Estimating Inequality Measures from Quantile Data

Enora Belz

Centre de Recherche en Économie et Management
Center for Research in Economics and Management

University of Rennes 1

University of Caen Normandie

Working Paper

Estimating Inequality Measures from Quantile Data

Enora Belz
Univ Rennes, CNRS, CREM UMR 6211, F-35000 Rennes, France

Octobre 2019 - WP 2019-09

UNIVERSITÉ CAEN NORMANDIE

UNIVERSITÉ DE RENNES 1

Estimating Inequality Measures from Quantile Data

Enora Belz[1,*]

[1]Univ Rennes, CNRS, CREM - UMR 6211, F-35000 Rennes, France
[*]enora.belz@univ-rennes1.fr

Abstract

This article focuses on the problem of dealing with aggregate data. It proposes an innovative method for modelling Lorenz curves and estimating inequality indices on small populations, when (only) quantiles are available. When dealing with small population areas and due to privacy restrictions, individual or income share data are often not available and only quantiles are reported. The method is based on conditional expectation in order to find the different income shares and thus model a Lorenz curve with the functional forms already proposed in the literature. From this Lorenz curve, inequality indices (Gini, Pietra, Theil indices) can be derived. A simulation study is performed to evaluate this method and compare it with the other methods used. An example based on real Parisian data is presented to illustrate the method. A R package was written with all functions used in this article[1].

Keywords: inequalities; income; distribution; aggregated data; Lorenz curve; Gini; Pietra; Theil.

1 Introduction

A suitable and widely used approach to depicting economic inequalities is to provide indices to measure the degree of inequality. These measures enable comparisons of living standards between different countries, regions or among time. Descriptive measures could be useful in understanding economic relationship (Kaplow, 2005). In this regard, several indices have been described in the literature, including the Gini coefficient, the Pietra index or the Theil indexes. Similarly, the Lorenz curve is a relevant indicator of income distribution and most indices are related to it. The modelling of a Lorenz curve then provides a useful tool. Two main strategies have been developed for modelling a Lorenz curve, either by approximating the empirical Lorenz curve or either by modelling an income distribution and deriving the Lorenz curve from it.

The first is focused on the parametric approximation of the size distribution of income. Based on this estimation, a Lorenz curve and the several indices can be derived (see Champernowne and Cowell (1998) for a survey). Several relevant functional forms have been used to describe the income distribution (see Chotikapanich (2008) for an exhaustive list). The most popular forms are the so-called generalized distribution of the second kind (GB2) proposed by McDonald (1984) and its special and limited cases. Special and limited cases include Pareto distributions, Log-normal distributions, Gamma-type size distributions (Gamma, Generalized Gamma, Weibull) and Beta-type size distribution (GB2, Singh-Maddala, Dagum, Fisk, Beta distribution of the first and the second kind) (see Kleiber and Kotz (2003) for a survey). Alternatively, an approach is to apply on shares the generalized non-parametric Pareto interpolation technique developed by Blanchet et al. (2017).

The second literature examines the parametric approximation of a Lorenz curve. In the same way, several functional forms have been developed to fit Lorenz curves. The pioneers are Kakwani and Podder (1973) who proposed a functional form consistent with the distribution of Australian data. Many others authors

[1]See https://github.com/EnoraBelz/Inequality.

suggested functional forms, in particular, Rasche et al. (1980), Pakes (1981), Gupta (1984), Arnold (1986), Villaseñor and Arnold (1989), Basmann et al. (1990), Ortega et al. (1991), Chotikapanich (1993), Sarabia (1997) or Rohde (2009). Several methods allow the estimation of the functional form parameters. The earlier models are based on linear or non-linear least squares. Castillo et al. (1998) and Sarabia et al. (1999) proposed an alternative approach by considering the median or least median square. More recently, Chotikapanich and Griffiths (2002) developed a maximum likelihood estimator based on a Dirichlet distribution to capture the cumulative nature of the shares. The selection of the best functional form is subsequently based on an adjustment criterion. Chotikapanich and Griffiths (2005) suggested an alternative way by averaging functional forms using a Bayesian model averaging approach.

However, problems arise due to the lack of income data at the individual level. Data are usually aggregated and provide less information than individual data. Institutions sometimes report inequality indices with aggregate data, although most only the Gini index. In such cases, an interesting approach is to reconstruct the proposed index and provide alternative indices. Analysis is straightforward when individual data are accessible. The income distribution and the Lorenz curve can be estimated directly from their empirical forms. However, personal income data are not widely available. They are usually aggregated, whether census or survey data. Census data appear to be more relevant to provide an overview over time or across sub-regions. Nevertheless, the institutions provide several different forms of available data. At the national or regional scale, data are mostly reported in the form of income shares. Institutions may provide data as class mean income and deciles or quintiles group shares like on World Income Inequality Database (WIID), World Inequality Database (WID) or even World Bank. The available information thus depicts points on the Lorenz curve. For example, the poorest 10% hold 3.6% of total income in France in 2016 (WIID). When dealing with smaller areas, the available data may be only quantiles. The income quartiles, quintiles or deciles can be provided but not income shares or class mean income. The available information does not depicts points on the Lorenz curve. For example, the poorest 10% receive less than €10,860 in France in 2016 (INSEE). Unfortunately, their share in the total income is not available. The information given is therefore nearby but not precisely the same. In other words, when focusing on a high geographical level (region or country), income shares can be found, although when dealing with smaller areas such as cities, the data provided are frequently reported as income quantiles.

The purpose of this paper is to determine a methodology for estimating a Lorenz curve and associated indices with only quantile data. The methods need to be adjusted to obtain a Lorenz curve from these quantiles. Quantiles can easily be transformed into tabulated data. The classic method with tabulated data is to assume the midpoint of each class as the class mean. This assumption is rather strong. For this reason, we develop an innovative method based on conditional expectations to compute class means. We will illustrate this method by measuring inequalities within the city of Paris. French municipalities can be subdivided into narrower areas called iris (Ilots Regroupés pour l'Information Statistique). The iris generally have between 1800 and 5000 inhabitants and are built in relation to large sections of the urban network. The narrowness of the areas limits the data available due to the privacy of personal data and INSEE (Institut National de la Statistique et des Études Économiques) provides only quartiles and deciles of income. This methodology is also applicable to other quantile data such as ZIP code income data in the United States.

In Section 2, we will present the proposed methodology for modelling a Lorenz curve from quantile data. Section 3 will illustrate the method with an application on simulated income data and Parisian iris data. A R vignette with codes is available from https://github.com/EnoraBelz/Inequality.

2 Methods

2.1 Definition of a Lorenz curve

The Lorenz curve, introduced by Lorenz (1905), is the curve defined by the points $(p, L(p))$ where p is the cumulative proportion of the income-receiving units sorted from the poorest to the richest and $L(p)$ the cumulative proportion of income received by these units. Gastwirth (1971) proposed a general definition of the Lorenz curve such that, if X is the income of a member of the population and assumed to be a random variable with cumulative distribution function $F(x)$, quantile function $F^{-1}(x)$ and mean $\mu = \int x dF(x)$, then the Lorenz curve is the mapping

$$p \mapsto L(p) = \frac{1}{\mu} \int_0^p F^{-1}(t) dt \tag{1}$$

A Lorenz curve will have some expected properties: (1) if $p = 0$ then $L(p) = 0$, (2) if $p = 1$ then $L(p) = 1$, (3) $L(p) \leq p$ and (4) $L(p)$ is continuous and differentiable and the slope of the curve increases monotonically. Several authors developed functional forms to fit a Lorenz curve (see Table 1).

Table 1. Lorenz curve functional forms

	Functional form $L(p)$	Parameter constraints
Kakwani and Podder (1973)	$p^\alpha e^{-\beta(1-p)}$	$\beta > 0, \alpha \geq 1$
Rasche et al. (1980)	$(1-(1-p)^\alpha)^\beta$	$\beta \geq 1, 0 \leq \alpha \leq 1$
Arnold (1986)	$\frac{p(1+(\alpha-1)p)}{1+(\alpha-1)+\beta(1-p)}$	$\alpha, \beta > 0, \alpha - \beta < 1$
Ortega et al. (1991)	$p^\alpha(1-(1-p)^\beta)$	$\alpha \geq 0, 0 < \beta < 1$
Chotikapanich (1993)	$\frac{e^{kp}-1}{e^k-1}$	$k > 0$
Sarabia (1997)	$\pi_1 p + \pi_2 p^{\alpha_1} + (1-\pi_1-\pi_2)(1-(1-p)^{\alpha_2})$	$0 \leq \pi_1, \pi_2 \leq 1, \alpha_1 \geq 1, 0 < \alpha_2 < 1$
Rohde (2009)	$p(\frac{\beta-1}{\beta-p})$	$\beta > 1$

2.2 From quantile data to tabulated data

These parametric forms require grouped or individual data which allow the derivation of an empirical Lorenz curve. Quantiles are not sufficient in their current form. Information on the cumulative proportion of the population is given, but the cumulative proportion of income is unknown. Indeed, a quantile indicates the income level such that $p\%$ is below this threshold. The information concerning the total income received by $p\%$ is not available and therefore the share in the total income of $p\%$ is undefined. Nevertheless, quantiles can be transformed into tabulated data. Quantiles become the boundaries of income ranges and the size of the subset becomes the share of individuals. In the case of deciles, the boundaries are the deciles and the proportion of individuals in each range is 10 percent. The use of both quartiles, deciles or other q-quantiles is also feasible. The proportion of individuals depends on being between two same subdivisions or between two different subdivisions.

Suppose X a random variable of income and different associated quantiles Q_k where $0 < k < K$ is the rank of the quantile and s_k the population share below this quantile so that $\mathbb{P}(X < Q_k) = s_k$. The ranges

are therefore $[0, Q_1[$ with proportion s_1 for the first range, $[Q_k, Q_{k+1}[$ with proportion $s_{k+1} - s_k$ from the second to the penultimate range and $[Q_K, +\infty[$ with proportion $1 - s_K$ for the last range (see Table 2 for an example).

Table 2. From quantile data to tabulated data with quartiles and deciles

From quantile data...											
	D_1	D_2	Q_1	D_3	D_4	Q_2/D_5	D_6	D_7	Q_3	D_8	D_9
	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%
...to tabulated data											
$[0, D_1[$	$[D_1, D_2[$	$[D_2, Q_1[$	$[Q_1, D_3[$	$[D_3, D_4[$	$[D_4, Q_2[$	$[Q_2, D_6[$	$[D_6, D_7[$	$[D_7, Q_3[$	$[Q_3, D_8[$	$[D_8, D_9[$	$[D_9, \infty[$
10%	10%	5%	5%	10%	10%	10%	10%	5%	5%	10%	10%

2.3 Income shares

The first step is to determine the total income received by each income bin and thereby the shares of income received. This requires determining the mean income within each income bin. In many studies, the mean income of each class is assumed to be the midpoint of the interval and a Pareto-tail for the open-ended interval (*Midpoint method*). In other words, it means that, somehow, we assume that the income is uniformly distributed in the ranges and consequently the center of mass (i.e. the mean) equals the center of the interval. This assumption is rather strong. Indeed, this assumes that each individual in a group has the same income level and therefore neglects intra-group variations. In addition, some income levels are more predominant in a population (e.g. minimum income, agreement income).

It can be noticed that quantiles enable to fit an income distribution function. Several different functional forms have been developed to model the distribution of income. McDonald (1984) suggested the Generalized Beta distribution of the second kind (GB2) which includes most of the previous functional forms as limited or special cases. For this reason, a GB2 distribution is assumed to describe the size of income and the four parameters are optimizied on quantile data. Suppose X a random variable of income with known density $f(x)$, the conditional expectation of being in a range $[a, b[$ can be determined as

$$E(X|x \in [a, b[) = \frac{\int_a^b x f(x) dx}{\int_a^b f(x) dx} \tag{2}$$

where $f(x)$ is the density of X. Therefore, knowing the size distribution of income allows to determine the means of the income per bins. The conditional means between each range can be computed by plugging-in the (four) estimated parameters of the GB2 distribution (*Conditional Expectation method*). Using the class mean income and the class population share , the income share of each bin in the total income can then be determined and consequently the cumulative income shares.

2.4 Functional form optimization

Quantile data become similar to grouped data and classical functional forms of Lorenz curves can be applied. In this way, seven different forms are used : Kakwani and Podder (1973) , Rasche et al. (1980), Arnold (1986),

Ortega et al. (1991), Chotikapanich (1993), Sarabia (1997) and Rohde (2009). The parameters are optimized by non-linear least squares (NLS) estimator as

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} (F(p_k, \theta) - s_k)^2 \tag{3}$$

where $F(., \theta)$ is the functional form to be optimized according to the parameter(s) θ, p_k is the population share of the k^{th} income range and s_k is the income share of the k^{th} income range. Thereafter, the different functional forms are compared according to a goodness-of-fit measure, especially the value of the Chi-squared statistic

$$\chi^2 = \sum_{k=1}^{K} \frac{(s_k - F(p_k, \hat{\theta}))^2}{F(p_k, \hat{\theta})} \tag{4}$$

where $F(., \hat{\theta})$ is the functional form with the optimal parameter(s) $\hat{\theta}$, p_k is the population share of the k^{th} income range and s_k is the income share of the k^{th} income range.

The algorithm is therefore as follows:

- Convert quantiles into tabulated data
- Estimate a GB2 family distribution on quantiles using ML estimation techniques
- Compute conditional expectations of each bin using the fitted distribution
- Calculate income shares and cumulative income shares
- Estimate a functional form on shares using a NLS estimator
- Compare the functional forms with goodness-of-fit measures.

2.5 Inequality measures

The Lorenz curve enables to determine the inequality indices. Gini coefficient G (Gini, 1914) can be evaluated as

$$G = \frac{E(|Y - X|)}{2E(X)} \tag{5}$$

where X and Y are i.i.d with common distribution $F(x)$. Alternatively, the Gini coefficient can be determined in terms of the Lorenz curve, corresponding to two times the area between $L(p)$ and the egalitarian line as

$$G = 2 \int_0^1 (p - L(p)) dp \tag{6}$$

As a result, Gini coefficients can be derived from the functional forms as a function of parameters (Table 3).

Table 3. Gini coefficient of the functional forms

	Gini Index
Kakwani and Podder (1973)	$1 - \frac{2e^{-\beta}}{1+\alpha}{}_1F_1(1+\alpha; 2+\alpha; \beta)$
Rasche et al. (1980)	$1 - \frac{2}{\alpha}B(\frac{1}{\alpha}; \beta+1)$
Arnold (1986)	$\begin{cases} \frac{\beta}{\beta-\alpha+1} + \frac{2\alpha\beta}{(\beta-\alpha+1)^2}[1 + \frac{\beta+1}{\beta-\alpha+1}\log(\frac{\alpha}{\beta+1})] & \text{if } \beta-\alpha+1 \neq 0 \\ \frac{\beta}{3(1+\beta)} & \text{if } \beta-\alpha+1 = 0 \end{cases}$
Ortega et al. (1991)	$\frac{\alpha-1}{\alpha+1} + 2B(\alpha+1; \beta+1)$
Chotikapanich (1993)	$\frac{(k-2)e^k + (k+2)}{k(e^k-1)}$
Sarabia (1997)	$\pi_2(1 - \frac{2}{1+\alpha_1}) - (1-\pi_1-\pi_2)(1-\frac{2}{1+\alpha_2})$
Rohde (2009)	$2\beta[(\beta-1)\log(\frac{\beta-1}{\beta}) + 1] - 1$

Note: B is the Beta function. ${}_1F_1$ is the confluent hyper-geometric function.

The Pietra index P (Pietra, 1932) is also widely used to measure inequality, it can be defined as

$$P = \frac{E(|X - E(X)|)}{2E(X)} \tag{7}$$

The Pietra index can also be defined in terms of the Lorenz curve, it corresponds to the maximum deviation between $L(p)$ and p as

$$P = \max_{p\in[0,1]}\{p - L(p)\} \tag{8}$$

In addition, it corresponds to two times the area of the largest triangle that can be inscribed between $L(p)$ and the equalitarian line (Arnold, 2008).

In his book, Theil (1967) also suggests inequality indices, Theil's indexes T_L (low) and T_H (high), based on the generalized entropy (see Cowell (2011) for a discussion on generalized entropy measures).

$$T_L = GE_0 = -E(\log(\frac{X}{\mu})) \tag{9}$$

$$T_H = GE_1 = E(\frac{X}{\mu}\log(\frac{X}{\mu})) \tag{10}$$

Rohde (2008) relates the concept of Generalized Entropy (GE) to the Lorenz curve and provides mathematical expressions to define the GE measures in terms of the Lorenz curve.

$$T_L = -\int_0^1 \log(L'(p))dp \tag{11}$$

$$T_H = \int_0^1 L'(p)\log(L'(p))dp \tag{12}$$

where $L'(p)$ is the first derivative of the Lorenz Curve.

3 Applications

In a first part, we will simulate samples to evaluate the performance of the methodology and compare it to other methods, then, we will use it to get local inequality indices in Paris.

3.1 Simulated data

Simulations are performed both to evaluate the performance of the method and to compare with the midpoint method. The different methods are applied on income distribution simulations. Several known distributions are selected to perform the simulations such as GB2, Log-normal and Singh-Maddala. The parameters of the distributions are determined to ensure realistic income levels. For each simulation, the income of 2000 individuals is simulated according to a defined distribution. Two areas are considered, one with a low Gini index and one with a high index. Three distributions[2] are used: for the High Gini Index case, $LN(10.6, 1.01)$, $GB2(40000, 1.7, 0.98, 1.02)$ and $SM(30000, 1.9, 0.7)$, respectively with Gini index 0.52, 0.58 and 0.69, and for the Low Gini Index case, $LN(10.5, 0.7)$, $GB2(35000, 2.5, 0.95, 1.02)$ and $SM(50000, 2.2, 1.8)$, respectively with Gini index 0.37, 0.39 and 0.35.

For each of them,

(i) we generate a sample of size 2000 and calculate Gini index from those individual observations directly - called Ind.

(ii) we can get the shares and partial information (quantiles), and then compute Gini index - called Shares

we can use only quantiles - and estimate the shares

(iii) the first step of estimating a GB2 distribution allows to compute a Gini index - called First Step

(iv) with the conditional expectation method - called Cond. Expectation

(v) with the midpoint method - called Midpoint.

Figure 1 provides a comparison of the different methods in terms of estimated Gini index.

Method (i) is the one with detailed data, with the other four are based on aggregated quantities. The distribution of the Gini estimate based on shares data (ii) is very close to that observed on the individual data (i). Then, when data are in the form of shares, having aggregate or individual data gives a fairly close result. However, if we make as if the shares are not available (as in the dataset we have) and only quantiles are available, the lack of information leads to estimating a Gini index less close to the true value.

Admittedly, the first step of the Conditional Expectation method (iii) also allows to compute a Gini index. This consists of fitting a GB2 distribution on the quantiles to determine conditional expectations, however, indices and the Lorenz curve can be computed using the GB2 distribution. Nevertheless, the underlying income distribution is not necessarily GB2 and, in addition, the GB2 distribution is a four-parameter distribution estimated on only a few points, and can therefore be misspecified, especially on the distribution tail. The second step (iv) should adjust the method by improving the results. The results suggest that the Conditional Expectation method (iv) computes a Gini index close than the GB2 one (iii). When the underlying distribution is GB2, both methods provide the same value close to the true value. When the

[2]Lognormal distribution $LN(\mu, \sigma)$, GB2 distribution $GB2(\mu, \sigma, \nu, \tau)$ and Singh-Maddala distribution $SM(\mu, \sigma, \nu)$.

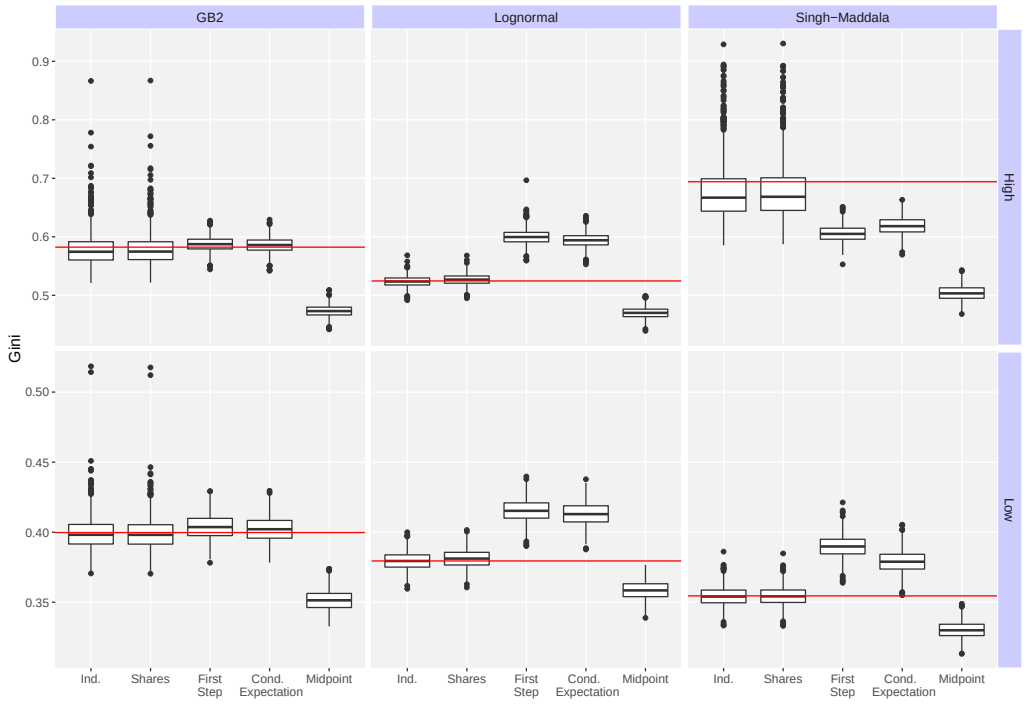

Figure 1. Comparison of the different methods in terms of estimated Gini

Note: The results are based on 1000 simulations for each cases. The cases are: *Ind.* the Gini on the individual data, *Shares* the Gini estimated on shares, *First Step* the Gini estimated with the GB2 fitting, *Cond. Expectation* the Gini estimated on quantiles with the Conditional Expectation method and *Midpoint* the Gini estimated on quantiles with the Midpoint method. The red line corresponds to the true value of the Gini index of the simulated distribution (obtained by $1,000,000$ simulations). The parameters of the simulated distributions are: *High* $LN(10.6, 1.01)$, $GB2(40000, 1.7, 0.98, 1.02)$ and $SM(30000, 1.9, 0.7)$ and *Low* $LN(10.5, 0.7)$, $GB2(35000, 2.5, 0.95, 1.02)$ and $SM(50000, 2.2, 1.8)$.

distribution is not GB2, the results indicate an overestimation of Gini coefficient (or underestimation in the case of Singh-Maddala *High Gini Index*), however, the Gini index of the second step is more accurate. As a result, the Conditional Expectation method adjusts the estimate of Gini index.

Conditional expectation (iv) and Midpoint (v) methods can also be compared. The Midpoint method underestimates the true values in all cases, while the conditional expectation method only underestimates the true values for Singh-Maddala *High Gini Index* and overestimates for Singh-Maddala *Low Gini Index* and Log-normal. If the distribution is GB2 or Singh-Maddala, the Conditional Expectation method significantly outperforms the Midpoint method. When the distribution is Log-normal, the result is less clear. The Conditional Expectation method overestimates the true value while the midpoint method underestimates it. The performance of both methods is then related to the underlying distribution.

3.2 Parisian income data

This paper focuses on the measure of inequality within Paris. The Institut National de la Statistique et des Études Économiques (INSEE) provides income data at a very narrow scale named iris (see Figure 2 for a

visualisation of median income in the iris). The deciles and quartiles of income[3] in these areas are available. Paris includes 966 iris which are either living (861), activity (88) or miscellaneous (17). For some iris with insufficient population (especially activity or miscellaneous iris), income information is not available. Thus, 865 iris are reported with income data. Using these data, a Lorenz curve can then be modelled, the Gini index can be determined and compared with that provided by INSEE, and the Pietra and Theil indices not included can be calculated.

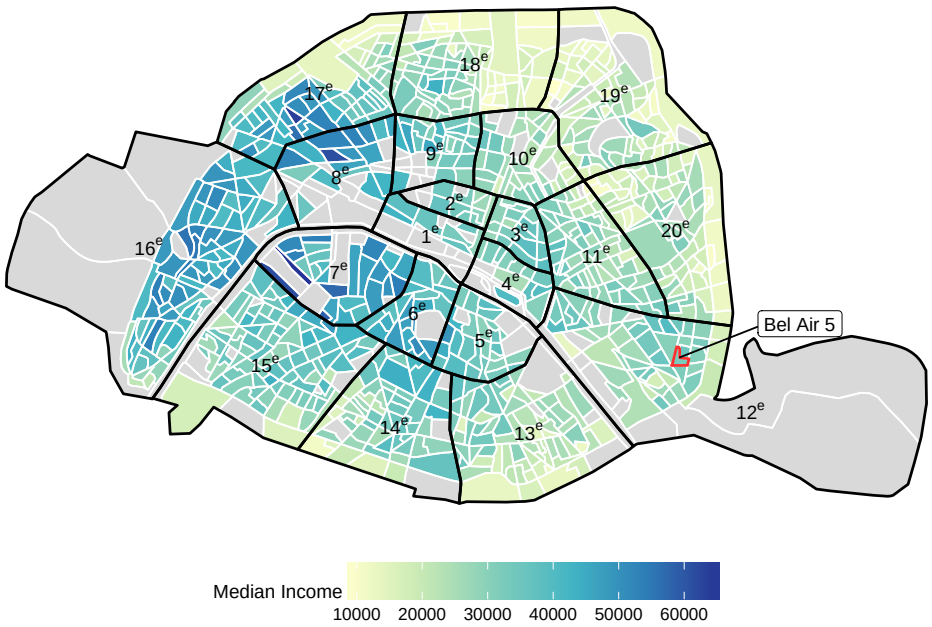

Figure 2. Median income of the Parisian iris

The first step is to determine the parameters of the GB2 distribution of each area by MLE (see dotted line in Figure 3). The parameters are used to identify the density and the distribution of income in the area. Conditional means of income bins can then be calculated (see blue lines in Figure 3). Those means are used to compute the cumulative shares of income and the cumulative shares of population are the shares between two quantiles. Table 4 displays an example of the data obtained for an iris ("Bel Air 5" in the 12th arrondissement). In this area, the first decile is at €10,570, implying that 10% of the population have an income below €10,570. In this range, the conditional mean is found to be at €7,403. The cumulative share of income can be derived from the total income (weighted sum of the conditional means). In this area, the poorest 10% of the population receive 2% of the total income. By plotting the cumulative share of income in relation to the cumulative share of the population, the points of the empirical Lorenz curve are observed (Figure 4).

[3]Quantiles of declared household income per consumption unit for the year 2014. The used scale (OECD scale) has the following weighting: 1 UC for the first adult in the household, 0.5 UC for other persons aged 14 or over and 0.3 UC for children under 14 years.

Table 4. Tabulated data obtained using the conditional expectation method for Bel Air 5 iris

	Quantile (€)	Income Range (€)	Proportion	Mean Income (€)	Income share	Population share
D_1	10,570	Below 10,570	0.10	7,403.64	0.02	0.10
D_2	18,696	10,570-18,696	0.10	14,947.50	0.06	0.20
Q_1	21,558	18,696-21,558	0.05	20,135.63	0.08	0.25
D_3	24,344	21,558-24,344	0.05	22,948.68	0.12	0.30
D_4	28,298	24,344-28,298	0.10	26,296.94	0.18	0.40
Q_2	32,626	28,298-32,626	0.10	30,412.82	0.26	0.50
D_6	37,782	32,626-37,782	0.10	35,113.38	0.36	0.60
D_7	43,444	37,782-43,444	0.10	40,488.39	0.46	0.70
Q_3	46,088	43,444-46,088	0.05	44,737.59	0.52	0.75
D_8	50,926	46,088-50,926	0.05	48,410.77	0.58	0.80
D_9	61,920	50,926-61,920	0.10	55,936.68	0.73	0.90
		61,920 and over	0.10	102,703.30	1	1

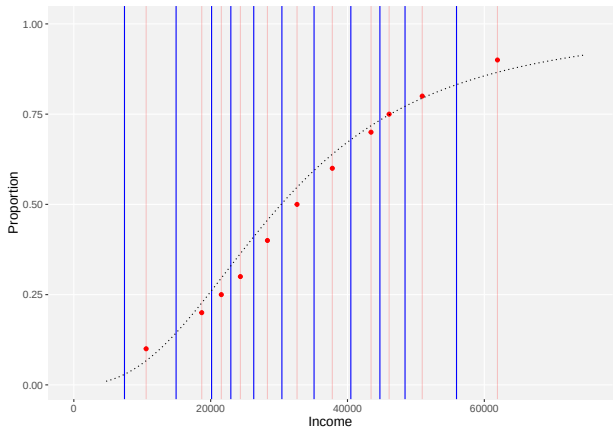

Figure 3. Estimated means of income per bins of Bel Air 5 iris

Note: The red lines and dots represent the observed quantiles. The blue lines represent the estimated means between two quantiles. The dotted line is the estimate of the GB2 cumulative distribution function.

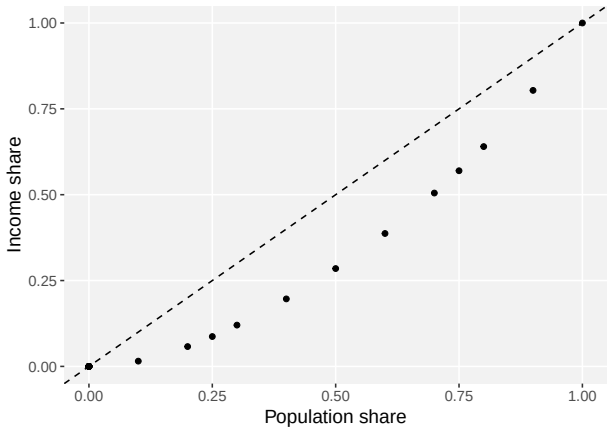

Figure 4. Empirical Lorenz curve of Bel Air 5 iris

To approximate the empirical Lorenz curve, all alternative functional forms can be applied on the cumulative shares of income and population. The parameters are optimized by NLS estimator. For each functional form, the most optimal parameters and Lorenz curve are obtained. The different forms could also be compared with each other using goodness-of-fit measures (see column 2 of the Table 5).

Table 5. Inequality measures for the different functional forms of the Bel Air 5 iris

Functional form	χ^2	Rank	Gini	Pietra	T_L	T_H
Kakwani and Podder (1973)	0.01028	4	0.345	0.261	0.205	0.185
Rasche et al. (1980)	0.00154	3	0.351	0.248	0.221	0.214
Arnold (1986)	0.01493	6	0.342	0.259	0.189	0.184
Chotikapanich (1993)	0.01102	5	0.344	0.261	0.199	0.184
Sarabia (1997)	0.00036	1	0.356	0.244	0.24	0.287
Ortega et al. (1991)	0.00112	2	0.352	0.247	0.225	0.221
Rohde (2009)	0.02149	7	0.340	0.259	0.183	0.183

Figure 5 represents the best functional form according to the chi-squared of each Parisian iris. 58% of the iris tend to be in favour of the Sarabia (1997) form, 38% for Ortega et al. (1991) form and 3% for Rasche et al. (1980) form. A rather notable pattern appears, the iris of the center and 16th arrondissements are generally in favour of an Ortega et al. (1991) form, while the iris of the outlying arrondissements are in favour of Sarabia (1997) form. This spatial pattern is similar to the one of inequality indices, and hence the functional form choice appears to be related to the income inequalities of the area.

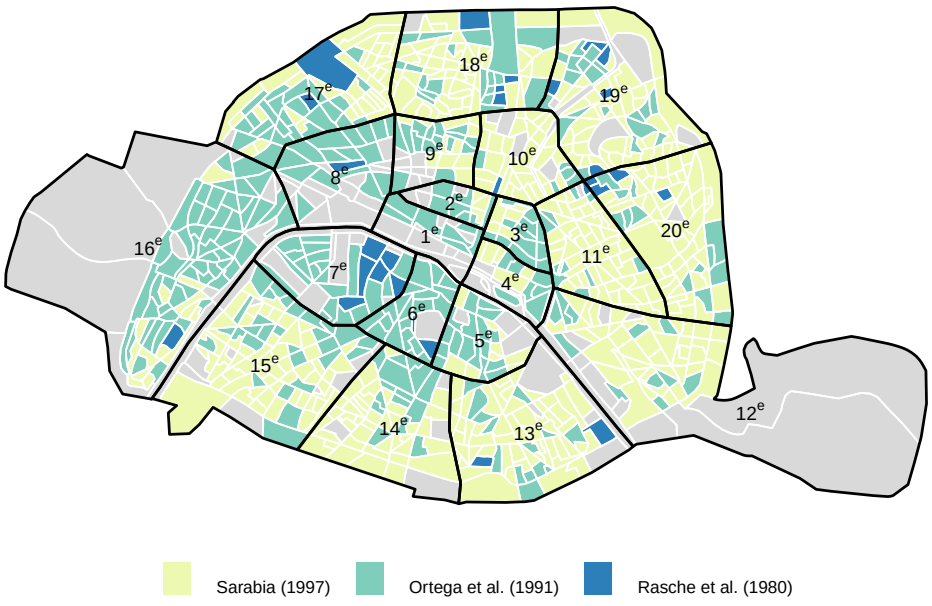

Figure 5. Best choice of functional forms obtained for each Parisian iris

The Lorenz curve is used to compute inequality indices (see Table 5 for Bel Air 5 iris). The Gini index can be defined as a function of each functional form parameters. The iris database provides the Gini index calculated from individual data. Then, a comparison between the one found with the functional form and the individual one should be considered. Figure 6 depicts the Gini derived from functional forms as a function of the one given by INSEE and Figure 7 the difference between them for each functional form. The method tends to slightly overestimate the Gini index. The difference remains quite small, in most cases ranging from 0 to 0.02. In addition, inequality indices not provided by INSEE, such as the Pietra or Theil indices, can be calculated according to the best functional form (see Figures 8 for Gini index, 10 for Pietra index, 9 for Theil L index and 11 for Theil H index).

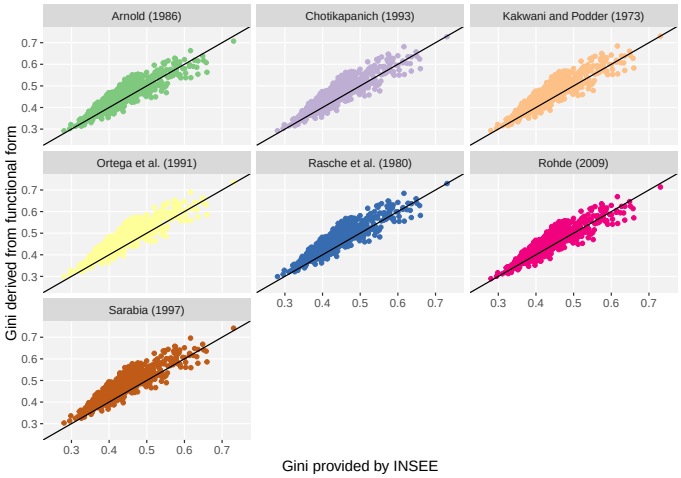

Figure 6. Comparison between estimated and observed Gini index by functional form

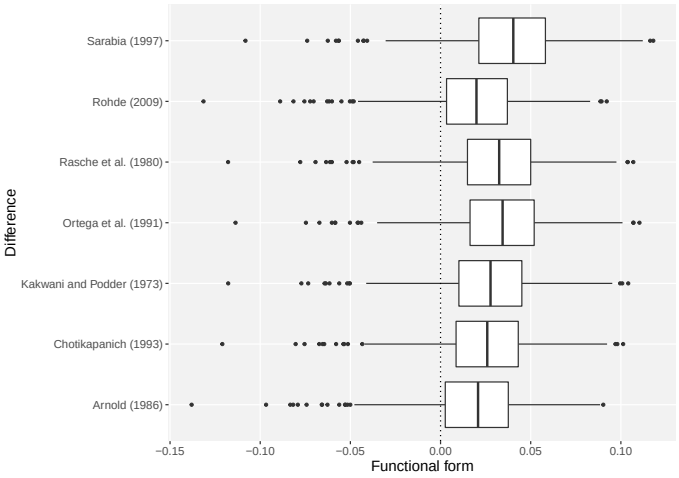

Figure 7. Difference between estimated and observed Gini index

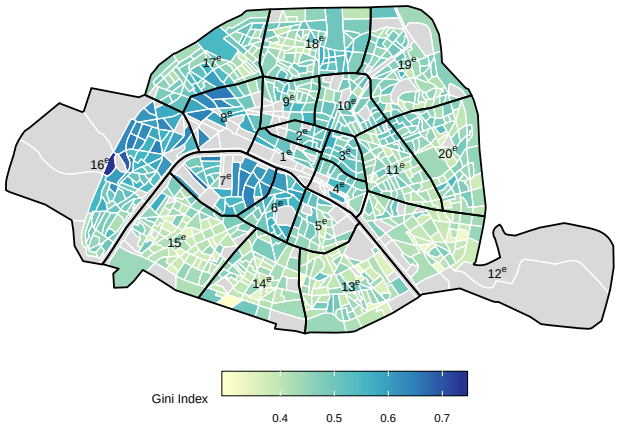

Figure 8. Gini Index

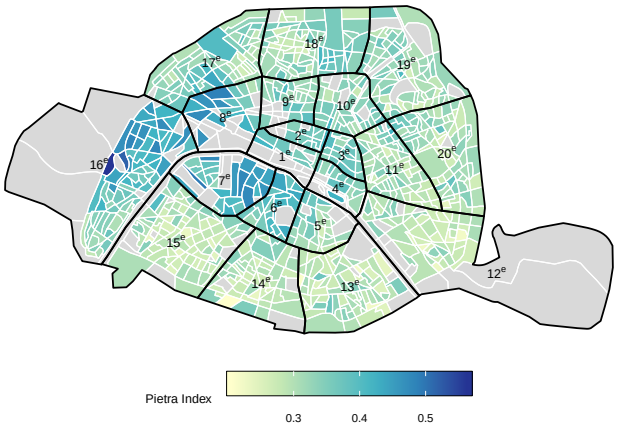

Figure 10. Pietra Index

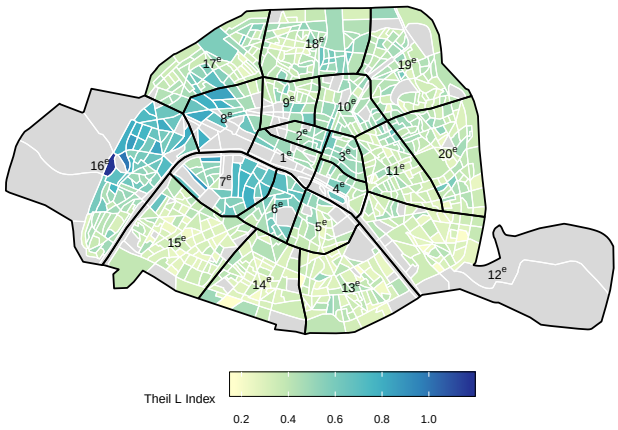

Figure 9. Theil L Index

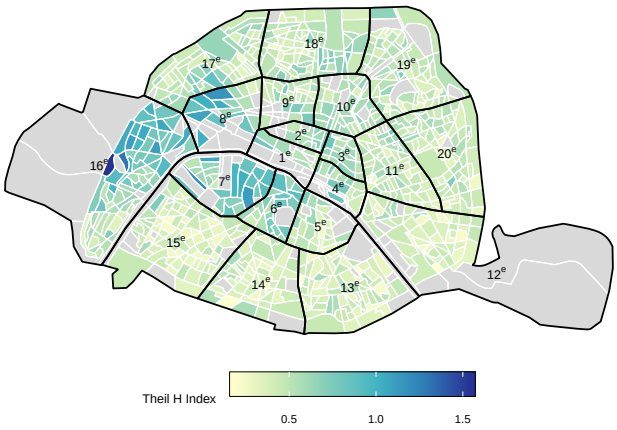

Figure 11. Theil H Index

4 Conclusion

The objective of this paper is to propose an innovative method to model Lorenz curves and estimate inequality indices on small populations, when only quantiles are available. The method is based on conditional expectation in order to find the different income shares and thus model a Lorenz curve with the functional forms already proposed in the literature. Real and simulated data are used to evaluate the proposed method and compare it to other methods used. We note from simulated data that it is more difficult to estimate a Gini index when shares are not available and only quantiles are available. However, the proposed Conditional Expectation method outperforms the traditional Midpoint method. Similarly, the method applied to the Parisian iris data provides a Gini index very similar to the true value. Finally, the proposed methodology enables to model a Lorenz curve and hence to estimate inequality indices with quantile data.

Therefore, this method is useful for measuring inequalities when data are limited. This approach can be applied on income data in quantile form. However, it can also be used for data in class form with an underlying distribution to find conditional expectations.

Acknowledgements

I would like to thank the Brittany region and the ACTINFO chair for their support. I also thank Olivier L'Haridon, Benoît Tarroux, Arthur Charpentier, Emmanuel Flachaire and the participants of the CNRS Thematic School on Public Policy Evaluation (ETEPP), the annual conference of the Société Canadienne de Science Économique (SCSE) and the R conference in Quebec 2019 for their constructive suggestions and discussions.

References

B. C. Arnold. A class of hyperbolic Lorenz curves. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 427–436, 1986.

B. C. Arnold. Pareto and generalized pareto distributions. In *Modeling income distributions and Lorenz curves*, pages 119–145. Springer, 2008.

R. L. Basmann, K. J. Hayes, D. J. Slottje, and J. Johnson. A general functional form for approximating the Lorenz curve. *Journal of Econometrics*, 43(1-2):77–90, 1990.

T. Blanchet, J. Fournier, and T. Piketty. Generalized pareto curves: theory and applications. 2017.

E. Castillo, A. S. Hadi, and J. M. Sarabia. A method for estimating Lorenz curves. *Communications in statistics-theory and methods*, 27(8):2037–2063, 1998.

D. G. Champernowne and F. A. Cowell. *Economic inequality and income distribution*. Cambridge University Press, 1998.

D. Chotikapanich. A comparison of alternative functional forms for the Lorenz curve. *Economics Letters*, 41(2):129–138, 1993.

D. Chotikapanich. *Modeling Income Distributions and Lorenz Curves*, volume 5. Springer Science & Business Media, 2008.

D. Chotikapanich and W. E. Griffiths. Estimating Lorenz curves using a Dirichlet distribution. *Journal of Business & Economic Statistics*, 20(2):290–295, 2002.

D. Chotikapanich and W. E. Griffiths. Averaging lorenz curves. *The Journal of Economic Inequality*, 3(1): 1–19, 2005.

F. Cowell. *Measuring inequality*. Oxford University Press, 2011.

J. L. Gastwirth. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039, 1971.

C. Gini. Sulla Misura della Concentrazione e della Variabilità dei Caratteri. *Atti del Reale Istituto veneto di scienze, lettere ed arti*, 73:1203–1248, 1914.

M. R. Gupta. Functional form for estimating the Lorenz curve. *Econometrica*, 52(5):1313–1314, 1984.

N. C. Kakwani and N. Podder. On the estimation of Lorenz curves from grouped observations. *International Economic Review*, pages 278–292, 1973.

L. Kaplow. Why measure inequality? *The Journal of Economic Inequality*, 3(1):65–79, 2005.

C. Kleiber and S. Kotz. *Statistical size distributions in Economics and Actuarial Sciences.*, volume 470. John Wiley & Sons, 2003.

M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219, 1905.

J. B. McDonald. Some Generalized Functions for the Size Distribution of Income. *Econometrica: journal of the Econometric Society*, pages 647–663, 1984.

P. Ortega, G. Martin, A. Fernandez, M. Ladoux, and A. Garcia. A new functional form for estimating Lorenz curves. *Review of Income and Wealth*, 37(4):447–452, 1991.

A. G. Pakes. *On Income Distributions and their Lorenz Curves.* Department of Mathematics, University of Western Australia, 1981.

G. Pietra. *Nuovi contributi alla metodologia degli indici di variabilita e di concentrazione.* Ferrari, 1932.

R. Rasche, J. Gaffney, A. Koo, and N. Obst. Functional forms for estimating the Lorenz curve: comment. *Econometrica*, 48(4):1061–1062, 1980.

N. Rohde. Lorenz curves and generalised entropy inequality measures. In *Modeling Income Distributions and Lorenz Curves*, pages 271–283. Springer, 2008.

N. Rohde. An alternative functional form for estimating the Lorenz curve. *Economics Letters*, 105(1):61–63, 2009.

J.-M. Sarabia. A hierarchy of Lorenz curves based on the generalized Tukey's lambda distribution. *Econometric Reviews*, 16(3):305–320, 1997.

J.-M. Sarabia, E. Castillo, and D. J. Slottje. An ordered family of Lorenz curves. *Journal of Econometrics*, 91(1):43–60, 1999.

H. Theil. Economics and information theory. Technical report, 1967.

J. Villaseñor and B. C. Arnold. Elliptical Lorenz curves. *Journal of Econometrics*, 40(2):327–338, 1989.