WPS7727

Policy Research Working Paper 7727

Is Inequality Underestimated in Egypt?

Evidence from House Prices

Roy van der Weide
Christoph Lakner
Elena Ianchovichina

WORLD BANK GROUP
Development Research Group
Poverty and Inequality Team
&
Middle East and North Africa Region
Office of the Chief Economist
June 2016

Policy Research Working Paper 7727

Abstract

Household income surveys often fail to capture top incomes which leads to an underestimation of income inequality. A popular solution is to combine the household survey with data from income tax records, which has been found to result in significant upward corrections of inequality estimates. Unfortunately, tax records are unavailable in many countries, including most of the developing world. In the absence of data from tax records, this study explores the feasibility of using data on house prices to estimate the top tail of the income distribution. In an application to Egypt, where estimates of inequality based on household surveys alone are low by international standards, the study finds strong evidence that inequality is indeed being underestimated by a considerable margin. The Gini index for urban Egypt is found to increase from 36 to 47 after correcting for the missing top tail.

This paper is a product of the Poverty and Inequality Team, Development Research Group and the Office of the Chief Economist, Middle East and North Africa Region. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at rvanderweide@worldbank.org.

Is Inequality Underestimated in Egypt? Evidence from House Prices

Roy van der Weide, Christoph Lakner and Elena Ianchovichina*

JEL classification: D31.

Keywords: inequality; top incomes; house prices; Egypt.

*All authors are with the World Bank. Contact information: rvanderweide@worldbank.org, clakner@worldbank.org and eianchovichina@worldbank.org. This is a background paper for the report entitled "Inequality, Uprisings, and Conflict in the Arab World" led by the World Bank's Chief Economist Office for the Middle East and North Africa region. The authors wish to thank Guoliang Feng and Youssouf Kiendrebeogo for excellent research assistance. We would like to thank Facundo Alvaredo, Francisco Ferreira, Nadine Ghobrial, Vladimir Hlasny, Aart Kraay, Peter Lanjouw, Branko Milanovic, Thomas Piketty, Martin Ravallion, Paolo Verme and participants of the World Bank workshop on the Arab Inequality Puzzle and the IARIW-CAPMAS Conference "Experiences and Challenges in Measuring Income, Wealth, Poverty and Inequality in the Middle East and North Africa" for useful comments. We are grateful to the UK Department for International Development for financial assistance through its Strategic Research Program.

1 Introduction

Estimates of income inequality are conventionally derived from household income and expenditure surveys. Due to the sizeable cost of collecting accurate data on household standards of living the sample size of these surveys generally constitutes less than half a percent of the total population. Unfortunately, the rich are often missing or under-covered, either due to non-response or under-reporting of income or both; see the recent literature on top income shares (e.g. Atkinson et al., 2011). Surveys still permit accurate estimation of median income and measures of poverty, even when data on top incomes are poor or are missing altogether, since the rich make up a small percentage of the total population. For the estimation of income inequality however, having good data on top incomes is crucial.

A remedy that has gained considerable traction recently is to estimate the top tail of the income distribution using data from income tax records. This estimate of the top tail can then be combined with an estimate of the bottom part from the household survey to obtain an estimate of the complete income distribution (Atkinson, 2007; Alvaredo, 2011; Alvaredo and Londoño Vélez, 2013; Diaz-Bazan, 2014; Anand and Segal, 2015).[1] Income tax records denote the ideal source of data as far as top incomes are concerned. For lower incomes tax records may be less reliable, here the household income survey arguably denotes the ideal data. When household survey and tax data are combined in this way, the Gini index for (i) the United States in 2006 increases from 59 to 62 (Alvaredo, 2011), (ii) Colombia in 2010 from 55 to 59 (Alvaredo and Londoño Vélez, 2013), and (iii) Korea in 2010 from 31 to 37 (Kim and Kim, 2013).

For all the pros of income tax records, the availability of the data is unfortunately rather limited, particularly in developing and emerging economies. The World Top Incomes Database (Alvaredo et al., 2015) for example includes no countries from the Middle East and North Africa region. Furthermore, data derived from tax records are less useful in places where tax evasion is more pervasive, as is the case in many developing countries. It should also be noted that combining household survey data and tax records is not without complications because the two data sources use different income definitions (disposable versus taxable) and have different units of analysis (households versus tax units, which could be individuals).

In the absence of data from tax records, this study explores the feasibility of using data on house prices to estimate the top tail of the income distribution.

[1] Diaz-Bazan (2014) generalizes the method of Atkinson (2007) and Alvaredo (2011) by allowing for a more general choice of the cut-off level for joining up the distributions. Morelli et al. (2015) review the literature attempting to combine household surveys and tax data in rich countries.

Market house price data can often be obtained more easily and, most importantly, tend to be available in the public domain, in contrast to tax administration data which are subject to important confidentiality concerns and require cooperation from governments. Also, house sellers have no incentive to understate the value of their homes, in contrast to the income they report on their tax returns.

Using house prices as an alternative to income tax records demands two methodological innovations to the study of top incomes. Firstly, we will not be observing actual household income or expenditure (as is the case with tax record data), but rather a predictor of income. Secondly, a database with house price listings is generally not obtained using a particular sampling design. Therefore, the data are not guaranteed to provide a nationally representative sample, they will arguably be biased towards large urban centers. We will propose workable solutions to both these challenges that will hopefully contribute to a wider use of this approach. Note that the methodology is not restricted to the use of house prices, it can be applied to any database containing predictors of top incomes.

We illustrate our approach with an empirical application to Egypt which provides a good testing ground for our method. In addition to being a major Arab country, inequality in Egypt is of considerable interest not least because it has been cited as one of the factors behind the Egyptian revolution (Hlasny and Verme, 2013). Estimates of inequality based on household surveys suggest that inequality is low in Egypt and that it has declined in the last decade to a Gini of around 31 in 2009. Using house prices to capture top incomes we find that inequality may be significantly underestimated in Egypt. The Gini for urban Egypt in 2009 is estimated at 47.0 compared to a survey-only estimate of 36.4. Our results are in contrast with other studies using different methods of adjusting for top incomes in Egypt (Hlasny and Verme, 2013), which report a more modest effect.[2] Their correction however does not consult a second source of data. If the main problem is that high income earners are simply missing from the survey, then no adjustment that relies soley on the survey will resolve the downward bias in estimates of inequality. The only way to obtain a meaningful correction is to bring in a second source of data that carries the necessary information on top incomes and hence will permit for the consistent estimation of income inequality. This reasoning is shared by Alvaredo and Piketty (2014) who similarly argue that the household survey data by themselves are insufficient to estimate top incomes in Egypt. While they make an appeal for making data on income tax records available, we propose to work with house price data instead. It should be noted that relying on predictors of top incomes rather than actual incomes derived from

[2]The Gini coefficient of household expenditure per capita in 2009 increases from 30.5 to 31.8 which is found to be statistically significant, but not economically significant.

tax records is not without caveats. For example, we need to make assumptions about the functional form of the relationship between the house price and household income, and about the functional form of the upper tail of the house price distribution. In addition it is assumed that one house constitutes one household and that all houses are domestically owned. Therefore, in cases where tax record data are available these should undoubtedly be considered first. However, we certainly believe that our approach provides more reliable estimates of inequality than estimates obtained using survey data alone. The perfect should not be the enemy of the good.

This paper is related to a number of other studies which have tried to correct household surveys for the problem of missing or underreported top incomes.[3] Korinek et al. (2006) exploit geographic variation in response rates to correct for selective non-response in the United States. Lakner and Milanovic (2015) exploit the gap between household surveys and national accounts to adjust the top end of the income distribution.[4]

This paper is structured as follows. The methodology is presented in Section 2. In Section 3 we introduce the data used in the empirical application to Egypt. The empirical application itself is presented in Section 4. Finally, Section 5 concludes.

2 Methodology

2.1 Combining income survey with top income data

The objective is to estimate the level of income inequality for a given population. We will refer to database 1 (DB-1) as the primary data source for the estimation of inequality. It is assumed that top incomes are mostly missing from this database. Database 2 (DB-2), which we will refer to as the secondary data source, primarily contains data on top incomes but generally not on lower incomes. Estimates of income inequality will be biased if computed using any single one of these databases. It takes a combination of the two to obtain consistent estimates of inequality. DB-1 commonly represents a household income or expenditure survey. For DB-2 researchers often look at tax record data, as is discussed in the introduction.

[3]Recently, the EU-SILC survey in some countries began using register-based information (including tax records) for some questions (Jäntti et al, 2013). This is of course preferable to any ex-post combination of these different data sources, as we use in this paper. In the year after the introduction of the register data, the Gini index for France increased from 39 to 44, which is consistent with the previously used household data underestimating top incomes (Burricand, 2013).

[4]See also the study on global interpersonal inequality by Anand and Segal (2015) who append for every country the estimated top 1% share to the household survey distribution. The latter is assumed to represent the bottom 99%. For the majority of countries, the top 1% share is predicted from a cross-country regression using the top 10% share in the household survey.

Let us denote household income by y and its cumulative distribution function by $F(y)$. Let τ denote the income threshold above which we will refer to incomes as "top incomes", and let λ measure the share of the population enjoying a top income, i.e. $\lambda = Pr[Y > \tau] = 1 - F(\tau)$. It is assumed that DB-1 permits a consistent estimator for $F_1(y) = Pr[Y \leq y | Y \leq \tau]$, and that DB-2 permits a consistent estimator for $F_2(y) = Pr[Y \leq y | Y > \tau]$. By the same token it is assumed that DB-1 does not permit a consistent estimator for $F_2(y)$, while DB-2 does not permit a consistent estimator for $F_1(y)$. Suppose also that an estimate of λ is available.[5] Given estimates of $F_1(y)$, $F_2(y)$ and λ, an estimator for the complete income distribution function $F(y)$ can be obtained as follows:

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

Given $F(y)$, any measure of income inequality can readily be computed. Alternatively, one may appeal to the sub-group decomposition of one's inequality measure of choice, which would by-pass the need for evaluating the income distribution for the population ($F(y)$). We have two sub-groups, those with income below τ (sub-group 1) and those with income above τ (sub-group 2). Let P_k denote the population share of sub-group k, and let S_k denote their corresponding income shares, i.e. $S_k = P_k \mu_k / \mu$, where μ_k and μ measure average income in sub-group k and the total population, respectively. Note that $P_1 = 1 - \lambda$ and $P_2 = \lambda$. Let us also define $S_1 = 1 - s$ and by extension $S_2 = s$. It can be verified that income inequality as measured by the Gini coefficient satisfies the following decomposition (see e.g. Alvaredo, 2011):

$$\begin{aligned} Gini &= P_1 S_1 Gini_1 + P_2 S_2 Gini_2 + S_2 - P_2 \\ &= (1-\lambda)(1-s)Gini_1 + \lambda s Gini_2 + s - \lambda, \end{aligned}$$

where $Gini_k$ measures the Gini coefficient for population sub-group k. A similar decomposition can be obtained for the mean-log-deviation MLD (see e.g. Shorrocks, 1980):

$$\begin{aligned} MLD &= P_1 MLD_1 + P_2 MLD_2 + P_1 \log\left(\frac{P_1}{S_1}\right) + P_2 \log\left(\frac{P_2}{S_2}\right) && (2) \\ &= (1-\lambda)MLD_1 + \lambda MLD_2 + (1-\lambda)\log\left(\frac{\mu}{\mu_1}\right) + \lambda \log\left(\frac{\mu}{\mu_2}\right) && (3) \\ &= (1-\lambda)MLD_1 + \lambda MLD_2 + \log(\mu) - \log\left(\mu_1^{1-\lambda}\mu_2^{\lambda}\right), && (4) \end{aligned}$$

[5] It is generally assumed that DB-2 contains the total number of units (i.e. households or tax units) whose income is above τ. Combined with the total population this yields an estimator for λ.

and for the Theil index T (see e.g. Shorrocks, 1980):

$$\begin{aligned} T &= S_1 T_1 + S_2 T_2 + S_1 \log\left(\frac{S_1}{P_1}\right) + S_2 \log\left(\frac{S_2}{P_2}\right) && (5) \\ &= (1-s)T_1 + sT_2 + (1-s)\log\left(\frac{\mu_1}{\mu}\right) + s\log\left(\frac{\mu_2}{\mu}\right) && (6) \\ &= (1-s)T_1 + sT_2 + \log\left(\mu_1^{1-s}\mu_2^{s}\right) - \log(\mu), && (7) \end{aligned}$$

where MLD_k and T_k measure the mean-log-deviation and Theil index for population sub-group k, respectively. Note that the between-group inequality components of both the mean-log-deviation and the Theil index equal the difference between the arithmetic- and the geometric mean income levels. They differ only in the weights used in the geometric mean; the mean-log-deviation weighs the sub-group means by their population shares whereas the Theil index weighs them by their incomes shares.

An inspection of the three sub-group decompositions tells us that the Theil index will be most sensitive to the top tail of the income distribution.[6] To illustrate the significance of the top tail to total inequality consider the limit where the population share of top income earners tends to zero ($\lambda \to 0$) while their income share tends to some positive value ($s > 0$). It can readily be seen that the between-group inequality component of the Gini coefficient tends to $s > 0$ in that case, while the within-group inequality among top income earners tends to zero, i.e. $G \to (1-s)Gini_1 + s$. It follows that the between-group inequality component for the mean-log-deviation tends to $\log(1-s)^{-1}$, while also here (as with the Gini) the within-group inequality among top earners tends to zero (yet it does not discount the contribution of within-group inequality among non-top earners), i.e. $MLD \to MLD_1 - \log(1-s)$. The Theil index stands out as the only of the three inequality measures where the within-group inequality among top earners does not vanish (i.e. makes a positive contribution to total inequality) while the between-group inequality component will tend to infinity (when μ_2 tends to infinity as $\lambda \to 0$ while $s > 0$).

2.2 An alternative to top income data: Challenges

As stated in the previous section, DB-2 (the top income database) typically takes the form of tax record data. These data have at least two advantages: (1) it directly observes realized incomes (which makes the estimation of $F_2(y)$ or any income statistics such as inequality among top earners rather straightforward), and (2) it provides a count of the number of top income earners, which makes for

[6] Hence it is expected that any efforts made to fix the top tail of the income distribution by bringing in complementary data (top income database) will be rewarded the most by the Theil index.

a straightforward estimation of λ. A key disadvantage of tax record data is that it is often difficult to obtain access to them. Moreover, they are more likely to be available in developed countries with good quality data systems in place, and less likely to be available in developing countries.

This paper explores the feasibility of using an alternative to tax record data that are more readily available. The empirical application presented in Section 4 considers data on house prices compiled from publicly available real estate property listings as the alternative.[7] The advantage of these data is that their availability extends to developing countries. The flip-side is that they also introduce a number of key methodological challenges due to the fact that the alternative database (a) observes predictors of income, not actual incomes, and (b) need not constitute a proper sample, so that it is unclear what population is being represented by the data.

The following two subsections aim to provide workable solutions to these two challenges that will hopefully contribute to a wider application of this approach.

2.2.1 A database of predictors of top incomes

Let us first focus on the challenge posed by observing a predictor of household income rather than actual income. Consider the following assumption.

Assumption 1 *Suppose that household income per capita can be described by:*

$$\log(Y_h) = m(x_h;\beta) + \varepsilon_h \tag{8}$$

$$= \beta_0 + \beta_1 \log(x_h) + \varepsilon_h, \tag{9}$$

where x_h denotes the predictor of household income, ε_h denotes a zero expectation error term, subscript h indicates the household, and where β denotes a vector of model parameters.

The assumption of a log-linear model is motivated by ease of exposition and by the fact that it fits our empirical data remarkably well. This assumption can be relaxed however by accommodating flexible functional forms for $m(x_h;\beta)$ if the data call for it. In our application the value of the household's house (or rental value) will serve as the predictor x_h.

To obtain some intuition for the implications of Assumption 1 it may be helpful to verify what it implies for the relationship between the expenditure share on housing and household income. Note however that we are concerned with predicting household income per capita rather than household income. (In

[7] Alternatively one could for example also look to data on mortgages or credit card statements. However, this approach may not be feasible in countries with underdeveloped or non-existing mortgage markets.

fact our dependent variable denotes household expenditure per capita; we will get back to this.) Let us abstract away from this distinction, for this thought experiment only, by considering the household size fixed (so that it is absorbed in the constant β_0). It can be verified that the assumed functional form implies that the expenditure share on housing is a convex declining function of income when $\beta_1 > 1$. The expenditure share is constant for $\beta_1 = 1$ and an increasing function of income for $0 < \beta_1 < 1$. More specifically, it is a concave increasing function for $\beta_1 \in (\frac{1}{2}, 1)$, a linear increasing function for $\beta_1 = \frac{1}{2}$, and a convex increasing function for $\beta_1 \in (0, \frac{1}{2})$. Despite its simplicity, the log-linear assumption permits a reasonable degree of flexibility in how the expenditure share on housing varies with income. Our prior would be that $\beta_1 > 1$, which is consistent with the empirical evidence that is available for the Engel curve on housing expenditure, see e.g. Larsen (2014). $\beta_1 > 1$ also ensures that the expenditure share stays below 1 when incomes tend to extreme values. It is unclear however whether this carries over when we substitute household expenditure for household income. We could not find any empirical study that investigates how expenditure on housing varies with total household expenditure. It is conceivable that one might find values of β_1 that are below 1. We would however still expect it to be above $\frac{1}{2}$. (Note that the model in that case does not rule out expenditure shares above 1 for extreme household expenditure values.)

Let $F_\varepsilon(e;\sigma)$ denote the distribution function of ε_h with unknown parameter vector σ. We will assume that ε_h is identically distributed across households, although this assumption can easily be relaxed. Note that the unknown parameter vectors β and σ both have to be estimated. In our empirical application, where the value of housing is considered as a predictor of income, the two can be estimated using the household income survey, since it includes both data on household incomes and data on the value of housing.

It will be convenient to define $n(\tau, y)$ as the number of households with income between τ and y, $n(\tau)$ as the number of households with income exceeding τ, and n as the total number of households in the population. For ease of exposition we will ignore the fact that the data may constitute a sample with sampling weights. $F_2(y)$ $(= Pr[Y \leq y | Y > \tau])$ and λ $(= Pr[Y > \tau])$ are seen to solve:

$$F_2(y) = \frac{n(\tau, y)}{n(\tau)} \tag{10}$$

$$\lambda = \frac{n(\tau)}{n}. \tag{11}$$

When DB-2 does not contain data on household incomes but data on a predictor of household incomes instead, we have that $n(\tau, y)$ and $n(\tau)$ can no longer be observed with certainty and so have to be estimated. Consider first an estimator

for $n(\tau)$:

$$\begin{aligned}
\hat{n}(\tau) &= \sum_h E[1(Y_h > \tau)|x_h] \\
&= \sum_h E[1(m(x_h;\beta) + \varepsilon_h > \log\tau)|x_h] \\
&= \sum_h Pr[\varepsilon_h > \log\tau - m(x_h;\beta)] \\
&= \sum_h (1 - F_\varepsilon(\log\tau - m(x_h;\beta);\sigma)),
\end{aligned}$$

where $1(a > b)$ denotes the indicator function that equals 1 if $a > b$ and 0 otherwise. In practice of course β and σ will have to be replaced with their respective estimators $\hat{\beta}$ and $\hat{\sigma}$. Similarly, an estimator for $n(\tau, y)$ can be obtained:

$$\begin{aligned}
\hat{n}(\tau,y) &= \sum_h E[1(\tau < Y_h \leq y)|x_h] \\
&= \sum_h E[1(m(x_h;\beta) + \varepsilon_h \leq \log y)|x_h] - E[1(m(x_h;\beta) + \varepsilon_h \leq \log\tau)|x_h] \\
&= \sum_h Pr[\varepsilon_h \leq \log y - m(x_h;\beta)] - Pr[\varepsilon_h \leq \log\tau - m(x_h;\beta)] \\
&= \sum_h F_\varepsilon(\log y - m(x_h;\beta);\sigma) - F_\varepsilon(\log\tau - m(x_h;\beta);\sigma).
\end{aligned}$$

Given $\hat{n}(\tau, y)$ and $\hat{n}(\tau)$, we may construct the estimators $\hat{F}_2(y) = \hat{n}(\tau, y)/\hat{n}(\tau)$ and $\hat{\lambda} = \hat{n}(\tau)/n$. Combined with the estimator for $F_1(y)$, which is estimated using DB-1 (i.e. the household income survey), we have all we need to estimate $F(y)$ (see eq. 1), the income distribution for the complete population. This in turn is all we need to compute any inequality measure of choice.

No assumptions have been made about the distribution of x_h at this point. Let us assume that the top end of the distribution of x_h can be described by a Pareto distribution.

Assumption 2 *Let $G_2(x)$ denote the distribution function of x conditional on $x > x_0$. It is assumed that $G_2(x)$ follows a Pareto distribution with shape parameters α:*

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

For ease of exposition let us also assume that the income threshold τ is set sufficiently high such that $Y > \tau$ implies $X > x_0$.

Assumption 3

$$Pr[Y \leq y|Y > \tau] = Pr[Y \leq y|Y > \tau,\ X > x_0].$$

It then follows that top incomes, exceeding the income threshold τ, too are Pareto distributed.

Proposition 4 *Given Assumptions 1, 2 and 3, $F_2(y)$ follows a Pareto distribution with shape parameter $\theta = \alpha/\beta_1$:*

$$F_2(y) = Pr[Y \leq y | Y > \tau] = 1 - \left(\frac{y}{\tau}\right)^{-\theta}. \tag{12}$$

Proof By Assumption 3 we have:

$$Pr[Y \leq y | Y > \tau] = Pr[Y \leq y | Y > \tau,\ X > x_0].$$

This is equivalent to:

$$\begin{aligned} Pr[Y \leq y | Y > \tau,\ X > x_0] &= \frac{Pr[\tau < Y \leq y | X > x_0]}{Pr[Y > \tau | X > x_0]} & (13) \\ &= \frac{Pr[Y \leq y | X > x_0] - Pr[Y \leq \tau | X > x_0]}{Pr[Y > \tau | X > x_0]}. & (14) \end{aligned}$$

Appealing to Assumptions 1 and 2, the term $Pr[Y \leq y | X > x_0]$ is seen to solve:

$$\begin{aligned} Pr[Y \leq y | X > x_0] &= Pr[\exp(\beta_0 + \varepsilon) X^{\beta_1} \leq y | X > x_0] \\ &= Pr[X \leq \exp(-\varepsilon/\beta_1) \left(\frac{y}{\exp(\beta_0)}\right)^{1/\beta_1} | X > x_0] \\ &= E_\varepsilon[G_2 \left(\exp(-\varepsilon/\beta_1) \left(\frac{y}{\exp(\beta_0)}\right)^{1/\beta_1}\right)] \\ &= E_\varepsilon[1 - \exp(\alpha\varepsilon/\beta_1) x_0^\alpha \left(\frac{y}{\exp(\beta_0)}\right)^{-\alpha/\beta_1}] \\ &= 1 - E_\varepsilon[\exp(\alpha\varepsilon/\beta_1)] x_0^\alpha \left(\frac{y}{\exp(\beta_0)}\right)^{-\alpha/\beta_1} \\ &= 1 - y_0^\theta y^{-\theta}, \end{aligned}$$

with $\theta = \alpha/\beta_1$ and $y_0 = M_\varepsilon^{1/\theta}(\theta) \exp(\beta_0) x_0^{\beta_1}$, where $M_\varepsilon(t)$ denotes the moment generating function of ε, i.e. $M_\varepsilon(t) = E[\exp(t\varepsilon)]$. By extension we have $Pr[Y \leq \tau | X > x_0] = 1 - y_0^\theta \tau^{-\theta}$.

Substituting the expressions for $Pr[Y \leq y | X > x_0]$ and $Pr[Y \leq \tau | X > x_0]$ into eq. (14) yields:

$$\begin{aligned} \frac{Pr[Y \leq y | X > x_0] - Pr[Y \leq \tau | X > x_0]}{Pr[Y > \tau | X > x_0]} &= \frac{1 - y_0^\theta y^{-\theta} - (1 - y_0^\theta \tau^{-\theta})}{1 - (1 - y_0^\theta \tau^{-\theta})} \\ &= 1 - \tau^\theta y^{-\theta}, \end{aligned}$$

which completes the proof. □

Note that θ controls the thickness of the top end of the income distribution, which is a key determinant of income inequality; the smaller the value of the tail index θ, the larger the proportion of high incomes, the higher the value of inequality. Under the assumption that top incomes are Pareto distributed, the mean top income level takes on the following form:

$$E[Y|Y > \tau] = \left(\frac{\theta}{\theta - 1}\right)\tau. \tag{15}$$

This mean top income level features in the computation of the top income shares as well as the computation of the between-inequality components.[8]

2.2.2 Population underlying top income database is unclear

Let us next address the challenge that emerges when the data underlying DB-2 are not necessarily representative of the whole population (i.e. households with incomes exceeding τ). Consider the possibility that DB-2 has "over-sampled" some and "under-sampled" other households among the top earners, such that DB-2 no longer yields a consistent estimator for $F_2(y)$ unless some corrective efforts are made. This is a rather realistic scenario as the data may constitute a series of transactions or listing prices rather than a proper sample drawn from the target population. For ease of exposition we will assume that DB-2 observes actual household incomes and not predictors of income, so that we may focus exclusively on the challenges presented in this section.

We will assume that the data are representative for selected sub-populations and that a representative "sample" can be obtained by anchoring DB-2 to some known population totals. Suppose that the target population can be sub-divided into D districts with $d = 1, \ldots, D$ indicating the district. The top income distribution for district d will be denoted by $F_{2,d}(y) = Pr[Y \leq y|Y > \tau, \textit{district}\, d]$. By extension, let $F_{1,d}(y) = Pr[Y \leq y|Y \leq \tau, \textit{district}\, d]$. Using this notation the complete income distribution for district d, denoted $F_d(y)$, satisfies:

$$F_d(y) = \begin{cases} (1 - \lambda_d)F_{1,d}(y) & y \leq \tau \\ (1 - \lambda_d) + \lambda_d F_{2,d}(y) & y > \tau, \end{cases} \tag{16}$$

where $\lambda_d = Pr[Y > \tau|\textit{district}\, d]$. The density functions corresponding to $F_{1,d}(y)$, $F_{2,d}(y)$ and $F_d(y)$ will be denoted by $f_{1,d}(y)$, $f_{2,d}(y)$ and $f_d(y)$, respectively.

[8] As an alternative to assuming a Pareto distribution for the top tail, and estimating the tail index parameter, one could also appeal to multiple imputation methods, see e.g. Douidich et al. (2015). This approach might in fact be more practical in case a more flexible functional form for $m(x_h; \beta)$ is being considered.

By definition the distribution of top incomes for the whole population solves:

$$F_2(y) = \sum_d F_{2,d}(y) P_{2,d}, \tag{17}$$

with $P_{2,d} = Pr[Y > \tau, \textit{district } d]$. These mixing probabilities permit the following decomposition:

$$P_{2,d} = \lambda_d \pi_d, \tag{18}$$

where π_d denotes the share of the total population (regardless of income) residing in district d. We make the following assumption.

Assumption 5 *It is assumed that:*

- *The data at hand permit consistent estimation of* $(F_{2,d}, f_{2,d})$ *and* $(F_{1,d}, f_{1,d})$ *for all* d.
- *The district population shares* $\{\pi_d\}$ *are known.*

That leaves $\lambda_d = Pr[Y > \tau | \textit{district } d]$ as the only unknown that needs to be estimated. One way to estimate λ_d is to impose the assumption that $f_d(y)$ is a continuous function.

Assumption 6 $f_d(y)$ *is a continuous function of* y.

Let $\hat{f}_{1,d}(\tau)$ and $\hat{f}_{2,d}(\tau)$ denote the estimators for $f_{1,d}(\tau)$ and $f_{2,d}(\tau)$, respectively. Assumption 5 ensures that these are consistent estimators. The following proposition derives an estimator for λ_d by appealing to Assumption 6.

Proposition 7 *Let* $\hat{f}_{k,d}(y)$ *denote a consistent estimator for* $f_{k,d}(y)$ *for* $k = 1, 2$. *Under Assumption 6,* $\hat{\lambda}_d$ *presented below provides a consistent estimator for* λ_d*:*

$$\hat{\lambda}_d = \frac{\hat{f}_{1,d}(\tau)}{\hat{f}_{1,d}(\tau) + \hat{f}_{2,d}(\tau)}. \tag{19}$$

Proof Evaluating the first-order derivative of $F_d(y)$ from eq. (16) with respect to y yields:

$$f_d(y) = \begin{cases} (1 - \lambda_d) f_{1,d}(y) & y \leq \tau \\ \lambda_d f_{2,d}(y) & y > \tau \end{cases} \tag{20}$$

By Assumption 6, $f_d(y)$ is continuous in y, which imposes that $(1 - \lambda_d) f_{1,d}(y) = \lambda_d f_{2,d}(y)$ for $y = \tau$. Rearranging the terms in this equality gives us the following solution for λ_d:

$$\lambda_d = \frac{f_{1,d}(\tau)}{f_{1,d}(\tau) + f_{2,d}(\tau)}. \tag{21}$$

The estimator for λ is obtained by replacing $f_{1,d}(\tau)$ and $f_{2,d}(\tau)$ with their estimators. Provided that all terms on the right-hand side of eq. (21) are consistently

estimated, which is guaranteed by Assumption 5, it follows that the estimator for λ_d will be consistent. □

Finally, note that the sub-group inequality decompositions presented in Section 2.1 can readily be extended to accommodate the sub-division of the top tail into D districts. (Note that the bottom segment can in principle stay as is, i.e. need not to be sub-divided into districts.) Let us denote the income share going to the top tail from district d by $s_d = P_{2,d}(\mu_{2,d}/\mu)$, where $\mu_{2,d} = E[Y|Y > \tau, \textit{district } d]$. Note that the population- and income shares corresponding to the bottom segment now solve $1-\sum_d \lambda_d$ and $1-\sum_d s_d$, respectively. Similarly, let us denote the Theil index or the mean-log-deviation for the top incomes from district d by $T_{2,d}$ and $MLD_{2,d}$, respectively. Using this notation, the decomposition of the Theil index and the mean-log-deviation into the $1+d$ sub-groups is seen to solve:

$$\begin{aligned} MLD &= (1-\Sigma_d\lambda_d)MLD_1 + \Sigma_d\lambda_d MLD_{2,d} + \log(\mu) - \log\left(\mu_1^{(1-\sum_d \lambda_d)}\Pi_d\mu_{2,d}^{\lambda_d}\right) \\ T &= (1-\Sigma_d s_d)T_1 + \Sigma_d s_d T_{2,d} + \log\left(\mu_1^{(1-\sum_d s_d)}\Pi_d\mu_{2,d}^{s_d}\right) - \log(\mu). \end{aligned}$$

3 Data

This paper uses two different types of data-sets: (1) Household Income, Expenditure and Consumption Survey (HIECS) data, and (2) listings of homes for sale derived from (large) real-estate databases. All data used in this study are for Egypt. The HIECS is from 2008/9. The house price data are slightly more recent, covering the period early 2013 to 2015, and come from two different real-estate firms. Details are given below.

3.1 Egyptian Household Income, Expenditure and Consumption Survey

The Egypt HIECS 2008/9 is conducted by the Central Agency for Public Mobilization and Statistics (CAPMAS). We were given a 50% sample of the survey (approximately 24,000 observations).[9] Throughout the paper, our welfare aggregate is expenditure per capita which is consistent with standard practice in most developing countries. Household expenditures have been adjusted for spatial differences in prices by deflating nominal values by a spatial price index following

[9]Hlasny and Verme (2013) were able to access the 100% sample on site at CAPMAS.

Belhaj Hassine (2015).[10]

Compared to income, consumption expenditure typically produces lower estimates of inequality, especially at the top. This can be explained by a declining marginal propensity to consume and by the fact that consumption surveys tend to understate the spending on durables at the top (e.g. Aguiar and Bils (2015) for the United States). For their study of top incomes in Egypt, Hlasny and Verme (2013) used income as their welfare measure. An argument for using consumption instead of income is that data on the former are often of a higher quality in developing and emerging economies and are less vulnerable to idiosyncratic noise as households tend to smooth their consumption over time. In what follows we will be abusing terminology by often referring to income inequality and the income distribution even though our data measure expenditures, not income.

As discussed in detail in Verme et al. (2014), inequality in Egypt as assessed from household surveys is low and has even declined in the decade before the 2011 revolution. The Gini coefficient of consumption expenditure declined by around 2pp from 32.8 in 2000 to 30.8 in 2009.[11] Our paper tests whether the low estimate in 2009 is robust to replacing the top tail of the income distribution with an estimate that is obtained using a combination of household expenditure- and house price data.

3.2 Real Estate Data

In late 2014/early 2015 we obtained data on houses and apartments for sale from two Egyptian real estate firms: Betak-online and Bezaat.[12] The two rank among the larger real-estate firms whose listing database can be accessed online; analogous to Redfin and Zillow in the United States. The data differ in detail but a listing typically consists of the asking price, the location (the city or a further subdivision), and the date when it was listed. Interviews with the Ministry of Housing in Cairo confirmed that the listing price provides a good approximation to the actual sales price.[13] We keep listings classified as houses, apartments, flats or villas, since these refer to private housing. There are a number of other types of listings which we exclude, the three largest groups being land, shop, and chalet.

The model that relates the value of the house to household expenditure (per capita) is estimated using the household survey data, which report (imputed) rents not property prices. We will be assuming that rent- and sale (or listing) prices are proportional to each other, which is sufficient for our needs.

[10]For a recent discussion of challenges with real consumption measurement, see e.g. Van Veelen (2002) and Van Veelen and Van der Weide (2008).

[11]Source: PovcalNet, accessed 31 October 2015.

[12]The URLs are respectively: www.betakonline.com; and www.bezaat.com.

[13]For our purposes it is sufficient that the actual price is proportional to the listing price.

The household survey is from 2009, while the rents derived from the real estate data refer to late 2013 - early 2015. There is no real need however to express the values in prices from the same year, i.e. to inflate the 2009 expenditures to 2014 prices or to deflate the house prices to 2009 prices. Instead we will be assuming that the Pareto tail index associated with the top tail of the income distribution is stable over the 2009-2014 period.

3.3 Does the household survey indeed omit the rich?

One way of illustrating whether the household data under-represent the top part of the distribution is to compare some of the characteristics of the top 1 percent the household survey with those of senior Egyptian executives. For the purpose of this exercise, household income is imputed from household expenditures in the survey using the average savings rate in Egypt for 2009.[14] The data on executive pay come from Payscale, an online information company providing current information on salary, benefits, and compensation by type of job, location, and other characteristics. The numbers are presented in Table 1.[15]

	% surveyed population	Minimum	Median	Maximum
Household income	Top 1%	11,995	14,666	98,080
CEO total pay	Top 1.2%	23,723	68,970	168,545
CFO total pay	Top 0.8%	22,551	54,563	212,393

Table 1: Annual income of top earners in Egypt (USD, nominal, 2009 prices)

We focus on the total compensation of senior executives, who represent 2 percent of survey participants and have the highest reported median compensation among survey participants.[16] Therefore, in principle, these households should be in the top 1% of households in Egypt's household survey. However, since the median senior executive income is closer to the maximum income than to the median income of the richest 1 percent in the household survey, and the maximum income earned by senior executives is much higher than the maximum income in

[14]We assume that household income reflects mainly the income of the household head and that the top households save at the average rate. The source for the average savings rate is the World Bank's World Development Indicators (WDI).

[15]Household income is imputed based on information on household expenditures in Egypt's 2009 household survey and the average saving rate in Egypt in 2009. The total pay of senior executives in Egypt is obtained from a global database of salaries and compensation for 2015. If anything we expect the data on executive pay to be on the conservative side. The values in the table are deflated and converted from EGP into USD using annual average inflation and exchange rate data from the World Bank's World Development Indicators.

[16]The senior executives surveyed by Payscale are either chief executive officers (CEOs) or chief financial officers (CFOs) in Egyptian firms.

the household survey, it appears that the household survey under-represents the top earning households, particularly the top earning senior executive households.

Similarly, in Vietnam the top salaries recorded in their household survey are less than half of average executive salaries obtained from corporate salary surveys (World Bank, 2014). In the case of Argentina, Alvaredo (2010) finds that while the tax data have almost 700 observations with incomes exceeding 1 million USD, there are none in the Argentine household survey. In a comparison of 16 Latin American household surveys, the ten richest households have incomes similar to a managerial wage, which is arguably substantially smaller than the incomes of top capital owners (Székely and Hilgert, 1999).

4 Empirical application

This section presents our empirical application to Egypt. As outlined in the methodology section we combine data on household expenditures with data on house prices. The household expenditures are obtained from the 2009/10 Egypt Household Income, Expenditure and Consumption Survey (HIECS), which is also used for Egypt's official estimates of poverty and inequality. The house prices represent listing prices for houses that have been put up for sale via two large real estate firms operating in Egypt. We use the real estate database to estimate the top end, defined as the top 5 percent, of the income distribution. The "bottom" 95 percent of the income distribution is estimated using the HIECS.

The following practical decisions and assumptions are made: (a) we restrict the analysis to urban Egypt only (this can be extended to apply to all of Egypt under the assumption that rural households do not rank in the top of the income distribution in Egypt), (b) it is assumed that house price quotes are proportional to (imputed) rental values (as the household expenditure survey contains data on rents only, and we rely on the survey to identify the relationship between house value and household income),[17] (c) it is assumed that the Pareto tail index of the income distribution has been stable between 2009/10 (the time of the survey) and 2013/14 (the time of the house price database), (d) it is assumed that one house constitutes one household (the fact that top income households could be associated with multiple houses may lead us to under-estimate inequality) and that all houses are domestically owned, and (e) we will only be using house price data for Cairo and Alexandria to estimate the top tails of their respective income distributions. For the rest of urban Egypt the entire income distribution will be estimated using the HIECS. The latter decision is motivated by the fact that: (i)

[17] Under a non-arbitrage condition, house prices and rental values are expected to move in parallel, see e.g. Himmelberg et al. (2005).

the lion-share of the "rich" that are missing or whose incomes are understated in the HIECS arguably reside in either Cairo or Alexandria, and (ii) the real estate markets are most developed in Cairo and Alexandria such that the coverage and the quality of the house price data are highest for these two cities (henceforward we will refer to these as districts).

Table 2 provides some basic statistics on the number of observations available to us. For the house price databases we only counted observations above the median house price value (which practically coincides with the mode of the house price density). Since we are interested in the top tail behavior of the house price distribution, we do not use the lower house price values.

	Database		
sub-group	Betak-online	Bezaat	HIECS
Cairo	5772	8475	1289
Alexandria	1293	2012	767
Urban Egypt			6935

Table 2: Number of observations used

The following sections proceed with the empirical application that combines the household expenditure survey and the house price data. A validation of our methodology in a controlled setting where only the survey data are used can be found in the Annex.

4.1 Pareto tail index estimated on income survey data

This subsection presents first estimates of the Pareto tail index of Cairo's and Alexandria's income distributions by using household survey data only. These estimates will serve as a reference point. Under the assumption of Pareto distributed top tails we have that: $1 - F_2(y) = \left(\frac{y}{\tau}\right)^{-\theta}$. Rearranging terms yields:

$$\log(y) = \log(\tau) - \frac{1}{\theta}\log(1 - F_2(y)). \tag{22}$$

If this assumption holds true, a plot of $\log(y)$ against $-\log(1 - F_2(y))$ should reveal a linear relationship with a slope parameter equal to $\frac{1}{\theta}$. Figure 1 provides this plot using the top 10 percent of the household expenditure data from the HIECS. For the majority of data points a linear relationship seems to provide a reasonable fit. A deviation from linearity can be observed however toward the far end of the income spectrum, where the slope appears to fall. Consequently, we should expect estimates of θ to come out higher if we were to increase the income threshold above which observations are included.

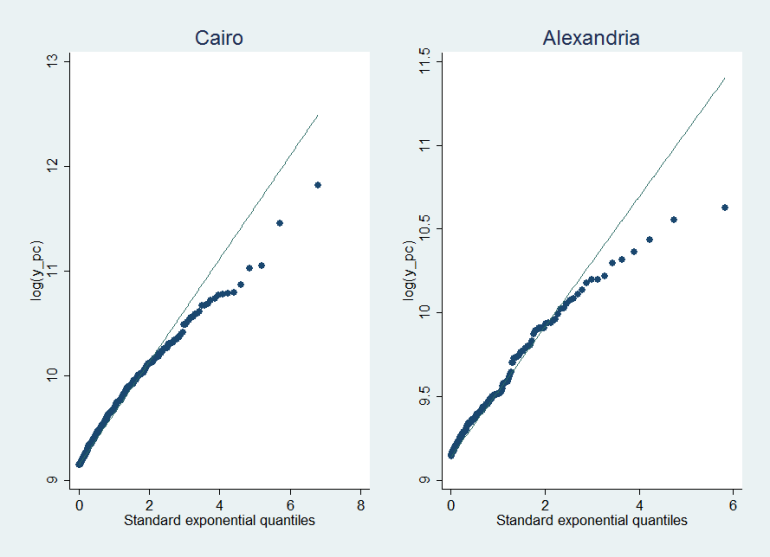

Figure 1: Pareto quantile plot for household expenditure per capita (household survey)

Figure 2 plots the maximum-likelihood (ML) estimates of θ for different values of the number of top observations used, ranging from the top 15 percent (85th percentile and up) to the top 5 percent of income observations (95th percentile and up). The grey area indicates the 95 percent confidence interval, which is seen to widen as the number of observations is reduced. It is also confirmed that for both Cairo and Alexandria the tail index is estimated to be higher at higher income thresholds (i.e. when the number of observations is reduced toward the top end), which is consistent with what we observed in Figure 1. The dotted line indicates the median level of the tail index (taken over all estimates within the plotted range) which roughly corresponds to the level where the estimates establish a plateau, most noticeably in the case of Alexandria. These will serve as our benchmark estimates of θ.

Observe that the HIECS estimates the top tail of the income distribution to be heavier (lower tail index) in Cairo than in Alexandria. Put differently, top income shares and income inequality are estimated to be highest in Cairo, which is arguably what one would expect. Relative ordering put aside, the question is whether the tail indices are being over-estimated, i.e. whether the thickness of the top tails are being under-estimated. The next sub-section will address this question by consulting data on house prices.

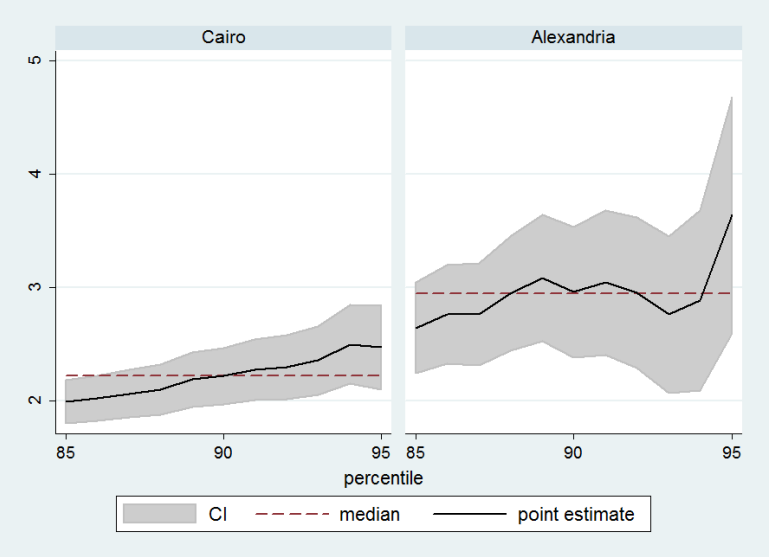

Figure 2: Pareto tail index estimates for household expenditure per capita (household survey)

4.2 Estimating the tail index using both income and house price data

We will go through the following steps in order to estimate the Pareto tail index θ by combining data on household expenditure from the HIECS with data on house prices. First we estimate the tail index associated with the top end of the house price distributions in Cairo and Alexandria, which we denoted α (see Assumption 2). Next we estimate the model from Assumption 1 that provides a link between house prices and household expenditures, where it is particularly parameter β_1 that we are interested in. With the estimators $\hat{\alpha}$ and $\hat{\beta}_1$ in hand, for Cairo and Alexandria separately, we apply Proposition 4 and obtain $\hat{\theta}_{mix} = \hat{\alpha}/\hat{\beta}_1$ as an alternative estimator for θ.

Figure 3 plots $\log(x)$ against $-\log(1 - G_2(x))$, analogous to Figure 1 but now using data on house prices (i.e. x denotes the listing price of a house). This plot uses the top 5 percent of above median value house prices from the respective house price databases (Betak-online and Bazaat). While a linear model appears to fit the data reasonably well, which supports the Pareto assumption, a deviation from linearity can be observed toward the top of the house price distribution. This non-linearity at the top is also observed for the household expenditure data from the HIECS (see Figure 1), albeit more pronounced for the house price data. The pattern is most noticeable for Cairo.

Figure 4 gives us an idea of the range of values α might attain by plotting estimates of the tail index as we vary the database and the number of top observations used for estimation. Note that this figure is analogous to Figure 2.

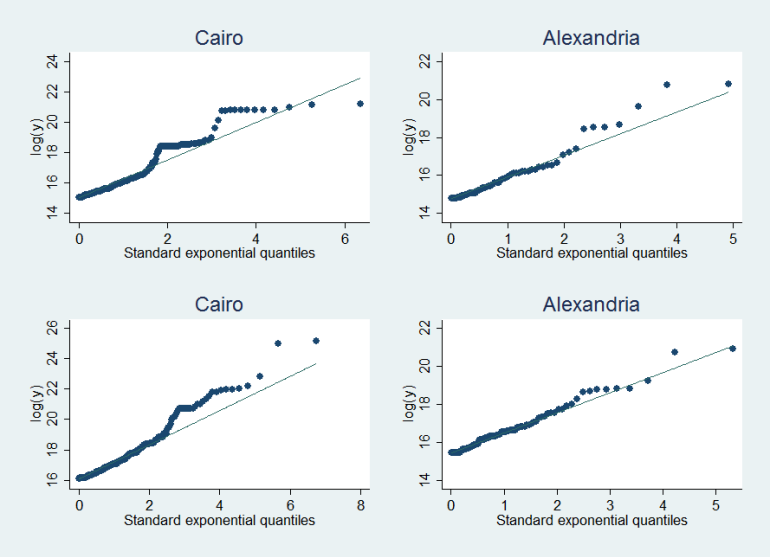

Figure 3: Pareto quantile plot for house prices (real-estate data): (a) Betak-online (top half), and (b) Bezaat (bottom half)

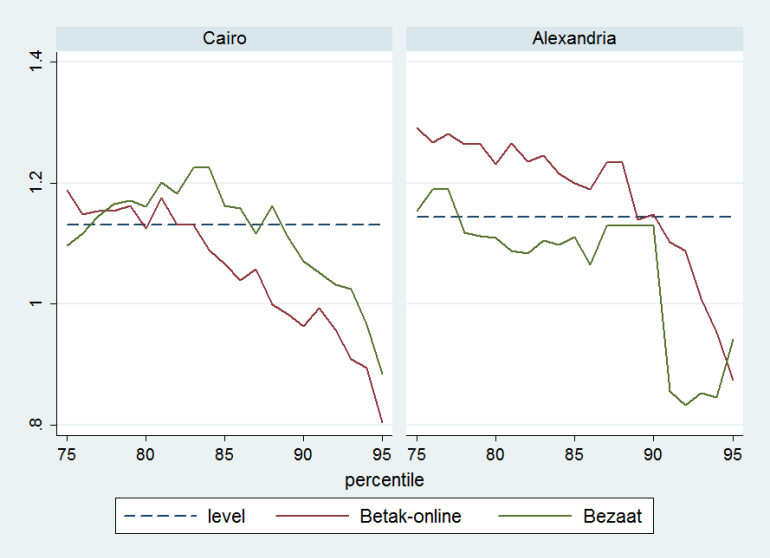

Figure 4: Pareto tail index estimates for house prices (real-estate data)

We omitted the confidence intervals in this case as they are small in comparison to the differences observed between the databases. The dotted line indicates our estimate of α; it is obtained as the median value of $\hat{\alpha}$ obtained over the two databases and between the percentiles 75 and 92 (i.e. between the top 25 and 8 percent). In the case of Alexandria the estimate roughly corresponds to a range where $\hat{\alpha}$ is found to level off. For Cairo it proved harder to find such a range. Our estimator is arguably on the conservative side in this case; our data appear to indicate that the tail index for Cairo is more likely to be lower than higher.

In other words, if anything, we may be slightly under-estimating the top income share (and hence inequality) for Cairo. Obviously, where we draw the line for $\hat{\alpha}$ is to a certain degree arbitrary. Toward the end of Section 4.3 we will briefly comment on how the range of α observed here may transalte into a range for θ and by implication a range for estimated levels of inequality.

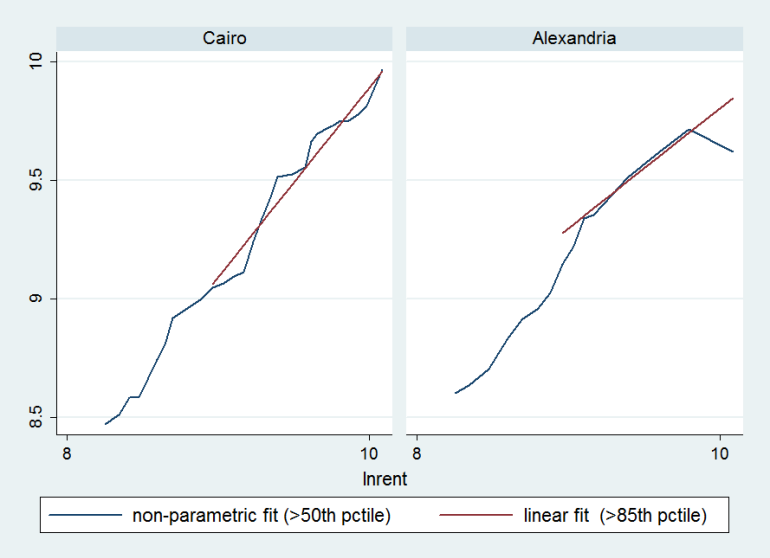

Figure 5: Household expenditure per capita versus imputed rent (log-log, household survey)

Next we need estimates of β_1. Here we fully rely on data from the HIECS. Before we imposed a functional form on $m(x)$, which describes the relationship between household expenditure per capita and the value of the houseold's house (captured by imputed rent), we first fitted a non-parametric kernel regression to the data (for Cairo and Alexandria separately). The results are presented in Figure 5. It is found that a linear model captures the relationship between log of household expenditure and log of (imputed) rent reasonably well, particularly in the case of Cairo. Alexandria shows a degree of concavity but also here a linear model arguably provides a good fit for high values of rent and household expenditure; see the fitted linear lines included in the figure.

Estimates of β_1 appear to be less sensitive to where we place the cut-off for the data included in the estimation when compared to estimates of α. See Figure 6 which investigates how $\hat{\beta}_1$ varies with the number of top observations included in the regression. The grey area indicates the 95 percent confidence interval. Notice how $\hat{\beta}_1$ is reasonably stable across the different cut-offs considered, which is consistent with the degree of linearity observed in Figure 5. The dotted lines denote the estimates that will be used in our analysis (see the values reported the first column of Table 3), which are obtained as the value of $\hat{\beta}_1$ for the top 10

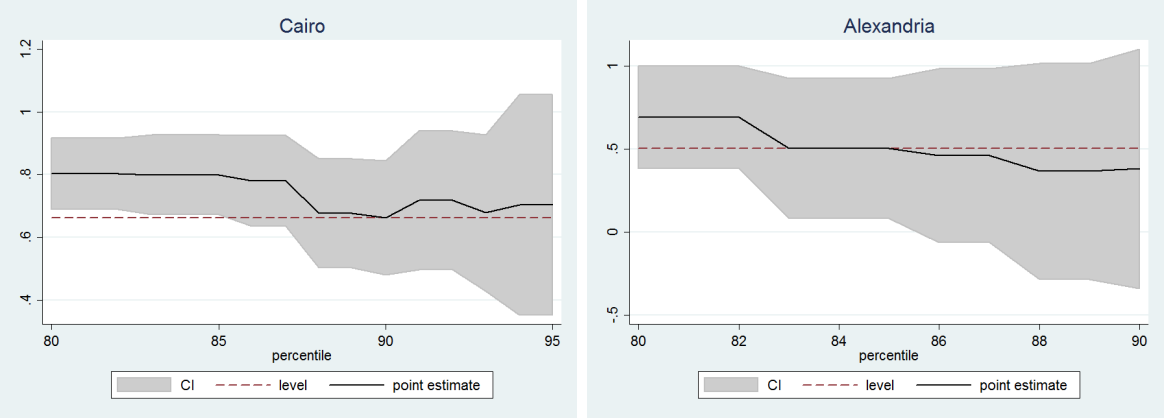

Figure 6: Estimates of β_1 estimated using increasingly smaller number of top observations (household survey)

percent (90th percentile) for Cairo and for the top 15 percent (85th percentile) for Alexandria.[18]

sub-group	$\hat{\beta}_1$	$\hat{\alpha}$	$\hat{\theta}_{mix}$	$\hat{\theta}_{svy}$	$\hat{\gamma}_{mix}$	$\hat{\gamma}_{svy}$
Cairo	0.662	1.131	1.708	2.216	2.412	1.822
Alexandria	0.505	1.144	2.267	2.958	1.789	1.511

Table 3: Estimates of β_1, α, θ and γ

What does this mean for θ? Our findings are summarized in Table 3, which shows the estimator $\hat{\theta}_{mix} = \hat{\alpha}/\hat{\beta}_1$ as well as the individual components $\hat{\alpha}$ and $\hat{\beta}_1$ that go into the estimator ($\hat{\gamma}$ denotes the inverted Pareto Lorenz coefficient, defined as $\hat{\gamma} = \frac{\hat{\theta}}{\hat{\theta}-1}$). For comparison we also include the estimator $\hat{\theta}_{svy}$ that is obtained using data from the HIECS only (see section 4.1). Two observations stand out. Firstly, the data on house prices give us reason to believe that the top tail of the income distribution is under-estimated in Egypt when relying on household survey data only, as is evidenced by the fact that $\hat{\theta}_{mix}$ is visibly smaller than $\hat{\theta}_{svy}$. Secondly, both the estimators $\hat{\theta}_{mix}$ and $\hat{\theta}_{svy}$ confirm that top income shares are largest in Cairo.

4.3 Main results: Re-estimating inequality for Egypt

Having new estimates of the Pareto tail indices for the respective income distributions of Cairo and Alexandria is not enough. To see what this means for total inequality for (urban) Egypt we also need estimates of the share of the population that resides in the respective metropolitan areas and enjoys incomes above τ, i.e.

[18]Notice that these estimates too are on the conservative side; lower values for $\hat{\beta}_1$ yield higher estimates of $\hat{\theta}_{mix}$ and hence lower estimates of inequality.

estimates of $Pr[Y > \tau,\ district\ d]$ for $d = Cairo,\ Alexandria$. We estimate these by: $Pr[Y > \tau,\ district\ d] = Pr[Y > \tau | district\ d]Pr[district\ d]$, where $Pr[district\ d]$ (the share of the urban population residing in district d) is obtained from the most recent population census and where $Pr[Y > \tau | district\ d]$ is estimated using Proposition 7. For comparison the latter is also estimated using data from the HIECS only. The two different estimators are denoted by $\hat{\lambda}_{prop7}$ and $\hat{\lambda}_{svy}$, respectively. $Pr[Y > \tau,\ district\ d]$ and $Pr[district\ d]$ are denoted by P and π, respectively, such that $\hat{P}_{prop7} = \pi\hat{\lambda}_{prop7}$ and $\hat{P}_{svy} = \pi\hat{\lambda}_{svy}$ (where we suppressed the subscript d for ease of notation). The estimates are presented in Table 4.

sub-group	π	$\hat{\lambda}_{prop7}$	$\hat{\lambda}_{svy}$	$\hat{P}_{prop7}$	$\hat{P}_{svy}$
Cairo	0.251	0.116	0.101	0.029	0.025
Alexandria	0.130	0.079	0.048	0.010	0.006
Other urban	0.619	0.028	0.028	0.017	0.017

Table 4: Estimates of π, λ and P

Notice that our estimate of λ finds that the percentage of households residing in Cairo and Alexandria with incomes exceeding τ is larger than what the HIECS alone would have us believe. This combined with the earlier observation that $\hat{\theta}_{mix} < \hat{\theta}_{svy}$ leads us to believe that relying on survey data alone will arguably under-estimate both the number of households with high incomes as well as the size of their incomes (either because top income earners are missing in the survey or because they under-report their incomes, or both). Table 5 compares estimates of top income shares S obtained using the HIECS to those obtained using both the HIECS and the house price data. The additional columns compare estimates of inequality among top income households (i.e. only including households whose income exceeds τ) for three different measures of inequality.

sub-group	S_{mix}	S_{svy}	$Gini_{mix}$	$Gini_{svy}$	MLD_{mix}	MLD_{svy}	$Theil_{mix}$	$Theil_{svy}$
Cairo	0.159	0.118	0.414	0.227	0.295	0.087	0.532	0.107
Alexandria	0.042	0.036	0.283	0.156	0.141	0.038	0.208	0.041
Other	0.063	0.066	0.223	0.223	0.082	0.082	0.097	0.097

Table 5: Estimates of S, $Gini$, MLD and $Theil$ (for the top tail)

Estimates of total inequality for (urban) Egypt are obtained by adding estimates of bottom- and between inequality to the estimates of top inequality reported in Table 5. Bottom inequality (i.e. inequality among households with income below τ) is estimated using the HIECS only. The between inequality component is estimated using data from both sources as it is a function of average income among top earners (which is a function of θ; see eq. 15) as well as a

function of λ (in the case of MLD) and of the top income share S (in the case of the Theil index), see equations (4) and (7). In the case of the Gini coefficient we implement the approximate decomposition that is also used by Alvaredo (2011): $Gini \approx (1 - \sum_d \lambda_d)(1 - \sum_d s_d)Gini_1 + \sum_d s_d$.

	Survey and House prices	Survey only
Gini	0.470	0.364
MLD	0.278	0.217
Theil	0.420	0.258

Table 6: Estimates of inequality for (urban) Egypt in 2009/10: Survey-only versus Survey+House prices

The total inequality estimates are presented in Table 6. The survey-only estimate of the Gini coefficient for (urban) Egypt in 2009/10 stands at 36.4. This is relatively low by international standards and hence would suggest that Egypt ranks among lower inequality countries. Our estimate of the Gini coefficient is 47.0 which is considerably higher than the official estimate. The level of top incomes recorded in the HIECS is found to be at odds with house prices observed toward the top end of the market in Cairo and Alexandria. Our estimates represent an attempt to correct for this. We repeated the analysis for other choices of inequality measures, specifically for the MLD and Theil measures. Noticeable increases in inequality can be observed for all measures considered. The magnitude of the adjustment is largest for the Theil index which is consistent with the fact that the Theil index is most sensitive to the top tail of the income distribution when compared to the other two choices of inequality measures.

The precision of our estimate of inequality is largely determined by the precision with which we are able to estimate α and β_1 (provided that the assumptions under which the estimators have been derived reasonably apply to the data at hand). It is instructive to verify what level of inequality would be obtained using rather conservative values for θ. Note that a most conservative estimate of θ can be obtained by combining a value of α from the top end of the estimated range with a value of β_1 from the low end of the estimated range (but taking $\beta_1 \geq \frac{1}{2}$ which rules out housing expenditure shares that are convex increasing functions of household expenditure; see the discussion following Assumption 1). For Cairo this gives us a value of around 2.4 (1.2/0.50; see Figures 4 and 6). For Alexandria we obtain a value that is just over 2.5 (1.25/0.5; see Figures 4 and 6). Note that these values are in the range of the respective survey-only estimates of θ (see Figure 2). In other words, it would take a very conservative estimate for $\hat{\theta}_{mix}$ to reproduce the survey-only estimate of inequality. The estimate we consider most reasonable finds a Gini coefficient for (urban) Egypt of 47.0, which is roughly

10 points higher than the survey-only estimate. Of course, by the same token, we may also be under-estimating inequality. Working with values of θ toward the lower end of our estimated range yields estimates of inequality that are noticeably higher than the Gini coefficient of 47.0. Using our estimates for θ (and λ) in a back-of-the-envelope calculation, we find that there are approximately 170 households in Cairo whose household expenditure exceeds 1 million USD per year. Although no other information on the number of millionaires in Egypt is currently available, this estimate seems rather conservative.

5 Concluding remarks

A growing literature has shown that household surveys provide only limited information about top incomes and therefore underestimate income inequality. This paper presents a method that corrects for this underestimation. We use the household survey for the bottom part of the distribution and combine it with another data source that provides a better coverage of the top tail. The existing literature has restricted itself to the use of tax record data to capture the top tail. Unfortunately income tax records are unavailable in many countries, including most of the developing world. Our method permits a much larger set of data for the top tail; the only requirements are that the data (i) contain a good predictor of household income, and (ii) provide a good coverage of the top tail.

We apply this method to Egypt, where estimates of inequality based on household surveys alone are low by international standards. Using publicly available data from real estate listings to estimate the top tail of the income distribution, we find strong evidence that inequality in Egypt is being underestimated. The Gini index for urban Egypt is found to increase from 36 to 47 after correcting for the missing top tail. A natural next step would be to use data on house prices to estimate the top tail of the wealth distribution, and extend the analysis to other countries.

References

Aguiar, M. and Bils, M. (2015). Has consumption inequality mirrored income inequality? *American Economic Review*, **105,** number 9, 2725–56.

Alvaredo, Facundo (2010). The rich in argentina over the twentieth century, 1932-2004. In *Top Incomes: A Global Perspective* (eds Anthony B. Atkinson and Thomas Piketty), pp. 253–298. Oxford University Press.

Alvaredo, Facundo (2011). A note on the relationship between top income shares and the gini coefficient. *Economics Letters*, **110,** number 3, 274–277.

Alvaredo, Facundo, Atkinson, Anthony B., Piketty, Thomas and Saez, Emmanuel (2015). The world top incomes database. http://topincomes.gmond.parisschoolofeconomics.eu/.

Alvaredo, Facundo and Londoño Vélez, Juliana (2013). High incomes and personal taxation in a developing economy: Colombia 1993-2013. Working Paper 12. Commitment to Equity-CEQ.

Alvaredo, Facundo and Piketty, Thomas (2014). Measuring top incomes and inequality in the middle east: Data limitations and illustration with the case of egypt. Working Paper 832. ERF.

Anand, Sudhir and Segal, Paul (2015). The global distribution of income. In *Handbook of Income Distribution* (eds Anthony B. Atkinson and Franois Bourguignon), volume 2A. Elsevier.

Atkinson, Anthony B. (2007). Measuring top incomes: Methodological issues. In *Top Incomes over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries* (eds Anthony B. Atkinson and Thomas Piketty). Oxford University Press.

Atkinson, Anthony B., Piketty, Thomas and Saez, Emmanuel (2011). Top incomes in the long run of history. *Journal of Economic Literature*, **49,** number 1, 3–71.

Belhaj Hassine, Nadia (2015). Economic inequality in the arab region. *World Development*, **66,** 532 – 556.

Burricand, Carine (2013). Transition from survey data to registers in the french silc survey. In *The Use of Registers in the Context of EU-SILC: Challenges and Opportunities* (eds Markus Jäntti, Veli-Matti Törmälehto and Eric Marlier). European Union.

Diaz-Bazan, Tania (2014). Measuring inequality from top to bottom. Working Paper.

Douidich, Mohamed, Ezzrari, Abdeljaouad, van der Weide, Roy and Verme, Paolo (2015). Estimating quarterly poverty rates using labor force surveys: a primer. *World Bank Economic Review*; Advance Access pulished 2015.

Himmelberg, Charles, Mayer, Christopher and Sinai, Todd (2005). Assessing high house prices: Bubbles, fundamentals and misperceptions. *Journal of Economic Perspectives*, **19,** number 4, 67–92.

Hlasny, Vladimir and Verme, Paolo (2013). Top incomes and the measurement of inequality in egypt. Policy Research Working Paper Series 6557. The World Bank.

Jäntti, Markus, Törmälehto, Veli-Matti and Marlier, Eric (2013). *The Use of Registers in the Context of EU-SILC: Challenges and Opportunities*. European Union.

Kim, Nak Nyeon and Kim, Jongil (2013). Sodok jipyo ui jaegumto [reexamining income distribution indices of korea] (in korean). *Journal of Korean Economic Analysis*, **19,** number 2, 1–57.

Korinek, Anton, Mistiaen, Johan A. and Ravallion, Martin (2006). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality*, **4,** number 1, 33–55.

Lakner, Christoph and Milanovic, Branko (2015). Global income distribution: From the fall of the berlin wall to the great recession. *World Bank Economic Review*; Advance Access published August 12, 2015.

Larsen, Erling (2014). The Engel curve of owner-occupied housing consumption. *Journal of Applied Economics*, **17,** number 2, 325–352.

Morelli, Salvatore, Smeeding, Timothy and Thompson, Jeffrey (2015). Post-1970 trends in within-country inequality and poverty: rich and middle-income countries. In *Handbook of Income Distribution* (eds Anthony B. Atkinson and Franois Bourguignon), volume 2A. Elsevier.

Székely, Miguel and Hilgert, Marianne (1999). What's behind the inequality we measure: An investigation using latin american data. *Research Department Working Paper Inter-American Development Bank*.

van Veelen, Matthijs (2002). An impossibility theorem concerning multilateral international comparison of volumes. *Econometrica*, **70,** number 1, 369–375.

van Veelen, Matthijs and van der Weide, Roy (2008). A note on different approaches to index number theory. *American Economic Review*, **98,** number 4, 1722–1730.

Verme, P., Milanovic, B., Al-Shawarby, S., Tawila, S. El, Gadallah, M. and A.El-Majeed, E. A. (2014). *Inside Inequality in the Arab Republic of Egypt: Facts and Perceptions across People, Time, and Space*. World Bank.

World Bank (2014). *Taking stock: an update on Vietnam's recent economic development*. World Bank.

A A small validation exercise: Re-estimating inequality in the survey after dropping top incomes

This section presents a modest validation of our approach in a controlled setting. Suppose that the survey provides a representative sample of the population. We will simulate a situation where top incomes are missing by dropping some of the top incomes from our sample. The data on imputed rents could then serve as our second data source (we drop only the household income observations but keep the imputed rent data for all households in the survey including the top income earners), i.e. they would stand in for the house price database. Dropping top income households from the sample will predictably lower the estimate of inequality. It would be encouraging if an application of our approach, where the top tail of the income distribution is estimated using the imputed rent data, will get us close to the original survey-direct estimate of inequality that is based on the full sample.

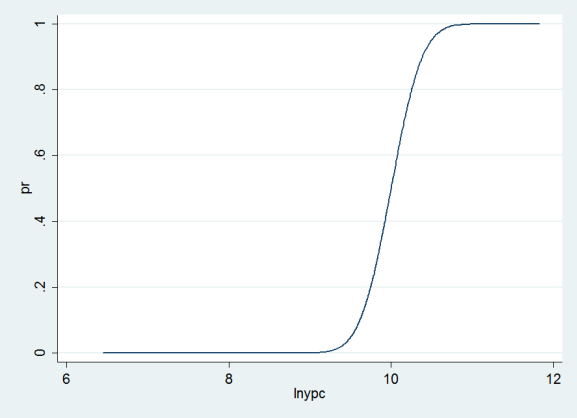

Figure 7: Simulated non-response probabilities as a function of household (log) expenditure per capita

Simply dropping the top $x\%$ of top incomes from the survey sample would leave us with a truncated distribution of household incomes which would be rather unrealistic. Instead we simulate non-response behavior by dropping households with a probability that increases with household income. This will preserve the smooth decay of the top tail of the income distribution, provided that the simulated non-response function does not abruptly jump from zero to one. Individual household observations are dropped at random with probability $Pr[non-response] = \Phi((\log(y_h) - 10)/0.3)$, where y_h denotes household income per capita and where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Figure 7 plots the corresponding probabilities for our sample. The

income distribution for the sub-sample (not reported here) does not show any signs of truncation. Less than 3 percent of the observations are dropped which is sufficient to produce a sizable decline in inequality (see Table 8): the estimated Gini drops from 36.4 (full sample) to 31.9 (sub-sample).[19]

Table 7 presents the estimates of β_1, α, θ, λ and P. Note that β_1 is estimated using the sub-sample of households, α is estimated using the imputed rent data for the full sample (which now stand in for the house price database), while θ, λ and P are estimated using a combination of the two data sources. Observe that the estimates of the Pareto tail index α obtained using the imputed rent data from the survey are markedly higher than the estimates obtained using the house price database (see Table 3 in the main text). This is consistent with the hypothesis that top income households are under-represented in the original survey. The estimates for β_1 are in the same range as those obtained using the full sample.

sub-group	$\hat{\beta}_1$	$\hat{\alpha}$	$\hat{\theta}_{mix}$	$\hat{\theta}_{svy}$	$\hat{\lambda}_{prop7}$	$\hat{P}_{prop7}$
Cairo	0.692	1.385	2.001	2.910	0.058	0.015
Alexandria	0.676	2.567	3.795	3.900	0.021	0.003

Table 7: Estimates of β_1, α, θ, λ and P

The corresponding estimates of total inequality for urban Egypt are shown in Table 8. While we do not perfectly reproduce the original estimates of inequality, we find the results encouraging. All measures of inequality considered show a clear upward adjustment. We slightly over-estimate the Gini, and slightly under-estimate the MLD and Theil indices. This may be due to the fact that we are using an approximation to the sub-group decomposition for the Gini (see Alvaredo, 2011), while the sub-group decompositions for the MLD and Theil indices are exact.[20]

	Survey only (full-sample)	Survey only (sub-sample)	Survey + imputed rents
Gini	0.364	0.319	0.387
MLD	0.217	0.163	0.201
Theil	0.258	0.175	0.253

Table 8: Estimates of inequality for (urban) Egypt in 2009/10: Survey-only versus Survey+imputed rents

[19]Note that we have not adjusted the sampling weights after dropping the selected households.
[20]There is no additive sub-group decomposition for the Gini that is exact.