Is Inequality Underestimated in Egypt? Evidence from House Prices

Elena Ianchovichina (World Bank)
Christoph Lakner (World Bank)
Roy van der Weide (World Bank)

Paper prepared for the IARIW-CAPMAS Special Conference "Experiences and Challenges in Measuring Income, Wealth, Poverty and Inequality in the Middle East and North Africa"

Cairo, Egypt
November 23-25, 2015

Session 5: Inequality I
Tuesday, November 24, 2015
08:30-10:00

Is Inequality Underestimated in Egypt? Evidence from House Prices

Elena Ianchovichina, Christoph Lakner and Roy van der Weide*

November 5, 2015

Abstract

Due to data constraints, most of the recent evidence on the rise of top income shares stems from developed countries. This paper presents a methodology for correcting the information in household surveys by imputing the top tail of the distribution of consumption expenditure. Our imputation uses data extracted from real estate listings which over-represent the top and, in contrast to the conventionally used tax record data, are in the public domain in most countries. We apply this methodology to Egypt, where the recent Arab Spring revolution brought issues of equity to the fore. However, according to the standard household survey analysis, consumption inequality in Egypt has been low and has even declined in the decade leading up to the revolution. Our imputation methodology does not change substantially this result a finding that challenges the notion that inequality of consumption was a factor behind the Egyptian revolution.

*All authors are with the World Bank. Contact information: eianchovichina@worldbank.org, clakner@worldbank.org, rvanderweide@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 Francisco Ferreira, Peter Lanjouw, Branko Milanovic, Martin Ravallion, Paolo Verme and participants of the World Bank workshop on the Arab Inequality Puzzle for useful comments. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors, and do not necessarily represent the views of the World Bank and its affiliated organizations.

1 Introduction

The recent literature on top income shares has argued that top incomes, and hence inequality, are underestimated in household surveys (Atkinson et al., 2011).[1] This literature suggests that administrative data, typically from tax records, provide a more accurate description of the top tail. These estimates of the top tail could be combined with estimates of the bottom part from the household survey to get closer to the true distribution (Alvaredo, 2011; Alvaredo and Londoo Vlez, 2013; Diaz-Bazan, 2014; Anand and Segal, 2015). However, the availability of tax record data, particularly in poor countries, is still limited. For example, the World Top Incomes Database (Alvaredo et al., 2015) includes no countries in the Middle East and North Africa region. Furthermore, data derived from tax records is less useful in places where tax evasion is more pervasive, as is the case in many developing countries, including some Arab states.

In the absence of fiscal data, this paper uses house prices to estimate the upper tail of the distribution of consumption expenditure. Market house price data are obtained relatively easily and, most importantly, are in the public domain, in contrast to tax administration data which raise important confidentiality concerns and require cooperation from governments. Furthermore, house sellers have no incentive to understate the value of their homes, in contrast to the income they report on their tax returns, and market price information is applicable to homes owned by the top end of the household consumption distribution.

We use this method to impute the top tail of the consumption distribution in Egypt. We find that it has a very small effect on the national estimates of inequality. The Gini coefficient remains practically unchanged at a low level, and although the Theil index increases by around a quarter from 21.9 to 26.3, it is still relatively low. Our results confirm earlier studies which showed a modest effect of adjusting for top incomes in Egypt using alternative methods (Hlasny and Verme, 2013). This contrasts with the evidence from a number Latin American countries that points to substantial inequality effects after correcting household surveys with tax record data (Alvaredo, 2011; Alvaredo and Londoo Vlez, 2013).

Because of our use of non-standard data for measuring the top tail, we need to bring two methodological innovations to the study of top incomes. First, we do not observe consumption or income (as is the case with tax record data), but need to impute this from the distribution of market house prices. We allow for errors in this consumption model using a multiple imputation framework. Second, our

[1] While the tax data have almost 700 observations with incomes exceeding 1 million USD, there are none in the Argentine household survey (Alvaredo, 2010). The ten richest households in a comparison of 16 Latin American household surveys have incomes similar toa managerial wage, which is probably substantially smaller than the incomes of top capital owners (Szkely and Hilgert, 1999).

house price data are not a nationally representative sample, but biased towards large urban centers. We account for this by making imputations separately by city and then weighting them accordingly.

In addition to being a major Arab country, Egypt provides a good testing ground for our method, as real estate data can be obtained from online sources. Furthermore, inequality in Egypt is of considerable interest not least as it has been cited as one of the factors behind the Egyptian revolution (Hlasny and Verme, 2013). Household surveys suggest that the inequality of consumption expenditure is low in Egypt and that it has declined in the last decade to a Gini of around 0.31 in 2009. Two other papers have discussed the issue of missing top incomes in Egypt. Hlasny and Verme (2013) find that the Gini does not increase substantially when they adjust the upper tail using a Pareto imputation.[2] However, their Pareto distribution is fitted using the household survey, which underestimates the top tail. While acknowledging the general problem, Alvaredo and Piketty (2014) argue that at present the data are insufficient to estimate top incomes in Egypt and that one would need reliable data from tax records.

This paper is related to a number of 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 US. Closely related to the data source in our paper, World Bank (forthcoming) uses data on mortgages and car loans in Indonesia to estimate the upper tail of the distribution which is combined with the household survey for the bottom tail. In their study on global interpersonal inequality, Anand and Segal (2015) append for every country the estimated top 1% share to the household survey distribution, which is assumed to represent the bottom 99%.[4]

Our method for correcting the underreporting of top incomes relies on combining the top tail of imputed consumption (obtained from the observed distribution of house prices) with the bottom distribution obtained from the household survey. A recent body of literature combines tax records and household surveys in a similar fashion. Building on Atkinson (2007), Alvaredo (2011) shows how the true Gini coefficient can be derived by combining the Gini coefficient of the top

[2]The Gini coefficient of consumption expenditure per capita in 2009 increases from 0.305 to 0.318, which is small although statistically significant.

[3]Recently, the EU-SILC survey in some countries began using register-based information (including tax records) for some questions (Jntti 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]For the majority of countries, this top 1% share is predicted from a cross-country regression using the top 10% share in the household survey. In an earlier attempt to correct global inequality estimates for missing top incomes, Lakner and Milanovic (2015) had relied on the gap between household surveys and national accounts together with a parametric top tail.

tail with the Gini coefficient for the bottom group from the household survey. When household survey and tax data are combined in this way, the Gini index in Colombia increases from 55 to 59 in 2010 (Alvaredo and Londoo Vlez, 2013). Diaz-Bazan (2014) generalizes this method by allowing for a more general choice of the cut-off level for joining up the distributions.

This paper is structured as follows. Section II describes the two main datasets used in the paper, namely the Egyptian Household Income, Expenditure and Consumption Survey (HIECS) 2008/2009, and our dataset of house prices which has been extracted from listings of homes for sale (and rent) on a number of Egyptian websites. Section III explains the methods we use to (1) impute consumption from the distribution of rents obtained from the house prices observed; and (2) combine the bottom and top estimates. Section IV presents the results and section V concludes.

2 Data

This paper uses two main datasets. First, we use the HIECS 2008/2009 to obtain the distribution of household consumption. We expect that this survey underrepresents the top, as discussed above. Hence we use this survey for the bottom part of the true distribution. The consumption model which imputes consumption expenditure into the distribution of market house prices is also based on this household survey. Second, from a number of Egyptian websites we extracted listings of homes for sale (and rent), as described in more detail below.

2.1 Household Survey: Egyptian Household Income, Expenditure and Consumption Survey

We use the Egypt HIECS 2008/2009 which is conducted by the Central Agency for Public Mobilization and Statistics (CAPMAS). We only have access to the 50% sample of the survey (approximately 24,000 observations).[5] Furthermore, at the time of writing, we did not have access to the latest 2010-11 round of the HIECS which would be closer to the real estate data. However, this later round of the survey was conducted during a revolution, so the data may be less stable (Hlasny and Verme, 2013).

Throughout the paper, our welfare aggregate is consumption expenditure per capita which is consistent with standard practice in most developing countries. Compared to income, consumption expenditure produces lower estimates of inequality, especially at the top. This can be explained by a declining marginal

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

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 USA). For their study of top incomes in Egypt, Hlasny and Verme (2013) used income as their welfare measure. However, it makes sense to use consumption expenditure in our paper because we link the house price data to the household survey data using the implied rents which are a component of consumption expenditure.

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 0.328 in 2000 to 0.308 in 2009.[6] Our paper tests whether these conclusions are robust to imputing top incomes based on external data on house prices and an innovative multiple imputation methodology.

2.2 Real Estate Data

In late 2014/early 2015, we obtained data on houses and apartments for sale (and rent) from a number of Egyptian real estate websites.[7] The websites differ in the exact details, but a listing typically consists of the asking price, the location (the city or a further subdivision), and the date when it was listed. This paper uses data extracted from Betakonline which was the largest database and also includes real estate for rent. In the future we plan on checking the robustness of our results to the use of data from the alternative websites.

We keep listings classified as apartments, flats or villas, since these clearly 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 website contains a variable which describes whether the listing is for rent or sale, which we use to generate the price-to-rent ratios (see below). We correct for any misclassifications in this variable by using the median sales prices for the city: We exclude records listed for rent (sale), but whose price is more than 10% (less than 5%) of the median sales price in the city. Records with a missing rent/sale variable are reclassified depending on whether the listed price is more or less than 10% of the median sales price in the city.

The consumption model is based on the household survey data, which reports rents not property prices. Therefore, we need to convert the asking prices observed in the real estate data into rents.[8] We calculate city-level price-to-rental ratios

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

[7]These websites include 2olli, Aqar-estate, Betakonline, Bezaat and Dubizzle.

[8]Himmelberg et al. (2005) describe a method for calculating the annual user cost of housing, and thus the annual rent, under a non-arbitrage condition. Because of widespread distortions in the Egypt real estate market, the non-arbitrage condition might not hold. Therefore, we use observed

from the median values observed in Betakonline, thereby also allowing for spatial variation. For cities for which we do not observe rents and sales, we use the national ratio. The national ratio is around 5, ranging from 4.5 to 6.2 across the cities included in the database.

The household survey is from 2009, while the rents derived from the real estate data refer to the end of 2014/early 2015. Therefore, we need to convert the rent prices back to 2009 values, which is not straightforward, especially given Egypts tumultuous history during this time. We have decided to use the real-estate price index by Aqarmap. Unfortunately, this index reaches back only as far as January 2011. We have used the national CPI to extend the index back to 2009.

3 Methodology

3.1 Combining income survey with top income database

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 no data 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 unbiased estimates of inequality. DB-1 commonly represents a household income survey. For DB-2 researchers often look at tax record data, as is discussed in the introduction.

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 $F_1(y)$.) It is also assumed that DB-2 contains the total number of units (i.e. households) whose income is above τ. Combined with the total population this yields an estimator for λ. 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:

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

price-to-rental ratios.

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 $S_1 = 1 - \alpha$ and by extension $S_2 = \alpha$. It can be verified that income inequality as measured by the Gini coefficient G satisfies the following decomposition (see e.g. Alvaredo, 2011):

$$\begin{aligned} G &= P_1S_1G_1 + P_2S_2G_2 + S_2 - P_2 && (2)\\ &= (1-\lambda)(1-\alpha)G_1 + \lambda\alpha G_2 + \alpha - \lambda, && (3) \end{aligned}$$

where G_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):

$$\begin{aligned} MLD &= P_1MLD_1 + P_2MLD_2 + P_1\log\left(\frac{P_1}{S_1}\right) + P_2\log\left(\frac{P_2}{S_2}\right) && (4)\\ &= (1-\lambda)MLD_1 + \lambda MLD_2 + (1-\lambda)\log\left(\frac{\mu}{\mu_1}\right) + \lambda\log\left(\frac{\mu}{\mu_2}\right) && (5)\\ &= (1-\lambda)MLD_1 + \lambda MLD_2 + \log(\mu) - \log\left(\mu_1^{1-\lambda}\mu_2^{\lambda}\right), && (6) \end{aligned}$$

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

$$\begin{aligned} T &= S_1T_1 + S_2T_2 + S_1\log\left(\frac{S_1}{P_1}\right) + S_2\log\left(\frac{S_2}{P_2}\right) && (7)\\ &= (1-\alpha)T_1 + \alpha T_2 + (1-\alpha)\log\left(\frac{\mu_1}{\mu}\right) + \alpha\log\left(\frac{\mu_2}{\mu}\right) && (8)\\ &= (1-\alpha)T_1 + \alpha T_2 + \log\left(\mu_1^{1-\alpha}\mu_2^{\alpha}\right) - \log(\mu), && (9) \end{aligned}$$

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.[9] 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 ($\alpha > 0$). It can readily be seen that the between-group inequality component of the Gini coefficient tends to $\alpha > 0$ in that case, while the within-group inequality among top income earners tends to zero, i.e. $G \to (1-\alpha)G_1 + \alpha$. It follows that the between-group inequality component for the mean-log-deviation tends to $\log(1-\alpha)^{-1}$, 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. $MLD \to MLD_1 - \log(1-\alpha)$. 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 $\alpha > 0$).

3.2 Replacing the top income database: Challenges

As stated in the previous section, DB-2 (the top income database) typically takes the form of tax record data. This data has 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 a straightforward estimation of λ. A key disadvantage of tax record data is that it is often difficult to obtain access to. Moreover, it is more likely to be available in developed countries with good quality data systems in place, and by extension less likely to be available in developing countries.

This paper explores the use of alternatives for the top income database DB-2 (i.e. alternatives to tax record data) that are more readily available. We can think of a number of such alternatives. The empirical application presented in Section 4 works with a house price database that has been compiled from publicly available real estate property listings. But one could also look to data on mortgages or the consumption of selected luxury goods. 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 that complicate the estimation of income inequality (and the income distribution), namely they: (a) observe predictors of income, not actual incomes, and (b) do not constitute a proper sample, so that it is unclear what population is being represented by the

[9] 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.

data.

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

3.2.1 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 can be described by:*

$$\log(Y_h) = m(x_h;\beta) + \varepsilon_h, \tag{10}$$

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, such that $E[\log(y_h)|x_h] = m(x_h;\beta)$.

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:

$$F_2(y) = \frac{n(\tau, y)}{n(\tau)} \tag{11}$$

$$\lambda = \frac{n(\tau)}{n}. \tag{12}$$

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:

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

Being the top income database, DB-2 will generally not cover the total population but the top tail of the income distribution instead. But when DB-2 does not in fact contain any data on household incomes it cannot be verified whether all relevant households with incomes exceeding the given income threshold τ are being represented. We observe $m(x_h;\beta)$ which gives us the probability that the household is part of the top tail. In practice DB-2 may only cover households with $m(x_h;\beta) > \theta$ for some (income predictor) threshold θ that will serve as an analogue to the actual income threshold τ. In order to obtain consistency of the estimators for $n(\tau, y)$ and $n(\tau)$ in that case, we will have to assume that relevant households are excluded with probability zero, i.e. $Pr[Y > \tau | m(x_h;\beta) \leq \theta] = 0$.

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.

Alternatively, the top tail of the income distribution and by extension selected measures of income inequality can be estimated by appealing to multiple imputations (see e.g. Rubin). To illustrate this approach let us focus on the estimation of a given measure of inequality, say the Theil index, where we will also be taking

advantage of the sub-group decomposition for this inequality measure.

This approach will simulate R random income values for all households included in DB-2 using the data-generating-process (DGP) from Assumption 1. Each simulation $r = 1, \ldots, R$ executes the following steps:

1. Draw the model parameters $\tilde{\beta}^{(r)}$ and $\tilde{\sigma}^{(r)}$, either from the corresponding estimated asymptotic distributions or by bootstrapping the database that is used to estimate these parameters
2. Draw $\tilde{\varepsilon}_h^{(r)}$ for all h from $F_\varepsilon(e; \tilde{\sigma}^{(r)})$
3. Compute $\tilde{Y}_h^{(r)} = \exp\left(m(x_h; \tilde{\beta}^{(r)}) + \tilde{\varepsilon}_h^{(r)}\right)$ for all h

This gives us R replications of DB-2 complete with (simulated) income data for all households included. Next we compute the relevant statistics for each of these simulations by going through the following steps:

4. Compute $\tilde{n}^{(r)}(\tau) = \sum_h 1\left(\tilde{Y}_h^{(r)} > \tau\right)$, and $\tilde{\lambda}^{(r)} = \tilde{n}^{(r)}(\tau)/n$
5. Compute $\tilde{\mu}_2^{(r)} = \frac{1}{\tilde{n}^{(r)}(\tau)} \sum_h \tilde{Y}_h^{(r)}$, and $\tilde{\mu}^{(r)} = (1 - \tilde{\lambda}^{(r)})\hat{\mu}_1 + \tilde{\lambda}^{(r)}\tilde{\mu}_2^{(r)}$, where $\hat{\mu}_1$ is estimated using DB-1
6. Compute $\tilde{\alpha}^{(r)} = \tilde{\lambda}^{(r)}\tilde{\mu}_2^{(r)}/\tilde{\mu}^{(r)}$
7. Compute $\tilde{T}_2^{(r)}$ using only simulated incomes that satisfy $\tilde{Y}_h^{(r)} > \tau$ (or another inequality measure of choice)
8. Finally, draw $\tilde{T}_1^{(r)}$ either from its estimated distribution (estimated using DB-1) or by re-estimating T_1 to a bootstrapped version of DB-1

When all steps have been completed we have R simulated values of λ, α, T_2, and T_1. Inserting these into to the sub-group decomposition for the Theil index (see eq. 9), yields R simulated values of the Theil index for the complete population. Let us denote these simulated values by $\tilde{T}^{(r)}$ for $r = 1, \ldots, R$. The average $\tilde{T}^{(r)}$ gives us a point estimate of the Theil index, while the corresponding standard error can be obtained by computing the standard deviation of $\tilde{T}^{(r)}$ over the R replicated values. (Note that standard errors are more easily obtained using this multiple imputation approach as it circumvents the need to derive the variance of the various estimators analytically.)

3.2.2 Population underlying top income database is unclear

Let us next address the challenge that emerges when the data underlying DB-2 is not representative of the target population (i.e. households with incomes exceeding τ). We will consider two scenarios corresponding to different degrees in which the data is not representative. For ease of exposition we will abstract away here from the challenges presented in the previous section, 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.

In the first scenario the only limitation of DB-2 is that it does not provide an accurate count of the number of households with top incomes. We will assume here that the estimation of the top tail of the income distribution is not compromised.

Assumption 2 *Let $f_k(y)$ denote the probability density function corresponding to $F_k(y)$, i.e. the first-order derivative of $F_k(y)$ with respect to y, for $k = 1, 2$. It is assumed that both $f_k(y)$ and $F_k(y)$ can be consistently estimated using DB-k.*

This means that the only innovation needed at this stage is to find another way of estimating λ. One way to identify λ is to impose the assumption that the probability density function $f(y)$ for the complete population, the first derivative of $F(y)$, is a continuous function of y.

Assumption 3 *Let $f(y)$ denote the probability density function corresponding to $F(y)$, i.e. the first-order derivative of $F(y)$ with respect to y. It is assumed that $f(y)$ is a continuous function of y, specifically around the threshold $y = \tau$.*

The proposition below derives an estimator for λ by appealing to Assumptions 2 and 3.

Proposition 4 *Let $\hat{f}_k(y)$ denote a consistent estimator for $f_k(y)$ for $k = 1, 2$. Under Assumptions 2 and 3, $\hat{\lambda}$ presented below provides a consistent estimator for λ:*

$$\hat{\lambda} = \frac{\hat{f}_1(\tau)}{\hat{f}_1(\tau) + \hat{f}_2(\tau)}. \tag{13}$$

Proof Evaluating the first-order derivative of $F(y)$ from eq. (1) with respect to y yields:

$$f(y) = \begin{cases} (1-\lambda) f_1(y) & y \leq \tau \\ \lambda f_2(y) & y > \tau \end{cases} \tag{14}$$

By Assumption 3, $f(y)$ is continuous in y, which imposes that $(1-\lambda) f_1(y) = \lambda f_2(y)$ for $y = \tau$. Rearranging the terms in this equality gives us the following solution for λ:

$$\lambda = \frac{f_1(\tau)}{f_1(\tau) + f_2(\tau)}. \tag{15}$$

The estimator for λ is obtained by replacing $f_1(\tau)$ and $f_2(\tau)$ with estimators. Provided that all terms on the right had side of eq. (15) are consistently estimated, which is guaranteed by Assumption 2, it follows that the estimator for λ will be consistent. □

Note that the same estimator for λ may be adopted when the top income database

does provide a count for the number of top income households, as is the case in Piketty/Alavaredo.

The second scenario adds a second limitation; 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, say, and not a proper sample drawn from the target population.

We will assume that a representative "sample" can be obtained by anchoring DB-2 to some known population totals. Specifically, we make the following assumptions.

Assumption 5 *The target population can be divided into D districts, with $d = 1, \ldots, D$ indicating the district. It is assumed that:*

- *The share of the total population residing in each of the districts is known, which will be denoted by $\{\pi_d\}$.*
- *DB-2 provides a representative "sample" for any given district d, such that it permits a consistent estimation of $F_{2,d}(y) = Pr[Y \leq y | Y > \tau, \textit{district}\, d]$.*

Under Assumption 5, a consistent estimator for $F_2(y)$ can be obtained by appealing to the following identity:

$$F_2(y) = \sum_d F_{2,d}(y) P_{2,d}, \tag{16}$$

where $P_{2,d} = Pr[Y > \tau, \textit{district}\, d]$. The task at hand is to find a way to estimate $P_{2,d}$ for $d = 1, \ldots, D$. The following decomposition for $P_{2,d}$ takes advantage of the known population shares π_d:

$$\begin{aligned} P_{2,d} &= Pr[Y > \tau | \textit{district}\, d]\, \pi_d & (17) \\ &= \lambda_d \pi_d. & (18) \end{aligned}$$

That leaves $\lambda_d = Pr[Y > \tau | \textit{district}\, d]$ as the only unknown that needs to be estimated. Note that λ_d is the equivalent of λ if district d would be the target population. Hence, one way to estimate λ_d is to apply Proposition 4 where $\hat{f}_1(\tau)$ and $\hat{f}_2(\tau)$ are replaced by the district d analogues which we shall denote by $\hat{f}_{1,d}(\tau)$ and $\hat{f}_{2,d}(\tau)$. Assumption 5 ensures that DB-2 yields a consistent estimator for $F_{2,d}(y)$ and by extension $f_{2,d}(y)$. In addition to this we will have to assume that DB-1 gives us consistent estimators for $F_{1,d}(y)$ and $f_{1,d}(y)$.

Finally, note that the sub-group inequality decompositions presented in Section 3.1 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 $\alpha_d = P_{2,d}(\mu_{2,d}/\mu)$, 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 $1-\sum_d \alpha_d$, respectively. Similarly, let us denote the Theil index or the mean-log-deviation, say, 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:

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

4 Empirical application

4.1 Overview of data

To be completed

4.2 Main results

This section presents the main findings of our empirical application to Egypt. As stated earlier we combine household expenditure data with a database on house prices for Egypt. We use the latter database to estimate the top tail of the income distribution. The "bottom" part of the income distribution is estimated using the household survey data. The following practical decisions and assumptions were made: (a) we use data for urban Egypt only (the analysis 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) all house prices are converted to rental values as the household expenditure survey contains data on rents only (the survey data is used to identify the relationship between housing value (i.e. rental value) and household income; this relationship is then used to impute household incomes into the house price database), and (c) 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).

Table 1 presents the inequality estimates we obtained for urban Egypt in 2009. The first column in the table shows the survey direct estimates (where the entire income distribution is estimated using the survey data, i.e. these estimates discard the house price database). The estimates that are obtained by combining the survey and the house price database (using multiple imputations) can be found

in the second column. For the Gini coefficient and the mean-log-deviation (MLD) it is found that the two different approaches yield remarkably similar estimates of the level of inequality. Among the three different inequality measures considered here the Theil index is seen to stand out as the only measure for which we observe an increase in the level of inequality when the top tail of the income distribution is estimated using the house price database. This 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.

	Survey direct	Imputed
Gini	33.7	33.6
MLD	18.7	19.4
Theil	21.9	26.3

Table 1: Estimates of inequality for (urban) Egypt: Survey direct- versus Multiple Imputation estimates

All things considered, the observed changes in the estimates of inequality obtained by correcting the top tail of the income distribution are underwelming. We are tempted to conclude that Egypt does not rank as a high inequality country. Further work is underway to test the robustness of this finding.

5 Concluding remarks

This paper has presented a method for imputing the top tail of an income or consumption distribution, and how this can be used to correct for the underreporting of top incomes in household surveys. The objective of this paper is to offer a method that could be used in settings where tax record data are not available, or of low quality. At present, the coverage of tax record data is limited especially in developing countries. In our application, the data underlying the imputation of the top tail was derived from real estate listings in Egypt, but in principle this method could be applied to other data which over-represents the top tail.

We find no evidence that the Egypt household survey underestimates inequality by underestimating the top tail. This is consistent with other evidence on Egypt, but contradicts studies on many other countries. Furthermore, it challenges the notion that inequality was one of the driving forces behind the Egypt revolution of 2011. However, it is important to bear in mind that the data we have used to approximate the top tail could produce a lower bound on the true inequality, e.g. if many households own multiple houses or if they purchase houses through alternative channels.

The main motivation for this paper was to measure top incomes accurately,

which has important implications for public policy, such as fiscal policy. Furthermore, our application using real estate data has a direct policy implication. Our analysis uses the idea that property wealth has some explanatory power for the top tail of the consumption distribution. This is a justification for making more extensive use of property taxation especially in settings where top incomes are hard to tax, e.g. due to evasion or the presence of income sources which are hard to tax such as capital incomes. While implementing property taxes is not without difficulties, it has been recognised that they could be a first step towards a more progressive taxation of wealth and income especially in developing countries (Norregaard, 2015).

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 Londoo Vlez, 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.

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 Jntti, Veli-Matti Trmlehto and Eric Marlier). European Union.

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

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

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

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.

Norregaard, J. (2015). Taxing immovable property: Revenue potential and implementation challenges. In *Inequality and fiscal policy* (eds B. Clements, R. de Mooij, S. Gupta and M. Keen). International Monetary Fund.

Szkely, 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*.

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 (forthcoming). Estimating the upper end of indonesias income distribution. forthcoming.