimilarity index distribution but a generalization for other segregation indices is not feasible. In a more recent paper, Ransom (2000) examines the sampling distributions of dissimilarity and Gini indexes by deriving their exact sampling distributions, and developing asymptotic inference procedures. Allen, Burgess, Davidson, and Windmeijer (2015) developed this framework further, and show that the use of bootstrap methods can improve test procedures.

Resampling methods are valid, nonparametric alternatives to conventional inferential statistics. These methods are particularly interesting in the context of segregation analysis, because the data used often are a sample of the total population, and even if the analyst has data on the entire population, there is a risk of data collection and manipulation errors. Resampling allows us to create simulated distributions of the indices as the basis for testing different null hypotheses which depend on the resampling technique and the sampling unit (individual or spatial unit).

*Randomization tests*

Permutation tests (also called randomization tests or exact tests) are statistical significance tests in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the statistic under rearrangements of the labels on the observed data. If the number of possible combinations is too high, we can use asymptotically equivalent tests such as Monte Carlo permutation (or approximate permutation or random permutation) tests. First, we generate random population distributions, and for each replicate we compute the index which gives us the simulated reference distribution. This allows us to compute a pseudo $p$ value that is equal to 1 minus the relative rank of the index in the reference distribution (Anselin 2003) which can be interpreted as a statistical significance test of the hypothesis that the segregation index is the result of random processes.

In the case of individual sampling units, each individual makes a random draw without replacement among spatial units, according to a probability vector. Location probability can be identical or constrained, e.g., by unit population (Cortese *et al.* 1976), or by area (Tivadar *et al.* 2014). As in Findlay and Findlay (1984) and Carrington and Troske (1997), we generate random localization patterns by resampling directly from the original data (aggregation of individual independent random draws) instead of sampling with theoretical distribution (Boisso *et al.* 1994; Tivadar *et al.* 2014). If we set sampling on spatial units (Feitosa *et al.* 2007; Tivadar *et al.* 2014), we have a similar framework to the permutation test developed in spatial auto-correlation analysis (Anselin 1995). Random localization is obtained from per-

mutations of entire populations among spatial units which allows us to test the significance of the spatial component of the index.

### *Bootstrapping*

Because segregation measures are often computed from sample data, distributional information required to test hypotheses can be obtained by the bootstrap method (Efron 1979) used to estimate the distributions of a statistic by resampling with replacement from the data set. These distributions can be examined in order to establish a probability that the statistic's value will include the value implied under the null hypothesis. Thus, bootstrap techniques allow us to construct confidence intervals around the original point estimate, using Efron's percentage method (Efron 1979).

If we consider individual sampling, the construction of bootstrap distributions is achieved by using resampling directly from the original data – the alternative being to use draws from a theoretical distribution as in Boisso *et al.* (1994). The method is appropriate especially if the initial data are based on a population sample. If the initial data are based on a sample of units, then bootstrap sampling should be applied at the unit level. This technique can be employed for spatial segregation analysis (Lee *et al.* 2015), but seems more appropriate for the analysis of organizational segregation (Carrington and Troske 1997).

### *Jackknife*

Although temporally jackknife preceded bootstrap (Quenouille 1956; Tukey 1958), the method is similar to the bootstrap, and is used in statistical inference mainly to estimate the bias and standard error (variance) of a statistic. The simulated index distribution is obtained by systematically recomputing the statistic, excluding one or more observations at a time from the sample set.

Jackknife with individual sampling seems less useful in the context of segregation analysis, but is a particularly interesting method in the case of unit sampling because it allows detection of replicates that represent outliers. If a significant inferior outlier is found in the index reference distribution, this means that without a specific spatial unit, the segregation level would be significantly lower and implies that this spatial unit is playing a significant role in segregation. To our knowledge, the jackknife technique has been used in the segregation literature only by Massey (1978) in order to estimate dissimilarity index variances. **OasisR** allows several automatic standard outlier detection techniques, such as boxplot and standard deviation methods, and different score methods (normal, $t$ Student and chi-squared scores) and the median absolute deviation method, based on functions developed in the **outliers** package (Komsta 2011).

### *Bayesian inference*

Lee *et al.* (2015) proposed a new method for estimating the dissimilarity index and quantifying its uncertainty, based on a Bayesian hierarchical modeling approach (with inference based on Markov chain Monte Carlo simulation). The authors consider two distinct models: a globally smooth model (binomial generalized linear mixed model, where the set of random effects are spatially auto-correlated) and a locally smooth model which allows geographically adjacent areal units to have very similar or very different minority proportions. In both cases an estimate and a 95% credible interval for the dissimilarity index can be obtained, by

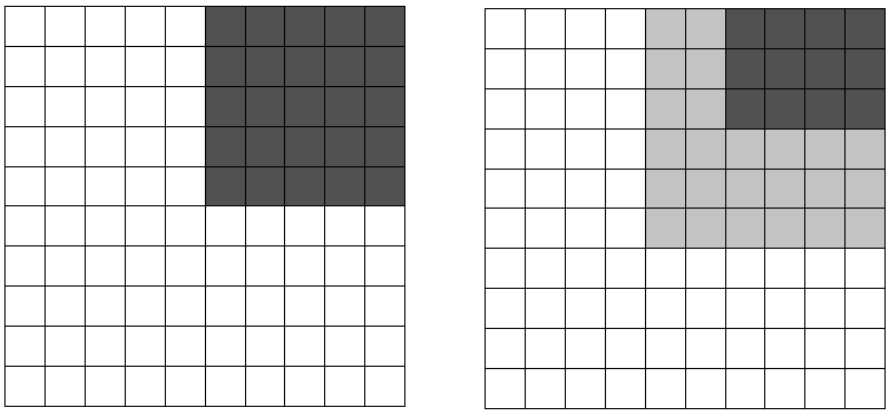

Figure 1: Theoretical examples.

computing the posterior predictive distribution of the index. The model was implemented using the **CARBayes** package (Lee 2013) in the R software environment. This methodology represents an opportunity for future developments applying the Bayesian spatial modeling approach to other segregation measures.

## 3. Using OasisR

The R functions developed were designed to address many different situations in the easiest way possible. For all **OasisR** aspatial segregation functions, the only input required is the population distribution table within units. The table should not include row totals (unit total populations) which could be interpreted as a supplementary social group. For spatial indices, a second necessary input is spatial data, which can be provided in three ways (see the example below). Many functions have specific parameters, but their input is not obligatory since by adopting their default values, functions compute the usual form of indices. For a detailed description of the parameters, see the **OasisR** manual (Tivadar 2019). Generally, the output of segregation functions is a numeric value for multi-group indices, a vector of each group's index value for one-group indices, and a numeric matrix for between group indices. Values are rounded to four digits.

As support, we use a very simple 10×10 grid theoretical example which is used in many studies (Morrill 1991; Wong 1993; Lee *et al.* 2015; Hong *et al.* 2014); the data are provided by the package. From the various available distributions (Hong 2014), we chose two situations: one with complete segregation of the minority, and one with a particular social groups mix. The population in the dark gray cells is formed only of minority members, in the white cells it is formed only of majority members, and in the light gray ones the population contains a mix of the two groups. In each cell, we consider that the total population number is 100 individuals. For more complex examples, see Appendix B.

To compute an aspatial segregation index, the script is basic since we need to use the function name and a distribution table:

```
R> A <- segdata@data[, 1:2]
```

```
R> DIDuncan(A)

     [,1] [,2]
[1,]    0    1
[2,]    1    0
```

In the case of spatial indices, there are three ways to introduce spatial data in **OasisR**. The first solution is to provide a spatial R object, using the `spatobj` argument. It is possible also to import a shapefile, by using two arguments: (1) `folder`, to provide the path where the shapefile is located on the drive, and (2) `shape`, the name of the shapefile (without an extension). The shape import uses the `readOGR` function in the **rgdal** package (Bivand, Keitt, and Rowlingson 2019). Finally, spatial information can be provided directly as vectors/matrices of contiguity, common boundaries, areas, distances, etc. This spatial information can be computed within **OasisR** because the package includes geographical functions based on the **spdep** (Bivand, Pebesma, and Gómez-Rubio 2013; Bivand 2019) and **rgeos** (Bivand and Rundel 2019) packages, with appropriate output for the segregation functions.

```
R> foldername <- system.file("extdata", package = "OasisR")
R> shapename <- "segdata"
R> areavector <- area(segdata)
R> Delta(A, spatobj = segdata)

[1] 0.75 0.25

R> Delta(A, folder = foldername, shape = shapename)

OGR data source with driver: ESRI Shapefile
Source: "C:/R/win-library/3.3/OasisR/extdata", layer: "segdata"
with 100 features
It has 19 fields
[1] 0.75 0.25

R> Delta(A, a = areavector)

[1] 0.75 0.25
```

Certain supplementary arguments emerge for specific indices. For example, in the `Atkinson` function, inequality aversion can be set via the `delta` argument. The argument `variant` specifies which variant of Wong's indices is chosen: `variant = "w"` or `variant = "s"`. Section 2.3 showed that certain proximity measures can be computed as multi-group, between group or one-group indices. The user can choose the index type via the argument `itype`. In the case of exposure indices $xPx^k$ and $xPy^{k_1k_2}$, the logical argument `exact` determines whether indices are computed using the approximate or exact definition. The functions `spatmultiseg` and `rankorderseg` have several arguments specific to the **seg** package. For rank-ordered measures, `polorder` gives the order of the of polynomial regression model.

```
R> B <- segdata@data[, 7:8]
R> xPy(B)
```

```
       [,1]   [,2]
[1,] 0.7500 0.2500
[2,] 0.0789 0.9211

R> xPy(B, exact = TRUE)

       [,1]   [,2]
[1,] 0.7475 0.2525
[2,] 0.0797 0.9203
```

Spatial segregation functions based on distance have particular arguments. Spatial interactions can be defined via the `fdist` argument: `"l"` for the linear and `"e"` for the exponential inverse function of distance. Other distance functions can be used by introducing a user distance matrix in the R functions, and by setting a linear function. Metric conversions are based on the `conv_unit` function in the **birk** package (Birk 2016), with `distin` and `distout` the respective arguments for the input and output measures. Argument `diagval` defines the distance within a spatial unit: `"0"` for the null diagonal and `"a"` for White's formula (0.6 square root of the area). Other versions can be used by introducing a user distance matrix in the function. Examples of how to use these functions are provided in Appendix B. Indices based on the contiguity matrix have a supplementary logical argument `queen`, to choose the criterion used for contiguity matrix computation: `TRUE` for queen and `FALSE` for rook (by default). For centralization indices, the user must introduce the argument `center` which is the number of the spatial unit in the table representing the area's center . For polycentric versions, the input must be a vector. For measures based on the generalized contiguity matrix ($K$-order matrix), two arguments can shape spatial interactions: argument `K` represents the order of the contiguity matrix (equal to 2 by default), and argument `f` designates the function used for the distance decay effect, the negative exponential (by default) or reciprocal function. Argument `ptype` determines whether Wong's indices are computed using only internal boundaries (`ptype = "int"`) or all the borders of the spatial units (`ptype = "all"`). For the absolute clustering index $ACL$, it is possible to define the spatial interactions matrix that will be used, based on the `spatmat` argument: `"c"` for contiguity matrix (by default) and `"d"` for the distance matrix.

With the help of function `ResampleTest` the user can conduct all the statistical tests based on sampling, as described in the previous section. The main inputs of the function are the population distribution table `x`, the name of the function to be tested `fun`, the simulation type `simtype` (`"Boot"` to generate bootstrap replications, `"Jack"` to generate jackknife replications and `"MonteCarlo"` for a randomization test using Monte Carlo simulations), the number of simulations `nsim` (equal to 99 by default), the sampling unit used: `sampleunit = "unit"` when the sampling is based on spatial/organizational units and `sampleunit = "ind"` for individual sampling). In the bootstrapping technique, the argument `perc` is a vector with the percentiles to be displayed in the output, and the argument `samplesize` gives the size of the sample used for bootstrapping. If null, the sample size equals the number of spatial units (in the case of unit sampling), or the total population (in the case of individual sampling). For jackknife simulations, there are two specific arguments. When the argument `outl` is `TRUE` the function provides the outliers obtained by jackknife iterations. Argument `outmeth` defines the outlier detection method: boxplot, standard deviation, normal scores, $t$ Student scores,

chi-squared scores and median absolute deviation. Estimations based on scoring methods are obtained from the **outliers** package. If outliers are detected, the argument `sdtimes` is used as a multiplication factor of the standard deviation used to detect outliers, and `QRrange` determines the boxplot thresholds as the multiplication of IQR (inter quartile range). The argument `proba` is used for random location processes that are not equiprobable (a vector of probabilities should be provided). If the jackknife technique is employed, `proba` indicates the probability (confidence interval) for scoring tests. If the argument `setseed` is set to `TRUE`, a zero seed is set for the random number generator, which is useful to have replicable simulations. In addition, specific arguments such as geographical data and other arguments presented above, should be introduced to allow the segregation function to be tested.

The `ResampleTest` output is a list of several objects: index name, simulation type, summary statistics of the simulations, simulated values of the index, simulated population distribution. If outliers detection is used, additional objects are included: outliers matrix and outliers values as list and plot. The `ResampleTest` output can be used by the `ResamplePlot` function to plot the main results. Certain additional graphic arguments can be used to customize the output: the colors and the legend (position, format and character size). Here we provide a simple script to test the spatial component of Morrill's index; for more examples, see Appendix B.

```
R> ISDuncan(A)

[1] 1 1

R> ISMorrill(A, spatobj = segdata)

[1] 0.9444 0.9444

R> set.seed(1234)
R> test <- ResampleTest(A, fun = "ISMorrill", spatobj = segdata,
+    simtype = "MonteCarlo", sampleunit = "unit", nsim = 999)
R> test$Summary

  Var ISMorrill   Mean Rank P.Value
1   1    0.9444 0.6206 1000   0.001
2   2    0.9444 0.6206 1000   0.001

R> ResamplePlot(test)
```

# 4. Conclusions

**OasisR** is a package implemented in the R software which allows computation of many segregation indices. It was designed to respond to a range of applications in the easiest way possible. This package has the advantage that it is implemented in R which allows total control of the input arguments, most of which have default values that correspond to the standard use of indices. This feature enables less experienced users to conduct segregation analysis with ease. Moreover, there is the possibility to develop further analysis within R, to automate the scripts

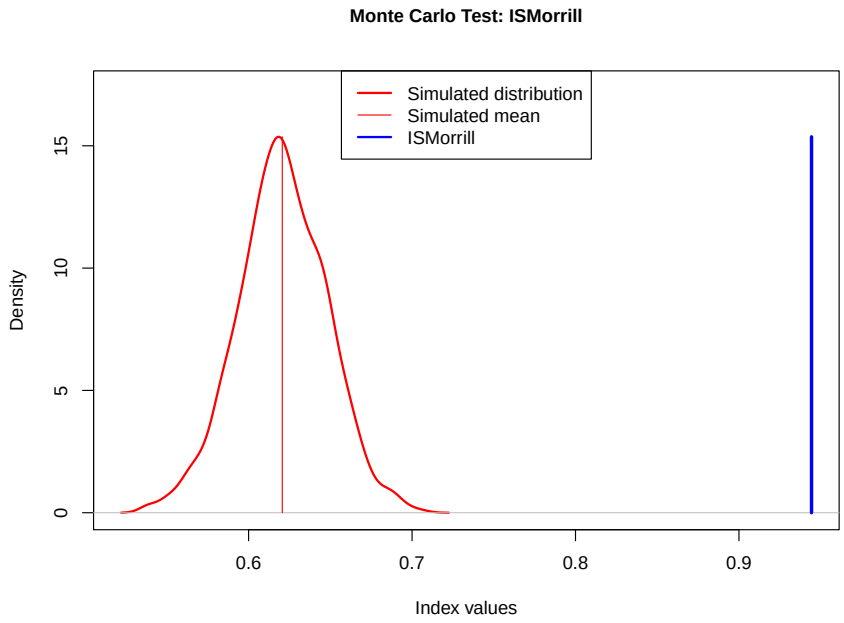


Figure 2: Permutation test.

and to integrate the analysis into other software packages. Another advantage compared to other segregation tools, is its noncommercial use which makes it available to a wide range of individuals who can download it for free directly from the Comprehensive R Archive Network (CRAN) where it is available at https://CRAN.R-project.org/package=OasisR. Since it is an open source package, it can be improved by the scientific community.

As we saw in Section 2, one of the most important benefits of **OasisR** is that it clarifies many ambiguities concerning definition of the segregation indices by providing proper computation and the possibility to choose among the different forms of the indices in the literature.

Another important contribution is the development of several resampling methods which allow testing of the statistical significance of indices. Three distinct types of simulations are provided: randomization tests, bootstrapping, and jackknife. Additionally, graphic functions are provided for a better visualization of results. It is clear that is still work to do in this field, and especially concerning the application of Bayesian inference.

The **OasisR** package is currently in its third version. Despite optimization efforts, more work is needed to improve this aspect regarding certain complex indices whose computation can take time, especially in the case of big study zones.

## References

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. doi:10.1111/ectj.12039.

Anselin L (1995). "Local Indicators of Spatial Association—LISA." *Geographical Analysis*, **27**(2), 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x.

Anselin L (2003). ***GeoDa*** *0.9 User's Guide*. University of Illinois, Urbana-Champaign.

Apparicio P (2000). "Les Indices de Ségrégation Résidentielle: Un Outil Intégré Dans un Système d'Information Géographique." *Cybergeo: European Journal of Geography*, **134**, 1–17. doi:10.4000/cybergeo.12063.

Apparicio P, Martori JC, Pearson AL, Fournier É, Apparicio D (2014). "An Open-Source Software for Calculating Indices of Urban Residential Segregation." *Social Science Computer Review*, **32**(1), 117–128. doi:10.1177/0894439313504539.

Apparicio P, Petkevitch V, Charron M (2008). "Segregation Analyzer: A C#.Net Application for Calculating Residential Segregation Indices." *Cybergeo: European Journal of Geography*, **414**, 1–27. doi:10.4000/cybergeo.16443.

Atkinson AB (1970). "On the Measurement of Inequality." *Journal of Economic Theory*, **2**(3), 244–263. doi:10.1016/0022-0531(70)90039-6.

Bell W (1954). "A Probability Model for the Measurement of Ecological Segregation." *Social Forces*, **32**(4), 357–364. doi:10.2307/2574118.

Birk MA (2016). ***birk***: *MA Birk's Functions*. R package version 2.1.2, URL https://CRAN.R-project.org/package=birk.

Bivand R (2019). ***spdep***: *Spatial Dependence: Weighting Schemes, Statistics and Models*. R package version 1.0-2, URL https://CRAN.R-project.org/package=spdep.

Bivand R, Keitt T, Rowlingson B (2019). ***rgdal***: *Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.4-3, URL https://CRAN.R-project.org/package=rgdal.

Bivand R, Rundel C (2019). ***rgeos***: *Interface to Geometry Engine – Open Source (**GEOS**)*. R package version 0.4-3, URL https://CRAN.R-project.org/package=rgeos.

Bivand RS, Pebesma E, Gómez-Rubio V (2013). *Applied Spatial Data Analysis with R*. 2nd edition. Springer-Verlag. doi:10.1007/978-1-4614-7618-4. URL http://www.asdar-book.org/.

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. doi:10.1016/0304-4076(94)90082-5.

Brown LA, Chung SY (2006). "Spatial Segregation, Segregation Indices and the Geographical Perspective." *Population, Space and Place*, **12**(2), 125–143. doi:10.1002/psp.403.

Carlson SM (1992). "Trends in Race/Sex Occupational Inequality: Conceptual and Measurement Issues." *Social Problems*, **39**(3), 268–290. doi:10.2307/3096962.

Carrington WJ, Troske KR (1997). "On Measuring Segregation in Samples with Small Units." *Journal of Business & Economic Statistics*, **15**(4), 402–409. doi:10.1080/07350015.1997.10524718.

Coleman JS (1966). "Equal Schools or Equal Students?" *Public Interest*, **4**, 70–75.

Cortese CF, Falk RF, Cohen JK (1976). "Further Considerations on the Methodological Analysis of Segregation Indices." *American Sociological Review*, **41**(4), 630–637. doi:10.2307/2094840.

Cowgill DO, Cowgill MS (1951). "An Index of Segregation Based on Block Statistics." *American Sociological Review*, **16**(6), 825–831. doi:10.2307/2087511.

Duncan OD, Cuzzort RP, Duncan B (1961). *Statistical Geography: Problems in Analyzing Area Data*. Free Press, Glencoe.

Duncan OD, Duncan B (1955a). "A Methodological Analysis of Segregation Indexes." *American Sociological Review*, **20**(2), 210–217. doi:10.2307/2088328.

Duncan OD, Duncan B (1955b). "Residential Distribution and Occupational Stratification." *American Journal of Sociology*, **60**(5), 493–503. doi:10.1086/221609.

Efron B (1979). "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, **7**(1), 1–26. doi:10.1214/aos/1176344552.

Egan KL, Anderton DL, Weber E (1998). "Relative Spatial Concentration among Minorities: Addressing Errors in Measurement." *Social Forces*, **76**(3), 1115–1121. doi:10.2307/3005705.

Falk RF, Cortese CF, Cohen J (1978). "Utilizing Standardized Indices of Residential Segregation: Comment on Winship." *Social Forces*, **57**(2), 713–716. doi:10.2307/2577693.

Farber S, O'Kelly M, Miller HJ, Neutens T (2015). "Measuring Segregation Using Patterns of Daily Travel Behavior: A Social Interaction Based Model of Exposure." *Journal of Transport Geography*, **49**, 26–38. doi:10.1016/j.jtrangeo.2015.10.009.

Farber S, Páez A, Morency C (2012). "Activity Spaces and the Measurement of Clustering and Exposure: A Case Study of Linguistic Groups in Montreal." *Environment and Planning A: Economy and Space*, **44**(2), 315–332. doi:10.1068/a44203.

Feitosa FF, Câmara G, Monteiro AMV, Koschitzki T, Silva MPS (2007). "Global and Local Spatial Indices of Urban Segregation." *International Journal of Geographical Information Science*, **21**(3), 299–323. doi:10.1080/13658810600911903.

Findlay A, Findlay A (1984). "A Monte Carlo Approach to Estimating the Significance of Segregation." *Environment and Planning A: Economy and Space*, **16**(2), 225–231. doi:10.1068/a160225.

Folch DC, Rey SJ (2016). "The Centralization Index: A Measure of Local Spatial Segregation." *Papers in Regional Science*, **95**(3), 555–576. doi:10.1111/pirs.12145.

Gini C (1921). "Measurement of Inequality of Incomes." *The Economic Journal*, **31**(121), 124–126. doi:10.2307/2223319.

Gorard S, Taylor C (2002). "What Is Segregation?: A Comparison of Measures in Terms of "Strong" and "Weak" Compositional Invariance." *Sociology*, **36**(4), 875–895. doi:10.1177/003803850203600405.

Hong SY (2014). *R Package* ***seg***. URL https://sites.google.com/site/hongseongyun/seg.

Hong SY, O'Sullivan D (2018). ***seg****: A Set of Tools for Measuring Spatial Segregation*. R package version 0.5-5, URL https://CRAN.R-project.org/package=seg.

Hong SY, O'Sullivan D, Sadahiro Y (2014). "Implementing Spatial Segregation Measures in R." *PLoS ONE*, **9**(11). doi:10.1371/journal.pone.0113767.

Hoover EM (1941). "Interstate Redistribution of Population, 1850–1940." *The Journal of Economic History*, **1**(2), 199–205. doi:10.1017/s0022050700052980.

Hornseth RA (1947). "A Note on "The Measurement of Ecological Segregation"." *American Sociological Review*, **12**(5), 603–604.

Iceland J, Weinberg DH, Steinmetz E (2002). *Racial and Ethnic Residential Segregation in the United States: 1980–2000*. U.S. Census Bureau, Series CENSR-3. U.S. Government Printing Office, Washington, DC.

Isard W (1960). *Methods of Regional Analysis: An Introduction to Regional Science*. The Technology Press of MIT, New York.

Jahn J, Schmid CF, Schrag C (1947). "The Measurement of Ecological Segregation." *American Sociological Review*, **12**(3), 293–303. doi:10.2307/2086519.

Jakubs JF (1981). "A Distance-Based Segregation Index." *Socio-Economic Planning Sciences*, **15**(3), 129–136. doi:10.1016/0038-0121(81)90028-8.

James DR, Taeuber KE (1985). "Measures of Segregation." *Sociological Methodology*, **15**, 1–32. doi:10.2307/270845.

Johnston R, Poulsen M, Forrest J (2007). "Ethnic and Racial Segregation in U.S. Metropolitan Areas, 1980–2000." *Urban Affairs Review*, **42**(4), 479–504. doi:10.1177/1078087406292701.

Komsta L (2011). ***outliers****: Tests for Outliers*. R package version 0.14, URL https://CRAN.R-project.org/package=outliers.

Konstantinidis S, Townshend I (1999). ***SEGCALC*** *– A Program for Simultaneous Computation of Multiple Indices of Segregation*. URL http://people.uleth.ca/~towni0/segcalcreadme.txt.

Lee D (2013). "**CARBayes**: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software*, **55**(13), 1–24. doi:10.18637/jss.v055.i13.

Lee D, Minton J, Pryce G (2015). "Bayesian Inference for the Dissimilarity Index in the Presence of Spatial Autocorrelation." *Spatial Statistics*, **11**, 81–95. doi:10.1016/j.spasta.2014.12.001.

Lieberson S (1981). "An Asymmetrical Approach to Segregation." In VR C Peach, S Smith (eds.), *Ethnic Segregation in Cities*, pp. 61–82. Croom Helm, London.

Massey DS (1978). "On the Measurement of Segregation as a Random Variable." *American Sociological Review*, **43**(4), 587–590. doi:10.2307/2094781.

Massey DS, Denton NA (1988). "The Dimensions of Residential Segregation." *Social Forces*, **67**(2), 281–315. doi:10.1093/sf/67.2.281.

Maurin L, Schneider V (2015). *Rapport sur les Inégalités en France*. Observatoire des Inégalités.

Morgan BS (1975). "The Segregation of Socioeconomic Groups in Urban Areas: A Comparative Analysis." *Urban Studies*, **12**, 47–60. doi:10.1080/00420987520080041.

Morgan BS (1983a). "An Alternate Approach to the Development of a Distance-Based Measure of Racial Segregation." *American Journal of Sociology*, **88**(6), 1237–1249. doi:10.1086/227802.

Morgan BS (1983b). "A Distance-Decay Based Interaction Index to Measure Residential Segregation." *Area*, **15**(3), 211–217.

Morrill R (1991). "On the Measure of Geographic Segregation." *Geography Research Forum*, **11**, 25–36.

Poulsen M, Johnston R, Forrest J (2010). "The Intensity of Ethnic Residential Clustering: Exploring Scale Effects Using Local Indicators of Spatial Association." *Environment and Planning A: Economy and Space*, **42**(4), 874–894. doi:10.1068/a42181.

Poulsen M, Johnston R, Forrest J (2011). "Using Local Statistics and Neighbourhood Classifications to Portray Ethnic Residential Segregation: A London Example." *Environment and Planning B: Urban Analytics and City Science*, **38**(4), 636–658. doi:10.1068/b36094.

Quenouille MH (1956). "Notes on Bias in Estimation." *Biometrika*, **43**(3–4), 353–360. doi:10.1093/biomet/43.3-4.353.

Ransom MR (2000). "Sampling Distributions of Segregation Indexes." *Sociological Methods & Research*, **28**(4), 454–475. doi:10.1177/0049124100028004003.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reardon SF (1998). "Measures of Racial Diversity and Segregation in Multigroup and Hierarchical Structured Population." In *Annual Meeting of the Eastern Sociological Society*. Philadelphia.

Reardon SF (2009). "Measures of Ordinal Segregation." In Y Flückiger, SF Reardon, J Silber (eds.), *Occupational and Residential Segregation*, volume 17 of *Research on Economic Inequality*, pp. 129–155. Emerald Group Publishing Limited, London.

Reardon SF (2011). "Measures of Income Segregation." The Stanford Center on Poverty and Inequality.

Reardon SF, Bischoff K (2011). "Income Inequality and Income Segregation." *American Journal of Sociology*, **116**(4), 1092–1153. doi:10.1086/657114.

Reardon SF, Firebaugh G (2002). "Measures of Multigroup Segregation." *Sociological Methodology*, **32**(1), 33–67. doi:10.1111/1467-9531.00110.

Reardon SF, Firebaugh G, O'Sullivan D, Matthews S (2006). "A New Approach to Measuring Socio-Spatial Economic Segregation." In *29th General Conference of The International Association for Research in Income and Wealth*. URL http://www.iariw.org/papers/2006/reardon.pdf.

Reardon SF, O'Sullivan D (2004). "Measures of Spatial Segregation." *Sociological Methodology*, **34**(1), 121–162. doi:10.1111/j.0081-1750.2004.00150.x.

Sakoda JM (1981). "A Generalized Index of Dissimilarity." *Demography*, **18**(2), 245–250. doi:10.2307/2061096.

Shannon CE (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Shevky E, Williams M (1949). *The Social Areas of Los Angeles. Analysis and Typology*. University of California Press.

Simpson EH (1949). "Measurement of Diversity." *Nature*, **163**, 688. doi:10.1038/163688a0.

StataCorp (2017). *Stata Statistical Software: Release 15*. StataCorp LLC, College Station. URL http://www.stata.com/.

Stearns LB, Logan JR (1986). "Measuring Trends in Segregation: Three Dimensions, Three Measures." *Urban Affairs Review*, **22**(1), 124–150. doi:10.1177/004208168602200107.

Theil H (1972). *Statistical Decomposition Analysis*. North-Holland, Amsterdam.

Theil H, Finizza AJ (1971). "A Note on the Measurement of Racial Integration of Schools by Means of Informational Concepts." *The Journal of Mathematical Sociology*, **1**(2), 187–193. doi:10.1080/0022250x.1971.9989795.

Tivadar M (2019). ***OasisR**: Outright Tool for the Analysis of Spatial Inequalities and Segregation*. R package version 3.0.1, URL https://CRAN.R-project.org/package=OasisR.

Tivadar M, Schaeffer Y, Torre A, Bray F (2014). "**OASIS** – Un Outil d'Analyse de la Ségrégation et des Inégalités Spatiales." *Cybergeo: European Journal of Geography*, **699**, 1–17. doi:10.4000/cybergeo.26579.

Tukey JW (1958). "Bias and Confidence in Not-Quite Large Samples." *The Annals of Mathematical Statistics*, **29**(2), 614. doi:10.1214/aoms/1177706647.

White MJ (1983). "The Measurement of Spatial Segregation." *American Journal of Sociology*, **88**(5), 1008–1018. doi:10.1086/227768.

White MJ (1986). "Segregation and Diversity Measures in Population Distribution." *Population Index*, **52**(2), 198–221. doi:10.2307/3644339.

Williams JJ (1948). "Another Commentary on So-Called Segregation Indices." *American Sociological Review*, **13**(3), 298–303. doi:10.2307/2086569.

Winship C (1977). "A Revaluation of Indexes of Residential Segregation." *Social Forces*, **55**(4), 1058–1066. doi:10.2307/2577572.

Wong D (1996). "Enhancing Segregation Studies Using GIS." *Computers, Environment and Urban Systems*, **20**(2), 99–109. doi:10.1016/s0198-9715(96)00003-8.

Wong D, Shaw SL (2011). "Measuring Segregation: An Activity Space Approach." *Journal of Geographical Systems*, **13**(2), 127–145. doi:10.1007/s10109-010-0112-x.

Wong DW (2016). "From Aspatial to Spatial, from Global to Local and Individual: Are We on the Right Track to Spatialize Segregation Measures?" In FM Howell, JR Porter, SA Matthews (eds.), *Recapturing Space: New Middle-Range Theory in Spatial Demography*, pp. 77–98. Springer International Publishing, Cham. doi:10.1007/978-3-319-22810-5_5.

Wong DWS (1993). "Spatial Indices of Segregation." *Urban Studies*, **30**(3), 559–572. doi:10.1080/00420989320080551.

Wong DWS (2003). "Implementing Spatial Segregation Measures in GIS." *Computers, Environment and Urban Systems*, **27**(1), 53–70. doi:10.1016/s0198-9715(01)00018-7.

Wong DWS, Chong WK (1998). "Using Spatial Segregation Measures in GIS and Statistical Modeling Packages." *Urban Geography*, **19**(5), 477–485. doi:10.2747/0272-3638.19.5.477.

Yao J, Wong DWS, Bailey N, Minton J (2019). "Spatial Segregation Measures: A Methodological Review." *Tijdschrift voor Economische en Sociale Geografie*. doi:10.1111/tesg.12305.

Zoloth BS (1976). "Alternative Measures of School Segregation." *Land Economics*, **52**(3), 278–298. doi:10.2307/3145527.

# A. Indices definition and use in OasisR

The table in this appendix gives an overview on indices definition and use. The following notation is used in the table:

$n$ – number of spatial units;

$x_i^k$ – population of group $k$ in spatial unit $i$;

$X^k$ – population of group $k$ in the area;

$t_i$ – total population in spatial unit $i$;

$T$ – total population in the area;

$t_i^{k_1,k_2} = x_i^{k_1} + x_i^{k_2}$ – population of groups $k_1$ and $k_2$ in spatial unit $i$;

$T^{k_1,k_2} = X^{k_1} + X^{k_2}$ – population of groups $k_1$ and $k_2$ in the area;

$p_i^k$ – proportion of population $k$ in spatial unit $i$;

$P^k$ – proportion of population $k$ in the area;

$p_i^{k_1,k_2} = \frac{x_i^{k_1}}{x_i^{k_1}+x_i^{k_2}}$ – proportion of group $k_1$ in the population $k_1$ and $k_2$ in spatial unit $i$;

$P^{k_1,k_2} = \frac{X^{k_1}}{X^{k_1}+X^{k_2}}$ – proportion of group $k_1$ in the population $k_1$ and $k_2$ in the area;

$\delta$ – inequality aversion parameter for Atkinson index;

$Per_i$ – perimeter of spatial unit $i$;

$A_i$ – area of spatial unit $i$;

$A$ – total area of the zone;

$c_{ij}$ – elements of contiguity matrix;

$\bar{c}_{ij}$ – elements of contiguity matrix, where $\bar{c}_{ii} = 1$;

$c_{ij}^{\lambda}$ – elements of $\lambda$ order contiguity matrix;

$b_{ij}$ – elements of shared boundaries matrix;

$d_{ij}$ – elements of distance matrix;

$f(d_{ij})$ – function of spatial interaction, based on the distance between centroids of spatial units $i$ and $j$. Usually two forms are used: linear $f(d_{ij}) = d_{ij}$ and exponential $f(d_{ij}) = \exp(-\beta d_{ij})$, where $\beta$ is a distance decay parameter;

$f(\lambda)$ – function of contiguity interaction, similar to distance interaction. We propose two forms: reciprocal $f(\lambda) = 1/\lambda$ and exponential $f(\lambda) = \exp(-\beta\lambda)$, where $\beta$ is a distance decay parameter;

$R$ – a spatial region;

$q$ – points within the region;

$\tau_q$ – population density at point $q$;

$\tau_q^k$ – population density of group $k$ at point $q$;

$\tilde{\tau}_q^k$ – population density of group $k$ in the local environment of point $q$;

$\pi_q^k$ – proportion in group $k$ at point $q$;

$\tilde{\pi}_q^k$ – proportion in group $k$ in the local environment of point $q$;

$\hat{p}, \hat{q}$ – percentile ranks in the population of interest corresponding to a continuous variable (such as income);

$n_1^k$ – rank of spatial unit where the sum of all $t_i$ equals or exceeds $X^k$ (from 1 to $n_1^k$), spatial units being ordered by geographic size;

$n_2^k$ – rank of spatial unit where the sum of all $t_i$ equals or exceeds $X^k$ (from $n$ to $n_2^k$), spatial units being ordered by geographic size;

$T_1^k$ – sum of all $t_i$ from spatial unit 1 to spatial unit $n_1^k$;

$T_2^k$ – sum of all $t_i$ from spatial unit $n_2^k$ to spatial unit $n$;

$\tilde{X}_i^k$ – the cumulative percentage of the $k$ group population through the $i$th spatial unit, spatial units being ordered by distance to the center (standard version of centralization) or to the closest center (polycentric version);

$\tilde{t}_i^k$ – the cumulative percentage of total population through the $i$th spatial unit, spatial units being ordered by distance to the center (standard version of centralization) or to the closest center (polycentric version);

$\tilde{n}$ – the magnitude of the centrality effect: $\tilde{n} = n$ for unconstrained centralization, and $\tilde{n} < n$ for local/constrained centralization indices.

| Index | Definition | **OasisR** function |
|---|---|---|
| *Evenness* | | |
| Segregation | $IS^k = \sum_{i=1}^{n} \left[ \frac{t_i \left\lvert p_i^k - P^k \right\rvert}{2TP^k(1-P^k)} \right]$ | `ISDuncan(x)` |
| Dissimilarity | $D^{k_1,k_2} = \frac{1}{2} \sum_{i=1}^{n} \left\lvert \frac{x_i^{k_1}}{X^{k_1}} - \frac{x_i^{k_2}}{X^{k_2}} \right\rvert$ | `DIDuncan(x)` |
| Gini | $G^k = \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \frac{t_i t_j \left\lvert p_i^k - p_j^k \right\rvert}{2T^2 P^k (1-P^k)} \right]$ | `Gini(x)` |
| Gini2 | $G_2^{k_1,k_2} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left[ t_i^{k_1,k_2} t_j^{k_1,k_2} \left\lvert p_i^{k_1,k_2} - p_j^{k_1,k_2} \right\rvert \right]}{2\left(T^{k_1,k_2}\right)^2 P^{k_1,k_2}(1-P^{k_1,k_2})}$ | `Gini2(x)` |
| Atkinson | $A^k = 1 - \left( \frac{P^k}{1-P^k} \right) \left[ \sum_{i=1}^{n} \left( \frac{\left(1-p_i^k\right)^{1-\delta} \left(p_i^k\right)^{\delta} t_i}{P^k T} \right) \right]^{\frac{1}{1-\delta}}$ | `Atkinson(x, delta)` |
| Gorard | $GS^k = \frac{1}{2} \sum_{i=1}^{n} \left\lvert \frac{x_i^k}{X^k} - \frac{t_i}{T} \right\rvert$ | `Gorard(x)` |
| Entropy | $H^k = \sum_{i=1}^{n} \left[ \frac{t_i \left(E^k - E_i^k\right)}{E^k T} \right]$ with<br>$E^k = P^k \ln\left(\frac{1}{P^k}\right) + \left(1-P^k\right) \ln\left(\frac{1}{1-P^k}\right)$<br>$E_i^k = p_i^k \ln\left(\frac{1}{p_i^k}\right) + \left(1-p_i^k\right) \ln\left(\frac{1}{1-p_i^k}\right)$<br>$p_i^k \ln\left(\frac{1}{p_i^k}\right) = 0$ if $p_i^k = 0$<br>$\left(1-p_i^k\right) \ln\left(\frac{1}{1-p_i^k}\right) = 0$ if $p_i^k = 1$ | `HTheil(x)` |
| Adjusted contiguity segregation index | $IS^k(adj) = IS^k - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \left\lvert p_i^k - p_j^k \right\rvert}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}}$ | `ISMorrill(x, c, queen, spatobj, folder, shape)` |
| $K$th order contiguity segregation index | $IS^k(Kadj) = IS^k - \sum_{\lambda=1}^{K} f(\lambda) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^{\lambda} \left\lvert p_i^k - p_j^k \right\rvert}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^{\lambda}}$ | `ISMorrillK(x, ck, queen, K, f, spatobj, folder, shape)` |
| Adjusted contiguity dissimilarity index | $D^{k_1,k_2}(adj) = D^{k_1,k_2} - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \left\lvert p_i^{k_1,k_2} - p_j^{k_1,k_2} \right\rvert}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}}$ | `DIMorrill(x, c, queen, spatobj, folder, shape)` |

| | | |
|---|---|---|
| $K$th order contiguity dissimilarity index | $D^{k_1,k_2}(Kadj) = D^{k_1,k_2} - \sum_{\lambda=1}^{K} f(\lambda) \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^{\lambda} \left\lvert p_i^{k_1,k_2} - p_j^{k_1,k_2} \right\rvert}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^{\lambda}}$ | `DIMorrillK(x, ck, queen, K, f, spatobj, folder, shape)` |
| Adjusted boundary segregation index | $IS^k(w) = IS^k - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} \lvert p_i^k - p_j^k \rvert}{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}}$ | `ISWong(x, b, a, p, ptype, variant = "w", spatobj, folder, shape)` |
| Adjusted boundary dissimilarity index | $D^{k_1,k_2}(w) = D^{k_1,k_2} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} \left\lvert p_i^{k_1,k_2} - p_j^{k_1,k_2} \right\rvert}{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}}$ | `DIWong(x, b, a, p, ptype, variant = "w", spatobj, folder, shape)` |
| Adj. bound. & perimeter area ratio segregation index | $IS^k(s) = IS^k - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} \lvert p_i^k - p_j^k \rvert \left( \frac{\frac{Per_i}{A_i} + \frac{Per_j}{A_j}}{2 \max_{l=[1:n]} \left( \frac{Per_l}{A_l} \right)} \right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}}$ | `ISWong(x, b, a, p, ptype, variant = "s", spatobj, folder, shape)` |
| Adj. bound. & perimeter area ratio dissimilarity index | $D^{k_1,k_2}(s) = D^{k_1,k_2} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} \left\lvert p_i^{k_1,k_2} - p_j^{k_1,k_2} \right\rvert \left( \frac{\frac{Per_i}{A_i} + \frac{Per_j}{A_j}}{2 \max_{l=[1:n]} \left( \frac{Per_l}{A_l} \right)} \right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}}$ | `DIWong(x, b, a, p, ptype, variant = "s", spatobj, folder, shape)` |
| *Exposure* | | |
| Isolation index (exact version) | $xPx^{*k} = \sum_{i=1}^{n} \left( \frac{x_i^k}{X^k} \frac{x_i^k - 1}{t_i - 1} \right)$ | `xPx(x, exact = TRUE)` |
| Isolation index (approximate version) | $xPx^k = \sum_{i=1}^{n} \left( \frac{x_i^k}{X^k} \frac{x_i^k}{t_i} \right)$ | `xPx(x, exact = FALSE)` |
| Interaction index (exact version) | $xPy^{*k_1,k_2} = \sum_{i=1}^{n} \left( \frac{x_i^{k_1}}{X^{k_1}} \frac{x_i^{k_2}}{t_i - 1} \right)$ | `xPy(x, exact = TRUE)` |
| Interaction index (approximate version) | $xPy^{k_1,k_2} = \sum_{i=1}^{n} \left( \frac{x_i^{k_1}}{X^{k_1}} \frac{x_i^{k_2}}{t_i} \right)$ | `xPy(x, exact = FALSE)` |
| Eta2 | $Eta2^k = \frac{xPx^k - P^k}{1 - P^k}$ | `Eta2(x)` |
| Spatial isolation/exposure index. See Reardon and O'Sullivan (2004) for details. | $x\tilde{P}^*x^k = \int_{q \in R} \frac{\tau_q^k}{X^k} \tilde{\pi}_q^k \mathrm{d}q$<br>$x\tilde{P}^*y^{k_1,k_2} = \int_{q \in R} \frac{\tau_q^{k_1}}{X^{k_1}} \tilde{\pi}_q^{k_2} \mathrm{d}q$ | `spatinteract(x, spatobj, folder, shape, ...)` |

| | | |
|---|---|---|
| Distance-decay isolation index | $DPxx^k = \sum_{i=1}^{n} \frac{x_i^k}{X^k} \sum_{j=1}^{n} \frac{K_{ij} x_j^k}{t_j}$ | `DPxx(x, d, distin, distout, diagval, spatobj, folder, shape)` |
| Distance-decay interaction index | $DPxy^{k_1,k_2} = \sum_{i=1}^{n} \frac{x_i^k}{X^k} \sum_{j=1}^{n} \frac{K_{ij} y_j^k}{t_j}$ with $K_{ij} = \exp(-\beta d_{ij}) t_i \sum_{j=1}^{n} \exp(-\beta d_{ij}) t_j$ | `DPxy(x, d, distin, distout, diagval, spatobj, folder, shape)` |
| *Clustering* | | |
| One-group mean proximity | $Pxx^k = \frac{1}{(X^k)^2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i^k x_j^k f(d_{ij})$ | `Pxx(x, d, fdist, distin, distout, diagval, spatobj, folder, shape)` |
| Between group mean proximity | $Pxy^{k_1 k_2} = \frac{1}{X^{k_1} X^{k_2}} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i^{k_1} x_j^{k_2} f(d_{ij})$ | `Pxy(x, d, fdist, distin, distout, diagval, spatobj, folder, shape)` |
| Multi-group mean proximity | $Poo = \frac{1}{T^2} \sum_{i=1}^{n} \sum_{j=1}^{n} t_i t_j f(d_{ij})$ | `Poo(x, d, fdist, distin, distout, diagval, itype = "multi", spatobj, folder, shape)` |
| Multi-group mean proximity (between group version) | $Poo^{k_1 k_2} = \frac{1}{X^{k_1} + X^{k_2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(x_i^{k_1} + x_i^{k_2}\right) \left(x_j^{k_1} + x_j^{k_2}\right) f(d_{ij})$ | `Poo(x, d, fdist, distin, distout, diagval, itype = "between", spatobj, folder, shape)` |
| Spatial proximity index (multi-group version) | $SP = \frac{\sum_{k=1}^{n} X_k Pxx^k}{TPoo}$ | `SP(x, d, fdist, distin, distout, diagval, itype = "multi", spatobj, folder, shape)` |
| Spatial proximity index (one-group version version) | $SP^k = \frac{Pxx^k}{Poo}$ | `SP(x, d, fdist, distin, distout, diagval, itype = "one", spatobj, folder, shape)` |
| Spatial proximity index (between group version) | $SP^{k_1,k_2} = \frac{X^{k_1} Pxx^{k_1} + X^{k_2} Pxx^{k_2}}{\left(X^{k_1} + X^{k_2}\right) Poo^{k_1,k_2}}$ | `SP(x, d, fdist, distin, distout, diagval, itype = "between", spatobj, folder, shape)` |
| Absolute clustering index | $ACL^k = \frac{\left[\sum_{i=1}^{n} \frac{x_i^k}{X^k} \sum_{j=1}^{n} (w_{ij} x_j^k)\right] - \left[\frac{X^k}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right]}{\left[\sum_{i=1}^{n} \frac{x_i^k}{X^k} \sum_{j=1}^{n} (w_{ij} t_j)\right] - \left[\frac{X^k}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right]}$ with $w_{ij} = \bar{c}_{ij}$, or $w_{ij} = f(d_{ij}) = \exp(-\beta d_{ij})$ | `ACL(x, spatmat, c, queen, distin, distout, diagval, spatobj, folder, shape)` |
| Relative clustering index | $RCL^{k_1,k_2} = \frac{Pxx^{k_1}}{Pxx^{k_2}} - 1$ with $f(d_{ij}) = \exp(-\beta d_{ij})$ | `RCL(x, d, fdist = "e", distin, distout, diagval, spatobj, folder, shape)` |

| | | |
|---|---|---|
| Relative clustering index (linear version) | $RCL^{k_1,k_2} = 1 - \frac{Pxx^{k_1}}{Pxx^{k_2}}$ with $f(d_{ij}) = d_{ij}$ | `RCL(x, d, fdist = "l", distin, distout, diagval, spatobj, folder, shape)` |
| *Concentration* | | |
| Delta | $\Delta^k = \frac{1}{2}\sum_{i=1}^{n}\left\lvert\frac{x_i^k}{X^k} - \frac{A_i}{A}\right\rvert$ | `Delta(x, a, spatobj, folder, shape)` |
| Absolute concentration index | $ACO^k = 1 - \frac{\sum_{i=1}^{n}(x_i^k A_i/X^k) - \sum_{i=1}^{n_1^k}(t_i A_i/T_1^k)}{\sum_{i=n_2^k}^{n}(t_i A_i/T_2^k) - \sum_{i=1}^{n_1^k}(t_i A_i/T_1^k)}$ | `ACO(x, a, spatobj, folder, shape)` |
| Relative concentration index | $RCO^{k_1,k_2} = \frac{\left[\sum_{i=1}^{n}(x_i^{k_1}A_i/X^{k_1})/\sum_{i=1}^{n}(x_i^{k_2}A_i/X^{k_2})\right]-1}{\left[\sum_{i=1}^{n_1^{k_1}}(t_i A_i/T_1^{k_1})/\sum_{i=n_2^{k_2}}^{n}(t_i A_i/T_2^{k_2})\right]-1}$ | `RCO(x, a, spatobj, folder, shape)` |
| *Centralization* | | |
| Relative centralization | $RCE^{k_1,k_2} = \left(\sum_{i=2}^{\tilde{n}}\tilde{X}_{i-1}^{k_1}\tilde{X}_i^{k_2}\right) - \left(\sum_{i=2}^{\tilde{n}}\tilde{X}_i^{k_1}\tilde{X}_{i-1}^{k_2}\right)$ | `RCE(x, dc, center, spatobj, folder, shape)` `RCEPoly(x,...)` `RCEPolyK(x, K, kdist, ...)` |
| Duncan's absolute centralization | $DCE^k = RCE^{k,-k} = RCE^{k,\text{Total}}/(1-P^k) = = \frac{\left(\sum_{i=2}^{\tilde{n}}\tilde{X}_{i-1}^k\tilde{t}_i\right) - \left(\sum_{i=2}^{\tilde{n}}\tilde{X}_i^k\tilde{t}_{i-1}\right)}{(1-P^k)}$ | `ACEDuncan(x, dc, center, spatobj, folder, shape)` `ACEDuncanPoly(x,... )` `ACEDuncanPolyK(x, K, kdist, ...)` |
| Massey's absolute centralization | $ACE^k = \left(\sum_{i=2}^{n}X_{i-1}^k A_i\right) - \left(\sum_{i=2}^{n}X_i^k A_{i-1}\right)$ | `ACE(x, a, dc, center, spatobj, folder, shape)` `ACEPoly(x, a,...)` |
| *Multi-group indices* | | |
| Multi-group dissimilarity | $D^* = \frac{1}{2TI_S}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(t_i\left\lvert p_i^k - P^k\right\rvert\right)$ | `DMulti(x)` |
| Multi-group Gini | $G^* = \frac{1}{2T^2 I_S}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(t_i t_j\left\lvert p_i^k - p_j^k\right\rvert\right)$ | `GiniMulti(x)` |
| Multi-group normalized exposure | $P^* = \frac{1}{T}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{t_i\left(p_i^k - P^k\right)^2}{1-P^k}$ | `PMulti(x)` |
| Multi-group information theory | $H^* = \frac{1}{TE}\sum_{k=1}^{K}\sum_{i=1}^{n}t_i p_i^k \ln\left(\frac{p_i^k}{P^k}\right)$ with $E = \sum_{k=1}^{K}P^k\ln\left(\frac{1}{P^k}\right)$ | `HMulti(x)` |

| | | |
|---|---|---|
| Multi-group relative diversity | $RD^* = \frac{1}{TI_S}\sum_{k=1}^{K}\sum_{i=1}^{n} t_i\left(p_i^k - P^k\right)^2$ | `RelDivers(x)` |
| Squared coefficient of variation | $C^* = \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{t_i}{T}\frac{\left(p_i^k - P^k\right)^2}{(K-1)P^k}$ | `CMulti(x)` |
| Spatial multi-group dissimilarity<br>See Reardon and O'Sullivan (2004) for details | $\tilde{D} = \frac{1}{2TI_S}\sum_{k=1}^{K}\int_{q\in R}\tau_q\lvert\tilde{\pi}_q^k - P^k\rvert\mathrm{d}q$ | `spatmultiseg(x, spatobj, folder, shape)[1]` |
| Spatial multi-group relative diversity<br>See Reardon and O'Sullivan (2004) for details | $\tilde{R} = 1 - \int_{q\in R}\frac{\tau_q\tilde{I}_q}{TI}\mathrm{d}q$<br>where $\tilde{I}_q = \sum_{k=1}^{K}\tilde{\pi}_q^k(1-\tilde{\pi}_q^k)$ | `spatmultiseg(x, spatobj, folder, shape)[2]` |
| Spatial multi-group information theory<br>See Reardon and O'Sullivan (2004) for details | $\tilde{H} = 1 - \frac{1}{TE}\int_{q\in R}\tau_q\tilde{E}_q\mathrm{d}q$<br>with $\tilde{E}_q = \sum_{k=1}^{K}\tilde{\pi}_q^k\ln\tilde{\pi}_q^k$ | `spatmultiseg(x, spatobj, folder, shape)[3]` |
| Ordinal segregation indices<br>$\Lambda_1$ - ordinal information theory index<br>$\Lambda_2$ - ordinal variation ratio index<br>$\Lambda_3$ - ordinal square root index<br>$\Lambda_4$ - ordinal absolute difference index<br>See Reardon (2009) for details | $\Lambda_\lambda = \sum_{i=1}^{n}\frac{t_i}{T\nu_\lambda}(\nu_\lambda - \nu_{\lambda i})$<br>with $\nu_\lambda = \frac{1}{K-1}\sum_{k=1}^{K-1} f_\lambda(\mu^k)$<br>$\mu_i^k = \sum_{j=1}^{k}\frac{x_i^j}{t_i}$<br>$f_1(\mu) = -[\mu\ln\mu + (1-\mu)\ln(1-\mu)]$<br>$f_2(\mu) = 4\mu(1-\mu)$<br>$f_3(\mu) = 2\sqrt{\mu(1-\mu)}$<br>$f_4(\mu) = 1 - \lvert 2\mu - 1\rvert$ | `ordinalseg(x)` |
| Rank-order segregation indices<br>$\Lambda_1^R = H^R$ – rank-order information theory index<br>$\Lambda_2^R = R^R$ – rank-order variation ratio index<br>$\Lambda_3^R = S^R$ – rank-order square root index<br>See Reardon (2011) for details | $\Lambda_\lambda^R = \int_0^1 \frac{f(\hat{p})}{\int_0^1 f(\hat{q})\mathrm{d}\hat{q}}\Lambda_\lambda(\hat{p})\mathrm{d}\hat{p}$ | `rankorderseg(x, polorder, pred)` |
| *Social diversity indices* | | |
| Diversity index | $H_{SW} = -\sum_{k=1}^{K} P^k\ln P^k$ | `HShannon(x)` |
| Normalized diversity | $\bar{H}_{SW} = -\sum_{k=1}^{K}\frac{P^k\ln P^k}{\ln K}$ | `NShannon(x)` |
| Simpson's | $I_S = \sum_{k=1}^{K} P^k\left(1 - P^k\right)$ | `ISimpson(x)` |

| *Local indices* | | |
|---|---|---|
| Location quotient | $LQ_i^k = \left(\frac{x_i^k}{t_i}\right) \Big/ \left(\frac{X^k}{T}\right)$ | `LQ(x)` |
| Local diversity | $L_{SW}^i = -\sum_{k=1}^K p_i^k \ln p_i^k$ | `LShannon(x)` |
| Local entropy | $H2^i = -\sum_{k=1}^K \frac{p_i^k \ln p_i^k}{\ln K}$ | `HLoc(x)` |
| Local Simpson | $I_S^i = \sum_{k=1}^K p_i^k \left(1 - p_i^k\right)$ | `LSimpson(x)` |

# B. Examples of OasisR functions

To allow a better understanding of the **OasisR** functions, we provide several examples which demonstrate certain subtleties of their use. Similarly to the main article, we use the same theoretical examples based on a 10×10 checkboard, with two possible population distributions. We start with a short script which analyzes the Atkinson index sensitivity to inequality aversion parameter changes (distribution B):

```
R> delta.list <- seq(0.1, 0.9, by = 0.1)
R> result <- rep(0, 9)
R> for (i in 1:length(delta.list)) {
+    result[i] <- Atkinson(B, delta = delta.list[i])[1]
+  }
R> result

[1] 0.8538 0.8672 0.8827 0.9005 0.9211 0.9442 0.9687 0.9901 0.9997
```

The next example illustrates the variation of the generalized adjusted dissimilarity index as a function of the contiguity order $k$ (distribution A).

```
R> result <- rep(0, 11)
R> for (k in 1:11) {
+    result[k] <- DIMorrillK(A, spatobj = segdata, K = k)[1, 2]
+  }
R> result
R> plot(result, type = "l")

 [1] 0.9444 0.9022 0.8775 0.8646 0.8581 0.8552 0.8541 0.8536 0.8535 0.8534
[11] 0.8534
```

The perimeter definition (`ptype = "int"` for internal boundaries – by default, vs. `ptype = "all"` for entire borders of spatial units) matters for the adjusted boundary and perimeter area ratio dissimilarity index $D^{k_1 k_2}(s)$.

```
R> DIWong(A, spatobj = segdata, variant = "s", ptype = "int")

       [,1]   [,2]
[1,] 0.0000 0.9472
[2,] 0.9472 0.0000

R> DIWong(A, spatobj = segdata, variant = "s", ptype = "all")

       [,1]   [,2]
[1,] 0.0000 0.9444
[2,] 0.9444 0.0000
```

Similarly, the diagonal definition of the distance matrix has an impact on the value of distance-based indices:

```
R> DPxy(A, spatobj = segdata, diagval = "0")

       [,1]   [,2]
[1,] 0.7882 0.2118
[2,] 0.0706 0.9294

R> DPxy(A, spatobj = segdata, diagval = "a")

       [,1]   [,2]
[1,] 0.7688 0.2312
[2,] 0.0771 0.9229
```

As we reduce the effect of the distance by increasing the distance decay parameter, the minority spatial isolation reaches its maximum:

```
R> result <- rep(0, 11)
R> for (k in 1:11) {
+    result[k] <- DPxy(A, spatobj = segdata, diagval = "0", beta = k)[1, 1]
+  }
R> plot(result, type = "l")
R> result

 [1] 0.7882 0.9234 0.9725 0.9905 0.9968 0.9989 0.9996 0.9999 0.9999 1.0000
[11] 1.0000
```

Certain clustering indices, such as the spatial proximity index, can be defined as multi-group, one-group and between group measures:

```
R> SP(A, spatobj = segdata, diagval = "a", itype = "multi")

[1] 1.6636

R> SP(A, spatobj = segdata, diagval = "a", itype = "one")

[1] 2.9909 1.2212

R> SP(A, spatobj = segdata, diagval = "a", itype = "between")

       [,1]   [,2]
[1,] 1.0000 1.6636
[2,] 1.6636 1.0000
```

To compare gravitational distance-based measures, we need to pay attention to distance units which strongly influence results:

```
R> ACL(A, spatobj = segdata, spatmat = "d", diagval = "a")
```

```
[1] 0.6636 0.6636
```

```
R> ACL(A, spatobj = segdata, spatmat = "d", diagval = "a", distin = "m",
+    distout = "km")
```

```
[1] 9e-04 9e-04
```

Suppose that the area has two centers, located in the 3rd row and 8th column (28th polygon in the data base), and symmetrically in the 8th row and 3rd column (73rd polygon). The local effect of the constrained polycentric relative centralization index will depend on the area included by the number of closest neighbors to the center:

```
R> result <- rep(0, 25)
R> for (k in 1:25) {
+    result[k] <- RCEPolyK(A, spatobj = segdata, center = c(28, 83),
+      K = k)[1, 2]
+  }
R> plot(result, type = "l")
```

After seeing some interesting examples of segregation functions, we show some statistical tests enabled by the **OasisR** package, using resampling methods. We provide an example of the permutation test in the main paper; here we develop examples of bootstrapping and jackknife techniques. If the `outl` parameter is set to `TRUE`, the `ResampleTest` function automatically produces a boxplot for outlier detection.

```
R> xtest <- ResampleTest(B, fun = "ISDuncan", simtype = "Boot",
+    sampleunit = "unit", spatobj = segdata)
R> xtest$Summary
```

```
  Var ISDuncan 5th_percentile Median 95th_percentile      BootSE
1   1   0.8421         0.7847 0.8414          0.8909 0.003246122
2   2   0.8421         0.7847 0.8414          0.8909 0.003246122
```

```
R> xtest <- ResampleTest(B, fun = "xPx", simtype = "Jack",
+    sampleunit = "unit", spatobj = segdata, outl = TRUE)
R> xtest$Summary
```

```
  Var    xPx 5th_percentile Median 95th_percentile  JackBias     JackSE
1   1 0.7500         0.7391   0.75          0.7553 -0.003564 0.04559334
2   2 0.9211         0.9200   0.92          0.9238 -0.005544 0.01579764
```

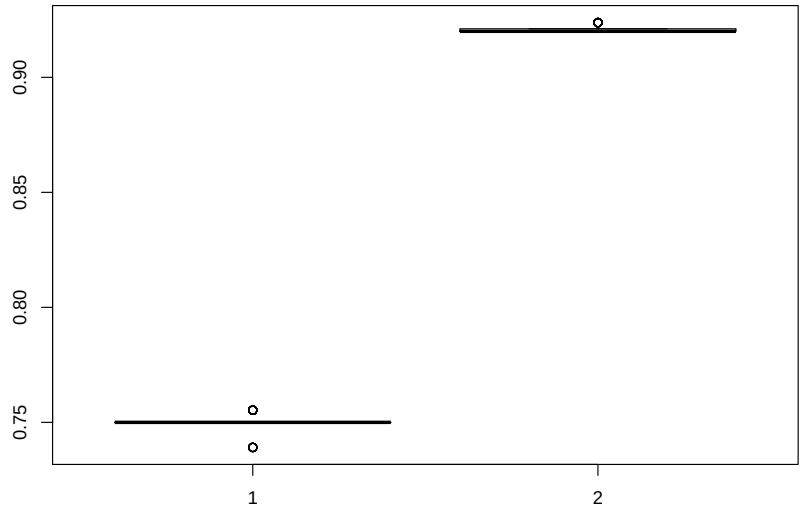


Figure 3: Outliers detection.

**Affiliation:**

Mihai Tivadar
UR DTGR
Université Grenoble Alpes, Irstea, Centre de Grenoble
2 rue de la Papeterie-BP 76
38402 , Saint-Martin-d'Hères, France
Telephone: +33 4 76 76 28 43
E-mail: mihai.tivadar@irstea.fr
URL: http://www.irstea.fr/

*Journal of Statistical Software* http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
May 2019, Volume 89, Issue 7 *Submitted:* 2016-07-29
doi:10.18637/jss.v089.i07 *Accepted:* 2018-02-13