

# DEMOGRAPHIC RESEARCH

*A peer-reviewed, open-access journal of population sciences*

---

## **DEMOGRAPHIC RESEARCH**

**VOLUME 44, ARTICLE 49, PAGES 1165–1184**

**PUBLISHED 11 JUNE 2021**

<https://www.demographic-research.org/Volumes/Vol44/49/>

DOI: 10.4054/DemRes.2021.44.49

*Formal Relationships 34*

**Measuring the concentration of urban population in the negative exponential model using the Lorenz curve, Gini coefficient, Hoover dissimilarity index, and relative entropy**

**Joel E. Cohen**

This article is part of the Special Collection on “Formal Relationships.”  
Guest Editors are Joshua R. Goldstein and James W. Vaupel.

© 2021 Joel E. Cohen.

*This open-access work is published under the terms of the Creative Commons Attribution 3.0 Germany (CC BY 3.0 DE), which permits use, reproduction, and distribution in any medium, provided the original author(s) and source are given credit.*

*See <https://creativecommons.org/licenses/by/3.0/de/legalcode>.*

## Contents

1	Introduction	1166
2	The negative exponential model of urban population density	1169
3	Analysis	1170
3.1	Area	1170
3.2	Population	1171
3.3	Lorenz curve	1172
3.4	Gini coefficient $G$	1173
3.5	Hoover index $H$	1174
3.6	Relative entropy	1175
4	Numerical example: Chicago 1900	1177
5	Discussion and conclusions	1180
6	Acknowledgments	1182
	References	1183

# **Measuring the concentration of urban population in the negative exponential model using the Lorenz curve, Gini coefficient, Hoover dissimilarity index, and relative entropy**

**Joel E. Cohen<sup>1</sup>**

## **Abstract**

### **BACKGROUND**

Stewart (1947) and Clark (1951) proposed that urban population density is a negative exponential function of the distance from a city's center. This model of the spatial distribution of urban population density has been influential in urban economics, transportation planning, and urban demography. Duncan (1957) suggested characterizing the inequality in the distribution of urban population density in this model by using standard economic measures of concentration or unevenness: the Lorenz curve, the Gini coefficient, and the Hoover dissimilarity index. Batty (1974) advocated measuring concentration using relative entropy.

### **OBJECTIVE**

We execute Duncan's (1957) and Batty's (1974) suggestions using mathematical analysis, not simulations.

### **METHODS**

We modify the negative exponential model to recognize that any city has a finite radius.

### **RESULTS**

Mathematical analysis reveals that all four measures of concentration depend sensitively on the finite radius of the city in the negative exponential model. We give a numerical example of the sensitivity of the concentration measures to the boundary radius.

### **CONTRIBUTION**

In empirical applications of the negative exponential model of urban population density, it is important to have clear, consistent standards for defining urban boundaries. Otherwise, differences between cities or over time within the same city in these four and

---

<sup>1</sup> Laboratory of Populations, Rockefeller University, 1230 York Avenue, Box 20, New York, NY 10065-6399 USA; Earth Institute and Department of Statistics, Columbia University, New York, NY 10027 USA; Department of Statistics, University of Chicago, Chicago, IL 60637 USA. Email: [cohen@rockefeller.edu](mailto:cohen@rockefeller.edu). ORCID: 0000-0002-9746-6725.

perhaps other measures of concentration could be due at least in part to differences in defining the radius or other boundaries of the city.

## 1. Introduction

The purpose of this note is to analyze mathematically the concentration of the spatial distribution of people in the negative exponential model of urban population density (Stewart 1947; Clark 1951) using three measures suggested by Duncan (1957) (the Lorenz curve, the Gini coefficient, and the Hoover index) and the relative entropy suggested by Batty (1974). Calculating these measures numerically in any particular instance is straightforward but yields no general insight. Mathematical analysis of these measures of concentration in the negative exponential model may not have been reported previously and yields some unexpected insight.

Stewart (1947: 180) analyzed the 60 “leading cities” in the 1940 US Census for which the distribution of population by census tracts was available. He concluded, “There is strong evidence for the following standard internal pattern, as a first approximation: The normal city, regardless of size, has roughly the same density of population at its edges, averaging there about 3 people per acre or 2,000 per square mile [ $\approx 770/\text{km}^2$ ]. From edge to center, the density tends to increase exponentially with the distance, reaching a peak density in some inner census tract which usually is adjacent to others having densities nearly as great. The peak density tends to increase with the size of the city.”

Without reference to Stewart (1947), Clark (1951) proposed “two generalizations the validity of which is now universally recognized: 1. In every large city, excluding the central business zone, which has few resident inhabitants, we have districts of dense population in the interior, with density falling off progressively as we proceed to the outer suburbs. 2. In most (but not all) cities, as time goes on, density tends to fall in the most populous inner suburbs, and to rise in the outer suburbs, and the whole city tends to ‘spread itself out.’ The evidence assembled below appears to be sufficient to show that, in practically every case, the falling off of density, as we proceed to the outer suburbs, follows a simple mathematical equation of exponential decline.” Stewart (1953) published a generous report of Clark (1951).

Duncan (1957: 27) summarized “the major techniques of describing and measuring population distribution, indicating some unresolved problems of method that might well be the focus of further research.” He presented the Lorenz curve, the Gini “concentration ratio” or Gini coefficient  $G$ , and an “index of concentration” or “index of dissimilarity”  $H$ , sometimes called the dissimilarity index, Hoover index, Robin Hood index, or Schutz

index. After mentioning Clark (1951), Amos Hawley, and Donald J. Bogue, but not Stewart (1947, 1953), as prior authors who “have studied density gradients according to distance from centres,” Duncan (1957: 44) remarked, “It is also possible to generalize the Lorenz curve and concentration ratio technique for use in this connection.”

Clark (1951) has been cited at least 1,509 times, Duncan (1957) at least 145 times, and Stewart (1947) 107 times (Google Scholar, 2021-01-13). Clark (1951) has been much more influential than Stewart (1947), even though Stewart specified a crucial boundary condition (where does the city end?) that Clark neglected. The difference in impact could have at least three causes: Clark (1951), unlike Stewart (1947), included detailed graphics and statistical tables for multiple cities at multiple points in time; Clark published in a journal more likely to be seen by urban economists, geographers, and demographers; and Stewart cloaked his findings in the language of “social physics.”

It is virtually impossible to be sure that the mathematical analysis that Duncan (1957) suggested has not already been carried out, but there is no indication of such mathematical analysis in at least three classic reviews (Massey and Denton 1988; McDonald 1989; Smith 1997).

Focusing on measures of segregation of ethnic or cultural groups, Massey and Denton (1988) collected 20 quantitative indices and proposed to classify them as measuring five dimensions of the spatial variation of a population: unevenness, exposure, clustering, concentration, and centralization. Massey and Denton (1988) computed the values of the 20 indices for comparisons of the spatial distributions of the census-defined Hispanics, Blacks, and Asians (one group at a time) with the census-defined non-Hispanic white population in 60 Standard Metropolitan Statistical Areas (SMSAs; the US Census Bureau used this term from 1959 to 1983) in the 1980 US Census. The 60 SMSAs were the 50 with the largest populations plus 10 others with many census-defined Hispanics. To uncover the underlying principal dimensions of variation of the 20 indices, Massey and Denton (1988) did two kinds of principal components analysis: one with five orthogonal (uncorrelated) dimensions and one allowing correlations among the five principal dimensions.

Massey, White, and Phua (1996) executed a similar analysis on 58 of the 60 SMSAs that Massey and Denton studied in 1988 and then extended the analysis to 318 Metropolitan Statistical Areas (MSAs) in the 1990 US Census. Massey, White, and Phua (1996) reached similar but not identical conclusions: The first four principal components were more important and the fifth principal component played only a small role. Massey and Denton (1998) corrected an error of Massey and Denton (1988) while confirming the earlier conclusions. In Massey’s (2012) retrospective summary, “We obtained a factor pattern matrix that yielded a robust interpretation across different rotation and extraction methods, one that confirmed the conceptual structure we had hypothesized.”

Rugh and Massey (2014) and Massey and Tannen (2015) improved on previous analyses of Massey and colleagues by using commercially developed data sets with temporally consistent boundaries of 287 metropolitan areas from 1970 to 2010. Rugh and Massey (2014: 211) distinguished metropolitan inhabitants from “urban residents (living in census tracts with greater than 1,000 persons per square mile).” This threshold differs from Stewart’s (1947) urban boundary estimate for 1940 by a factor of 2. These analyses eliminated any contribution that changes in boundaries over time or place may have made to variation over time or place in measures of segregation, including some of the measures analyzed here.

Smith (1997) reviewed many theoretical models and empirical analyses of what she called “monocentric urban density,” not including any works by J.Q. Stewart. She surveyed models of the intensity of land use around a center from the 1826 agricultural economic model of Johann Heinrich von Thünen (1783–1850) to the models of Clark (1951), Newling (1969), Bussi ere and Snickars (1970), Batty (1974), and many others in the second half of the twentieth century. She reported, “During the decade following Clark’s [1951] assertion, researchers tested nearly one hundred cities in the developed and developing world, covering over 150 years. The results were consistent with Clark’s negative exponential density function” (Smith 1997: 118).

In Stewart’s (1947) and Clark’s (1951) negative exponential model, the logarithm of population density is a linearly decreasing function of distance from the city’s center. Newling (1969) modeled the logarithm of population density as a quadratic function of distance from the city’s center, with a reduced density at the center, a peak density at some intermediate distance from the center, and progressively decreasing density at further distances from the center. Following earlier suggestions of Clark (1951) and Newling (1969), Smith (1997: 119) suggested “a dynamic interpretation – the [Stewart–] Clark model reflects the North American city in its early stages and the Newling model portrays the city in a later developmental stage. The central density becomes less over time and the peak density shifts outward from the city center.”

Bertaud and Malpezzi (2014: 16) analyzed the spatial distribution of population in 57 mostly large, metropolitan areas from 1990 to 2009 in “rich countries (e.g., France, Germany, the United States), poor countries (e.g., Vietnam, Afghanistan, Ethiopia), and emerging markets (e.g., China, Brazil, Mexico) [27 countries in all]. All continents are represented: 21 in Asia, 17 in Europe, 10 in North America (including Mexico), 5 in Africa, and 4 in Latin America.” Bertaud and Malpezzi (2014) did not specify how they established the boundaries of cities. “The first important finding is that in many cities – perhaps a surprisingly large number, to some – the negative exponential density gradient implied by the standard urban model fits the data quite well. On the other hand, in a number of cities, population density departs a *lot* from the standard model” (Bertaud and Malpezzi 2014: 34; italics in original).

Bussi re and Snickars (1970) derived a negative exponential model of urban population density by maximizing the continuous entropy of the probability density function of population density subject to constraints of total population size and costs of traveling to the center of the city, which they assumed proportional to the radial distance to the center. Continuous entropy is highly problematic as a measure of concentration or inequality. Several problems with it will be mentioned when the preferable concept of relative entropy is defined below.

The four measures of population concentration analyzed here for the negative exponential model of urban population density are a small proportion of the measures that have been proposed and used, and the negative exponential model is only one of many monocentric density models. The scarcity of mathematical analysis of many of these measures and models leaves ample room for this note to illustrate by example how such mathematical analysis might proceed and what insight such analysis might yield.

This note measures the concentration of urban population density in the negative exponential model by calculating analytically the Lorenz curve, Gini coefficient  $G$ , and Hoover index  $H$ , as Duncan (1957) suggested, and the relative entropy, as Batty (1974) suggested. The model and these measures are defined below. A numerical example follows this analysis.

## 2. The negative exponential model of urban population density

We summarize the negative exponential model of urban population density using metric units of measurement. The model assumes a city with a single center. Let  $r$  be the radial distance in kilometers from the city's center and  $D(r)$  be the population density of resident population at radius  $r$ . Then, except possibly in the central business district,

$$(1) \quad D(r) = ae^{-br}, a > 0, b > 0, 0 \leq r \leq R \leq \infty.$$

The dimension of  $D(r)$  and  $a$  is number/km<sup>2</sup>, of  $b$  is 1/km, and of  $r$  is km. Thus  $br$  is a dimensionless pure number independent of the unit in which distance is measured. According to Clark (1951: 490–491), “That the falling off [of] density is an exponential function, as in the above equation, appears to be true for all times and all places studied, from 1801 to the present day, and from Los Angeles to Budapest.” Further studies did not support this claim of universality but did confirm that the model is widely useful (McDonald 1989; Bertaud and Malpezzi 2014).

McDonald (1989: 363) pointed out that (1) refers to *gross* population density, defined as the ratio of resident population to land area, including “land in all uses, not just residential use. Net density refers to population per unit land in residential use.”

McDonald (1989: 364, his reference [43], which is apparently unpublished gray literature by C. Kramer in 1955) reported “that the standard negative exponential function fits net density patterns better than it fits gross density patterns.” Our subsequent analysis will not distinguish gross from net density because it is applicable to both.

Clark interpreted  $a$  as the hypothetical maximal population density that would be observed at the center of the city if the central business district had no reduction in the exponential trend of the population density of the districts close to the center, and  $b$  as the rate of decrease in population density per unit of increasing distance from the center. Clark pointed out that the exponential formula (1) allows neither for the reduction of residences in the city’s central business zone nor for water bodies, parks, and other areas not available for residences.

No city has infinite radius. Beyond some finite radius  $R$ ,  $0 < R < \infty$ , it is not plausible to consider the population as connected to the central city, as Stewart (1947) noted. All of Clark’s (1951) empirical examples had finite radii determined by some standard Clark did not specify. Clark (1951) did not formally recognize this constraint of finite size. Newling (1969: 245, 249) defined the radius of an urbanized area as the distance from the center at which the population density (in his quadratic exponential model) falls to some specified minimum value. Batty (1974: 7) introduced a maximal radius  $R$  into the negative exponential model and determined  $R$  as the radius required to reduce the difference between the entropy of a hypothetical city of infinite radius and the entropy of a city of radius  $R$  (with both such cities obeying the negative exponential distribution of population density with the same central density  $a$  and same exponent  $b$ ) to an arbitrarily prespecified threshold such as 5%. Our analysis follows Stewart (1947), Newling (1969), and Batty (1974) in assuming a finite boundary radius  $R$ . We focus on the sensitivity to  $R$  of four measures of urban population concentration.

### 3. Analysis

The analysis of the negative exponential model (1) proceeds in three steps, dealing with area, population, and measures of concentration. The analysis assumes that the outer boundary of the city lies at some finite radius  $R$  from the central city.

#### 3.1 Area

A circle at exactly radius  $x$  from the city center has circumference  $2\pi x$ . Hence the cumulative area  $A(r)$  (in square kilometers) within radius  $r$  (in kilometers) from the city center is



$$(2) \quad A(r) = \int_0^r (2\pi x) dx = \pi r^2, \quad 0 \leq r \leq R < \infty.$$

So the fraction  $F_A(r)$  of the city's total area  $A(R) = \pi R^2$  within distance  $r$  from the center,  $0 \leq r \leq R$ , or equivalently the cumulative distribution function (cdf) of area  $F_A(r)$  within each radius  $r$ ,  $0 \leq r \leq R$ , from the city's center is

$$(3) \quad F_A(r) = \frac{\pi r^2}{\pi R^2} = \frac{r^2}{R^2}, \quad 0 \leq r \leq R, \quad 0 \leq F_A(r) \leq 1, F_A(0) = 0, F_A(R) = 1.$$

$F_A(r)$  does not depend on the scale on which distance is measured. That is,  $F_A(r)$  does not change if  $r$  and  $R$  are measured in miles, inches, or any other constant multiple of kilometers.

The probability density function of area at distance  $r$  from the center is

$$(4) \quad f_A(r) = \frac{dF_A(r)}{dr} = \frac{2r}{R^2}, \quad 0 \leq r \leq R, \text{ with } \int_0^R f_A(r) dr = \int_0^R \frac{2r}{R^2} dr = 1.$$

The distance  $r_A$  from the city center that includes any cumulative fraction

$$(5) \quad p_A := F_A(r_A), \quad 0 \leq p_A \leq 1,$$

of the total urban area is given by the inverse function of  $F_A$ , which solves  $p_A = r^2/R^2$  for  $r$ , namely

$$(6) \quad r_A := F_A^{-1}(p_A) = R\sqrt{p_A} = R\sqrt{F_A(r_A)}.$$

### 3.2 Population

Clark (1951) observed that the exponential model of density (1) implies that the cumulative population  $P(r)$  (in thousands) within radius  $r$  (in kilometers) from the city center is

$$(7) \quad P(r) = \int_0^r D(x) \frac{dA(x)}{dx} dx = \int_0^r a e^{-bx} (2\pi x) dx = \frac{2\pi a}{b^2} \{1 - e^{-br} (1 + br)\}.$$

As the city boundary recedes to infinity, Clark (1951) observed that a city's total population approaches a finite limit,

$$(8) \quad P(\infty) = \lim_{R \rightarrow \infty} P(R) = \frac{2\pi a}{b^2}.$$

$P(r)$  and  $a/b^2$  are dimensionless pure numbers. Because  $1 + x < e^x$  for any  $x > 0$ , we have  $0 < P(r) < P(R) < \infty$  for any  $a > 0, b > 0, 0 < r < R \leq \infty$ .

If the city has finite radius  $R < \infty$ , then the total population  $P(R)$  of the city is, from (7):

$$(9) \quad P(R) = \frac{2\pi a}{b^2} \{1 - e^{-bR}(1 + bR)\}.$$

So the cumulative fraction  $F_P(r)$  of the city's population within radial distance  $r$ ,  $0 \leq r \leq R < \infty$ , from its center, or equivalently the cdf of population  $F_P(r)$  within radius  $r$ ,  $0 \leq r \leq R < \infty$ , from the city's center is

$$(10) \quad F_P(r) = \frac{\frac{2\pi a}{b^2}\{1 - e^{-br}(1 + br)\}}{\frac{2\pi a}{b^2}\{1 - e^{-bR}(1 + bR)\}} = \frac{1 - e^{-br}(1 + br)}{1 - e^{-bR}(1 + bR)}, \quad 0 \leq r \leq R < \infty, \quad 0 \leq F_P(r) \leq 1,$$

with probability density function of the population (number of people, not the population density) at radius  $r$  equal to

$$(11) \quad f_P(r) = \frac{dF_P(r)}{dr} = \frac{b^2 r e^{-br}}{1 - e^{-bR}(1 + bR)}, \quad 0 \leq r \leq R < \infty, \quad \int_0^R f_P(r) dr = 1.$$

$F_P(r)$  does not depend on the scale on which distance is measured, but  $f_P(r)$  has the dimension of  $b$ , namely, 1/kilometers (in general, 1/distance).

### 3.3 Lorenz curve

The Lorenz curve, being a function and not a single number, is not among the 20 indices studied by Massey and Denton (1988). The scalar indices  $G$  and  $H$  can be calculated from the Lorenz curve. The Gini coefficient  $G$  is the second "measure of evenness" in the classification of Massey and Denton (1988: 285). The Hoover index  $H$  as defined by Duncan (1957: 30, denoted  $\Delta$ ) equals the "dissimilarity index"  $D$  of Massey and Denton (1988: 284, their (1)) when  $D$  is applied to comparing a population's distribution with the distribution of land area.  $D$  is their first and preferred "measure of evenness" (Massey and Denton 1988: 308). The Hoover index  $H$  (Duncan's  $\Delta$ ) is identical to the index of spatial concentration DEL of Massey and Denton (1988: 289, their (10)). Hence two measures analyzed here,  $G$  and  $H$ , correspond to three measures analyzed by Massey and Denton (1988); Massey, White, and Phua (1996); Massey and Denton (1998); Rugh and Massey (2014); Massey and Tannen (2015); and many other students of population concentration and segregation.

Both cdfs of area  $F_A$  and of population  $F_P$  are functions of the radius  $r$ . Both are independent of the units in which distance is measured. The Lorenz curve  $L(p_A)$  expresses the cdf  $F_P$  of population, not as a function of  $r$  directly but indirectly as a function of the cdf of area  $p_A := F_A(r_A)$ ,  $0 \leq p_A \leq 1$ , without explicit reference to the radius  $r$ . Given any fraction  $p_A$ ,  $0 \leq p_A \leq 1$ , we find the corresponding  $r_A = R\sqrt{p_A}$  from (6), compute the corresponding value of  $F_P$  evaluated at  $r_A$  from (10), and set the Lorenz curve  $L(p_A)$  to equal

$$(12) \quad L(p_A) := F_P(r_A) = \frac{1 - e^{-bR\sqrt{p_A}(1+bR\sqrt{p_A})}}{1 - e^{-bR(1+bR)}}, \quad 0 \leq p_A \leq 1, \quad 0 \leq L(p_A) \leq 1.$$

The Lorenz curve is dimensionless, i.e., independent of the units in which distance is measured, and satisfies  $L(0) = 0, L(1) = 1$ . Because the negative exponential model assumes that population density decreases from the center outward as the cumulative area increases, the highest population densities come first with increasing cumulative area. Therefore the Lorenz curve lies above the line of equality that would hold if population density were constant at every radius  $r$  up to  $R$ . This location of the Lorenz curve above the diagonal line is common in urban transportation studies. In the hypothetical case of constant density, the cumulative population would be directly proportional to the cumulative area starting from the center and the Lorenz curve would be a diagonal straight line of slope 1 passing through the origin.

An unexpected result of this analysis is that even though  $L(0) = 0$ ,

$$(13) \quad \lim_{bR \rightarrow \infty} L(p_A) = 1 \text{ for every } 0 < p_A \leq 1 \text{ and for every } 0 < b < \infty.$$

This limit arises because  $\min(b, R) > 0$  and for any  $p > 0$ ,  $\lim_{x \rightarrow \infty} e^{-xp}(1 + xp) = 0$ . Hence the numerator and denominator of the right side of (12) both approach 1. Thus for a very rapid decline in density from the center (large  $b$ ) or for a very large radius of the city (large  $R$ ) or both, the Lorenz curve approaches 1 for any positive proportion of area  $p_A > 0$  (but  $L(0) = 0$ ).

### 3.4 Gini coefficient $G$

The Gini coefficient  $G$  is defined as

$$(14) \quad G := \frac{\int_0^1 |L(p_A) - p_A| dp_A}{\frac{1}{2}}.$$

Because  $F_p(r) \geq F_A(r)$ ,  $0 \leq r \leq R$ , and hence  $L(p_A) \geq p_A$ ,  $0 \leq p_A \leq 1$ , we have  $|L(p_A) - p_A| = L(p_A) - p_A$ . (If population were described in areas of *increasing* density, the Lorenz curve would fall *below* the diagonal line of constant density, and then we would have  $|L(p_A) - p_A| = p_A - L(p_A)$ .) Thus

$$\begin{aligned}
 G &= 2 \int_{p_A=0}^1 (L(p_A) - p_A) dp_A = 2 \int_0^1 L(p_A) dp_A - 2 \int_0^1 p_A dp_A \\
 &= \frac{2}{1 - e^{-bR}(1 + bR)} \int_0^1 [1 - e^{-bR\sqrt{p_A}}(1 + bR\sqrt{p_A})] dp_A - 1 \\
 (15) \qquad &= 2 \frac{6(bR - e^{+bR} + 1) + b^2 R^2 (e^{+bR} + 2)}{b^2 R^2 (e^{+bR} - bR - 1)} - 1.
 \end{aligned}$$

$G$  is a dimensionless number between 0 and 1, inclusive. For fixed  $R > 0$ ,

$$(16) \qquad \lim_{b \rightarrow 0} G = 0.$$

When  $b \rightarrow 0$ , population density becomes constant at every radius and  $G$  shows that the distribution of population density is as even as possible. Since  $\min(b, R) > 0$  and the dominant term in both the numerator and denominator in the fraction in (15) is  $b^2 R^2 e^{+bR}$ , it follows that

$$(17) \qquad \lim_{bR \rightarrow \infty} G = 1.$$

This unexpected finding shows that, for fixed  $b > 0$ ,  $G$  depends on (in fact, increases with) the outer radius  $R$  of the city.  $G$  is not invariant with respect to where the city is defined to end, even when the parameters  $a, b$  remain constant.

### 3.5 Hoover index $H$

The Hoover index  $H$  is defined as

$$(18) \quad H := \sup\{F_p(r) - F_A(r) : 0 \leq r \leq R\} = \sup\{L(p_A) - p_A : 0 \leq p_A \leq 1\}.$$

$H$  is a dimensionless number between 0 and 1, inclusive. The supremum is attained when the slope of  $L(p_A)$  equals the slope of  $p_A$  as a function of  $p_A$ , namely, when

$$(19) \quad L'(p_A) := \frac{dL(p_A)}{dt} = 1.$$

Differentiating  $L(p_A)$  from (12) gives condition (19) explicitly as

$$(20) \quad L'(p_A) = \frac{b^2 R^2 \exp(-bR\sqrt{p_A})}{2(1-\exp(-bR)(bR+1))} = 1.$$

Solving this condition for  $p_A$  gives  $p_A^*$  as the value of  $p_A$  at which  $L(p_A) - p_A$  is maximal:

$$(21) \quad p_A^* = \left( \frac{1}{bR} \log \left( \frac{2(1-e^{-bR}(bR+1))}{(bR)^2} \right) \right)^2.$$

Then

$$(22) \quad H = L(p_A^*) - p_A^*.$$

To obtain an explicit formula for  $H$ , we substitute (12) for  $L(p_A)$  in (22) and then replace  $p_A$  by  $p_A^*$  from (21). The resulting formula is opaque and not worth reproducing here.

However, the limiting behavior of  $H$  as  $bR \rightarrow \infty$  is easy to determine. As  $bR \rightarrow \infty$ , in (21) we have  $1/(bR) \rightarrow 0$  and  $2(1 - e^{-bR}(bR + 1))(bR)^{-2} \rightarrow 0$  and  $p_A^* \rightarrow 0$  (since  $\lim_{x \downarrow 0} x \log x = 0$ ). For any  $p_A^* > 0$ , (13) implies that  $L(p_A^*) \rightarrow 1$ . Thus as  $bR \rightarrow \infty$ , (22) implies that

$$(23) \quad \lim_{bR \rightarrow \infty} H = 1,$$

even though  $L(p) - p = 0$  when  $p = 0$ . This unexpected finding shows that for fixed  $b > 0$ , the Hoover index  $H$  depends on the outer radius  $R$  of the city and approaches 1 regardless of  $b > 0$  as  $R$  gets large. Like the Lorenz curve and the Gini coefficient, the Hoover index depends on where the city is defined to end.

### 3.6 Relative entropy

Batty (1974) proposed several measures, related to information-theoretic entropy, of the evenness of the spatial distribution of urban population. The continuous entropy function (Bussières and Snickars 1970: 297, eq. (18); Batty 1974: 50, eq. (19)) is problematic

conceptually because it is not invariant under change of coordinates, may fail to be nonnegative, and lacks other desired attributes of a measure of evenness (Marsh 2013: 16). The relative entropy, also known as the Kullback–Leibler divergence (Batty 1974: 42–43, eqs. (2)–(4)), avoids these problems. It measures how one probability density function differs from a second probability density function. It is not a metric on the space of probability density functions because it is not symmetric and does not obey the triangle inequality.

Here we calculate the (continuous) relative entropy  $D_{KL}(P||A)$  of the probability density of population (absolute numbers of people, not population density per unit of area) with respect to the probability density of area in the negative exponential model with a finite urban radius. Obviously, zero areas contain zero people (or, in mathematical jargon, the distribution of people is absolutely continuous with respect to the distribution of area). If  $D_{KL}(P||A) = 0$ , then equal fractions of the urban area have equal fractions of the urban population, or population is proportional to area. That is, population density is constant everywhere in the city. The more concentrated people are with respect to area, or the less evenly people are distributed in space, the larger the relative entropy of population with respect to area.

The probability density function  $f_P(r)$  of population at radius  $r \in [0, R]$  from the center is given by (11). The probability density function  $f_A(r)$  of area at radius  $r \in [0, R]$  from the center is given by (4). The relative entropy of population with respect to area is

$$\begin{aligned}
 (24) \quad D_{KL}(P||A) &:= \int_0^R f_P(r) \log\left(\frac{f_P(r)}{f_A(r)}\right) dr = \int_0^R \frac{b^2 r e^{b(R-r)}}{e^{bR} - (1+bR)} \log\left(\frac{\frac{b^2 r e^{b(R-r)}}{e^{bR} - (1+bR)}}{\frac{2r}{R^2}}\right) dr \\
 &= \frac{b^2 e^{bR}}{e^{bR} - (1+bR)} \int_0^R r e^{-br} \left[ \log\left(\frac{b^2 r e^{b(R-r)}}{2r} \cdot \frac{R^2}{e^{bR} - (1+bR)}\right) \right] dr \\
 &= \frac{b^2 e^{bR}}{e^{bR} - (1+bR)} \int_0^R r e^{-br} \left[ b(R-r) + \log\left(\frac{(bR)^2 e^{bR}}{2[e^{bR} - (1+bR)]}\right) \right] dr \\
 &= \frac{b^2 e^{bR}}{e^{bR} - (1+bR)} \left[ \frac{2(e^{-Rb} - 1) + Rb(e^{-Rb} + 1)}{b^2} + \log\left(\frac{(bR)^2}{2[1 - e^{-bR}(1+bR)]}\right) \right] \\
 &= \frac{(2(e^{-Rb} - 1) + bR(e^{-Rb} + 1))}{1 - e^{-bR}(1+bR)} \\
 &\quad + \frac{b^2}{1 - e^{-bR}(1+bR)} (2 \log(bR) - \log 2 - \log[1 - e^{-bR}(1+bR)]).
 \end{aligned}$$

As  $bR \rightarrow \infty$ , the first term on the right is asymptotic to  $bR$  and the second term is asymptotic to  $2b^2 \log(bR)$ , both of which go to infinity as  $bR \rightarrow \infty$ . Hence

$$(25) \quad \lim_{bR \rightarrow \infty} D_{KL}(P||A) = \infty.$$

For fixed  $b > 0$ , the relative entropy  $D_{KL}(P||A)$  depends on the outer radius  $R$  of the city and approaches  $\infty$  regardless of  $b > 0$  as  $R \rightarrow \infty$ . As with the Lorenz curve, Gini coefficient, and Hoover index, the relative entropy depends on where the city is defined to end.

#### 4. Numerical example: Chicago 1900

This example illustrates the mathematics. It is not intended as an adequate empirical analysis. For Chicago in 1900, Clark (1951: 492, 494) estimated  $a = 110$  thousand people per square mile and  $b = 0.45$  per mile. He did not state a value of  $R$ . His graph of population density as a function of radius ended around 7.5 miles from the city center. The population density (people per square mile) at the boundary  $R = 7.5$  miles from the center implied by the model (1) with these parameter values is  $D(7.5) = 110 \exp(-0.45 \times 7.5) = 3.76$  thousand people per square mile. This crude estimate differs from Stewart's (1947) estimated population density at the boundary of US cities in 1940 (namely, 2 thousand people per square mile) by less than a factor of 2. Clark's parameter estimates for Chicago in 1940 ( $a = 120, b = 0.3, R = 17$ ) imply that  $D(17) = 0.73$  thousand people per square mile, not far from the urban population density threshold of one thousand people per square mile used by Rugh and Massey (2014: 211).

Without converting to metric measurements, we use the numerical values  $a = 110$ ,  $b = 0.45$ , and, for simplicity,  $R = 10$  for the following illustrations.

**Figure 1:** Measures of population concentration in the negative exponential model, based on Clark's (1951) estimates for Chicago in 1900. See text for detailed explanations.

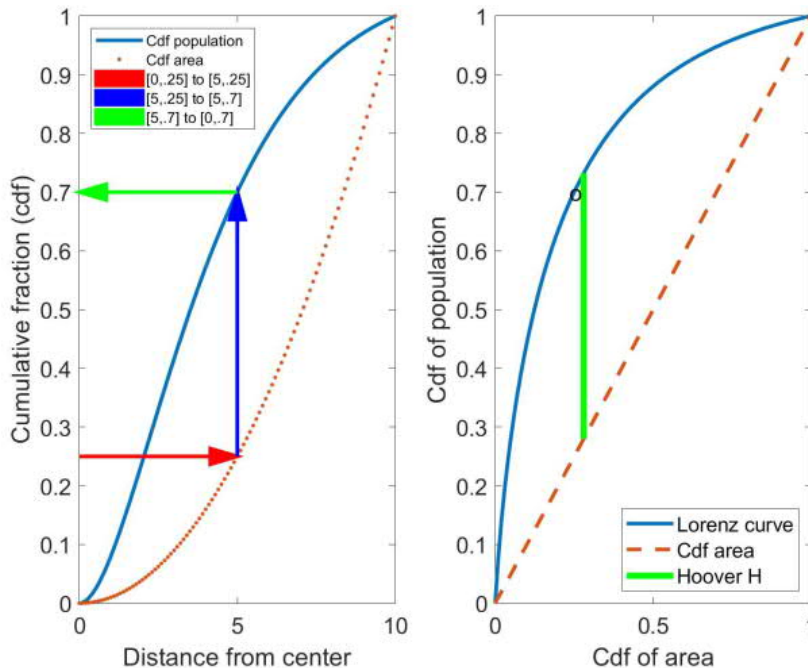


Figure 1(a, left) plots, on the vertical axis, the cdf of area  $A$  (red dots) and the cdf of population  $P$  (blue solid line) according to the negative exponential model at 101 equally spaced distances from 0 to 10 (horizontal axis). Figure 1(b, right) plots the Lorenz curve (blue solid line) on the vertical axis as a function of the cdf of area. In Figure 1(b), the dashed straight red line is the diagonal with slope 1 through the origin. Its height equals the cdf of area.

The three arrows in Figure 1(a) illustrate the construction of one point of the Lorenz curve in Figure 1(b), the small black oval with coordinates (cdf of area = 0.25, cdf of population = 0.7). Starting from the cumulative fraction of area = 0.25, the red arrow in the lower left of Figure 1(a) travels to the right to intersect the distance from the center where the cdf of area equals 0.25, which is exactly 5 miles from the center. This means that one-quarter of the city's total land area falls within a 5-mile radius from the center.



Then the blue arrow in the middle of Figure 1(a) travels straight upward to intersect the cdf of population at distance 5 miles from the center. Then the green arrow travels horizontally to the left to intersect the vertical axis at the cumulative fraction of population, approximately 0.70, within 5 miles from the center. The point (0.25, 0.70) on the Lorenz curve in Figure 1(b) is the illustrative point in the small black oval. It expresses the cdf of population as a function of the cdf of area.

The Hoover index  $H$  is the length of the thick vertical green bar in Figure 1(b). It is the maximal distance between the Lorenz curve and the diagonal line, or equivalently the maximal distance between the cdf of population and the cdf of area. It is located where the radius is approximately 5.4 miles from the center, the cdf of area is approximately 0.29, and the cdf of population is approximately 0.74, so that  $H \approx 0.74 - 0.29 = 0.45$ .

The Gini coefficient  $G$  equals the quotient of the area between the Lorenz curve and the diagonal line divided by the area above the diagonal line in Figure 1(b). Because the area above the diagonal line in Figure 1(b) must be 1/2 (because the diagonal line divides the unit square in half),  $G$  is also twice the area between the Lorenz curve and the diagonal line. In this example  $G \approx 0.58$ . These values satisfy the inequalities (Duncan 1957: 31)  $H \approx 0.45 \leq G \approx 0.58 \leq 2H - H^2 \approx 0.7$ .

In the negative exponential model, the Lorenz curve, the Gini coefficient  $G$ , the Hoover index  $H$ , and the relative entropy  $D_{KL}(P||A)$  are sensitive to the boundary radius  $R$ . For example, in Table 1, with  $b = 0.45$ , as  $R$  increases from 5 to 25,  $G$  almost triples,  $H$  more than triples, and the relative entropy  $D_{KL}(P||A)$  increases by a factor of more than 13.

**Table 1: Hypothetical boundary radii  $R = 5, 10, 15, \dots$  miles in the negative exponential model of population density in Chicago in 1900 using Clark's (1951) parameter estimates  $a = 110$  thousand people per square mile and decay exponent  $b = 0.45$ , boundary density  $D(R)$  thousand people per square mile, Gini coefficient  $G$ , Hoover index  $H$ , and relative entropy  $D_{KL}(P||A)$ . The measures of population concentration depend sensitively on the boundary radius  $R$  of the city.**

Radius $R$	Boundary density $D(R)$	Gini $G$	Hoover $H$	Relative entropy $D_{KL}(P  A)$
5	11.59	0.31	0.23	0.16
10	1.22	0.58	0.45	0.62
15	0.13	0.76	0.61	1.19
20	0.01	0.85	0.72	1.71
25	1.43E-03	0.91	0.78	2.15
30	1.51E-04	0.93	0.83	2.51
$\infty$	0	1	1	$\infty$

## 5. Discussion and conclusions

We analyzed mathematically four measures of urban population concentration in the negative exponential model of Stewart (1947) and Clark (1951), amended to recognize that any city has a finite radius. This analysis revealed that the Lorenz curve, Gini coefficient, Hoover dissimilarity index, and relative entropy of population with respect to area depend sensitively on the finite radius of the city. For any positive area, the first three measures approach 1 and the relative entropy diverges to infinity as the city's radius increases. These examples raise the disquieting possibility that, more generally, differences among cities in time or space in these, and perhaps other, measures of concentration could be due at least in part to differences in defining the radius or boundary of the city. Hence in empirical applications of these indices of population concentration, and perhaps in empirical applications of other measures that are sensitive to the radius or boundary of the city, it is important to have clear and consistent standards for defining urban boundaries.

This observation is related to the modifiable areal unit problem (MAUP) (Gehlke and Biehl 1934; Buzzelli 2020; Ye and Rogerson 2021). For example, one way to define urban boundaries, following Stewart (1947) and Newling (1969), is to estimate empirically a minimum population density at existing urban boundaries, where the boundaries are arrived at politically or otherwise. But estimates of population density depend on the size of spatial units used to estimate density, such as census blocks, census tracts, zip codes, counties, or rectangular grid cells of dimension  $100\text{ m} \times 100\text{ m}$  or  $1\text{ km} \times 1\text{ km}$ .

An alternative measure of concentration when the negative exponential model describes a city's spatial distribution of population density well is the value of the model's parameter  $b$ , which describes the rate of decay of population density with increasing distance from the center. If  $b = 0$ , population density is independent of distance from the center and is as evenly distributed as possible. The greater the positive value of  $b$ , the more rapidly population density falls with increasing distance from the center and the more concentrated the population density is at the center. The reciprocal  $1/b$  is the increase in distance from the center required to multiply population density by a factor of  $1/e \approx 0.3679$ . For example, with Clark's (1951) estimate  $b = 0.45/\text{mile}$  for Chicago in 1900, an increase in the radial distance from the center by  $1/b = 2.22$  miles is associated according to (1) with a 63% reduction in population density ( $0.63 \approx 1 - 1/e$ ). But estimates of  $b$ , like estimates of boundary densities, depend on the size of spatial units used to estimate density. In the extreme case, if the whole city, however defined, is treated as a single spatial unit, then the population density (total population divided by total area) is constant, regardless of distance from the city's putative center. Neither of

the two above approaches to estimating urban population concentration escapes the MAUP.

Massey and Denton (1988: 292) recognized that “the boundaries of a central city are political rather than natural creations. Central cities that were founded early have long been ringed by incorporated suburbs, while many newer cities continue to expand through incorporation.” From 1977 to 1981, the SMSAs studied by Massey and Denton (1988) were defined by the then Office of Federal Statistical Policy and Standards in the Department of Commerce, which had overall responsibility for federal statistical policy (US Bureau of the Census 1994, Ch. 13: 11). The Census Bureau’s definitions of “urban areas” and “urban” and methods of delineating boundaries have varied over time (US Bureau of the Census 1994, Ch. 12).

Massey and Denton (1988) did not investigate the consequences of the methods of delineating boundaries of their 60 cities for their 20 indices of segregation. In attempting to replicate results from the 1980 census using 1990 census data on 58 MSAs that were among the 60 SMSAs Massey and Denton (1988) analyzed from the 1980 census, Massey, White, and Phua (1996: 177) observed, “Because census tract boundaries and metropolitan area definitions inevitably change from census to census, our data set is not precisely comparable with the one [Massey and Denton (1988)] employed. ... In a few cases, new counties were added to metropolitan areas between 1980 and 1990; however, by far the most common change was an intercensal adjustment of tract boundaries.” Massey, White, and Phua (1996: 182) reported that “the 1990 solution [using 58 MSAs] is not as clean as that observed in 1980. Factorial complexity is greater and the five axes are not as well defined as before (i.e., indexes for conceptually distinct dimensions often display high loadings on the same underlying factor). In addition, the factor pattern matrix is dominated by the first factor [evenness] more in 1990 than in 1980, and the other four factors explain relatively less of the common variance.” Moreover, “clustering is no longer well established as an independent dimension in 1990” (Massey, White, and Phua 1996: 185). Rugh and Massey (2014) and Massey and Tannen (2015) avoided the problems of changing boundaries by using consistently defined metropolitan boundaries from 1970 to 2010.

Arcaute et al. (2015: 2) stated that “metropolitan statistical areas (MSAs) in the USA, and larger urban zones (LUZs) in Europe ... were designed to incorporate urbanized and economic functional areas, but they are not necessarily consistent with one another as no consensus exists on how cities should be defined.” Arcaute et al. (2015) showed that alternative methods of delineating boundaries of cities in England and Wales, which they devised, systematically affect the claimed power-law scaling of 30 urban attributes as a function of total urban population in the 2001 census. No measures of population concentration were included among their 30 indicators.

It appears that a gap, a research opportunity, remains. Arcaute et al. (2015) examined widely differing definitions of city boundaries at a given time but not their consequences for indicators of spatial concentration of population. A question remains open: How sensitive are measures of population concentration empirically to different methods of delineating the boundaries and the internal spatial units of cities?

## **6. Acknowledgments**

Constructive suggestions from three referees and Douglas S. Massey greatly improved the paper. Roseanne Benjamin provided assistance during this work.

## References

- Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., and Batty, M. (2015). Constructing cities, deconstructing scaling laws. *Journal of the Royal Society Interface* 12: 20140745. doi:10.1098/rsif.2014.0745.
- Batty, M. (1974). Urban density and entropy functions. *Journal of Cybernetics* 4(2): 41–55. doi:10.1080/01969727408546065.
- Bertaud, A. and Malpezzi, S. (2014). The spatial distribution of population in 57 world cities: the role of markets, planning, and topography. June 27, 2014. [https://coller.m.tau.ac.il/sites/nihul.tau.ac.il/files/media\\_server/Recanati/management/elrov/june2014\\_symposium/Malpezzi.pdf](https://coller.m.tau.ac.il/sites/nihul.tau.ac.il/files/media_server/Recanati/management/elrov/june2014_symposium/Malpezzi.pdf).
- Bussi re, R. and Snickars, F. (1970). Derivation of the negative exponential model by an entropy maximising method. *Environment and Planning* 2: 295–301. doi:10.1068/a020295.
- Buzzelli, M. (2020). Modifiable areal unit problem. In: Kobayashi, A. (ed.). *International encyclopedia of human geography* (2nd ed). 9: 169–173. New York: Elsevier. doi:10.1016/B978-0-08-102295-5.10406-8.
- Clark, C. (1951). Urban population densities. *Journal of the Royal Statistical Society* 114(4): 490–496. doi:10.2307/2981088.
- Duncan, O.D. (1957). The measurement of population distribution. *Population Studies* 11: 27–45. doi:10.1080/00324728.1957.10413229.
- Gehlke, C.E. and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association* 29(185A): 169–170. doi:10.2307/2277827.
- Marsh, C. (2013). Introduction to continuous entropy [electronic resource]. Princeton University. [https://www.crmarsh.com/static/pdf/Charles\\_Marsh\\_Continuous\\_Entropy.pdf](https://www.crmarsh.com/static/pdf/Charles_Marsh_Continuous_Entropy.pdf).
- Massey, D.S. and Denton, N.A. (1988). The dimensions of residential segregation. *Social Forces* 67(2): 281–315. doi:10.1093/sf/67.2.281.
- Massey, D.S. and Denton, N.A. (1998). The elusive quest for the perfect index of concentration: Reply to Egan, Anderton, and Weber. *Social Forces* 76(3): 1123–1133. doi:10.2307/3005706.

- Massey, D.S., White, M.J. and Phua, V.C. (1996). The dimensions of segregation revisited. *Sociological Methods and Research* 25(2): 172–206. doi:10.1177/0049124196025002002.
- Massey, D.S. (2012). Reflections on the dimensions of segregation. *Social Forces* 91(1): 39–43. doi:10.1093/sf/sos118.
- Massey, D.S. and Tannen, J. (2015). A research note on trends in black hypersegregation. *Demography* 52: 1025–1034. doi:10.1007/s13524-015-0381-6.
- McDonald, J.F. (1989). Econometric studies of urban population density: A survey. *Journal of Urban Economics* 26: 361–385. doi:10.1016/0094-1190(89)90009-0.
- Newling, B.E. (1969). The spatial variation of urban population densities. *Geographical Review* 59(2): 242–252. doi:10.2307/213456.
- Rugh, J.S. and Massey, D.S. (2014). Segregation in post-Civil Rights America: Stalled integration or end of the segregated century? *Du Bois Review: Social Science Research on Race* 11(2): 205–232. doi:10.1017/S1742058X13000180.
- Smith, B.E. (1997). A review of monocentric urban density analysis. *Journal of Planning Literature* 12(2): 115–135. doi:10.1177/088541229701200201.
- Stewart, J.Q. (1947). Suggested principles of “Social Physics.” *Science* 106(2748): 179–180. doi:10.1126/science.106.2748.179.
- Stewart, J.Q. (1953). Urban population densities. *Geographical Review* 43(4): 575–576.
- US Bureau of the Census (1994). *Geographic Areas Reference Manual (GARM)*. Washington, DC: US Department of Commerce, Economics and Statistics Administration. <https://www2.census.gov/geo/pdfs/reference/GARM/GARMcont.pdf>.
- Ye, X. and Rogerson, P. (2021). The impacts of the modifiable areal unit problem (MAUP) on omission error. *Geographical Analysis* doi:10.1111/gean.12269.