Exploring the correlation between city size and residential segregation: comparing Chilean cities with spatially unbiased indexes

Online supplementary information

Methodological annex: formulation of segregation indexes

This annex provides detailed mathematical formulations of the three segregation indexes that have been reported.

First, the Jargowsky[1] neighbourhood sorting index with individual data of household heads is defined as follows:

$$J = \frac{(H^{-1}\sum_n H_n(\bar{v}_n - \bar{v})^2)^{1/2}}{(H^{-1}\sum_h (v_h - \bar{v})^2)^{1/2}} \quad (1)$$

where H is the number of households, indexed by h; n is an index of zones; and v is the number of years of formal education of household heads (income in Jargowsky[2]). This index is a ratio of variation among zones over total variation among households, adjusted for a city's socioeconomic structure in a useful way for comparative urban studies.

In each city, Jargowsky indexes were calculated for each step in a process of hierarchical spatial clustering,[3] starting from blocks. In each iteration, two areas were joined following an objective function of the minimal increase in the internal sum of squares of education years among household heads. This is akin to Ward's[4] hierarchical clustering method, but including a neighbourhood restriction that is defined by the linkages of a Delaunay triangulation among block centroids. Thus, only contiguous units can be merged, in order to produce a continuous and exhaustive partitioning of urban space. For each city, a complete set of Jargowsky indexes was calculated for every scale of aggregation, from blocks to the partition of the city into two homogeneous urban areas.

In order to control spurious aggregation effects, at each aggregation level we calculated the difference between the Jargowsky indexes obtained with real data and the average of 50 random distributions of education years among household heads, for each city. This strategy is based on the "gap statistic",[5] an estimate of the optimal number of clusters based on the comparison between actual dispersion indexes and their expectations under a random distribution. In this Monte Carlo analysis, the values of study years were shuffled among fixed residences, in order to maintain identical statistical distributions of the main variable and stable patterns of population densities. The areal dissimilarity defined by the aggregation of random data is completely spurious and measures the spatial aggregation bias. Thus, we discounted these spurious effects in the calculation of two variants

---

[1] Jargowsky, P A (1996), "Take the money and run: economic segregation in U.S. metropolitan areas", *American Sociological Review* Vol 61, No 6, pages 984–998.

[2] See reference 1.

[3] Openshaw, Stan and P Taylor (1979), "A million or so correlation coefficients: three experiments on the modifiable areal unit problem", in N Wrigley (editor), *Statistical Applications in the Spatial Sciences* Vol 21, Pion, London, pages 127–144.

[4] Ward, J H (1963), "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association* Vol 58, No 301, pages 236–244.

[5] Tibshirani, R, G Walther and T Hastie (2001), "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* Vol 63, No 2, pages 411–423.

of Jargowsky's index. First were the Jargowsky differences between real data and shuffled data, allowing us to determine the strongest segregation scale for each city (Figure 2 in the main text). Second was the Jargowsky integral (*JI*), the average of the Jargowsky differences among all aggregation levels, formulated as follows:

$$JI = n^{-1} \sum_{1}^{n} JR_n - \overline{JS_n} \quad (2)$$

where *JR* and *JS* are Jargowsky's indexes (defined in formula 1), respectively calculated with real and shuffled data at every aggregation step from 1 to *n*, with *n* being equivalent to the number of city blocks. *JS* values were averaged for 50 random permutations of education years among household heads.

Second, we measured evenness from an egocentric spatial perspective, calculating the Jargowsky index (formula 1) for concentric areas around each block, with 300-, 600- and 1,000-metre radiuses. This is another way to develop a multiscalar segregation analysis that avoids arbitrary boundary definitions.[6]

Third, we calculated the Global Moran index[7] for each city in order to measure the spatial autocorrelation of high and low values of the average number of education years among blocks, providing a measure of the concentration of residences with higher and lower socioeconomic levels. In order to avoid underestimating segregation in small cities, we considered the 12 nearest neighbours to each block, which roughly corresponds to a 300-metre radius, although there are greater distances between large blocks in urban peripheries. The resulting spatial relationships were weighted by the number of households in each block. This index is formulated as follows:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \quad (3)$$

where z_i is the deviation of average years of formal education in a block from the city-level mean, $w_{i,j}$ is the spatial weight between blocks *i* and *j*, *n* is the number of blocks and S_0 is the aggregate of all the spatial weights.

[6] Reardon, S F and D O'Sullivan (2004), "Measures of spatial segregation", *Sociological Methodology* Vol 34, No 1, pages 121–162.

[7] Anselin, L (1995), "Local indicators of spatial association—LISA", *Geographical Analysis* Vol 27, No 2, pages 93–115.

TABLE S1
Regression results for segregation indexes with city size and control variables

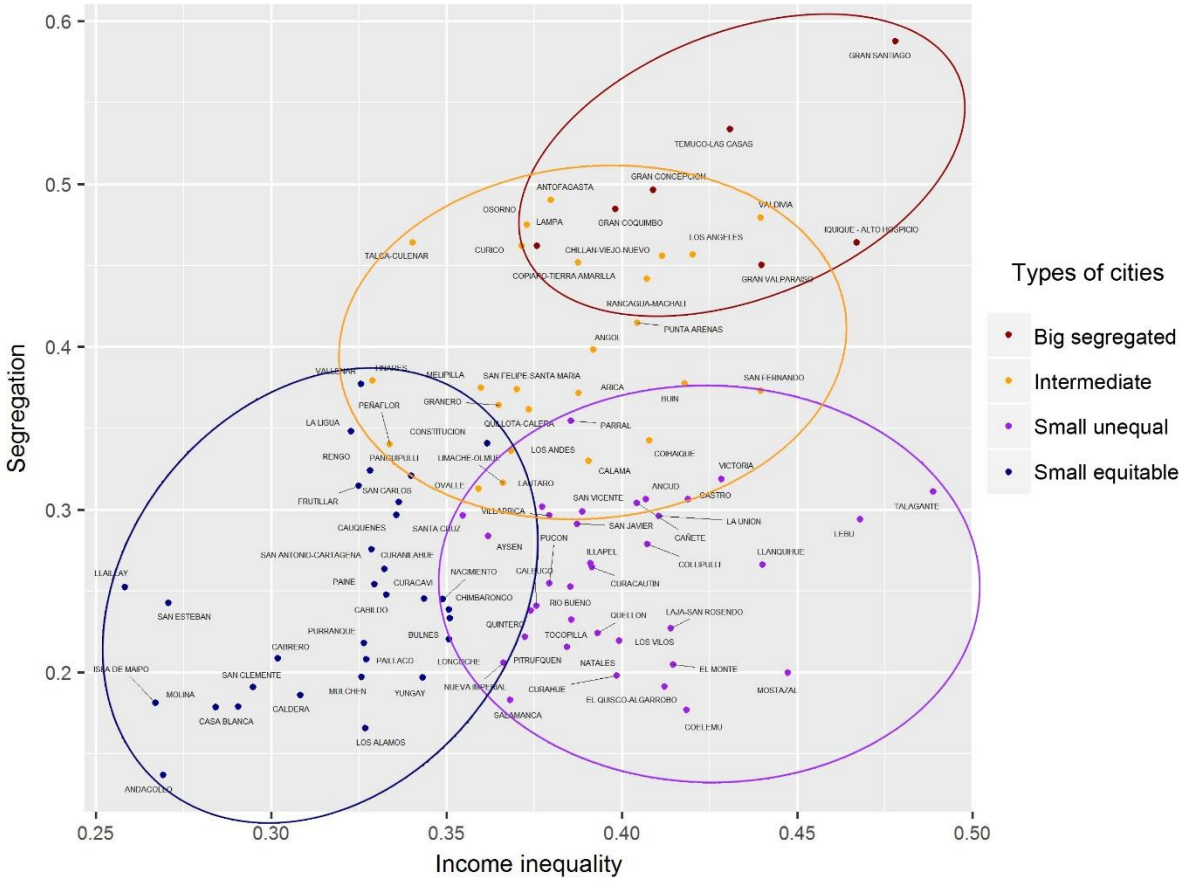
Dependent variable: Jargowsky difference			
Adj. R^2: 0.755		p-value: < 0.000	
Variables	Estimate	Std. error	Pr(>|t|)
(Intercept)	-0.367	0.064	0.000
Log households	0.070	0.005	0.000
Income inequality (Gini)	0.188	0.092	0.045
Median income	0.000	0.000	0.211
Multidimensional poverty rate	-0.112	0.089	0.211
New residents (%)	0.283	0.133	0.036
Adjacency to sea or lake (dummy)	-0.010	0.012	0.416
Topographic roughness	0.000	0.000	0.188

Dependent variable: egocentric evenness 300m			
Adj. R^2: 0.749		p-value: < 0.000	
Variables	Estimate	Std. error	Pr(>|t|)
(Intercept)	-1.372	0.151	0.000
Log households	0.172	0.013	0.000
Income inequality (Gini)	0.233	0.217	0.286
Median income	0.000	0.000	0.615
Multidimensional poverty rate	0.094	0.209	0.652
New residents (%)	0.102	0.312	0.745
Adjacency to sea or lake (dummy)	-0.024	0.029	0.409
Topographic roughness	-0.001	0.001	0.112

Dependent variable: global Moran			
Adj. R^2: 0.597		p-value: < 0.000	
Variables	Estimate	Std. error	Pr(>|t|)
(Intercept)	-0.563	0.135	0.000
Log households	0.104	0.012	0.000
Income inequality (Gini)	0.178	0.195	0.362
Median income	0.000	0.000	0.856
Multidimensional poverty rate	-0.111	0.187	0.553
New residents (%)	0.229	0.280	0.414
Adjacency to sea or lake (dummy)	-0.014	0.026	0.594
Topographic roughness	-0.001	0.000	0.255

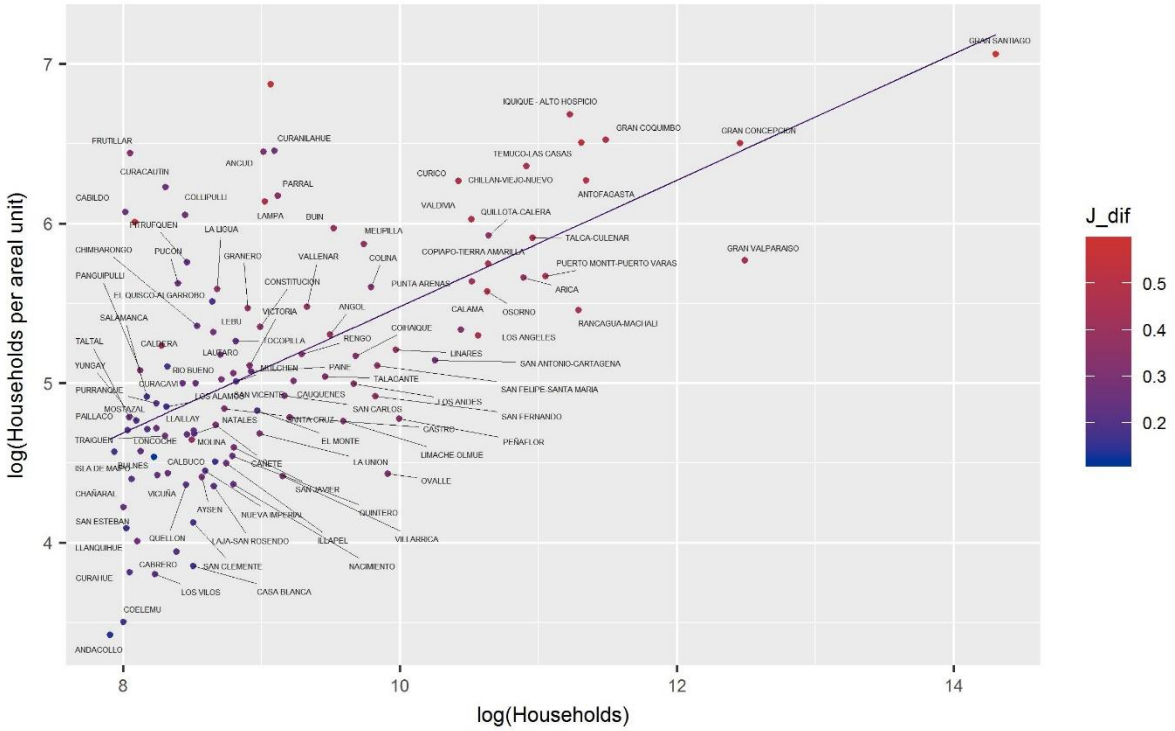
NOTE:
The model for the Jargowsky integral index is not shown, as it has very similar significant variables and coefficients to the Jargowsky difference model.

FIGURE S1
Cities by segregation, size, inequality and residential mobility



SOURCE: authors.

FIGURE S2
Size of clusters by city size



NOTE:
J_dif = Jargowsky difference.

SOURCE: authors.

END REFERENCES

Anselin, L (1995), "Local indicators of spatial association—LISA", *Geographical Analysis* Vol 27, No 2, pages 93–115.

Jargowsky, P A (1996), "Take the money and run: economic segregation in U.S. metropolitan areas", *American Sociological Review* Vol 61, No 6, pages 984–998

Openshaw, Stan and P Taylor (1979), "A million or so correlation coefficients: three experiments on the modifiable areal unit problem", in N Wrigley (editor), *Statistical Applications in the Spatial Sciences* Vol 21, Pion, London, pages 127–144.

Reardon, S F and D O'Sullivan (2004), "Measures of spatial segregation", *Sociological Methodology* Vol 34, No 1, pages 121–162.

Tibshirani, R, G Walther and T Hastie (2001), "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* Vol 63, No 2, pages 411–423.

Ward, J H (1963), "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association* Vol 58, No 301, pages 236–244.