A pattern-based approach to analysis and visualization of spatio-racial distribution

Anna Dmowska[1]
Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznan, Poland
http://socscape.edu.pl
dmowska@amu.edu.pl

Tomasz F. Stepinski
Space Informatics Lab, Department of Geography and GIS, University of Cincinnati, USA
stepintz@uc.edu

Jakub Nowosad
Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznan, Poland
nowosad.jakub@gmail.com

Abstract

Racial geography in US urban areas is extensively studied with the emphasis on assessing the extent of racial segregation. However, the used methodology has not changed for at least two decades; it relies on calculating ratios of population counts in the entire city and its subdivisions – census aggregation areas. This has a number of limitations; the two most important are: assessment of segregation depends on the subdivisions used, segregation can only be calculated for regions with census subdivisions. Here we present a different conceptualization of racial geography, which leads to a new method called racial landscape (RL). We use block-level census data to construct a high-resolution grid where each cell represents single race inhabitants. The result is a spatial, racial pattern; a degree of spatial autocorrelation of this pattern is a measure of segregation that does not require using subdivisions. We shortly describe the RL method and its application to Cook County, IL. We also describe here its implementation in the R computational environment.

1 Introduction

Populations in North America's cities and increasingly in Europe are mixtures of several distinct groups (frequently racial or ethnic groups). Spatial distributions of various groups differ, leading to a spatially variable mix of the population. There are locations where a single group dominates, and there are locations where they mix to a smaller or larger extent. This results in a complex spatio-racial distribution that needs to be visualized and quantified in order to compare different cities or perform a longitudinal study of a single city.

Currently, multi-racial distribution is quantified in terms of racial diversity and/or segregation [2, 5]. Both indices are calculated from ratios of population counts in the entire city and its subdivisions (for example, census tracts in the US). This follows the format of the census data that is reported at the subdivision level. Diversity is conceptualized as the evenness of the city-wide histogram of different races counts and is calculated using the entropy

[1] Corresponding author

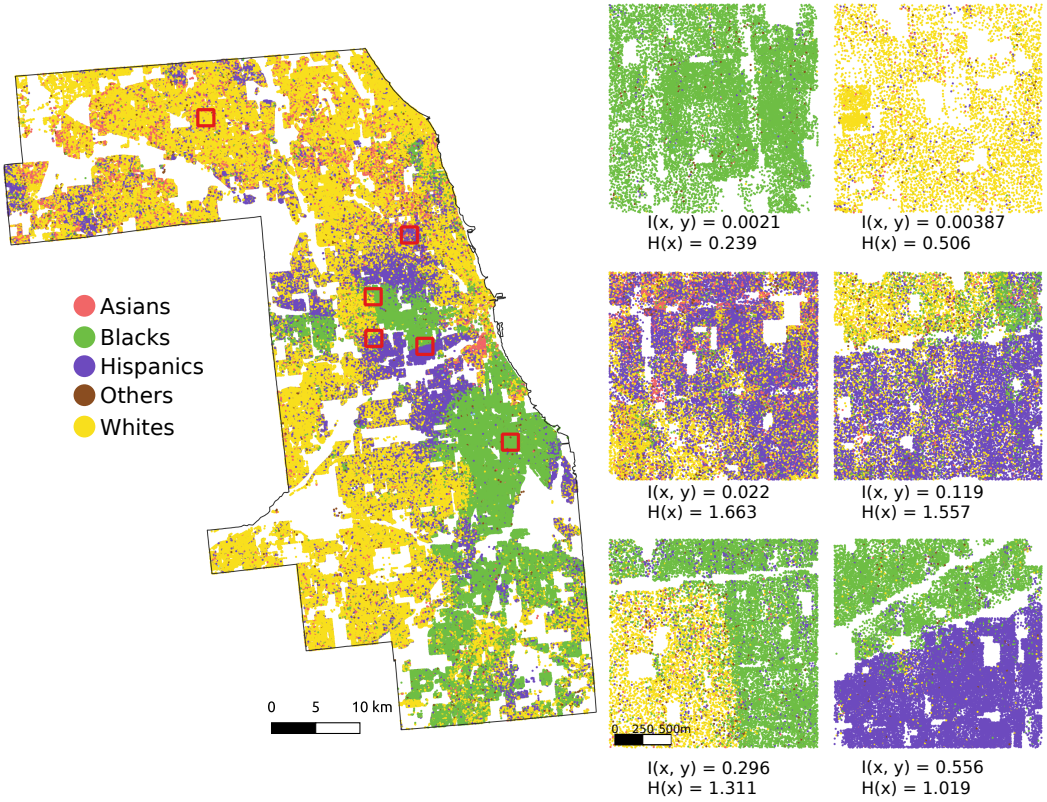

Figure 1 (Left) A spatio-racial pattern in Cook County, IL, in 2010 under the assumption of constant population density. (Right) Pattern magnified to the extents of six small regions selected from different locations in Cook County. Local values of $H(x)$ (diversity) and $I(x;y)$ (segregation) are given.

of subpopulation shares $H = -\sum p_i \log_2 p_i$, where p_i is the share of ith of n subpopulations and $i = 1, \ldots, n$. Multi-racial segregation is commonly described as an average relative difference between the entropy of the entire city (H) and entropies of individual subdivisions (H_i). The problem with a so-defined index of segregation is it dependence on the sizes and particulars of subdivisions, so no unequivocal assessment of segregation can be calculated.

This situation can be alleviated by changing the conceptualization of the problem. Instead of thinking about the spatio-racial distribution of the data (subdivisions), we can think about individual inhabitants represented by points at the locations of their residences having colors corresponding to their races. This can be achieved by using the smallest subdivisions (census blocks) and distributing the inhabitants of each block randomly throughout its spatial extent. Figure 1 shows the resultant point pattern for Cook County, IL, which includes the city of Chicago. We can now measure the degree of segregation as the amount of positive spatial autocorrelation (clumping) of the point pattern. As this does not involve any subdivisions, the result is unequivocal. Point pattern can also provide an accurate depiction of racial geography (see Figure 1). However, analyzing the point pattern composed of millions of points is difficult and time-consuming. Here we present a methodology, referred to as the "racial landscape" (RL) method [1], which replaces inhabitants by their race-specific densities, thus expediting the calculations. After describing the RL method, we focus on its implementation in the R computational environment.

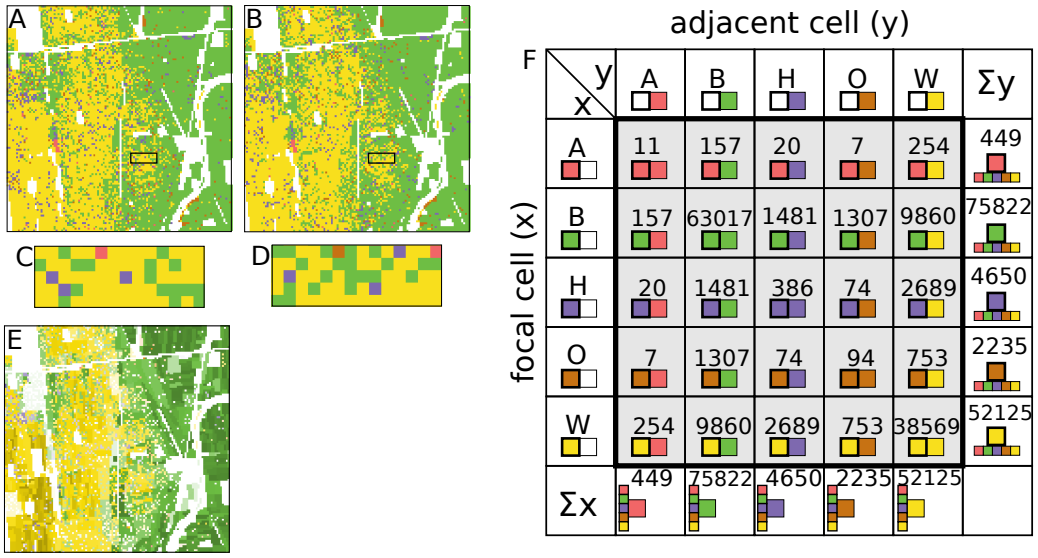

Figure 2 (A, B) Two different random realizations of a racial pattern, no perceivable differences on the scale of the entire area. (C, D) Magnification of patterns to a single block shows differences between two patterns on such a small length scale. (E) Racial map (landscape), which takes into consideration variable population density. (F) Exposure matrix for the pattern shown in (E); The EM is the part of the figure within a heavy line square. EM entries are rounded to the nearest integer. Capital letters indicate race category: A-Asians, B-Blacks, H-Hispanics, O-Others, W-Whites.

2 Method description

Consider a racially inhomogeneous cencus block divided into m cells and inhabited by k different races. We distribute block's inhabitants into $m_1, \ldots, m_k$, $\sum m_i = m$ race-specific types of equal-size cells. Numbers of cells of different race-type are in proportion to block's composition. Note that each cell is characterized by a race and a population *density* (not a population count). Thus, the size of the cell does not matter as long as it is smaller than the size of the block; larger cells will just be characterized by larger densities; the cells are not accumulation areas, instead they are elements of density sampling grid. Here we selected to use 30m×30m cells. Spatial assignment of different category cells *within a block* is stochastic, obtained by the Monte Carlo simulation. Thus, a racial pattern on the block scale is random, but the pattern on larger scales is not random. The result is a grid-based pattern, with each cell having a categorical label (race) and the value of local population density. Such a pattern is straightforward to visualize by assigning different cell colors to each race category and changing the intensity of the color depending on the cell's population density (see Figure 2E).

Consider a bi-variate distribution $f(x, y)$ – a probability of a randomly selected pair of adjacent (4-connectivity) pair of cells; a focus cell has a category x and an adjacent cell has a category y. $f(x, y)$ is calculated using the $n \times n$ normalized co-occurrence matrix (CM). From $f(x, y)$ we calculate the joint entropy $H(x, y)$, the marginal entropy $H(x)$, the conditional entropy $H(x|y)$, and the mutual information $I(x; y)$ (for details see [3]). The marginal entropy measures diversity while the mutual information measures spatial autocorrelation, and thus it measures segregation. Note that calculating $I(x; y)$ from the co-occurrence matrix quantifies segregation under an (unrealistic) assumption of constant population density.

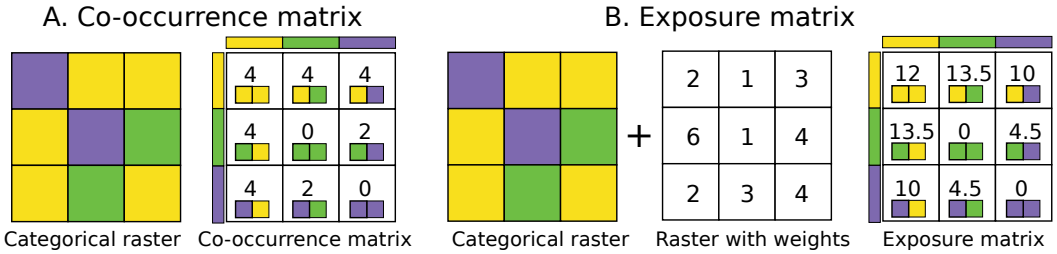

Figure 3 Quantification of spatio-racial pattern using co-occurrence (A) and exposure matrices (B). Each matrix need to normalized to obtain $f(x, y)$ from which $H(x)$ and $I(x; y)$ are calculated.

To measure segregation with actual densities, we introduce a modification of CM – each adjacency contributes a value of local (average over two adjacent cells) density to the matrix instead of the constant value of 1. We call such a modified matrix an exposure matrix (EM). The comparison between CM and EM is shown in Figure 3. Finally, we calculate $H(x)$ and $I(x; y)$ from EM instead of CM.

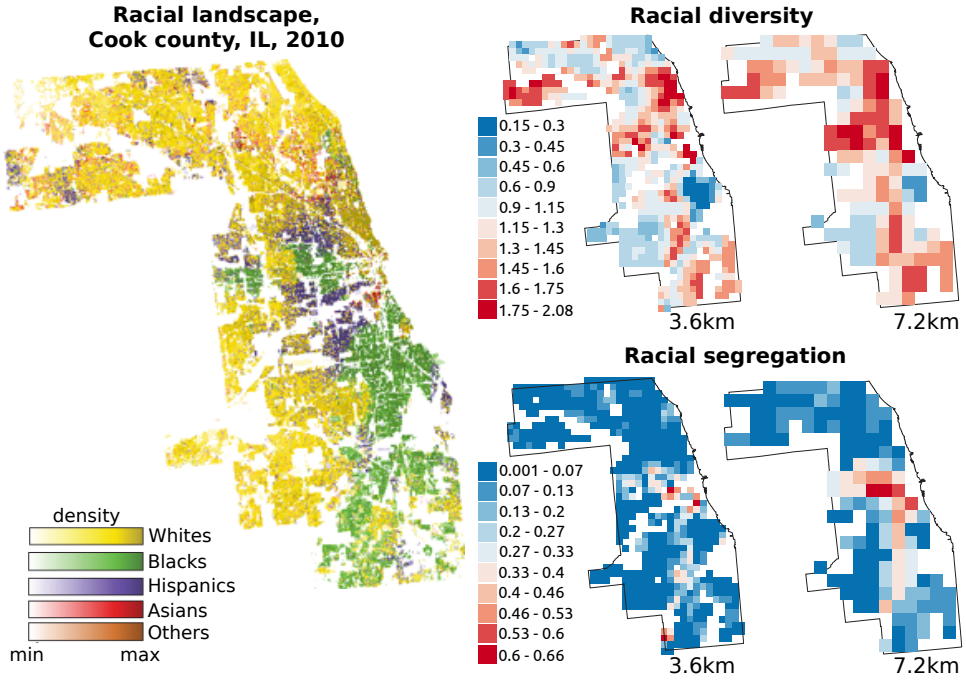

Figure 4 (Left) A spatio-racial pattern in the Cook county, IL in 2010 at the resolution of 30m. (Right) Maps of local diversity and segregation measured on spatial scales of 3.6 and 7.2km.

2.1 Example of application

Figure 4 shows some results of applying the RL method to Cook County in 2010. The results include the racial landscape visualization at the resolution of 30m and the racial diversity and segregation maps of local patterns having a scale of 3.6km and 7.2km, respectively. These scales were chosen because they are multiples of 30m. The racial landscape shows the distribution of 5 races, Whites, Blacks, Hispanics, Asians and Others. The racial diversity

and segregation maps show how segregation and diversity are measured locally. This shows spatial changes in local racial diversity and segregation throughout the extent of Cook County. Interestingly, notice that racial segregation is only noticeable on larger scales.

The diversity and segregation at the scale of the entire Cook County are 1.8393(5) and 0.4731(9), respectively. These numbers are ensemble averages from 100 realizations of Monte Carlo simulations. Numbers in brackets are the standard deviation calculated from 100 realizations. The segregation and diversity at the 3.6km scale are 1.133(8) and 0.075(6), and at the 7.2km are 1.29(4) and 0.14(2). These numbers are calculated in two steps: (1) the average value of metric for each local pattern is calculated from 100 realizations; (2) the average values from all local patterns are calculated. Note increase in values of uncertainties as the length scale decreases. Also, note that both diversity and segregation decrease with decreasing scale.

3 R implementation

The RL has been implemented in the R package – raceland (`https://cran.r-project.org/package=raceland`). The raceland package implements a complete computational framework that allows for: (1) Constructing racial landscape based on race-specific raster grids. (2) Quantifying the racial landscape pattern using metrics derived from the Information Theory concept (entropy and mutual information). (3) Visualizing the racial landscape, as well as racial diversity and segregation. These steps can be performed for the entire area or for any local pattern. The local pattern is a square-shaped block of cells that defines the spatial scale of the analysis (examples of a local pattern are shown in Figure 1).

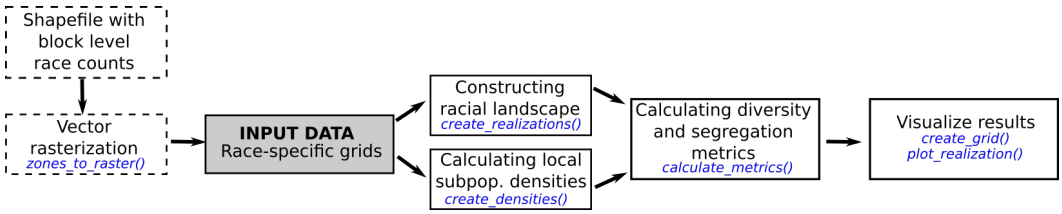

Figure 5 The racial landscape workfolow implemented in the R package raceland

The computational framework consists of a few steps (Figure 5):

1. ***Preprocessing census data.*** Racial landscape is calculated using high-resolution grids with each cell containing race-specific population density. Race-specific raster grids at the 30m resolution for each county in the United States can be downloaded from the SocScape project at `http://socscape.edu.pl`. Raceland package also allows performing calculation using a spatial vector file (e.g., shapefile) with attribute table containing race counts for aggregated data. In such case, a spatial vector object is first rasterized using the *zones_to_rasters()* function from the raceland package. People of a given race are redistributed to the cells by dividing the number of people by the number of cells belonging to the particular spatial units.
2. ***Constructing racial landscape.*** The racial landscape is a grid in each cell that has two attributes: a race category and a population density. The racial landscape is constructed based on race-specific grids. The method implemented in the raceland package (*create_realizations()* function) uses the cell's race probabilities and a random number generator to randomly assign a specific race label to each cell. The probabilities are established using the vector representing the racial composition in each cell. It is a

stochastic approach that yields to a series of realization with a slightly different pattern, but since all realizations have the same statistical properties, their metrics are similar. A single realization is sufficient for the visualization of the racial pattern. For increased accuracy, a spatio-racial pattern is quantified as an ensemble average from multiple realizations.

The second attribute of racial landscape constitutes of population density information calculated using the function *create_densities()*. Inclusion of population density is required in order to the correct assessment of racial segregation.

3. ***Quantifying racial landscape.*** The spatio-racial pattern is quantified using exposure matrix and further summarized by marginal entropy and mutual information. The marginal entropy characterized the level of diversity, whereas mutual information is associated with measuring racial segregation. In the raceland package, the calculation of the exposure matrix is build-in into the *calculate_metrics()* function. Exposure matrix alone can be calculated using the *get_wecoma()* function from the comat package [4].
4. ***Mapping racial diversity and segregation.*** The average value of entropy and mutual information calculated for each square-shaped local pattern from all realizations can be joined to the spatial grid object. Such operation allows for mapping metrics and shows how segregation and racial diversity change over the area. These two metrics can also be classified into nine classes and visualized as a single map. These classes represent the configuration types of the local racial pattern.

4 Conclusion

The RL method addresses the limitations of methods currently used by racial demographers. It yields an easy-to-understand racial map that quickly conveys the racial character of a city. Because the method does not use divisions into census aggregation areas, the value of city/region segregation is calculated unequivocally. The traditional method can only calculate segregation for a region that is divided into census aggregation areas. In practice, only segregation values of the entire city/metro area divided into census tracts are calculated. The RL method can calculate the value of segregation for any subset of the entire area. Therefore, it is possible to construct maps showing spatial variability of local segregation (see Figure 4). The RL method is more computationally complex than the traditional method. Thus, we have written the R package raceland that implements the RL method.

References

1 Anna Dmowska, Tomasz F. Stepinski, and Jakub Nowosad. Racial Landscapes–a pattern-based, zoneless method for analysis and visualization of racial topography. *Applied Geography*, 122:102239, 2020.

2 Douglas S. Massey and Nancy A. Denton. The dimensions of residential segregation. *Social forces*, 67(2):281–315, 1988.

3 Jakub Nowosad and Tomasz F Stepinski. Information theory as a consistent framework for quantification and classification of landscape patterns. *Landscape Ecology*, 34(9):2091–2101, 2019.

4 Jakub Nowosad and Tomasz F. Stepinski. Pattern-based identification and mapping of landscape types using multi-thematic data. *International Journal of Geographical Information Science*, Latest Articles:1–16, 2021.

5 Michael J. White. Segregation and diversity measures in population distribution. *Population index*, 52(2):198–221, 1986.