LISER

LUXEMBOURG INSTITUTE OF SOCIO-ECONOMIC RESEARCH

n° 2018-19 November 2018

WORKING PAPERS

Inference for the neighborhood inequality index

Francesco **ANDREOLI**[1]

[1] Luxembourg Institute of Socio-Economic Research (LISER), Luxembourg

Inference for the neighborhood inequality index*

Francesco Andreoli†

October 2018

Abstract

The neighborhood inequality (NI) index measures aspects of spatial inequality in the distribution of incomes within the city. The NI index is defined as a population average of the normalized income gap between each individual's income (observed at a given location in the city) and the incomes of the neighbors, living within a certain distance range from that individual. This paper provides minimum bounds for the NI index standard error and shows that unbiased estimators can be identified under fairly common hypothesis in spatial statistics. These estimators are shown to depend exclusively on the variogram, a measure of spatial dependence in the data. Rich income data are then used to infer about trends of neighborhood inequality in Chicago, IL over the last 35 years. Results from a Monte Carlo study support the relevance of the standard error approximations.

Keywords: income inequality, individual neighborhood, geostatistics, variogram, census, ACS, ratio measures, variance approximation, Chicago, Monte Carlo.

JEL codes: C12, C46, D63, R23.

*I am grateful to Eugenio Peluso for relevant comments and to the Department of Economics, University of Verona, for hospitality. This paper forms part of the research project *The Measurement of Ordinal and Multidimensional Inequalities* (grant ANR-16-CE41-0005-01) of the French National Agency for Research whose financial support is gratefully acknowledged. Replication code is made available on my web-page.

†Luxembourg Institute of Socio-Economic Research, LISER. MSH, 11 Porte des Sciences, L-4366 Esch-sur-Alzette/Belval Campus, Luxembourg. E-mail: francesco.andreoli@liser.lu.

1 Introduction

Cities are the most unequal places in America (Moretti 2013, Baum-Snow and Pavan 2013, Chetty and Hendren 2018), and increasingly so nowadays compared to the last few decades (Watson 2009). Income inequalities that arise from differences across neighborhoods and administrative areas of the city have received substantial attention in the literature (Massey and Eggers 1990, Jargowsky 1997, Reardon and Bischoff 2011). Much less is known about the extent of income inequality in the neighborhood (see Hardman and Ioannides 2004, Shorrocks and Wan 2005, Dawkins 2007, Wheeler and La Jeunesse 2008, Kim and Jargowsky 2009). The degree of inequality and poverty within the neighborhood of residence has been found to have an independent effect on important dimensions of quality of life, such as labor market attachment (Conley and Topa 2002), well-being (Ludwig et al. 2012) health (Ludwig et al. 2011, Ludwig et al. 2013, Chetty et al. 2016) and intergenerational mobility (Andreoli and Peluso 2018).

Existing measures of neighborhood inequality either allow to identify places in the city where the poor population is over-represented compared to the city average (Reardon and Bischoff 2011, Iceland and Hernandez 2017), or rely on variance decompositions methods based on administrative partitions of the urban territory (for instance, by census tract or school district) to measure value the contribution of inequality across areas to total inequality (Wheeler and La Jeunesse 2008, Shorrocks and Wan 2005). These approaches to spatial inequality raise concerns on their normative validity, as well as on their reliability from a measurement perspective, insofar inequality indices based on the urban space partition are inevitably affected by the Modifiable Areal Unit Problem (Openshaw 1983, Wong 2009).

In a recent contribution, Andreoli and Peluso (2018) have developed a new spatial measures of neighborhood inequality, the NI index, that addresses these critics by relying on the notion of individual neighborhood (Galster 2001, Clark et al. 2015), rather than administrative neighborhood, to estimate neighborhood inequality. The index is defined as follows. Consider a population of $n \geq 3$ individuals, indexed by $i = 1, ..., n$, and let y_i be the income of individual i and $\mathbf{y} = (y_1, y_2, ..., y_n)$ the sample income distribution with

average $\mu > 0$. The authors assume that information on incomes comes with information about their location on the city map. In this way, they can construct individual neighborhoods d_i, gathering n_{id} individuals living within a distance range d from any given location i. There are as many individual neighborhoods as individuals in the city, and each individual neighborhood can be characterized by an average income $\mu_{id} = \frac{\sum_{j \in d_i} y_j}{n_{id}}$, and a degree of inequality $\Delta_i(\mathbf{y}, d) = \frac{1}{\mu_{id}} \sum_{j \in d_i} \frac{|y_i - y_j|}{n_{id}}$. The NI index measures spatial inequality in the city as the degree of inequality in the average individual neighborhood. It is defines as

$$NI(\mathbf{y}, d) = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{n} \Delta_i(\mathbf{y}, d), \tag{1}$$

and it is shown to have solid normative and statistical properties (it can be related, for instance, to the probabilistic interpretation of the Gini coefficient in Pyatt 1976).

The NI index depends on d, which is a parameter chosen by the researcher. The plot of $NI(\mathbf{y}, d)$ against d defines a *neighborhood inequality curve*. The curve is expected to be close to the origin when $d = 0$ (individual neighborhoods are very small) or when there is high spatial dependence in incomes, with high and low income households segregated in space. When d reaches the size of the city, each individual neighborhood spans the whole city. In this case, neighborhood inequality converges to citywide inequality measured by the Gini index and the NI curve is flat. Andreoli and Peluso (2018) make use of the NI curve to assess neighborhood inequality in American metro areas. Their findings suggest that neighborhood inequality in American cities is: i) high and close to citywide inequality even when d is small; ii) on the rise since 1980s; iii) a predictor of children future income opportunities related to the place they were exposed to during youth (as estimated in Chetty and Hendren 2018).

These findings are based on the American census (STF 3A files) and the Community Survey (ACS) data. Both data sources only report statistical tables of demographics and of population counts at given income cutoffs that are representative at the block group level, the finest available statistical partition of the American territory. Incomes in the ACS are estimated from 5-years rotational samples based on 2010 census population counts. There is therefore a possibility that the NI index is imprecisely estimated and that

the trends of growing NI trends observed in Andreoli and Peluso (2018) are not robust from a statistical perspective.

In this paper, I derive minimum bounds for the standard error of the NI index and use these bounds to infer about various forms of dominance in NI curves, inspired by the stochastic dominance testing approach (Bishop, Chakraborti and Thistle 1989, Dardanoni and Forcina 1999, Andreoli 2018). I utilize some properties of the ratio estimators in Goodman and Hartley (1958) to derive bounds for the NI index variance when the data generating process is not i.i.d., accommodating for the possibility of spatial dependence. I then show (Sections 2 and 3) that under fairly common assumptions in spatial statistics (Cressie 1991, Chilès and Delfiner 2012), the estimators of the NI index standard error can be identified in terms of the distribution of locations on the map (non stochastic) and of the variogram, a measure of spatial dependence of the data (Matheron 1963). I use these results in Section 4 to infer about changes in NI in Chicago, IL, where I find statistical support for findings in Andreoli and Peluso (2018). A simulation study in Section 5 confirms the qualities of the standard error estimator I propose. Section 6 concludes.

2 Statistical properties of the NI index

2.1 Setting

Let $\mathcal{S}$ denote a random field. The spatial process $\{Y_s : s = 1, \ldots, n\}$ with $s \in \mathcal{S}$ is jointly distributed as $\mathcal{F}_{\mathcal{S}}$. This process is a collection of random variables Y_s located over the random field $\mathcal{S}$, which serves as a model of the relevant urban space. The joint distribution function $\mathcal{F}_{\mathcal{S}}$ combines information about the marginal income distributions in each location and the degree of spatial dependence of incomes on $\mathcal{S}$. Through geolocalization, it is possible to compute the distance "$||.||$" between locations $s, v \in \mathcal{S}$. Let $||s - v|| \leq d$ indicate that the distance between the two locations is smaller than d, or equivalently $v \in d_s$. The cardinality of the set of locations d_s is n_{d_s}, while n is the total number of locations. The observed income distribution $\mathbf{y}$ is a particular draw from $\mathcal{F}_{\mathcal{S}}$, where only one income realization is observed in location s.

Assume that data come equally spaced on a grid, so that for any two points $s, v \in \mathcal{S}$ such that $||v-s|| = h$ we write $v = s+h$. The process distributed as $\mathcal{F}_{\mathcal{S}}$ is said to display *intrinsic (second-order) stationarity* (see Chilès and Delfiner 2012) if $E[Y_s] = \mu$, $Var[Y_s] = \sigma^2$ and $Cov[Y_s, Y_v] = c(h)$ where the covariance function is isotropic and $v = s+h$. Under these circumstances, let $Var[Y_{s+h} - Y_s] = E[(Y_{s+h} - Y_s)^2] = 2\sigma^2 - 2c(h) = 2\gamma(h)$ denote the *variogram* of the process at distance range h (Matheron 1963). The function $2\gamma(h)$ is informative of the correlation between two random variables that are exactly d distance units away one from the other. The slope of the graph of the variogram function displays the extent to which spatial association affects the joint variability of the elements of the process. In general, $2\gamma(d) \to 0$ as d approaches 0, indicating that random variables that are very close in space tend to be strongly spatially correlated and variability in incomes at the very local scale is small. Conversely, $2\gamma(d) \to 2\sigma^2$ when d is sufficiently large, indicating spatial independence between two random variables Y_s and Y_v far apart on the random field.

Noticing that $E[Y_{s+h} \cdot Y_s] = \sigma^2 - \gamma(h) + \mu^2$, the covariance between differences in random variables can be written as $Cov[(Y_{s+h_1} - Y_s), (Y_{v+h_2} - Y_v)] = \gamma(s+h_1-v) + \gamma(s-(v+h_2)) - \gamma(s-v) - \gamma(s+h_1-(v+h_2))$ as in Cressie and Hawkins (1980) and Cressie (1991). Sine data are assumed to occur on a transect, let denote by s and v the position on the transect, and consider $s - v = h \geq 0$ where h indicates that the random variables are located within distance range h. The transect can be directional, implying that negative and positive distances carry relevant information when aggregated. Let $\delta_p = 1$ whenever $h_p > 0$ and $\delta_p = -1$ whenever $h_p < 0$, $p = 1, 2$. Under these circumstances: $Cov[(Y_{s+h_1} - Y_s), (Y_{v+h_2} - Y_v)] = \gamma(|h+\delta_1 h_1|) + \gamma(|h-\delta_2 h_2|) - \gamma(|h|) - \gamma(|h+\delta_1 h_1 - \delta_2 h_2|)$. Consider further the possibility of abandoning directional information by assuming that locations are arranged so that $h_1 > 0$ and $h_2 > 0$ and adopt the convention that $\gamma(-h) = \gamma(h)$ (i.e. only the order but not the direction on the transect matters), then the covariance is identified as $Cov[(Y_{s+h_1} - Y_s), (Y_{v+h_2} - Y_v)] = \gamma(h+h_1) + \gamma(h-h_2) - \gamma(h) - \gamma(h+h_1-h_2)$

Let now introduce one additional distributional assumption: Y_s is gaussian with mean μ and variance σ^2. The random variable $(Y_{s+h} - Y_s)$ is also gaussian with variance $2\gamma(h)$,

which implies $|Y_{s+h} - Y_s|$ is *folded-normal* distributed (Leone, Nelson and Nottingham 1961) and its first and second moments depend exclusively on the variogram, having expectation $E[|Y_{s+h} - Y_s|] = \sqrt{2/\pi Var[Y_{s+h} - Y_s]} = 2\sqrt{\gamma(h)/\pi}$ and variance $Var[|Y_{s+h} - Y_s|] = (1 - 2/\pi)2\gamma(h)$.

These results turn out to be useful in characterizing the NI index.

2.2 Properties

The NI index of the spatial process $\mathcal{F}_{\mathcal{S}}$ can be written in terms of first order moments of the random variables Y_s as follows:[1]

$$NI(\mathcal{F}_{\mathcal{S}}, d) = \sum_{s} \sum_{v \in d_s} \frac{1}{2n\, n_{d_s}} \frac{E[|Y_s - Y_v|]}{E[Y_v]}.$$

The degree of spatial dependence represented by $\mathcal{F}_{\mathcal{S}}$ enters in the NI formula through the expectation terms conditional on $\mathcal{S}$. Consider first the case in which $\mathcal{F}_{\mathcal{S}}$ displays no spatial dependence in incomes, that is, the random variables Y_s and Y_v are i.i.d. for any $s, v \in \mathcal{S}$. One direct implication is that $NI(\mathcal{F}_{\mathcal{S}}, d) = \frac{E[|Y_s - Y_v|]}{E[Y_v]}$, which coincides with the definition of the standard Gini inequality coefficient (see for instance Muliere and Scarsini 1989).

If, instead, spatial dependence is at stake, then the expectation $E[|Y_s - Y_v|]$ varies across locations and cannot be identified and estimated from the observation of just one data point in each location. More structure is needed.

I maintain the assumption that the spatial process is defined on the transect with equally spaced lags. For given d, I can thus partition the distance spectrum $[0, d]$ into B_d ordered intervals of fixed size d/B_d. Each interval is denoted by the index b with $b = 1, \ldots, B_d$. I further denote with d_{bi} the set of locations at interval b (and thus distant $b \cdot d/B_d$ from s_i) within the range d from location s_i. The cardinality of this set is $n_{d_{bi}} \leq n_{d_i} \leq n$. Assuming additionally the intrinsic stationarity of $\mathcal{F}_{\mathcal{S}}$ and normality,

[1] Biondi and Qeadan (2008) use a related estimator to assess dependency across time in paleorecords observed in a given location.

the NI index rewrites:

$$\begin{aligned} NI(\mathcal{F}_{\mathcal{S}}, d) &= \sum_i \sum_{j \in d_i} \frac{1}{2n\, n_{d_i}} \frac{E[|Y_{s_j} - Y_{s_i}|]}{\mu} \\ &= \sum_i \sum_{j \in d_i} \frac{1}{2n\, n_{d_i}} \frac{\sqrt{4\gamma(||s_j - s_i||)/\pi}}{\mu} \\ &= \sum_i \frac{1}{n} \sum_{b=1}^{B_d} \frac{n_{d_{bi}}}{n_{d_i}} \sum_{j \in d_{bi}} \frac{1}{2\, n_{d_{bi}}} \frac{\sqrt{4\gamma(s_i + b - s_i)/\pi}}{\mu} \\ &= \frac{1}{2} \sum_{b=1}^{B_d} \left(\sum_i \frac{n_{d_{bi}}}{n\, n_{d_i}} \right) \frac{\sqrt{4\gamma(b)/\pi}}{\mu}, \end{aligned} \tag{2}$$

This result, derived in Andreoli and Peluso (2018), shows that the NI index is fully characterized by the distribution of locations on the city map (non stochastic) and the degree of spatial dependence measured by the variogram. I exploit this property to obtain estimators for the NI index standard errors.

3 Variance bounds for the NI index

3.1 Main result

I derive minimum bounds for the SE of the NI index under three assumptions: 1) the underling spatial process is stationary; 2) the spatial process occurs on a transect at equally spaced points; 3) each element of the process is gaussian with expectation μ and variance σ^2.

Let assume that the random field $\mathcal{S}$ is limited to n locations. For simplicity, I denote these locations by i such that $i = 1, \ldots, n$ and $\{Y_i : i = 1, \ldots, n\}$. The joint distribution of the process is $\mathcal{F}$. Each location has weight $w_i \geq 0$ with $w = \sum_i w_i$, which might reflect the underling population density at a given location. These weights are assumed to be non-stochastic. The first implication is that, asymptotically, the random variable $\mu_{id} = \sum_{j \in d_i} \frac{w_j}{\sum_{j \in d_i} w_j} Y_j$ is equivalent in expectation to $\tilde{\mu} = \sum_i \frac{w_i}{\sum_i w_i} Y_i$, i.e., $E[\tilde{\mu}] = \mu$. The second implication is that the spatial correlation exhibited by $\mathcal{F}$ is stationary in d and

can be represented through the variogram of $\mathcal{F}$, denoted $2\gamma(d)$.

An asymptotically equivalent version of the weighted NI index of the process distributed as $\mathcal{F}$ is

$$NI(\mathcal{F}, d) = \frac{1}{2\mu} \sum_{i=1}^{n} \sum_{j \in d_i} \frac{w_i w_j}{2\, w \sum_{j \in d_i} w_j} |Y_i - Y_j| = \frac{1}{2\mu} \Delta_d. \tag{3}$$

The NI index can thus be expressed as a ratio of two random variables. Asymptotic approximations for the SE of ratios of random variables have been developed in Goodman and Hartley (1958, p. 496) and later by Koop (1964) and Tin (1965). I use these results to obtain minimum variance bounds for the NI index in (3) as follows:

$$\begin{aligned} Var\left[NI(\mathcal{F}, d)\right] &= \frac{1}{4n\mu^2} Var[\Delta_d] + \frac{(NI(\mathcal{F}, d))^2}{n\mu^2} Var[\tilde{\mu}] - \\ &\quad \frac{NI(\mathcal{F}, d)}{n\mu^2} Cov[\Delta_d, \tilde{\mu}] + O(n^{-2}), \end{aligned} \tag{4}$$

where the SE approximation is $SE_d = \sqrt{Var\left[NI(\mathcal{F}, d)\right]}$ at any d. The approximation converges quickly when the number of locations is large, as it the case in applications based on census micro data, and holds when income realizations are spatially correlated.[2] As suggested in Tin (1965), I use plug-in estimators for the SE.

The three assumptions stated above allow to identify the different elements in (4). Let scalars m, b, b' identify distances on the transect. Under assumption 1) and 2) the

[2]The sample counterpart of the NI index in (3) can be interpreted as a U-statistic. As shown by Hoeffding (1948), the variance bound in (4) converges to the asymptotic unbiased estimator of the NI index variance when the income observations are i.i.d. Under this specific circumstance, asymptotic normality is also granted both with simple and with complex sampling design (Xu 2007, Davidson 2009).

variance of $\tilde{\mu}$ writes

$$\begin{aligned} Var[\tilde{\mu}] &= \sum_i \frac{w_i}{w} \sum_j \frac{w_j}{w} E[Y_i Y_j] - \mu^2 \\ &= \sum_i \frac{w_i}{w} \sum_{m=1}^{B} \frac{\sum_{j \in d_{mi}} w_j}{w} \sum_{j \in d_{mi}} \frac{w_j}{\sum_{j \in d_{mi}} w_j} c(||s_i - s_j||) \quad (5) \\ &= \sum_{m=1}^{B} \left(\sum_i \frac{w_i}{w} \frac{\sum_{j \in d_{mi}} w_j}{w} (\sigma^2 - \gamma(m)) \right) \quad (6) \\ &= \sigma^2 - \sum_{m=1}^{B} \omega(m)\gamma(m), \quad (7) \end{aligned}$$

where (7) is obtained from (6) by renaming the weight scores so that $\sum_{m=1}^{B} \omega(m) = 1$, and by using the definition of the variogram and the fact that $s_j = s_i + m$.

The second variance component of (4) can be written as follows:

$$\begin{aligned} Var[\Delta_d] &= \sum_{i=1}^{n} \sum_{j \in d_i} \frac{w_i w_j}{w \sum_{j \in d_i} w_j} \sum_{\ell=1}^{n} \sum_{k \in d_\ell} \frac{w_\ell w_k}{w \sum_{k \in d_\ell} w_k} E[|Y_i - Y_j||Y_\ell - Y_k|] \\ &\quad - \left(\sum_i \frac{w_i}{w} \sum_{j \in d_i} \frac{w_j}{\sum_{j \in d_i} w_j} E[|Y_j - Y_i|] \right)^2 . \end{aligned}$$

The first component of $Var[\Delta_d]$ cannot be further simplified, as the absolute value operator enters the expectation term in a multiplicative way. Under assumption 3), the expectation can be simulated, since the random vector (Y_j, Y_i, Y_k, Y_ℓ) is jointly normally distributed with expectations (μ, μ, μ, μ) and variance-covariance matrix $Cov[(Y_j, Y_i, Y_k, Y_\ell)]$, with:

$$Cov[(Y_j, Y_i, Y_k, Y_\ell)] = \begin{pmatrix} \sigma^2 & c(||s_j - s_i||) & c(||s_j - s_k||) & c(||s_j - s_\ell||) \\ & \sigma^2 & c(||s_i - s_k||) & c(||s_i - s_\ell||) \\ & & \sigma^2 & c(||s_k - s_\ell||) \\ & & & \sigma^2 \end{pmatrix}.$$

Assume further that data occur on a transect with equally spaced intervals $s_j - s_i = b \geq 0$ and $s_k - s_\ell = b' \geq 0$ for the positive integers $b \leq B_d$ and $b' \leq B_d$. I aslo take the

(unrestrictive) convention that $s_i - s_\ell = m$ with $0 \leq m \leq B$. I can hence express the variance-covariance matrix as a function of the variogram

$$Cov[(Y_j, Y_i, Y_k, Y_\ell)] = \begin{pmatrix} \sigma^2 & \sigma^2 - \gamma(b) & \sigma^2 - \gamma(m+b-b') & \sigma^2 - \gamma(m+b) \\ & \sigma^2 & \sigma^2 - \gamma(m-b') & \sigma^2 - \gamma(m) \\ & & \sigma^2 & \sigma^2 - \gamma(b') \\ & & & \sigma^2 \end{pmatrix}.$$

The expectation $E[|Y_i - Y_j||Y_\ell - Y_k|]$ can be simulated from a large number S (with $S = 1,000$) of independent draws $(y_{1s}, y_{2s}, y_{3s}, y_{4s})$ with $s = 1, \ldots, S$, from the random vector (Y_j, Y_i, Y_k, Y_ℓ). The simulated expectation is a function of the variogram parameters m, b, b' and d and of σ^2. It is denoted $\theta(m, b, b', d, \sigma^2)$ and estimated as follows:

$$\theta(m, b, b', d, \sigma^2) = \frac{1}{S} \sum_{s=1}^{S} |y_{2s} - y_{1s}| \cdot |y_{4s} - y_{3s}|.$$

With some algebra, and using the fact that $E[|Y_\ell - Y_i|] = 2\sqrt{\gamma(m)/\pi}$ for locations ℓ and i at distance $m \leq B$ one from each other, it is then possible to write the term $Var[\Delta_d]$ as follows:

$$\begin{aligned} Var[\Delta_d] &= \sum_{m=1}^{B} \sum_{b=1}^{B_d} \sum_{b'=1}^{B_d} \omega(m, b, b', d) \theta(m, b, b', d, \sigma^2) \\ &\quad -4 \left(\sum_{m}^{B_d} \omega(m, d) \sqrt{\gamma(m)/\pi} \right)^2. \end{aligned} \tag{8}$$

In the formula, $\omega(m, b, b', d) = \sum_i \frac{w_i}{w} \sum_{j \in d_{bi}} \frac{w_j}{\sum_{j \in d_i} w_j} \sum_{\ell \in d_{mi}} \frac{w_\ell}{w} \sum_{k \in d_{b'\ell}} \frac{w_k}{\sum_{k \in d_\ell} w_k}$ and $\omega(m, d) = \sum_i \frac{w_i}{w} \sum_{j \in d_{mi}} \frac{w_j}{\sum_{j \in d_i} w_j}$ are calculated as before.

The third component of (4) is the covariance term. It can also be written as a function of the variogram. To show this, I maintain the convention that $s_i - s_\ell = m \geq 0$ and

$s_j - s_i = b \geq 0$. This gives the following equivalence:

$$\begin{aligned} E[|Y_j - Y_i|Y_\ell] &= E[|Y_jY_\ell - Y_iY_\ell|] = E[Y_jY_\ell] - E[Y_iY_\ell] - 2E[\min\{Y_jY_\ell - Y_iY_\ell, 0\}] \\ &= c(||s_j - s_\ell||) + \mu^2 - c(||s_i - s_\ell||) - \mu^2 - 2E[\min\{Y_jY_\ell - Y_iY_\ell, 0\}] \\ &= \gamma(m) - \gamma(m+b) - 2E[\min\{Y_jY_\ell - Y_iY_\ell, 0\}]. \end{aligned} \tag{9}$$

The expectation $E[\min\{Y_jY_\ell - Y_iY_\ell, 0\}]$ is non-liner in the underlying random variables. Under the Gaussian assumption, the expectation can be simulated from a large number S (with $S = 1,000$) of independent draws (y_{1s}, y_{2s}, y_{3s}) with $s = 1, \ldots, S$, from the random vector (Y_j, Y_i, Y_ℓ), which is normally distributed with expectations (μ, μ, μ) and variance-covariance matrix $Cov[(Y_j, Y_i, Y_\ell)]$. As the process occurs on the transect, the variance-covariance matrix writes

$$Cov[(Y_j, Y_i, Y_\ell)] = \begin{pmatrix} \sigma^2 & \sigma^2 - \gamma(b) & \sigma^2 - \gamma(m+b) \\ & \sigma^2 & \sigma^2 - \gamma(m) \\ & & \sigma^2 \end{pmatrix}$$

for given m, b and d. The resulting simulated expectation is denoted $\phi(m, b, d, \sigma^2)$ and computed as follows:

$$\phi(m, b, d, \sigma^2) = \frac{1}{S} \sum_{s=1}^{S} \min\{y_{1s}y_{3s} - y_{2s}y_{3s}, 0\}.$$

Based on this result, the covariance term in (4) becomes:

$$\begin{aligned} Cov[\Delta_d, \tilde{\mu}] &= \sum_i \frac{w_i}{w} \sum_{j \in d_i} \frac{w_j}{\sum_{j \in d_i} w_j} \sum_\ell \frac{w_\ell}{w} E[|Y_j - Y_i|Y_\ell] \\ &\quad -\mu \sum_i \frac{w_i}{w} \sum_{j \in d_i} \frac{w_j}{\sum_{j \in d_i} w_j} E[|Y_j - Y_i|] \\ &= \sum_{m=1}^{B} \sum_{b=1}^{B_d} \omega(m, b, d) \left[\gamma(m) - \gamma(m+b) - 2\phi(m, b, d, \sigma^2)\right] \\ &\quad -2\mu \sum_{m=1}^{B_d} \omega(m, d) \sqrt{\gamma(m)/\pi}. \end{aligned} \tag{10}$$

The weights in (10) coincide respectively with $\omega(m,b,d) = \sum_i \frac{w_i}{w} \sum_{\ell \in d_{mi}} \frac{w_\ell}{w} \sum_{j \in d_{bi}} \frac{w_j}{\sum_{j \in d_i} w_j}$ and $\omega(m,d) = \sum_i \frac{w_i}{w} \sum_{j \in d_{mi}} \frac{w_j}{\sum_{j \in d_i} w_j}$.

A consistent estimator for the SE, denoted $\hat{SE}_d$, is obtained by plugging into (4) the empirical counterparts of the variogram and the lag-dependent weights, using the formulas in (7), (8) and (10).

3.2 Implementation

Consider a sample of size n. Income realizations are denoted y_i, with $i = 1, \ldots, n$. The income vector $\mathbf{y} = (y_1, \ldots, y_n)$ is a draw from the spatial random process $\{Y_s : s \in \mathcal{S}\}$, where a location s identifies a precise point on a map. Information about location (latitude and longitude) of an observation i is denoted by $s_i \in \mathcal{S}$. Distance measures between locations can be easily constructed based on the geodesic formula. Furthermore, observed incomes are associated with weights $w_i \geq 0$ and are indexed according to the sample units, with $w = \sum_i w_i$. It is often the case that the sample weights give the inverse probability of selection of an observation from the population.

The mean income within an individual neighborhood of size d, denoted μ_{id}, is estimated by $\widehat{\mu}_{id} = \sum_{j=1}^{n} \hat{w}_j y_j$ where

$$\hat{w}_j := \frac{w_j \cdot \mathbf{1}(||s_i - s_j|| \leq d)}{\sum_j w_j \cdot \mathbf{1}(||s_i - s_j|| \leq d)}$$

so that $\sum_j \hat{w}_j = 1$, and $\mathbf{1}(.)$ is the indicator function. The estimator of the average neighborhood mean income is instead $\hat{\mu}_d = \sum_{i=1}^{n} \frac{w_i}{w} \widehat{\mu}_{id}$. The estimator of the NI index, denoted $\hat{NI}(\mathbf{y}, d)$, is the sample weighted average of the mean absolute deviation of the income realization in location s_i from the income realization in any other location s_j such that $||s_i - s_j|| \leq d$. Formally

$$\hat{NI}(\mathbf{y}, d) = \sum_{i=1}^{n} \frac{w_i}{w} \frac{1}{2\hat{\mu}_{id}} \sum_{j=1}^{n} \hat{w}_j \, |y_i - y_j|,$$

where $\hat{w}_j$ is defined as above.

The estimation is conditional on d, which is a parameter under control of the researcher. The distance cutoff d is conventionally reported in miles and is meant to capture a continuous measure of the size of an individual neighborhood. In empirical applications, one can estimate as many values of d as there are pairs of observations in distinct locations on the maps. For computational reasons, however, the NI index and its SE are estimated only for a finite number of distance cutoffs, identifying intervals of fixed length. The maximum number of cutoffs indicates the point at which distance between observations is large enough that the NI index converges to the Gini index and its SE is constant. For a given neighborhood of size d, I partition the distance interval $[0, d]$, defining the size of the individual neighborhood, into K intervals $d_0, d_1, \ldots, d_K$ of equal size, with $d_0 = 0$. I always use d_k to denote the distance between any pair of observations i and j located at distance $d_{k-1} < ||s_i - s_j|| \leq d_k$ one from the other. The pairs $(d_k, \hat{NI}(\mathbf{y}, d_k))$ for any $k = 1, \ldots, K$ can be hence plotted on a graph. The curves resulting by linearly interpolating these points are the empirical equivalent of the neighborhood inequality curves.

A plug-in estimator for the asymptotic standard error of the GINI indices can be derived under the assumptions listed in the previous sections. The SE estimator crucially depends on four components: (i) the consistent estimator for the average $\tilde{\mu}$, denoted $\hat{\mu}$, which coincides with the sample average; (ii) the consistent estimator for variance σ^2, denoted $\hat{\sigma}^2$, which is given by the sample variance; (iii) the consistent estimator for the variogram; (iv) the estimator of the weighting schemes.

Empirical estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are standard. The robust non-parametric estimator of the variogram proposed by Cressie and Hawkins (1980) can be used to assess the pattern of spatial dependency of georeferenced data on income realizations. The empirical variogram is defined for given distance ranges, meaning that it produces a measure of spatial dependence among observations that are located exactly at a given distance range one from the others. I use $b = 1, \ldots, B$ to partition the empirical distance range between any given pair of locations into equally spaced lags. Then, I estimate the variogram on each of these lags. This means that $2\gamma(b)$ refers to the correlation between incomes placed

at distance lags of exactly b intervals, each of size d/B.

It is understood that the size of the sample is large compared to B, in the sense that the sampling rate per unit area remains constant when the partition into lags becomes finer. This assumption allows to estimate a non-parametric version of the variogram at every distance cutoff. Following Cressie (1991), I use weighted least squares to fit a theoretical variogram model to the empirical variogram estimates. The theoretical model is a continuous parametric function mapping distance into the corresponding variogram level. In the application, I use the spherical (semi)variogram model (see Cressie 1985), denoted $\gamma(h) = \alpha + \beta(3/2 \min\{h/a, 1\} - 0.5 \min\{h/a, 1\})$, where $\alpha,\ \beta$ are parameters to be estimated and a is the so-called range level: beyond distance a, the random variables Y_{s+h} and Y_s with $h > a$ are assumed to be spatially uncorrelated. The variogram satisfies the condition $\gamma(0) \to 0$ and $\gamma(a) = \sigma^2$. The max number of intervals B is set so that $d = 2a$. The estimated parameters are then used to draw predictions for the estimator $2\hat{\gamma}$ of the variogram at each distance cutoff.[3] The predictions are then plugged into the SE estimators of the NI index.

Finally, SE estimation requires to produce reliable estimators of the weights ω. These are non-parametrically identified from the formulas provided above. In some cases, however, computation of the exact weights requires several iterations across observations. The overall computation time thus increases exponentially in the number of observations and the procedure becomes quickly unfeasible. I propose alternative, feasible estimators for these weights, denoted $\hat{\omega}$, that are expressed as linear averages. The computational time is, nevertheless, quadratic in the number of observations as it requires to construct a routine that first computes weights estimators for each observation separately, and the averages across all observations at given distance cutoffs.

I consider here only the weights that appear in the estimators $\hat{SE}_d$ in (4) that cannot be directly inferred (i.e., are computationally unfeasible) from observed weights. For a given observation i, define $w(b, i) = \sum_{j \in d_{bi}} w_j$ for any ring $b = 1, \ldots, B_d, \ldots, B$ of radius d_b the weight associated with income realizations that are exactly located b lags away from

[3]Cressie (1985) has shown that this methodology leads to consistent estimates of the true variogram function under the stationarity assumptions mentioned above.

i. Then, denote $w(d,i) = \sum_{j \in d_i} w_j = \sum_{b=1}^{B_d} w(b,i)$. I consider the following estimators

$$\text{for (8)} \quad : \quad \hat{\omega}(m,b,b',d) = \sum_i \frac{w_i}{w} \frac{w(b,i)}{w(d,i)} \frac{w(m,i)}{w} \frac{w(m+b',i)}{w(m+d,i)};$$

$$\text{for (10)} \quad : \quad \hat{\omega}(m,b',d) = \sum_i \frac{w_i}{w} \frac{w(m,i)}{w} \frac{w(b',i)}{w(d,i)}.$$

To compute these weights, each observation i has to be first assigned with the total weight $w(b,i)$ of those observations $j \neq i$ that are located exactly at distance b from i. Then, $\hat{\omega}(m,b,b',d)$ and $\hat{\omega}(m,b',d)$ are obtained by averaging these weights across i's. The key feature of these estimators is that weights of observations occurring at distance b' from an observation located at distance m from i are estimated by averaging across all observations the relative weight of observations at distance $m+b'$ from i.

3.3 Hypothesis testing

The NI index and the implied NI curves can be used to assess patterns and trends of neighborhood inequality. Various hypotheses are of interest. One might be interested in assessing the extent at which inequality in the average individual neighborhood of size d is different from citywide inequality measured by the Gini index. The relevant null hypothesis is $H_0^1 : \ NI(\mathbf{y},d) = G(\mathbf{y})$ against an unrestricted alternative (reflecting the fact that neighborhood inequality can be either larger or smaller than citywide inequality). A second concern may be on the way the patterns of neighborhood inequality are sensible to the size of individual neighborhoods. In presence of income sorting, one expects that inequality within neighborhoods of small size to be, on average, smaller than inequality in neighborhoods of larger size. Consequently, the NI curve is expected to be increasing in the individual neighborhood size. The relevant null here is $H_0^2 : \ NI(\mathbf{y},d') = NI(\mathbf{y},d)$ for $d' > d$, to be tested against a restricted alternative. Rejecting both null hypotheses H_0^1 and H_0^2 gives statistical support for the existence of a neighborhood component in the urban income distribution.

It is also of interest to study the dynamics of neighborhood inequality across income distributions $\mathbf{y}_t$ and $\mathbf{y}_{t'}$. For a given size d of the individual neighborhood, the relevant

null hypothesis is $H_0^3(d): \ NI(\mathbf{y}_t, d) = NI(\mathbf{y}_{t'}, d)$ against an unrestricted one. A growth or decline in neighborhood inequality is robust when it involves a dominance in neighborhood inequality curves. In this case, the relevant null hypothesis is: $H_0^3: \ H_0^3(d) \ \forall d$ against an unrestricted one. As for H_0^1 and H_0^2, the null hypotheses are expressed in the form of equalities to stress that one is compelled to conclude in favor of increasing or decreasing neighborhood inequality only if there is strong evidence against the null hypothesis.

The acceptance regions for the null hypotheses H_0^1, H_0^2 and $H_0^3(d)$ can se constructed using the confidence bounds implied by the SE approximations provided above. Under the normality assumption, confidence bounds for the NI index based on individual neighborhoods of size d take the form $\hat{NI}(\mathbf{y}, d) \pm z_\alpha SE_d$, where $\hat{NI}$ is a consistent estimator of the NI index and z_α is the standard normal critical value for confidence level $1 - \alpha$ (for instance, 95%). To test H_0^3, it is sufficient to plot the confidence bounds of $NI(\mathbf{y}_{t'}, d) - NI(\mathbf{y}_t, d)$ against d and verify that the horizontal orthant lies homogeneously in the implied rejection region.

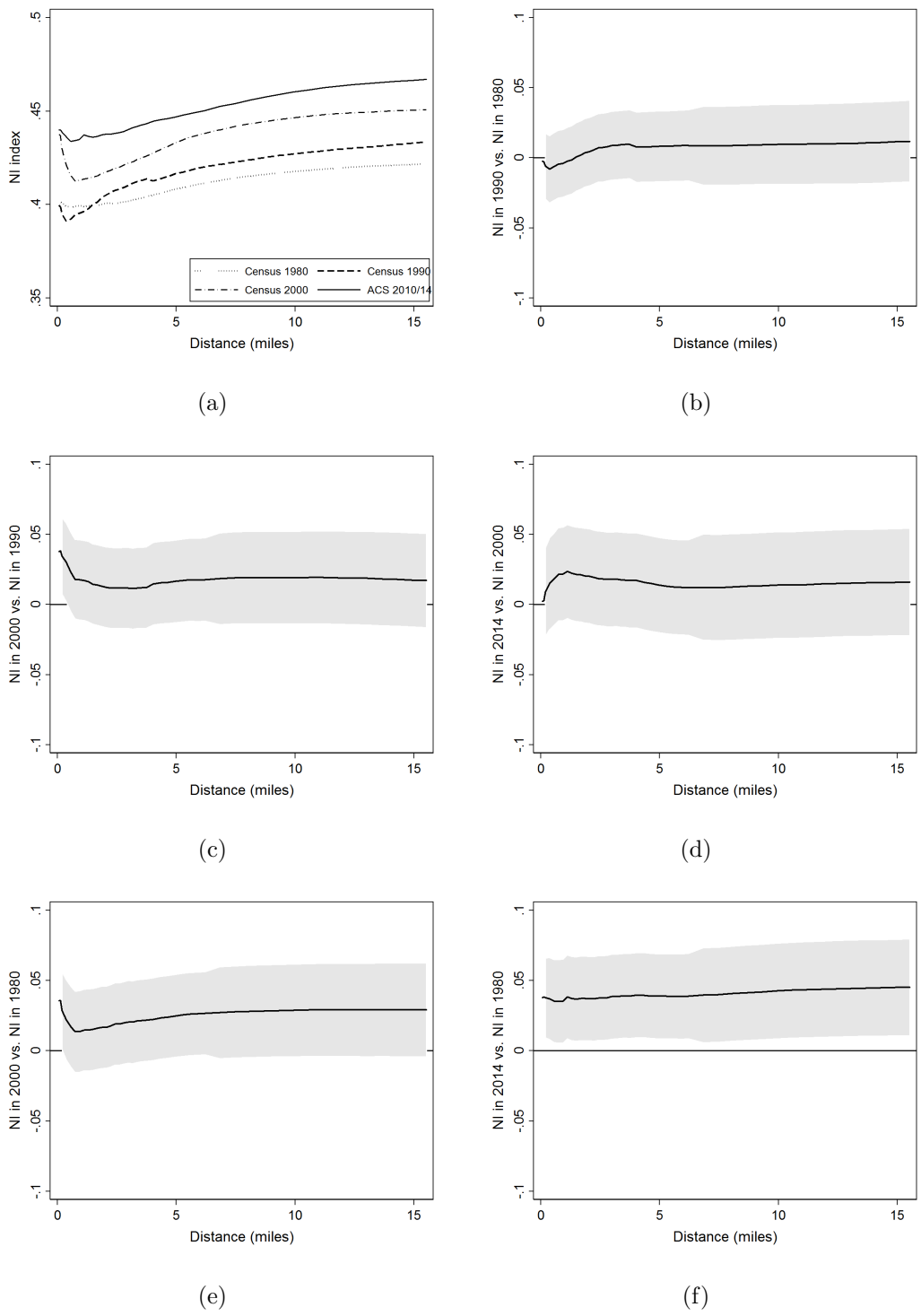
4 Inference for patterns and trends of neighborhood inequality in Chicago, IL, 1980-2014

Andreoli and Peluso (2018) provide robust evidence that neighborhood inequality is high in large American metro areas and almost converges to citywide income inequality, even when estimates are based on individual neighborhoods of small size (smaller than half a mile). They also find that neighborhood inequality has increased substantially over the last 35 years in virtually all largest cities. Are these patterns producing reliable evidence for the population? Is the growth in neighborhood inequality statistically significant?

I use the same data as in Andreoli and Peluso (2018) to draw inference about NI curves for the Metropolitan Statistical Area of Chicago, IL in the years 1980, 1990, 2000 and 2014.[4] Chicago has experienced large demographic growth over the last 35 years, with

[4]Data for 1980-200 are from the decennial US Cesus, STF 3A files. Data for 2014 are from the 2010-2014 sample estimates of the American Community Survey. In all cases, data on the spatial distribution of incomes in Chicago are reported in the form of tables, representative of the population living within

Figure 1: Trends in neighborhood inequality in Chicago, IL



Note: Author analysis of US Census and ACS data. Confidence intervals are at 95% level.

a given spatial partition. I use the census block group partition. For each block group in every year I obtain a synthetic equivalent gross household income distribution, reporting for each income level (estimated from a distribution model) a population weight. See Andreoli and Peluso (2018) for details about estimation of incomes and equivalence scales. I use these estimates as observations, each is assumed to be located at the block group's centroid. Distances can be recorded from georeferenced maps.

the number of inhabited census blocks (each gathering approximately 1000 households) increasing from 3756 in 1980 to more than 4700 in 2014. The growth in average equivalent income (in nominal terms), ranging from $13,794 in 1980 to $55,710 in 2014, has been followed by an expansion of inequality. The Gini index for the citywide income distribution has evolved steadily, from 0.434 in 1980 to 0.461 in 1990, then to 0.473 in 2000 and finally 0.486 in 2014, reflecting both demographic and economic changes.

Neighborhood inequality in Chicago mirrors the trends observed in other large American metro areas. As shown in panel a) of figure 1, in each year the NI index is high and close to the level of the citywide Gini index even in neighborhoods of relatively small size.[5] The NI estimates are always significant at all distance ranges. As table 1 shows (panel A), the hypothesis H_0^1 is rejected with p-values always close to zero when the individual neighborhood size is smaller than 5 miles. When the individual neighborhood is of 12 miles or above, neighborhood inequality is statistically indistinguishable from the level of inequality observed in the city at conventional levels of significance in 1980, 2000 and 2014. The same table, panel B, reports the evolution of the NI index at different distance thresholds compared to the level of neighborhood inequality in individual neighborhoods of size 0.4 miles. The gap in the NI index, in italics, is positive almost everywhere and always increasing with distance. Nonetheless, these differences are not statistically significant in a distance range smaller than 5 miles. At 12 miles, H_0^2 can be rejected in every year with p-values that are slightly larger than 5% (smaller in 1990). The patterns of p-values in the table confirm findings in Andreoli and Peluso (2018) that after 2000 the degree of neighborhood inequality registered in small neighborhoods has become more representative of the degree of inequality in the city, possibly reflecting the implications of the recent economic crises.

The trends of neighborhood inequality in Chicago resemble those observed in other large American metro areas. The year-to-year changes in NI, reported in panels b), c) and d) of figure 1, are always positive at every distance range. The magnitude of these changes is, however, too small to be statistically significant. Nonetheless, the cumulated change of

[5]The nature of the Census and ACS publicly accessible data does not allow to unbiasedly estimate NI in neighborhoods smaller than 0.3 miles. Confidence intervals are only reported for larger neighborhoods.

Years	Distance d in miles					
	0.4	1	2	3	5	12
Panel A: p-values for H_0^1						
1980	0.000	0.000	0.000	0.000	0.004	0.160
1990	0.000	0.000	0.000	0.000	0.000	0.003
2000	0.000	0.000	0.000	0.000	0.001	0.070
2014	0.000	0.000	0.000	0.000	0.002	0.108
Panel B: p-values for H_0^2						
1980	.	0.493	0.454	0.396	0.239	0.067
	0	*0.000*	*0.001*	*0.003*	*0.009*	*0.020*
1990	.	0.357	0.122	0.046	0.020	0.002
	0	*0.004*	*0.014*	*0.020*	*0.025*	*0.039*
2000	.	0.311	0.410	0.461	0.239	0.060
	0	*-0.008*	*-0.004*	*0.002*	*0.012*	*0.027*
2014	.	0.467	0.465	0.390	0.269	0.071
	0	*-0.001*	*0.002*	*0.005*	*0.011*	*0.027*

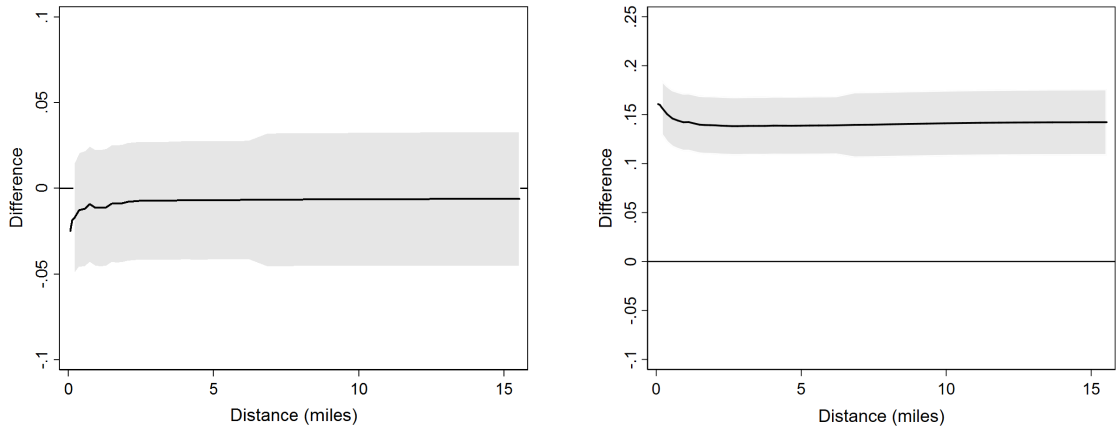
Table 1: P-values for null hypothesis of the type $H_0^1 : \ NI(\mathbf{y}_t, d) = G(\mathbf{y}_t)$ and $H_0^2 : \ NI(\mathbf{y}_t, d) = NI(\mathbf{y}_t, 0.4)$, with $t = 1980,\ 1990,\ 2000,\ 2014$ and $G(\mathbf{y}_{1980}) = 0.434$, $G(\mathbf{y}_{1990}) = 0.461$, $G(\mathbf{y}_{2000}) = 0.473$, $G(\mathbf{y}_{2014}) = 0.486$. Differences in levels of the NI index are in italic.

neighborhood inequality over the four decades turns out to be positive and significant at every distance range. As panel f) shows, the acceptance region for H_0^3 is always positive and never includes the horizontal axis, implying that the NI curve of Chicago for 2014 lies always above that of 1980 and the gap between the two is statistically different from (in fact, larger than) zero.

5 Monte Carlo study

The size and power properties of the estimators adopted to test dominance in NI curves are now assessed within the framework of a Monte Carlo study. The study reports simulated size and power for tests of differences of NI indices estimated at pre-determined distance cutoffs on samples of variable size n ($n = 1000,\ 2000,\ 5000,\ 8000$ observations), each draw from distinct known distributions. I calibrate the distributions to represent the actual and alternative counterfactual distributions of gross equivalent household income in Chicago, IL, in 2014. I first obtain these counterfactual distributions by applying suitable

Figure 2: Neighborhood inequality in Chicago, IL, 2014, versus two counterfactual distributions



(a) $NI(\mathcal{F}_0, d) - NI(\mathcal{F}_1, d)$ (b) $NI(\mathcal{F}_0, d) - NI(\mathcal{F}_2, d)$

Note: Author analysis of US Census and ACS data. Confidence intervals are at 95% level.

transformations to the actual ACS 2010-2014 module data, so that these distributions can be clearly ordered in terms of NI curves dominance. Then, I use moments of these population distributions to identify moments of the income data generating processes that I will use in the simulation study.

The first distribution $\mathcal{F}_0$ represents the spatial income distribution in Chicago, 2014, which I study in the previous section. This distribution has mean $\mu_0 = \$53,456$, standard deviation $\sigma_0 = \$55,310$ and spatial covariance structure $cov(s, v)$ across pairs of locations at distance h one from the other. I characterize the spatial covariance function, representing spatial dependence in the population model, through the variogram $\gamma_0(.)$, so that $cov(s, v) = \sigma_0^2 - \gamma_0(h)$. I fit the spherical model for the (semi)variogram function to the data, to obtain parameters $\alpha_0 = -327,203$, $\beta_0 = 21.14$ and range level $a = 10$ miles.

I produce two counterfactual population distributions $\mathcal{F}_1$ and $\mathcal{F}_2$ from the same data. The distribution $\mathcal{F}_1$ is obtained by adding noise to $\mathcal{F}_0$, so that $y_1 = y_0 + \varepsilon$ for $y_1 \sim \mathcal{F}_1$, $y_0 \sim \mathcal{F}_0$ and $\varepsilon \sim N(0, 6118.44)$, where the variance term of idiosyncratic disturbances is half a million time smaller than σ_0^2. This counterfactual distribution displays similar patterns of neighborhood inequality as $\mathcal{F}_0$. The null hypothesis H_0^3: $NI(\mathcal{F}_0, d) = NI(\mathcal{F}_1, d)$, $\forall d$

cannot be rejected, as shown in panel (a) of figure 2. This new population distribution has expectation $\mu_1 = \mu_0$, standard error $\sigma_1 = \$55,631 > \sigma_0$ and variogram $\gamma_1(.)$ with parameters $\alpha_1 = -69,660$ and $\beta_1 = 21.19$.

The second counterfactual distribution $\mathcal{F}_2$ is designed in a way that its NI curve lies always below that of $\mathcal{F}_0$. This distribution is obtained by simulating the effect of a redistributive linear income tax scheme applied to incomes distributed as $\mathcal{F}_0$. Andreoli and Peluso (2018) have demonstrated that only a basic income flat tax scheme guarantees that $\mathcal{F}_2$ dominates $\mathcal{F}_0$ in terms of NI curves. We hence use the transformation $y_2 = (1-t)y_0 + m$, for $y_2 \sim \mathcal{F}_2$, $y_0 \sim \mathcal{F}_0$, a flat tax rate $t = 0.3$ and basic income $m = 0.3\mu_0$. This counterfactual distribution displays different patterns of neighborhood inequality compared to $\mathcal{F}_0$. The null hypothesis H_0^3: $NI(\mathcal{F}_0, d) = NI(\mathcal{F}_1, d),\ \forall d$ is clearly rejected in favor of a restricted strong dominance alternative, as shown in panel (b) of figure 2. This new population distribution has expectation $\mu_2 = \mu_0$, standard error $\sigma_2 = \$38,716 < \sigma_0$ and variogram $\gamma_2(.)$ with parameters $\alpha_2 = -158,424.5$ and $\beta_2 = 20.43$.

The simulation study is based on models for the income process, denoted $\mathbf{Y}_f^n$ for $f = 0, 1, 2$, that replicate the population distributions $\mathcal{F}_0$, $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. As before, the income process is a collection of random variables indexed by n, a parameter controlled within the experiment, and defined over the random field $\mathcal{S}_n$. The first concern is to replicate the spatial structure of the data and construct a random field $\mathcal{S}_n$ that is representative of the map of Chicago in terms of distance scale and population density. To do so, I draw a random field $\mathcal{S}_n$ (reporting information about latitude and longitude as well as demographic weights of n locations) directly from the ACS 2010-2014 map of Chicago, by sampling n locations without replacement. This procedure should guarantee that the structure of ACS data for Chicago is always reflected in the outcomes of the simulation. These sampled locations are stored in a separate file for each n and used throughout the simulation experiment. Results will be conditional to the random field $\mathcal{S}_n$.[6]

My second concern is to model the spatial income process $\mathbf{Y}_f^n$ so that it represents

[6]The extracted coordinates of the random fields, alongside parameter estimates and replication code for this Monte Carlo study, are made available on the author web page.

income variability and spatial association underlying the population distributions $\mathcal{F}_f$. Given the random field $\mathcal{S}_n$, I maintain the assumption that the spatial income process satisfies intrinsic stationarity and the Gaussian hypothesis, implying that the process is fully characterized by known moments of the population distribution, so that $\mathbf{Y}_f^n \sim (\mu_f, \sigma_f^2, \gamma_f(.))$ for $f = 0, 1, 2$. The Monte Carlo experiment consists in randomly drawing realizations from $\mathbf{Y}_f^n$, each denoted $\mathbf{y}_{f,r}^n$ with $r = 1, \ldots, 200$, and assessing for each draw r if a certain null hypothesis about dominance in NI curves can or cannot be rejected, provided that the actual pattern of dominance in the populations is known. I use the SE approximations discussed in Section 3 to conclude about acceptance/rejections of the relevant null hypothesis. The decision outcome is registered with an indicator, which is then averaged across the 200 replicas to simulate size and power.

Each draw $\mathbf{y}_{f,r}^n$ from the spatial income process $\mathbf{Y}_f^n$, $f = 0, 1, 2$, should be representative of the degree of spatial association in the underlying population distribution $\mathcal{F}_f$. Coherently with previous assumptions, the spatial association between any pair of locations s, v on the random field $\mathcal{S}_n$ at geographic distance h (in miles) is provided by the covariance term $c_{s,v} = \sigma_f^2 - \gamma_f(h)$. The empirical estimates of the moments $(\mu_f, \sigma_f^2, \gamma_f(.))$ from the population distribution $\mathcal{F}_f$ identify the covariance matrix $\mathbf{C}_f$ of the spatial income process, with $\mathbf{C}_f = \{c_{s,v}\}$, $s = 1, \ldots, n$, $v = 1, \ldots, n$ and $c_{s,s} = \sigma_f^2$. I use decomposition methods to factorize the covariance matrix as $\mathbf{C}_f = \mathbf{D}_f \cdot \mathbf{D}'_f$, where $\mathbf{D}_f$ is a lower triangular matrix of size $n \times n$. This matrix conveys the information about the spatial association and variability in the population.[7] Each replica r of a distribution f (of size n) is then obtained as

$$\mathbf{y}_{f,r}^n = \mu_f \mathbf{e}_n \; + \; \mathbf{D}_f \cdot \boldsymbol{\nu}_r,$$

where $\mathbf{e}_n$ is a $n \times 1$ vector with all elements equal to one and $\boldsymbol{\nu}_r$ is a $n \times 1$ vector of standard normal distributed i.i.d. innovations. Throughout all replicas, values of the NI index and

[7] In many cases, the covariance matrix $\mathbf{C}_f$ is not positive semi-definite, implying that exact symmetric decompositions are not available. I use approximations based on the spectral theorem, as suggested in Bunch and Parlett (1971), to decompose $\mathbf{C}_f = \mathbf{X} \cdot diag(\mathbf{L}) \cdot \mathbf{X}'$, where $\mathbf{L}$ is a vector of eigenvalues and $\mathbf{X}$ collects the corresponding eigenvectors. I then set negative eigenvalues to zero to obtain $\mathbf{L}^*$ and produce an approximation of $\mathbf{C}_f$, denoted $\mathbf{C}_f^*$, such that $\mathbf{C}_f^* = \mathbf{X} \cdot \sqrt{diag(\mathbf{L}^*)} \cdot \sqrt{diag(\mathbf{L}^*)} \cdot \mathbf{X}'$. I then apply a Q-R decomposition of $\mathbf{C}_f^*$ to obtain $\mathbf{C}_f^* = \mathbf{R}'_f \cdot \mathbf{R}_f$ where $\mathbf{D}_f = \mathbf{R}'_f$ is a lower-triangular matrix.

n	Distance cutoffs (miles)									# Rej.	Rej.	Weak	Strong
	0.4	0.7	1	1.4	1.7	2	3	5	12				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Panel A : Size comparisons for the true null H_0^3 : $NI(\mathbf{y}_{0,r}^n, d) = NI(\mathbf{y}_{1,r}^n, d),\ \forall d$													
1000	.	0.00	.	0.24	.	0.19	0.16	0.03	0.01	1.1	0.46	0.60	0.00
2000	0.00	0.00	0.32	0.22	0.23	0.19	0.10	0.08	0.00	2.4	0.62	0.53	0.00
5000	0.00	0.00	0.33	0.15	0.17	0.09	0.01	0.01	0.00	1.1	0.52	0.47	0.00
8000	0.00	0.00	0.22	0.06	0.09	0.06	0.05	0.01	0.00	0.8	0.38	0.53	0.00
Panel B : Power comparisons for the true alternative H_a^3 : $NI(\mathbf{y}_{0,r}^n, d) \geq NI(\mathbf{y}_{2,r}^n, d),\ \forall d$													
1000	.	0.00	.	0.29	.	0.31	0.19	0.09	0.03	1.7	0.60	0.92	0.00
2000	0.00	0.00	0.40	0.31	0.44	0.26	0.13	0.08	0.00	3.3	0.85	0.88	0.00
5000	0.00	0.00	0.55	0.28	0.49	0.24	0.12	0.09	0.03	4.3	0.81	0.98	0.00
8000	0.00	0.05	0.57	0.34	0.53	0.32	0.30	0.22	0.15	8.7	0.82	0.99	0.00

Table 2: Monte Carlo simulations of the size and power of dominance tests for NI curves that are based on the NI index SE approximations.

of the SE approximations can be meaningfully computed only at some distance cutoffs. For samples of size $n = 2000,\ 5000,\ 8000$, distance cutoffs are set at approximately a third of a mile distance range increments within the first 7 miles, and then at about two third of a mile increments within the next 12 miles (at 19 miles range the NI index converges to citywide inequality). For the sample of size $n = 1000$, distance thresholds within 1 and 19 miles are set by looking at increments of three quarters of a mile exclusively. $H_0^3(d)$ is tested at each distance cutoff. The null hypothesis of the type H_0^3 is tested instead by looking at all distance cutoffs.

The first goal of this section is to infer about the *size* for the tests for various null hypothesis about NI curves. The size of the test corresponds to the share of simulated samples that allow to reject the relevant null hypothesis when the null hypothesis is true in the population. I use population distributions $\mathcal{F}_0$ and $\mathcal{F}_1$ as references, where it is known that H_0^3: $NI(\mathcal{F}_0, d) = NI(\mathcal{F}_1, d),\ \forall d$ is true. I draw replicas $\mathbf{y}_{0,r}^n$ and $\mathbf{y}_{1,r}^n$ from the underlying spatial models $\mathbf{Y}_0^n$ and $\mathbf{Y}_1^n$ for various levels of n. For each replica r I then test whether $H_0^3(d)$: $NI(\mathbf{y}_{0,r}^n, d) = NI(\mathbf{y}_{1,r}^n, d)$, as well as the implied null H_0^3, are rejected by the sample data. I record the rejections and store the average share of rejections over the 200 replicas in Panel A of table 2. Columns (1) to (9) report the size of test for null hypothesis $H_0^3(d)$ at well defined distance cutoffs. Column (10) reports the average number of rejections of $H_0^3(d)$ at available distance cutoffs across all replicas. Column (11)

reports the proportion of times that a null hypothesis $H_0^3(d)$ is rejected at least one. This figure likely estimates an upper bound for the size, insofar it is sufficient that there exist a d for which $H_0^3(d)$ is rejected to conclude that the null is rejected at stage r. Columns (12) and (13) report, respectively, the share of cases where the rejection entails a weak dominance in NI curves (i.e., all cases where multiple rejections of $H_0^3(d)$ occur within the same replica r and differences in NI curves have the same sign) and the proportion of the cases in (12) where dominance is strong (i.e., H_0^3 is rejected at every distance cutoff). The product of the coefficients in columns (13), (12) and (11) gives a good estimate of a lower bound for the size of the tests.

Overall, the tests based on the NI index SE bounds have larger size compared to the nominal 5% level. The size of tests carried out at fixed distance cutoffs is smaller than 10% when the sample size is at least of 5000 units, while it is much larger for samples of smaller size. The size of the test is virtually zero when the NI index estimates are based on individual neighborhoods of size smaller then 1 mile. This might reflect the consequences of imperfectly estimating the income distribution at the very local scale. A bit more expected is the fact that the size of the tests for $H_0^3(d)$ at distance ranges larger than 5 miles are below 5%. At these distance ranges, in fact, neighborhood inequality converges to the levels of citywide inequality measured by the Gini coefficient, and the SE approximation converges asymptotically (since the spatial association of incomes becomes negligible). There is on average less than 1 rejection of $H_0^3(d)$ across the distance cutoffs for which I test. The upper bound for the size is of 38% in the largest sample. The size of the test monotonically converges to this number as the sample size grows. A linear interpolation of size estimates in column (11) suggests that the upper bound for the size converges to its nominal value of 5% when the sample size is larger than 16,000 units. I also find that the number of rejections related to the weak form of dominance is about 50%, although no strong forms of dominance are registered.

The second goal of this section is to infer the *power* for the tests for various null hypothesis about NI curves. The power corresponds to the share of replicas that reject the relevant null hypothesis in favor of a specific alternative when the alternative is true

in the population. I use population distributions $\mathcal{F}_0$ and $\mathcal{F}_2$ as references, where it is known that H_0^3: $NI(\mathcal{F}_0, d) = NI(\mathcal{F}_1, d),\ \forall d$ is rejected in favor of (strong) dominance in NI curves. I draw replicas $\mathbf{y}_{0,r}^n$ and $\mathbf{y}_{2,r}^n$ from the underlying spatial models $\mathbf{Y}_0^n$ and $\mathbf{Y}_2^n$ for various n. For each replica r, I then test if $H_0^3(d)$: $NI(\mathbf{y}_{0,r}^n, d) = NI(\mathbf{y}_{2,r}^n, d)$ at each distance cutoff separately, as well as the implied null H_0^3, are rejected by the sample data. I find that the power of tests for $H_0^3(d)$ are relatively small for small and large distance cutoffs. Power estimates are instead always larger than 30% for distance cutoffs between 1 and 5 miles for which I test. I record larger power estimates for violations of H_0^3, with all cases displaying statistically significant differences in NI curves that are of the same sign. Tests for H_0^3 neglect the positive correlation between SE computed at different distance cutoffs, thus making rejections of the null hypothesis more likely (since part of the variability in NI curves estimates is neglected). Hence, rejections rates for H_0^3 in favor of (weak) dominance can only be interpreted as upper bounds for the power of the joint tests. I estimate these upper bounds by the product of the share of rejections (column (11)) times the proportions of rejections where weak dominance is detected (column (12)). The upper bound for the power of tests for H_0^3 is of 74.8% for samples of size 2000 units and grows to 81% in the largest samples. These power estimates are reasonable and, despite being only upper bound, support the validity of tests for NI curves dominance based on the SE approximations I propose even in relatively small samples. I also find that the average number of distance cutoffs where $H_0^3(d)$ is rejected at any given simulated sample grows steadily with the simulated sample size (column (10)), from 1.7 rejections when $n = 1000$ to 8.7 rejections on average when $n = 8000$, alongside larger chances that these rejections are in favor of a weak form of dominance in NI curves. Altogether, these figures confirm the relevance of the SE approximations for inferring about patterns and dynamics of neighborhood inequality.

6 Concluding remarks

This article provides variance bounds for the neighborhood inequality index proposed by Andreoli and Peluso (2018). These bounds are identified from the knowledge of the

variogram function which, under assumptions on the income generating process that are common in spatial statistics literature, fully characterizes the spatial income distribution.

An application to rich income data from the American Census and the Community Survey motivates the interest in using SE approximations for the NI index when assessing patterns and trends of neighborhood inequality across American cities. Focussing on the city of Chicago, IL, I find robust statistical evidence that neighborhood inequality is large even for individual neighborhoods of small size, but it is statistically different from city-wide inequality (as measured by the Gini index). The cumulated growth of neighborhood inequality over the period 1980-2014 is substantial and significant at standard confidence levels, reflecting a general trend in largest American cities documented in Andreoli and Peluso (2018). The Monte Carlo study shows that the tests for NI curves dominance based on the SE approximations I study have higher size than the nominal values, although the (upper bound) size estimates quickly converges when the sample size grows. I expect that a sample of 16,000 units, smaller than the sample used to obtain estimates on the 5-years ACS module, is sufficient to guarantee that the size of the test is consider converge to their nominal values. The power of these tests is relatively small for null hypotheses defined at given distance cutoffs (but larger than 30%), but power grows significantly to more than 80% when considering tests for NI curves (weak) dominance (although these are only upper bounds). Some of the null hypothesis I investigate require multiple testing, a factor I do not account for in the simulation exercise and that will be addressed elsewhere.

As a remark, the SE bounds for the NI index that I provide seem to be relevant for inferring about neighborhood inequality in samples of urban population of no less then 8000 individuals. Investigations about the appropriate testing procedure when placing dominance/non-dominance of NI curves under the null are also left for future research.

References

Andreoli, F. (2018). Robust inference for inverse stochastic dominance, *Journal of Business & Economic Statistics* **36**(1): 146–159.

Andreoli, F. and Peluso, E. (2018). So close yet so unequal: Neighborhood inequality in American cities, *ECINEQ Working paper 2018-477* .

Baum-Snow, N. and Pavan, R. (2013). Inequality and city size, *The Review of Economics and Statistics* **95**(5): 1535–1548.

Biondi, F. and Qeadan, F. (2008). Inequality in paleorecords, *Ecology* **89**(4): 1056–1067.

Bishop, J. A., Chakraborti, S. and Thistle, P. D. (1989). Asymptotically distribution-free statistical inference for generalized Lorenz curves, *The Review of Economics and Statistics* **71**(4): pp. 725–727.

Bunch, J. and Parlett, B. (1971). Direct methods for solving symmetric indefinite systems of linear equations, *SIAM Journal on Numerical Analysis* **8**(4): 639–655.

Chetty, R. and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects*, *The Quarterly Journal of Economics* **133**(3): 1107–1162.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A. and Cutler, D. (2016). The association between income and life expectancy in the United States, 2001-2014., *The Journal of the American Medical Association* **315**(14): 1750–1766.

Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, New York.

Clark, W. A. V., Anderson, E., Östh, J. and Malmberg, B. (2015). A multiscalar analysis of neighborhood composition in Los Angeles, 2000-2010: A location-based approach to segregation and diversity, *Annals of the Association of American Geographers* **105**(6): 1260–1284.

Conley, T. G. and Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment, *Journal of Applied Econometrics* **17**(4): 303–327.

Cressie, N. (1985). Fitting variogram models by weighted least squares, *Journal of the International Association for Mathematical Geology* **17**(5): 563–586.

Cressie, N. A. C. (1991). *Statistics for Spatal Data*, John Wiley & Sons, New York.

Cressie, N. and Hawkins, D. M. (1980). Robust estimation of the variogram: I, *Journal of the International Association for Mathematical Geology* **12**(2): 115–125.

Dardanoni, V. and Forcina, A. (1999). Inference for Lorenz curve orderings, *Econometrics Journal* **2**: 49–75.

Davidson, R. (2009). Reliable inference for the Gini index, *Journal of Econometrics* **150**(1): 30 – 40.

Dawkins, C. J. (2007). Space and the measurement of income segregation, *Journal of Regional Science* **47**: 255–272.

Galster, G. (2001). On the nature of neighbourhood, *Urban Studies* **38**(12): 2111–2124.

Goodman, L. A. and Hartley, H. O. (1958). The precision of unbiased ratio-type estimators, *Journal of the American Statistical Association* **53**(282): 491–508.

Hardman, A. and Ioannides, Y. (2004). Neighbors' incom distribution: Economic segregation and mixing in US urban neighborhoods, *Journal of Housing Economics* **13**(4): 368–382.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *The Annals of Mathematical Statistics* **19**(3): 293–325.

Iceland, J. and Hernandez, E. (2017). Understanding trends in concentrated poverty: 1980-2014, *Social Science Research* **62**: 75 – 95.

Jargowsky, P. A. (1997). *Poverty and Place: Ghettos, Barrios, and the American City*, New York: Russell Sage Foundation.

Kim, J. and Jargowsky, P. A. (2009). *The Gini-coefficient and segregation on a continuous variable*, Vol. Occupational and Residential Segregation of *Research on Economic Inequality*, Emerald Group Publishing Limited, pp. 57 – 70.

Koop, J. C. (1964). On an identity for the variances of a ratio of two random variables, *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(3): 484–486.

Leone, F. C., Nelson, L. S. and Nottingham, R. B. (1961). The folded normal distribution, *Technometrics* **3**(4): 543–550.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R. and Sanbonmatsu, L. (2012). Neighborhood effects on the long-term well-being of low-income adults, *Science* **337**(6101): 1505–1510.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R. and Sanbonmatsu, L. (2013). Long-term neighborhood effects on low-income families: Evidence from Moving to Opportunity, *American Economic Review* **103**(3): 226–31.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G. J., Katz, L. F., Kessler, R. C., Kling, J. R., Lindau, S. T., Whitaker, R. C. and McDade, T. W. (2011). Neighborhoods, obesity, and diabetes - A randomized social experiment, *New England Journal of Medicine* **365**(16): 1509–1519. PMID: 22010917.

Massey, D. S. and Eggers, M. L. (1990). The ecology of inequality: Minorities and the concentration of poverty, 1970-1980, *American Journal of Sociology* **95**(5): 1153–1188.

Matheron, G. (1963). Principles of geostatistics, *Economic Geology* **58**(8): 1246–1266.

Moretti, E. (2013). Real wage inequality, *American Economic Journal: Applied Economics* **5**(1): 65–103.

Muliere, P. and Scarsini, M. (1989). A note on stochastic dominance and inequality measures, *Journal of Economic Theory* **49**(2): 314 – 323.

Openshaw, S. (1983). *The modifiable areal unit problem*, Norwick: Geo Books.

Pyatt, G. (1976). On the interpretation and disaggregation of Gini coefficients, *The Economic Journal* **86**(342): 243–255.

Reardon, S. F. and Bischoff, K. (2011). Income inequality and income segregation, *American Journal of Sociology* **116**(4): 1092–1153.

Shorrocks, A. and Wan, G. (2005). Spatial decomposition of inequality, *Journal of Economic Geography* **5**(1): 59–81.

Tin, M. (1965). Comparison of some ratio estimators, *Journal of the American Statistical Association* **60**(309): 294–307.

Watson, T. (2009). Inequality and the measurement of residential segregation by income in American neighborhoods, *Review of Income and Wealth* **55**(3): 820–844.

Wheeler, C. H. and La Jeunesse, E. A. (2008). Trends in neighborhood income inequality in the U.S.: 1980–2000, *Journal of Regional Science* **48**(5): 879–891.

Wong, D. (2009). The modifiable areal unit problem (MAUP), *The SAGE handbook of spatial analysis* pp. 105–124.

Xu, K. (2007). U-statistics and their asymptotic results for some inequality and poverty measures, *Econometric Reviews* **26**(5): 567–577.

LUXEMBOURG INSTITUTE OF
SOCIO-ECONOMIC RESEARCH

11, Porte des Sciences
Campus Belval
L-4366 Esch-sur-Alzette

T +352 58 58 55-1
F +352 58 58 55-700

www.liser.lu