Rongen, Gerton

Working Paper

A new inequality estimate for urban India? Using house prices to estimate inequality in Mumbai

WIDER Working Paper, No. 2018/181

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Rongen, Gerton (2018) : A new inequality estimate for urban India? Using house prices to estimate inequality in Mumbai, WIDER Working Paper, No. 2018/181, ISBN 978-92-9256-623-4, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki, https://doi.org/10.35188/UNU-WIDER/2018/623-4

This Version is available at:
https://hdl.handle.net/10419/211219

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2018/181

A new inequality estimate for urban India?

Using house prices to estimate inequality in Mumbai

Gerton Rongen*

December 2018

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: This paper applies a novel inequality estimation method to household consumption expenditure in Mumbai, India. Since the richest households may be missing in survey data, this re-estimated inequality figure takes them into account by combining survey data with house price data. However, application of this method does not indicate that the survey-based Gini coefficient of 0.447 underestimates consumption inequality in Mumbai; none of the ten investigated scenarios yields a higher estimate. Further analyses are necessary to assess the robustness of estimates and the usefulness of applying this method to the whole of urban India.

Keywords: house prices, India, inequality, top incomes
JEL classification: C81, D31, O15

Acknowledgements: The author would like to thank Chris Elbers, Peter Lanjouw, Rinku Murgai, Roy van der Weide and the other participants in the India component of the UNU-WIDER project on 'Inequality in the Giants' for their very helpful comments.

* Amsterdam Institute for Global Health and Development, Amsterdam, Netherlands, g.rongen@aighd.org.

This study has been prepared within the UNU-WIDER project on 'Inequality in the Giants'.

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-623-4

Typescript prepared by Ans Vehmaanperä.

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

The Indian economy has grown enormously over the last twenty-five years, averaging real output growth of 7 per cent per year.[1] At the same time, the poverty headcount rate has decreased rapidly from 49 per cent in 1987/88 to 21 per cent in 2011/12. In absolute numbers, around 400 million Indians lived in extreme poverty in 1987/88, while that number had fallen to about 270 million people in 2011/12.[2]

The consequences of growth for overall income inequality are less clear. There is no consensus about the prevailing level of inequality or about inequality developments in the past decades. The main survey-based estimate for India as a whole is a Gini coefficient of 0.36 for the year 2011/12, and a figure of 0.38 for urban India (Himanshu, 2015). These values are comparable to estimates for neighbours Bangladesh and Pakistan, which had Gini coefficients of 0.32 and 0.31 respectively in 2010 and 2011/12. Nevertheless, inequality is low in comparison to countries like Brazil (0.53 in 2012), China (0.42 in 2012), and South Africa (0.63 in 2010). Except for Brazil, all these estimates give the dispersion of consumption expenditures instead of income inequality. Empirically, income inequality tends to be higher than consumption inequality. With India growing richer, some households are likely to save a larger part of their income. This means that the difference between consumption and income inequality could become larger.

Another possible limitation of this estimate is that it is survey-based. Several studies have shown that richer households are more likely to be missing or not covered by surveys. This can be due to non-response, underreporting, or a combination of both (see for example Korinek et al., 2006). The consequence is that inequality may be underestimated.

One alternative approach is to rely on tax data to estimate the incomes of the richest households, as has been done for India by Chancel and Piketty (2017). They estimate that the share of income that goes to the richest 1 per cent of households has increased to 22 per cent of total national income – the highest number since the start of availability of tax data in 1922. The figures they present have become part of the Indian debate on inequality and are not undisputed.

Furthermore, reliable tax data may not always be available. With the aim of including the richest households absent in surveys, van der Weide et al. (2018) developed a method that allows for the re-estimation of inequality by combining survey data with a second database containing predictors of income or consumption. Using the information from this second database, the distribution of top incomes can be estimated and added to the distribution inferred from the survey. New inequality measures, such as the Gini coefficient, mean log deviation and Theil index can subsequently be estimated. Van der Weide et al. illustrate the method by applying it to Egypt, using a database of house prices to estimate the top tail of the income distribution.

The aim of this paper is to re-estimate household consumption inequality in Mumbai, the most populous Indian city, following the method proposed by van der Weide et al. (2018). In addition, I will compare the resulting new figure to existing estimates based on survey data alone. In doing so, the paper also attempts to assess the feasibility of applying this method more broadly to inequality in urban India.

---

[1] World Development Indicators – annualized growth rate based on GDP in constant rupees 1993–2017.

[2] Poverty and inequality data in this introduction are from the World Bank PovcalNet database.

Following this method, I find that there is no indication that the National Sample Survey (NSS) data lead to underestimation of inequality in Mumbai. This stands in contrast with the empirical application to Egypt, where income inequality was estimated to be much higher than based on survey data alone, with Gini coefficients of 0.52 for the combined distribution and 0.39 for the survey-based estimate. Some challenges and practical obstacles arose during analysis. These issues need to be investigated further before we can assess if it is useful to apply the method to urban India as a whole.

The outline of this paper is as follows: the next section will explain the methodology used, also indicating where I deviate from the method proposed by van der Weide et al. (2018). Section 3 will give a brief overview of the two data sets used. In Section 4, I discuss the results of the empirical application to India. Finally, Section 5 provides concluding remarks and areas for further investigation.

2 Methodology

This section summarizes the approach by van der Weide et al. (2018), as applied in the current paper. I start by discussing the proposed distribution function of household income and its components, based on one database of actual observations (DB 1) and another of predictors (DB 2).[3] I will finish by describing how a new Gini coefficient can be estimated based on this combined distribution.

The complete income distribution can be characterized as follows:

$$F(y) = \begin{cases} (1-\lambda)\, F_1(y), & y \leq \tau \\ (1-\lambda) + \lambda\, F_2(y), & y > \tau \end{cases}$$

where household income is denoted by y and its cumulative distribution function by F(y). The income threshold for a top income is given as τ.[4] A household with income above τ has a 'top income'. The proportion of households with a top income is then described as λ. The cumulative distribution of their incomes is given by $F_2(y)$, which is derived from the predictors of income in DB 2 and starts at τ. In contrast, the cumulative distribution of households with an income of at most τ will be estimated based on the observations in DB 1 and denoted as $F_1(y)$.

One challenge of this approach is that DB 2 contains predictors of household income, not observations of the variable of interest itself. Van der Weide et al. (2018) propose a two-step approach that allows for estimation of $F_2(y)$. First, we need to describe the relationship between income and its predictor. A log-linear model is suited to that task, although the approach can be generalized to other functional forms. Thus we assume that the relationship can be given as:

$$\log(Y_h) = \beta_0 + \beta_1 \log(x_h) + \varepsilon_h$$

[3] In this section I will discuss how to obtain a combined distribution of income, following van der Weide et al. (2018). However, the available NSS data contain consumption expenditures, not income. Since consumption and income are expected to be highly correlated, the assumption is that the method can be applied, mutatis mutandis, to consumption expenditures as well.

[4] The value of tau obviously depends on the case at hand and can be set independently. I will return later to the implications of varying the value of tau for the result of the re-estimation procedure in this empirical application.

Y_h and x_h denote household income and a predictor of household income respectively, with ε_h being an error with expectation zero. β_1 is the parameter of interest, as it is used to transform the distribution of predictors into a distribution of income. It is estimated, however, by means of a regression using survey data from DB 1. In the application to Egypt and also for India, we are dealing with a database of house prices, although usually a household survey does not record the value of the house in which the household lives. Consequently, imputed rental values need to be used as predictor, which requires the additional assumption that these are proportional to actual house values.

Next we move to the top tail of the distribution of predictors, that is for values of predictor x larger than x_0, denoted by cumulative distribution function $G_2(x)$. It is assumed that this top tail follows a Pareto distribution that can be characterized as:

$$G_2(x) = 1 - \left(\frac{x}{x_0}\right)^{-\alpha} \text{ for } x > x_0$$

α being the Pareto tail index, also known as its shape parameter. Van der Weide et al. (2018) then show that this implies that also top incomes, i.e. those above τ, follow a Pareto distribution that can be written as

$$F_2(y) = 1 - \left(\frac{y}{\tau}\right)^{-\theta}$$

where the Pareto tail index θ is defined as α/β_1. This parameter is the crucial element in the re-estimation procedure, since it defines the thickness of the tail of the income distribution. A smaller θ indicates that there are more households in the tail, meaning more households with a top income. This in turn implies that inequality is higher compared to a distribution with a larger θ.

Having discussed how $F_2(y)$ can be estimated, the remaining unknown elements in our overall distribution function are $F_1(y)$ and λ. Since DB 1 contains actual observations of income, we simply use these to estimate $F_1(y)$ for incomes up to and including τ. In practice, it is quite possible that DB 1 will also include some observations above τ. These will be excluded from the estimation of $F_1(y)$. Finally, λ can be estimated by assuming that the probability density functions that correspond to $F_1(y)$ and $F_2(y)$ are continuous. The former are denoted by $f_1(y)$ and $f_2(y)$ respectively. Evaluating the estimated density functions at point τ gives the following estimator for λ:

$$\hat{\lambda} = \frac{\hat{f}_1(\tau)}{\hat{f}_1(\tau) + \hat{f}_2(\tau)}$$

The combined distribution function $F(y)$ consists of two non-overlapping sub-groups, which means we can decompose the Gini coefficient as follows (see for example Alvaredo, 2011):

$$\text{Gini} = P_1 S_1 \text{Gini}_1 + P_2 S_2 \text{Gini}_2 + S_2 - P_2$$

$$\text{Gini} = (1-\lambda)(1-s)\,\text{Gini}_1 + \lambda\, s\, \text{Gini}_2 + s - \lambda$$

P_k and S_k denote the population share and income share respectively of group k. Going back to our distribution function $F(y)$, P_2 is equal to λ and, conversely, P_1 equals $1 - \lambda$. If we define S_2 as s, S_1 equals $1 - s$. The income share of the top incomes, s, is estimated as:

$$s = \lambda \frac{\bar{y}_2}{\bar{y}}$$

where $\bar{y}$ denotes mean income of the overall distribution, and $\bar{y}_2$ the mean top income, given as:

$$\bar{y}_2 = \frac{\theta \tau}{\theta - 1}$$

which is the expected value of a Pareto distribution.

Gini_1 can be computed by using the survey data at hand. Gini_2 can be estimated on the basis of the shape parameter of the Pareto distribution, using the formula for the Gini coefficient of a Pareto distribution. This gives

$$\text{Gini}_2 = \frac{1}{2\theta - 1}$$

The paper continues in the next section by describing the data I have used for the empirical application of this method to India.

3 Data

This paper utilizes two databases. The first set of data, used to derive the distribution $F_1(y)$, is survey data collected by the Indian National Sample Survey Office. It contains data that is representative both on the national and district level. I use observations from the NSS Round 68, held in 2011/12, which are the most recent data available. The main variables of concern are monthly household consumption and imputed rent. In contrast to the survey setup, I include imputed rent in the consumption variable.[5]

For households in urban districts, the survey data include an estimate of imputed rent if the household lives in a property it owns. The value of imputed rent is an estimate that the enumerator and respondent arrive at by considering rental prices for similar properties in the neighborhood. I limit the analysis to Mumbai for practical reasons. The NSS contains 788 observations for the Mumbai district – for a subset of 499 households imputed rent was recorded.

The second database, on house prices in Mumbai, was obtained from public listings on the online platform Makaan in August 2018. The data set contains information on location, asking price, date of availability and type of property for 132,773 listings. A subset of these properties was not built yet, but was already offered for sale.

The two databases contain data collected at different points in time. I have not made a correction for temporal price differences, but assume instead that the Pareto tail index of the distribution of the highest house prices has not changed in the period 2011–18.

[5] For 103 households, both actual rent expenditures and imputed rent are recorded, one could imagine that the household rents an additional apartment next to the one it owns, or a parking space. In calculating total consumption expenditures, both categories of rent were included.

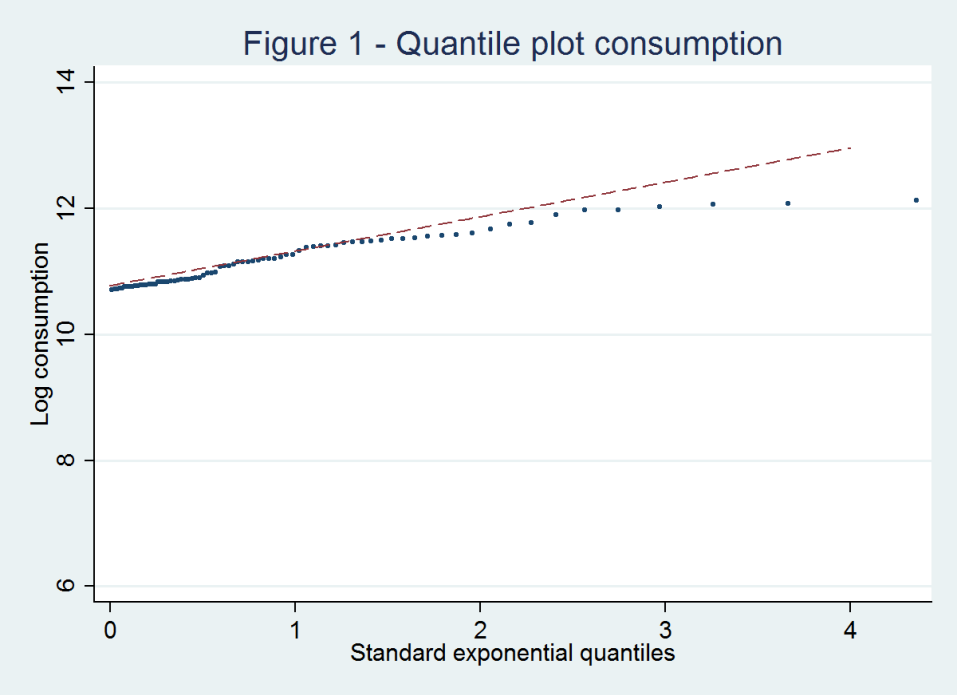
4 Application and results

Before I analyse how inequality estimates in India are influenced by the re-estimation method described above, I discuss a number of assumptions that are required. First, we assume that the quoted house prices in DB 2 are proportional to imputed rents. This is necessary because we will link the distribution of house prices to the distribution of consumption expenditures through a regression of consumption on imputed rents. Second, as noted above, we assume that the upper tail of the house prices distribution is constant over the period 2011/12 to 2018, which is when the data in DB 1 and DB 2 respectively were collected. A third assumption is that one house constitutes one household. It is quite possible that richer households own multiple properties, which would lead our method to underestimate inequality. Fourth, we also assume that all houses are domestically owned.

A *Distribution of survey data top consumption*

This subsection discusses the distribution of survey consumption found in DB 1. Although the method proposed by van der Weide et al. (2018) is generalizable to other types of distributions, the application to Egypt is based on (top) consumption and house prices being Pareto distributed. I verify whether this holds for Mumbai as well by plotting log consumption against the standard exponential quantiles ($-\log(1-p)$, where $p = F_1(y)$) using the top ten per cent of observations (Figure 1). We observe that the points are mostly close to the dashed line with slope parameter $1/\theta$, i.e. the inverse of the estimated Pareto tail index, except for the highest observations. This leads to the conclusion that top survey consumption is approximately Pareto distributed.

Figure 1: Quantile plot consumption



Notes: A plot of log consumption against $-log(1-F_1(y))$, y being survey consumption for Mumbai households. The dashed line has a slope equal to $1/\theta$.

Source: Author's illustration based on NSS Round 68.

Table 1 provides estimates of the Pareto tail index when using different numbers of top observations, starting from the 85th percentile upwards to the 95th percentile and up. The index seems reasonably stable, with an estimated value between 1.7 and 2.0 for most cases. As reference value I select the median value of these estimates, which gives a Pareto tail index of 1.855 for the upper part of the survey-based distribution of consumption expenditures. This value will be used to compare the re-estimated parameter against.

Table 1: Tail index survey consumption Mumbai

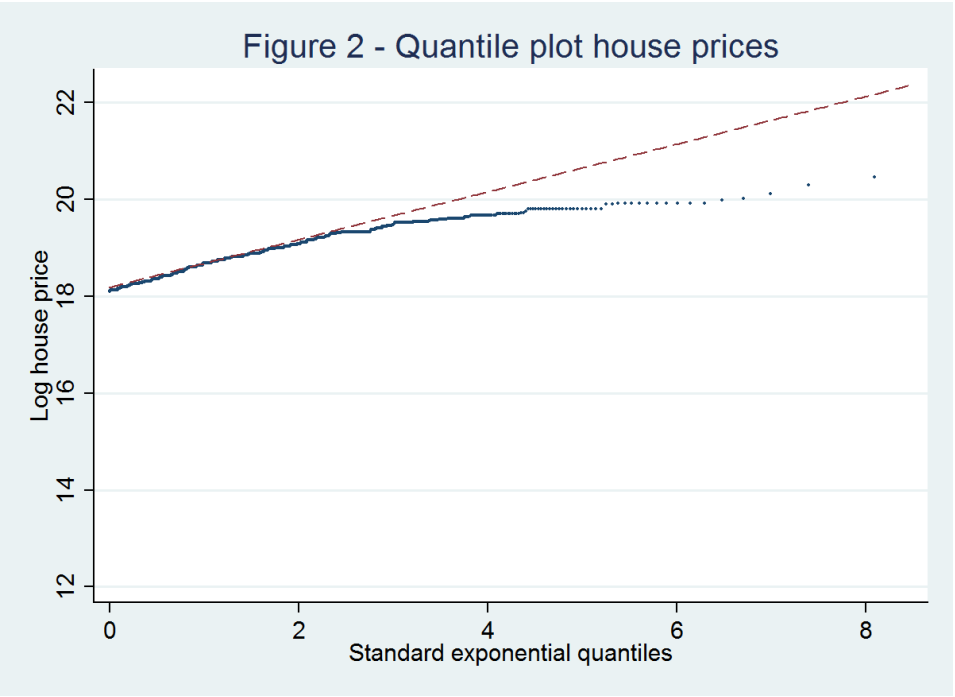
percentile	tail_index	st_error	N
85	1.970	0.218	118
86	1.895	0.207	110
87	1.942	0.224	102
88	1.766	0.184	94
89	1.855	0.209	86
90	1.831	0.210	78
91	1.865	0.225	72
92	1.698	0.186	63
93	1.747	0.201	55
94	1.788	0.196	47
95	2.159	0.285	39

Source: Author's calculation based on NSS Round 68.

B Distribution of house price data at the top

Next, I examine the distribution of house prices in DB 2. Analogous to the consumption observations, Figure 2 shows a plot of the log of house prices against standard exponential quantiles, now for the 2.5 per cent highest observations. We observe that linearity holds for the lower end, but that the observations toward the higher end depart further from the dashed line with slope parameter $1/\alpha$. This implies that, compared to the Pareto distribution, there are fewer extremely expensive houses in the tail of the actual distribution, i.e. the actual tail is thinner. This may lead us to underestimate the Pareto tail index, which would then lead to a higher inequality estimate. Nevertheless, the majority of observations lies reasonably close to the dashed line.

Figure 2: Quantile plot house prices

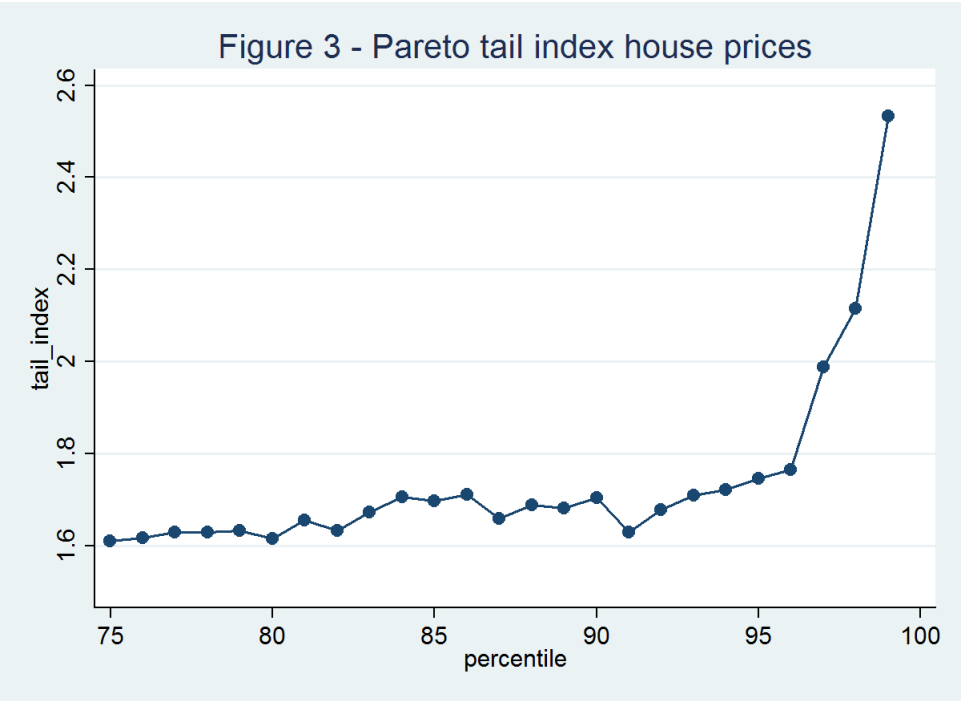


Notes: A plot of log house price against $-log(1-CDF(x))$, x being house price. The dashed line has a slope equal to $1/\alpha$.

Source: Author's illustration based on house price data from Makaan.

Figure 3 graphs the values of the estimated Pareto tail index based on different subsamples, from the 75th to the 99th percentile and up. It shows that the estimates do not vary much between the 75th and 96th percentile, remaining in between 1.6 and 1.8. For the most expensive houses, the value of the tail index shoots up, indicating that the tail is getting thinner. As input for the re-estimation exercise, I take the median value for the 75th to 96th percentile estimates, which is a Pareto tail index $\hat{\alpha}$ of 1.676.

Figure 3: Pareto tail index house prices



Notes: Estimated Pareto tail index using different upper parts of the house price sample.

Source: Author's compilation based on house price data from Makaan.

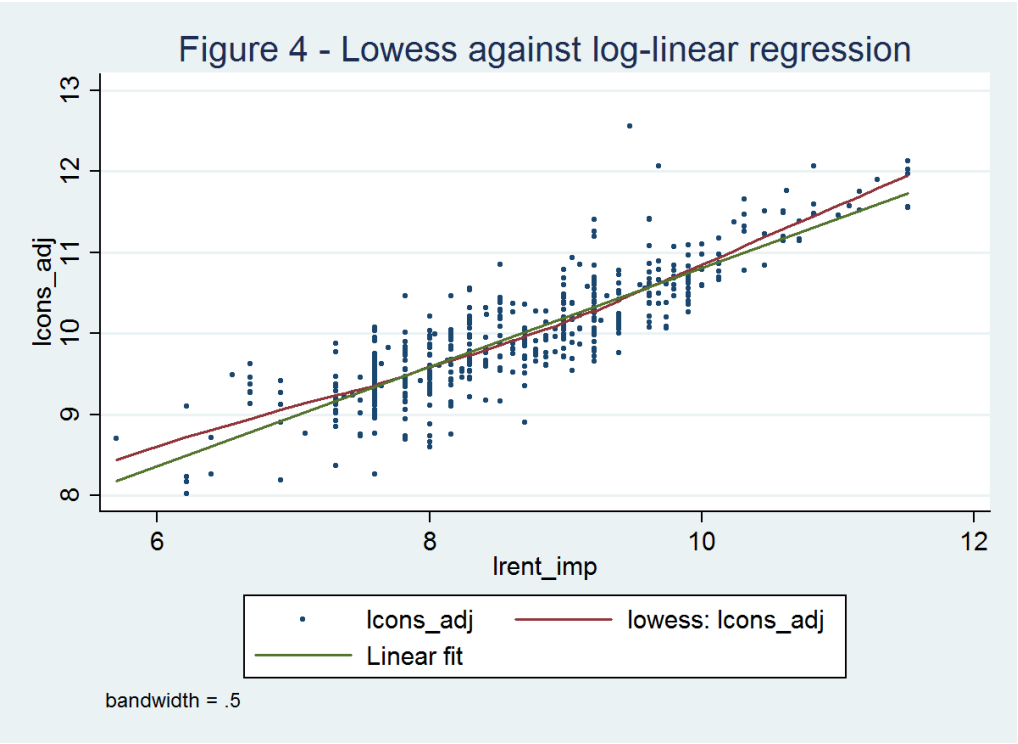
This estimate of α is robust to what upper part of the sample is used for estimation. Nevertheless, other factors may have a large effect on the estimate. One such factor is the selection of neighbourhoods included. Another is whether properties that still need to be built are included in the estimate. The current estimate includes all subdivisions of Mumbai and both readily available as well as future dwellings. This is an important area for further robustness checks.

C The relationship between consumption and rent

This subsection investigates the relationship between consumption expenditures and imputed rent, which is necessary to link the two distributions discussed above. First, I examine the functional form of the relationship between these two variables. I perform a locally-weighted regression of log consumption on log rent, which gives the curve displayed in Figure 4.[6] In addition, Figure 4 plots the observations and a line of linear fit obtained through a univariate OLS regression. For the full sample, the line with slope $\hat{\beta}_1$ derived from the linear regression is close to the locally-weighted regression. I conclude that the log-linear specification is adequate to predict consumption for the richer households.

[6] The bandwidth of the lowess regression is 0.5.

Figure 4: Lowess against log-linear regression

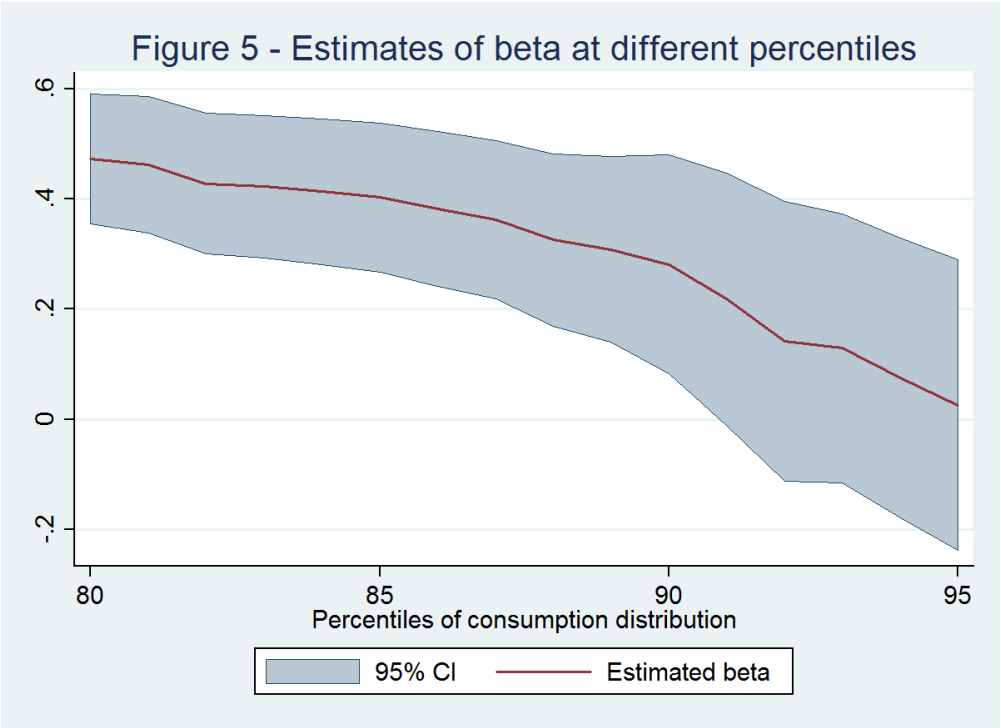


Notes: Scatter plot and fitted values of regressions of log consumption on log imputed rent for Mumbai households.

Source: Author's compilation based on NSS Round 68.

However, further analysis brings to light one complication, as illustrated by Figure 5. That figure shows the ordinary least squares (OLS) estimator of the log-linear regression based on different parts of the distribution, from the 80th percentile and up to the 95th percentile upwards. The shaded area gives the 95 per cent confidence interval, which is widening towards the upper end. We observe that the estimator tends towards zero if fewer observations are used. This also holds if a subset of observations from the middle of the distribution is used. At the same time, the precision of the estimate and the R^2 of the regression decrease fast.

Figure 5: Estimates of beta at different percentiles



Notes: Estimated values of β_1 from a OLS regression of log consumption on log imputed rent, based on increasingly fewer observations, starting at the 80th percentile and up.

Source: Author's illustration based on NSS Round 68.

Another problem is that the locally-weighted regression seems to indicate that the slope parameter is increasing towards the higher end of the consumption distribution. This also makes sense in practice: a one per cent increase in rent is associated with larger increases in consumption for higher amounts of rent paid. Turning this around, it means that—let's say at the 80th percentile of consumption—the amount of rent paid will increase more slowly than at the median. However, as Figure 5 shows, the OLS estimator becomes smaller when using fewer and fewer top observations.

This seems to be in contradiction and may indicate that the OLS regression suffers from attenuation bias. One possible explanation is that this is caused by disturbances in the measurement of the independent variable, imputed rent. During the survey that variable is estimated by the enumerator and respondent together. They compare the house of the respondent to similar rental properties in the neighbourhood and attribute a rental value based on that comparison.

This obviously requires further investigation. For the current purpose, however, I will stick to the log-linear regression, in strict keeping with the approach by van der Weide et al. (2018). To illustrate to what extent the final re-estimates of inequality depend on the value of β_1, I will compare two estimates: 1) an estimate obtained by using only the top 20 per cent of observations, which gives $\hat{\beta}_1 = 0.472$; and 2) an estimate based on the full sample, which yields $\hat{\beta}_1 = 0.614$.

As pointed out above, further analyses are needed regarding the specification of the relationship between consumption and rent. One option is to swap the dependent and independent variable, as this may prevent the attenuation bias.

D Distribution of top consumption based on house prices

The estimation procedures discussed above yield the inputs needed to characterize the top end of the consumption distribution on the basis of the predictors in DB 2. These parameters, $\hat{\beta}_1$ and $\hat{\alpha}$, are displayed in Table 1. Together, they give rise to $\hat{\theta}_{mix}$ (defined as $\hat{\alpha}/\hat{\beta}_1$), the Pareto tail index of the distribution of top incomes. We can compare this estimate to the index that we find for observed survey consumption, given in the table by $\hat{\theta}_{svy}$. For both estimates of β_1, $\hat{\theta}_{mix}$ is larger than $\hat{\theta}_{svy}$, indicating a thinner tail for the top end of the distribution. This already provides some indication that the re-estimated inequality indices will not be affected much by adding top incomes to the survey-based distribution.

Table 1 - Estimated parameters used in the re-estimation

β_1 estimated with:	$\hat{\beta}_1$	$\hat{\alpha}$	$\hat{\theta}_{svy}$	$\hat{\theta}_{mix}$
Full sample	0.614	1.676	1.855	2.730
Upper 20%	0.472	1.676	1.855	3.548

Notes: These are $\hat{\beta}_1$ from an OLS regression of log consumption on log rent, the Pareto tail index of the top house price distribution ($\widehat{\alpha}$) and the Pareto tail index for the top income distribution ($\hat{\theta}$). The latter is given both for the survey-based estimate and the combined distribution.

Source: Author's calculations based on NSS Round 68 and house prices Makaan.

E Re-estimating inequality in Mumbai

After having obtained estimates for α, β_1 and θ, the starting point for further analysis is selecting an appropriate value of τ, the threshold for a top income. A practical choice is to select one of the highest percentile values in the consumption survey. As a consequence, the observations in the survey above this value are excluded from the overall distribution, because, by construction, the distribution for consumption expenditures exceeding τ is given by $F_2(y)$, which is based on the distribution of house prices.

Since we started from the assumption that the household survey is missing the richest households, it makes sense to focus on only the highest three percentiles.[7] For comparison, however, I present estimates for the 95th to 99th percentiles, separately for the two values of $\hat{\beta}_1$ estimated earlier. Utilizing the lower value of $\hat{\beta}_1$, Table 2 gives two estimates each for the consumption share of the top incomes (s) and the Gini coefficient – one set based on the survey only, and another set based on our re-estimated overall distribution. Table 3 does the same for the higher value of $\hat{\beta}_1$, which was estimated above using the full sample of survey observations. The tables also provide an estimate for λ, the proportion of households with a top income. Obviously, its value depends on the value selected for τ. In addition, we need the probability density functions of $F_1(y)$ and $F_2(y)$ to estimate λ. For $f_1(y)$, I considered two methods, kernel density estimation and fitting a lognormal distribution. However, the kernel density estimates do not produce a smooth, non-increasing distribution of λ, while the estimates based on the lognormal distribution do. Another advantage of the lognormal distribution is that it uses all observations in the dataset to estimate the parameters needed to estimate λ. Therefore, I am presenting results obtained on the basis of a

[7] Theoretically, we would expect that none of the top consumption households is present in the survey. This is perhaps too strict an assumption, so I focus on the 97th, 98th and 99th percentile value to set tau.

fitted lognormal distribution of the survey data.[8] For the estimation of $f_2(y)$, I made use of the probability density function of the estimated Pareto distribution.

We observe that all ten re-estimated Gini coefficients, for both estimated β_1, are below the survey-based Gini of 0.447.[9] The differences between the estimated consumption shares are rather large, notably for the lower $\hat{\beta}_1$. The smallest difference is observed when I set τ to the consumption value observed at the 99th percentile and use the higher $\hat{\beta}_1$.[10] Here, the difference between the two estimated income shares of the top incomes is around five per cent, which implies a re-estimated Gini that is only about two and a half percentage points lower.[11]

Table 2 - Estimates of income share and Gini (for $\hat{\beta}_1 = 0.472$).

perc_svy	τ	$\hat{\lambda}$	s_svy	s_mix	Gini_svy	Gini_mix
95	69,161	0.020	0.229	0.091	0.447	0.386
96	75,802	0.016	0.198	0.078	0.447	0.391
97	91,592	0.009	0.162	0.054	0.447	0.396
98	101,744	0.007	0.120	0.042	0.447	0.407
99	148,183	0.002	0.070	0.016	0.447	0.418

Notes: τ and perc_svy give the threshold value for a household top income (in INR per month) at the survey percentile at which it was obtained. $\hat{\lambda}$ gives the estimated population share of households with a top income, while s gives the corresponding consumption share and Gini the Gini coefficient. The latter two measures are given both based on the survey alone and on the combined distribution.

Source: Author's calculations based on NSS Round 68 and house prices Makaan.

Table 3 - Estimates of income share and Gini (for $\hat{\beta}_1 = 0.614$)

perc_svy	τ	$\hat{\lambda}$	s_svy	s_mix	Gini_svy	Gini_mix
95	69,161	0.025	0.229	0.129	0.447	0.404
96	75,802	0.020	0.198	0.110	0.447	0.407
97	91,592	0.012	0.162	0.077	0.447	0.407
98	101,744	0.008	0.120	0.061	0.447	0.416
99	148,183	0.002	0.070	0.023	0.447	0.421

Notes: τ and perc_svy give the threshold value for a household top income (in INR per month) at the survey percentile at which it was obtained. $\hat{\lambda}$ gives the estimated population share of households with a top income, while s gives the corresponding consumption share and Gini the Gini coefficient. The latter two measures are given both based on the survey alone and on the combined distribution.

Source: Author's calculations based on NSS Round 68 and house prices Makaan.

[8] Estimating λ using kernel density estimation did generally produce higher estimates for the Gini coefficient. However, given the erratic behaviour of the estimated λ, I consider the estimates based on the lognormal distribution more reliable.

[9] The standard error of this estimated Gini coefficient is 0.020 based on a bootstrap method, and 0.029 when using a linearization method.

[10] The 99th percentile is a monthly household consumption expenditure of 148,183 Indian rupees, which is equivalent to around USD 7500 in purchasing power parity terms.

[11] Further investigation is needed into the precision of these results, for example using bootstrapping methods. Van der Weide et al. (2018) discuss the precision of the estimates obtained for Egypt by investigating what happens to the estimates if the most conservative estimates for α and β_1 are used.

In consequence, this re-estimation method does not indicate that the survey-based Gini is underestimating consumption inequality. Note, however, that the results are sensitive to a number of choices. Hence, they depend on what value of τ is chosen, on what method is selected to estimate λ and on which subset of the survey sample is taken to estimate the relationship between consumption and imputed rent. However, there is no set selection procedure. In the application to Egypt, for example, it is not clear how the value of τ was selected.

This finding for Mumbai stands in contrast with the empirical application to Egypt, where income inequality was estimated to be much higher than based on survey data alone, with Gini coefficients of 0.52 for the combined distribution and 0.39 for the survey-based estimate.

5 Conclusion

This paper has applied a novel method that could be used for the re-estimation of inequality in India, given the discussion around prevailing estimates. Combining survey data and information about the distribution of house prices, the method represents an alternative to survey-only or tax data-based estimates. Nonetheless, for the district of Mumbai I find that this method does not indicate that the survey-based Gini coefficient of household consumption is underestimated. Employing ten different scenarios, none of the re-estimated Gini coefficients is higher than the survey-only estimate of 0.447.

Nevertheless, the application of this method to Mumbai brought up a number of challenges and entailed making a number of practical decisions not all of which were encountered in the original application to Egypt (van der Weide et al., 2018). These require further investigation. First, the method hinges on being able to determine a good estimator for the relationship between consumption or income and (imputed) rent. That was not straightforward in this case. The log-linear relationship between consumption and rent is sensitive to the number of top observations included in the regression. This may be due to the nature of the imputed rent variable, which enumerator and interviewee estimate together based on a comparison with rental properties in the neighbourhood. The regression estimator may suffer from attenuation bias due to too much noise in this imputed rent variable. This requires further investigation into the regression specification. Second, the distribution of house prices warrants further research. Subsets of house prices, based for instance on availability or neighbourhoods, may follow a different distribution. Third, there is no clear criterion to set the threshold value for a top income. The choice of such threshold influences the final results of the re-estimation exercise. As a practical choice, I selected the highest percentile values observed in the survey data.

With these caveats in mind, the conclusion remains that there is no indication that the NSS survey data underestimate household consumption inequality in Mumbai. Further analysis is necessary before a decision can be made regarding the usefulness of applying this method to re-estimate inequality in urban India as a whole.

References

Alvaredo, F. (2011). 'A note on the relationship between top income shares and the Gini coefficient'. *Economics Letters*, 110, number 3, 274–77.

Chancel, L., and T. Piketty (2017). 'Indian income inequality, 1922–2014: From British Raj to Billionaire Raj?' *WID.world Working Paper Series* No 2017/11. Paris: World Inequality Database.

Himanshu (2015). 'Inequality in India'. *Seminar*, Issue 672 August 2015, 30–35. New Delhi: Seminar Publications.

Korinek, A., J. Mistiaen, and M. Ravallion (2006). 'Survey nonresponse and the distribution of income'. *Journal of Economic Inequality*. 4, 33–55.

Makaan (2018). Dataset compiled by author from www.makaan.com (accessed on 14 August 2018).

National Sample Survey Office (2013). India - Household Consumer Expenditure, NSS 68th Round Sch 1.0 Type 1: July 2011 – June 2012. New Delhi NSSO, Ministry of Statistics and Programme Implementation, Government of India.

PovcalNet (2018). Online analysis tool for global poverty monitoring. Washington, DC: World Bank. Available at http://iresearch.worldbank.org/PovcalNet/home.aspx (accessed on 29 November 2018).

van der Weide, R., C. Lakner, and E. Ianchovichina (2018). 'Is Inequality Underestimated in Egypt? Evidence from House Prices'. *Review of Income and Wealth*. Vol 64, Issue s1, S55–S79.

World Development Indicators (2018). Dataset. Washington, DC: World Bank. Available at https://datacatalog.worldbank.org/dataset/world-development-indicators (accessed on 29 November 2018).